“INTERDISCIPLINARY DATA SCIENCE METHODS USING MACHINE LEARNING FOR ENHANCED KNOWLEDGE ACQUISITION”

Paraskevas Koukaras

Supervisor: Dr. Christos Tjortjis (Associate Professor, Department Head and Dean of School of Science & Technology, International Hellenic University)

Members: Dr. Panagiotis Tsaparas (Associate Professor, Department of Computer Science and Engineering, University of Ioannina) and Dr. Spiros Denaxas (Professor, Institute of Health Informatics, University College London)

Submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy

School of Science and Technology
September, 2021
Thessaloniki - Greece

Dedications

To my beloved daughter and wife from whom I deprived family time and my respected parents and brother for their eternal, unconditional and ceaseless support…

Keywords

Association Rule Mining; COVID-19; Clustering; Data Analytics; Data Mining; Data Preprocessing; Energy Balancing; Energy Flexibility; Energy Load Forecasting; Ensemble Methods; Machine Learning; Microgrids; Optimal Energy Scheduling; Portfolio Optimization; Power Management; Sentiment Analysis; Short Term Prediction; Social Media; Supervised/Unsupervised Learning; Text Polarization; Timeseries; Topic Extraction;

Abstract

This thesis reports on a series of novel approaches enabling knowledge acquisition through exploiting various machine learning capabilities. Interdisciplinary approaches may expose new possibilities for data analytics. Two theoretical frameworks are conceptualized reporting on findings that relate to the research domains of Social Media (SM) and Energy. Common methods/algorithms/tools may be utilized for knowledge extraction considering specific mining tasks.

The first theoretical framework presents and combines three novel approaches in Social Media domain elaborating in mining tasks related with Social Media Types (SMTs), Social Media Topic Extraction (SMTE) and Social Media Sentiment Analysis (SMSA). SMTs are evaluated through a novel hypothesis-based data driven methodology that analyses Social Media Platforms (SMPs) and categorizes SMPs based on their services proposing new SMTs. The proposed methodology evaluates a new taxonomy, based on a mixture of hypothesis and data driven approach utilizing association rules and clustering algorithms. As a result, three new SMTs emerge, namely Social, Entertainment and Profiling networks, that update and capture emerging SMP services.

Regarding SMTE, this study utilizes Twitter data to mine association rules and extract knowledge about public attitudes. COVID-19 pandemic acts as the use case, analysing crawled tweets. The approach incorporates topic extraction and visualization techniques, to form word clusters that infer to themes of opinions. Association rule mining is utilized to improve the process of extracted topics, producing more accurate and generic results. For the examined period, out of 50 initially retrieved topics with common SMTE methods, the proposed novel approach manages to reduce topics to just a few ones.

SMSA relates to the identification and analysis of sentiment polarity in microblogging data. Such a mining task enables new possibilities for knowledge extraction and evaluation of public sentiment in response to global events, producing valuable insights. COVID-19 is the use case, gathering data from Twitter. The main objective in this topic is the evaluation of a possible correlation between public sentiment and the number of cases and deaths attributed to COVID-19. Findings

correlate sentiment polarity with announced deaths, starting 41 days and expanding up to three days prior to the count. Also a strong correlation is identified, between COVID-19 Twitter conversation polarity and reported cases, but a weak correlation between polarity and reported deaths.

The second theoretical framework presents and combines three novel approaches in Energy domain elaborating in mining tasks related with Energy Balancing (EB), Energy Load Forecasting (ELF) and Energy Optimal Day-Ahead Scheduling (EODS). Energy management may be improved by performing EB in both Peer-to-Peer (P2P) and Virtual Microgrid-to-Virtual Microgrid (VMG2VMG) level. This task yields an interdisciplinary analytics-based approach for the formation of VMGs achieving EB. Computer Science methods are incorporated for addressing an Energy sector problem, utilizing data preprocessing techniques and Machine Learning concepts. Each prosumer is perceived as a peer, while VMGs are perceived as clusters of peers. This approach incorporates clustering and binning algorithms for preprocessing Energy data (for 94 prosumers) producing options for generating VMGs. Then, a customized Exhaustive brute-force Balancing Algorithm (EBA) balances at the cluster-to-cluster level (VMG2VMG balancing) reporting outcomes and prospects for scaling up and expanding this work.

A novel approach in the task of ELF exposes improvements for residential house energy requirements. This task is crucial for Energy sector stakeholders (e.g., DSO, aggregators etc.) since they are able to plan in more efficient manner their Demand Response (DR) management strategies. The experimentation includes the retrieval of energy readings from a state-of-the-art nearly Zero Energy Building (nZEB). Focus is made on one step ahead ELF, producing an approach regardless the time resolution of available data while yielding high accuracy results. Ensemble methods and forecasting algorithms are utilized while the evaluation of forecasting results is performed with popular accuracy metrics (MAPE, SMAPE and RMSE) and an Execution Time (ET) metric.

Optimal energy management relates with the task of EODS. A novel approach is proposed in the form of a framework/tool for a multi-objective analysis comprising a decision-making system. Two distinct optimization problems for two actors (consumers and aggregators) are considered, with each solution completely or partly interacting with the other in the form of DR signal exchange. The overall optimization

is formulated by a bi-objective optimization problem for the consumer's side aiming at cost minimization and discomfort reduction; and a single objective optimization problem for the aggregator's side aiming also at cost minimization. Experimentation is conducted on a real pilot (Terni Distribution System portfolio). The framework performs decision making by forecasting the day-ahead energy management requirements while aiming at optimal management of energy resources considering both aggregator's and consumer's preferences and goals.

Achievements of this thesis highlight prospects for enhanced knowledge acquisition through the conception of two theoretical frameworks in the domains of Social Media and Energy while envisioning an interdisciplinary research design. The theoretical frameworks, "A Multi-Functional Framework for defining Social Media Types, extracting Topics and Inferences, and discovering Correlations based on Public Sentiment" and "A Novel Framework for P2P and VMG2VMG Energy Balancing, Incorporating One Step Ahead Load Forecasting and Optimization for Day-Ahead Energy Scheduling" incorporate common data mining methods/algorithms elevating the necessity for interdisciplinary novel approaches in multi-domain data analytics along with benefits they might yield.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ACID	Atomicity Consistency Isolation Durability
AI	Artificial Intelligence
API	Application Programming Interface
ARM	Association Rule Mining
CNN	Convolutional Neural Network
CPP	Critical Peak Pricing
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DER	Distributed Energy Resource
DM	Data Mining
DR	Demand Response
DRM	Demand Response Management
DSO	Distribution System Operator
DSS	Decision Support System
EAP	Ensemble utilizing Averaged Predictions
EB	Energy Balancing
EBA	Exhaustive (brute-force) Balancing Algorithm
ELF	Energy Load Forecasting
EODS	Energy Optimal Day-Ahead Scheduling
EPE	Ensemble utilizing Polynomial Exhibitor
ET	Execution Time
EWA	Ensemble utilizing Weighted Averages
GBT	Gradient Boosting Trees
LDA	Latent Dirichlet Allocation
LR	Linear Regression
LSTM	Long Short-Term Memory-based
LTLF	Long-Term Load Forecasting
MAPE	Mean Absolute Percentage Error
MG	Microgrid
ML	Machine Learning
MLP	Multi-Layer Perceptron
MTLF	Medium-Term Load Forecasting
NLP	Natural Language Processing
NN	Neural Network
nZEB	nearly Zero Energy Building
OSA-ELF	One Step Ahead Energy Load Forecasting
P2P	Peer-to-peer
PV	Photovoltaic
RES	Renewable Energy Source

RMSE	Root Mean Squared Error
RTP	Real Time Pricing
SG	Smart Grid
SM	Social Media
SMAPE	Symmetric Mean Absolute Percentage Error
SMP	Social Media Platform
SMPr	System Marginal Price
SMSA	Social Media Sentiment Analysis
SMT	Social Media Type
SMTE	Social Media Topic Extraction
SN	Social Network
STLF	Short-Term Load Forecasting
SVR	Support Vector Regression
VMG	Virtual Microgrid
VMG2VMG	Virtual Microgrid-to- Virtual Microgrid
VPP	Virtual Power Plant
VSTLF	Very Short-Term Load Forecasting
WCSS	Within-Cluster Sum of Squares

Statement of Original Authorship

The work contained in this thesis has not been previously submitted to meet requirements for an award at this or any other higher education institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except myself and where due reference is made.

Signature: 

Date: 7/9/21

Copyright Statement

Acknowledgements

I would first like to express my gratitude to my Thesis supervisor Dr. Christos Tjortjis for the patient mentoring, encouragement, advice and guidance during my PhD studies. I have been extremely lucky to have such a clear communication with my supervisor. He has always been there responding to any of my queries promptly and showing me the way to all stages of this research journey, enabling this research to be successfully conducted and submitted to international peer reviewed journals. I would also like to extend my sincere gratitude to other members of my Ph.D. committee, Dr. Panagiotis Tsaparas and Dr. Spiros Denaxas for their assistance, but also to Dr. Dimosthenis Ioannidis and Dr. Dimitrios Tzovaras for trusting me to take on research opportunities and funding certain parts of my research. I must also thank the International Hellenic University and the CERTH-ITI for supporting my research in all possible ways.

Finally, I must also express my deepest gratitude to my parents and family for their support, encouragement and understanding throughout all my years of studies and through the process of researching and writing this Thesis.

Chapter 1: Introduction

This chapter contextualises the conducted research. It outlines the background (Sect. 1.1) and context (Sect. 1.2) of the research, and its purposes (Sect. 1.3). Sect. 1.4 highlights the significance and scope that drives this research and elaborates on the utilized terms. Finally, Sect. 1.5 presents an outline of the remaining chapters. It also splits the content of this thesis into two parts (PART I and II) based on the investigated domain.

1.1 BACKGROUND

This thesis reports on an interdisciplinary approach for data analytics, exposing various machine learning capabilities. Novel research tasks take form engaging in the domains of Social Media (SM) and Energy, enabling knowledge acquisition by utilizing mutual methods/algorithms. This document narrates on two theoretical frameworks that envision novel approaches in SM Types (SMTs), SM Topic Extraction (SMTE), SM Sentiment Analysis (SMSA) and Energy Balancing (EB), Energy Load Forecasting (ELF), Energy Optimal Day-Ahead Scheduling (EODS), respectively.

People around the world use SM to communicate, connect and interact with other users, sharing and propagating information at a great rate (Chaffey, 2021). SM facilitate sharing information, ideas, interests and other forms of expression through virtual communities and networks (Kietzmann *et al.*, 2011). There is a great variety of services offered having many common features (Obar and Wildman, 2015). SM are considered interactive Internet-based applications (Kaplan and Haenlein, 2010). SM are full of user-generated data, such as posts, photos, videos and so on. They offer user accounts (profiles) on websites and mobile apps, facilitating the generation of web based social networks, connecting users or groups (Boyd and Ellison, 2007). A Social Network (SN) is a social structure consisting of several actors/entities/groups of entities, that describe a variety of interactions among them.

Studies like the one reported in (Pallis, Zeinalipour-Yazti and Dikaiakos, 2011) present taxonomies for SN, which describe the spectrum of attributes that relate to these systems. They provide a reference point for different system compositions,

aiming at capturing their building blocks, whilst examining the architectural designs and business models they might pose. SN offer different techniques for analysing the structure of social atoms (entities), as well as a set of theories for understanding and recognizing patterns hidden in them (Scott *et al.*, 1996). Such patterns can be local or global, which can be further analysed in order to mine special entities that might influence others or examine characteristics of parts or the whole network (Otte and Rousseau, 2002). During the early years of SM networking, Social Media Platforms (SMP) had a clear vision statement. Nowadays, most SM provide services and functionalities using different names. SM users take advantage of services such as connecting, sharing, entertaining, monetizing etc., seeking to detect brand awareness indicators, usage for sales, feedbacks, opinions and more, before approaching specific target groups.

Categorizing SMPs helps addressing appropriate groups and improves understanding regarding SM, whilst getting better results from each platform/site. New opportunities arise for research and improvements based on new data at disposal. Although SM networking is considered a new field of studies, more and more researchers work on it, due to its wide user adoption (Tankovska, 2021). SM data types are highly dependent on typical user activities. There are various characteristics and implications on SM that often lead to confusions regarding data handling (Richthammer *et al.*, 2013). Therefore, this work aims to elaborate on Social Media Types (SMTs), updating current literature, as well as to introduce new perspectives on SMPs multiple feature offerings. While reference is made to SMTs and networks, this thesis surveys and categorizes most common such types and researches on an update to their current standardization. To achieve that, SMTs features are extracted forming services that are labelled as "Utilities", developing a methodology based on an initial hypothesis H_0 ("standard SMTs can be narrowed down to a smaller number n") which is later backed up by further elaboration on the SM feature dataset. A report on SM evolution is made and how can a data-driven approach be used to generate a new SMTs taxonomy.

This is significant because SM offer an increasingly wider variety of services, making it difficult to determine their core purpose and mission; therefore, their type. This thesis assesses SMT evolution, presents and evaluates a novel hypothesis-based

data driven methodology for analysing SMPs and categorizing SMTs based on their services.

People have been using SM in an extensive global scale, exchanging messages, posting opinions, news and more. During recent years SM public usage has greatly increased. There are multiple functionalities on offer, rendering SM one of the most popular online activities (Koukaras, Tjortjis and Rousidis, 2020). In 2020, over 3.6 billion people worldwide engaged with SM, with a predicted number of around 4.41 billion users in 2025 (Tankovska, 2021). They became a great source of data for knowledge extraction. The COVID-19 pandemic constitutes a worldwide health crisis, which became one of the hottest discussion topics. People generate vast amounts of online data regarding a variety of issues pertaining economic, social, political or health implications. At the same time, individuals, organizations, corporations and governments use SM. They act as a medium for exchanging information or monitoring opinions and attitudes about this crisis.

The analysis of content derived from SM, such as Twitter, is a challenging task, since a large amount of data needs to be accumulated, summarized or aggregated. In general, SM platforms tend to generate text data that are sparse and noisy. Great endeavour is required to analyse them for knowledge extraction. Therefore, nowadays more than ever, there is a need for methods and techniques to handle these huge volumes of data and generate opportunities for mitigating global crises like COVID-19. Governments, companies, organizations, and other involved parties need tools for understanding the topics of discussions in such events. These opportunities may involve better decision support for policy makers, improved and useful online information retrieval and more. Data Mining (DM) techniques, such as Clustering, Classification and Association Rule Mining (ARM) are widely used for the extraction of knowledge from SM data in various domains, such as healthcare (Rousidis, Koukaras and Tjortjis, 2020).

Researchers often utilize such techniques for aggregating or summarizing information retrieved from online content. If grouped/modelled appropriately this content can generate topics or themes that effectively represent the essence of the data (Vayansky and Kumar, 2020). This thesis attempts to introduce improvements in topic extraction from SM data. It showcases a methodology that narrows down the selection of wordsets to be included in extracted topics. This is achieved by utilizing a topic

extraction method and ARM to identify word frequency appearances in these topics. Such methodology can lead to the generation of fewer topics, with stronger word inference that represent public attitudes as expressed by Twitter posts.

During the COVID-19 pandemic, traffic in almost all popular SM has increased over 10%, mainly due to regional lockdowns and the need for additional information regarding the virus. Posts on SM circulate expressing positive attitudes of those who believe that humanity deals indeed with a dangerous virus outbreak, negative attitudes of sceptics or even virus deniers, and neutral attitudes. An online article presented findings about increased internet traffic during the pandemic. Facebook reported bandwidth issues, Instagram stories were increased by 15%. Twitter and LinkedIn also demonstrated an increase, Tik Tok downloads went up by 18%. Even some alternative types of communication and interaction like Zoom, a video communication application, and podcasts, increased by 52% (*How social media traffic is adapting to COVID-19 - Marx Layne*, 2020). SM, such as Twitter, allow the global community to deal with COVID-19 by offering reliable information, online connectivity, and real-time event tracking. However, as later studies demonstrated, there are also disadvantages of using SM as a source of information. The online anti-vaccine movement is growing and shares many ties with COVID-19 deniers and conspiracy theorists (Twitter Inc., 2021).

Enhancing Microgrid (MG) management has recently drawn significant research attention (Cagnano, De Tuglie and Mancarella, 2020). MGs offer great opportunities for improving energy distribution through balancing, CO2 emissions reduction, energy production, cost reduction etc. (Tabatabaei, Kabalci and Bizon, 2019). The main challenges that the energy sector faces refer to securing seamless and reliable power system operation, the best way possible (Asmus, 2010). For example, Renewable Energy Source (RES) integration could aid reducing the fluctuations on daily energy loads paired with improvements in energy storage. However, RESs are intermittent in nature. Therefore, novel decision support systems are required to manage their smooth integration. Moreover, it is important to utilize them in various ways for minimizing power losses, as well as optimal energy management regarding peak demands (Badami *et al.*, 2019).

Energy sector professionals and researchers need ways to improve energy management in the presence of RES. Energy balancing at the P2P and VMG2VMG

level can act as a tool for that. This is important since it could cause: i) potential energy cost reduction in energy transactions between VMGs and aggregators, as well as aggregators and the Distribution System Operator (DSO), ii) main grid tariff costs reduction, iii) better energy distribution by enabling power trade-offs in-between prosumers and (iv) prosumer participation to energy-sharing programs. Energy can be transferred either uni- or bi-laterally. Opportunities for Demand Response (DR) power transfer at the P2P level arise. These points offer novel advancements in the energy sharing research domain. Reviewing the literature, shows that energy balancing draws the attention of many researchers from various fields. This can be attributed to the wide range of applications and domains required for developing the next generation of power systems, i.e. The Smart Grid (SG).

These requirements attempt to meet the increasing needs for technological implementations brought to the energy sector by the rapid spread of RES integration. Energy distribution/balancing is considered a challenging domain, as it involves multiple factors, such as RES intermittency, peak hour demands, DR signals, etc. Historical data can offer an insight into the times that energy demand is at its peak. Stochastic indicators can be retrieved, about the times during the day when grid functionality relaxes, or energy demands are nearly even. This thesis elaborates on forming clusters or bins of prosumers to achieve improved energy management for Virtual Microgrids (VMGs). It adopts an interdisciplinary approach, incorporating concepts from the domains of Computer Science and Energy. Energy management is an important task as confirmed by the review of the literature. Balancing energy inputs/outputs to the grid, the main aim of this work, is equally important. Also, it provides an analysis offering various functionalities for VMG balancing, that can be incorporated to a stand-alone toolkit.

These functionalities can be combined with external components. Alternatively, they can act as baseline architectural system components, such as Blockchain, Energy Business Intelligence. This way they can address other energy related problems, like minimizing costs/kWh and CO2 emissions. Thus, they can help to conceive a complete energy management system. The dataset utilized for simulations refers to 94 prosumers. The goal is to perform balancing at the Peer-to-peer (P2P) and VMG2VMG level. Such an achievement is verified by multiple simulations and visual presentations. The simulations incorporate clustering and binning algorithms to handle

energy prosumer data. An Exhaustive brute-force Balancing Algorithm (EBA) handles VMG2VMG energy balancing. Simulation outputs can be combined with various Use Cases. For example, i) cost minimization or profit maximization, during the energy transactions between VMGs and aggregators and aggregator and DSO, ii) opportunities arising from VMG2VMG energy transactions, when RES are used at a large scale and iii) VMGs capability to utilize P2P energy transactions, i.e., energy trade among prosumers.

Energy Load Forecasting (ELF) plays a major role in any power system operation and design. It allows utility providers to model electricity consumption and prepare for future power loads and the Distribution System Operators (DSOs) to manage and match future energy generation with consumption. The need for accurate predictions is evident given that the energy distribution infrastructure is undergoing a rapid transformation, expanding its capabilities, while envisioning faster response times regarding its elasticity for energy resource allocation. This can result in multiple benefits, such as production cost reductions, realistic energy price estimation (Grimes *et al.*, 2014), better power output scheduling management and better future system capacity planning.

RES become increasingly common, while introducing parameters previously unknown to the energy sector, such as the intermittent nature of renewable generation. These factors have different weights when trying to solve load forecasting problems, as they introduce restricting variables, such as seasonality (Javaid and Javaid, 2020), trends, weather conditions (Sharma *et al.*, 2014) etc. Research has been conducted to improve ELF accuracy, when contemplating particular load types, such as residential consumption, which was proven quite challenging. It involves great uncertainty, especially when compared to other load types, such as industrial or commercial (*SmartHome / ITI*, 2020).

Identified types of ELF are of paramount importance for the correct and unceasing functionality of power systems (Xia, Wang and McMenemy, 2010). Short-Term Load Forecasting (STLF) offers benefits regarding operational security and power system savings, Medium-Term Load Forecasting (MTLF) aids in planning and operation, whereas Long-Term Load Forecasting (LTLF) is suitable for planning future investments in power system infrastructure (Torkzadeh *et al.*, 2014). With technological advancements in IoT devices (Mukherjee *et al.*, 2020) it has become

easy to capture information from energy smart meters (Casado-Mansilla *et al.*, 2018) or sensor nodes and predict available energy reserves (Gebben, Bader and Oelmann, 2015). This allows practitioners and researchers to develop approaches for more reliable and efficient energy management solutions. It is now easier to produce accurate probabilistic models for individual ELF, improving the prospects for better smartgrid management. On the other hand, aggregated energy consumption can be easier, since it smooths irregularity, volatility, and spikes that individual energy profiles exhibit (Hsiao, 2015).

This work presents research results regarding One-Step Ahead ELF (OSA-ELF), a form of STLF, while surveying other state-of-the-art attempts in the field. A fine-tuned timeseries forecasting approach is proposed and implemented to predict OSA-ELF, utilizing ensemble and supervised learning techniques. This approach does not heavily depend on time resolution measurements, since the aim is to predict the next step. The fundamental logic of ensemble relates to merging/combining multiple predictions derived from different models and utilize the ones that yield better overall predictions. Then, the results can ensemble as averaged predictions or ensemble using weighted averages or ensemble utilizing polynomial exhibitors. This approach expands on the energy load prediction capabilities by implementing an OSA-ELF approach, developing a separate prediction model for each forecast time step.

In the energy sector, multi-objective optimization is a demanding problem involving many real-world parameters such as flexibility and DR management. These may pose as dependent, or conflicting objective problems. Addressing such problems with techniques like scalarization that transforms multi-objective into single-objective problems is quite common (Jahn, 1985). Other ways involve multiple objective functions that describe problems in detail. Optimization problems generate research challenges based on the solving approach. For example, real-world optimization problems are usually modelled as non-linear programming problems with many objectives. Conversions of such problems to single objective may cause practical issues, since it outputs a single optimal solution considering trade-offs identified on a single, transformed problem. In such cases, a certain degree of detail is omitted rendering the approach non-realistic.

New trends in multi-objective optimization attempt to retain compulsory problem details defining multiple objective functions to be solved in parallel. Solutions

are formulated as optimal Pareto fronts, generating many options for the best solution each acting as a trade-off for another. The state-of-the-art aims to develop methods that improve efficiency and speed of finding optimal solutions for forming Pareto fronts. Exhaustive approaches involve implementation of repetitive algorithms having each iteration outputs a solution closer to the optimal, making such problems time complexity dependent. On the other hand, heuristic approaches such as evolutionary algorithms that introduce population approaches mitigate the issue of time complexity offering optimal solutions on a single run. Any decision making related with multi-objective problems should focus on efficient and timely solutions integrating fine-tailored algorithmic implementations based on problem complexity.

Single-objective optimization finds one optimal solution optimizing only one objective function. Multi-objective optimization finds two or more optimal solutions optimizing many objective functions at the same time, while many optimal solutions derive from the objective space. Optimal solutions are often visualized forming a Pareto front aiding an enhanced decision-making process (Poli and Koza, 2014). This thesis, elaborates on the conception of a tri-layer optimization framework. To achieve day-ahead optimal energy scheduling, energy load is optimized for two actors (consumer, aggregator) while managing their interaction with a Decision Support System (DSS). The aggregator's problem poses as a single-objective optimization problem while consumer poses as a bi-objective optimization problem. The aggregator's optimization is expanded to implement a DR signal scheme that aims to optimize portfolio energy management while reducing overall cost. Therefore, after the performed analysis, cost is minimized and comfort is maximized, considering the profiles of all consumers involved, while offering optimization options for both consumers and aggregators.

The proposed tri-layer optimization framework offers autonomous consumer and aggregator optimization and the possibilities to collaborate, in case it is deemed profitable for either of them. It highlights consumer capabilities allowing the optimization of assets without the need of DR signals from the aggregator. This approach generates new options to relax the energy contract (between consumer and aggregator) and introduce elasticity on DR signal acceptance. It also increases the prospects for autonomous peer-to-peer (P2P) level energy optimization (Koukaras, Tjortjis, *et al.,* 2021). The proposed framework's validation on real-world pilots

showcase that optimization to minimize cost, while keeping occupant discomfort at acceptable levels through a flexible DR scheme enforcement.

The reason for addressing the abovementioned research tasks being contextualized into two frameworks; one narrating on three novel approaches in SM and one on three novel approaches in Energy has to do with the research prospects and opportunities that came up during the PhD studies. Combining findings of two domains while utilizing research methodologies based on common methods/algorithms for clustering, association rule mining etc. while conceptualizing interdisciplinary approaches may yield novel results in data analytics and knowledge acquisition.

1.2 CONTEXT HIGHLIGHTS

The Social Media (SM) theoretical framework assesses the evolution of Social Media Types (SMTs), proposes and evaluates a novel hypothesis-based data driven methodology for a new taxonomy on SMTs, based on:

i) The hypothesis that the number of SMTs is smaller than what current literature suggests,

ii) Observations on data regarding SM usage/services, and

iii) Experimentation using association rules and clustering algorithms.

Also, it extracts knowledge about public attitudes regarding worldwide crises.

i) It reviews recent advances on topic and opinion extraction methods from social media data.

ii) It performs and assesses topic extraction related with the COVID-19 use case.

iii) It proposes improvements to the Latent Dirichlet Allocation by incorporating Association Rule Mining to deal with common discrepancies.

iv) The proposed methodology reduces the number of topics, whilst retaining the core themes of user attitudes.

Finally, it evaluates public attitudes towards the COVID-19 spread, by performing sentiment analysis on over 2.1 million tweets in English. The tweets are globally gathered from February to August 2020. It investigates whether there is a correlation

between public sentiment and the number of cases and deaths attributed to COVID-19.

i) It Improves the sentiment polarization output by utilising text preprocessing techniques.

ii) It validates trends on twitter sentiment in relation with the actual number of COVID-19 Cases and Deaths.

iii) It develops an approach that creates opportunities for disease forecasting by discovering and monitoring multivariable correlations, such as polarity vs. cases or polarity vs. deaths.

The Energy theoretical framework, presents an interdisciplinary approach investigating the energy balancing problem. Its contributions can be summarized as follows:

i) A heuristic approach utilizing unsupervised learning to form and balance VMGs.

ii) A high-level, generic approach that addresses P2P and VMG2VMG balancing. The presentation of simulations combining clustering/binning for preprocessing is detailed. This approach can act as a standalone toolkit for baseline estimation, or as part of a comprehensive energy management system.

iii) A potential solution to the energy balancing problem which incorporates a Computer Science centric approach for an Energy sector problem. The investigation conducted emphasizes the need for multidisciplinarity for such problems, due to the increased complexity of fast evolving energy management systems.

Also, it deals with Energy Load Forecasting (ELF). A use case for testing improvements regarding load forecasting capabilities is the residential house energy requirements. This work utilizes historical data from a pilot performing multiple tests for improved ELF. It focuses on the aspects of one step ahead ELF, while aiming at producing an approach that can be utilized regardless of the time resolution of available data. It predicts the “next step” value regardless of the duration of the step (15-minutes, one-hour, one-day etc.) with relevantly high accuracy, offering a tool that can be used for a wide variety of forecasting applications. The evaluation of forecasting results is performed utilizing popular accuracy metrics.

Finally, it addresses optimal energy management. A multi-objective analysis may act as a tool that offers responses for optimal energy management through a decision-making system. Actors (consumers and aggregators) may completely or partly interact with the other in the form of Demand Response signal exchange. The overall optimization is formulated by a bi-objective optimization problem for the consumer's side aiming at cost minimization and discomfort reduction; and a single objective optimization problem for the aggregator's side aiming also at cost minimization. A Decision Support System (DSS) tool incorporates forecasting of day-ahead energy management requirements. The main goal is to achieve an overall optimal management of energy resources considering both aggregator's and consumer's preferences and goals.

1.3 PURPOSES

This section specifies interdisciplinary knowledge acquisition purposes, aims while emphasising on practical outcomes. Six research tasks address novel approaches in Social Media (SM) and Energy domains.

The first research task deals with SM types. SM have been established as multifunctional networking tools that tend to offer an increasingly wider variety of services, making it difficult to determine their core purpose and mission, therefore, their type. The first research task assesses this evolution of Social Media Types (SMTs), presents, and evaluates a novel hypothesis-based data driven methodology for analysing Social Media Platforms (SMPs) and categorizing SMTs. It reviews and updates literature regarding the categorization of SMPs, based on their services. It develops a methodology to propose and evaluate a new taxonomy. As a result, three SMTs are proposed, namely Social, Entertainment and Profiling networks, typically capturing emerging SMP services. The results show that the hypothesis is validated by implementing such a methodology and threats to validity are discussed.

The second research task retrieves data from Twitter to mine association rules and extract knowledge about public attitudes regarding worldwide crises. It exploits the COVID-19 pandemic as a use case, and analyses tweets gathered between February and August 2020. The proposed methodology comprises topic extraction and visualization techniques, such as wordclouds, to form clusters or themes of opinions. It uses association rule mining to discover frequent wordsets and generate rules that

infer to user attitudes. The goal is to utilize Association Rule Mining (ARM) as a postprocessing technique to enhance the output of any topic extraction method. Therefore, only strong wordsets are stored after discarding trivia ones. Also, frequent wordset identification is employed to reduce the number of extracted topics. Findings showcase that 50 initially retrieved topics are narrowed down to just a few ones, combining Latent Dirichlet Allocation with ARM. The methodology facilitates producing more accurate and generalizable results, whilst exposing implications regarding social media user attitudes.

The third research task deals with the identification and analysis of sentiment polarity in microblog data has been drawing increasing attention. Researchers and practitioners attempt to extract knowledge by evaluating public sentiment in response to global events. Tracking and alerting the public in case of a pandemic can produce valuable insights for reducing the negative impact in terms of human lives, policies, societies, economies and more. This study aims to evaluate public attitudes towards the COVID-19 spread, by performing sentiment analysis on over 2.1 million tweets in English. The tweets are globally gathered from February to August 2020. It investigates whether there is a correlation between public sentiment and the number of cases and deaths attributed to COVID-19. Findings highlight a correlation of sentiment polarity and deaths, starting 41 days and expanding up to three days prior the count. The conducted analysis also detects a strong correlation between COVID-19 twitter conversation polarity and reported cases, but a weak correlation between polarity and reported deaths.

According to the forth research task, a way to improve energy management is to perform balancing both at the Peer-to-peer (P2P) level and then at the Virtual Microgrid-to-Virtual Microgrid (VMG2VMG) level, while considering the intermittency of available Renewable Energy Source (RES). This thesis proposes an interdisciplinary analytics-based approach for the formation of VMGs addressing energy balancing. The proposed approach incorporates Computer Science methods to address an Energy sector problem, utilizing data preprocessing techniques and Machine Learning concepts. It features P2P balancing, where each peer is a prosumer perceived as an individual entity, and VMGs as clusters of peers. Several simulations were conducted utilizing clustering and binning algorithms for preprocessing energy data. This study offers options for generating VMGs of prosumers, prior to using a

customized Exhaustive brute-force Balancing Algorithm (EBA). EBA performs balancing at the cluster-to-cluster level, perceived as VMG2VMG balancing. To that end, the study simulates on data from 94 prosumers, and reports outcomes, biases, and prospects for scaling up and expanding this work. Finally, this thesis outlines potential ideal usages for the approach, either standalone or integrated with other toolkits and technologies.

The fifth research task reports on research regarding Energy Load Forecasting (ELF). Energy sector stakeholders, such as Distribution System Operators (DSO) or Aggregators take advantage of improved forecasting methods. Increased forecasting accuracy facilitates handling energy imbalances between generation and consumption. It also supports Smart Grid framework processes, such as Demand-Side or Demand Response Management (DRM). This task presents a novel approach for One-Step-Ahead Energy Load Forecasting (OSA-ELF), considering several techniques. It utilizes historical data from a state-of-the-art nearly Zero Energy Building (nZEB) smart home, performing multiple tests for improved ELF. It focuses on OSA aspects of ELF, yet it can be utilized regardless of the time resolution. It predicts the "next step" value, regardless of the step's duration (15-minutes, one-hour, one-day etc.) with high accuracy, and can be used for a wide variety of forecasting applications. To that end, fine-tuned ensemble methods and forecasting algorithms were utilized for experimenting with short term ELF. Forecasting evaluation produced good results with regards to popular accuracy metrics, Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), Symmetric Mean Absolute Percentage Error (SMAPE) and an Execution Time (ET) metric.

The sixth and final research task addresses the issue of energy optimization. Over the last decades, industry and academic communities have made great strides at improving aspects related with optimal energy management. These include better ways for efficient energy asset management, generating great opportunities for optimization of energy distribution, discomfort minimization, energy production, cost reduction and more. This task proposes a framework for a multi-objective analysis, acting as a novel tool that offers responses for optimal energy management through a decision-making system. The novelty resides to the structure of the methodology since it considers two distinct optimization problems for two actors (consumers and aggregators), that each solution may completely or partly interact with the other in the form of Demand

Response signal exchange. The overall optimization is formulated by a bi-objective optimization problem for the consumer's side aiming at cost minimization and discomfort reduction; and a single objective optimization problem for the aggregator's side aiming also at cost minimization. The framework consists of three architectural layers, namely the consumer, aggregator and DSS, forming a tri-layer optimization framework with multiple interacting objects (objective functions, variables, constants and constraints). The DSS layer is responsible for decision making by forecasting the day-ahead energy management requirements. The main goal is to achieve an overall optimal management of energy resources considering both aggregator's and consumer's preferences and goals. This is also conducted through exhaustive simulations using real data from a real pilot, that is part of Terni Distribution System portfolio.

1.4 SIGNIFICANCE, SCOPE AND DEFINITIONS

This section discusses the identified research problems contained in this thesis. It also offers a high-level presentation of the research design and developed methodologies for highlighting research significance and scope.

This thesis narrates on research achievements in interdisciplinary data analytics utilizing machine learning capabilities for enhanced knowledge acquisition. It splits into two theoretical frameworks split in six research tasks. Three address novel actions in the Social Media domain (PART I) and three associate with the Energy domain (PART II). Each task incorporates a mixed methodology process, merging the concepts of Experimental (E), Correlational (C), Discourse (D), Framework (F), Case Study (CS) and Analysis (A) methodological types (Fig. 1).

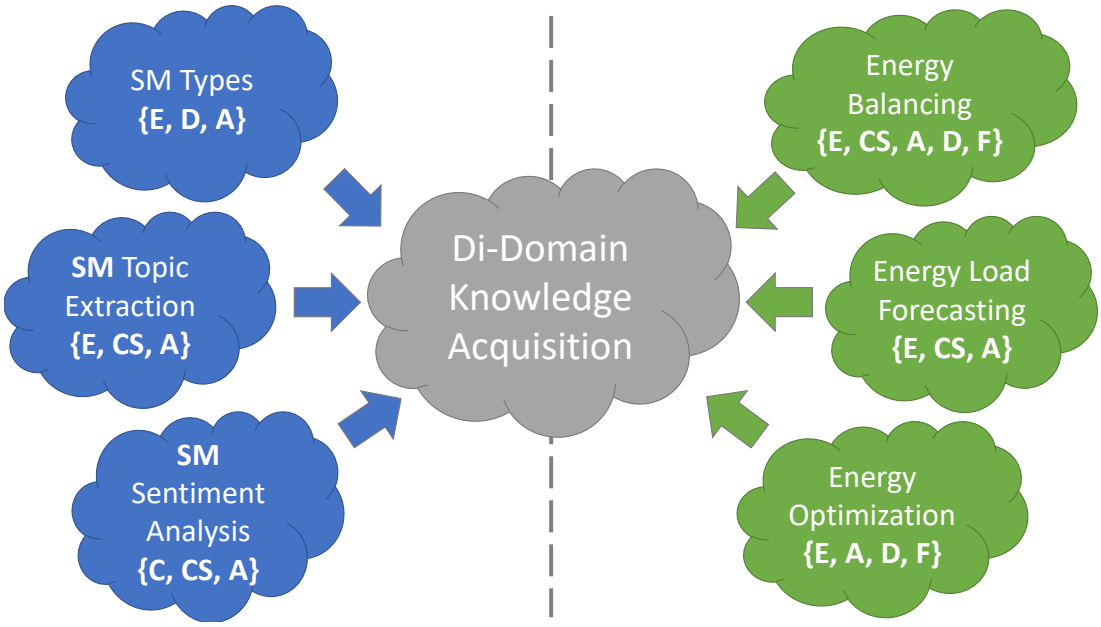

Fig. 1. Overview of Research.

For improving reference to each research task, a short title is applied along with a description and an informative sub-domain it belongs (Table 1).

Table 1. Research Tasks.

Task	Abbreviation	Description	Sub-Domain
1	SMTs	Research and update state-of-the-art on SM types.	SM platforms
2	SMTE	Improve SM topic extraction	Twitter, COVID-19
3	SMSA	SM sentiment analysis for insights generation	Twitter, COVID-19
4	EB	Balancing on a Cluster to Cluster and P2P level	MGs and P2P Energy transfer
5	ELF	Timeseries forecasting with ensemble methods	Occupant Energy consumption
6	EODS	Conception of a DSS for multi-objective optimization of Energy assets for day-ahead Energy Scheduling	Energy Portfolio management & Occupant profiling

Based to the first research task, the current standardization on categories of SMTs (like the ones presented in (Kietzmann *et al.*, 2011); (Kaplan and Haenlein, 2010); (Gundecha and Liu, 2012)) is considered decaying, since SMTs develop rapidly. There are platforms that offer various services and multiple features that are labelled as Utilities. The aim is to introduce a new taxonomy that narrows down the

current SMTs standardization, since most of the modern SMPs tend to offer multiple Utilities into a single platform/product. Therefore, this thesis investigates this issue, expecting to offer another option regarding SMTs. The developed methodology takes into consideration observations on a dataset that contains different SM alongside their official features. Two experiments are performed involving association rule mining and clustering in order to unfold a data-driven methodology that validates a summarized research question: "Can the current state-of-the-art on SMTs be updated by reducing the number of SMT standards; thus, better reflecting the current state of play?"

The second research task addresses identified problems with the topic of SM topic extraction. SM has become a great source for data retrieval for a variety of domains of interest. They offer new opportunities for knowledge extraction in worldwide scale especially for emerging topics like a pandemic. COVID-19 caused a variety of economic, political, social, and health effects. Governments, companies, organizations, and other involved parties need tools for understanding the topics of discussions in such events. This thesis experiments on COVID-19 Twitter data by introducing a novel methodology for identifying topics of discussions related with it.

It performs a combination of a common topic extraction technique, Latent Dirichlet Allocation (LDA) (Blei, Ng and Jordan, 2002) and then enhances its output with ARM. The methodology aims at preprocessing tweets for a specific period and then extracting knowledge related to public opinion. The goal is to mitigate issues that may arise during a topic extraction process and precisely identify topics. Topic extraction techniques can be generic, often generating wordsets that do not infer to topics in a clear manner. At the same time, quite often, words can appear in multiple extracted topics leading to redundant data that are not ACID (atomicity, consistency, isolation, durability) compliant (Haerder and Reuter, 1983). For example, LDA tends to appoint the same words within tweets to multiple topics. This occurs since the selection of words per topic is based on the probability of each word belonging to that topic.

Finding the same words along multiple topics means that these words have high probabilities in those topics. Moreover, LDA requires a fixed number of topics to be known ahead of time. This thesis overcomes this by performing topic coherence and topic stability analysis. It also tackles the non-hierarchical nature of most topic

extraction approaches by allowing the sharing of unique wordsets by creating a pool of words and retrieve the strongest of them. Finally, it mitigates issues regarding static topic extraction by predefining the period for retrieving COVID-19 related topics, but also by not considering time as an investigated feature.

The methodology consists of two main processes. First, topic extraction is performed using LDA. Then ARM takes place to extract the strongest wordset rules enhancing the overall topic extraction output. The resulting wordsets infer to discussion topics about the pandemic, whilst mitigating the issue of multiple word-to-topic assignments. The strongest rules can be visualized using graphs or WordClouds outlining in a more accurate and clear manner the frequently discussed topics. At this point, insights about SM user attitudes can be extracted. Possible implications are exposed, generating prospects for a Decision Support System (DSS). This can monitor single or multiple SMs, while reporting on public attitudes on a worldwide scale.

The third research task investigates the existence of correlations of tweet sentiment polarization with cases and deaths from COVID-19. The approach builds upon three contextual layers: the Preprocessing and Polarization layer, the Hypotheses Formation layer and the Statistical Analysis layer. The first layer attempts to assess the sentiment of posts on Twitter. The second layer aims to form hypotheses related to COVID-19 cases and deaths. The third layer validates the existence of correlations between the timeseries of these three variables (tweet sentiment, cases, deaths). The significance of this work is attributed to the fact that COVID-19 has caused economic and psychological problems to individuals and societies, on top of medical ones.

By addressing these points, knowledge extraction capabilities can be improved utilizing SM data and elaborating on trends or correlations between different data sources. This enhances the predictive, descriptive, and diagnostic analytics of this study (Koukaras and Tjortjis, 2019). This task addresses the following research questions: Can tweets act as an indicator for predicting polarity related with COVID-19 cases and deaths? Is there a correlation between these three data labels (tweets, cases, deaths)? If yes, to what extend? Are there any trends in these data labels? The developed methodology attempts to answer these questions envisioning future improvements involving association rule mining, forecasting and comparative reporting of various approaches.

The forth research task deals with the problem of Energy balancing in a portfolio of residential, commercial and industrial assets which can be quite challenging, due to the great intermittency and stochasticity involved i.e., diverse energy consumption behaviour, weather conditions, the intermittency of RES and more. Also, there are local ecosystem constraints, such as power loads, a variety of RES, costs, grid topology or local fees that ensure the fairness and stability of such an energy system (Cardoso *et al.*, 2018). Often the available assets form VMGs conceptualized as clusters, which have specific parameters i.e., Key Performance Indicators (KPIs), such as energy consumption or generation. In order to achieve better energy management, especially when extensive usage of RES can be exploited, new ways for Energy balancing at the P2P and VMG-2VMG level should be conceived. The DSO and aggregators need tools that aid accuracy regarding Distributed Energy Resource (DER) power operation information, while exposing new possibilities regarding portfolio management.

The information available should be presented in an aggregated form, enhancing monitoring capabilities of both DSO and aggregator; the former in terms of managing the grid, the latter in terms of managing its portfolio. Most of the approaches like in (Zhang *et al.*, 2018), (Anoh *et al.*, 2020) and (Vergados *et al.*, 2016) either attempt to generate VMGs to reduce the cost of energy or address the issue of energy balancing at the P2P level. Developing a tool addressing the points mentioned above may offer a long-term improvement on applied balancing strategies. At some point, these can be addressed by introducing the idea of local or global optimal balancing. Such balancing can expose diagnostic and prescriptive capabilities in data analytics, similar with other identified data science domains (Koukaras and Tjortjis, 2019).

This exposes new possibilities, since prosumer portfolios can be virtually clustered and balanced. Local balancing is perceived through balancing at the P2P level, while global balancing is perceived after performing balancing at the VMG2VMG level. Implications include the combination of such a tool with a cost minimization process that involves transferring energy from individual VMGs to the VMG portfolio of an aggregator or the DSO. Before energy transfer to the latter commences, a global VMG2VMG balancing could be performed. Further investigation, associated with VMG2VMG energy transfer during excessive usage of RES, could take place. VMGs enable trade-offs of energy in-between VMGs, offering optimal energy distribution in close proximity, and maximizing savings from main

grid tariffs. Such a tool should be provided as software, hardware or data as a service. For deployment and operation, cloud computing technologies may offer a complete information technology solution (Dikaiakos *et al.*, 2009).

The fifth research task deals with ELF that plays a vital role in the operation and management of power systems and can be classified into categories with different purposes and time-horizon as follows: Very Short-Term Load Forecasting (VSTLF), refers to predictions for very short time-windows. These fluctuate from one minute to one or several hours into the future (usually with five-minute steps), while training the prediction model utilizing historical energy loads, measured on the same day (Guan *et al.*, 2013). Short-Term Load Forecasting (STLF) refers to predictions that aim to present estimations of load levels for the next 30 minutes up to two weeks (Jacob, Neves and Vukadinović Greetham, 2020). Mid-Term Load Forecasting (MTLF), deals with predictions presenting estimations of the energy load levels for a few months up to a few years (Amjady and Daraeepour, 2011), a time-window subject to various similar research problem definitions (Bunnoon, Chalermyanont and Limsakul, 2009). Long-Term Load Forecasting (LTLF) may cover time horizons expanding between one to 10 years or even longer (Daneshi, Shahidehpour and Choobbari, 2008).

The approach proposed in this thesis can be used uniformly for any given energy load prediction task, regardless of load forecasting type. To achieve that, the term "one-step" ahead is employed, claiming that the performed prediction is contextually dependent and relative to the historical energy load input data. Yet, this approach performs well for STLF, as it aims to optimize the one-hour ahead ELF resolution, experimenting with real data (nZEB building loads) improving its energy load management and its power saving capabilities (Martín-Gómez, Vidaurre-Arbizu and Eguaras-Martínez, 2014).

In this context, this thesis also presents several state-of-the-art attempts in timeseries ELF. It considers the following ensemble methods: Ensemble utilizing Averaged Predictions (EAP), Ensemble utilizing Polynomial Exhibitor (EPE) and Ensemble utilizing Weighted Averages (EWA). It also contemplates widely used algorithmic options for modeling the forecasting process: Multi-Layer Perceptron (MLP), Long Short-Term Memory-based (LSTM), Gradient Boosting Trees (GBT), Support Vector Regression (SVR) and Linear Regression (LR). The performance evaluation of the proposed OSA-ELF approach is conducted utilizing the following

accuracy metrics: Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE) and Symmetric Mean Absolute Percentage Error (SMAPE).

The sixth research task offers a novel solution on Energy optimal day-ahead scheduling. Due to the heavy penetration of RES in energy grids, the flexibility parameter plays a paramount role for improving energy distribution, stability and reliability. Flexibility enables an improved management of any type of energy transfer related with the grid, according to an initiation signal. Such transfers include energy loads, energy generation within the grid or incoming and outgoing from the grid. It also creates new opportunities for energy profiling and portfolio management offering new capabilities for both consumers and aggregators. Both consumers and aggregators may monitor power exchanges and interactions more efficiently optimizing the performance of the power grid. Flexibility is dependent on various factors such as DR programs, RES, resource scheduling and more. DR programs should be enforced rigorously since non-compliance penalties render this energy concept inefficient.

Triggering flexibility resources without strict scheduling leads to non-viable costs. DR programs should be enforced in a way that energy sector stakeholders such as the aggregator can effectively handle available energy resources/reserves. That way, the aggregated monitoring and adjustment of flexibility aids in improving the exploitation and application of DR programs and the benefits they yield. There are various studies that contemplate the matter at hand.

There are those that include Battery Energy Storage Systems (BESS) that are managed by aggregators with the consent of the end users (Yang *et al.*, 2020). The use of BESS has the added benefit of not interfering with the end user's consumption and therefore not requiring any consideration on their comforts. Additionally, in this case the prosumers' view is included only as an input for the aggregator's portfolio optimization, mainly considering the aggregator's view. This view is adopted again in (Park, Lee and Won, 2020), where the uncertainty of the load is examined, when an aggregator participates in the DR market. In other cases, although the end user's view is considered, this is done at the expense of reducing the role of the aggregator to just sending the electricity price signals (Menniti *et al.*, 2013). There are also cases where the collaboration of the aggregator with Microgrid clusters incorporates multi-level chance-constrained programming (Daneshvar *et al.*, 2020). However, the main focus of that study is the transactive energy management trading among microgrids.

In this study, both the view of the aggregator and the end-user is examined in equal terms, conducting the optimization for each one semi-autonomously. While the objective remains the best outcome for each, the interests of the other are still taken into consideration. Especially, in the case of the end-user adopting a human-centric approach. Furthermore, since the energy consumption of the end user is implicated, not only its uncertainty is considered, but also the comfort of the user by managing his/hers energy consumption.

Moreover, this study uses pilot site real data regarding an aggregator's portfolio. For addressing, fine-tuning and combining these concepts a multi-objective optimization problem may concurrently handle the parameters of flexibility, consumer discomfort, energy cost and more. Solutions provide data for creating services that enable aggregators to post flexibility/DR signals or participate in electricity markets on a more informed manner. This study integrates both concepts of flexibility and DR and provides a solution in the form of an optimization problem for consumers and aggregators. It enables a DR strategy subject to specific constraints, to generate objective functions before running optimization algorithms for day-ahead energy scheduling. The algorithms tackle the issue of optimal energy distribution based on the combination of a single-objective and a bi-objective optimization. The single-objective optimization minimizes the portfolio cost for the aggregator while the bi-objective optimization optimizes cost minimization along with discomfort for the consumer. This problem approach generates a framework for multi-objective analysis.

The envisioned optimization framework improves portfolio management and DR functionality/efficiency. Also, while minimizing consumer cost it offers acceptable counterbalancing options for consumer discomfort. It poses as a long-term improvement for the applied DR strategies and optimal energy management highlighting an autonomous optimization for consumers and aggregators. It elaborates on possibilities for their collaboration, if it is deemed profitable. The consumer can optimize assets without the need of aggregator’s DR signals. That way, new capabilities for relaxing the contract and DR scheme arise as well as autonomous optimization on P2P level, envisioning an automated DR optimization and scheduling framework.

1.5 PART I & II OUTLINE

Chapters 2 and 6 contain the literature review related with each of the investigated topics relating with the domain of SM and Energy respectively. Also, they summarize identified research gaps and implications of this thesis. Chapters 3 and 7 narrate on research design and methodology including utilized data sources and limitations. Chapters 4 and 8 present the results of the interdisciplinary approach while Chapters 5 and 9 distil and analyse findings. Chapter 10 discusses overall accomplishments, take away message and future directions.

PART I

Chapter 2: Literature Review

This chapter reviews literature on the following topics: Social Media Types (Sect. 2.1); Social Media Topic Extraction (Sect. 2.2); Social Media Sentiment Analysis (Sect. 2.3). Sect. 2.4 highlights the implications from the literature and develops the conceptual framework regarding the Social Media domain.

2.1 SOCIAL MEDIA TYPES

There are various approaches when dealing with a new taxonomy proposal. For example, Engelbrecht et al. categorize data-driven business models based on three points: the data source, the target audience and the technological effort (Engelbrecht, Gerlach and Widjaja, 2016). Then, they propose eight categories of business models. This work aims to research categories of Social Media (SM), therefore SM Types (SMTs), a rather untapped topic regarding SM. Based on Social Theories, there is the Social Atom as an individual that interacts with the Social Molecule which is the community, constructing seven probable building blocks (Identity, Conversations, Sharing, Presence, Relationships, Reputation, Groups) of SM (Kietzmann *et al.*, 2011). A categorization of SM sites (and by extension SMTs) such as blogs, SM sites, and virtual game worlds can be found in (Kaplan and Haenlein, 2010). The classification is based on purpose and functionality. Nine types of SM are identified (Gundecha and Liu, 2012):

i) Online Social Networking Web-based services that allow individuals and communities to connect with real world friends and acquaintances online. Users interact with each other through status updates, comments, media sharing and messages. Examples: Facebook, Myspace, LinkedIn.

ii) Blogging Journal-like websites for users, to contribute textual and multimedia content, arranged in a reverse chronological order. Blogs are generally maintained by an individual or by a community. Examples: Huffington Post, Business Insider, Engadget, WordPress.com, Medium.

iii) Micro-blogging Same as blogs, but with limited content. Examples: Twitter, Tumblr, Plurk.

iv) Wikis Collaborative editing environment that allows multiple users to develop Web pages. Examples: Wikipedia, Wikitravel, Wikihow.

v) Social news Sharing and selection of news stories and articles by communities of users. Examples: Digg, Slashdot, Reddit, Quora.

vi) Social book-marking Allows users to bookmark Web content for storage, organization, and sharing. Examples: Delicious, StumbleUpon.

vii) Media sharing Sharing of media on the Web including video, audio, and photos. Examples: YouTube, Flickr, UstreamTV.

viii) Opinion, reviews and rating. The primary function of such sites is to collect and publish user submitted content in the form of subjective commentary on existing products, services, entertainment, businesses and places. Examples: Epinions, Yelp, Cnet, Zomato, TripAdvisor.

ix) Answers Platforms for users seeking advice, guidance or knowledge to ask questions. Other community users can answer these questions based on previous experiences, personal opinions or relevant research. Answers are generally judged using ratings and comments. Examples: Yahoo! answers, WikiAnswers.

2.2 SOCIAL MEDIA TOPIC EXTRACTION

In (Cataldi, Di Caro and Schifanella, 2010) authors highlight the primary role of Twitter to facilitate short text messages for a variety of purposes, whilst proposing a novel topic detection technique. This technique allows real-time retrieval of top emergent topics in communities. Text content is extracted from tweets and modelled according to a life circle introducing an aging theory. Identified terms are labelled as emerging, if they frequently occur in a specified time interval and have been rarely used in the past. The authors also used the Page Rank algorithm to rank the importance of the content based on its source. The study's findings are validated by a navigable graph. It connects emerging terms accompanied by keywords, under user specified time slots for various use cases.

Twitter offers microblogging services. It generates a great number of instant messages, creating opportunities for opinion summarization. Celebrities and brands

are entities that generate a large volume of tweets. A proposal about an entity-centric and topic-based opinion summarization framework is presented in (Meng *et al.*, 2012). The methodology for summarizing topics based on extracted opinions comprises of i) mined topics from hashtags, ii) grouping hashtags on a weekly basis, iii) similarity calculation among them and iv) utilization of the Affinity Propagation algorithm to group hashtags into coherent topics. Then a dependent sentiment classification approach identifies the opinion for a specific target of tweets generating insights. The integration of topic, opinion, insights and other factors (e.g., language styles), forms an optimization framework for extracting opinion summaries.

Analysing content from Twitter and attempting to summarize its information might become a quite challenging task. This process can be accomplished by extracting topical key phrases. More specifically, a context-sensitive PageRank method that considers locality, ranks keywords along with a probabilistic scoring function. It also deals with concepts such as relevance and interestingness of key phrases, before proceeding to ranking key phrases. The approach was validated by experimenting on a Twitter dataset, showing the efficacy of the topical key phrase extraction process (Zhao *et al.*, 2011).

SM pose a great source of user generated information context. That context exposes user interaction, streams of content, friendships and more. In Twitter, this user content can be extracted by exploiting conversation patterns and lists of user generated Twitter data. In such an approach, mined user context can generate user topics of interest. The validity of this attempt showcases an 84% precision regarding the indication that topic information can be extracted from just the user context (Pochampally and Varma, 2011).

Since SM contain rich and abundant information, there is also the need for successful filtering and extraction of trending topics and events. A variety of methods exists for this process, exposing different qualitative results. A comparative research identified six topic detection methods, validated with three Twitter datasets regarding events. Variables, such as the nature of the event, the activity over time, sampling and preprocessing related data affects each method. Standard Natural Language Processing (NLP) methods do well on very specific topics, while novel methods need to be employed for handling heterogeneous streams of concurrent events. A novel topic

detection method based on topic co-occurrence and ranking, seems to perform well, considering the aforementioned conditions (Aiello *et al.*, 2013).

A study about criminal incidents exploits SM data to predict such events that previously relied only on historic crime records, geospatial data, and demographics (Wang, Gerber and Brown, 2012). In SM there might be context that involves incidents of interest, inferring to possible upcoming events. This approach combines NLP of Twitter posts, dimensionality reduction with Latent Dirichlet Allocation (LDA) and a linear prediction model. The evaluation of results is performed attempting to predict hit-and-run crimes, showcasing that it performs better than a baseline model for multiple days.

For effective text mining in real-time data from SM, the process of stemming maybe valuable. New text mining challenges arise due to the great amount of text data produced by SM. Most of the text mining techniques apply to pre-defined datasets that are processed without applying limitations to computational complexity and execution times, while not paying attention to event triggers. A work that proposes a lightweight event detection method that uses wavelet signal analysis of hashtags is presented in (Cordeiro, 2012). LDA along with Gibbs Sampling become a part of a proposed strategy for detecting events while Continuous Wavelet Transformation identifies mention hypes for user specified hashtags. Results show that the proposed approach can summarize Twitter events in streaming environments.

A novel technique for topic modelling utilizing Twitter data is presented in (Alvarez-Melis and Saveski, 2016). It creates groups of tweets occurring in the same conversation between two users. Therefore, a new scheme takes form, allowing tweets and their replies to be aggregated into a document. The users who post them are marked as co-authors. An experimental dataset for topic modelling occurs by utilizing LDA and Author-Topic Model (ATM) to create pools of tweets. A comparative analysis shows that the proposed approach outperforms others in the quality of formed clusters and document retrieval.

Topic coverage and sentiment dynamics from Twitter and news publications can be used for a comparative source analysis regarding medical topics, such as Ebola. In (Kim *et al.*, 2016) the authors perform content and sentiment analysis by collecting 16,189 news articles from 1,006 different publication sources and 7,106,297 tweets from the Twitter stream Application Programming Interface (API) regarding the Ebola

virus. Results showcase that news media have greater topic coverage than Twitter. Also, the sentiment variety as well as the life span of coverage are shorter in Twitter than news media. News media may be labelled as more focused on event-related entities (e.g., person, organizations, locations), while Twitter is more time-oriented.

SM are being used extensively around the world and are a source of news, opinions and many more on a daily basis. Therefore, opportunities for an automated tool that identifies topics and user sentiment arise. A prototype tool that attempts to identify the pulse of Arabic users for a trending tool is introduced in (Rafea and Mostafa, 2013). Twitter data are used for extracting unigram words appearing more than 20 times in a given corpus. Then they are fed as features for grouping tweets with bisecting k-means clustering. Results show that the quality of identified topics reaches 72.5%.

Twitter is commonly used for the dissemination of online events that happen on real-time. Frequent Pattern mining was utilized to detect topics on Twitter data. In that case topics comprise groups of words; yet the possibilities for utility pattern generation are omitted. Such a method that attempts to detect emerging topics utilizing Utility Pattern mining along with Frequent Pattern mining is proposed in (Choi and Park, 2019). Tweets are grouped based on time windowing and a utility of words is defined based on the growth rate and frequency. Then, post-processing extracts topic patterns to be stored in a Topic-tree data structure. Experimental evaluation is conducted in three datasets, demonstrating better results (5% higher topic recall) and faster execution times compared to other topic detection techniques.

SM platforms such as Twitter generate sparse and noisy text that could be organized into an ontology containing multiple topics. The data can be processed in real-time generating new opportunities for an industrial application that deals with the topic modelling issue in a feasible and effective manner. Such a system providing functionalities like non-topical tweet detection, automatic labelled data acquisition, diagnostic and corrective learning and high precision topic inference is presented in (Yang *et al.*, 2014). Topic inference is achieved with a training algorithm for text classification. It uses a mechanism that associates text with external information sources with 93% precision and adequate topic coverage.

2.2.1 The COVID-19 Case

The COVID-19 pandemic impacts severely societies and people at the economic, psychological and health level. Governments, organizations, and individuals make use of SM attempting to mitigate this impact. There is a need to extract knowledge regarding content topics that emerge from SM platforms, to inform policy makers and health experts about public opinions and needs. A study utilized 2.8 million tweets identifying 12 topics that were clustered into four themes: i) the origin of the virus, ii) the source of the tweet, iii) its economic and society impact, and iv) mitigation of the risk infection options (Abd-Alrazaq *et al.*, 2020). Ten topics had positive sentiment, whilst two had negative, related to deaths and racism. The implications suggest that SM should be utilized to communicate useful health information to the public. Worldwide health systems should focus on disease detection, monitoring and surveillance systems that take advantage of the information generated by the SM.

On top of the societal and economic problems that COVID-19 has caused, there are also the psychological ones. Traditionally psychological evaluation requires time-consuming surveys that remain exposed to cognitive and sampling biases. At the same time, it cannot be conducted in large scale. In (Li, Chaudhary and Zhang, 2020) the authors propose an algorithm that combines a Correlation Explanation learning algorithm with a clinical Patient Health Questionnaire lexicon attempting to identify COVID-19 stress symptoms in the United States. The algorithm uses spatiotemporal data to overcome traditional topic detection models issues as well as ease severe complications arising from human intervention in DM. The experimental results show that there is a strong correlation of stress symptoms with the increasing number of virus cases for major US cities, such as Chicago, New York, Seattle etc. Also, public perceptions are susceptible to news and messages about COVID-19. During the first months of the pandemic, people started stockpiling resources due to the fear of infection, but after a few months concerns shifted, due to more long-term financial anxieties.

COVID-19 was first identified in December 2019 as a virus that causes pneumonia in China. SM and more specifically Sina Weibo acted as a place for the public to communicate and seek healthcare information about this crisis. A study reports on demographic and other findings associated with people who live in Wuhan, the city where COVID-19 was first identified and used Sina Weibo for help. These

include the median age, fever as the most common symptom, relying on family for help and more. The average illness onset time to seek help online was 10 days, while the average time from illness onset to reverse transcription-polymerase chain reaction testing (RT-PCR) was eight days. These findings attempted to highlight the benefits and usefulness that SM can offer in healthcare (Huang *et al.*, 2020).

COVID-19 is a trending topic on SM and more specifically on Twitter. Governments and policy makers take measures attempting to contain the virus spread. Authors in (Noor *et al.*, 2020) attempt to analyse and report on public reactions according to data extracted from Twitter. They utilized VOSviewer for extracting clusters of COVID-19 related tweet sentiment that form topics, labelled as public sentiments for i) USA, ii) Italy, Iran and a vaccine, iii) doomsday and science credibility, iv) India, v) COVID-19's emergence, vi) Philippines, and vii) US Intelligence Report. In addition, the most frequent itemsets were synonyms of COVID-19 while the most frequent and confident association rules involved words related with testing, lockdown, and China.

Lockdowns around the world were imposed against the pandemic. SM have been paramount for sharing information about this crisis. A mixed-methods analysis of tweets for the period of May 10 to May 24, 2020 was conducted utilizing the MAXQDA software along with the Twitter Application Programming Interface (API) for COVID-19 related data for New York. Content analysis unveiled primary topics from unstructured textual data, exposing six themes. These are: surveillance, prevention, treatments, testing and cure, symptoms and transmission, fear, and financial loss. Accessing public concerns in real-time during the pandemic exposes new fears regarding public health (Osakwe *et al.*, 2021).

Another work attempted to identify discourses regarding the pandemic and the corresponding implemented policies (Lopez, Vasu and Gallemore, 2020). Methods from NLP, Text Mining and Network Analysis tried to extract a corpus of tweets and exploit the most common replies. Also, these replies are evaluated regarding their time intervals. Therefore, information and misinformation transmitted during the start of the pandemic can be presented. The dataset utilized reports on entries from January 2020 when worldwide reported COVID-19 cases were less than 1000. The results of this study can support governments to make better decisions about future pandemics.

During the pandemic, SM has been crowded by COVID-19 relevant context. All SM platforms provide medical content, but Twitter has been the leader on medical relevant content. Therefore, emergency medicine has increased its user penetration since thousands of tweets per hour regarding COVID-19 are sent. Medical practitioners use Twitter to communicate their ideas, comments, and information. In (Rosenberg, Syed and Rezaie, 2020) authors attempt to deal with these free-flow messages and the possibility of generating harmful outcomes as they are not reviewed by experts. Also, it presents the benefits that the emergency medicine community gets from Twitter, as well as ways for the average emergency physician to take advantage of this opportunity.

According to the literature, there is a variety of attempts for topic and opinion extraction from Twitter with various problem applications, themes and use cases (healthcare, celebrities, crime, event detection etc.). In addition, there is a wide mix of methods used. For example, Page Rank (Cataldi, Di Caro and Schifanella, 2010; Zhao *et al.*, 2011), Affinity Propagation (Meng *et al.*, 2012), LDA and linear prediction model (Wang, Gerber and Brown, 2012), LDA with Gibbs Sampling and Continuous Wavelet Transformation (Cordeiro, 2012), LDA with Author-Topic Model (Alvarez-Melis and Saveski, 2016), k-means clustering (Rafea and Mostafa, 2013), Utility Pattern mining along with Frequent Pattern mining (Choi and Park, 2019), VOSviewer (Noor *et al.*, 2020) and MAXQDA software (Osakwe *et al.*, 2021). Moreover, the evaluation is performed in different datasets under different circumstances, rendering a comparative validation/analysis impossible. In general, topic extraction techniques generate wordsets that do not infer to topics in a clear manner. At the same time, words can often appear in multiple extracted topics. Topic extraction involving LDA frequently requires to predefine the number of topics, whilst enforcing a non-hierarchical analysis which does not allows data sharing among internal algorithmic iterations. Finally, most LDA related attempts address static topic extraction, i.e. topics do not evolve over time.

This thesis envisions an approach that enhances the output of any topic extraction technique by performing ARM on its output. It tackles the issues of: i) appointing the same words within tweets to multiple topics by replacing a probabilistic identification of strong words with a rule-based identification. ii) requiring a predefined number of topics by performing topic coherence and topic stability

analysis. iii) non-hierarchical analysis by considering unique wordsets in a uniform manner (throughout the whole dataset) by creating a single pool of words resulting from multiple LDA executions, retrieving the strongest of them for generating topics (using ARM on the output of LDA). iv) issues related with static topic extraction with no evolution of topics over time by predefining the period for retrieving topics related with COVID-19, but also not considering time as an investigated feature in line with the research design.

To the best of my knowledge, this approach is unique. For validating its results, LDA is utilized, a common topic extraction method, combined with ARM for identifying the strongest wordsets that form topics.

2.3 SOCIAL MEDIA SENTIMENT ANALYSIS

The foundations for Opinion Mining and Sentiment Analysis were laid by Pang and Lee (Pang and Lee, 2008). According to them, opinion mining, emotion analysis and / or subjectivity analysis are defined as: "the areas that deal with the computational processing of opinion, emotion and subjectivity in the text".

Sentiment Analysis is utilized for extracting knowledge from Internet and SM data (Paltoglou and Thelwall, 2010; Thelwall *et al.*, 2010; Thelwall, Buckley and Paltoglou, 2011, 2012; Thelwall, 2014, 2018). The SentiStrength tool analyses social web texts with human-level accuracy for English and other languages. The power of positive and negative emotion in short texts is calculated even in informal language. Two strong emotions are reported: -1 (negative) to -5 (extremely negative) and 1 (positive) to 5 (extremely positive). The reason it uses two scores is that according to psychology research (Berrios, Totterdell and Kellett, 2015), people process positive and negative emotions in parallel; hence the mixed emotions they may have. Alternatively, it can also report binary (positive / negative), ternary (positive / negative / neutral) and simple scale (-4 to +4) results.

Text sentiment analysis is also utilized to recognize opinions in microblogs. In (Korenek and Šimko, 2014), authors propose a methodology that combines sentiment analysis and appraisal theory. They aim to improve sentiment classification of text forming opinions, which are particularly connected with a specific entity.

Sentiment analysis may be improved by integrating identified emoticons and words into a two-layer graph. This approach enables the extraction of the top ranked

words to a sentiment lexicon utilizing a random walk algorithm. Graphical emoticons act as labels for a word-emoticon mutual reinforcement ranking model that is evaluated on microblog posts (Feng *et al.*, 2015).

Other works involving sentiment analysis on twitter data include the prediction chart position for songs (Tsiara and Tjortjis, 2020). Authors investigated the relation between the number of mentions of songs and artists along with their semantic orientation and the performance on charts. They employed regression analysis for calculating the difference between actual and predicted positions and moderated results. The focus of the predictions was on finding the top 5, 10 or 20 songs with best predictions featuring the top 20.

A Context-aware microblog sentiment classification attempt proposes a Context Attention based Long Short-Term Memory network. It considers microblog conversations as hierarchical sequences, allocating words and tweets, and applying a variety of weights through an attention mechanism (Feng *et al.*, 2019).

Sentiment analysis has also been widely used for predicting election outcomes. A hybrid method for sentiment analysis was utilized for election related tweets, combining Greek lexicons with classification methods (Beleveslis *et al.*, 2019). Probabilistic classification was applied, along with hashtag-based filtering analyses of public sentiment, related with events that took place during the pre-election period.

Moreover, a method for predicting the winner of the 2016 USA presidential elections focusing on observations for three key states was presented (Oikonomou and Tjortjis, 2018). Two methodological steps were employed: preprocessing and analysis of Twitter data to detect negative and positive sentiment towards the candidates, for accurately predicting the winner of the presidency.

SM sentiment analysis is often used to specifically extract negative opinions. These opinions may relate to market demands, enterprises or politics. A novel approach for extracting negative-sentiment-oriented features from SM data is presented in (Hsu, Hsu and Tseng, 2019). This approach utilizes text mining, machine learning and a web crawler for data collection.

2.3.1 The COVID-19 Case

COVID-19 research is one of the hottest trends since 2020. Multiple studies related with SM data and potential capabilities, such as forecasting outcomes have

been reported and categorized according to the application domain (e.g., healthcare, politics etc.) (Rousidis, Koukaras and Tjortjis, 2020). Other studies combined sentiment analysis, mainly using data from SM Platforms (SMPs) and COVID-19. The most representative ones related with this study are reported here. However, based on the conducted research there are no published similar efforts, which use sentiment analysis on data from Twitter in order to find potential links between the polarity of sentiments and the number of reported COVID-19 cases or deaths.

In (Abd-Alrazaq *et al.*, 2020), the authors retrieved the main topics related to COVID-19 posted on Twitter. From 2.8 million tweets, they identified around 167k related posts. According to their categorization, there were four main themes about COVID-19: i) its origin (location), ii) its source (causes that lead to its transmission to humans), iii) its impact on people and countries and iv) methods for controlling its spread.

Park et al. investigated news-sharing behaviour, along with information transmission networks from data related to COVID-19, gathered from 44k Korean Twitter users (Park, Park and Chong, 2020). They identified more than 78k relationships and found that the communication regarding COVID-19 amongst users was more frequent and faster. Additionally, they classified top news channels that were shared through tweets and concluded "that the spill over effect of the news articles that delivered medical information about COVID-19 was greater than that of news with nonmedical frames".

Cinelli et al. investigated the diffusion of COVID-19 information by using data from five different SMPs (Gab, Instagram, Reddit, Twitter, and YouTube) (Cinelli *et al.*, 2020). They gathered 1.3m posts in total, with 7.4m comments from 3.7m users for a 45-day period from 01/01/2020 until 14/02/2020. From these posts, 88.5% came from Twitter. 94.5% of comments and 85.7% of users came from YouTube. After analysing the spread of debatable information or even misinformation, they concluded that Gab is the SMP most prone to misinformation spread. Finally, they concluded that the channels of information dissemination and their contents depend on two factors: i) the SM itself and ii) the interaction patterns of groups of users that discuss the topic.

In a similar study, Singh et al. attempted to investigate the sharing of COVID-19 information and misinformation on Twitter (Barkur and Vibha, 2020). For a two-month period, from January 16, 2020 to March 15, 2020, they collected 2.8 million

tweets, along with 457k quotes and 18.2 million retweets. The language of tweets was predominately English (55.2%), followed by Spanish (12.5%) and French (7.4%). Overall, 32 languages were recognized. The countries that suffered the most during the investigated period, demonstrated an increase to COVID-19 related tweets compared to pre-COVID-19 use. They also provided a worldwide tweet geo-location distribution and compared it to the reported cases. The authors provided a summary of the themes identified using most frequent words in tweets and found that just 0.6% of the tweets are discussing myths or conspiracy theories.

Sentiment analysis on tweets from India after the lockdown was conducted using a dataset of 24k tweets for a 4-day period, from 25/03/2020 until 28/03/2020 (Lopez, Vasu and Gallemore, 2020). The most dominant stems were "consult", "manag" and "disast". Analysis was conducted by categorizing the sentiments via 10 words ("anger", "anticipation", "disgust", "fear", "joy", "negative", "positive", "sadness", "surprise" and "trust"). The most popular tweet sentiments were "positive", "trust" and "negative" with a count of 24k, 16k and 9.5k respectively.

Lopez et al. used Text Mining, Natural Language Processing and Network Analysis to investigate the perception of COVID-19 policies by mining a COVID-19 related multi-language Twitter dataset (Singh *et al.*, 2020). From January the 22nd to March the 13th 2020 (a 52-day period) they collected around 6.5 million tweets. 63.4% were written in English, 12.7% in Spanish/Castilian and the rest in 64 other languages. Extreme retweet bursts were observed in Europe in late February and early March. Finally, a geo-located distribution of 1,625 Tweets was provided.

Samuel et al. tried to identify public sentiment associated with the pandemic using COVID-19 related Tweets in the US, and the R statistical software (Samuel *et al.*, 2020). They downloaded Tweets from February to March 2020, a period when the pandemic hit the USA and by using Geo-Tagged Analytics, association with non-textual variables, Sentiment Analysis, and classification methods, they found that Naive Bayes performs better (91%) for sentiment classification of short (in length) COVID-19 tweets compared to logistic regression (74%). Better performance was also identified for long tweets, but with worse accuracy (57% and 52% respectively).

Finally, another study by Hamzah et al. introduced the CoronaTracker, a world-wide COVID-19 outbreak data analysis and prediction tool (Hamzah *et al.*, 2020). They utilized the Susceptible-Exposed-Infectious-Recovered (SEIR) predictive

modelling to forecast the COVID-19 outbreak based on daily observations. In their methodology they included sentiment analysis on articles from news (561 positive and 2,548 negative) to further understand public reaction towards the pandemic. They provided a word-tag cloud for both sentiments, and they displayed the top five positive and negative articles.

2.4 CONCEPTUAL FRAMEWORK AND IMPLICATIONS

This section distils key findings from literature to elaborate and present possible implications of the conducted research on the Social Media domain.

2.4.1 Introducing new Social Media Types

As Valentini and Kruckeberg (Valentini and Kruckeberg, 2012) stated: "Within this digital environment, it is extremely important to have a clear understanding of the meaning, use, and implication of new/digital and social media". Along with the rise of the number of Social Media (SM) and their users, the ambiguity of their features rises, too. According to the same study it is vital to distinguish digital technologies from their social functionality and to understand the SM use in order to evaluate user behaviour and attitudes. This study can aid researchers, SM users and professionals by facilitating i) SM selection, ii) identification of new trends and iii) collaborations and acquisitions.

i) Despite the fact that there is a clear preference over SM that users and professionals use (Lua, 2019); and with the top-10 SM having 500+ million users each, there is still some confusion over their role. This thesis aims at selecting the most popular and representative SM in terms of features, yet this selection is not exhaustive. Rideout (Rideout, 2015) demonstrated that teen SM users spend around seven hours per day using screen media, whilst three of these hours are spend in social networking websites. According to (Leiner *et al.*, 2018) "Social media pose serious challenges for uses-and-gratifications research, such as the entangled use of contemporary media services". There are indeed detailed features and characteristics for each SM, although many of them are overlapping, as they are similar. At the same time, there is a great number of volatile features and there are dissimilarities that may not seem to be so distinct; yet they create a chaotic environment that can confuse the users. The proposed categorization of SM might help the stakeholders to select the optimum SM that best meets their needs, since

50% of the respondents of Copp's survey agree that the need to personalize content and experiences is a major challenge (Copp, 2018). An appropriate SM selection can support and reinforce public communication activities and social connection.

ii) Teague mentions that around half of business marketers are still making up social media plans on the fly without proper marketing strategies, whilst most of them (~65%) are valuing likes, comments and shares as extremely important for their strategies (*Your 2019 Social Media Strategy: 4 Trends You Can't Ignore*, 2019). According to (Paterl, 2019; Goodwin, 2020) the new trends in SM for 2019 are: (1) Rebuilding trust in SM platforms, (2) Storytelling, (3) Building a brand narrative, (4) Quality and creativity over quantity, (5) Put Community and Socialization back in SM, (6) Influencers continue to grow their communities, (7) Selfies, videos and branding (LiveVideos, Vertical videos, Interactive videos, more smartphone-quality videos), (8) Earn, rebuild, or keep the trust of your followers, (9) Hyper-targeted personalization, and (10) Know your platforms. The proposed hybrid SMTs' conceptualization can facilitate the identification of new trends in the future, since they incorporate the features and suggest more functional, well-structured and up-to-date SM that marketers and researchers could use.

iii) There are constantly buyouts between SM platforms and applications. For instance, even back in 2014, around 26 billion USDs were spent during the seven most important buyouts in SM (Forrest, 2014): 1. Google buying YouTube for $1.65 billion, 2. Facebook buying Instagram for $1 billion, 3. Facebook buying WhatsApp for $19 billion, 4. Google buying Waze for $966 million, 5. Twitter buying Vine for $970 million, 6. Microsoft buying Yammer for $1.2 billion and 7. Yahoo buying Tumblr for $1.1 billion. Facebook for instance has acquired around 80 other companies (WikiPedia, 2019). Finally, index.co has accumulated the acquisitions in SM per year (*Top 50 2019's Acquisitions in Social Media - Index*, 2019). More than 423 billion USDs has been spent for approximately 700 acquisitions in SM. Therefore, this work, which documents features from more than 100 SM, classified and suggested new hybrid categories, can facilitate collaborations and acquisitions between SM. For instance, SM with complementary features can be merged or collaborate. Similarly, a popular SM

that lacks a specific feature, can acquire a SM with this distinct feature, like in the case of Facebook and WhatsApp.

2.4.2 Improving Topic Extraction on Social Media

The proposed topic extraction methodology enhances the extraction of insights regarding SM user opinions, attitudes, and discussions. These may refer to worldwide events such as a pandemic, and it could be implemented in a plethora of topics since there is not a predefined ontology or vocabulary. The proposed methodology utilizes as a use case the Twitter platform although it could be expanded to additional SM platforms. This would generate better prospects for a holistic compare and contrast analysis on topic extraction for multi-SM applications. For example, the proposed topic extraction methodology could be implemented for Facebook, Instagram, Pinterest and more.

During the COVID-19 pandemic SM were flashed with unstructured and unfiltered messages presenting opinions and ideas often resulting in negative outcomes and actions. It is usually due to the lack of a mechanism that allows the review of all this free flow information by experts. This highlights the need for a methodology/tool being able to identify and categorize these data flows under generic topics. At the same time, it may be useful to have a more precise overview of the discussions of topics in SM.

On the other hand, SM may also provide valuable medical relevant content. Policy makers such as governments or medical parties can take advantage of a more accurate topic extraction method to engage with the SM public opinion. With such a tool they could grasp in a more precise yet generalized manner the themes of SM discussions. Then if deemed necessary they can intervene by communicating with the public with information and guidelines from experts.

2.4.3 Discovering Correlations of Sentiment analysis with Exogenous Variables

Nowadays, knowledge extraction is also feasible from other sources such as SM, since they are characterised by high user penetration and great traffic regarding posted text. These online posts may be utilized for discovering correlations between attitudes of masses in times of need. A process with prospects for generating insights for worldwide events is Sentiment Analysis, calculating how positive or negative is the online text content.

In case of COVID-19, tools that enable timely tracking and alerting of the public may become very useful for tackling issues (economic, social etc.) related with worldwide events such as a pandemic. This thesis envisions such functionalities by identifying trends on Twitter and correlating them with COVID-19 posts. It also exposes valuable insights for predicting disease outbreaks through monitoring and evaluation of multivariable correlations. These are, text sentiment polarity with COVID-19 numbers of Cases and Deaths.

A tool that combines sentiment analysis with exogenous variables yields great prospects for mitigating drawbacks related with a health crisis. Exploitation parties involve healthcare/medical professionals, research community or governments retrieving useful indicators for psychological correlations resulting from SM. Also, SM posts offer new capabilities for retrieving or publishing information with health context to the communities, enhancing the online presence of public health policy makers (e.g. governments).

Implications of this part of the thesis, envision a system that makes decisions about a worldwide healthcare/medical crisis (such as COVID-19) (Koukaras, Rousidis and Tjortjis, 2020) utilizing online content. Numerous parameters could be retrieved such as population characteristics (e.g. age, gender), indexes (e.g. economic, regional), vaccination programs, government policies (e.g. lockdowns) and more, along with posts' sentiment analysis from multiple SM.

Chapter 3: Research Design

This chapter describes the design adopted by this research to achieve the aims and objectives stated in Sect. 1.3. Sect. 3.1 discusses the methodology of this study, and the stages by which the research design was implemented. Sect. 3.2 elaborates on the data sources and preprocessing steps. Sect. 3.3 describes how the data were analysed (methods/algorithms). Finally, Sect. 3.4 discusses the limitations, ethical considerations and potential threats to validity of the conducted research.

3.1 METHODOLOGY AND RESEARCH DESIGN

The research design involves a mixed quantitative and qualitative approach. Although most of the research tasks experimented incorporating algorithmic outputs and mathematical calculations (quantitative) for reporting on results, there are qualitative parts yielding non-calculable elements and attributes. For example, in Tasks 1 and 2 there are research design features that exposed qualitative characteristics (Table 2).

Table 2. PART I Types of primary research design per Research Task.

Task	Type	Description
1	Mixed	Primarily qualitative since SM utilities result from observations and grouping of official features per SM, yet also quantitative since the methodology involves hypotheses, rules and statistical inferences.
2	Mixed	Primarily quantitative yet also qualitative due to methodological inference to themes per topic.
3	Quantitative	There is hypothesis formation and significant statistical inferences.

The data collection methods involve a variety of primary data sources, highlighting the extended endeavour for data retrieval and procurement requirements of this study (Table 3). The detailed presentation of research data can be found in Sect. 3.2.

Table 3. PART I Data collection methods per Research Task.

Task	Method(s)	Description
1	Documents and Records, Observations	Observation and retrieval of SM utilities and generation of a dataset

2	Online Data crawling	Utilization of a crawler for grabbing tweets forming Twitter dataset
3	Online Data crawling	Utilization of a crawler for grabbing tweets forming Twitter dataset

3.1.1 Task 1

The proposed research process can be divided into seven steps (Fig. 2). A brief description of the proposed steps follows: Step 1 entails data collection to form a dataset of features from 112 SMPs. Step 2 combines preprocessing by data normalization, transformation and reduction along with missing values and duplicate removal. In Step 3, observations are recorded and finalize the dataset based on SM utilities. Step 4 defines the axioms to follow for enlisting and shifting between the proposed SMTs. Step 5 involves experiments by using: i) FP-Growth, an association rules algorithm in Experiment#1, and ii) three Clustering algorithms (DBSCAN, k-medoids, Random Clustering) in Experiment#2. Step 6 uses experimental results to propose a new SMTs taxonomy. Finally, Step 7 examines whether the proposed taxonomy is viable by testing the hypothesis and comparing results with related work.

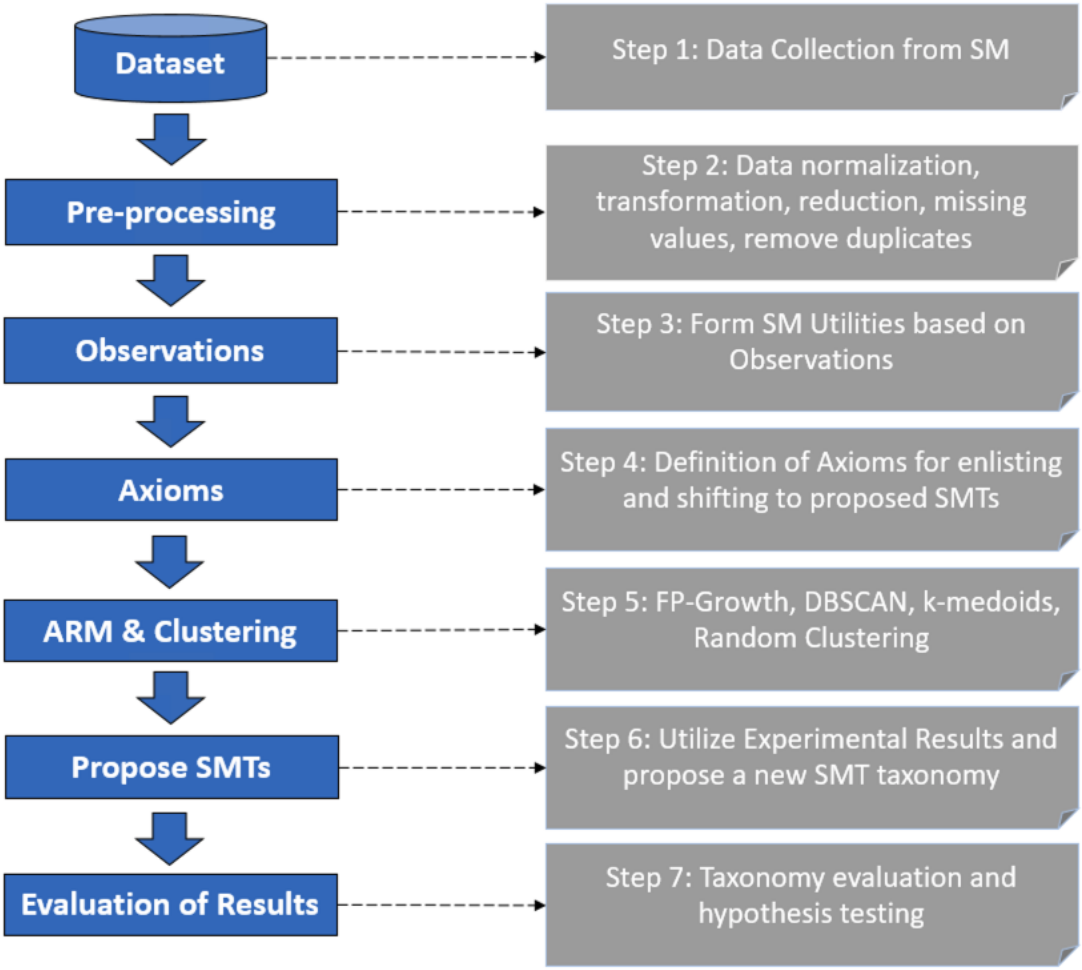

Fig. 2. Task 1 methodology flowchart.

Since it is implied that SMPs can form hybrid types based on their common Utilities, the effort is extended to introduce a new taxonomy. The process is a mixture of data-driven and hypothesis-based approaches emphasizing on the data-driven aspect, meaning that the feature dataset will be more decisive and act as a validator for the initial hypothesis H0 when forming the proposed taxonomy.

3.1.2 Task 2

This work utilizes Twitter data retrieved over a period of 153 days, from 27th of February 2020 up to 28th of August 2020, in a worldwide scale. It aims to provide insights about the pandemic, exposing capabilities related with topic extraction from text data. The proposed research methodology consists of four steps: i) Preprocessing, ii) Topic Extraction, iii) Association Rule Mining (ARM) and iv) Knowledge Extraction. Fig. 3 outlines the key components of this methodology, as well as the process flow for the methodology. Each of these steps are detailed in the following

sections. Knowledge extraction is elaborated separately as it culminates the results of this study.

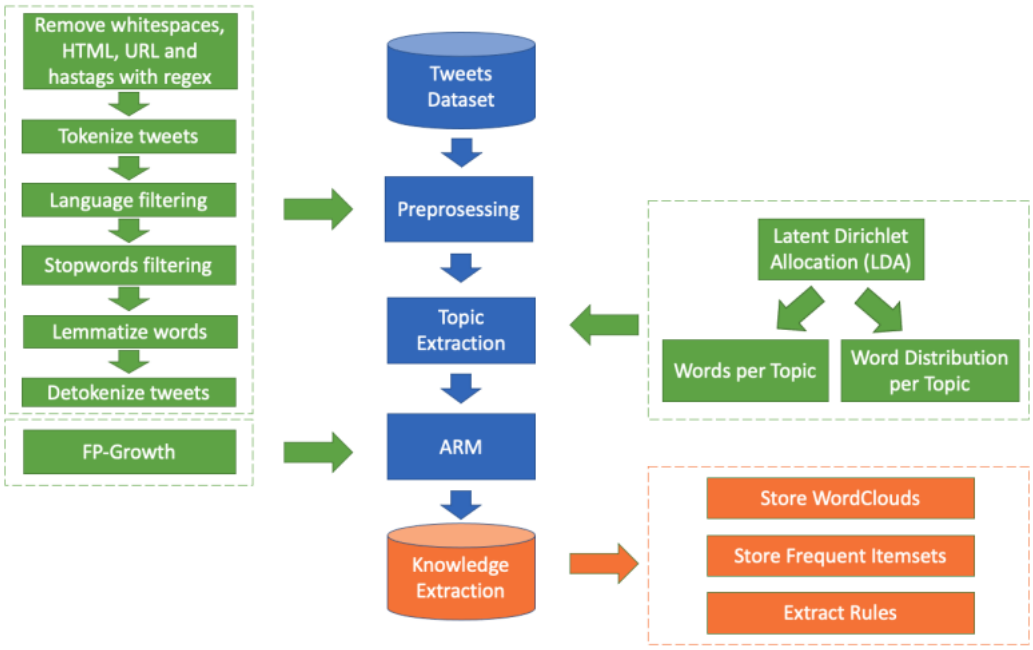

Fig. 3. Task 2 methodology flowchart.

3.1.3 Task 3

This thesis aims at calculating the polarization of around 2.1 million tweets that have been retrieved throughout the period from 27th of February 2020 until 28th of August 2020, a period of 153 days, and then analyse possible correlations between COVID-19 Cases and Deaths, as reported in a worldwide scale, for the same period. The proposed approach can be split into three contextual layers. Data Preprocessing and Polarization layer, the Hypotheses Formation layer and the Statistical Analysis layer are depicted in Fig. 4 along with the research flow. More details about each of these layers follow in the next section.

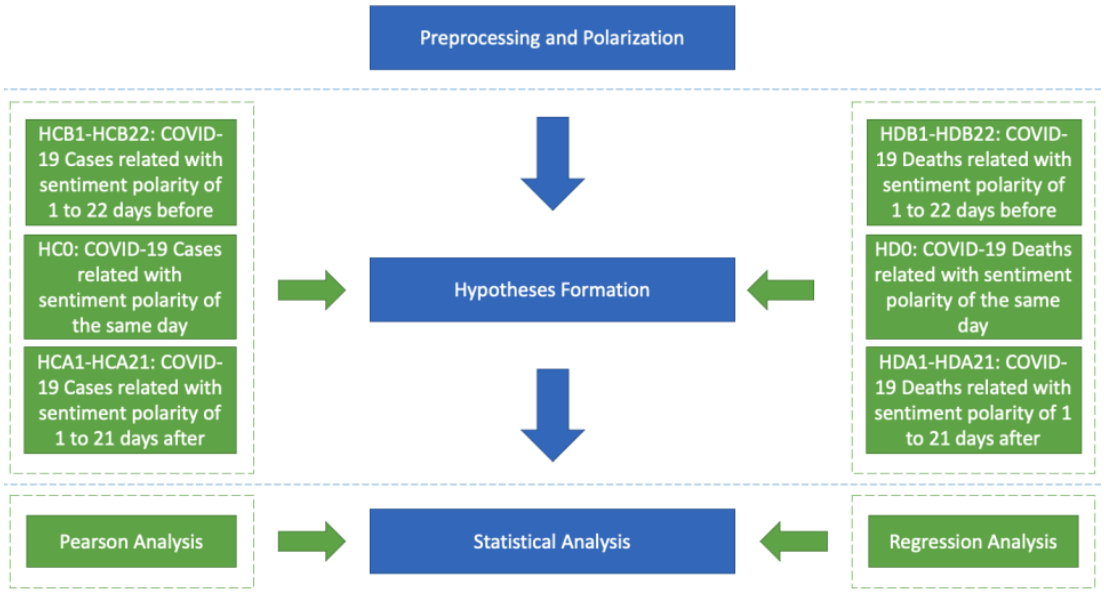

Fig. 4. Task 3 methodology flowchart.

3.2 DATA SOURCES

3.2.1 Task 1

The dataset used for the proposed methodology contains various SMPs; the choice is based on ranking regarding active monthly users, using the expanded and merged version of Table 5 and Table 6. A platform's user penetration is considered, as well as the variety of its official features, as the most important attributes when enlisting a candidate platform to the methodology. It is built and populated by data retrieved from the official sites of each of the reviewed 112 SMPs . Some platforms with smaller user penetration implement fewer features. Clearly the list is not exhaustive, given the volatile nature of SM popularity and feature base. Data preprocessing techniques are utilized such as removing duplicates and missing values, or data transformation and reduction as needed to normalize the research dataset (further explained in Observation#1 below). Having presented the most common SMTs in Sect. 2.1, Table 4 summarizes the top fifteen ranked SMPs with regards to active users (*Statistics for Social Networks: Top 15 Most Popular Social Networking Sites, Active users [2017] - DreamGrow*, 2017).

Table 4. SMP ranking by active users.

SMP	Number of active users (millions)
Facebook	2,010
YouTube	1,500

Instagram	800
Twitter	328
Reddit	250
Vine	200
Pinterest	175
Ask.fm	160
Tumblr	115
Flickr	112
Google+	111
LinkedIn	106
VK	95
ClassMates	57
Meetup	32

The mapping of features to Utilities is described step-by-step by Observations #1–4 below. All in all, each feature is examined, and grouped these logically, according to their semantic meaning in context. Each group was then labelled by a term, signifying the corresponding utility.

Observation#1 Mapping of platform features into Utilities takes place, using common sense, semantics and denotation forming Table 6, in line with similar research (Kaplan and Haenlein, 2010; Kietzmann *et al.*, 2011; Gundecha and Liu, 2012). This mapping is heuristic, not guaranteed to be the optimal, but it is suitable for practically appointing each feature (described by a word or a sentence) to a Utility. For example, Facebook, LinkedIn and VK implement the "Messaging" feature, which can be grouped under the Utility called "Connecting".

The most representative official features for SMPs are shown in Table 5 (data retrieved from the official documentation for each platform). Nowadays, the majority of SM support multimedia sharing, posting, hash-tagging features and more, under different feature labelling. An expanded form of the current standardized types is utilized, as used in (Kaplan and Haenlein, 2010; Kietzmann *et al.*, 2011; Gundecha and Liu, 2012), to assign relevant feature labels into conceptually compliant Utilities.

Table 5. Official features for the 15 top ranked SMPs.

SMP	Feature
Facebook	Friends, Fans, Wall, News Feed, Fan Pages, Groups, User Groups, Apps, Live Chat, Likes, Photos, Videos, Text, Polls, Links, Status, Pokes, Gifts, Games, Messaging, Classified Section, Upload and Download Options for Photos.

YouTube	Playback Upload Quality and Formats, Live Streaming, 3d Videos, 360° Videos, Post Text, Images (Including Gifs), Live Video (On Channel).
Instagram	Explore, Photographic Filters, Video, Photos, Instagram Direct, Instagram Stories, Monetization, Stand-Alone Apps, Third-Party Services.
Twitter	Tweet, Retweet, Direct Messaging, Follow People & Trending Topics, Links, Photos, Videos.
Reddit	Social News Aggregation, Web Content Rating, Discussion Website, Content Sharing, Links, Text Posts, Images, Voting.
Vine	Record short Video Clips, Ability to "Revine" Videos on a Personal Stream, Protected Posts.
Pinterest	Pins, Boards, Exploring, Following.
Ask.fm	Profiles, Send Each Other Questions.
Tumblr	Dashboard (Blog Posts), Queue, Tags, Html Editing, Messaging to Blogs, Questions.
Flickr	Accounts, Organization, Access Control, Interaction and Compatibility, Filtering, Licensing.
Google+	User Profiles, Circles, Stream, Identity Service, Privacy, +1 Button, Google+ Pages, Communities, Locations, What's Hot, Google Local, Photography, Additional Features, Collections, Deprecated Features.
LinkedIn	User Profile Network, Security and Technology, Messaging, Applications, External, Third Party Applications, Embedded In Profile, Mobile, Groups, Job Listings, Online Recruiting, Skills, Publishing Platform, Influencers, Advertising and for-Pay Research.
VK	Messaging, News, Communities, Like buttons, Privacy, Synchronization with other Social Networks, SMS Service.
ClassMates	Privacy, Post to and read Community Boards and view Information about upcoming Reunions, Emails.
Meetup	Groups, Members, Organize meetups.

Observation#2 A transformation of features takes place so that each attribute in the dataset represents a semantically equivalent specific Utility in the real-world. Examples: feature “Messaging” becomes “Connecting”, users exchange text, voice and/or video etc. which is a means for establishing social connections. Feature “Tags” becomes “Sharing”, feature “wall” becomes “Profile” etc.

Based on Observation#1 and Observation#2 fourteen distinct Utilities (Connecting, Sharing, Multimedia, Privacy, News, Promoting, Voting, Publishing, Schedule, Profile, Applications, Professional, Opinions, Entertainment) emerge that group up unique official SM features under a single conceptual label (Utility). Appendix A showcases the feature transformations for the complete dataset (112 SM sites).

Observation#3 By using the map in Appendix B and grouping features under the Utility label, it is observed that different SMPs utilize common Utility instances, as shown in Table 6.

Table 6. SMP grouping based on common Utility.

SMP	Utility	Number of SMPs (max=15)
Facebook, Instagram, Twitter, Reddit, Pinterest, Ask.fm, Tumblr, Flickr, Google+, LinkedIn, VK, Classmates, Meetup	Connecting	13
YouTube, Instagram, Twitter, Reddit, Vine, Pinterest, Ask.fm, Tumblr, Google+, Classmates	Sharing	10
Facebook, YouTube, Instagram, Twitter, Reddit, Vine, Pinterest, Flickr, Google+	Multimedia	9
Facebook, Flickr, Google+, LinkedIn, VK, Classmates	Privacy	6
Facebook, Twitter, Reddit, Pinterest, VK	News	5
Facebook, Twitter, Ask.fm, LinkedIn,	Promoting	4
Facebook, Reddit, Google+, VK	Voting	4
Tumblr, Google+, LinkedIn	Publishing	3
Flickr, Classmates, Meetup	Schedule	3
Facebook, Google+, LinkedIn	Profile	3
Facebook, Instagram, LinkedIn	Applications	3
Instagram, LinkedIn	Professional	2
Facebook	Opinions	1
Facebook	Entertainment	1

Observation#4 By further observing Observation#3 and Table 6 it is inferred that various hybrid SMTs can be formed, characterized by specific Utilities. For example, hybrid type#1 [Pinterest, Reddit, Facebook, Twitter] that characterizes SMPs that offer News, Multimedia, and Connecting capabilities, hybrid type#2 [Instagram, LinkedIn] that offers Professional, Connecting and Application capabilities.

Table 7. Fraction of each Utility in dataset.

Utility	Absolute count c	Fraction of dataset
Connecting	85	0.2178
Multimedia	78	0.2
Professional	50	0.1282
Sharing	35	0.0897
Entertainment	28	0.0717
Opinions	21	0.0538
Profile	17	0.0435
Publishing	17	0.0435
Applications	14	0.0358
Schedule	12	0.0307
Privacy	11	0.0282
Voting	9	0.0230
News	7	0.0179

Promoting	6	0.0153

This section records observations from the dataset regarding 112 SM. Table 7 shows the absolute count (c) of occurrences of each Utility, along with the proportion of c as a fraction of c over the total number of Utility occurrences in the dataset.

Table 8. Top 15 SMPs with their Utilities.

Facebook Utility	Utility
Facebook	Connecting, Profile, News, Promoting, Applications, Voting, Multimedia, Opinions, Entertainment, Privacy
YouTube	Multimedia, Sharing
Instagram	Connecting, Applications, Multimedia, Sharing, Professional
Twitter	Connecting, News, Multimedia, Sharing
Reddit	Connecting, News, Voting, Multimedia, Sharing
Vine	Multimedia, Sharing
Pinterest	Connecting, News, Multimedia, Sharing
Ask.fm	Connecting, Promoting, Sharing
Tumblr	Connecting, Sharing, Publishing
Flickr	Connecting, Multimedia, Schedule, Privacy
Google+	Connecting, Profile, Voting, Multimedia, Privacy, Sharing, Publishing
LinkedIn	Connecting, Profile, Promoting, Applications, Privacy, Professional, Publishing
VK	Connecting, News, Voting, Privacy
Classmates	Connecting, Privacy, Sharing, Schedule
Meetup	Connecting, Schedule

Appendix C shows the complete set of Utility occurrences for each SM whilst Table 8 summarizes the utilities of the top fifteen SMPs. Using Appendix C, the research effort is extended to support H0 with the inception of generalized axioms for enlisting and shifting between the Proposed Social Media Types (taxonomy) as follows:

i) Axiom 1 (A1): Primary Utility (P) for each SM platform is its Utility with the highest count of occurrences, c.

ii) Axiom 2 (A2): Secondary Utility (S) for each SM platform is its Utility with the second highest count of occurrences, c.

iii) Axiom 3 (A3): Trivia Utility (T) for each SM platform is its Utility with the lowest count of occurrences, c.

iv) Axiom 4 (A4): If there is a tie in calculating P among two or more Utilities in a SM entry, $\sum_1^c P$ utilities are considered.

v) Axiom 5 (A5): If there is a tie in calculating S among two or more Utilities in a SM entry, $\sum_1^c P$ utilities are considered.

vi) Axiom 6 (A6): When none of A1–A5 apply, a platform is categorized by its official goals.

Based on axioms A1–A6 and the dataset observations, each of the proposed SMT is characterized by Primary, Secondary, and Trivia Utilities, as presented in Appendix D.

Some examples of applying the rules to the top populated SM are presented in Table 9 and Table 10.

Table 9. Facebook break-down of Utility occurrences.

Facebook Utility	Count	Type
Connecting	7	Primary
Multimedia	4	Secondary
Professional	-	-
Sharing	-	-
Entertainment	1	Trivia
Opinions	1	Trivia
Profile	1	Trivia
Publishing	-	-
Applications	1	Trivia
Schedule	-	-
Privacy	1	Trivia
Voting	1	Trivia
News	2	Trivia
Promoting	2	Trivia

For further clarification of the mapping process, Appendix C appoints the features to Utilities, thus Table 9 refers to Facebook and counts seven occurrences of Connecting since its seven features: Fans, Groups, Live Chat, Pokes, Gifts, Messaging, User Groups are grouped under the Utility Connecting (refer to Observation#1).

Table 10. YouTube break-down of Utility occurrences.

Facebook Utility	Count	Type
Connecting	-	-

Multimedia	6	Primary
Professional	-	-
Sharing	1	Secondary
Entertainment	-	-
Opinions	-	-
Profile	-	-
Publishing	-	-
Applications	-	-
Schedule	-	-
Privacy	-	-
Voting	-	-
News	-	-
Promoting	-	-

On the same context, in Table 10 YouTube scores one on Sharing since the feature "Post Text" is semantically linked with the Utility "Sharing". Having examined Appendix C and Appendix D, the research effort is extended trying to prove H0 by mining the dataset using RapidMiner.

3.2.2 Task 2

The dataset for this research task was gathered using a crawler to retrieve tweets from Twitter's search functionality[1]. The search keywords included COVID-19 common synonyms such as "coronavirus", "covid", "covid-19" and "corona". The tweets under scrutiny date from 27/2/2020 until 28/8/2020, summing up to 2,146,243 unique tweets. The tweets have been filtered so that they only contain English text. Although crawling bypasses some of Twitter API's[2] drawbacks, such as the max number of retrieved tweets, it generates other issues like the need for better text preprocessing.

The implemented method crawled tweets including nine labels. These labels can be used for future expansion of this work. For the purposes of this research, the label "text" is utilized, marked with bold in Table 11, and the rest are discarded. This is because there was no research intention to associate tweets with users or with datetimes or tweet post counts. The main concern is to extract themes or topics for the investigated period of 153 days in a uniform manner.

[1] https://twitter.com/explore
[2] https://developer.twitter.com

Table 11. Tweet description.

No	Attribute	Description
1	Username	Twitter account username.
2	Tweet_id	Unique id of the tweet.
3	**Text**	The tweet's text.
4	Tweet_url	The tweet's URL.
5	Retweet_n	The sum of times the tweet is retweeted.
6	Like_n	The sum of times the tweet is liked.
7	Reply_n	How many replies the tweet has.
8	Datetime	The datetime of the tweet.
9	User_id	Unique id of the user.

Data preprocessing is a step that typically involves data transformations prior to analysis. This work deals with text preprocessing in English, in order to prepare the data for topic extraction and ARM. The employed sub-processes of preprocessing include incorporation of regex for removing whitespaces, HTML, URL and hashtag elements, Language text filtering (English), stopword removal, tokenization, and detokenization.

More specifically, regular expressions are employed to remove whitespaces, HTML, URL and hashtags with the re library (Regular expression operations[3]). The tokenization of text is employed by utilizing the nltk.tokenize package[4]. Next, stopword removal is implemented utilizing in a sequential manner the build-in English lexicon versions from three libraries/packages/modules spacy[5], genism[6] and nltk.corpus.stopwords[7]. The reason for doing so, is to increase the overall stopword lexicon without having to manually append words. To perform a morphological analysis of the words, lemmatization from nltk.stem.wordnet[8] is utilized. That way the lemma of the words in most of the cases is extracted. Detokenization takes place by incorporating the Penn Treebank detokenization implementation from nltk.tokenize.treebank[9]. After preprocessing the tweets take the form shown in Table

[3] https://docs.python.org/3/library/re.html
[4] https://www.nltk.org/_modules/nltk/tokenize.html
[5] https://spacy.io/
[6] https://pypi.org/project/gensim/
[7] https://www.nltk.org/_modules/nltk/corpus.html
[8] http://www.nltk.org/_modules/nltk/stem/wordnet.html#WordNetLemmatizer
[9] https://www.nltk.org/_modules/nltk/tokenize/treebank.html

12 while the overall usable tweets were reduced from 2.146.243 to 2.062.864 (96,12% of the initial dataset).

Table 12. Tweet Examples of original and pre-processed tweets.

Original Tweet	Pre-processed tweet
Thank you for taking the time to consume, digest, and distill this information for all of us. Just want to give you some positive reinforcement from a fan of Boomers and ultimately history and proven experience. #knowledge #coronavirus	thank take time consume digest distill information want positive reinforcement fan boomers ultimately history prove experience
No I am more worried about the 1% fatality rate of the #coronavirus	worry fatality rate
Which zombie apocalypse film does our government's response to #Coronavirus most closely resemble?	zombie apocalypse film governments response closely resemble
# Coronavirus Vaccine 'at Least a Year' Away, Health Official Says #USA #Republicans President Donald #Trump told reporters we were "very close" to a coronavirus vaccine, causing confusion as to the state of a vaccine. https://www. newsweek.com/anthony-fauci- coronavirus-vaccine-year-away-public-availability-1489214	vaccine year away health official say president donald tell reporters close coronavirus vaccine cause confusion state vaccine
So let me get this straight, world governments can't stop #coronavirus from spreading but if we pay more tax we can change the the planets temperature #ClimateChangeHoax	let straight world governments stop spread pay tax change planets temperature
The #coronavirus is not a time for politics. #COVID19	time politics
"Radiologists understanding of clinical and chest CT imaging features of coronavirus disease 2019 (COVID-19) will help to detect the infection early and assess the disease course. https:// pubs.rsna.org/doi/10.1148/radiol.2020200490 #Corona #coronavirus #CoronaOutbreak #COVID19 #COVID-19	radiologists understand clinical chest ct image feature coronavirus disease covid help detect infection early assess disease course
Coronovirus could be answer we've been looking for to tame medicare costs and prevent the collapse of #SocialSecurity #coronavirususa #coronavirus #CoronavirusOutbreak #CoronaVirusUpdates #boomers pic.twitter.com/EX5sQ6XXX7	coronovirus answer weve look tame medicare cost prevent collapse
'It's post-apocalyptic': how #coronavirus has altered day-to-day life https://www. theguardian.com/world/2020/feb/21/post-apocalyptic-how-coronavirus-has-altered-day-to-day-life-wuhan-north-england?CMP=share_btn_tw	postapocalyptic alter daytoday life

3.2.3 Task 3

The same Twitter dataset and preprocessing steps (Fig. 5) with the previous research task (Sect. 3.2.2) were utilized, making the necessary adjustments (choosing different data properties) according to this task's research design. The data contain

nine properties per tweet, as shown in Table 11. Just the attributes text and datetime were utilized. The rest of the attributes (e.g username, tweet_id or user_id) were discarded, since they offer no added value to this approach.

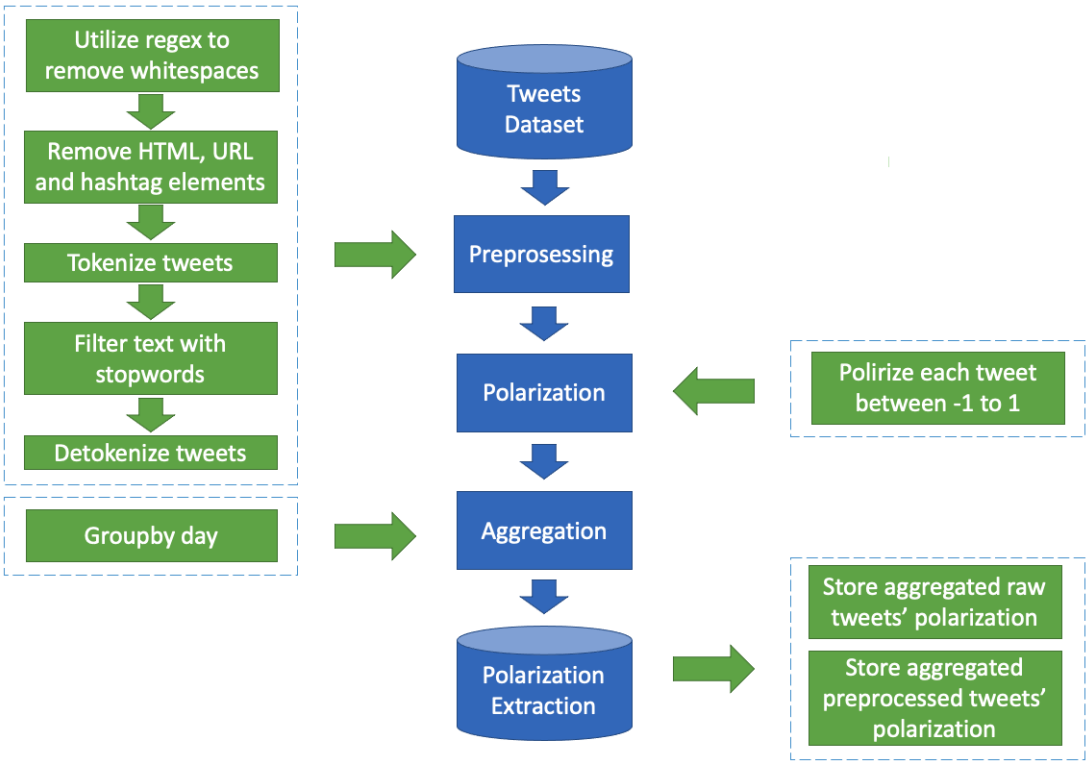

Fig. 5. Task 3 dataset preprocessing.

For sentiment polarization the TextBlob[10] library was incorporated, which retrieves values per tweet rated between -1 (most negative) to +1 (most positive). Table 13 illustrates examples of raw and processed tweets and their retrieved polarity.

Table 13. Examples of polarity values of raw and processed tweets.

Date	Raw tweet	Raw tweet polarity	Processed tweet	Processed tweet polarity
14-03-2020	...It's very difficult to get ANY news on the #coronavirus in #China - and what is happening there now?	-0.65	difficult news happening	-0.5

[10] https://textblob.readthedocs.io/en/dev/

23-03-2020	So what's going to be your newspaper of choice for when the time comes? Stay safe, stay isolated & save lives #Coronavirus #Healthcare #QBAIN https://lnkd.in/dzViDaX	+0.5	whats going newspaper choice time comes stay safe stay isolated save lives	+0.5
26-03-2020	Looking for peace of mind in these uncertain times because of the #Coronavirus ?! We're offering #mobile #patrols to businesses in #Cardiff , #Newport , #Bristol and around- such a cost effective #security #solution which could save you £££ from theft and vandalism. pic.twitter.com/idA1Xe4wZl	+0.3	looking peace mind uncertain times offering businesses cost effective save theft vandalism	+0.6
01-04-2020	Stranded at sea: Coast Guard says cruise ships must stay offshore with sick onboard. https://krld.radio.com/articles/ap-news/coast-guard-cruise-ships-must-stay-at-sea-with-sick-onboard … via @KRLD #CruiseShips #COVID19 #coronavirus #Florida	-0.7142	stranded sea coast guard says cruise ships stay offshore sick onboard	-0.7142

Finally, the sentiment polarization values of all tweets were grouped by day, retrieving their average value. These text transformations were executed in a sequential manner attempting to improve the accuracy of sentiment polarity reporting.

3.3 METHODS & ANALYSIS

This section discusses how the data were processed and analysed. It offers enough detail for the readers to replicate the analysis. Also, it justifies the reasoning for choosing specific algorithms/methods. Table 14 outlines all methods and algorithms utilized in this study and categorizes them to (Mining) Tasks.

Table 14. PART I Outline of Mining Tasks and Methods.

Research Task	(Mining) Task	Method

1	Clustering, Association Rule Mining	DBSCAN, k-medoids, Random-Clustering, FP-Growth, Association Rules Mining Metrics (Support, Confidence)
2	Topic Extraction, Association Rule Mining	Latent Dirichlet Allocation, FP-Growth, Association Rule Mining Metrics (Support, Confidence, Lift, Leverage)
3	Statistical Analysis, Hypothesis formation	Pearson correlation, Regression Analysis (p-values), Hypothesis formation

The following sections offer a presentation of methods per (Mining) Task with brief descriptions of key algorithms, as well as details about the methods employed for the experiments. There are numerous data mining functions to choose from, implemented by a variety of algorithms (Kanellopoulos *et al.*, 2011; Yakhchi *et al.*, 2017). RapidMiner[11] (Rittho *et al.*, 2001) was employed for Research Task 1 experimentation (SMTs), because it contains all the algorithms utilized for the experiments. For Research Tasks 2 and 3 (SMTE and SMSA) a Python implementation was employed utilizing the necessary packages/libraries.

3.3.1 Clustering

Clustering is an unsupervised learning method, which creates groups from datasets that consist of objects or entities that are characterized by similar or identical attribute values but are adequately different from entities that belong to other clusters (Kanellopoulos *et al.*, 2011). For running a clustering algorithm, the distance measure (e.g., Euclidean, Manhattan, Jaccard, Cosine distances) (Choi, Cha and Tappert, 2009) needs to be specified. After that, clustering methods often continue with the process of object selection and a method for evaluating the results (Jain, 2010). For evaluation quality measures can be used like cohesiveness (measure for object-to-object distance), separateness (measure for cluster-to-cluster distance) and silhouette index (mix of cohesiveness and separateness) (Zafarani, Abbasi and Liu, 2014).

DBSCAN

This clustering method was incorporated in Research Task 1 (SMTs). Density-based spatial clustering of applications with noise (DBSCAN) (Ester *et al.*, 1996) is

[11] RapidMiner is a software suite that provides an integrated environment for data preparation, machine learning (Bishop, 2007), deep learning, text mining (Aggarwal and Zhai, 2012), and predictive analytics. It supports all steps of the data mining process including data preparation, results visualization, model validation and optimization.

density-based, meaning that given a set of points in some space, it tries to group together points that are packed together, labelling outlying points that are alone in low-density regions. It functions on three abstract steps (Schubert *et al.*, 2017):

i) Find the points in the ε (eps) neighbourhood of every point and identify the core points with number of neighbours more than minPts.

ii) Find the components that are connected with core points on the neighbouring graph, without taking into consideration non-core points.

iii) Assign every non-core point to a nearby cluster if the cluster is an ε (eps) neighbour, else assign it to noise.

For the RapidMiner (Rittho *et al.*, 2001) implementation of this algorithm, parameters used are: epsilon = 1: (Range:real; 0.0±∞; default:1), which specifies the size of the neighborhood and min points= 5: (Range:integer; 1±∞; default:5), which specifies the minimum number of points forming a cluster. As for measure types, there are four options: Mixed Measures, Nominal Measures, Numerical Measures and Bregman Divergences. The last two cannot be used since the dataset does not contain numerical attributes. So, out of the remaining two groups of measure types Mixed Measures, and specifically the Mixed Euclidean Distance for two reasons were chosen: i) Nominal Measures contain, Nominal Distance, Dice Similarity, Jaccard Similarity, Kulczynski Similarity, RogersTanimoto Similarity, RussellRao Similarity and Simple Matching Similarity which all form two clusters with no reasonable results except from Nominal Distance. which produces exactly the same results as Mixed Euclidean Distance, and ii) according to RapidMiner user statistics, 79% of users utilize the Mixed Euclidean Distance measure which in this case outperforms the rest of the measures.

k-medoids & k-means

k-medoids is a clustering algorithm related to k-means (Arthur and Vassilvitskii, 2007) and the medoidshift algorithm (Kaufman and Rousseeuw, 2002). k-medoids was used in Research Task 1 (SMTs). Both k-means and k-medoids partition the dataset, and attempt to minimize the distance between points labelled to belong to a cluster and a point designated as the epicentre of the cluster. Running k-medoids in RapidMiner the following default parameter values were used: max runs10, max optimization step 100. Other values were also set, but they produced the same or poorer results.

Regarding the measure type, Mixed Euclidean Distance was used, for the reasons explained in the DBSCAN section.

Random-Clustering

This clustering method was incorporated in Research Task 1 (SMTs). Random-Clustering (Sibuya, 1993) generates simple and uniform random partitions. It has a single parameter controlling the partition of a random permutation into its cycles. The limit distribution of the size index of the generated partition is the join of the independent Poisson distributions with means determined by the size and the parameter. As for RapidMiner's parameters, in this algorithm the only one required is the number of clusters to be formed.

3.3.2 Topic Extraction

This mining task was employed for Research Task 2 (SMTE). To perform topic extraction, document (tweet) clustering needs to be employed. This process allows for the abstraction and analysis of lots of data. Once the tweets are clustered, new tweets can be allocated to existent clusters (topics) based on a text similarity metric. Latent Dirichlet Allocation (LDA) (Blei, Ng and Jordan, 2002) is a topic modelling algorithm that stores and identifies topics and text distribution from a pool of existent text documents, such as tweets. Therefore, once tweet preprocessing concludes, LDA can extract topics. The LDA steps are:

i) Set a number k of topics to be identified.

ii) Place randomly each tweet word in one temporal topic.

iii) Iterate, processing all tweets and words computing a) the probability the currently indexed tweet to be appointed to a specific topic according to how many of the words of this tweet are already appointed to the same topic as the currently indexed word and b) the percentage of the tweets appointed to the same topic as the currently indexed word.

Steps 1-3 are executed n times, with n being a predefined value before the algorithm commences. Then, each tweet is appointed to a unique topic according to its

most dominant words. For LDA parameter tuning the default parameters of scikit-learn[12] were utilized. For n, 10 iterations were set, the default parameter value.

The value for k is usually set by calculating topic coherence for a variety of topic numbers. HDP-LDA can be used instead of LDA, despite both being subject to misinterpretation (Wang, Paisley and Blei, 2011). LDA requires user specified k, while HDP-LDA operates with an unbounded number k, defined by data. For the purposes of this study, multiple LDA models were created with different k values, utilizing the pre-processed tweets dataset (Sect. 3.2.2). Jaccard similarity (Jaccard, 1912) was employed to perform a topic stability analysis (Greene, O'Callaghan and Cunningham, 2014), but also performed topic coherence analysis (Röder, Both and Hinneburg, 2015). This combinatorial approach is exploited presenting results in the form of two metrics. This process acts as an indicator for identifying semantic similarity across high scoring words, within all possible k number of topics. According to Fig. 6, the optimal number of topics k=10 is derived from the maximum difference between coherence and similarity.

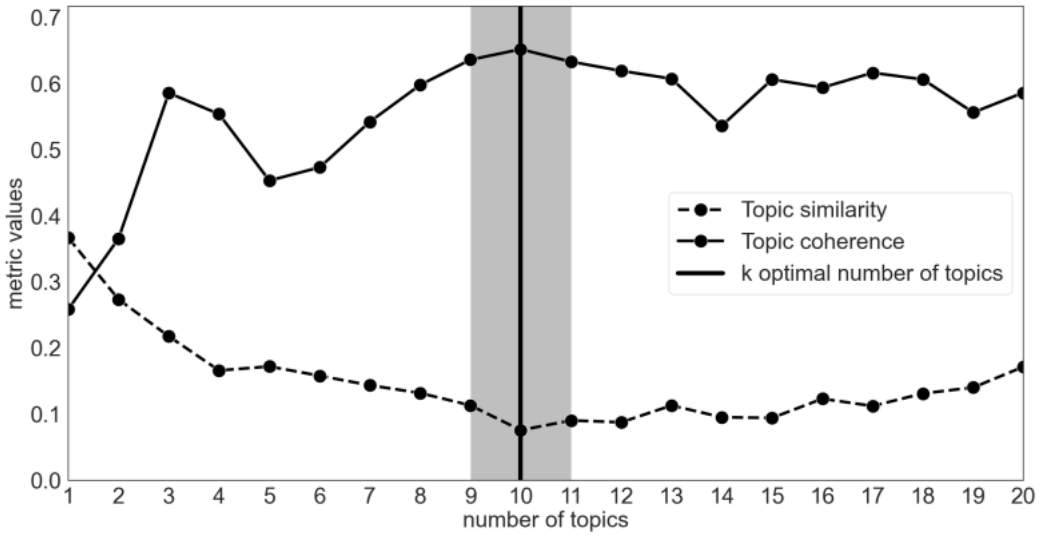

Fig. 6. Similarity, coherence and optimal number of topics.

To sum up, first LDA is incorporated to the proposed methodology to extract topics. This allows a high-level analysis that narrows down public attitudes during the COVID-19 epidemic period under scrutiny (27/2/2020 up to 28/8/2020). That way, strong keywords were extracted with a commonly used method (LDA) representing

[12] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html

public attitudes/opinion. Next ARM is performed on these strong words attempting to enhance the topic extraction process while mitigating issues of the LDA approach.

3.3.3 Association Rule Mining

Association Rule Mining (ARM) is a Data Mining (DM) technique that allows the discovery of relationships between variables in databases. In general, ARM refers to items and itemsets (sets of items) and associations among them. This thesis utilizes two notations of ARM one for Research Task 1 (SMTs) and one for Research Task 2 (SMTE). This section describes ARM based on the latter, since the core notation is the same. Therefore, Research Task 2 (SMTE) customizes and employs ARM notation and terminology to fit the SM domain. It substitutes transactions with tweets, items with words and itemsets with wordsets (i.e., sets of words). Thus, it searches and identifies the strongest word rules within tweets by using metrics such as support and confidence (Piatetsky-Shapiro, 1991). There is a variety of ARM algorithms to choose from, such as Apriori (Agrawal and Srikant, 2013) or FP-Growth (Li and Deng, 2007) that require to manually set the minimum support levels for extracting frequent wordsets. Other implementations, such as ARMICA (Yakhchi *et al.*, 2017), utilize techniques like the heuristic Imperialism Competitive Algorithm (ICA) to extract frequent wordsets without the need to specify support levels.

For the purposes of this thesis, four measures for extracting interesting word rules from tweets were utilized, minimum support, confidence as well as lift and leverage.

Support

Let $I = \{i1, i2, i3 \dots , in\}$ be a set of n binary attributes called words. Let $P = \{t1, t2, t3 \dots , tn\}$ be a set of transactions called the pool of tweets. Each transaction in P has a unique id and within it resides a subset of the words in I. A generated rule is a suggestion of the form $X \Rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = \emptyset$. In that case the wordset X is the antecedent or left-hand-side (LHS) and Y the consequent, or right-hand-side (RHS) of the rule (Hipp, Güntzer and Nakhaeizadeh, 2000).

For selecting the most interesting rules from a set of rules for a pool of tweets, thresholds regarding significance and interestingness are appointed. The metrics to set these thresholds are minimum support and minimum confidence, respectively.

Support $Supp\ (X)$ of a wordset X is the proportion of the tweets within a dataset that contains that set of words (Hahsler, Grün and Hornik, 2005).

Confidence

Confidence is an estimate of the probability of identifying words in the RHS of a rule within the tweet given that these words also satisfy the LHS. Confidence is essentially a metric that defines whether a rule is true and its frequency. It is given by formula (1) (Hahsler, Grün and Hornik, 2005):

$$Conf(\mathrm{X} \Rightarrow \mathrm{Y}) = Supp(\mathrm{X} \cup \mathrm{Y}) \,/\, Supp(X) \quad (1)$$

Lift

Lift calculates the degree of communality of antecedent and consequent wordsets that appear concurrently in case they are statistically independent. Lift values range between [0, ∞]. When lift is equal to 1, the wordsets (antecedent and consequent) are not related. A value between (0, 1) suggests a negative association, while values ranging between [1, ∞] indicate even greater association as the value increases (2) (Brin *et al.*, 1997).

$$Lift(\mathrm{X} \Rightarrow \mathrm{Y}) = Conf(\mathrm{X} \Rightarrow \mathrm{Y}) \,/\, Supp(Y) \quad (2)$$

Leverage

Leverage calculates how different is the probability of a rule compared to the probability in case the antecedent and consequent wordsets are statistically independent. Its value ranges between [-1, 1]. When the value is equal to 0 this indicates that the wordsets are independent. The greater positive value of leverage suggests an even stronger positive relation, while more negative values indicate stronger negative relations of the wordsets (antecedent and consequent) (3) (Piatetsky-Shapiro, 1991).

$$Leverage(\mathrm{X} \Rightarrow \mathrm{Y}) = Supp(\mathrm{X} \Rightarrow \mathrm{Y}) - Supp(\mathrm{X}) * Supp(Y) \quad (3)$$

FP-Growth

FP-Growth was utilized for Research Tasks 1 (SMT) and 2 (SMTE). It is an exhaustive ARM algorithm able to produce the same results as Apriori, yet it is much faster (Han, Kamber and Pei, 2012). That is a positive characteristic, since the experimentation involves thousands of unique words inside millions of tweets. FP-Growth finds how many times words appear in the dataset of tweets and places them

to a header table. A FP-tree structure is constructed through the insertion of these instances. The words in each of the instances get sorted in a descending order, based on the frequency of appearance in the dataset, enabling a faster tree traversal. At this point a threshold for traversal is set and all words that do not match the required conditions are discarded. That way large wordsets can be constructed faster; by processing the compressed dataset version in a recursive manner, without creating candidate words and validating them throughout the full dataset. The recursive process concludes finding the longest sets of words pertaining the minimum coverage and the generation of association rules starts (Han, Pei and Yin, 2000).

3.3.4 Hypothesis Formation

This approach was extensively utilized for Research Task 3 (SMSA). In order to help with all hypothesis statements for hypothesis formation, two terms are introduced. The term "before" refers to tweets that prelude the actual COVID cases or deaths, whilst the term "after" refers to tweets that follow the actual COVID cases or deaths. It should also be noted that the aggregated number of cases and deaths always increase during the pandemic period under scrutiny (27/2/2020 to 28/8/2020). For that reason, the hypotheses associate COVID-19 daily cases and deaths as retrieved from the WHO[13] database on a worldwide scale.

Therefore, 44 null hypotheses were tested (HCB1-HCB22, HC0, HCA1-HCA21) related with COVID-19 cases in total; indicatively, three are presented, since the rest are generated in the same manner: one with polarization lead (tweets before cases), one with no lead or lag (tweets in the same day with cases) and one with a polarization lag (tweets after cases). For instance, "three days before" means that tweets published at 4/3/2020 are matched with the number of new cases and deaths three days later 7/3/2020. "Three days after" means that tweets published at 4/3/2020 are matched with the number of new cases and deaths three days earlier at 1/3/2020.

HCB22: The number of cases on a day is not correlated with a higher on average positive sentiment polarity 22 days before.

HC0: The number of cases on a day is not correlated with a higher on average positive sentiment polarity during the same day.

[13] https://covid19.who.int/

HCA21: The number of cases on a day is not correlated with a higher on average positive sentiment polarity 21 days after.

Similarly, 44 null hypotheses were tested (HDB1-HDB22, HD0, HDA1-HDA21) related with COVID-19 deaths in total; three are presented: one with polarization lead (tweets before deaths), one with no lead or lag (tweets at the same day with deaths) and one with a polarization lag (tweets after deaths).

HDB22: The number of deaths on a day is not correlated with a higher on average positive sentiment polarity 22 days before.

HD0: The number of deaths on a day is not correlated with a higher on average positive sentiment polarity during the same day.

HDA21: The number of deaths on a day is not correlated with a higher on average positive sentiment polarity 21 days after.

3.3.5 Statistical Analysis

Statistical analysis methods (Pearson correlation and p-value) were utilized in Research Task 3 (SMSA).

Pearson correlation

Pearson correlation between sentiment polarity and COVID-19 cases or deaths was utilized, with different combinations of tweets before and tweets after. Experiments for a 44-days period attempt to determine if a relationship between these two variables exist. The Pearson product-moment correlation coefficient is a statistical measurement of the correlation (linear association) between two sets of values. The Pearson product-moment correlation coefficient for two sets of values, x and y, is given by the formula (4):

$$r = \frac{\sum(x - \bar{x})\,(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}} \quad (4)$$

where, x and y are the sample means of the two arrays of values.

If the value of r is close to +1, this indicates a strong positive linear correlation. For example (adjusted to this study's context), this means that when cases increase, polarity increases or when cases decrease, polarity decreases. On the other hand, if r is close to -1, this indicates a strong linear negative correlation. For example (adjusted to

this study's context), this means that when cases increase polarity decreases or vice versa (polarity increases, cases decrease) (5).

$$Coefficient\ (r) = \ Pearson\ (array1; array2)\ (5)$$

The strength of correlation can be described by the absolutes for r as follows: 0.00-0.19 very weak, 0.20-0.39 weak, 0.40-0.59 moderate, 0.60-0.79 strong, 0.80-1.00 very strong (Evans, 1996).

p-value

As an extra statistical measure, p-value calculation was performed. P-value is the probability to get results that are at least as extreme as the results that are observed, considering the assumption that the null hypothesis is correct (Ioannidis, 2018). This methodological addition attempts to double check hypotheses that connect sentiment polarization with the number of the cases or deaths. Therefore, the p-values per case (polarization-cases and polarization-deaths) are extracted. As for the actual p-value, the alpha (α) value a.k.a. level of significance, it is set to 0.05. This means that for a p-value < 0.05 it is considered that the null hypothesis is rejected, while for p-value > 0.05 the null hypothesis is accepted due to lack of enough proposition against it. Yet, in general there is always a chance that a null hypothesis is true (Nickerson, 2000).

3.4 LIMITATIONS, ETHICS AND THREATS TO VALIDITY

This section outlines any ethical consideration of the research as well as problems, limitations and threats to the validity of the results.

3.4.1 Social Media Types

Data were gathered from the official SM descriptions. Their features were recorded and processed to generate a dataset by grouping under adjective comprehension, removing duplicates and missing values when necessary. The SM used were chosen taking into consideration user penetration and available features. Some SM implement fewer features than others (e.g., Facebook compared with Tinder), thus the analysis might be impaired by this disparity.

Frequent itemsets were extracted in order to produce generalized rules for forming new SMTs, yet with relatively high confidence, but rather low support. Ideally, strong rules should be reported (high confidence and support), but due to the nature of the dataset explained in Sect. 3.2.1 (a simple grouping was implemented

although the results might be considered ambiguous, due to the general subjectivity of grouping features as they are matched with a specific Utility), it is not possible to do so at the required extend. This perceived threat to validity was the primary reason for pursuing further experimental validation by clustering.

The second experiment offers more positive results, since the number of categories was further reduced. In order to generate fewer clusters, experimentation continued by removing dominant utilities during the analysis. The assumption was made that by removing one by one the three most frequent utilities, while presenting and analysing the output in a sequential manner, will enhance results.

3.4.2 Topic Extraction on Social Media

Limitations are discussed that may introduce bias to the proposed methodology. For performing the proposed analysis, data were retrieved by using a Twitter crawler. This is just one SM, over a specific period resulting in a number of tweets, since the COVID-19 outburst. More specifically, from 27/2/2020 until 28/8/2020, 2.146.243 unique tweets were collected in a worldwide scale. Inevitably this dataset (at some point) generates restrictions for a more robust analysis. More SM data sources could be utilized, for example retrieving COVID-19 data from other SM platforms and combining them into one dataset.

There are also data biases due to the preprocessing techniques utilized. Although multiple data preprocessing steps were implemented, it is almost impossible to completely clean the text. For example, based on the proposed approach it is not possible to ascertain synonyms of strong words that are used in slang or jargon. The most representative such word in this analysis was COVID- 19 and its synonyms. Although the hashtags have been filtered within the tweets, people use words such as “corona”, “covid”, “coronavirus”, or simply “virus” and numerous other miss-spellings. Also, there are always missing data that may cause faulty representation of results. This issue could be mitigated by further improving the implemented preprocessing process.

The decisions made regarding the minimum thresholds of support, confidence, lift and leverage, could also introduce bias. During the proposed analysis, the most frequent wordsets were extracted attempting to form new topics, yet with the higher possible leverage and support. Preferably, the strongest rules should be identified (high

confidence and support values) while taking into consideration leverage and lift. For that reason, the resulting grouping of wordsets to topics may be ambiguous, due to the general subjectivity that this methodology introduces. This threat to validity generates concrete reasons for pursuing extended experimental validation of the proposed methodology regarding ARM. Yet, the abovementioned limitations and assumptions motivate points for future work.

3.4.3 Finding Correlations Between Public Sentiment and COVID-19 Cases and Deaths

This research task is subject to possible biases and threats to validity, reported here. There were no geolocation restrictions when crawling the data, therefore Twitter data are considered to span all over the world. As already mentioned, the language for tweets retrieved was English. Also, findings are subject to limited generalization since data from only one SM platform, i.e., Twitter were utilized.

The data retrieval script (crawler) did not retrieve all the tweets related with COVID-19 available for the period of 27/2/2020 to 28/8/2020. This is due to minor technical issues encountered during that period. These issues relate with the server that runs the script and its uptime. Nevertheless, around 14k English tweets per investigated day on a global scale were gathered.

The sentiment analysis methods have certain issues that may result in poor validity of the polarization process. The polarization process was not enhanced, but the overall preprocessing was addressed. Therefore, the output of the sentiment polarization may be subject to disputes. The proposed improvements focused on text preprocessing techniques.

The hypotheses generated, and their validation process (Pearson correlations and p-values) contain certain arbitrary concepts that cannot guaranty absolute result validity. For example, for the p-value the alpha value was set to 0.05 adopting the empirical cut off threshold typically used by the statistics community (Kennedy-Shaffer, 2019).

Chapter 4: Results

This chapter presents results of the three novel approaches in Social Media domain elaborating in novel research tasks related with Social Media Types (SMTs), Social Media Topic Extraction (SMTE) and Social Media Sentiment Analysis (SMSA).

4.1 SOCIAL MEDIA TYPES

Two experiments using RapidMiner were conducted on the dataset. In the first experiment, FP-Growth was used, an exhaustive Association Rule Mining (ARM) algorithm, which produces the same results as Apriori, but is faster (Han, Kamber and Pei, 2012). In the second experiment, a progressive approach was followed using three different heuristic clustering algorithms, DBSCAN, k-medoids, Random Clustering, running twelve experiments, organized in four steps as explained later, because there was the need to compare intermediate results at each step. The research experiments do not exclusively deal with the association rule concepts, but also with clustering. A "learn-by-data" based approach was used to reduce the possible number of clusters on SMTs. This means that after having experimented with FP-Growth, but results were not satisfactory. Then experiments continued using clustering algorithms that seemed to have better results than association rules. These experiments are detailed in the remaining of this section.

4.1.1 Experiment#1

FP-Growth executions aimed at generating strong association rules for the Utility entries for each SM on the dataset. Fig. 7 presents all the association rules when using min confidence = 100%, min items per itemset = 1, and max items per itemset = 3. 100% confidence guarantees that the rule is always true. Regarding the support level, experimentation took place with a variety of values based on the data of each experiment. As a starting value, minimum support was set to 2.7% and later raised up to 10%. The aim was to retrieve the greatest values possible (driven by data) both in confidence and support, in order to find strong rules.

```
Association Rules
[Applications] --> [Connecting] (confidence: 1.000)
[News] --> [Connecting] (confidence: 1.000)
[Multimedia, Privacy] --> [Connecting] (confidence: 1.000)
[Multimedia, Applications] --> [Connecting] (confidence: 1.000)
[Multimedia, News] --> [Connecting] (confidence: 1.000)
[Professional, Applications] --> [Connecting] (confidence: 1.000)
[Sharing, Applications] --> [Connecting] (confidence: 1.000)
[Sharing, News] --> [Connecting] (confidence: 1.000)
[Profile, Privacy] --> [Connecting] (confidence: 1.000)
[Profile, Applications] --> [Connecting] (confidence: 1.000)
[Publishing, Privacy] --> [Connecting] (confidence: 1.000)
[Publishing, Applications] --> [Connecting] (confidence: 1.000)
[Privacy, Voting] --> [Connecting] (confidence: 1.000)
[Privacy, Applications] --> [Connecting] (confidence: 1.000)
[Voting, Applications] --> [Connecting] (confidence: 1.000)
[Voting, News] --> [Connecting] (confidence: 1.000)
[Sharing, News] --> [Multimedia] (confidence: 1.000)
[Voting, Applications] --> [Multimedia] (confidence: 1.000)
[Opinions, Schedule] --> [Professional] (confidence: 1.000)
[Publishing, Privacy] --> [Profile] (confidence: 1.000)
[Profile, Applications] --> [Privacy] (confidence: 1.000)
[Voting, Applications] --> [Profile] (confidence: 1.000)
[Voting, Applications] --> [Privacy] (confidence: 1.000)
```

Fig. 7. Association Rules retrieved from dataset.

It was found that some utilities form strong rules with high values for support and 100% confidence. For example:

i) When an SM platform provides the Applications utility, it is sure to contain Connecting (support=6.2%). This suggests that based on the data “Applications” and “Connecting” can be part of the same meta-utility, meaning that in essence “Applications” are never provided unless “Connecting” is.

ii) In the same manner, when a platform provides the News utility, it is sure to contain Connecting (support = 5.4%).

iii) When it provides the Multimedia and Privacy utilities, it is sure to contain Connecting (support = 5.4%).

iv) When it provides the Multimedia and Applications utilities, it also contains Connecting (support = 4.5%).

v) When it provides the Multimedia and News utilities, it also contains Connecting (support = 3.6%).

When it provides the Professional and Applications utilities, it also contains Connecting (support=3.6), and so on. However, if the 23 rules shown in Fig. 7 were to be used to formulate groups of utilities, 16 rules would be of the form X => Connecting. In other words, 10 utilities including Connecting would form one big group, whilst the remaining four utilities will be standalone, producing a taxonomy of five new SMTs. The complete list of rules with confidence = 100% is shown in Fig. 7. For further reference, Appendix E displays all frequent itemsets with min. support = 2.7%, including itemsets producing the rules presented in Fig. 7 with confidence = 100%.

Fig. 8. Venn Diagram for Support=10%.

At first, experimentation commenced in order to create rules with min. confidence = 100%, yet they proved to be too strict, so thresholds were lowered by including all results with confidence ≤100%, but with min support = 10%. Based on these frequent itemsets a basic grouping is performed, aiming to produce results that better back the stated hypothesis H0. Applying a threshold of 10% Support on Appendix E it is observed that eight groups of utilities are created as shown in Fig. 8.

As shown in Fig. 9, it is implied that Connecting, Professional, Multimedia and Sharing belong to the same group while Entertainment, Profile, Publishing and Opinions form standalone groups.

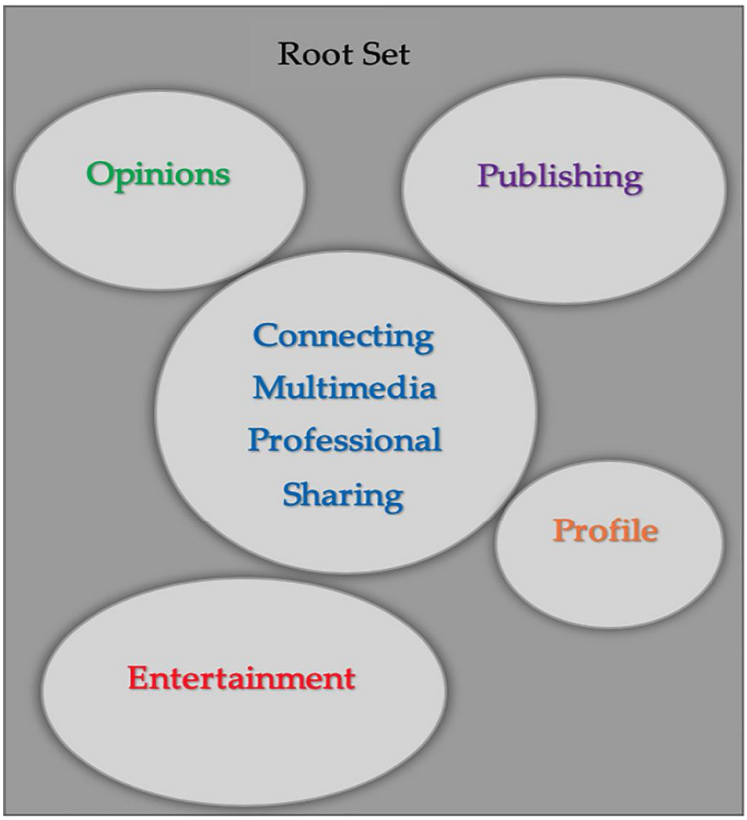

Fig. 9. Venn Diagram with five groups.

Grouping Utilities based on this approach means that itemsets with lower support levels are neglected while leading to the generation of one big group and four smaller ones. Despite the positive results, association rules could be considered biased since some utilities appear more often than others in the dataset. To address that Experiment #2 was conducted.

4.1.2 Experiment#2

The dataset was clustered in a sequential way by excluding one by one the top three dominant utilities (Connecting, Multimedia, Professional). At this point, taxonomies can be generated using clustering as shown in the Tree Diagram in Fig. 10.

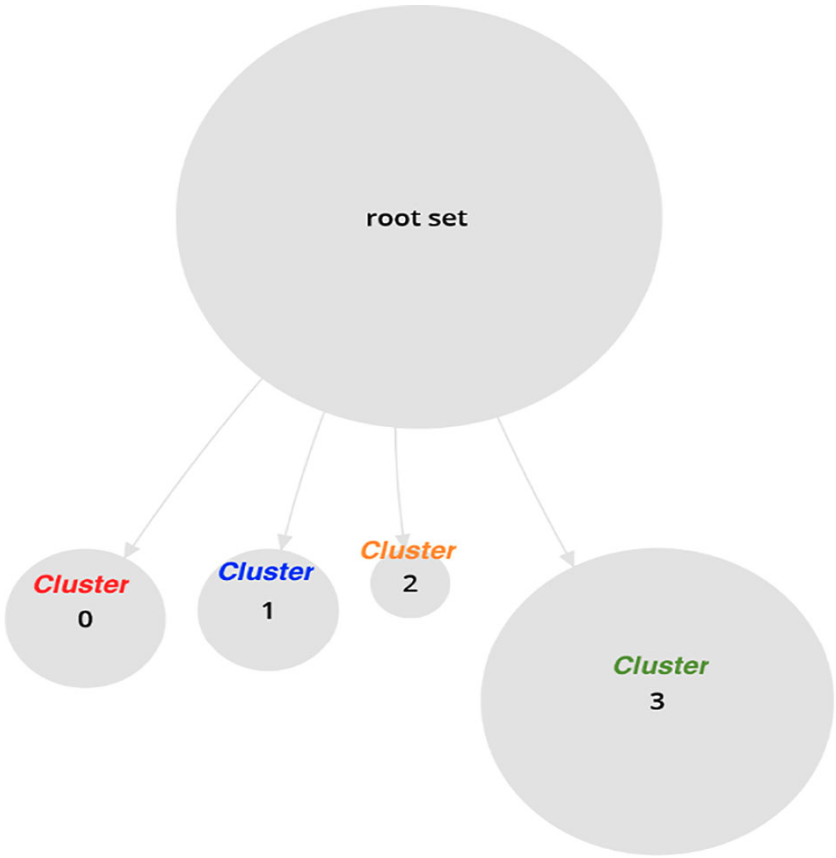

Fig. 10. Tree diagram for k-medoids results.

Experimentation commenced by executing clustering algorithms aiming to generate groups that could help on forming new SMTs. Table 15 lists results after running three different clustering algorithms: Density-Based Spatial Clustering of Applications with Noise (DBSCAN), k-medoids and Random Clustering on the dataset, before removing the dominant utilities (Connecting, Multimedia, Professional). For DBSCAN the default parameters from RapidMiner were utilized, which are: epsilon = 1, min points = 5. DBSCAN does not need to be given the number of clusters. It automatically produced k = 6 clusters. For k-medoids k = 6, max runs = 10, max optimization steps = 100 were used and for Random Clustering, k = 6. Each of the algorithms produced six clusters of variable composition. Given the lack of a ground truth and the unsupervised nature of clustering these results cannot be meaningfully evaluated in a standalone basis.

Table 15. Clustering results including dominant attributes.

Clustering Method	C0	C1	C2	C3	C4	C5
DBSCAN	69	7	12	10	7	7
k-medoids	25	9	36	17	9	16

Random Clustering	19	23	16	14	19	21

Next, three algorithms were executed removing one by one the most dominant Utilities from the dataset. First, the experiment was executed with the same parameters having removed the top ranked of the biased Utilities: Connecting (Table 16).

Table 16. Clustering results without Connecting Utility.

Clustering Method	C0	C1	C2	C3	C4
DBSCAN	61	14	16	11	10
k-medoids	28	29	25	20	10
Random Clustering	19	28	25	18	22

DBSCAN produced k = 5 clusters which is an output that is closer to validate the hypothesis (H0). For the next experiments, k was reduced according to the number of clusters produced by DBSCAN, since it is an algorithm that determines the number of clusters. The reason so doing so was for comparing the output for each run of the three clustering algorithms. The goal was to find the point at which two or more algorithms produce the same number of clusters.

Then, experimentation continued with the same parameters having removed the top two ranked of the biased utilities: “Connecting” and “Multimedia” (Table 17). DBSCAN again produced k = 5 clusters.

Table 17. Clustering results without Connecting and Multimedia Utility.

Clustering Method	C0	C1	C2	C3	C4
DBSCAN	53	5	24	17	13
k-medoids	23	52	11	24	2
Random Clustering	19	28	25	18	19

Finally, experimentation involved the removal of all dominant utilities: Connecting, Multimedia, Professional, with the same parameters, except this time, given that DBSCAN produced k = 4 clusters, Random Clustering also used k =4 in order to compare the results for the same number of clusters. It is evident that, DBSCAN reduces the number of clusters from six to four, so does k-medoids since for k = 6 it creates two clusters (C4 and C5) that each contains zero items and for k = 4 it simply swaps the items in C3 with the ones in C2 (Table 18).

Table 18. Clustering results without all biased attributes.

Clustering Method	C0	C1	C2	C3	C4	C5
DBSCAN	51	6	41	14	-	-
k-medoids (k=6)	23	20	58	11	0	0
k-medoids (k=4)	23	20	11	58	-	-
Random Clustering	26	29	28	29	-	-

After examining Appendix F it was found that the generated clusters were formed based on the presence of specific utilities in each cluster. In particular, SM with the Entertainment Utility belong to C0. SM with the Sharing Utility belong to C1. SM with the Profile Utility belong to C2. All the remaining SM which do not have any Utility, or they have any Utility except from Entertainment or Sharing, or Profile belong to C3.

Table 19 shows a part of the results (see the complete cluster analysis Appendix F) from the last step of the sequential execution of the clustering algorithms.

Table 19. Proposed clusters posing as sample taxonomies with k-medoids (k=4).

id	C0	id	C1	id	C2	id	C3
1	Facebook	2	YouTube	12	LinkedIn	10	Flickr
25	WeChat	3	Instagram	18	Snapchat	13	VK
39	Kiwibox	4	Twitter	19	Quora	15	Meetup
46	DevianArt	5	Reddit	30	Telegram	16	WhatsApp
56	Last.fm	6	Vine	38	Pinboard	17	Messenger
58	Flixster	7	Pinterest	86	LiveJournal	21	Nextdoor
59	Gaia Online	8	Ask.fm	88	Qzone	22	ProductHunt
67	Goodreads	9	Tumblr	94	Xing	23	AngelList
79	Wayn	11	Google+	101	Solaborate	24	Kickstarter
80	CouchSurfing	14	ClassMates	103	Xanga	26	Skype
81	TravBuddy	20	GirlsAskGuys	110	MyHeritage	27	Viber
82	Tournac	34	Stumbleupon	-	-	28	Viadeo
83	Cellufun	35	Foursquare	-	-	29	Gab
89	QQ	53	43Things	-	-	31	Tagged
92	YY	55	Uplike	-	-	32	Myspace
95	VampireFreaks	65	Tinder	-	-	33	Badoo
98	ASmallWorld	85	Plurk	-	-	36	MeetMe
99	ReverbNation	87	Weibo	-	-	37	Skyrock A192
100	SoundCloud	90	Baidu	-	-	40	Twoo
105	Zynga	97	Ravelry	-	-	41	Yelp
106	Habbo	-	-	-	-	42	Snapfish
107	FunnyOrDie	-	-	-	-	43	Photobucket

111	MocoSpace	-	-	-	-	44	Shutterfly

General Purpose Networks: SM which are mainly described by Connecting, Multimedia, Professional and Sharing Utilities belong to this set.

Entertainment Networks: This set describes SM that have to do with Entertainment. Gaming, Shopping, Sports, Travel, Movies etc.

Publishing Networks: This set contains SM with blogging, general form of publishing and microblogging being their main functionality.

Profiling Networks: This set comprises SM that offer functions promoting skills, goals, personal journals, etc.

Opinion Networks: The final set contains SM that mainly deal with recommendations, reviews, discussions, polls etc.

A taxonomy for SMTs was created based on a set of generalized axioms produced after running Experiment#2:

Axiom 7: Any SM that provides at least the Entertainment Utility alone, or Entertainment along with Profile, or Entertainment along with Sharing, is assigned to Cluster0.

Axiom 8: Any SM that provides at least the Sharing Utility alone, or Sharing along with Profile, is assigned to Cluster1.

Axiom 9: Any SM that provides at least the Profile Utility alone is assigned to Cluster2.

Axiom 10: If none of axioms 7–9 above stands, the SM belongs to Cluster3.

This leads to the conclusion that a new Taxonomy for SMTs can be proposed:

Entertainment Networks: The first cluster showcases results that are similar to Experiment# 1 generating a SM category which describes SM that have to do with general entertainment, gaming, shopping, sports, travel, movies etc.

Sharing Content Networks: This cluster contains SM that support features that prompt content sharing, hashtags, quotes, location sharing, any kind of posts etc.

Profiling Networks: This cluster produces the same results with Experiment#1, forming a category that describes SM that offer functions that promote skills, goals, personal journals, etc.

General Purpose Networks: The final cluster has all the remaining SM that did not enroll on one of the above Networks (Entertainment, Sharing, Profiling).

Moving on to the evaluation of the two experiments (Experiment#1, Experiment#2), the aim was to produce a methodology that reduces the number of SMTs. Current literature proposes nine SMTs (Gundecha and Liu, 2012) or seven SMTs (Kietzmann *et al.*, 2011). In comparison with this work, it is noted that by running clustering methods on the dataset, the output is better than that of association rules, since the formed clusters (taxonomies) were reduced from five to four moving closer to proving the initial hypothesis H0. However, in both of the experiments fewer SMTs were produced.

By examining results from Experiment#1 and Experiment#2 an insight for a proposed new taxonomy on SMTs may be provided, motivated and reasoned by the dataset observations and experiments:

Entertainment networks: This cluster of SM appears in both Experiments#1 and #2 and it consists of SM that have to do with general entertainment, such as games, sports, cinema, travel, and so on. By further analysing the data it was identified that this SMT offers the following Utilities:

Primary: Entertainment.

Secondary: Connecting, Multimedia, Opinions.

Trivia: Sharing, Privacy, News, Promoting, Voting, Publishing, Schedule, Profile, Applications, Professional.

Profiling Networks: This cluster also appears in both Experiment#1 and #2, and forms an SMT describing SM that offer functions promoting skills, goals, personal journals, etc. By analysing the data, it is observed that such SM offer the following Utilities:

Primary: Profiling.

Secondary: Connecting, Multimedia, Professional, Opinions, Publishing, Privacy, Voting, Applications, Promoting.

Trivia: Sharing, News, Schedule, Entertainment.

Social Networks: This SMT is generated by merging General Purpose Networks as described by findings from Experiments#1 and 2. Such SM offer the following Utilities:

Primary: Connecting, Multimedia, Professional, Sharing.

Secondary: Publishing.

Trivia: Privacy, News, Promoting, Voting, Schedule, Profile, Applications, Opinions, Entertainment.

On all of the three proposed SMTs, as secondary Utilities were labelled the ones that are found to be paired with the Primary Utility of each proposed SMT, without considering the support level of the association rule and as trivia were labelled the ones that do not display any association rule at all (Appendix E). This proposed taxonomy verifies the initial hypothesis (H0). Evaluating results, source (Gundecha and Liu, 2012) essentially concludes with nine SMTs, source (Kietzmann *et al.*, 2011) with seven SMTs and source (Kaplan and Haenlein, 2010) with three yet not operationally representing based on the current evolution of SM. By consolidating results from Experiments 1 and 2 this thesis produces an updated version of SMTs as described in this section.

4.2 SOCIAL MEDIA TOPIC EXTRACTION

This section presents results related to retrieved WordClouds, Frequent Wordsets and Association Rules. As discussed in Sect. 3.3.2, LDA was used to extract topics from the database of pre-processed 2,062,864 tweets. The goal was to form clusters of topics that represent public feelings, opinions, attitudes or discover inferences regarding them. For that reason, LDA was performed several times. Table 20 shows results from these simulations.

Table 20. Summary of simulations, 10 topics and extracted words.

Simulation	min_threshold	Number of extracted words
1	1,000	2,713
2	2,000	1,597
3	3,000	1,153
4	4,000	886
5	5,000	713

A minimum threshold is set, ensuring that if words appear in less than 1000, 2,000, 3,000, 4,000 and 5,000 tweets respectively, they are ignored. These thresholds are empirically set, ranging around 0.05% and 0.25% of the overall sum of tweets (2,062,864). As a result, words that seem to be too common to have a strong meaning for the topics are discarded. The number of extracted words per execution is reported. That way, it is possible to observe how the number of extracted words decreases as the minimum threshold increases. Also, an analysis is performed to identify the number of extracted topics to be 10, as explained in Sect. 3.3.2. Next, the 20 most common words for each topic are retrieved, along with their weights (number of appearances) to form WordClouds.

Table 21. 10 most common words per topic (T) for simulation#1.

Rank	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9
1	covid	coronavirus	pandemic	home	like	case	need	people	virus	close
2	hand	outbreak	impact	stay	time	new	help	die	mask	school
3	corona	read	covid	work	know	test	fight	days	trump	services
4	wash	pandemic	business	safe	go	deaths	people	infect	spread	measure
5	vaccine	say	market	time	social	report	health	patient	china	order
6	ms	covid	new	test	look	total	world	symptoms	stop	public
7	dr	question	company	get	think	number	crisis	kill	people	lockdown
8	clean	ill	crisis	family	watch	positive	pandemic	covid	don't	spread
9	virus	amid	help	quarantine	good	confirm	time	virus	news	open
10	drug	uk	team	individuals	day	coronavirus	support	disease	corona	pay

This is done in order to rank and visualize the most frequently appearing words per generated topic. Table 21 indicatively shows the top 10 words from LDA simulation#1. The complete ranking list per LDA simulation along with the weights per word (number of appearances) can be found in Appendix G.

4.2.1 WordClouds

It is quite common to use WordClouds to visualize the strongest words per topic. For example, Fig. 11 depicts T3 topic results from simulation#1.

Fig. 11. LDA simulation#1, topic T3 WordCloud.

Similar figures are employed to visualize the words for each topic, such as T4 shown in Fig. 12.

Fig. 12. LDA simulation#1, topic T4 WordCloud.

For enhancing the visualization of results, instead of just using WordClouds for depicting the topics retrieved, a graph for each LDA simulation is utilized (Fig. 13). Each topic is depicted as a hub linked with nodes corresponding to words belonging to that topic generating a network of words and topics. The size of each node depends on the word's weight, i.e., its number of appearances in the dataset. The size of each hub results from the average number of appearances of its words. A graph visualization for each simulation is presented in Appendix H.

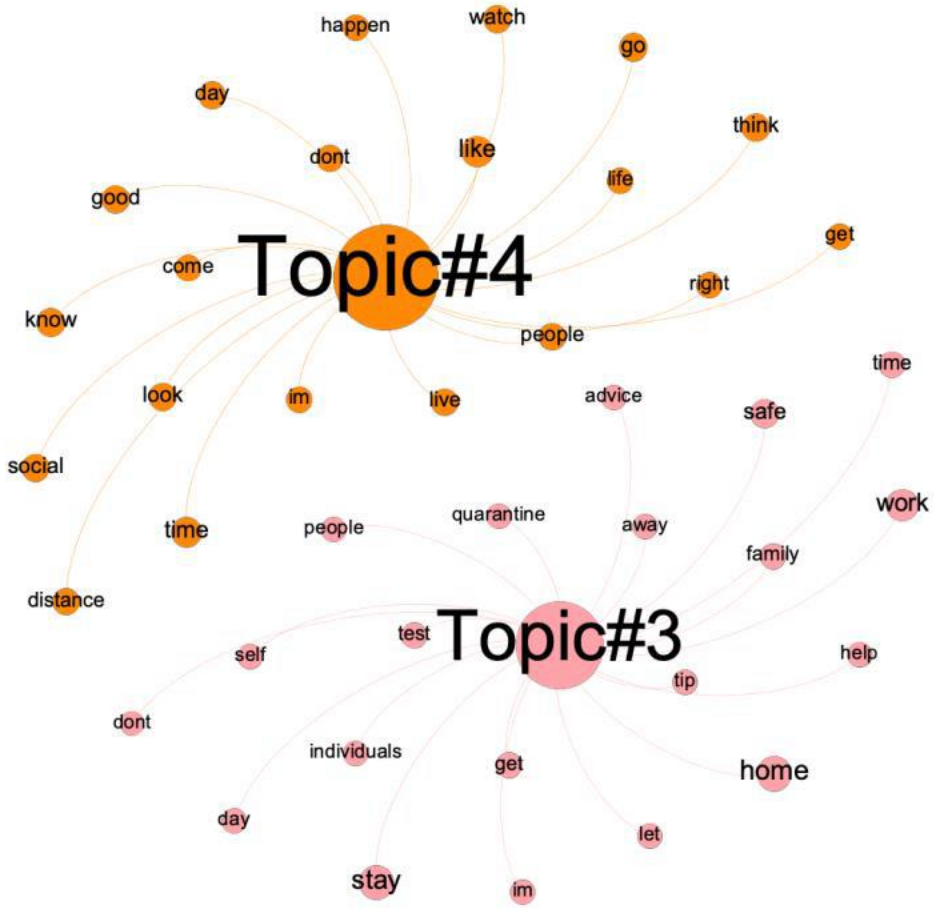

Fig. 13. Graph visualization of simulation#1, T3 and T4 topics.

Five simulations were performed attempting to generate topics or clusters that contain the stronger words retrieved from the tweet's dataset. Yet, it is unclear what the theme of each topic is or whether some topics infer to the same theme.

One way to retrieve the theme for each topic is to set a threshold for appearances beyond which words can be considered influential to set the theme of the topic. For example, according to Table 22 T3 has a sum of 431,579 word appearances, while T4 has a sum of 547,267 word appearances. If a min threshold of 10% is arbitrary set for considering a word important, this would mean that any words that have less appearances than 43,158 should be omitted. Respectively, for T4 this number would be 54,727 appearances.

Table 22. SMTE simulation#1 T3 and T4 word ranking.

Rank	T3	T3 word appearances	T4	T4 word appearances
1	home	78711	like	53379
2	stay	70744	time	44102

3	work	56798	know	34043
4	safe	36521	go	33540
5	time	20496	social	31223
6	test	16538	look	31207
7	get	14618	think	30573
8	family	13733	watch	30250
9	quarantine	11941	good	29796
10	individuals	11262	day	26768
11	people	11107	people	26198
12	let	11001	live	24323
13	help	10832	dont	22888
14	im	10715	distance	22381
15	day	10506	come	20529
16	advice	10039	life	19270
17	self	9910	im	17741
18	dont	8797	happen	17246
19	away	8742	right	16013
20	tip	8567	get	15797

In that case, the extracted themes for each topic should be: T3 (home, stay, work) and T4 (null). If the threshold is lowered to 5%, many words that might introduce noise to the theme inference may be included.

At 5% the following outcome is reported: T3 (home, stay, work, safe) and T4 (like, time, know, go, social, look, think, watch, good). Therefore, a mechanism is required that searches and decides which words are considered more important amongst all the retrieved topics, and clarifies this selection based on the occurrences within the dataset. Also, it is observed that words such as “people” or “get” may appear in multiple topics. To that end, ARM is performed on the set of strong words per topic. The aim is to narrow these down possibly to single words, and identify, in a more precise manner, what is the public attitude regarding the pandemic during the investigated period.

4.2.2 Frequent Wordsets

As a first step, all extracted words appointed to topics resulting from all simulations get appended to a list. That way all simulation results become merged into a dataset that contains the strongest words that can be exploited from the tweets dataset. The top-30 most frequent words in this dataset are shown in Fig. 14. The top five words in a descending order are “people”, “coronavirus”, “covid”, “need” and “pandemic”.

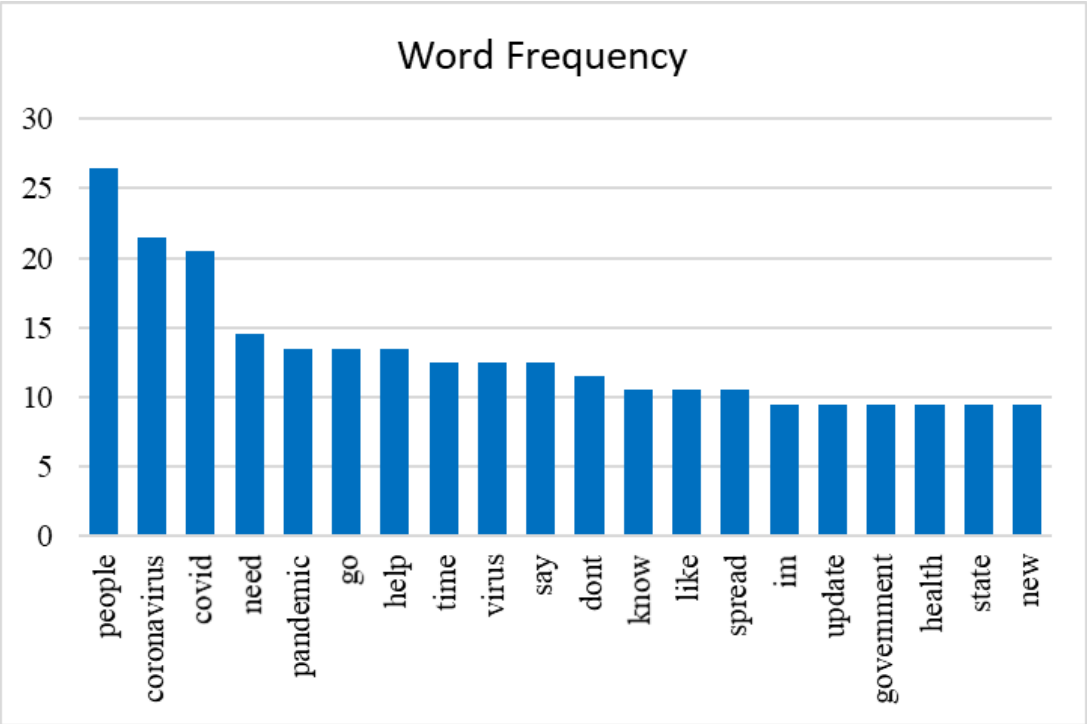

Fig. 14. Top 30 frequent words.

For identifying frequent wordsets FP-Growth was utilized (Sect. 3.3.3). Experimentation took place with a variety of values for the minimum support level in order to extract frequent wordsets.

Table 23. Number of frequent wordsets with different support values.

Minimum support value (%)	Number of frequent wordsets
50	1
40	3
30	4
20	19
9-10	16622
7-8	33480
5-6	135696
3-4	570924

According to Table 23 there is only one frequent wordset when setting minimum support to 50%. The level was further dropped down to 3% and stopped there, since the identified wordsets with such low support levels become too many. The aim is to identify the most frequent wordsets with the stronger possible associations.

4.2.3 Association Rules

Lift and leverage were utilized, as described in Sect. 3.3.3, to extract the strongest rules. These measures show how the probability of a rule to hold can relate to the expected probability, when the antecedent and consequent wordsets are independent from each other. The difference is that lift can compute the strongest wordset associations for wordsets with lower support (less frequent wordsets), whilst leverage prioritises wordsets with higher support levels. As aim is to extract the strongest rules within the dataset, focus was made on ranking the strongest rules, based on high leverage values, shown in Table 24.

Table 24. Number of association rules based on leverage.

Minimum leverage value (%)	Number of association rules
10	58
9.4-9.5	102
9.3	114
9.2	130
9.1	138
9	4,598,926

Words were grouped into topics by combining results from the top-50 most common words and association rules with the strongest support, confidence, lift and leverage. Strong rules were identified by filtering out rules with leverage less than 9%, as they tend to have the same support, confidence and lift, rendering the rules identical. The strongest 138 association rules can be found in Appendix I.

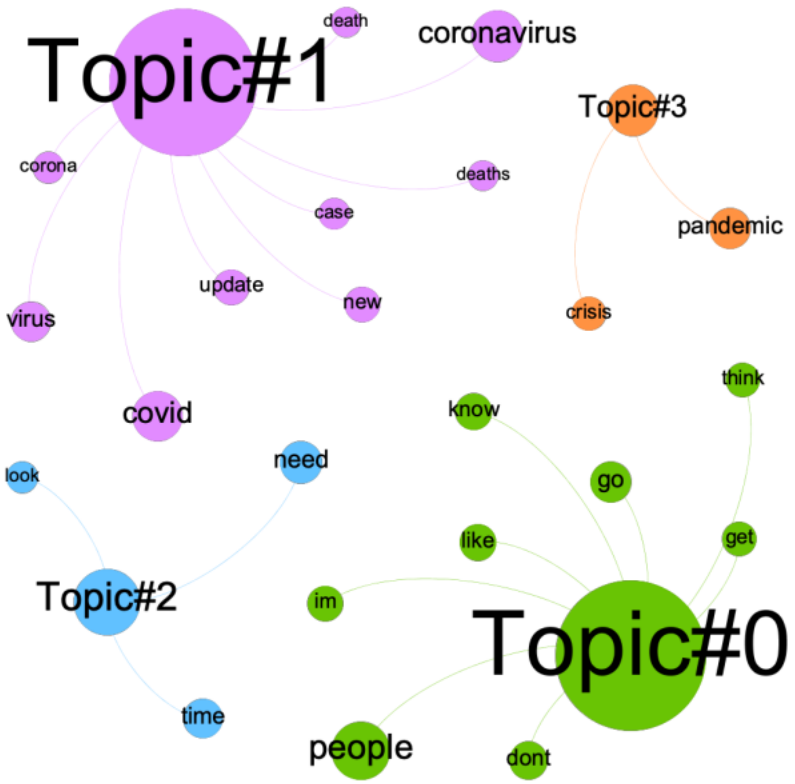

Fig. 15. Final Topic Extraction, resulting from ARM utilization.

The strongest wordsets were grouped observing these rules and used graphs to visualize the final topics extracted from the dataset. This process forms four topics, depicted in Fig. 15. Node size for topics signifies its weight or word count of appearances, while node size for words signifies its frequency.

Topic T0 infers that people (“people”) discuss personal opinions (“im”, “think”, “like”, “know”), negative attitude (“dont”) and inferenced motivations such as “go”, or “get”.

Topic T1 infers that users appointed to this cluster post tweets using COVID-19 synonyms, such as “corona”, “coronavirus”, “covid”, “virus” to comment about updates (“update”) on new (“new”) cases (“case”) or death(s). It should be noted that COVID-19 hashtags were filtered during data preprocessing. In addition, according to leverage values of rules associated with “death” and “deaths”, it is more likely that “deaths” appear instead of “death”. If stemming was applied during preprocessing “death” and “deaths” would stand as the same lemma: “death”. Stemming was not applied by choice, as this level of information detail matters, for example, if a tweet refers to a single death or its plural form; yet lemmatization was applied.

Topic T2 has a more arbitrary inference, since it contains the words “need”, “look” and “time”. It might be the case that users discuss the time needed for developing a cure. Yet, this cannot be proven by this analysis since these associations were not identified within the results.

Topic T3 has a rather clear inference containing the words “pandemic” and “crisis”. Therefore, people label or mention in posts COVID-19 as a pandemic and probably comment on the numerous drawbacks of a crisis.

It is also evident that tweets that belong to T0 and T1 appear much more often than T2 and T3, as indicated by the size of their nodes. When the filters for convergence and lift lower, other topics appear (Fig. 17), yet the analysis was stopped at this point. This has to do with the leverage threshold as reported earlier.

The analysis focuses on Twitter data related with COVID-19 retrieved between February and August of 2020. Data preprocessing was performed and then a common topic extraction technique (LDA) was applied to retrieve the themes of discussions regarding the pandemic. Next, an attempt to improve topic extraction by narrowing down the number of top retrieved topics utilizing ARM commences. The aim was to discover the greatest possible values, while being driven by data observation. These values are related with the interpretation of ARM measures, each used for a different purpose to identify the strongest rules between wordsets inside topics. The experimentation was conducted with various values for these measures as presented in this section. The results showcase that the task of topic extraction can be enhanced and further generalized by combining LDA with ARM leading to less yet more representative topics.

4.3 SOCIAL MEDIA SENTIMENT ANALYSIS

This section presents the outputs from Preprocessing and Polarization layer, as well as the Statistical Analysis layer for validation of Hypotheses formation layer (described in Fig. 4). Fig. 16 reports on the output of the Preprocessing and Polarization layer, highlighting the performed sentiment polarization analysis, on raw tweets with no preprocessing at all and processed tweets. By observing the output, it is noted that the last layer (Statistical Analysis) receives as input the Processed Tweets’ values.

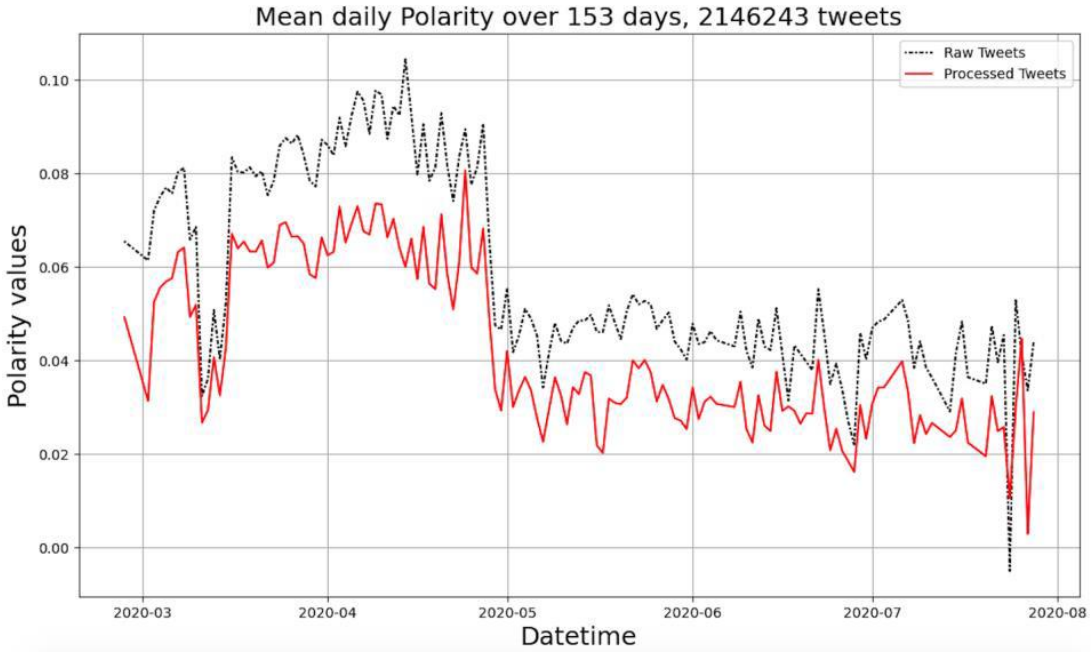

Fig. 16. Changes in sentiment polarity after preprocessing.

For statistical analysis, the Pearson correlation is calculated for the timeseries of the Processed Tweets. This process attempts to distinguish possible relationships between variables Polarity vs. Cases and Polarity vs. Deaths for the 153 investigated days.

Table 25. Pearson analysis and p-values for hypotheses.

	Pearson		p-value	
Before or After (tweets)	New Cases	New Deaths	New Cases	New Deaths
22 days before	-0.682397393	0.064007689	2.656751E-22	0.4318411072
21 days before	-0.668832958	0.094788235	3.457032E-21	0.2438260698
20 days before	-0.659785755	0.114805283	1.778129E-20	0.1576280727
19 days before	-0.654462348	0.077009274	4.540812E-20	0.3440739433
18 days before	-0.647419723	0.098690523	1.525454E-19	0.2248666403
17 days before	-0.648581248	0.111132148	1.251861E-19	0.1714386103
16 days before	-0.642864119	0.110891371	3.284914E-19	0.1723742716
15 days before	-0.637800131	0.100506525	7.590937E-19	0.2164085749
14 days before	-0.631030263	0.104354511	2.270785E-18	0.1992412055
13 days before	-0.623964697	0.114745846	6.926604E-18	0.1578446973
12 days before	-0.615586662	0.099098688	2.506834E-17	0.2229455290
11 days before	-0.608527263	0.072126649	7.195522E-17	0.3756198922
10 days before	-0.612990331	0.064721794	3.705770E-17	0.4267106953
9 days before	-0.609080218	0.060639591	6.631372E-17	0.4565107126
8 days before	-0.602979915	0.052506207	1.618103E-16	0.5191964112
7 days before	-0.602403416	0.032609932	1.758695E-16	0.6890378327
6 days before	-0.593575808	0.032416871	6.169950E-16	0.6907843868
5 days before	-0.596424444	**0.022436303**	4.132437E-16	0.7831007437
4 days before	-0.597113974	-0.02443483	3.748100E-16	0.7643233905
3 days before	-0.594898888	-0.03989054	5.124412E-16	0.6244401717
2 days before	-0.601584888	-0.035023577	1.978978E-16	0.6673433965

1 day before	-0.598194551	-0.08960527	3.214907E-16	0.2706860939
0 days	**-1**	-0.102988581	5.370485E-16	0.2052186403
1 day after	**-0.589327552**	-0.103525311	1.113581E-15	0.2028546523
2 days after	-0.590766552	-0.105324435	9.125971E-16	0.1950738874
3 days after	-0.598896707	-0.147560598	2.908906E-16	0.0687235141
4 days after	-0.602324548	-0.179448648	1.778832E-16	0.0264525313
5 days after	-0.60894604	-0.187501582	6.764163E-17	0.0202938748
6 days after	-0.616741202	-0.210136379	2.104434E-17	0.0091311637
7 days after	-0.613129478	-0.216184442	3.629279E-17	0.0072760296
8 days after	-0.61366015	-0.250020096	3.351459E-17	0.0018278124
9 days after	-0.62382306	-0.279359638	7.081152E-18	0.0004704534
10 days after	-0.627804213	-0.297903763	3.791896E-18	0.0001840738
11 days after	-0.635435763	-0.334763794	1.116439E-18	0.0000234179
12 days after	-0.647962474	-0.368159086	1.391019E-19	2.840249E-06
13 days after	-0.64904319	-0.41192768	1.156945E-19	1.221343E-07
14 days after	-0.649966799	-0.432694794	9.877983E-20	2.326802E-08
15 days after	-0.657188862	-0.452799342	2.815904E-20	4.188228E-09
16 days after	-0.657710312	-0.456969479	2.568535E-20	2.893469E-09
17 days after	-0.671844689	-0.490331816	1.978649E-21	1.246703E-10
18 days after	-0.682830135	-0.525601893	2.442358E-22	3.037625E-12
19 days after	-0.695486107	-0.546111834	1.952164E-23	2.853218E-13
20 days after	-0.705188049	-0.577610178	2.570777E-24	5.431783E-15
21 days after	-0.712609124	**-0.596110089**	5.153229E-25	4.320166E-16

The goal is to discover a qualitative measure of whether SM data can be correlated with COVID-19, and therefore expose predictive opportunities regarding cases and deaths. For each column, the maximum and minimum correlation absolute values are in bold. Based on p-values (Table 25), Table 26 depicts the status of the conceived hypotheses.

Table 26. HCB1-HCB22, HC0, HCA1-HCA21 & HDB1-HDB22, HD0, HDA1-HDA21 status.

Before/After (tweets)	H (New Cases)	Status	H (New Deaths)	Status
22 days before	HCB22	Rejected	HDB22	Accepted
21 days before	HCB21	Rejected	HDB21	Accepted
20 days before	HCB20	Rejected	HDB20	Accepted
19 days before	HCB19	Rejected	HDB19	Accepted
18 days before	HCB18	Rejected	HDB18	Accepted
17 days before	HCB17	Rejected	HDB17	Accepted
16 days before	HCB16	Rejected	HDB16	Accepted
15 days before	HCB15	Rejected	HDB15	Accepted
14 days before	HCB14	Rejected	HDB14	Accepted
13 days before	HCB13	Rejected	HDB13	Accepted
12 days before	HCB12	Rejected	HDB12	Accepted
11 days before	HCB11	Rejected	HDB11	Accepted
10 days before	HCB10	Rejected	HDB10	Accepted
9 days before	HCB9	Rejected	HDB9	Accepted
8 days before	HCB8	Rejected	HDB8	Accepted
7 days before	HCB7	Rejected	HDB7	Accepted

6 days before	HCB6	Rejected	HDB6	Accepted
5 days before	HCB5	Rejected	HDB5	Accepted
4 days before	HCB4	Rejected	HDB4	Accepted
3 days before	HCB3	Rejected	HDB3	Accepted
2 days before	HCB2	Rejected	HDB2	Accepted
1 day before	HCB1	Rejected	HDB1	Accepted
0 days	HC0	Rejected	HD0	Accepted
1 day after	HCA1	Rejected	HDA1	Accepted
2 days after	HCA2	Rejected	HDA2	Accepted
3 days after	HCA3	Rejected	HDA3	Accepted
4 days after	HCA4	Rejected	HDA4	Rejected
5 days after	HCA5	Rejected	HDA5	Rejected
6 days after	HCA6	Rejected	HDA6	Rejected
7 days after	HCA7	Rejected	HDA7	Rejected
8 days after	HCA8	Rejected	HDA8	Rejected
9 days after	HCA9	Rejected	HDA9	Rejected
10 days after	HCA10	Rejected	HDA10	Rejected
11 days after	HCA11	Rejected	HDA11	Rejected
12 days after	HCA12	Rejected	HDA12	Rejected
13 days after	HCA13	Rejected	HDA13	Rejected
14 days after	HCA14	Rejected	HDA14	Rejected
15 days after	HCA15	Rejected	HDA15	Rejected
16 days after	HCA16	Rejected	HDA16	Rejected
17 days after	HCA17	Rejected	HDA17	Rejected
18 days after	HCA18	Rejected	HDA18	Rejected
19 days after	HCA19	Rejected	HDA19	Rejected
20 days after	HCA20	Rejected	HDA20	Rejected
21 days after	HCA21	Rejected	HDA21	Rejected

For each row in Table 26 several findings can be extracted. For example, for the first row related with hypotheses HCB22 and HDB22 (22 days before), it is concluded that:

i) New Cases exhibit strong negative correlation with polarity (cases increase while polarity drops).

ii) New Deaths exhibit very weak positive correlation with polarity (deaths increase and polarity increases, as well as deaths decrease and polarity decreases).

iii) New Cases p-value<0.05, therefore the null hypothesis (HCB22) is rejected. HCB22: The number of cases on a day is not correlated with a higher on average positive sentiment polarity 22 days ahead.

iv) New Deaths p-value>0.05, therefore the null hypothesis (HDB22) is accepted. HDB22: The number of deaths on a day is not correlated with a higher on average positive sentiment polarity 22 days ahead.

For entries related with HC0 and HD0 (0 days):

i) New Cases exhibit very strong negative correlation with polarity (cases increase while polarity drops).

ii) New Deaths exhibit very weak negative correlation with polarity (deaths increase and polarity decreases, as well as deaths decrease, and polarity increases).

iii) New Cases p-value<0.05, therefore the null hypothesis (HC0) is rejected. HC0: The number of cases on a day is not correlated with a higher on average positive sentiment polarity 0 days before/after.

iv) New Deaths p-value>0.05, therefore the null hypothesis (HD0) is accepted. HD0: The number of deaths on a day is not correlated with a higher on average positive sentiment polarity 0 days before/after.

4.3.1 Interpretation of results

This section discusses knowledge extracted by result interpretation. It informs about insights regarding the COVID-19 crisis by utilizing data from Twitter. Focus is made on extracting and reporting on the polarity from tweets and examine the correlation strength of the polarity with the number of COVID-19 cases and deaths.

According to Fig. 16, it is evident that the overall polarity of evaluated Tweets from 27/2/2020 to 28/8/2020 with 2.146.243 harvested tweets, shows a negative trend, as polarity values drop. More specifically, polarity values start dropping after mid-April 2020. This can be attributed to the fact that since the start of COVID-19 official data reporting in February, people have been reluctant to accept that there is indeed a pandemic (Islam *et al.*, 2020). Yet, at 17/4/2020 there was the greater number of reported daily deaths (12430) for the evaluated period of this study. So, this is the triggering point for an established long term negative overall polarity trend.

According to Table 25, tweets before and after a day are strongly negatively correlated with COVID-19 cases on average. Although, tweets "after" correlation with COVID-19 cases exhibit a slightly stronger negative correlation. Average correlation values for tweets "before" and tweets "after" with new cases are -0.624 and -0.640 respectively. For both cases, the findings are sound since new cases increase while polarity drops (negative trend).

It was also found that the strongest correlation between polarity and new cases is on the same day with Pearson value -1, while the weakest correlation is 1 day after,

with Pearson value -0.589327552. Overall, there is a strong correlation between COVID-19 twitter conversations' polarity with reported cases.

Moreover, tweets before and after a day are very weakly to weakly correlated with COVID-19 deaths on average. Tweets "before" correlation with COVID-19 deaths exhibit a very weak positive correlation while tweets "after" exhibit a weak negative correlation with COVID-19 deaths. The average correlation values for these periods are +0.056 and -0.341 respectively. These findings can be interpreted as, tweets' "before" polarity increases related with a day's deaths and then there is a trend reversal for tweets' "after" with a much greater negative correlation increase (after deaths are announced, the polarity decreases further). More precisely, this trend reversal happens 4 days "before" and negatively increases until 21 days "after".

It is also evident that the strongest correlation between polarity and new deaths is at 21 days after while the weakest is at five days before. Overall, there is weak correlation between COVID-19 twitter conversations' polarity with reported new deaths.

In case the utilization of p-values as expressed in Sect. 3.3.5 are not disputed, all hypotheses associating New Cases with tweets' polarization have a p-value <0.05 therefore they are rejected (Table 26). This is expected since new cases have an overall increasing trend, while sentiment polarity exhibits a negative trend. As for hypotheses associating New Deaths with tweets' polarization values diversify, yet there is cohesion in observed trends. From HDB1-HDB22, HD0 and HD1-HD3 there is p-value>0.05 therefore these hypotheses are accepted. The remaining ones are rejected. This observation suggests that as the number of deaths on a date increases there is higher on average positive sentiment polarity for 22 days before to three days after (a sum of 25 days).

Table 27. Hypotheses' status for New Cases and New Deaths up to 50 days before.

Before (tweets)	H (New Cases) p-values	Status	H (New Deaths) p-values	Status
50 days before	1.29E-23	Rejected	5.67727E-05	Rejected
45 days before	2.19E-21	Rejected	0.018876142	Rejected
43 days before	2.19E-21	Rejected	0.015088417	Rejected
42 days before	3.27E-22	Rejected	0.032555018	Rejected
41 days before	5.48E-22	Rejected	**0.078846068**	**Accepted**
40 days before	2.48E-22	Rejected	0.125211833	Accepted
30 days before	1.22E-24	Rejected	0.921742577	Accepted
27 days before	1.18E-23	Rejected	0.706649347	Accepted
23 days before	1.31E-22	Rejected	0.362643608	Accepted

As an expansion to Table 26, the p-values were checked for New Cases and New Deaths up to 50 days before, to identify the threshold of the day when the Hypotheses' status change. According to Table 27, Hypotheses regarding New Cases remain "Rejected" while Hypotheses regarding New Deaths change from "Accepted" to "Rejected" 42 days before tweets. Therefore, the previously mentioned period of 25 days (according to Table 26) is expanded to a period of 6.5 weeks or 45 days starting from three days after tweets and ending at tweets 41 days before the reported deaths. This suggests that based on this research findings, people tend to realize four days after a day when deaths increase and post tweets with diminished yet positive polarization. Also, it means that the negativity in tweets remains connected with the daily deaths; the hypotheses remain accepted for a very long period of time (45 days). While compared with the daily cases, the hypotheses remain rejected for the initial and the expanded period under scrutiny.

Chapter 5: Analysis

This chapter discusses, interprets and evaluates results. It comments on and analyses the achievements of each methodological stage of this thesis in the domain of Social Media (Sect. 5.1). The research tasks are combined and reported with a representative title based on the final evaluation of accomplishments. Therefore, a theoretical framework integrates a variety of (common) mining tasks highlighting the possibilities for knowledge acquisition utilizing interdisciplinary approaches within the context of the first part (PART I) of this thesis.

5.1 A MULTI-FUNCTIONAL FRAMEWORK FOR DEFINING SOCIAL MEDIA TYPES, EXTRACTING TOPICS AND INFERENCES, AND DISCOVERING CORRELATIONS BASED ON PUBLIC SENTIMENT

This theoretical framework integrates research accomplishments from Research Tasks 1, 2 and 3 relating to SMTs, SMTE and SMSA (according to Table 1).

5.1.1 Social Media Types: Introducing a Data Driven Taxonomy

Literature review reveals that SMTs are in a rapid stage of evolution. SMPs integrate multiple user services; thus, a variety of SMTs tend to offer conceptual Utilities instead of being "single minded". This is due to the accelerated spread and absorption of various SM services. Users require all-in-one platforms easy to use, that satisfy their needs holistically (Asur *et al.*, 2012; Perrin, 2015).

This thesis researches on this issue, aiming to offer an alternative regarding SMTs. The proposed methodology is based on observations on a dataset that contains various SM along with their descriptions. Two experiments were performed using association rule mining and clustering algorithms in order to implement a data-driven approach that proves the initial hypothesis (H0) stating that current standardization on SMTs can be updated, thus reducing the number of SMTs.

Table 28. SMTs comparison of research outcomes with literature.

Source	Description	Number n of SMTs
(Gundecha and Liu, 2012)	Online Social Networking, Blogging, Micro-blogging, Wikis, Social news, Social book-	9

	marking, Media sharing, Opinion, reviews and rating, Answers	
(Kietzmann *et al.*, 2011)	Identity, Conversations, Sharing, Presence, Relationships, Reputation, Groups	7
(Kaplan and Haenlein, 2010)	Blogs, social media sites, virtual games worlds	3
Experiment #1	General purpose, Entertainment, Publishing, Profiling, Opinion	5
Experiment #2	Entertainment, Sharing content, Profiling, General purpose	4
Proposal consolidating results from Experiment #1 & #2	Entertainment networks, Profiling networks, Social networks	3

Table 28 summarizes the outcomes of existing research on SMTs, as well as the output of this work. By empirically observing the results, it may be noted that the first experiment (Experiment #1) produces five SMTs which is perceived to be better and more synched with the current state of play in SM than categorizations proposing nine (Gundecha and Liu, 2012) or seven (Kietzmann *et al.*, 2011) SMTs respectively. Yet, when comparing this early result with a work proposing three SMTs (Kaplan and Haenlein, 2010), despite this referring to a different time period (2010), it seemed appropriate that a tighter categorization scheme was needed. Thus, further research was conducted, striving for better results. With Experiment #2, four clusters were discovered, i.e., four SMTs, which seems more semantically appropriate and representative than five produced by Experiment #1. Finally, an insight of the consolidated version of the two experiments is presented, as discussed in Sect. 4.1, typically capturing emerging SMP services.

5.1.2 Mining Association Rules from Twitter data to Discover Word Patterns, Topics and Inferences

An analysis of a total of 2,146,243 unique tweets for a period of 153 days, from 27/2/2020 up to 28/8/2020 is made, focusing on discussions during the pandemic. Thus, a public dataset for a similar task is not available. Data preprocessing and topic extraction is performed with various simulation parameters. For topic extraction the LDA method was used. The output of LDA simulations showcased the existence of discrepancies regarding topic extraction. For example, the same strong words can appear in multiple topics at the same time or extracted topics may contain many trivia words that make topic theme inference more difficult.

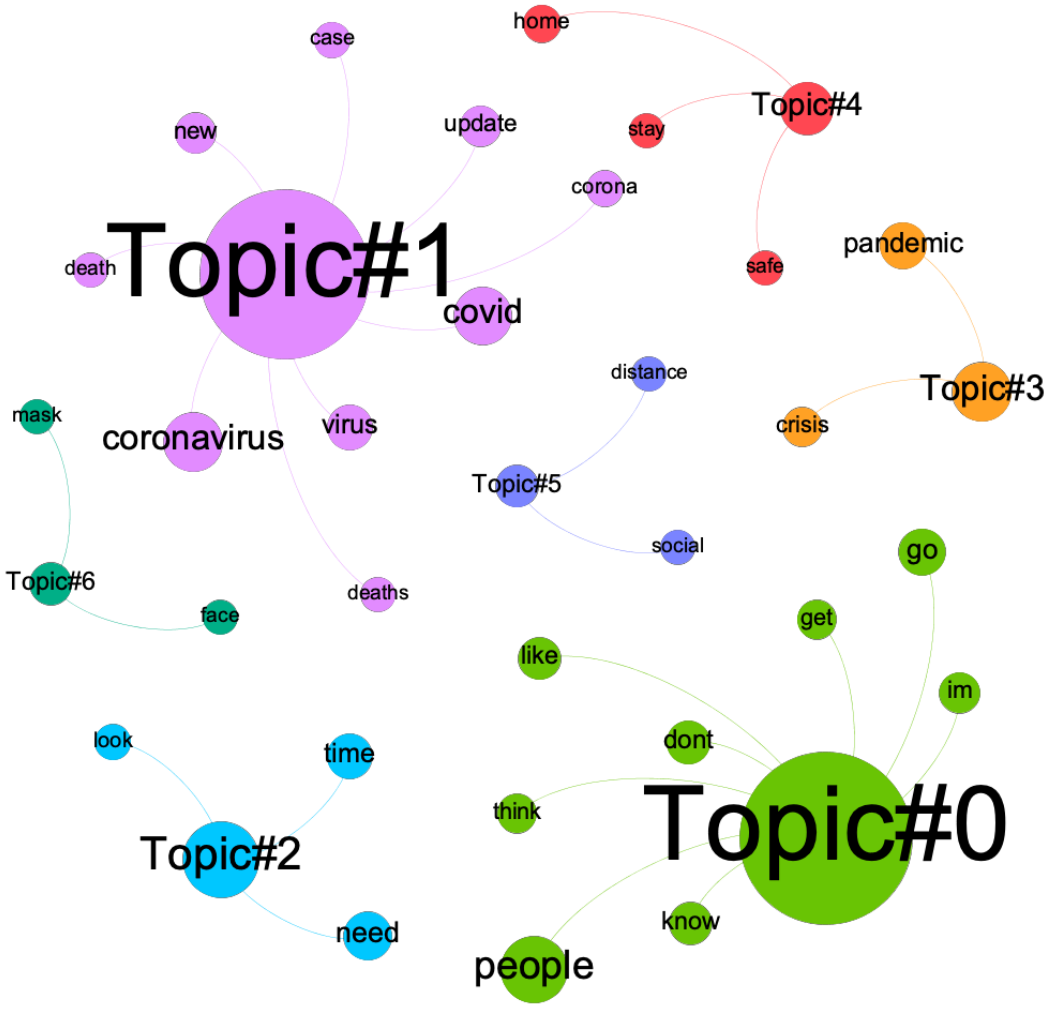

Fig. 17. Final Topic Extraction resulting from ARM with relaxed filtering.

To address these issues, ARM was performed, to identify new topics that suggest SM user attitudes in a more precise and distinct manner. For this purpose, wordset rules were generated utilizing common ARM measures, such as support, confidence, lift and leverage. Out of the 50 topics retrieved by the LDA topic extraction method, they are narrowed down to four topics (Fig. 17) by removing trivia wordsets, while showcasing few and more representative strong wordsets that infer to each topic theme. The utilization of ARM as an extra filter for the results of common topic extraction techniques, such as LDA, can aid at producing more representative and accurate results regarding SM user attitudes.

5.1.3 Sentiment Analysis of Twitter: Correlation Between Public Sentiment and COVID-19 Cases and Deaths

COVID-19 is a pandemic accompanied by an increased traffic in SM, offering great opportunities for knowledge extraction. The text contained in COVID-19 posts can act as a data source for reaching conclusions or finding correlations between attitudes or reactions of masses. Sentiment analysis is a process that measures the negativity or positivity of posts enabling practitioners and researchers to take actions for generating assertions regarding worldwide events.

There is great need for effective tools that allow timely tracking and alerting of the public in case of worldwide healthcare events such as a pandemic. Mitigating or reducing losses in terms of human lives is critical, as well as ways for reducing the impact to societies and the economy. This thesis attempts to offer such functionalities by validating trends on Twitter and correlating them with COVID-19. It also produces new insights for timely disease outbreak prediction by monitoring and evaluating multivariable correlations, such as sentiment polarity text from SM with COVID-19 actual numbers of Cases and Deaths.

The key finding of this thesis is that SM users react to increased numbers of COVID- 19 related deaths after four days by posting tweets with fading polarization. The negative trend in tweet sentiment polarity associated with COVID-19 deaths expands to 45 days, unlike that associated with COVID-19 cases, while the negative correlation becomes even stronger (polarity drops) after death numbers get published. The overall polarity of 2.146.243 tweets for the period under scrutiny (27/2/2020 to 28/8/2020) has a negative trend; people have been posting less positive/ more negative tweets as the COVID-19 pandemic kept spreading.

Further findings include that tweets one day before or after are strongly negatively correlated with COVID-19 cases on average, while tweets “after” correlation with COVID-19 cases exhibit a slightly stronger negative correlation. New cases increase while polarity drops forming a negative trend. The strongest correlation between polarity and new cases is for the same day, and the weakest correlation is one day after. Overall, there is a strong correlation between COVID-19 tweet polarity with reported cases. Also, tweets a day before or after are on average very weakly / weakly correlated with COVID-19 deaths. Tweets “before” correlation with COVID-19 deaths exhibit a very weak positive correlation, while tweets “after” showcase a weak

negative correlation with COVID-19 deaths. Overall, there is weak correlation between COVID-19 tweet polarity with reported new deaths.

PART II

Chapter 6: Literature Review

This chapter reviews literature on the following topics: Energy Balancing (Sect. 6.1); Energy Load Forecasting (Sect. 6.2) and Energy Optimal Day-Ahead Scheduling (Sect. 6.3) . Sect. 6.4 highlights the implications from the literature and develops the conceptual framework on the Energy domain.

6.1 ENERGY BALANCING

6.1.1 Terminology

Peer-to-peer (P2P) energy sharing provides a modern way to exchange energy, in a distributed manner that is more consistent with the Smart Grid (SG) concept. It allows MGs or prosumers, possibly consisting of various Distributed Energy Resources (DER), to exchange surplus energy. This is more appealing than conventional Peer-to-Grid (P2G) trading, since it is handled in a distributed manner, among peers, rather than in a centralized one, as has been the case in P2G. However, for P2P to function, a detailed IT infrastructure is needed, including sensors, energy meters and communication systems, called the Internet of Energy (IoE) (Kafle *et al.*, 2016), to establish a management tool that effectively offers the necessary interaction within the energy market (Tzovaras *et al.*, 2019). At the same time, great improvements to cost effectiveness of P2P mechanisms are required, so that they can offer financial advantages (Long *et al.*, 2018).

MGs have been tested and evaluated over the past decades through lab work, pilot demonstrations and expos, ultimately debuting in energy markets. Their operation and implementation have been radically improved, offering technological novelties, cost optimizations and other benefits. Such benefits include better electricity infrastructure, reliability and better green energy resource integration, such as Photovoltaic (PV), wind turbine etc. (Hirsch, Parag and Guerrero, 2018), or even underwater kite systems (Paiva and Fontes, 2018) through improved local and more immediate control and coordination among them. At the same time, they have contributed to reducing CO2 emissions via optimally utilizing RESs. Moreover, they

aided electricity provision to isolated areas, where a connection to the rest of the grid was previously considered either impossible, or unaffordable.

VMGs emerge profoundly as they can expand Microgrid (MG) capabilities, by releasing them from their physical form. Research attempts aim at expanding capabilities through novel hybrid designs for improving the desired control of MGs (Bintoudi *et al.*, 2018). They conceptually target solving the problem of optimized DR, scheduling within limited geographical areas with integrated RES (PV installations, Wind systems, etc.), enhancing the current functionality of MGs.

The electricity sector is being transformed as DER integration rises (Balijepalli *et al.*, 2011). At the same time, price of energy storage drops daily, enabling prosumers to be transformed into prosumagers (Sioshansi, 2019). Prosumagers can produce, consume, and store their own energy using an effective management system. Each prosumager can be considered a MG. New opportunity arise regarding P2P energy transactions, DR optimization and the concept of Virtual Power Plants (VPPs). VPP is another form of VMGs, where loads, generation and energy storage are managed optimally and efficiently, regardless of physical proximity.

Regarding power distribution, current directives from the European Commission promote the transit from a conventional power system model. To that end, the generation/load balance needs to be improved (Charalambous *et al.*, 2019). Especially under high RES integration, or when there are optimization requirements with regards to DR signals. Therefore, the energy management system needs to incorporate a variety of Computer Science related technologies. For example, Artificial Intelligence (AI) and Machine Learning (ML) techniques have proved useful for forecasting techniques. These are in turn required for various optimized schedules of DERs either in long-term basis (i.e., a day) or short-term basis (i.e., a few hours or minutes ahead) (Khuntia, Rueda and van der Meijden, 2016; Ellahi *et al.*, 2019). Internet of Things (IoT), Blockchain and Big Data have been useful for the P2P energy transactions (Livingston *et al.*, 2018).

For DR programs, there are two main categories related with this work: incentive and time-based programs (Nosratabadi, Hooshmand and Gholipour, 2017). There are time-based programs for kWh pricing. This study distinguishes the Time of Use (ToU), Real Time Pricing (RTP) and Critical Peak Pricing (CPP), where the price per kWh depends on the supply cost during different periods. According to ToU a period of a

day is divided into various pricing periods that correspond to different electricity prices, e.g., low, medium and high.

According to RTP, the electricity prices in the retail market, follow mostly the price fluctuations of the wholesale electricity market. Therefore, prices are adjusted dynamically. In both cases the fluctuation of the prices, either hourly in RTP, or in a period of hours in ToU, reflect the fluctuations of the load, that is the demand of electric energy.

6.1.2 State-of-the-art

This section is devoted to a review of the research relevant to the proposed approach, incorporating concepts from the research fields of Computer Science and Energy. More specifically, it showcases innovative attempts regarding MGs while presenting state-of-the-art implementations in DR programs. The aforementioned concepts are widely used in this work.

Work on energy management strategies attempts to improve the MG energy exchange mechanism by utilizing their DR and energy storage capacity (Akbary *et al.*, 2019). It focuses on measuring three KPIs: i) load or energy consumption patterns, ii) the energy by distributed generation resources and iii) an electricity cost reduction system.

The work in (Liu *et al.*, 2017) elaborates on energy sharing mechanisms inside a MG that comprises P2P prosumers. It takes into consideration the willingness of prosumers to shift their loads to the grid. The cost per kWh is calculated after a consensus is reached, among all the prosumers involved in this energy sharing. They also propose a day- or an hour-ahead pricing model. This is achieved utilizing a distributed iterative algorithm. For testing their framework, they use real data and compare it to using the traditional trade via feed-in-tariff, while improving the local consumption of PV generated energy.

Another general evaluation methodology targets P2P cost-efficiency energy sharing models (Zhou *et al.*, 2017). It identifies the values, estimates billing, and presents the performance index value of P2P sharing models. It comes up with the result that the Supply and Demand Ratio (SDR) model's economic performance is better than that of Mid-Market Rate (MMR), whilst both are much better than Bill Sharing (BS) models.

The work in (Moura and de Almeida, 2010) addresses the intermittence of renewable sources generation. There are various options regarding the utilization of energy sources. Also, more than one energy sources can be exploited simultaneously. This way, greater prospects are provided, since there is no unique dependency on renewables, but rather on other sources as well. In order to develop such a multi-objective system, the authors use historical data to forecast the renewables' complementary energy output. DR and Demand-Side Management (DSM) are necessary to calculate the need of intermittent capacity and to schedule production variations. They optimize the RES mix, as well as the minimum peak load share and minimum global cost, by calculating RES complementarity.

The study in (Blake and O'Sullivan, 2018) elaborates on industrial MGs. They incorporate and examine the concepts of grid connectivity, dispatchable and non-dispatchable energy sources and storage control systems. They highlight the benefits of operating DERs in MGs, such as CO2 emissions and cost reduction during times of peak loads on the grid.

Another study deals with an energy management system based on high renewable energy usage on MGs (Zhang, Gatsis and Giannakis, 2013). They deal with the intermittent nature of RES generation considering various attributes, such as loads, DR/generation costs, worst-case transaction costs etc. They use a dual decomposition method to break the problem into smaller chunks of problems, outputting the optimal solution.

Quiggin et al. use a linear programming approach in order to model a MG as a mix of RESs, storage and DR system. Their work resulted in an improved DR with balanced fluctuations (Quiggin *et al.*, 2012). Similarly, Ding et al. propose a mixed-integer linear-programming management system for industrial DERs. They optimize DER functionality by using day-ahead electricity prices to manipulate peak demands lowering the energy costs of the industrial facility (Ding, Hong and Li, 2014). Hawkes and Leach proposed a linear programming cost minimization model for designing MG with DERs. They concluded that grid-connected MGs can be more economical than isolated, or islanded ones (Hawkes and Leach, 2009).

A P2P energy trading in a MG is presented in (Zhang *et al.*, 2018). The authors consider small scale DERs for local energy trading between prosumers and consumers. A hierarchical system architectural model is proposed to identify the key elements and

technologies required for P2P energy trading. Tests show that P2P energy trading can improve the local balancing, regarding energy generation and consumption, especially when low voltage MGs can be utilized.

Another approach to energy P2P energy trading utilizing VMGs is developed with a game theoretic approach. The results show that by utilizing VMGs, improvements regarding CO2 reduction along with energy cost reduction can be achieved (Anoh *et al.*, 2020).

Vergados et al. address the issue of generating clusters of prosumers, thus forming VMGs, in order to participate to the energy market as a single entity, reducing total energy cost and forecasting inaccuracies. They used a dataset of 33 prosumers, and experimented with different clustering algorithms (spectral, genetic and an adaptive algorithm). The results show that a cost reduction can be achieved by grouping prosumers to VMGs given that VMG aggregators emerge (Vergados *et al.*, 2016).

A scalable DR framework featuring a Blockchain based near real time DR validation scheme is presented in (Cioara *et al.*, 2018). The proposed scheme allows the involved prosumers to verify the authenticity and integrity of all the DR events. Ongoing progress regarding a decentralized energy flexibility system is presented focusing on the requirements and the use cases were developed using ledger technologies (Droriano *et al.*, 2019). The usage of decentralized Blockchain mechanisms is evaluated in (Pop *et al.*, 2018), as well as their capabilities to deliver reliable, timely and secure energy flexibility in energy demand profiles of prosumers. These will be utilized by stakeholders of the energy flexibility markets, such as DSO, aggregators etc.

The state-of-the-art regarding energy forecasting requirements evaluates when prosumers are to be integrated with SGs. Energy demand is assessed individually (per prosumer) and in an aggregated way (sum of prosumers) forecasting the energy demand in order to inform the DSO about possible grid imbalances (Petrican *et al.*, 2018). A platform called FUSE, integrates multi-purpose functionalities regarding the energy sector. This platform is designed based on the state-of-the-art and the most recent real-life requirements and expectations from energy distribution systems (Stecchi *et al.*, 2019).

Another paper envisions the usage of 2nd life Li-ion batteries for enhancing distribution network operations. Batteries are to be utilized as "Storage as a service" for two use cases. In the first case the DSO manages the battery storage system aiming to increase power quality and improve Low Voltage management efficiency. In the second use case the batteries are managed by a District Energy Manager (DEM) performing peak saving and power smoothing according to a profile that is requested by the DSO. Results show that the usage of 2nd life Li-ion batteries looks promising for the power distribution network. Experimentation on DEM is ongoing (Bragatto *et al.*, 2019).

In (Pop *et al.*, 2019) researchers, address the issues regarding slow adoption of Blockchain technologies in the energy sector. Such issues are related to the low scalability and high processing overhead when dealing with near real time data from smart energy meters. A scalable 2nd tier solution is proposed, combining Blockchain ledger with distributed queuing systems and NoSQL databases. Thus, energy transactions are registered less frequently on the system, while retaining the tamper-evident benefits of Blockchain technology. Also, researchers propose a tamper-evident registration methodology for energy data transactions. Results show that the prototype looks promising regarding scalability and tampering of energy data.

The problem of inefficient and unsustainable operation of data centres with regards to their optimal fusion with the local energy grids is addressed in (Vesa *et al.*, 2020). A solution is proposed by dealing with matters related to the active participation in DR programs.

Another study reviews the most recent state-of-the-art regarding Blockchain initiative incorporating a 3-tier architecture enabling the classification of all technological solutions based on decentralized applications in the energy sector (Pop *et al.*, 2020).

In (Kong *et al.*, 2020) a particular interconnected MG system is proposed, with the aim to optimally coordinate the operation of several MGs within a market environment. It considers the technical constraints of the distribution network.

In (Liu *et al.*, 2018) authors propose a game theoretic non-cooperative distributed coordination control scheme. It addresses the multi-operator energy trading and facilitates a powerful control structure for MMGs. The difficulty of such a task is

identified, when multiple MGs are contemplated and especially when multiple beneficiaries coexist.

This proposed approach tries to facilitate such a process by identifying VMGs with similar needs. Clustering and binning techniques are utilized attempting to generate a methodology that exchanges energy more effectively, leading to improved energy sharing and balancing.

6.2 ENERGY LOAD FORECASTING

This section reviews the state-of-the-art regarding ELF (focusing on STLF) utilizing ensemble methods, combined with Machine Learning (ML) algorithms. There is a variety of use cases, conditions, and categories for ELF (VSTLF, STLF, MTLF and LTLF), as well as diverse utilized datasets and projected implications.

Researchers in (Amjady and Daraeepour, 2011), propose a method for STLF utilizing wavelet transformation, extreme learning machine and partial least squares regression. The functionality of this method involves a wavelet-based ensemble strategy that utilizes individual forecasting models. There is a mother wavelet that is combined with individual forecasting models forming various decomposition levels. The focus is on predicting hourly loads for the next day, incorporating 24 extreme learning machines, each corresponding to the hours of the previous day. Thus, it focuses on day-ahead forecast. The results/forecasts are combined into an ensemble forecast utilizing the partial least squares regression method.

A deep Neural Network (NN) methodology utilizing LSTM algorithms is presented investigating two LTSM architectures, while incorporating tests on data from a single residential costumer (Marino, Amarasinghe and Manic, 2016). The architecture under evaluation is LSTM. The tests were conducted on one-minute and one-hour resolution to perform predictions. LSTM outperformed common LSTM in both evaluation tests. However, both algorithms require testing with real-world datasets to compare their effectiveness.

Researchers in (Amarasinghe, Marino and Manic, 2017), investigate ELF performance using Convolutional Neural Networks (CNN) for some building. The conceived methodology uses the output from CNN to feed connected layers with the relevant information. The output of this approach is compared with Sequence-to-sequence (S2S) LSTM, Factored Restricted Boltzmann Machines (FCRBM),

"shallow" Artificial NN and Support Vector Regression (SVR), showing that CNN only outperforms SVR, while producing comparable results with the other methods. However, further testing in terms of datasets and accuracy is required to validate the performance of the proposed approach.

Other research attempts focus on predicting one-week ahead energy loads. They often incorporate ensemble forecasting, Artificial NNs and Deep Learning (DL). In (Sideratos, Ikonomopoulos and Hatziargyriou, 2020) they first cluster the energy consumptions with a fuzzy clustering algorithm to generate an ensemble prediction. Then, for the generated clusters, they utilize a regression method applied for modeling the local forecasting problem. A Radial Basis Function NN (RBFNN) is trained with a three-fold cross-validation and the hidden layers of the best three performing RBFNNs are utilized in order to transform the data to a four dimensional dataset. Then a CNN is employed utilizing the Adam optimization algorithm. The proposed model seems to outperform other state-of-the-art ones in two test scenarios for energy load prediction for the next seven days. That is, it focuses mainly on week-ahead forecasting.

The work in (Ribeiro, Mariani and Coelho, 2019), envisions a framework for STLF. The input data are preprocessed utilizing trend removal and normalization incorporating an optimal time window for choosing a subset of specific features. For the ensemble wavenet method learners, a variety of algorithms are used involving, bootstrapping, cross validation, stacked generalization etc. The proposed framework's output is compared with existing similar forecasting techniques, such as MLP, single wavenet etc. For the validation of the results a 10-fold cross-validation is employed to highlight the performance of the proposed framework. However, the results from the tested datasets seem confining or inconclusive. The implementation on one dataset shows improved performance over other techniques, but on another one shows higher accuracy, but only for a specific time window.

In another study, the researchers try to solve the problem of low interpretability of confidence level that accompanies the STLF, deriving from the great uncertainty of real-life individual energy loads (Li, Che and Yang, 2018). The Subsampled Support Vector Regression Ensemble (SSVRE) is utilized, aiming to offer improvements regarding computational accuracy, naturally enhancing the efficiency of the load forecast attempt. Once SSVRE model diversifies each of the SVR individual

ensembles, a novel swarm optimization learning based model is utilized. SSVRE has good results in terms of performance and uncertainty when attempting STLF. For this, however, certain conditions are required, such as to develop an effective strategy for building the SVR-based short-term forecasting model.

In (Yang, Hong and Li, 2019) the authors propose a deep ensemble learning probabilistic ELF framework aiming to address the aforementioned issues regarding the individualized load profiles. Individual load profiles are clustered and then a multi-task representation learning process is utilized. The results indicate that this framework incorporates improvements regarding the feature learning process of the formed load profile groups. Although, the issues of requiring a large amount of data and the high complexity of the approach remain.

An incremental heterogeneous ensemble model is incorporated in (Grmanová *et al.*, 2016). This model takes into consideration features of the smart meter energy data, such as seasonality and drifting. An incremental training of the model is performed, with the ability for parallelization and great robustness, highlighting the proposed ensemble method suitability for the specific use case. Vast amounts of data from smart meters are required though and a time-frame of 24 hours is considered.

Another paper uses Takagi-Sugeno-Kang neuro-fuzzy model to train and then forecast a day-ahead energy load profile, utilizing the LOcally LInear MOdel Tree (LOLIMOT) model. The study uses a real load profile aiming to refer to geographically partitioned case studies. Also, compartmentation of power system advantages is presented to highlight the importance of load forecasting in smaller regions. It takes into consideration other data features, such as temperature data etc. This approach presents data preprocessing techniques and concludes that running LOLIMOT is able to determine all the necessary parameters for testing the Takagi-Sugeno-Kang neuro-fuzzy model. Therefore, a flexible network topology that results from the trained model eases the task of load profiling extraction (Malekizadeh *et al.*, 2020). A particular preprocessing on historical data is required though.

Ensemble methods are also widely used in other energy domains. In (Khairalla *et al.*, 2018) the researchers investigate the Stacking Multi-Learning Ensemble (SMLE) model for a variety of time horizons for improving accuracy of predictions for Global Oil Consumption (GOC). They attempt to predict future values based on non-linear timeseries by combining SVR, Back Propagation NN (BPNN) and Linear

Regression (LR) under a four-phase ensemble method. It involves a generation, pruning, integration and an ensemble prediction phase. The proposed method is evaluated in a single and multistep time prediction horizon. The results indicate that the proposed method outperforms most similar state-of-the-art techniques. It should be stated that although it is a robust approach with prospects to outperform the average model, it is not possible to know apriori which model will perform best.

Researchers in (Wang, Wang and Srinivasan, 2018), exploit and test the Ensemble Bagging Trees (EBT) method utilizing data from building occupancy levels, energy meters and meteorological conditions. Testing is conducted using hourly electricity values and the metric for evaluating the accuracy of the proposed method is Mean Absolute Error (MAE). The results show that there are improvements both in computation times and in prediction accuracy, rendering this method a good choice for applications related with system fault detection diagnostics. However, the level of improvement varies over the various testing cases. The quality of the dataset may significantly impact the improvement of the proposed approach. Many researchers attempted to utilize ensemble learning for load forecasting in power systems in urban areas. For example, a two-level learning framework is proposed in (Wang *et al.*, 2020), combining clustering with LSTM and Fully Connected Cascade (FCC). Multiple LSTM models are trained with clustered data and then the FCC model is used to merge these LSTM models. The FCC model is trained and converged by a modified Levenberg-Marquardt (LM) algorithm. The framework was tested on two publicly available datasets used for short-term and mid-term predictions. MAPE is utilized to highlight the best performance from the proposed framework when compared with several other schemes. Still, the drawbacks of the applied first and second-order gradient methods are inherited in the proposed approach.

Another work aims to perform STLF for two-days ahead, while utilizing functional clustering and ensemble learning. Functional clustering is utilized for grouping similar daily load values. Ensemble learning of machine models is utilized to integrate all the models resulting from several load groups in order to improve the prediction accuracy of the proposed method. The results are presented through a real case study concerning load consumption patterns, showing that two-days ahead load forecasting can have a practical use when related to economic activities (Rodrigues

and Trindade, 2018). Yet, a large amount of data up to two years in advance is required for this method.

6.3 ENERGY OPTIMAL DAY-AHEAD SCHEDULING

This section presents the state-of-the-art on energy optimization and commonly utilized methods for minimizing variables such as operational cost while including models such as MGs, renewable energy sources, or parameters such as flexibility and more.

Multi-objective optimization problems usually involve many objectives to be optimized with many inter-dependencies. It is hard to discover optimal solutions that satisfy all objectives. Analytical and classic numerical methods entail mathematical calculations and search values that are clearly defined. On the other hand, heuristic methods negate these requirements attempting to find global optimal solutions. Real-work multi-objective optimization problem optimal solutions require utilizing a variety of methods to output the final results. These include Apriori, Pareto-dominated, interactive and new dominance methods. In energy sector multi-objective optimization is quite common for solving problems in environmental protection, energy saving, cost reduction, emissions reduction and more. For each case multi-objective optimization methods yield benefits and drawbacks generating prospects for future work (Cui *et al.*, 2017).

Global warming and environmental parameters introduce various constraints regarding the absorption of Distributed Energy Resources (DER) as well as in the economic aspects. A multi-objective optimization model is developed to analyse the optimal operating strategy of a DER system, while combining minimizations of energy cost and environmental impact with the latter assessed in terms of CO2 emissions. The pilot for validating this model is an eco-campus in Japan and the Pareto front of optimal solutions results from the compromise programming method. The electrical and thermal demands of the eco-campus consider the existence of photovoltaics (PVs), fuel cells and gas engines. Results showcase that when minimizing energy costs, the CO2 emissions increase, while the DER system's functionality becomes sensitive when more weight is applied to the environmental objectives. In addition, considering options such as bilateral electricity exchange (buy-back programs), utilization of

biogas or taxation on carbon emissions, affects the DER system's operation accordingly (Ren *et al.*, 2010).

On the consumer's side, the concept of zero energy and low energy consumption buildings has become a research field with many applications and ecological benefits. Both researchers and professionals agree that the effectiveness of these structures is quite often defined by the level of Renewable Energy Resources (RES) utilization. A comparative study deals with the design optimization techniques for integrating RES systems in such structures. The research approach considers the Genetic Algorithm (GA) for solving a single objective optimization problem and Non-Dominated Sorting Genetic Algorithm (NSGA-II) for a multi-objective optimization problem. Principles and parameters of building energy and renewable energy systems interact with one another generating variables and constraints for the optimization process. The pilot building for this approach is the Hong Kong Zero Carbon building. The results from optimization improvements showcase that when a RES system exists, the optimization process yields better results than the current configuration of the pilot building. Also, when a single objective needs to be strictly fulfilled, the single objective optimization yields the best results. When a variety of design options with or without compromises should be presented, the multi-objective optimization becomes ideal (Lu *et al.*, 2015).

Continued supply and fulfilment of local requirements in heating, cooling and electricity demand plays a vital role in the modern energy management systems. This should happen in an environmental and economic manner. DER systems, if efficiently utilized, may output great results in terms of carbon emissions reduction and optimal energy management aiding in tackling arising issue with climate change. A novel multi-objective framework compares two methods for effective design of DER systems considering Total Annual Cost (TAC) and carbon emissions. The first method yields a parallel sizing of the two objectives, while the second method incorporates predefined technologies and system capacity. The optimization process is evaluated in three scenarios integrating technologies, the two methods and a case study. Findings reveal that when a DER system connects with a Microgrid (MG), energy storage and a heating network outperforms the other two scenarios. Also, the first method yields better results regarding environmental emissions and cost reduction while offering more options in problem design (Karmellos and Mavrotas, 2019).

Modern buildings integrate many intelligent control systems, enhancing the occupants' household experience and comfort. The main issue regarding the best possible occupant experience includes the correlation of energy consumption with discomfort. Each of these two variables counterbalances the other. A multi-agent based control framework attempts to enhance smart building management defining energy consumption and occupant comfort levels as the two objective functions of the problem. The results form Pareto optimal solutions utilizing Multi-Objective Particle Swarm Optimization (MOPSO) and Weighted Aggregation (WA). The variety of trade-off options with regards to energy cost and occupant comfort generate opportunities for making better decisions on building design and management (Yang and Wang, 2012).

MG operation often involves the existence of RES. A scalable quantitative framework attempts to deal with the intermittent nature of RES on MG integration. When RES is heavily utilized a novel chance-constrained stochastic programming model considers three policies. One of the policies utilizes a fixed amount of RES output during the whole time of examination, while the other two utilize certain hours and all operating hours. A combined Sample Average Approximation (SAA) algorithm solves the problem, showcasing that the policy utilizing all operating hours introduces more restrictions for optimization, although there is peak RES utilization. In addition, this research presents possible energy management improvements, when PVs are combined with or covering demands of other fuel-based, or DER units in power outage periods. The minimization of operational cost results from sending dispatch signals to each existent resources (fuel-based or any DER units) (Marino *et al.*, 2018).

Energy optimal scheduling on MGs is a topic that lately attracts much research attention. Important issues related with that topic include power balance on normal and peak demand periods (outages). A novel approach considers the intermittency of RES generation and that of the demand, outages of distributed generators and cases of islanding. The problem is solved with a multiple chance-constrained scheduling model. The model along with the parameters mentioned, also considers, outages on energy storage mediums such as batteries. Chance-constraints transform utilizing control variables attempting to decrease the model complexity, while probability distribution functions handle the available energy reserve in a variety of conditions.

For example, when there is battery outage, or islanding. These functions introduce an index for Probability of Reserve Sufficiency (PRS). The model is validated and evaluated for a MG under different conditions recording PRS readings (Sefidgar-Dezfouli, Joorabian and Mashhour, 2019).

Introducing storage in MGs plays a paramount role in generating a variety of extra services. Yet, offering such storage services along with effective grid management is a demanding task. A chance constrained optimization approach considers electrical and thermal battery sources for improving grid reliability. The testing incorporates loads of 5-minute intervals for introducing randomness in energy loads/fluctuations. Batteries can charge and discharge fast. This characteristic makes them appropriate for managing energy flexibility constraints that may involve photovoltaic generation or random peak demand. Chance constraints mitigate issues with flexibility reducing errors while common variables state dependencies of thermal and electrical storage systems. Findings show that this approach manages energy fluctuations taking advantage of flexibility for a more reliable grid operation (Ciftci *et al.*, 2019).

Extensive usage of RES such as photovoltaic or wind power are essential for the transition to a more sustainable MG operation. A novel probabilistic optimization framework envisions a more efficient MG management. It utilizes chance constrained programming and a bi-objective approach involving RES integration and customer load profiles. Jointly distributed arbitrary variables capture the probabilities of reaching the expected energy load while forcing the operational cost below a certain threshold. This approach utilizes an improved hybrid Artificial Bee Colony (ABC) and Differential Evolution (DE) algorithm to optimize energy management of a MG. Results are validated with a sample average approximation technique that compares findings with a scenario and Monte Carlo stochastic programming approaches (Zare *et al.*, 2016).

Pollution distribution and reduction are parameters that should be optimally handled in hybrid energy systems. Economical and other environmental aspects can be enhanced with the incorporation of a DR program. The cost minimization of a hybrid energy system constitutes an objective function, while minimization of CO2emissions constitutes another. Common constraints or variables in functions may cause counterbalancing effects on the final optimization process. A multi-objective

optimization problem is solved outputting the most efficient solutions considering trade-offs. These may be reported in the form of DR signals. To validate the results and expose the benefits of this approach, a fuzzy satisfying technique chooses the optimal solution and a DR program outputs possible benefits with environmental and economic indicators (Nojavan *et al.*, 2017).

6.4 CONCEPTUAL FRAMEWORK AND IMPLICATIONS

This section distils key findings from literature to elaborate and present possible implications of the conducted research on the Energy domain.

6.4.1 Performing Energy Balancing on P2P and VMG to VMG level

The approach presented can be implemented as a standalone toolkit offering the functionalities described in this study. Based on the current state of ongoing research on this domain, possible use cases are identified that can act as an expansion, or can be incorporated as external components, utilizing or being dependent on the output of the conducted research.

The first use case envisions potential cost minimization or profit maximization when transferring energy from Virtual Microgrids (VMGs) handled by aggregators to the virtual portfolio of the Distribution System Operator (DSO). These grids are conceptualized as clusters, which have specific parameters regarding consumptions, productions, KPIs etc. Then the DSO can be more accurate regarding Distributed Energy Resource (DER) load distribution for calculating profits or minimizing functional costs.

The second use case addresses the opportunities arising from VMG2VMG transfer of energy, when Renewable Energy Sources (RESs) can be utilized in a large scale. VMGs are conceived while enabling trade-offs of energy in-between prosumers, offering optimal energy distribution in close proximity. They are also considered for maximizing savings from main grid tariffs and reducing green gas emissions. Local ecosystem VMGs constraints can be enforced to ensure fairness and stability of such an energy management system.

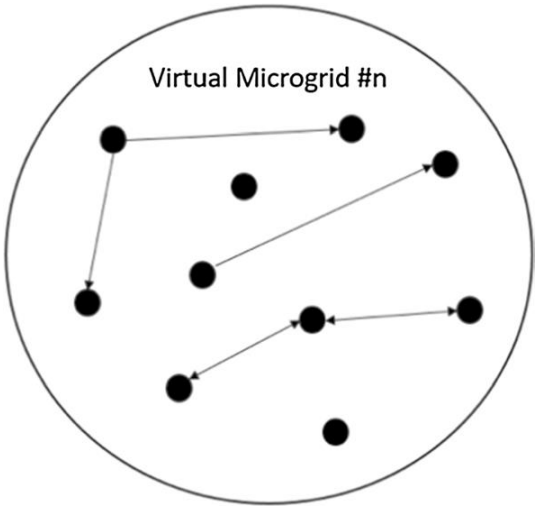

Fig. 18. DR Distribution on P2P level, optimized single VMG schema.

The third use case refers to VMGs capability to utilize P2P transfer, meaning that energy can be traded from prosumer to prosumer a.k.a., energy trading to cover Demand Response (DR) needs of prosumers. Fig. 18 shows a VMG perceived as a cluster #n while black dots indicate prosumers. There are directed edges that inform the direction of the DR energy transfer on P2P level between various nodes. Each node represents a prosumer that agreed to participate to an energy-sharing program. Energy can be exchanged uni- or bilaterally.

6.4.2 One Step Ahead Energy Load Forecasting

Experimentation was conducted in all cases under the same conditions regarding forecasting horizons, model configurations, metrics and objectives. This study forecasts energy loads in the short term, utilizing a dataset that contains real-world entries (energy load measurements from a smart home) associated with energy load values. The approach narrates a process that is executed under the same characteristics. After reviewing the literature, it distinguishes and presents commonly used prediction algorithms and ensemble methods.

There are many attempts for performing more accurate STLF. For example, the improved complete ensemble empirical mode decomposition with adaptive noise (ICEEMDAN), Grey Wolf Optimization (GWO) and Multiple Kernel Extreme Learning Machine (MKELM), called ICEEMDAN-GWO-MKELM (Li, Qian and He, 2020) that is tested with datasets of half-hour load data of four months from five regions in Australia or (Bendaoud and Farah, 2020), that utilizes a CNN using a two-

dimensional input showcasing results of 0.80% MAPE for OSA-ELF using four years (2013-2016) of 15-min national electricity load data from Algeria. Another study (Li *et al.*, 2021) develops a CNN LSTM with Selected Autoregressive Features (SAF) named CLSAF that employs autoregressive feature selection, exogenous features selection, and a "default" state for handling overfitting when there is high energy volatility. It achieves up to 25% higher forecasting accuracy compared to a persistence model and uses a dataset of 59 individual apartments, eleven floors, and one building, for three time periods in 2019 of a building in Manhattan, New York with one-hour aggregated energy loads. Finally, in (Farsi *et al.*, 2021) authors parallelize LSTM and CNN models to conceive a model named PLCNet. They utilize two real-world data sets, one containing hourly load consumptions from Malaysia and one daily power electric consumptions from Germany. Performance was evaluated with RMSE, MAPE and R-squared. PLCNet improved the accuracy from 83.17% to 91.18% for the German dataset and achieved 98.23% accuracy for the Malaysian dataset.

It should be noted that a comparative analysis with other STLF attempts for evaluating the performance of the proposed approach for OSA-ELF is not possible. Yet, it was chosen to present the results in SMAPE. When values of that metric reside between 0 and 10% are at its optimal indications (Botchkarev, 2018). According to Appendix R that contains the summarized performance evaluation of the proposed OSA-ELF, it may reach SMAPE values of up to 6.489%. In addition, for a comparative analysis it requires to run experiments on the same datasets under the same conditions, a case that would force a deviation from the goals and motivation of this work.

The main aim is to present an approach that can address in a uniform manner any given energy load prediction task, regardless the ELF category (VSTLF, STLF, MTLF and LTLF). To that end, OSA-ELF is utilized, claiming that the performed prediction depends on the resolution and context of the historical energy load input data experimenting and presenting results for STLF. More specifically, this study implements and tests five forecasting models (MLP, LSTM, XGBOOST, SVR and LR) combined with three ensemble forecasting models (EAP, EWA and EPE) under one proposed single-step forecasting approach. The experiments were tested using three widely used metrics (RMSE, MAPE and SMAPE) for timeseries forecasting accuracy evaluation, while recording their execution times (ET). The findings that arise from field data, and the approach's exploitation are presented in detail, aiming at

constituting work that can be utilized by researchers and practitioners alike, as a point of reference regarding ELF.

The implications of this study foresee more accurate power system operation, design and organization. ELF allows Energy sector stakeholders (aggregators, utility providers etc.) to model electricity consumption and prepare for future power loads. The energy distribution infrastructure rapidly evolves requiring faster response times and accelerated energy resource allocation. ELF can benefit energy production cost reductions, energy price estimation and system capacity planning. Moreover, improving accuracy in STLF can lead to better energy scheduling considering the intermittent nature of RES and their energy capacity restrictions. ELF is important for the precise and uninterrupted operation of power systems regardless of their scale (household, industrial facility etc.). This research has been conducted to improve household ELF accuracy in the short-term. To that end, it may benefit energy operational security and power system savings by enhancing planning and operation of any power system infrastructure.

6.4.3 A tri-layer Energy Optimization Framework

This study aims to produce a framework able to optimize the day-ahead energy load scheduling as a holistic approach, taking into account both the aggregator and the consumer's view. At the moment, it considers a DR scheme that manages the interaction of two stakeholders (aggregator and consumer). In its expanded version it should be able to handle also the DSO and validate a three-stakeholder interaction through various use cases. In that case, this framework will be able to identify, test and evaluate a wider range of requirements for a modern energy distribution network. The proposed framework models and offers insights for optimal energy management acknowledging the possible integration of a variety of parameters such as RES, storage units and more. Integration of such parameters may offer enhanced energy grid elasticity, since there are more options for handling issues related with peak loads, broken power grid links etc. Such parameters have been identified and presented in literature and can be added in an expanded version of this work adjusting the objective functions per stakeholder.

Solving each problem in a stand-alone manner allows interactions between consumers (P2P level) while allowing DSO to validate DR programs for the aggregator. Constraints and constants are combined for each case affecting

individualized solutions. Novel solutions for a more efficient monitoring and management of energy grid are very important. Nowadays, aggregators have to manage hundreds of thousands of consumers constituting very large portfolios showcasing the importance for more efficient energy load scheduling. DSOs also need reliable tools for monitoring, validating and enforcing when deemed necessary a constant and reliable energy grid functionality.

In detail, aggregators should be able to efficiently broadcast a day-ahead energy load schedule for their portfolios, based on SMPr, flexibility setpoints for consumers and more. The consumer requires 24/7 asset management and an efficient framework that enables a reliable DR signal confirmation scheme, ideally with dynamic participation. The DSO should retrieve that information from the aggregator and approve the DR portfolio schedule. In case of an energy grid issue, DSO should be able to identify the problem with the power distribution network and notify the aggregator in order to comply with the adjusted day-ahead energy load and flexibility requests. On that respect, this part of the thesis will be able to assist or become a point of reference for both the industry and the academia, when referring to DERs, P2P energy transfer models and multi-objective optimization on distribution of energy considering multi energy assets and actors within an energy ecosystem.

Chapter 7: Research Design

This chapter describes the design adopted by this research to achieve the aims and objectives stated in Sect. 1.3. Sect. 7.1 discusses the methodology of this study, and the stages by which the research design was implemented. Sect. 7.2 elaborates on the data sources and preprocessing steps. Sect. 7.3 describes how the data was analysed (methods/algorithms). Finally, Sect. 7.4 discusses the limitations, ethical considerations and potential threats to validity of the conducted research.

7.1 METHODOLOGY AND RESEARCH DESIGN

The research design involves a mixed quantitative and qualitative approach. Although most of the research tasks experimented incorporating algorithmic outputs and mathematical calculations (quantitative) for reporting on results, there are qualitative parts yielding non-calculable elements and attributes. For example, in Task 6 there are research design features that exposed qualitative characteristics (Table 29).

Table 29. PART II Types of primary research design per Research Task.

Task	Type	Description
4	Quantitative	Straightforward experimentation on data regarding energy readings.
5	Quantitative	Straightforward experimentation and comparison analysis of performance on data regarding energy readings.
6	Mixed	Straightforward experimentation on power grid data. Yet, there are features such as occupant discomfort that are subject to arbitrary conception.

The data collection methods involve a variety of primary data sources, highlighting the extended endeavour for data retrieval and procurement requirements of this study (Table 30). The detailed presentation of research data can be found in Sect. 7.2.

Table 30. PART II Data collection methods per Research Task.

Task	Method(s)	Description
4	Documents and records	Retrieval of timeseries dataset from a pilot power grid.
5	Documents and Records	Retrieval of timeseries dataset from a pilot building.

6	Documents and Records	Retrieval of timeseries dataset from a pilot power grid and API calls.

7.1.1 Task 4

This study elaborates on an interdisciplinary approach for energy balancing, based on Energy and Computer Science introducing a methodology and testing the validity of its results. The approach utilizes ML preprocessing concepts, such as clustering, binning, and a customized Exhaustive Balancing Algorithm (EBA). It exploits clusters and bins of energy consumption.

Fig. 19 demonstrates the conceptual layers of the architectural interaction between VMG2VMG, VMG to aggregator/s, and aggregators to DSO. The approach focuses on the VMG and Prosumer layers running multiple simulations to support a methodology for local and global balancing.

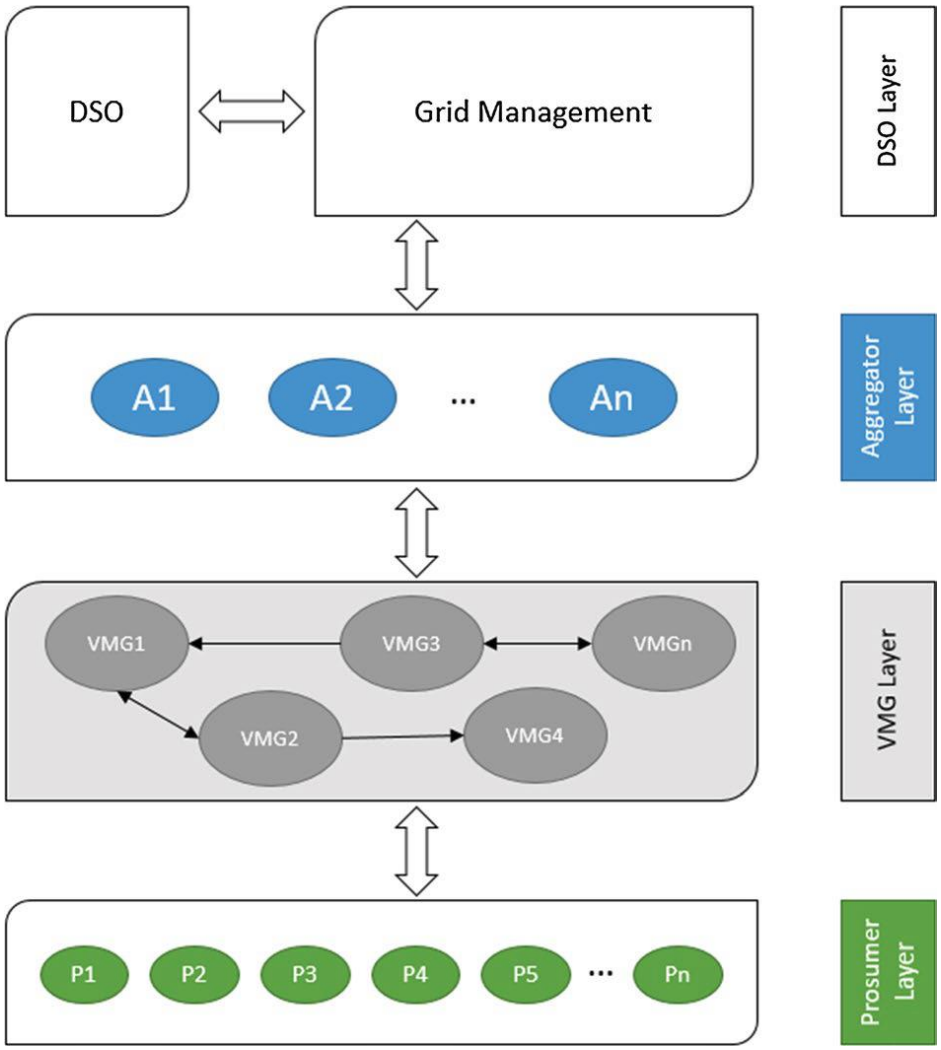

Fig. 19. Task 4 methodology flowchart.

By showcasing a solution in the form of a VMG formation problem, this study also envisions to reduce costs or increase profits. To that end, the proposed methodology enables a possible DR strategy that can be exploited given that VMG aggregators emerge in the energy market/ecosystem. This tackles the issue of energy balancing, based on unsupervised learning. The goal is to improve portfolio management, using a visualization-clustering/binning approach offered as a service, while enhancing DR functionality and efficiency for energy stakeholders i.e., aggregators and DSOs (Papazoglou and Georgakopoulos, 2003).

The methodology addresses local P2P Balancing and global VMG2VMG Balancing with four simulations that test and validate such an approach. These simulations, involve data preprocessing, rule enforcement and decision making, using an EBA. Simulations exploit data from a full day of energy usage, splitting load-demand measurements, as well as grid inputs/outputs, in the 24-h scale. Thus, a "timestamp" variable is utilized, ranging between 1 and 24.

7.1.2 Task 5

This study envisions an approach by combining and experimenting on timeseries ensemble methods and forecasting algorithms. The proposed approach builds upon the knowledge acquired aiming to improve the forecasting efficiency for STLF. The testing and evaluation of the results is limited to the day-ahead ELF since the focus is made on expanding the state–of-the-art on this domain (Koukaras, Bezas, *et al.*, 2021).

A new approach for OSA-ELF is conceived by generating a formula that utilizes multiple steps backwards (1 to n steps) generating n forecasting models. This approach aims to evaluate the backwards steps; thus, exhaustively utilize the historical data through generating different n models that all perform forecasting to the same one-step ahead time horizon (6).

$$forecast\ (t+1) = model1\big(value(t-1)\big),$$

$$forecast\ (t+1) = model2(value(t-1), value(t-2)),$$

$$forecast\ (t+1) = model3(value(t-1), value(t-2), value(t-3)),$$

...

$$forecast\ (t+1) = modeln(value(t-1), \dots, value(t-n)) \quad (6)$$

where t the time step and n the maximum available time steps.

Then, the resulting multiple forecasted values for the same time step can be ensembled, choosing any of the aforementioned ensemble methods but not limited to them; to produce a single forecast value. All that mentioned, the conducted experimentation implements the aforementioned new approach for OSA-ELF by incorporating the ensemble methods combined with the prediction algorithms showcased in Sect. 7.3.4. The resulting implementations of ensemble methods utilizing prediction algorithms are evaluated in a comparable way using accuracy metrics; the ones also detailed in Sect. 7.3.4.

According to Sect. 6.2, there are various state-of-the-art attempts for timeseries prediction on ELF.

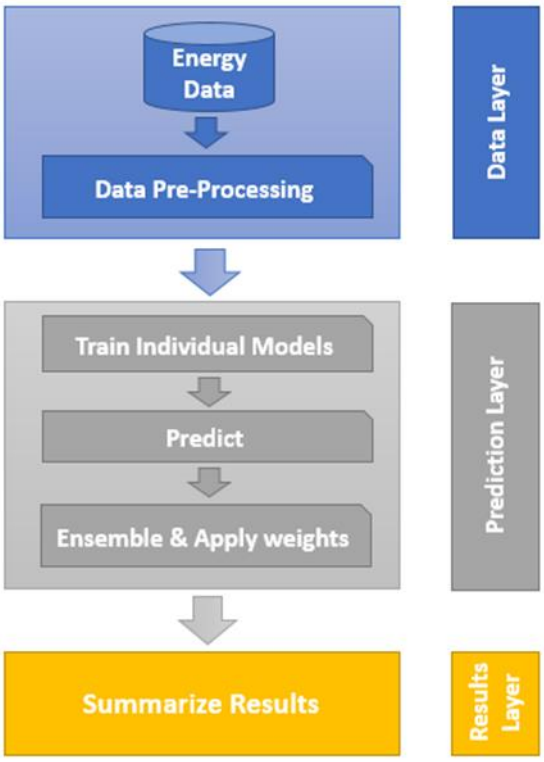

Fig. 20. Task 5 methodology flowchart.

The architectural design of the proposed method is depicted in Fig. 20. Three layers are distinguished that constitute the overall context of this study's approach. In the Data Layer three processes are performed, data loading, data pre-processing (filling missing values with linear interpolation (Noor *et al.*, 2014)) and supervised transformation turning a timeseries problem into a supervised learning problem. Such

a problem is formed by a series of input variables (X) and an output (Y), with the utilized algorithms attempting to learn the mapping function from X to Y. In the Prediction Layer, three processes are also performed, the training of individual models utilizing various types of supervised learning algorithms/methods (MLP, LSTM, XGBOOST, SVR, LR), then the prediction of energy values and finally the ensemble on the output values of the individual prediction in accordance with historical values.

The machine learning algorithms were implemented with Scikit-learn (Pedregosa *et al.*, 2015) and Tensorflow (Abadi *et al.*, 2016). In the Results Layer, a results summarization is performed for all models presenting a single prediction result for each forecasted timestamp. In the case of the OSA-ELF for a day's values, 24 separate models are developed for each hour by using LR, SVR, GBT, MLP, and LSTM as prediction models. For this approach, a separate model is developed for each forecast time step, which is applied to all the aforementioned combinations of experimentation process. The detailed steps of the experimentation process are outlined below:

i) Read three months energy load data for spring months of year 2018 (March, April, and May) from the nZEB Smart Home.

ii) Utilize a function that transforms the sequential timeseries data into a specific format.

iii) Separate data to training and testing sets.

iv) Scale the training and test data for each of the 24 hours with MinMaxScaler function (Pedregosa *et al.*, 2015).

v) Create a different model for each hour of the day utilizing MLP, LSTM, XGBOOST, SVR and LR abiding with the proposed timeseries forecasting approach. Append and save the predictions, inverse the results from scaling.

vi) Apply ensemble methods to the results of all the predictions and get a single final value.

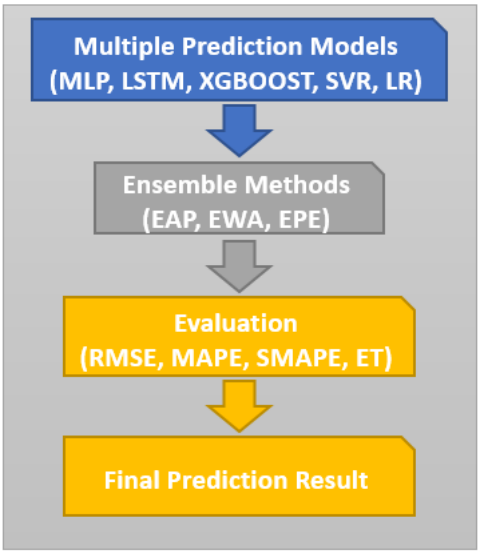

Fig. 21. Process flowchart of forecasting methodology.

By observing Fig. 21, it is evident that separate models are being developed that expose different forecasting characteristics. Some of the models are dependent on the different time horizons of the values while others on weights which are applied depending on the past time horizon that each of the individual models are based on. In addition, the proposed timeseries approach ensures the further the time intervals are back in the time scale, their applied weight values diminish.

7.1.3 Task 6

This section states the goals of the methodology and the utilized methods/algorithms. It also presents an overview of the steps of the proposed methodology.

The topic of advanced DR optimization is addressed, introducing a methodology and testing the validity of its results. Tasks such as the distinction of optimal load dispatch can be quite demanding, since they involve great randomness of events. For that reason, this optimization issue is approached by decomposing the initial problem and distinguishing two scenarios of functionality for the optimization methodology. The first scenario explores the possibilities of cost minimization along with maintaining acceptable levels of discomfort for the consumers. The second scenario enables portfolio cost minimization for the aggregators.

Multi-objective optimization problem solutions should be dependent on the type of problem and the envisioned output. In case problems are small and can be expressed

in a linear way, any solver can compute the optimal solution relatively quickly. If that applies, a good practice is to search for precise solutions. If all the non-dominated outputs can be retrieved, e-constraint method (Palli *et al.*, 1998) is an appropriate choice, otherwise, an algorithm with weighted sums can be used instead. On the other hand, if problems are large and can be expressed in a non-linear way, solvers take too long, and it becomes difficult and slow to extract the optimal solution even for single-objective problems. To address these issues, metaheuristics come into play, such as MOPSO (Alvarez-Benitez, Everson and Fieldsend, 2005), Non-Sorted GA type three (NSGA-III) (Yuan, Xu and Wang, 2014), or Strength Pareto Evolutionary Programming (SPEA2+) (Kim *et al.*, 2004) and more.

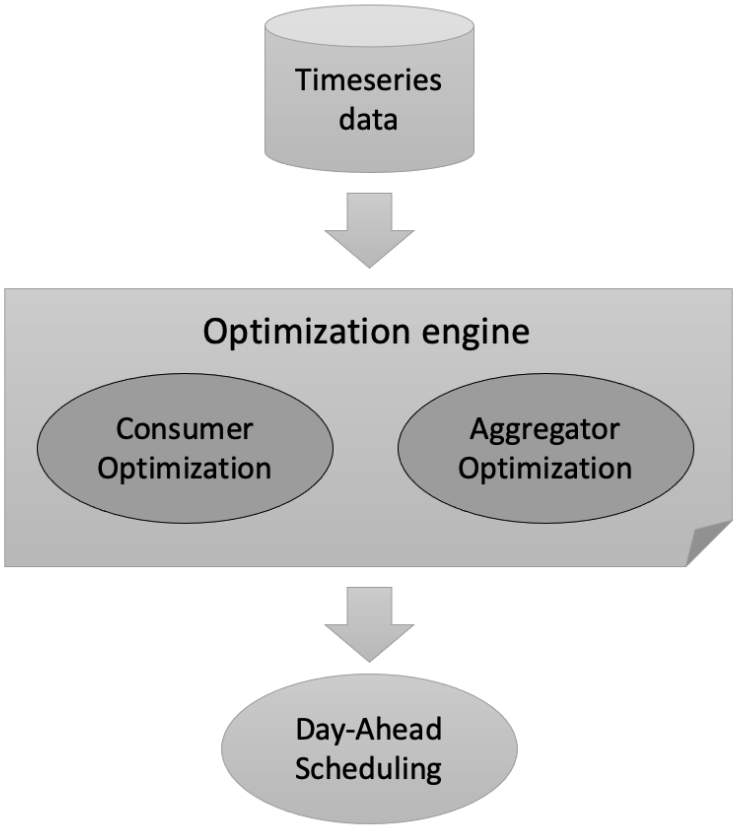

Fig. 22. Task 6 methodology flowchart.

According to Fig. 22, the proposed optimization engine consists of two optimization problems. The optimization engine takes as input historical time series data regarding consumers and outputs a day-ahead optimized energy schedule.

i) Consumer that poses as a bi-objective minimization problem. The minimization of cost and the minimization of consumer discomfort;

ii) Aggregator that is a single-objective minimization problem. The minimization of portfolio cost;

7.2 DATA SOURCES

7.2.1 Task 4

The dataset to test the proposed approach contains records corresponding to prosumers with various attributes, as listed in Table 31. These data were generated by capturing various measurements from CERTH/ITI nearly Zero Energy Building (nZEB) Smart Home (*SmartHome / ITI*, 2020). This nZEB incorporates a wide variety of novel technologies. It is a rapidly evolving prototyping infrastructure equipped with IoT devices and technologies that allow to measure and validate many experimental conditions attempting to imitate real life living conditions.

Table 31. Task 4 initial (raw) dataset description.

Attribute	Type of value
A	Meter Id
B	Quarter-hour progressive number in the day
C	Latitude of prosumer and longitude
D	Total Active Energy (TAE) delivered from the prosumer, injected into the grid
E	TAE absorbed by the prosumer from the grid
F	Fixed value (15), indicating that data gathering is carried out every 15 minutes
G	Datetime

The data for experimentation were retrieved from the extended nZEB API. That way access is granted to measurements from hundreds of devices within the Smart Home. The energy measurements (power generation and consumption to and from the grid, respectively) were collected for the period between 2017 and 2020. The rationale for choosing this data source for simulations resides in that nZEB excels at mitigating residential housing energy profiles by incorporating distinct housing units and rooms. That way distinct measurements can be retrieved, being either energy generation or consumption (Vafeiadis *et al.*, 2018). Therefore, according to preliminary results from ongoing research, this study distinguishes around 100 possible residential energy profiles along the observation of four years of energy generation and consumption patterns based on past research results (Adamopoulou, Tryferidis and Tzovaras, 2016). This admission forms the baseline approach for considering these energy residential

profiles to mimic individualized prosumers to be used for the simulations. To that end, a conceptual, yet very close to real life conditions, prosumer dataset emerges.

Furthermore, python's pandas library (McKinney, 2010) is utilized for enforcing various data preprocessing techniques, like removing duplicates and missing values as well as data transformation and reduction needed to normalize and clean the dataset. The raw dataset is a 980 MB.csv file with attributes listed in Table 32.

The pre-processed dataset is a 33.781 MB.csv file that contains 815.208 rows of data, which represents entries for 94 conceptual prosumers. The raw dataset time resolution was in quarters of an hour.

Table 32. Task 4 experimentation dataset description.

Attribute	Description
A (ID)	Prosumer id
B (Datetime)	Datetime in format ("%Y-%m-%d %H:%M:%S")
C (E0a)	TAE absorbed (in Watts) by the prosumer from the grid is signed as negative, while the TAE delivered from the prosumer to the grid is signed as positive.

Yet, it is aggregated to match the one-hour resolution, for enhancing the observation clarity for the presentation of the results. The dataset is transformed to match the requirements of this approach. Measurement units are also added. As an extra preprocessing step, coordinates attributes are discarded since the approach uses data from various prosumers in an assumed close vicinity. Finally, attributes C (E0a) with D (E0) are merged into attribute C (E0a) (as shown in Table 32).

7.2.2 Task 5

The testing was conducted based on the electricity consumption data from CERTH/ITI nZEB Smart Home. This Smart Home is a rapidly prototyping infrastructure that utilizes multiple novel technologies. It tries to resemble the majority of real domestic life experience while implementing experiments of actual living conditions. To achieve that it incorporates a number of Internet of Things (IoT)-based technologies that elaborate on and form Energy, Health, Big Data, Robotics and Artificial Intelligence services (*SmartHome / ITI,* 2020). Since, it is constructed by state-of-the-art materials and is developed as a low-carbon emissions building (Luo and Chen, 2020) accompanied by novel Information and Communication Technology

(ICT) solutions it constitutes one of the best use case scenarios for active testing, validation and evaluation of the proposed ELF approach.

The sampling time of data are per quarter and for the training and testing are aggregated per hour. A graphical representation of the Smart Home's electricity consumption during the spring season is represented in Fig. 23.

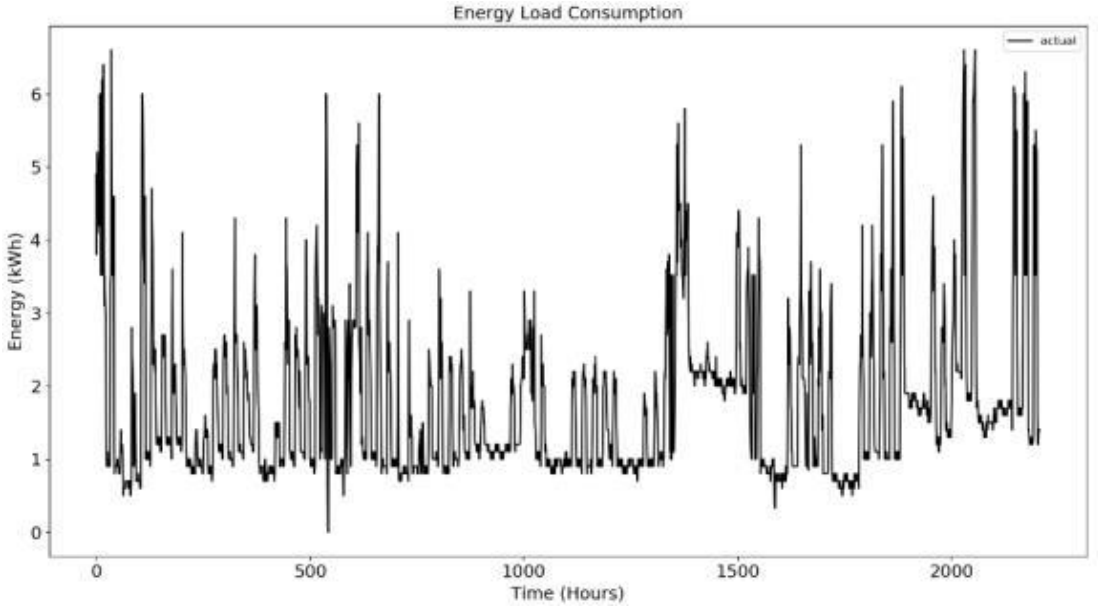

Fig. 23. Aggregated energy load data from CERTH/ITI nZEB Smart Home (spring season[14]).

A timeseries is a collection of observations of well-defined data items resulting from the gathering of the same repeated measurements over a period of time. One of the most common assumptions that take place to a variety of timeseries techniques is the stationary nature of data. In general, stationary procedures are characterized by the fact that their mean, fluctuation and autocorrelation structures stay unchanged as the time passes. They imply a flat-level looking arrangement of observations, without recognizable patterns, steady changes after some time and no intermittencies at all. When timeseries do not present the characteristics of stationarity, various methods can be utilized for that purpose. Some of them are the following.

[14] The x axis labeled as Time (Hours) depicts the seasonal timeseries in incremental hourly observations (0 to 2208), while the y axis labeled as Energy (kWh) depicts the actual energy load in kWh.

• The data can be differentiated. That is, given the series zi , the new series $yi = zi - zi{-}1$ is created.

• If the data contain a trend, some type of curve can be fitted to the data and then model the residuals from that fit. Since the purpose of the fit is to simply remove a long-term trend, a simple fit, such as a straight line, is typically used.

• For non-constant variances, by utilizing the logarithm or square root of the timeseries usually stabilizes the resulting variance.

Generally, the timeseries can be separated in three coordinates as follows (7):

$$x_t = m_t + s_t + y_t \ (7)$$

Where mt and st indicates the trend and seasonality respectively and yt the timeseries, which includes all the useful information. The energy data required to feed the proposed method, are historical data associated with energy load values requested from the nZEB Smart Home. A real-time data feed is also possible but will not be incorporated for the purposes of this experimentation. nZEB integrates a variety of APIs that allow various energy measures to be retrieved. This is achieved through a swagger API description which is private allowing for POSTs and GETs for hundreds of IoT measurements from devices within the smart building.

Table 33. Task 5 experimentation dataset description.

Attribute	Type of value
A	Datetime index in hourly timestamps
B	one-hour progressive number starting from 1 to 2208
C	Aggregated energy load: active energy measured by the smart meter in an hourly resolution.

Therefore, the dataset for experimenting on consists of hourly energy consumption values as they were measured by IoT smartgrid devices during the months of a spring season. Testing is conducted for spring months March (31 days), April (30 days) and May (31 days) 2018 summing up to 92 days of energy load measurements (Table 33). Testing is conducted in the last 20 days for evaluation of the forecasting results, while the training set is the first 72 days. The ambiguous standard rule for training and testing sets is 70-30 or 80-20 proportion; for the purposes of this study, 21.74% of the available measurements were used for testing and 78.26% of the available measurements were used for training. This was done to sustain the

chronological sequence of the timeseries data, including readings from entire days, for both training and testing sets.

7.2.3 Task 6

This section presents the dataset for development, operation and testing the proposed approach. It is a timeseries dataset and contains rows of attributes representing entries for consumers that belong to a portfolio of an aggregator. Various data preprocessing techniques were utilized, such as handling missing values, or data transformation and reduction, as needed to normalize the dataset. Experimentation data refer to 348 consumers with both energy prediction and flexibility readings over one-hour time intervals retrieved from ASM Terni pilot[15] in Italy. ASM Terni S.p.A. is an Italian multi-utility operating in the centre of Italy, notably it is the Distribution System Operator (DSO) of the city of Terni.

The local power distribution network covers a surface of 211 km2 and delivers about 400 GWh to 65,500 customers annually. The ASM distribution network is connected to the High Voltage (HV) grid through three primary substations and supplies electricity to residential and business customers through 60 Medium Voltage (MV) lines (10 kV to 20 kV) and about 700 secondary substations. The peak power is about 70 MW and the total length of the power lines in the grid is about 2,400 km (600km MV lines and 1,800 km LV lines). Currently the energy customers are about 65,500, 98% of which have an electronic meter. The grid is characterized by a large number of distributed renewable energy sources embedded in the MV and LV distribution networks reaching the total installed capacity of around 70 MW. In this regard, it is worth pointing out that, based on this energy mix, 200 GWh of the 400 GWh absorbed yearly, are produced by DER systems connected to the MV/LV grid of ASM, 70 GWh of which are from intermittent RES.

In 2019 the energy consumption reached 347 GWh, while the distributed production units connected to the MV / LV network (DER) generate 178 GWh (i.e. approximately 49% of the total demand). Therefore, in 2019 about 50% of the total consumption was covered by RES. In fact, in 2019 the local power network received renewable energy from: i)1325 PV plants, ii) one waste-to-energy, ii) eight hydro

[15] https://www.wisegrid.eu/pilot-sites/terni

power plants. In 2019 the total electric power generated (energy mix variation) from RES was as follows: i) 34 GWh from Solar Energy; ii) 68 GWh from Hydropower; iii) eight GWh from Biomass and waste – to – energy.

For this thesis the energy consumption and production of a cluster of 348 consumers have been used for the evaluation purposes. This cluster consists of residential, commercial and industrial end users, characterized by high level of auto-consumption rate. Although almost all the electricity users of the ASM's power distribution network have smart meters installed in their premises, for many of them, monthly values are collected. On the other hand, the data of the cluster are collected every 15 minutes and aggregated in one-hour resolution for experimentation; these data are gathered through the Advanced Metering Infrastructure (AMI) which consists of a Smart Meter, Current Transformers and GPRS modem that enables data transfer to central servers. After a consistency check data are stored in ASM Terni servers for 5years.

For further enhancing the dataset, yet another consumer is added to the portfolio of consumers, that is then ovel CERTH/ITI nZEB Smart Home[16] which is located in Thessaloniki, Greece. It is a rapid prototyping infrastructure incorporating various novel technologies. This structure imitates real domestic conditions experimenting on actual habitat conditions. Since it integrates a variety of Internet of Things (IoT) technologies and Information and Communication Technology (ICT) solutions, it stands as an ideal consumer pilot for this study, presenting its side and its interaction with the aggregator. Data observations expand from 2019-02-01 up to 2019-02-27. Whenever timeseries are utilized, the timestamp is in coordinated universal time (UTC). System Marginal Price (SMPr) is retrieved from entsoe transparency platform[17] for both pilot areas, Italy (348 consumers) and Greece (one consumer). Other important parameters that have been taken into consideration are temperature and operating reserves. The energy load prediction is based on an ensemble of a set of weak learners such as Multilayer Perceptron (MLP), Long Short-Term Memory (LTSM), Gradient Boosting trees (GBT) and Support Vector Regression (SVR) (Petrican *et al.*, 2018; Ves *et al.*, 2019). Their energy forecasting

[16] https://s3platform.jrc.ec.europa.eu/en/digital-innovation-hubs-tool
[17] https://transparency.entsoe.eu/

results are combined using a weighted average with the weights being dynamically computed as a function of the input features of the prediction process according to (8):

$$\omega_{MLP} * E_{MLP} + \omega_{LSTM} * E_{LSTM} + \omega_{GBR} * E_{GBR} + \omega_{SVR} * E_{SVR} \quad (8)$$

The weights associated with the prediction results of each of the four weak learners (MLP, LSTM, GBT and SVR) are computed using a genetic algorithm. The solution is composed of the four weights and the fitness function is based on the prediction error obtained by applying the weights on the predictions according to the test data. Such an approach is presented in (Vesa *et al.*, 2020), allowing for a dynamic weight computation combined with a weak learner configuration.

The energy demand prediction considers both energy and contextual features. The energy features are determined from the historical energy data acquired by consumers on-site smart meters. The contextual features represent data that are not specific to power but is correlated to context, such as season, day of the week, day of the month, etc. The energy flexibility prediction model uses two MLP neural networks to predict the flexibility lower bound (i.e., below the baseline demand) and upper bound (i.e. above the baseline demand) (Vesa *et al.*, 2020). The baseline energy demand shows the electricity would have been consumed by a consumer in the absence of DR and to determine it the X of Y method was used. Flexibility prediction features reflect the differences between the monitored energy profiles and the baseline, either above or below. The neurons used are of type ReLU (Nair and Hinton, 2010), the metric for training the network was MSE (Botchkarev, 2018) and the optimizer used for determining the weights was ADAM (Kingma and Ba, 2015). The energy load and flexibility predictions have a good accuracy featuring a MAPE < 10% (Botchkarev, 2018).

Table 34. Task 6 dataset description for two types of optimization.

Optimization type	Description
Consumer	Energy load forecast (kWh); System marginal price (€/kWh).
Aggregator	Portfolio load forecast (kWh); Portfolio upper bound flexibility of each consumer (kWh); Portfolio lower bound flexibility of each consumer (kWh); System marginal price (€/kWh).

Table 34 offers a summary and description of the data used in experimentation. For the bi-objective optimization timeseries data per consumer are utilized. These are energy load forecast and the SMPr. For the single-objective optimization, timeseries data for portfolio load forecast were utilized, upper and lower bounds of flexibility and the SMPr.

7.3 METHODS & ANALYSIS

This section discusses how the data were processed and analysed. It offers enough detail for the readers to replicate the analysis. Also, it justifies the reasoning for choosing specific algorithms/methods. Table 35 outlines all methods and algorithms utilized in this study and categorizes them to (Mining) Tasks.

Table 35. PART II Outline of Mining Tasks and Methods.

Research Task	(Mining) Task	Method
4	Classification, Clustering, Searching and Balancing	Custom rule-based classification, k-means, QCUT and CUT, EBA (custom brute force search and balancing algorithm)
5	Forecasting	Ensemble Methods (EAP, EWA, EPE), Prediction Algorithms (MLP, LSTM, GBT, SVR, LR), Evaluation Measures (RMSE, MAPE, SMAPE, ET)
6	Multi-objective Optimization	Interior Point Optimizer (Ipopt) for large-scale non-linear optimization and GNU Linear Programming Kit package (glpk) as a Mixed Integer Programming (MIP) solver

The following sections offer a presentation of methods per (Mining) Task with brief descriptions of key algorithms, as well as details about the methods employed for the experiments. There are numerous data mining functions to choose from, implemented by a variety of algorithms (Kanellopoulos *et al.*, 2011; Yakhchi *et al.*, 2017). For Research Tasks 4-6 (EB, ELF and EODS) a Python implementation was employed utilizing the necessary packages/libraries.

7.3.1 Clustering

Clustering is an unsupervised learning method, which creates groups from datasets that consist of objects or entities that are characterized by similar or identical attribute values but are adequately different from entities that belong to other clusters (Kanellopoulos *et al.*, 2011). For running a clustering algorithm, the distance measure

(e.g., Euclidean, Manhattan, Jaccard, Cosine distances) (Choi, Cha and Tappert, 2009) needs to be specified. After that, clustering methods often continue with the process of object selection and a method for evaluating the results (Jain, 2010). For evaluation quality measures can be used like cohesiveness (measure for object-to-object distance), separateness (measure for cluster-to-cluster distance) and silhouette index (mix of cohesiveness and separateness) (Zafarani, Abbasi and Liu, 2014).

k-medoids & k-means

k-medoids is a clustering algorithm related to k-means (Arthur and Vassilvitskii, 2007) and the medoidshift algorithm (Kaufman and Rousseeuw, 2002). k-means was used in Research Task 4 (EB). Both k-means and k-medoids partition the dataset, and attempt to minimize the distance between points labelled to belong to a cluster and a point designated as the epicentre of the cluster. Running k-medoids in RapidMiner the following default parameter values were used: max runs10, max optimization step 100. Other values were also set, but they produced the same or poorer results. Regarding the measure type, Mixed Euclidean Distance was used, for the reasons explained in the DBSCAN section (Sect. 3.3.1).

CUT and QCUT (Binning)

These methods were used in Research Task 4 (EB). CUT and QCUT (McKinney, 2010) bin prosumers a process that may be considered a form of clustering.

QCUT is based on the enforcement of the quantile-based discretization function. It is used on the raw data to discretize the consumptions/productions into equal-sized bins (buckets), based on ranks or based on the dataset entries' quantiles.

CUT is based on the enforcement of bin values into discrete intervals. This method is used on the raw data attempting to segment and sort the consumptions/productions into bins. It performs better when the aim is to transpose from continuous variables to categorical. For example, 100 production/ consumption values for five bins produce groups of production/ consumption value ranges.

7.3.2 Classification

This mining task relates with Research Task 4 (EB). A custom rule-based classification was conceived to formulate the proposed approach of VMG generation, enforcing rules that create classes of prosumers. That way, an initial categorization of

the prosumers within the dataset is enabled. Therefore, three district classes of prosumers are generated, each class conforming to Rules#1–3. Each rule, when satisfied assigns a prosumer to a specific class. These classes of prosumers are:

i) The ones whose energy generation is less than their consumption, i.e. they draw energy from the grid (class1).

ii) The ones who balance their generation and consumption levels, i.e. a minimal interaction with the grid (class2).

iii) The ones whose energy generation is greater than their consumption, i.e. they inject energy to the grid (class3).

Therefore, three classes are distinguished associating with any entries indexed by timestamp and other related values within the dataset. In that way, the retrieval of the aggregated values of generation/production for each class is straightforward. Fig. 24 shows the energy data from all prosumers at a specific timestamp, before enforcing the rules. Fig. 25, Fig. 26 and Fig. 27 depict the formation of each class on a specific timestamp.

Applying Rule#1, a class of prosumers (class1) is generated, those who draw power from the grid. More precisely, all prosumers with E0a < -1 (Watt) enroll to this class for the specific timestamp, as shown in Fig. 24.

Applying Rule#2, a class of prosumers (class2) is created, those who virtually neither draw from nor inject power to the grid.

$$-1 \leq E0_a \leq +1\ (Watt)\ (9)$$

To be precise, when (9) stands, all prosumers enroll to this class for the specific timestamp (Fig. 25). These value boundaries for Rule#2 are set because injection to or draws from the grid between -1 and $+1$ Watt can be considered negligible.

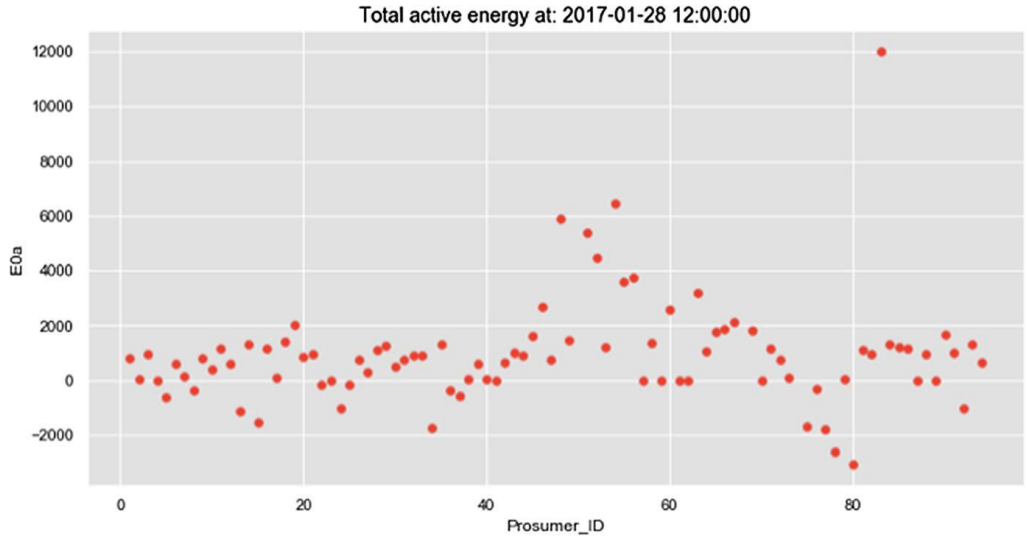

Fig. 24. Total Active Energy (TAE) per prosumer on a specific timestamp.

Applying Rule#3, a class of prosumers (class3) is created, those who inject power into the grid. To be precise, all prosumers with E0a > + 1 (Watt) are enrolled to this class, for that specific timestamp, shown in Fig. 27. At this point, various measurements can be utilized regarding the information contained in each class at a specified timestamp, such as the exact sum/max/mean/std (standard deviation) draw from/ injection to the grid, the number of prosumers etc. Next, this information is used for the process of VMG formation.

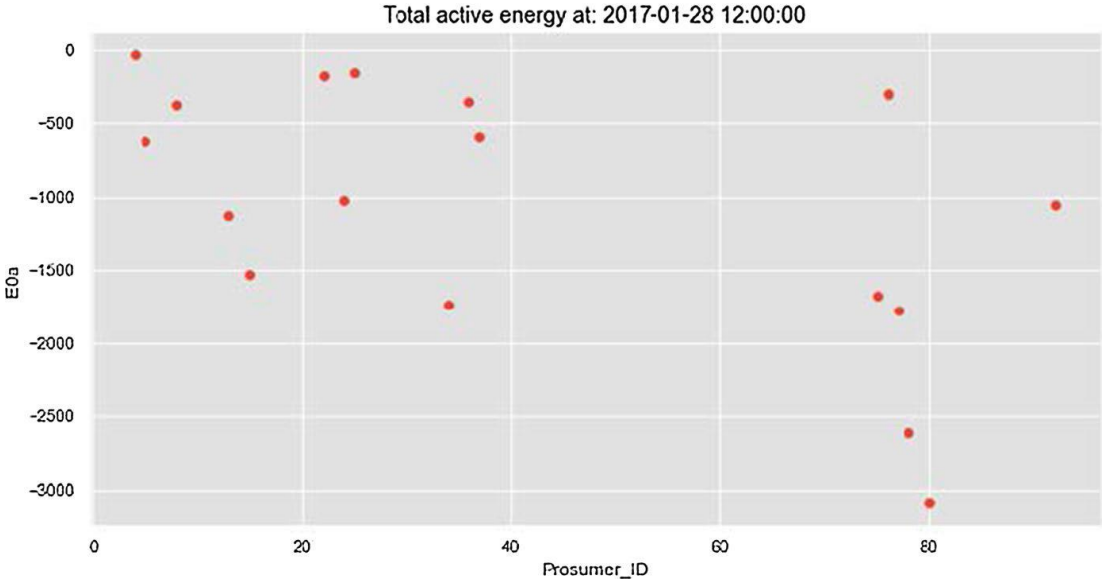

Fig. 25. TAE per prosumer abiding with Rule#1.

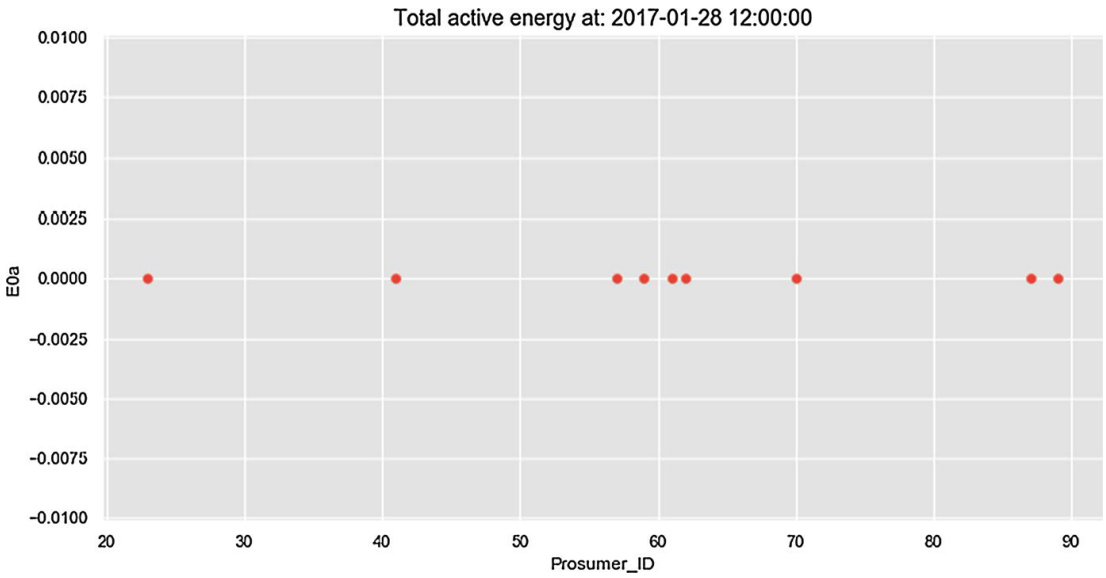

Fig. 26. TAE per prosumer abiding with Rule#2.

For example, 100 production/consumption values for five quantiles produce a categorical object indicating quantile membership for each of the consumption values.

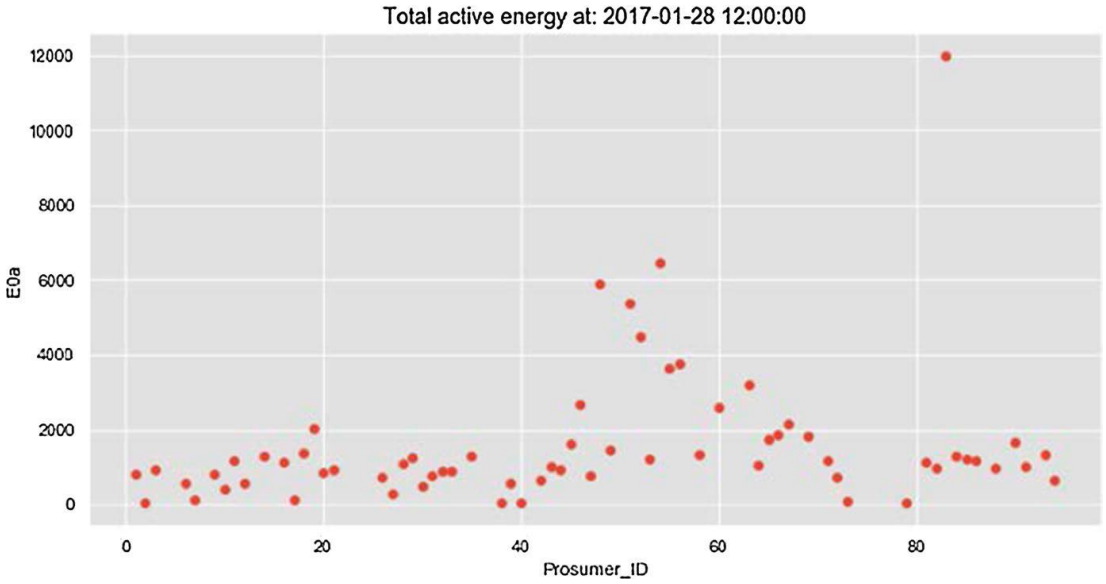

Fig. 27. TAE per prosumer abiding with Rule#3.

In order to achieve P2P balancing towards the envisioned VMG2VMG balancing, this study utilizes a hybrid approach performing clustering through rule enforcement, seeking to attain and verify achievements. A detailed presentation of this mixture of techniques is presented in Sect. 8.1.

7.3.3 Searching and Balancing

This mining task relates with Research Task 4 (EB). The EBA algorithm balances individual, or aggregated consumption/generation energy values resulting from prosumer clusters or bins. Also, Within-Cluster Sum of Squares (WCSS) (Pedregosa *et al.*, 2015) and the silhouette score (Pedregosa *et al.*, 2015) determine the best value k for clusters, as an input to the k-means algorithm. There are many methods for defining the number of clusters. For example, the silhouette score calculates the average similarity of objects in a cluster and the distance from objects in other clusters. As a method for defining number of clusters, it involves more computations, but it is more informative (Rousseeuw, 1987). The elbow method iterates on an algorithm that attempts to measure a point where a clustering score starts to decrease the most. In practice when a sharp elbow appears it may cause misinterpretations (Ketchen and Shook, 1996). For the purposes of this study, although initial experimentation involved elbow method and WCSS, the silhouette score was finally selected since it is more thorough, and less prone to wrong inspection of results.

7.3.4 Forecasting

This section describes methods/algorithms for timeseries forecasting, utilized for the Research Task 5 (ELF). These include, prediction algorithms, ensemble methods as well as evaluation measures. They are fine-tuned and incorporated into the methodology for evaluating the proposed single step forecasting approach, presented in Sect. 7.1.2.

Prediction Algorithms

This section presents the algorithmic options that are utilized for performing the actual forecasting modelling described in this study. An outline of the usage of each algorithm is presented aiming to generate a comprehensive demonstration of their functionalities.

Neural Networks (NN) utilizing Multi-Layer Perceptron (MLP)

This is a NN supervised learning algorithm that is able to learn a function by training on a supplied dataset, according to a number of dimensions for input and a number of dimensions for output. By feeding a set of features and a target value, it can learn function approximators for regression or classification tasks. In general, MLPs are networks with specific architectures tailored to the requirements for various

forecasting tasks (Murtagh, 1991). They usually exhibit many layers that contain interconnected neurons that each layer's output is fed to another layer resulting in a prediction layer that yields the final prediction values. These networks predict with just the input factors, without considering any time variations, usually resulting in regression predictions that are independent with one another.

For regression purposes, the MLPs essentially convey the data to a set of non-linear functions mapping the labels to sets of numbers. A generic MLP architecture with one hidden layer is demonstrated in (Pedregosa *et al.*, 2015). Moreover, MLPs train utilizing backpropagation of two stages comprising many executions. First, the outputs are validated against the coefficients of the actual values and the expected values are assessed utilizing a loss function. Next, the backward pass commences attempting to adjust the neurons' coefficients for outputting better predictions. This is done by finding the correlation at which the coefficients affect the value of the utilized loss function. That way, a learning rate process is incorporated that updates in an iterative way (utilizing the concept of epochs) each coefficient value.

Neural Networks (NN) utilizing Long Short-Term Memory (LSTM)

This type of NN is specifically conceived for handling sequence dependencies and is characterized as an artificial Recurrent NN (RNN) architecture (Hochreiter and Schmidhuber, 1997). Unlike MLP, this type of NN exhibits feedback connections expanding its processing capabilities from single data points to complex data sequences. LSTMs promise to deal with the problem of vanishing gradient (Kolen and Kremer, 2010) by incorporating the backpropagation of non-altered gradients within the network. The LSTMs define four distinct architectural components. A cell, an input gate, an output gate and a forget gate each one of them designed for a specific reason (Van Houdt, Mosquera and Nápoles, 2020). The cell acts as the memory of the network as it handles the data in different time intervals, while the rest of the components are responsible for the management and flow of information and processes inside the LSTM network.

Therefore, the input gate decides which and if new values enter the LSTM unit; the output gate which of these values will be used to feed the final output of the LSTM unit; and finally the forget gate decides on matters about the values that will stay within the unit. An overview of various LSTM types can be found in (Greff *et al.*, 2017).

LSTMs train utilizing supervised learning and sets of training sequences. To that end, optimization algorithms are used such as gradient descent along with back-propagation for deciding on the values of gradients recalibrating the LSTM weights. LSTMs address the abovementioned problem of vanishing gradient in two ways. First, while the values of errors are backpropagated returning to the input, their values remain unaltered and as a result their values do not diminish in an exponential way. Second, each LSTM unit essentially decides which values to keep and disregard, utilizing procedures involving sigmoids that introduce an extra technique for enhancing diversity. That mitigates the issue of non-altered gradients within the network.

Gradient Boosting Trees (GBT) utilizing XGBOOST

Gradient boosting utilizes a framework for implementing the following points. A framework allows great mobility for performing an analysis either it is a regression or classification, since a variety of methods can be used for any subsequent process step. First, there is a loss function that is to be optimized. For that purpose, a weak learner is utilized to perform the prediction, then an additive model that adds the weak performing learners, attempting to minimize the loss function. For the loss function, in regression specifically, squared error is used. For the weak learner, regression trees are utilized generating subsequent models thought splitting for prediction outputs. For example, AdaBoost can be used for short tree splits to ensure that the learner will remain weak. Finally, for the additive model, the generated trees are added one by one to the existing tree model. A gradient descent process is also utilized to minimize any losses during the tree addition procedure. The GBT being a type of gradient boosting models, attempts to reduce the output of an object function by relating a convex loss function with a term that applies penalties in model complexity. This happens through an iterative process that adds new trees to an already existing tree structure while predicting the residual errors of the previous trees combining them with the new one for outputting a final forecast value (Chen and Guestrin, 2016). A rather abstract version of a formula for forecasting with GBT is the following (10). It attempts to minimize an objective function for a new tree T_{min_t}.

$$\sum_{i=1}^{n}\left[gsf_i Tmin_t(x_i) + \frac{1}{2} gss_i {Tmin_t}^2(x_i)\right] \quad (10)$$

where, gsf_i and gss_i describing the first and second gradient statistics for the loss functions respectively and x_i a variable that attempts to apply weights for enforcing an over-fitting protection to the minimization process.

Support Vector Regression (SVR)

The SVR is a generalization of the support vector machines aiming to solve classification problems with regression analysis. For its implementation SVR uses kernels, sparse solution and Vapnik-Chervonenkis theory for controlling the margin and the number of support vectors. SVR over the years has proven to be a very good estimator of real-value functions. SVR in its supervised learning implementation trains utilizing a symmetrical loss function that applies penalties to high and low wrong estimates in an equal way. Essentially, an estimated function becomes surrounded in a symmetric way by a flexible "tube" of minimal values, with all absolute error values above and beyond this tube being ignored. The greatest advantage of SVR resides to the fact that its computation complexity is not dependent on the diversity of the input space related with its dimensionality (Awad and Khanna, 2015). An abstracted and rather simplistic equation that describes a generic SVR forecasting formula follows (11):

$$forecast_i = (p, x_i) + b \quad (11)$$

where x_i an input vector to the SVR model, p the parameter vector for the model, while b a bias. In addition, various SVR implementations allow the introduction of variables that enforce thresholds that control the flow of forecasting errors in and out of the accepted values of the abovementioned "tubing" concept.

Linear Regression (LR)

This model is utilized for adding a more basic implementation of a simple tool for running the experimentation and generate enough comparable outputs in timeseries forecasting. As a very basic model LR is chosen since it is widely used in various regression domains and exerts in producing very fast and relatively reliable results (Wetherill and Seber, 1977). Therefore, it is considered an appropriate model for unveiling a baseline estimator for any forecast attempt. Usually, when other forecasting models fall behind the accuracy output of LR, are essentially marked as poor performing. Its functionality can be described by a mathematical formula that takes into consideration and associates the value of forecast of a dependent variable in

a way that equals to the actual or current value of the variable. A rather simplistic equation that describes the abovementioned relationship is the following (12):

$$forecast_{time+horizon} = actual_{time} \quad (12)$$

where $horizon$ the requested forecast horizon, $time$ the timesteps and $actual$ the observed values.

Ensemble Methods

This section is dedicated to presenting forecasting ensemble method options while focusing on the identified ones that are used for performing the energy load forecast experimentations of this work. In addition, a step-by-step process for each ensemble method is presented.

In a more generic context, any ensemble method is a meta-algorithm that can combine multiple statistical and ML techniques producing one predictive model. The ensemble methods incorporated for the experimentation are the following:

Ensemble utilizing Averaged Predictions (EAP)

The basic idea of an ensemble is to combine predictions from several models averaging out errors for outputting better overall predictions. In order to keep track of the progress, it is helpful to formalize the ensemble as n models $forecast_i$ averaged into an ensemble (13):

$$ensemble(t) = \frac{1}{n}\sum_{i=1}^{n} forecast_i(t) \quad (13)$$

The detailed steps of the initial process followed for predicting for n steps ahead the energy load are outlined below:

i) Read data.

ii) Utilize a function that transforms the sequential timeseries data into suitable format [e.g. feature vectors, labels].

iii) For each different prediction point p, train the model with any supervised learning algorithm/method (MLP, LSTM, XGBOOST, SVR and LR) and get the coefficients of each model. Predict for each different model using the test set. Begin from the same $forecast_i$. For example, if the range of forecast is (1, n) all

the predictions should start from the nth row, because the prediction should be t_n. Use the t_n coefficients that correspond to the specific model.

iv) The final step is to calculate the average of all predicted values, starting from the first position until the length of $forecast_i$.

The starting point depends on the length of $forecast_i$ and the horizon. For example, if the range of $forecast_i$is (1, n) and the horizon is n.

Ensemble utilizing Weighted Averages (EWA)

By attempting to combine predictions, a weighted average is set in a way that for each model fi, there is a weight parameter $weighti \in (0,1)$ that assigns a weight value to that model's forecast. Weighted averaging requires the summation of all weights to be equal to 1.

The ensemble function is defined as (14):

$$ensemble(t) = \frac{1}{n}\sum_{i=1}^{n} weight_i * forecast_i(t) \quad (14)$$

This is a minor change from the previous definition, since once the models have generated predictions $f_i = forecast_i(t)$, learning the weights is the same as fitting any supervised learning algorithm/method (MLP, LSTM, XGBOOST, SVR and LR) on those predictions (15):

$$ensemble(f_i, \ldots, f_i) = weight_1 f_1 + \ldots + weight_n f_n \quad (15)$$

With some constraints on the weights. The ensemble would then take local averages based on the nearest neighbours of a given observation, empowering the ensemble to adapt to changes in model performance as the input varies.

The detailed steps of the modified process followed for predicting the n steps ahead energy load are outlined below:

i) Read data.

ii) Generate a function that transforms the sequential timeseries data into the format [e.g. feature vectors, labels].

iii) For each different prediction point p train the model with any supervised learning algorithm/method (MLP, LSTM, XGBOOST, SVR and LR) and get the coefficients of each model. Predict for each different model using the test set.

Begin from the same prediction point (For example if the range of $forecast_i$ is (1, n) all the forecasts should start from n row, because the prediction should be t_n. Use the right t_n coefficients that correspond to the right model. The main difference with the averaging method is that after the prediction, weights are applied depending on how old the data are. If the predictions are based on lots of past data, the weight which is applied has a smaller value the further to the past they belong.

iv) The final step is to calculate the average of all predicted values, starting from the first position until the length of $forecast_i$.

Ensemble utilizing Polynomial Exhibitor (EPE)

This ensemble method's implementation attempts to further improve the forecasting capabilities by combining predictions with polynomial regression. The process performed is similar to the previous one with a weighted parameter being assigned to the prediction model. A minor modification to the process takes place. Once the data on which the predictions are based get older, their polynomial exhibitor grows and at the same time a smaller weight is assigned to each one of them. The slightly modified process followed for predicting the p steps ahead energy load are similar with the weighted averaged method. The main and unique difference is the data transformation of the features. For example, if the prediction of the model is based on n past values, the features will be adjusted according to the following formula representing an input vector (16):

$$x_n^{n-1}, x_{n-1}^{n-2}, \ldots, x_1^2, x_0^1 \ (16)$$

Evaluation Measures

The evaluation of results is demonstrated by a detailed representation of the outputs of the utilized techniques. These are tailored to the proposed timeseries approach and ensemble models incorporating multiple forecasting algorithms for modelling. For evaluating the results of the proposed methodology three metrics are utilized namely RMSE, MAPE and SMAPE which are widely used in ML tasks such as regression, forecasting and prognostics (Botchkarev, 2018). Next, the generic mathematical formulation of these evaluation metrics takes place.

Root Mean Squared Error (RMSE)

This metric calculates the square root of the average squared difference of actual value and prediction value. According to its mathematical definition, RMSE applies more weight on larger errors. Furthermore, it is best suited for performing comparisons for diverse data types since, it is greatly affected and reliant on the scaling of the values under comparison. The RMSE generic mathematical formula is stated as (17):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n}} \quad (17)$$

where $x_i$the actual energy load value, y_i the energy load prediction value and n the total number of prediction values.

Mean Absolute Percentage Error (MAPE)

This metric involves the calculation of absolute value of the difference of actual and predicted values. According to its mathematical definition (18), MAPE applies more weight (penalize) on positive errors in relation with the negative ones. Also, an interesting point is that it is a metric that is nondependent on the scale of the data utilized.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i - y_i}{y_i}\right| * 100 \quad (18)$$

where x_i the actual energy load value, y_i the energy load prediction value and n the total number of prediction values.

Symmetric Mean Absolute Percentage Error (SMAPE)

This metric is calculated involving the actual values subtracted from predicted values expressed as division with the sum of both of them. One of SMAPE advantages is that it makes it easier to analyse the output in a statistical way. According to its mathematical definition (19), SMAPE is nondependent on the scale of the data since it enforces symmetry and properties that showcase diminishing high levels of biases.

$$SMAPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|y_i - x_i|}{(|x_i| + |y_i|)/2} \quad (19)$$

where $x_i$the actual energy load value, y_i the energy load prediction value and n the total number of prediction values.

As a generic rule of thumb when the SMAPE value fluctuates below the 10% threshold, the forecasting performance is to its optimal levels. Values between 10% and 20% thresholds indicate forecasting performance that is considered at acceptable limits. Finally, values above 20% indicate forecasting that should be forfeited since it indicates high levels of inaccuracy.

The abovementioned metrics were chosen based on their high popularity for evaluating forecasting tasks (Botchkarev, 2018). According to this study's approach these metrics are utilized for evaluating timeseries forecasting with various ensemble and prediction algorithms on the domain of ELF. The goal is to evaluate their accuracy by using the same measures form each case.

The RMSE metric output is extracted in the measurement units of the forecasting variable which is kilowatts per hour (kWh). The output of MAPE and SMAPE metrics is extracted in percentage scaling, enabling more comprehensive and easier to interpret values. In addition to the performance metrics utilized for evaluating the accuracy of the forecasts, this study also reports on the Execution Time (ET) for running the forecasting experiments. This aims at enhancing the reporting process, generating an additional comparison feature for the final report and discussion on overall performance.

Experimental parameters

This section reports on algorithmic parameters for running the experiments. It focuses on the parameters used for prediction algorithms. Grid Search[18] (Pedregosa *et al.*, 2015) was performed for MLP, LSTM, XGBOOST and SVR resulting in the following parameter and hyperparameter tuning.

Table 36. Generic parameters.

Parameter	Value	Description
testing_days	20	It defines the number of days for the testset.
horizon	1	The time horizon for performing predictions.
backwards_steps	24	The steps to check backwards.

[18]This method performs exhaustive search over specified parameter values for an estimator. The parameters are optimized by cross-validated grid-search over a parameter grid.

There are generic parameters (Table 36) that apply for all approaches and other specific parameters per utilized method/algorithm.

Table 37. Hyperparameters for MLP (Chollet, 2015).

Parameter	Value	Description
activation_function	'relu', 'linear'	It defines the standard neuron activation function per layer.
model_optimizer	'rmsprop'	Specifies the gradient descent optimizer.
epochs	10	Defines the number of epochs for the NN or else how many passes over the data entries.
loss_function	'mse'	This parameter calculates the prediction error of a Neural Net.
batch_size	1	Defines a hyperparameter that sets the number of samples to work; before updating the rest of the internal model parameters. The value 10 which is smaller than the size of the training set indicates a mini-batch gradient descent.
units	[256, 64, 1]	Defines a positive integer for dimensionality of the output space of each NN layer.

Hyperparameters of LSTM (Table 38) implementation are identical with the MLP (Table 37) implementation but with a different value in batch_size value.

Table 38. Hyperparameters for LSTM (Chollet, 2015).

Parameter	Value	Description
activation_function	'relu', 'linear'	It defines the standard neuron activation function per layer.
model_optimizer	'rmsprop'	Specifies the gradient descent optimizer.
epochs	10	Defines the number of epochs for the NN or else how many passes over the data entries.
loss_function	'mse'	This parameter calculates the prediction error of a Neural Net.
batch_size	10	Defines a hyperparameter that sets the number of samples to work; before updating the rest of the internal model parameters. The value 10 which is smaller than the size of the training set indicates a mini-batch gradient descent.
units	[256, 64, 1]	Defines a positive integer for dimensionality of the output space of each NN layer.

Table 37 and Table 38 present the significant hyperparameters for NNs implementation, MLP and LTSM. In both cases a sequential model is defined, therefore a sequence of layers. For both implementations one input layer is utilized for the historical data entries, two hidden layers with 256 memory units and 64 memory

units respectively with the ‘relu’ activation function and one output layer with linear activation function.

Table 39. Parameters for XGBOOST (Chen and Guestrin, 2016).

Parameter	Value	Description
n_estimators	[50, 60, 80, 100, 150]	It defines the number of trees in an XGBoost model within the XGBRegressor class.
max_depth	[3, 4, 5, 6, 7, 8]	It sets the depth for searching the GBT trees.

No specific parameters required for the LR model despite the generic parameters as stated in Table 36.

Table 40. Parameters for SVR (Pedregosa *et al.*, 2015).

Parameter	Value	Description
kernel	‘rbf’	It defines the type of the kernel to be used in the algorithm.
gamma	‘auto’	It defines the kernel coefficient for kernel parameter.

Finally, Table 39 and Table 40 presents the parameter setting for executing the XGBOOST and SVR algorithms respectively.

7.3.5 Multi-objective Optimization

This mining task was addressed for Research Task 6 addressing EODS. Therefore, methods/algorithms were utilized depending on the Scenario considered, achieving and verifying results. For Scenario #1 Interior Point Optimizer (Ipopt[19]) was used an open-source software package for large-scale non-linear optimization. For Scenario #2 the GNU Linear Programming Kit package (glpk[20]) was used as a Mixed Integer Programming (MIP) solver. These scenarios deal with consumers and aggregators, respectively. The usage of metaheuristics and evolutionary approaches were examined for the problem at stake, yet it was decided that using mixed integer programming and large-scale nonlinear programming solvers is more appropriate, since they produce optimal solutions in a quick and precise manner. Also with this

[19] https://coin-or.github.io/Ipopt/
[20] https://www.gnu.org/software/glpk/

approach, parameter fine-tuning like population size, number of function evaluations etc., commonly required in metaheuristic approaches is omitted.

Single-objective optimization

The optimal day-ahead schedule for the aggregator utilizes lower and upper bounds of flexibility for each consumer within the portfolio and the energy load forecast for a specific day and in one-hour resolution. Since, the aim is to minimize the aggregator's net cost for trading energy in the day-ahead market, the SMPr per market (Italian and Greek) is considered, exposing capabilities for optimizing overall portfolio economic benefits and the portfolio's day-ahead energy load forecast based on historical data.

Therefore, this study incorporates a single-objective modelling approach for minimizing the operational costs for the aggregators. Due to the abstract level of input data, a sub-process of the proposed framework is exploited for testing and validation of a single objective function for this use case. It yields the "on demand" algorithmic results output and the optimal contribution of the consumer to the grid (for a specific timestamp) while abiding with design constraints.

Objective function (20) manages the aggregator's day-ahead optimal energy resource scheduling, aiming to reduce its overall operational costs and maximize profits (Koukaras, Gkaidatzis, *et al.*, 2021).

$$fa_{cost}(t) = min \sum_{i=1}^{N_{tot\_C}} \sum_{t}^{t_{total}} E_{i,imp}(t) * SMP(t) * \Delta t \ (20)$$

Subject to constraints (21), (22) and (23):

$$E_{i,imp}(t) \geq flex_{i,lb}(t) \ (21)$$

$$E_{i,imp}(t) \leq flex_{i,ub}(t) \ (22)$$

$$\sum_{i=1}^{N_{tot\_C}} \sum_{t}^{t_{total}} E_{i,imp}(t) * \Delta t = \sum_{i=1}^{N_{tot\_C}} \sum_{t}^{t_{total}} E_{i,for}(t) * \Delta t \ (23)$$

With constraints (21) and (22), referring to the maximum and minimum amounts of energy that each consumer individually can reach and consequently the portfolio as a whole. Constraint (23) actually states that the sum of the optimized energy

scheduling must be equal to that of the initially predicted one. Therefore, taking into account that regardless of the optimized schedule finally proposed to the consumers, their overall energy consumption stays the same, and thus their daily consumption habits.

Bi-objective optimization

For optimizing the day-ahead energy scheduling for the consumer, the day-ahead energy load forecast in one-hour resolution is utilized. That way the overall energy load consumption can be retrieved. Using historical data from the said pilots and the knowledge of SMPr in both areas, the cost of imported energy for the consumer is calculated. In bi-objective optimization, reference is made to the consumer use case distinguishing two objectives, the operation cost minimization resulting from reducing energy consumption at occupant acceptable levels and the minimization of consumer's discomfort. These two objective functions are solved simultaneously considering, constraints and common variables. The objective function related to operation cost reduction is the following (24) (Koukaras, Gkaidatzis, *et al.*, 2021):

$$fp_{cost}(t) = min \sum_{t}^{t_{total}} E_{imp}(t) * EIT(t) * \Delta t \quad (24)$$

Subject to constraints (25) and (26):

$$E_{i,imp}(t) \geq 0, \forall\, t \in \Delta t \quad (25)$$

$$E_{i,imp}(t) \leq E_{i,max} * E_{i,for}(t), \forall\, t \in \Delta t \quad (26)$$

Effective energy management involves a certain level of load manipulation for facilities. For households consumer loads for home appliances are distinguished which are categorized in fixed, regulatable and deferrable loads. Lightning, cooking and electronic devices belong to the fixed load that should be used anytime, on request. Water heaters and Heat, Ventilation and Air-Conditioning (HVAC) systems belong to the regulatable loads, being subjects to usage delay or rescheduling. Appliances such as dishwashers and dryers, belong to the deferrable loads, since their operation can be deferred.

Consumer thermal comfort considering these parameters poses a challenging task for effective and targeted DR scheme implementation (Zhang *et al.*, 2016). Consumer comfort is subjective and poses significant difficulties for a realistic

assessment and quantification. According to ISO7730 thermal comfort standard (ISO 7730, 2005) the Predicted Mean Value (PMV) index may be utilized to calculate human perception of comfort. The PMV index ranges between [-3,3], where: 0 means neither hot, nor cold; +/-1 slightly warm (+), or slightly cold (-); +/-2 means heat (+), or cold (-); and +/-3 means very hot (+), or very cold (-). Yet, for retrieving PMV index values a certain number of extra data attributes is required, such as air temperature, air velocity, air humidity, mean radiant temperature, clothing insulation, occupant activity etc. Since this study has no access to such data, consumer (occupant) discomfort is calculated by adjusting a custom index that considers the PMV index value range and discomfort calculation, as presented in (Lu and Hong, 2019). That way, the output of discomfort is normalized across all consumers between values [0, 1], distinguishing five discomfort profiles. An example of such profiles is presented in Sect. 8.3. The objective function related to the minimization of the consumer discomfort is described as (27) (Koukaras, Gkaidatzis, *et al.*, 2021):

$$fp_{discomfort\_reduction}(t) = min \sum_{i=1}^{t_{total}} \frac{C_{pref}}{2} * E_{imp_{red}}(t)^2 + aux_{coef} * E_{imp_{red}} \quad (27)$$

Subject to constraints (25), (26), (28) and (29).

$$C_{pref} > 0 \quad (28)$$

$$aux_{coef} > 0 \quad (29)$$

The minimization of consumer discomfort function (27) models the degree of discomfort a consumer feels when reducing the energy consumption. The greater the energy consumption reduction, the greater discomfort (Yu, Lu and Hong, 2016). C_{pref} and aux_{coef} pose as customer-specific variables. A high value of C_{pref} expresses a consumer's preference to lower energy consumption also decreasing discomfort, while a greater aux_{coef} value infers to a coefficient expressing more discomfort (Yu and Hong, 2017).

7.4 LIMITATIONS, ETHICS AND THREATS TO VALIDITY

This section outlines any ethical consideration of the research as well as problems, limitations and threats to the validity of the results.

7.4.1 Energy Balancing utilizing Machine Learning techniques

The main identified biases of this interdisciplinary energy balancing approach refer to:

i) It poses as a Computer Science centric approach attempting to investigate an Energy sector problem, prescribing the requirements for a multidisciplinary approach on a problem that requires merging concepts from multiple research fields. This statement imposes that certain levels of domain details are disregarded. For example, the study does not consider constraints, such as power transmission loss or heating loss for calculating the VMG formulation. The values for PV production on the dataset represent the final product that interacts with the grid, whether injecting to, or drawing energy from it.

ii) Observing the outputs after running the VMG balancing process multiple times, it is evident that despite any balancing approach on VMGs, every attempt is still biased by the nature of RES. That is to say, a high production is expected during peak hours while the sun is up, and zero production after the sun sets. To resolve this issue, possible solutions could be either a) sufficient energy storage means or capacity for saving the excess produced energy, or b) the implementation and enforcement of a globally accepted VMG scheme that allows energy transfer from geographical areas with sunlight to areas without sunlight during the same timestamps.

7.4.2 One Step Ahead Energy Load Forecasting

The identified biases resulting from this research task can be summarized to the following points.

i) The first one is related with the proposed approach's context and its correlation with the concept of STLF. The proposed approach utilizes multiple models that act separately, nullifying any chances for existent dependency modelling between consequent prediction steps. For example, the prediction output of t+2 time step cannot be dependent on the prediction output from t+1 time step. In line with the performed experimentation, for predicting the hourly energy load for the next day, 24 separate models were created for each hour independent with one another.

ii) Regarding the metrics for evaluating the overall forecasting performance, it is preconceived to use RMSE, MAPE and SMAPE based on their high usability by

the academic and practitioner community. To that end, the accuracy of findings is validated being aware that there are numerous other metrics that can be incorporated, possibly mitigating any unidentified inconsistencies, or tampering on the final evaluation of results.

iii) This study acknowledges that there are numerous other ensemble methods, prediction algorithms and strategies for timeseries forecasting. Yet, the ones chosen, presented and implemented in this study are considered to be representative and educative options in conjunction with all around "fine performers" for approaching timeseries predictions. For example, in case of prediction models options from NNs (MLP and LSTM) are shown but also from widely used statistical models (LR) and common ML models (GBT and SVR).

7.4.3 Optimizing Day-ahead Energy Scheduling

The limitations of the proposed optimization framework can be attributed to the fact that this research does not consider constraints such as power transmission loss, heating losses, inventors, distances etc. for calculating the objectives. That is because the final product of energy load forecast, flexibility etc. was retrieved. from API pilots. Therefore, that information detail is already taken into consideration and pre-calculated rendering this research design into a high-level experimental approach. This research task aims to optimize day-ahead energy scheduling incorporating a DR scheme and two system stakeholders, the consumer and the aggregator. For a more holistic approach, elaboration on a framework that also includes DSO as the third system stakeholder should be made, experimenting on their possible interactions in different scenarios. An analytical comparative analysis regarding the options of solvers for producing optimization results for each scenario was not performed. Instead, large-scale nonlinear optimization and a mixed integer programming solver were used for outputting results for consumer and aggregator respectively. Metaheuristics involving genetic algorithms could be utilized for both scenarios while offering a thorough comparative analysis on solvers. These limitations generate directions for improvements, thus future work.

Chapter 8: Results

This chapter details results of the three novel approaches in Energy domain elaborating in novel research tasks related with Energy Balancing (EB), Energy Load Forecasting (ELF) and Energy Optimal Day-Ahead Scheduling (EODS).

8.1 ENERGY BALANCING

To test the proposed approach, tests are carried out (Simulations #1–4) on data regarding hourly timestamps for a specific day (Fig. 28). The prosumers retrieved from the utilized dataset have PVs for their generation. Therefore, for improved prosumer profile the time horizon of (08:00–17.00) for a specific day has been chosen; that is the time horizon of PV generation. Between these hours there is sunlight in the nZEB region. Appendix J depicts how the initial data are clustered by utilizing WCSS and silhouette score for defining the number of k clusters and k-means (Arthur and Vassilvitskii, 2007) algorithm for clustering. The reason for preferring the silhouette score is explained in Sect. 7.3.3. An example of k-means clustering for a specific timestamp, using the elbow method with WCSS, as well as the silhouette score is presented in this section. The same process was repeated for all timestamps when clustering was involved.

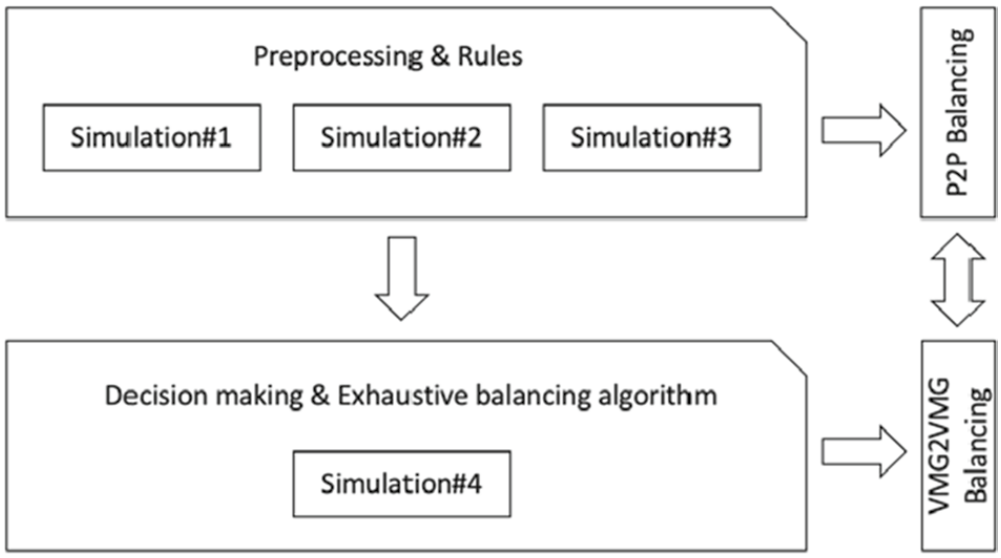

Fig. 28. Overview of results for Energy Balancing task.

As described in Sect. 7.2.1 various data preprocessing techniques were performed (removing duplicates, dealing with missing values, data transformation and reduction) to prepare the initial dataset for the simulations. Then, to get a glimpse of the initial state of the data, clustering is performed, before conducting the simulations. The reason for initiating this process is to ensure that this analysis is not impaired by any disparities. At this point it is clarified that the clustering is performed on each timestamp to create clusters of the different values of the C attribute (Table 32).

Rules detailed in Sect. 7.3.2, classify prosumers according to their energy profile. For example, some inject energy to the grid, some draw energy and some are virtually idle. It is assumed that all prosumers can generate energy through PV panels. The energy for each timestamp is the absolute value of generation minus consumption.

In Simulation#1, the rules aid P2P level balancing by linking prosumers with maximum generation with prosumers with maximum consumption through an iterative process.

In Simulation#2 and Simulation#3 this approach expands its analytical capabilities by forming VMGs for the stakeholders. Therefore, binning on the prosumer dataset is performed with a view to present the data in a quantitative, as well as a qualitative way.

The last simulation essentially presents the final stage of this approach, supporting decision making for balancing the formed clusters. Simulations 1–3 produce clusters and Simulation 4 inputs aggregated energy values (resulting from consumption or injection of energy to the grid) per cluster and utilizes EBA with user defined target values, as shown in Fig. 28.

8.1.1 Simulation#1

For the first simulation, the three rules are utilized to perform balancing for a specific timestamp, linking the prosumer with the maximum production value to the prosumer with the maximum consumption value. The methodology for this simulation comprises four steps depicted in Fig. 29 and explained below:

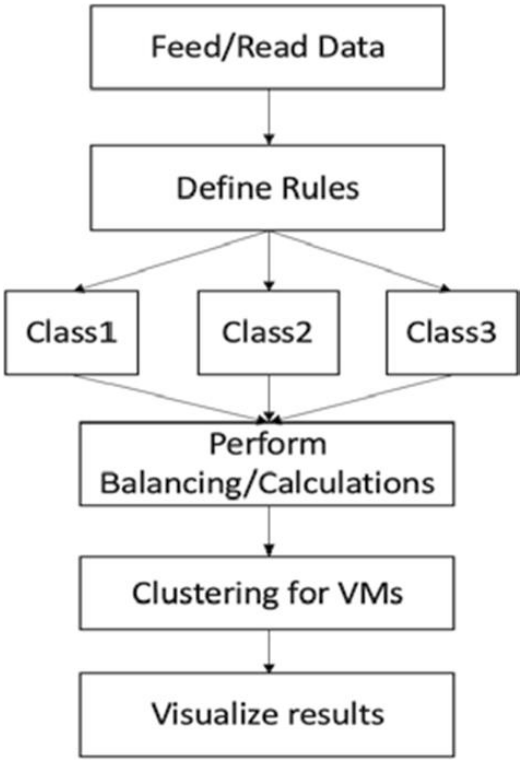

Fig. 29. EBA Simulation#1 methodology and results.

Step 1: Using the rules to transform and define the sum (injection or draw from grid) of intermittent resources for each formed class per requested timestamp (based on historical or live-feed data[21]).

Step 2: Once precise mapping of intermittent resources injection/draw of energy is finalized, perform balancing of injection/draw by connecting the prosumer with maximum injection value to the prosumer with the maximum draw value, aiming to balance the highest energy generation with the highest energy consumption. Iterate as long as there are loads (generation/consumption) to balance.

Step 3: Generate a new dataset that contains all the data resulting from Step 2. Then cluster the results to form VMGs that can be characterized by their levels of positive or negative energy injected or drawn from the grid. These VMGs can interact with the VMG layer to exchange energy with the rest of the VMG layer structure.

Step 4 Visualize and elaborate on the results. At this point plots are generated to offer an analytical explanation of the experimentation performed. Balancing is

[21] For simulations, historical data are utilized to have a realistic measure of outcomes and balancing goals.

performed with an iterative procedure. Subtracting from the prosumers with the maximum injection values to the grid (class3) from prosumers with the maximum draw values from the grid (class1), essentially enforcing a virtual link between them. Since rules are set, the balancing process takes into consideration just class1 and class3 prosumers because class2 prosumers are already balanced. Table 41 shows an example of this process by linking Prosumer's 83 injection to Prosumer's 80 draw resulting in a balanced value. The same applies for Prosumer 48 and Prosumer 77, as well as Prosumer 54 and Prosumer 78, respectively.

Table 41. Example after balancing for timestamp: '2017-01-28 12:00:00'.

ID	ID2	E0class3	E0ckass1	balancedE0a
83	80	11,998	-3,089	8,909
48	77	5,906	-1,774	4,132
54	78	6,450	-2,605	3,845

Once the balanced dataset is formed, preprocessing is performed to standardize (Pedregosa *et al.*, 2015) the dataset to be fed into two algorithms for deciding the best possible choice for the number k of clusters. The algorithms used for comparison reasons is WCSS (Pedregosa *et al.*, 2015) and the silhouette score (Pedregosa *et al.*, 2015).

The data are standardized so that all input values are to be centred on the value zero. This would indicate a perfect balance since it means that the energy generated and the energy consumed, are exactly equal. This is achieved by enforcing the following formula (30):

$$y = \frac{x_i - k}{s} \quad (30)$$

where y is the data which are rescaled in a way that k=0 and s=1.

WCSS is used to measure the sum of distances of the available observations from the cluster centroid as stated in (31):

$$WCSS = \sum_{i \in n} (X_i - Y_i)^2 \quad (31)$$

where Y_i is the centroid for observation X_i.

The silhouette coefficient for all data points I in a partition of k clusters is given by (32):

$$\bar{s}_k = \frac{1}{|I|} \sum_{i=1}^{|I|} s(i) \quad (32)$$

The higher the value of $\bar{s}_k$the better the quality of clustering.

After performing WCSS to the data, it is observed that the ideal number of clusters is three or four, as shown in Fig. 30.

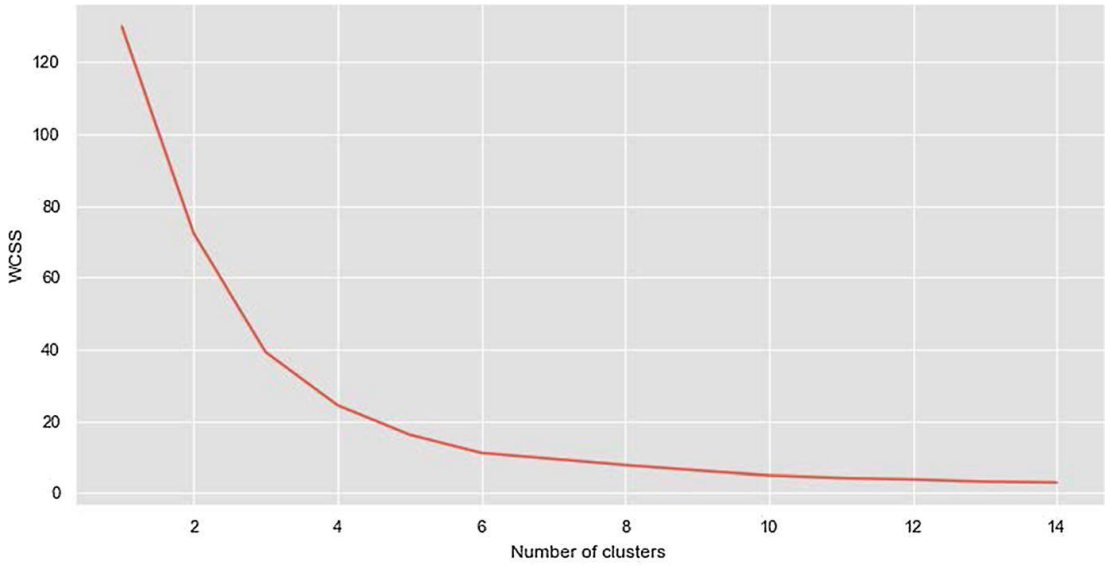

Fig. 30. WCSS for defining number of clusters after balancing for timestamp: '2017-01-28 12:00:00'.

To conclude on the number k of clusters required by the clustering method, the silhouette score is also utilized. It results in four as the ideal number of clusters, as shown in Table 42. Silhouette coefficients for defining the number of clusters after balancing for timestamp: '2017-01-28 12:00:00'., where the maximum value for the silhouette coefficient is 0.5183 for four clusters.

Table 42. Silhouette coefficients for defining the number of clusters after balancing for timestamp: '2017-01-28 12:00:00'.

Number of clusters	Silhouette coefficient
2	0.4722
3	0.4934
4	**0.5183**
5	0.4514

6	0.4360
7	0.4134
8	0.4255
9	0.4189
10	0.4516
11	0.4435
12	0.4384
13	0.4551
14	0.4386

Next, k-means clustering algorithm (Arthur and Vassilvitskii, 2007) is used to form clusters that will conceptually form the P2P balanced VMGs abiding with the data feed to the approach of this simulation. k-means partitions n objects into k clusters in a way that each object belongs to the cluster with the nearest mean (33).

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} x_i^{(j)} - C_j^2 \quad (33)$$

where k is the number of clusters, n the number of cases, x_i case i and C_j the centroid for cluster j. The distance function is calculated as: $x_i^{(j)} - C_j^2$.

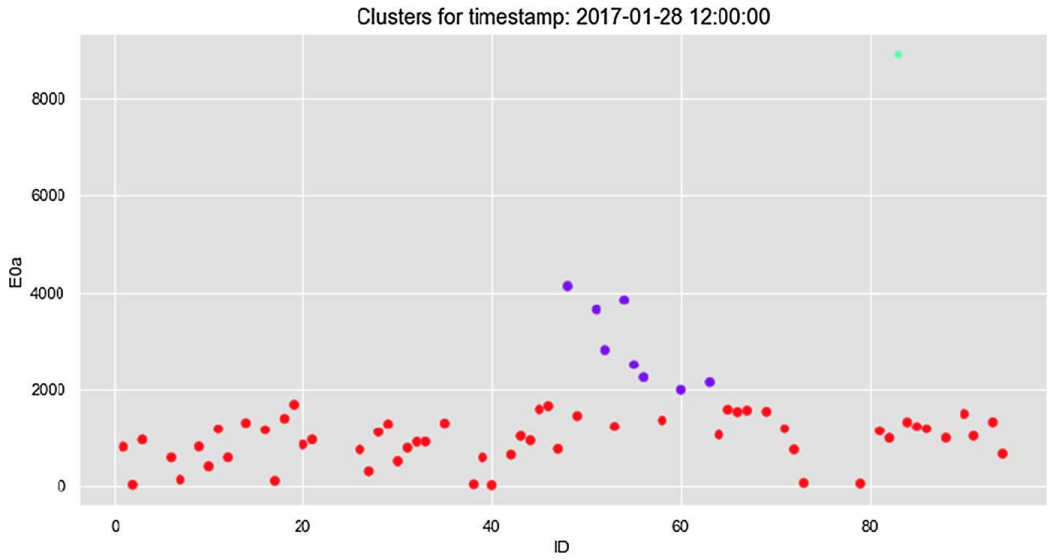

Fig. 31. VMGs formed for k=3.

Fig. 31 and Fig. 32 depict the output of k-means for k = 3 and k = 4 respectively. It is evident that at this specific timestamp, even after balancing, there is an outlier due to the PV production from Prosumer #83 (Table 41).

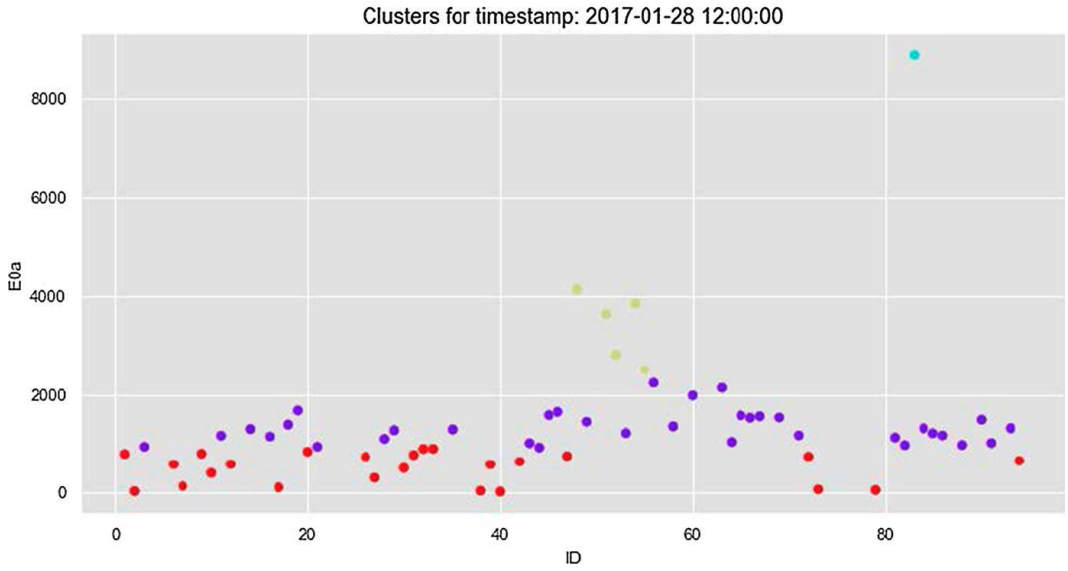

Fig. 32. VMGs formed for k=4.

The captions in Appendix K show how clustering is performed after linking for each timestamp the prosumer with the max value with the consumer with the min value. This simulation is labelled “V1 of energy balancing”.

In the next two simulations, binning is performed on the raw dataset, according to two distinct binning methods QCUT and CUT, respectively (McKinney, 2010). The output of these methods is depicted in Appendix L. It shows the overview of prosumer allocation after using binning with both techniques. The reason for doing so, is to enhance the visualization and analytical capabilities of the proposed approach, offering more options for the stakeholders when considering forming VMGs.

8.1.2 Simulation#2

This simulation describes binning as it is performed incorporating QCUT (McKinney, 2010) on raw data. This method attempts to create bins with almost the same number of consumptions/ productions entries on each of the bins (Fig. 33).

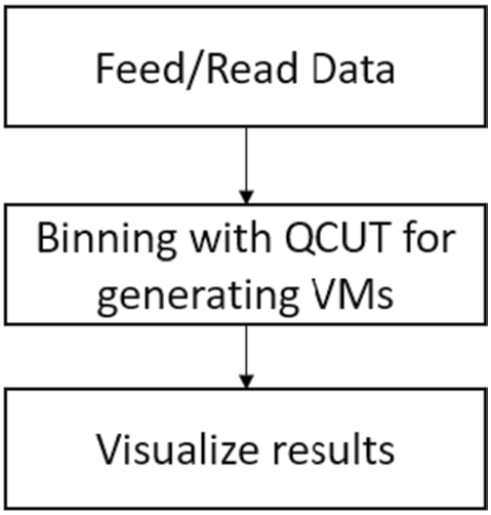

Fig. 33. EBA Simulation#2 methodology & results.

Step 1: Given a specific timestamp, the QCUT method is performed on data, resulting in five bins of prosumers.

Step 2 The aggregated consumption/generation for each bin is calculated. As stated earlier, the values signed as negative represent energy drawn from the grid, while values signed as positive are values that inject energy to the grid.

Step 3 Visualize and elaborate on the results. At this point plots are generated to offer an analytical explanation of the conducted experimentation. Fig. 34 showcases the process and output of Step 1 (each bin contains nodes with the same colour), while Table 43 showcases the process and output of Step 2.

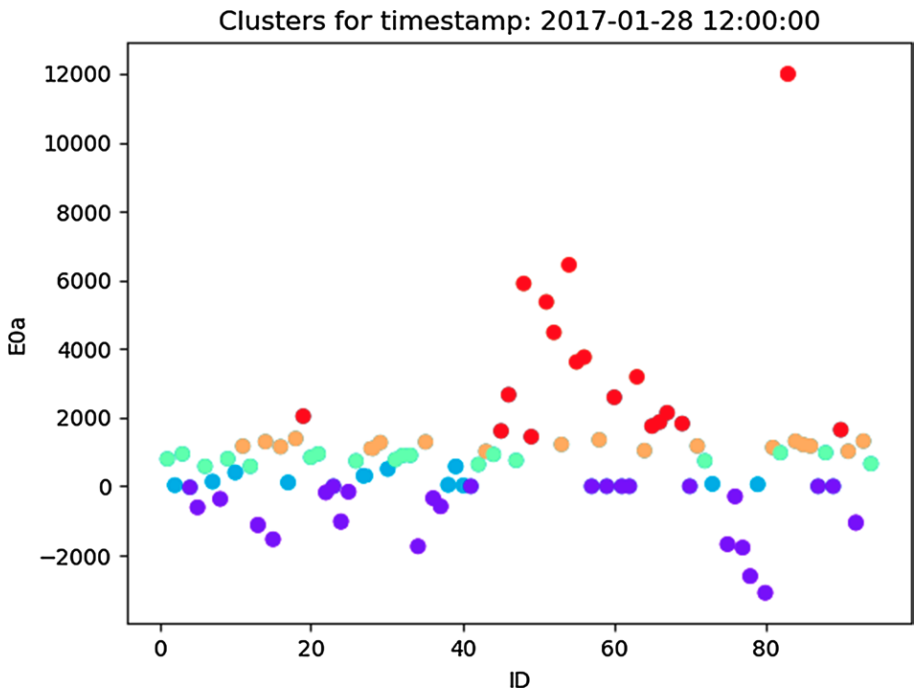

Fig. 34. QCUT on raw data.

Table 43 refers to Fig. 34 as it describes the count of prosumers that were placed on each of the bins along with the overall consumption/production. Appendix M shows how QCUT binning performs on specific timestamps within a single day (sun hours). This simulation is labelled “V2 of energy balancing”.

Table 43. Prosumer QCUT binning with sums of consumption/production per bin.

Bin label	E0a count	Sum
1	26	-18,197
2	11	2,255
3	18	14,458
4	18	21,635
5	18	64,359

8.1.3 Simulation#3

This simulation describes binning with the CUT (McKinney, 2010) method on raw data. This method tries to create bins that contain consumptions/ productions entries within similar ranges i.e., belonging to the same value group, but with different bin frequency (Fig. 35).

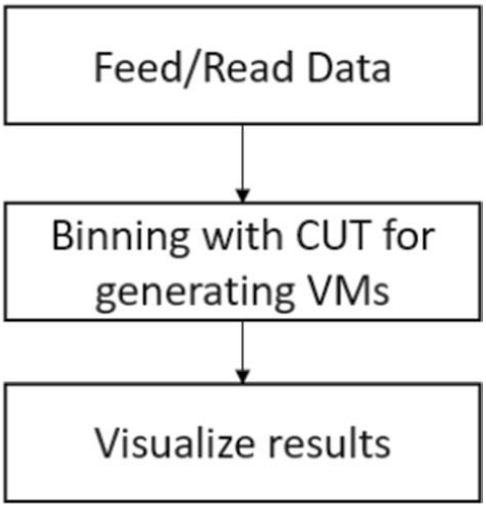

Fig. 35. EBA Simulation#3 methodology & results.

Step 1: Similarly, with the previous simulation, given a specific timestamp, CUT is performed on data resulting in five bins of prosumers.

Step 2 The aggregated consumption/ production is calculated for each bin. As stated earlier, the values signed as negative represent energy drawn from the grid while values signed as positive inject energy to the grid.

Step 3 The results are visualized and elaborated on. At this point plots are generated to offer an analytical explanation for this simulation. Fig. 36, explains the process and output of Step 1 (each bin contains nodes with the same color), while Table 44 showcases the process and output of Step 2.

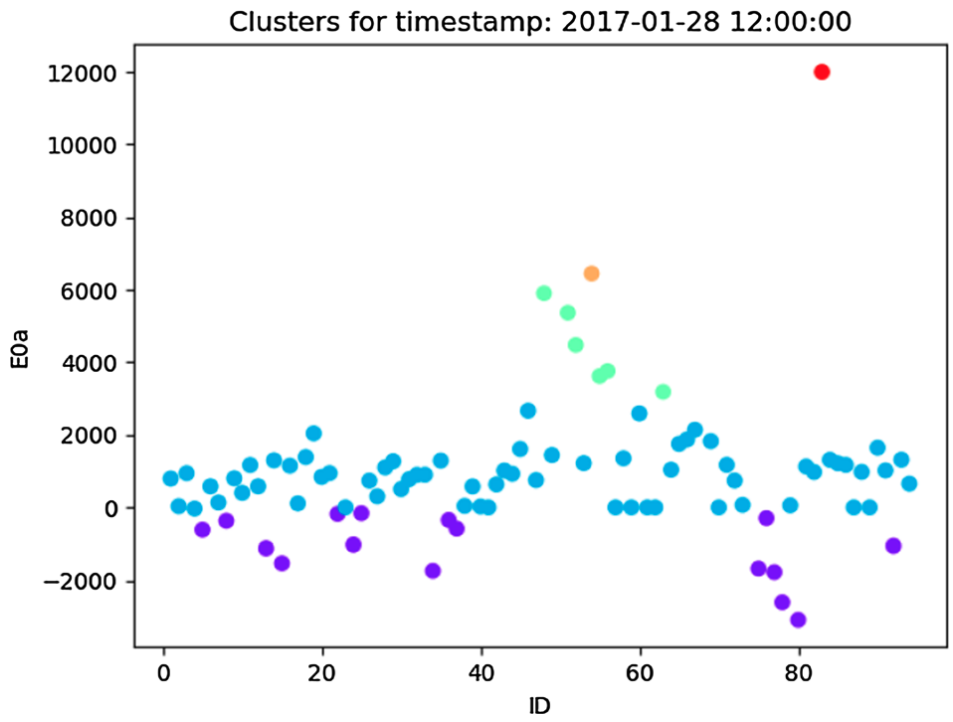

Fig. 36. CUT on raw data.

Table 44 refers to Fig. 36, as it describes the number of prosumers placed in each bin along with the overall consumption/production.

Table 44. Prosumer CUT binning with sums of consumption/production per bin.

Bin label	E0a count	Sum
1	16	-18,165
2	67	57,902
3	6	26,325
4	1	6,450
5	1	11,998

Appendix N shows how CUT binning performs on specific timestamps within a single day (sun hours). This simulation is labeled “V3 of energy balancing”.

8.1.4 Simulation#4

The last simulation outlines the final stage of the approach (Fig. 37); the final definition of the VMGs, incorporating decision making and EBA. This involves two steps.

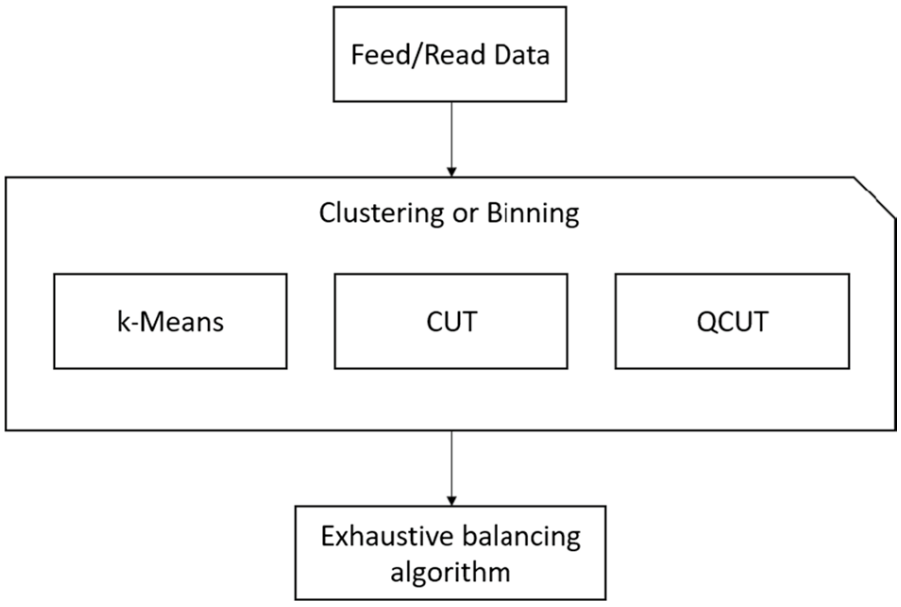

Fig. 37. EBA Simulation#4 methodology & results.

Step 1: At this point, a decision on which clustering, or binning process is more suitable for balancing (balanced VMG formation), needs to be taken. In case a naïve P2P balancing and clustering of prosumer VMGs is requested to act as a baseline estimator for an energy balancing operation, the output of Simulation#1 is utilized. In case the formation of VMGs based on a relative measure of consumptions/ productions is requested, the CUT method for binning is employed i.e., the output of Simulation#2. Finally, in case the absolute measure of consumptions/ productions is requested to form VMGs, the QCUT method for binning is employed i.e., the output of Simulation#3. These options are offered aiming to increase the number of possible outcomes of this simulation, whilst acting as individual and distinct sub-results that feed the main EBA algorithm, expanding the analytical capabilities of the proposed approach, for evaluation purposes.

Step 2: Once the binning measure or clustering technique is determined, balancing between the available bins or clusters (VMGs) that are formed is performed. This is achieved by utilizing a method that effectively decides which bins or clusters need to be linked with one or more additional bins/clusters. A customized exhaustive (brute-force) searching algorithm (referred as EBA) is implemented, that creates lists of the best matching values (consumptions/ productions) it receives as input, which are closer to user defined numeric target values.

The following example uses as input the output of one of the binning approaches (Simulations#2 or #3) and utilizes EBA, described in Step 2, to perform balancing on a VMG level (VMG2VMG balancing). Consider the following arrays:

Numbers= [−12277, −14027, +33929, −33000, +12630, −21612, +520, +24000, + 14000] and Targets = [+500, +1500, −1000],

where each entry on the Numbers array refers to the aggregated values of either draw (signed with "-") or injection (signed with "+"), and each entry of the Targets array refers to the requested by the stakeholder (DSO, aggregator) energy target value for the VMG (or cluster of prosumers).

Executing the balancing algorithm, inputting as arguments the two arrays, search_closest_to_target (Numbers, Targets), the EBA output is the following (Table 45).

Table 45. EBA output examples.

Arithmetic target value to approach for balancing	List of combined values forming a VMG
500	[-14027, 520, 14000]
1500	[33929, -12277, 12630, -33000]
-1000	[24000, -21612]

8.2 ENERGY LOAD FORECASTING

In this section the evaluation techniques and simulation results for the proposed approach are presented. To that end, the results are demonstrated in a comparative way highlighting the findings during the experimentation phase. Once the presentation of findings concludes a discussion follows attempting to highlight the identified biases of the proposed approach.

The proposed novel approach attempts to forecast energy load retrieving historical data from spring 2018 from a novel nZEB smart Home (Koukaras, Bezas, *et al.*, 2021). The performance of all methods and algorithms utilized in this study are summarized in Table 46, Table 47 and Table 48 respectively. The best performing predictions resulting from ensemble and prediction algorithms under the novel timeseries forecasting approach are emphasized (per metric) using bold.

Table 46. Ensemble utilizing Averaged Prediction (EAP)[22].

Prediction Model	MAPE (%)	SMAPE (%)	RMSE (kWh)	ET (seconds)
MLP	**12.962**	**6.489**	**0.651**	238
LSTM	14.093	6.885	0.668	825.5
GBT[23]	15.373	7.652	0.699	72.6
SVR	14.354	7.415	0.723	2.05
LR	14.506	7.062	0.664	**0.37**

Table 47. Ensemble utilizing Weighted Averages (EWA).

Prediction Model	MAPE (%)	SMAPE (%)	RMSE (kWh)	ET (seconds)
MLP	13.234	**6.612**	0.668	232
LSTM	15.191	7.236	**0.663**	960
GBT	15.116	7.488	0.698	72
SVR	**13.201**	6.706	0.684	2.13
LR	14.239	6.915	0.674	**0.42**

Table 48. Ensemble utilizing Polynomial Exhibitor (EPE).

Prediction Model	MAPE (%)	SMAPE (%)	RMSE (kWh)	ET (seconds)
MLP	17.298	8.113	0.686	227.9
LSTM	17.138	7.986	**0.679**	858.73
GBT	**15.113**	7.485	0.699	83.64
SVR	17.323	9.507	0.981	6.88
LR	15.293	**7.433**	0.686	**2.6**

Observing Table 46, Table 47 and Table 48 is evident that according to the examined scenarios the most optimistic forecasting accuracy per ensemble method is attained from the combination of:

i) EAP ensemble with the MLP prediction algorithm with the values of 12.962% for MAPE, 6.489% for SMAPE and 0.651 kWh for RMSE.

[22] Enforcing the same numeric precision structure along this section, MAPE, SMAPE and RMSE values are rounded off to three digits. ET values are rounded off to two digits.
[23] Grid Search algorithm (Rodrigues and Trindade, 2018) chooses the best parameters of GBT (n_estimators and max_depth) for each model step of the ensemble methodology.

ii) EWA ensemble with the SVR prediction algorithm with the values of 13.201% for MAPE, MLP with 6.612% for SMAPE and LSTM with 0.663 kWh for RMSE.

iii) EPE ensemble with the GBT prediction algorithm with the values of 15.113% for MAPE, LR with 7.433% for SMAPE and LSTM with 0.679 kWh for RMSE.

It is noted that, ET is calculated aiming to present the process time required for performing the predictions in all the experimental combinations. Generally, NNs (MLP, LSTM) take more time to execute, since they are much more complex approaches and by nature integrate more computationally expensive calculations. Therefore, MLP and LSTM implementations are expected to display high values of ET. On the other hand, ensemble methods with LR and SVR predict much faster. For example, the EPA with LR executes in 0.37 seconds compared to 238 seconds with EPA with MLP (Table 46), while outputting similar accuracy output as measured with the SMAPE metric (7.062 and 6.489 respectively).

This is expected, since NNs involve much more computationally expensive processes compared to the simpler statistical models like LR, as reported in Sect. 7.3.4. In this study the focus is made on the aspects of accuracy of forecasting and secondarily on the time required for executing the experiments. More precisely, ET acts as an extra generic estimator or parameter for deciding on the options on how to run the experimentations. All that mentioned, the three ensemble methods combined with LR can act as the baseline estimator for accuracy that can be executed very fast with good output results, allowing for a more established comparison with other more "sophisticated" estimators, such as NNs.

Table 49. Summary of top three best performing ensemble methods and algorithms.

Ensemble & Prediction Model	MAPE (%)	SMAPE (%)	RMSE (kWh)
EAP + MLP	12.962	6.489	0.651
EWA + MLP	13.234	6.612	0.668
EWA + GBT	13.201	6.706	0.684

A separate table (Table 49) is generated for highlighting these top three performing combinations of ensemble methods and prediction algorithms abiding with the proposed methodology for one step ahead load forecasting that eases on evaluating the overall output in a ranked manner. The ranking fails to be absolutely consistent since EWA with MLP shows a greater value than EWA with GBT (it is expected to

have a smaller one) yet is evident from the rest of the observations (SMAPE and RMSE values) that the overall top three ranked remain consistent.

For clarification purposes, the smaller values approaching zero on each of the metrics is considered the best one, therefore the minimum value close to zero is considered to be the best. A summarized table of values for all accuracy metrics and methods can be found in Appendix R.

8.3 ENERGY OPTIMAL DAY-AHEAD SCHEDULING

This section presents results as scenarios for the optimization actors. These are consumers and aggregators. The two optimizations, single- and bi-objective, interact with each other, based on the following architectural scheme forming a tri-layer interaction (Fig. 38).

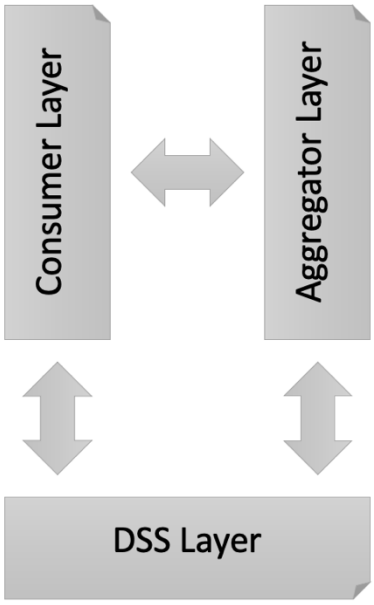

Fig. 38. Architectural layer interaction, Decision Support System and multi-objective optimization.

8.3.1 Scenario #1: Consumer

The Consumer layer optimizes consumer's energy requirements for day-ahead while integrating five personalized options for occupant discomfort. For presentation clarity reasons, just the results for a single consumer are presented. The same optimization process along with its results applies for all the consumers in the aggregator's portfolio (as described in Sect. 7.3.5).

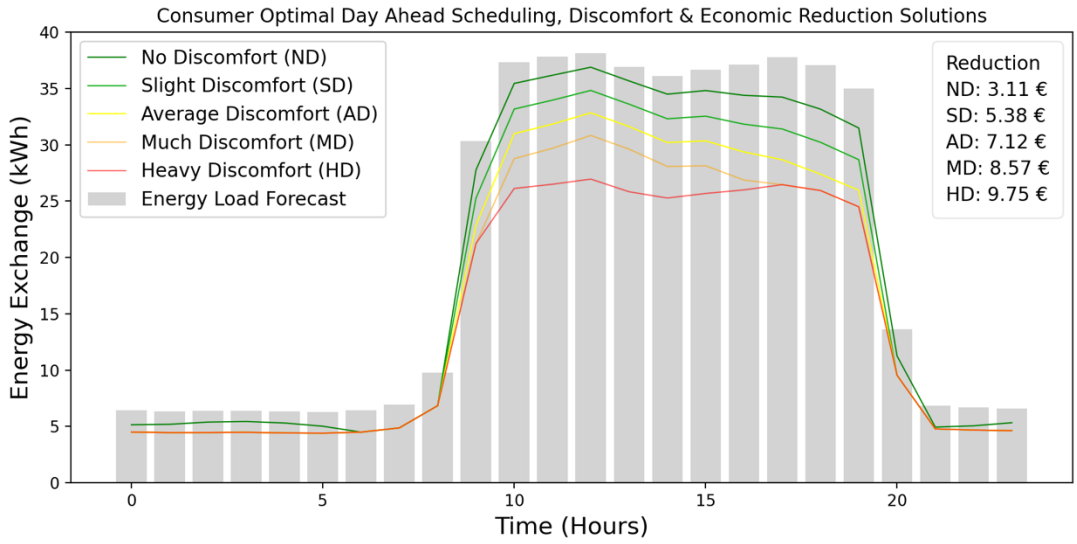

Fig. 39. Optimization for a single prosumer.

Fig. 39 shows the bi-objective optimization output of the Consumer layer for day-ahead scheduling along with discomfort options and economic reduction solutions. The five reduction solutions reflect five distinct levels of discomfort, considered in this work, that is:

i) No Optimization (NO): the initially state, prior to any optimization. Serves as a baseline scenario.

ii) No Discomfort (ND): the case, where the minimum amount of discomfort is considered.

iii) Slight Discomfort (SD): the case, where a slightly level of discomfort is considered.

iv) Average (AD): the case, where a medium level of discomfort is considered.

v) Much Discomfort (MD): the case, where a greater level of discomfort than in the above case is considered.

vi) Heavy Discomfort (HD): the case, where the greatest amount of discomfort is considered.

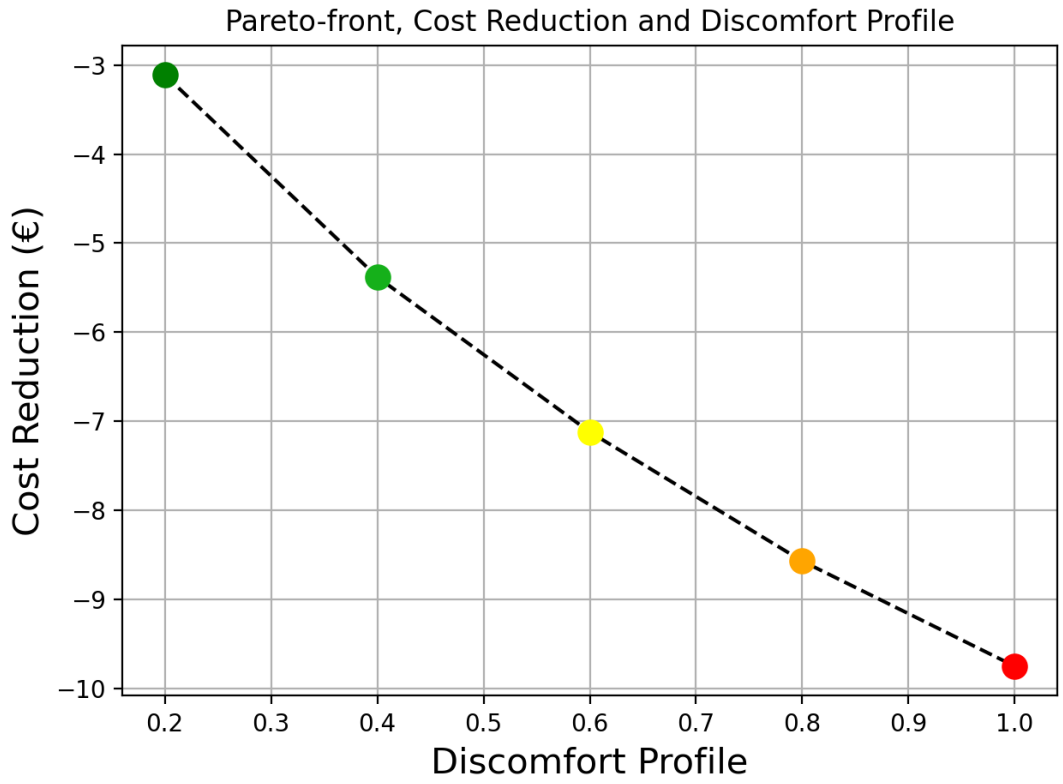

Fig. 40. Pareto front of optimal solutions.

These solutions being the outcome of the bi-objective optimization approach are presented in the Pareto front in Fig. 40. Appendix S outlines the cost minimization for consumer, based on those five available options. The values presented are for a period of 24 hours, i.e., a day, expressed in kWh and in relative change form (in %), in terms of the initial value, i.e. No Opt Scenario. Fig. 39 and Appendix S present results for one day within the examined period of 2019-02-01 to 2019-02-27, that is 2019-02-07. Similar results have been produced for every day of the examined period.

8.3.2 Scenario #2: Aggregator

As in the case of the consumer, cost minimization is the objective, with the difference only in scale and constrained by the flexibility of the consumer portfolio as described more thoroughly in Sect. 7.3.5. The DSS layer optimization outcome for the aggregator's consumer portfolio for the day-ahead energy scheduling, can be seen in Fig. 41.

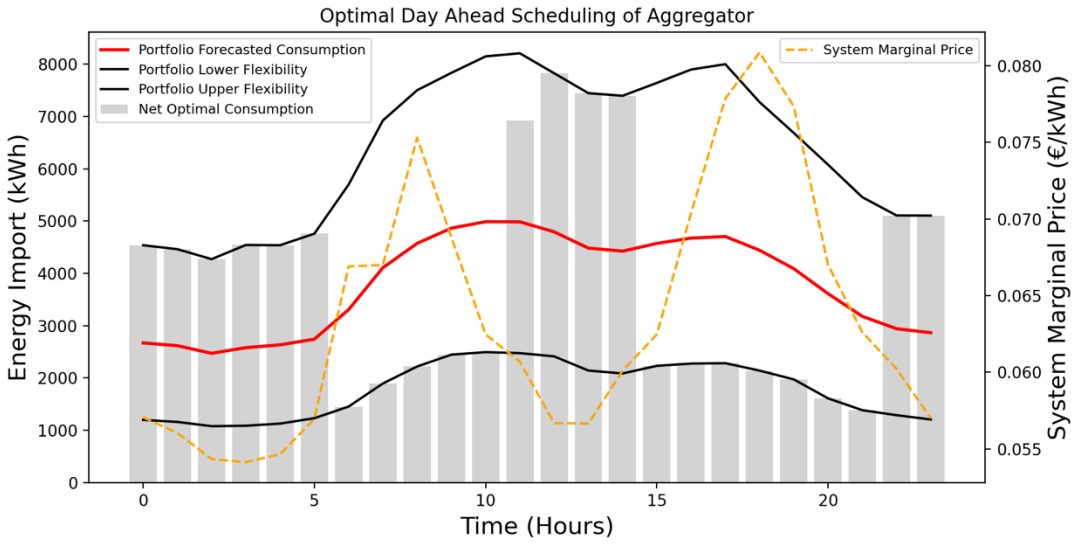

Fig. 41. Optimization for aggregator portfolio.

The upper and lower flexibility bounds can be seen, as a set for the whole portfolio in black, and the forecasted values in red. The orange dashed line depicts the SMPr values and grey bars show the final optimized energy consumption of the portfolio per hour. The way the actual energy consumption is modified is depicted by the forecasted red line to the grey bars according to the SMPr values; whenever the SMPr is high, the consumption is lowered to the lowest possible value; whenever the SMPr is low, the energy consumption is regained, reaching the highest values possible, in order to have the same total daily energy consumption in the end. According to this scheme, the DR signals to be sent to each of the consumers are calculated. Every time a deviation between the forecasted energy consumption and the optimize done exists, DR signals are to be sent to the consumers. Since the whole DR portfolio is included, the idea is that each consumer should contribute according to his/her capabilities, that is their attributed flexibility, in order to have a fair strategy among them. First, their contribution to the total portfolio flexibility is calculated for every hour as formulated in the following equations (34) and (35) (Koukaras, Gkaidatzis, *et al.*, 2021):

$$flex_{i,lb_{\%}}(t) = \frac{E_{i,for}(t) - flex_{i,lb}(t)}{\sum_{i=1}^{N_{tot\_C}} E_{i,for}(t) - \sum_{i=1}^{N_{tot\_C}} flex_{i,lb}(t)} \quad (34)$$

$$flex_{i,ub_{\%}}(t) = \frac{flex_{i,ub}(t) - E_{i,for}(t)}{\sum_{i=1}^{N_{tot\_C}} flex_{u,lb}(t) - \sum_{i=1}^{N_{tot\_C}} E_{i,for}(t)} \quad (35)$$

Then, this is applied to the difference between the optimized and originally forecasted energy consumption for that hour, and the required amount of energy in kWhs to be reduced, or increased. In that case, it is calculated for every consumer, as formulated in the following equations (36) and (37) (Koukaras, Gkaidatzis, *et al.*, 2021), that is:

$$flex_{i,lb}(t) = \begin{cases} flex_{i,lb_{\%}}(t) * \left[\sum_{i=1}^{N_{tot\_C}} E_{i,for}(t) - \sum_{i=1}^{N_{tot\_C}} E_{opt}(t)\right], \sum_{i=1}^{N_{tot\_C}} E_{i,for}(t) > \sum_{i=1}^{N_{tot\_C}} E_{i,opt}(t) \\ 0 \qquad , \sum_{i=1}^{N_{tot\_C}} E_{i,for}(t) < \sum_{i=1}^{N_{tot\_C}} E_{i,opt}(t) \end{cases} \tag{36}$$

$$flex_{i,ub}(t) = \begin{cases} 0 \qquad , \sum_{i=1}^{N_{tot\_C}} E_{i,for}(t) > \sum_{i=1}^{N_{tot\_C}} E_{i,opt}(t) \\ flex_{i,ub_{\%}}(t) * \left[\sum_{i=1}^{N_{tot\_C}} E_{i,for}(t) - \sum_{i=1}^{N_{tot\_C}} E_{opt}(t)\right], \sum_{i=1}^{N_{tot\_C}} E_{i,for}(t) < \sum_{i=1}^{N_{tot\_C}} E_{i,opt}(t) \end{cases} \tag{37}$$

Fig. 42 demonstrates an example, highlighting a single optimal solution for part of the portfolio posing as DR signals, e.g. four consumers out of the whole portfolio. Appendix T offers a more detailed representation, each column presenting a consumer and each row showing the lowest and upper bounds of flexibility contribution per timestamp. These values stand for the individualized deviations (per consumer) from the Portfolio Forecasted Consumption depicted in Fig. 41. The aforementioned results for the same day period are presented, as those in Fig. 39 and Fig. 41.

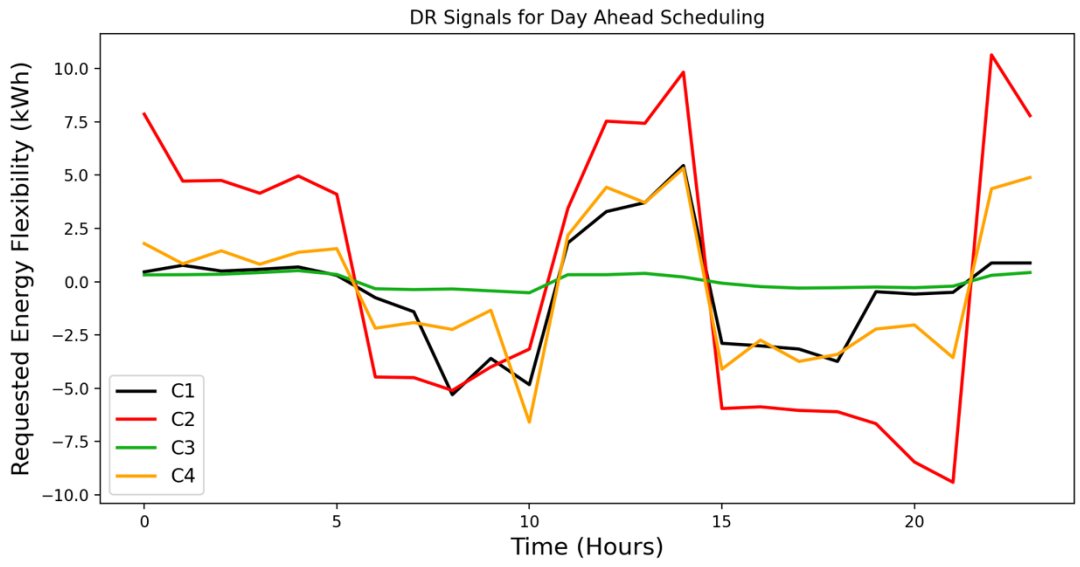

Fig. 42. DR signals sent to consumers C1, C2, C3 and C4.

Finally, in Table 50 a representative sample of portfolio savings for the whole examined period between 2019-02-01 and 2019-02-07 is presented.

Table 50. Portfolio savings for a week (2019-02-01 up to 2019-02-07).

Timestamp	Portfolio Initial Cost (€)	Optimization Savings (€)	Final Cost (€)	Cost Savings (%)
2019-02-01	6,282.314	441.9311	5,840.0329	-7.0345
2019-02-02	4,799.9383	315.214	4,484.7243	-6.567
2019-02-03	4,393.4721	332.1233	4,061.3488	-7.5594
2019-02-04	6,114.7628	498.09	5,616.6728	-8.1456
2019-02-05	5,715.1828	235.5469	5,479.6359	-4.1214
2019-02-06	5,424.4863	351.2224	5,073.2639	-6.4747
2019-02-07*	5,892.0956	344.1329	5,547.9627	-5.8406

* This timestamp relates to portfolio savings resulting from optimization output presented in Fig. 41.

8.3.3 Proposed Demand Response Scheme

The proposed DR scheme envisions an autonomous and dynamic improvement for optimal energy management involving two or more actors. At its current form, it utilizes the optimization outputs for consumer and aggregator attempting to enhance their collaboration while pertaining certain levels of freedom of choice for managing assets.

Both the consumer and the aggregator may choose their own optimized schedules, according to the electricity price and in parallel. The conception of such a DR scheme generates new capabilities for energy contract relaxation, while allowing optimization on the P2P level. More specifically, since optimization for day-ahead energy load scheduling takes place, the aggregator publishes an hourly DR schedule for all portfolio assets, that is consumers in his/her case, at a fixed timestamp the day before. For the examined use cases, this means on the 2019-02-06 at 18:00, since in most markets the SMPr for the day-ahead is determined around noon on the previous one. Then an assumption is made that consumers have to accept the DR signals and a two-hour period is provided for them to respond, that is 19.00-21.00. The decision of whether to accept or reject the DR signals received, is based mainly on their alignment with the optimized day-ahead schedule already selected by each consumer. Thus, some DR signals may only be rejected, while others may be either Accepted or Rejected.

This scheme is implemented by the DSS layer and generates default options for the involved actors (Fig. 43).

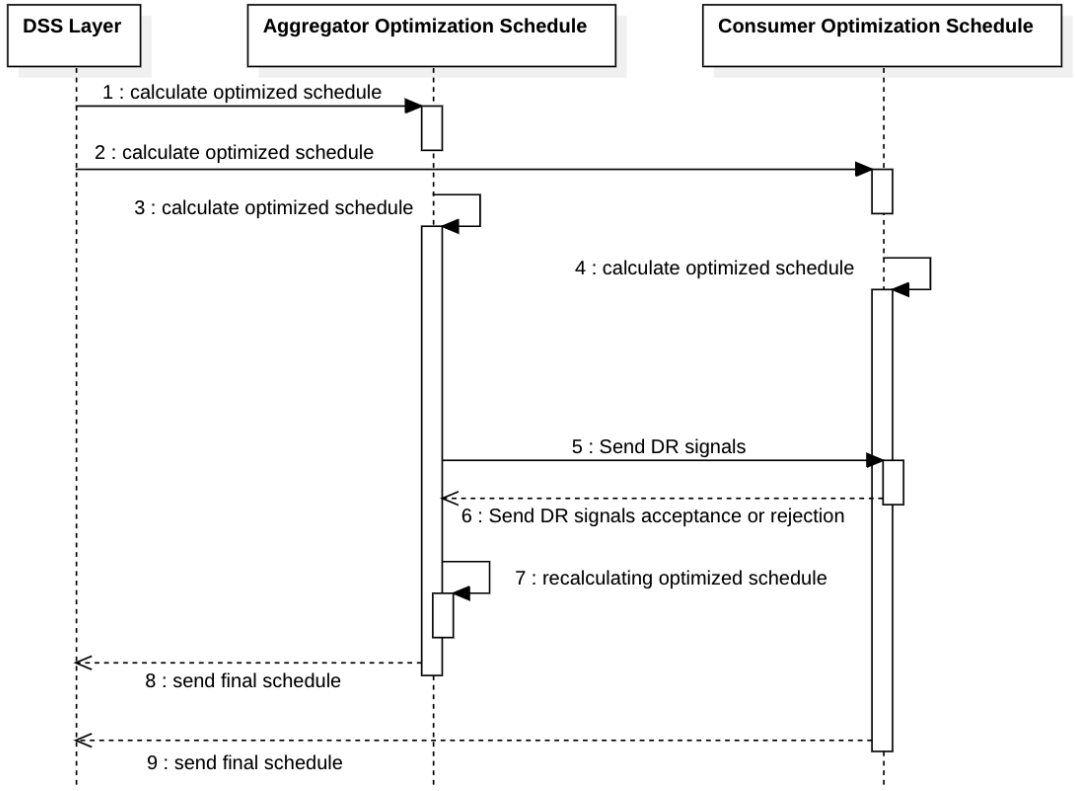

Fig. 43. Proposed DR scheme.

Next, three use cases of the proposed DR Scheme for a single consumer (C1) are presented, considering the results from Appendix S and Appendix T. Table 51, Table 52 and Table 53 show the envisioned DR scheme functionality for consumer (No Opt, SD and HD) to aggregator interaction, allowing consumers to choose which DR signals to accept, or decline. This type of interaction allows a more dynamic adjustment of the energy load while retaining certain levels of discomfort or economic benefits.

Table 51. Envisioned DR scheme and interaction between consumer and aggregator. No optimization for consumer.

t	Energy Load Forecast (No Opt)	DRS (C1)	Requested Energy Load (C1)	Status
0	6.418	0.46	6.878	Reject
1	6.347	0.77	7.117	Reject
2	6.356	0.5	6.856	Reject
3	6.386	0.58	6.966	Reject
4	6.309	0.69	6.999	Reject

5	6.276	0.3	6.576	Reject
6	6.409	-0.75	5.659	Accept/Reject
7	6.951	-1.41	5.541	Accept/Reject
8	9.763	-5.3	4.463	Accept/Reject
9	30.334	-3.6	26.734	Accept/Reject
10	37.325	-4.83	32.495	Accept/Reject
11	37.855	1.82	39.675	Reject
12	38.13	3.29	41.420	Reject
13	36.906	3.71	40.616	Reject
14	36.11	5.45	41.560	Reject
15	36.691	-2.89	33.801	Accept/Reject
16	37.139	-3.01	34.129	Accept/Reject
17	37.803	-3.16	34.643	Accept/Reject
18	37.065	-3.74	33.325	Accept/Reject
19	34.999	-0.47	34.529	Accept/Reject
20	13.621	-0.58	13.041	Accept/Reject
21	6.819	-0.5	6.319	Accept/Reject
22	6.673	0.88	7.553	Reject
23	6.593	0.88	7.473	Reject

Table 52. Envisioned DR scheme and interaction between consumer and aggregator. Slight Discomfort (SD) for consumer.

t	Energy Load Forecast (No Opt)	Min Energy Load (SD)	DRS (C1)	Requested Energy Load (C1)	Status
0	6.418	4.493	0.46	4.953	Accept/Reject
1	6.347	4.443	0.77	5.213	Accept/Reject
2	6.356	4.449	0.5	4.949	Accept/Reject
3	6.386	4.47	0.58	5.050	Accept/Reject
4	6.309	4.416	069	5.106	Accept/Reject
5	6.276	4.393	0.3	4.693	Accept/Reject
6	6.409	4.486	-0.75	3.736	Accept/Reject
7	6.951	4.866	-1.41	3.456	Accept/Reject
8	9.763	6.834	-5.3	1.534	Accept/Reject
9	30.334	25.262	-3.6	21.662	Accept/Reject
10	37.325	33.18	-4.83	28.350	Accept/Reject
11	37.855	33.97	1.82	35.790	Accept/Reject
12	38.13	34.831	3.29	38.121	Accept/Reject
13	36.906	33.611	3.71	37.321	Reject
14	36.11	32.312	5.45	37.762	Reject
15	36.691	32.546	-2.89	29.656	Accept/Reject
16	37.139	31.834	-3.01	28.824	Accept/Reject
17	37.803	31.408	-3.16	28.248	Accept/Reject
18	37.065	30.227	-3.74	26.487	Accept/Reject
19	34.999	28.68	-0.47	28.210	Accept/Reject
20	13.621	9.535	-0.58	8.955	Accept/Reject
21	6.819	4.773	-0.5	4.273	Accept/Reject
22	6.673	4.671	0.88	5.551	Accept/Reject
23	6.593	4.615	0.88	5.495	Accept/Reject

Table 53. Envisioned DR scheme and interaction between consumer and aggregator. Heavy Discomfort (HD) for consumer.

t	Energy Load Forecast (No Opt)	Min Energy Load (HD)	DRS (C1)	Requested Energy Load (C1)	Status
0	6.418	4.493	0.46	4.953	Accept/Reject
1	6.347	4.443	0.77	5.213	Accept/Reject
2	6.356	4.449	0.5	4.949	Accept/Reject
3	6.386	4.47	0.58	50.050	Accept/Reject
4	6.309	4.416	0.69	5.106	Accept/Reject
5	6.276	4.393	0.3	4.693	Accept/Reject
6	6.409	4.486	-0.75	3.736	Accept/Reject
7	6.951	4.866	-1.41	3.456	Accept/Reject
8	9.763	6.834	-5.3	1.534	Accept/Reject
9	30.334	21.234	-3.6	17.634	Accept/Reject
10	37.325	26.127	-4.83	21.297	Accept/Reject
11	37.855	26.499	1.82	28.319	Accept/Reject
12	38.13	26.96	3.29	30.250	Accept/Reject
13	36.906	25.834	3.71	29.544	Accept/Reject
14	36.11	25.277	5.45	30.727	Accept/Reject
15	36.691	25.684	-2.89	22.794	Accept/Reject
16	37.139	25.997	-3.01	22.987	Accept/Reject
17	37.803	26.462	-3.16	23.302	Accept/Reject
18	37.065	25.945	-3.74	22.205	Accept/Reject
19	34.999	24.499	-0.47	24.029	Accept/Reject
20	13.621	9.535	-0.58	8.955	Accept/Reject
21	6.819	4.773	-0.5	4.273	Accept/Reject
22	6.673	4.671	0.88	5.551	Accept/Reject
23	6.593	4.615	0.88	5.495	Accept/Reject

Chapter 9: Analysis

This chapter discusses, interprets and evaluates results. It comments on and analyses the achievements of each methodological stage of this thesis in the domain of Energy (Sect. 9.1). The research tasks are combined and reported with a representative title based on the final evaluation of accomplishments. Therefore, a theoretical framework integrates a variety of (common) mining tasks highlighting the possibilities for knowledge acquisition utilizing interdisciplinary approaches within the context of the second part (PART II) of this thesis.

9.1 A NOVEL FRAMEWORK FOR P2P AND VMG2VMG ENERGY BALANCING, INCORPORATING ONE STEP AHEAD LOAD FORECASTING AND OPTIMIZATION FOR DAY-AHEAD ENERGY SCHEDULING

This theoretical framework integrates research accomplishments from Research Tasks 4, 5 and 6 relating to EB, ELF and EODS (according to Table 1).

9.1.1 An Interdisciplinary Approach on efficient Virtual Microgrid to Virtual Microgrid Balancing incorporating Data Preprocessing techniques

This research action proposes an approach which implements high-level energy balancing at the VMG and P2P level, while reviewing some state-of-the-art computer science and energy attempts, highlighting the necessity for an interdisciplinary approach in this problem domain. The approach utilizes heuristics through rules and unsupervised learning to form and balance VMGs. Biases in the proposed methodology are stated in Sect. 7.4.1. These biases inferably generate directions for improvements. The methodology and its connection with result outline are depicted in Fig. 44.

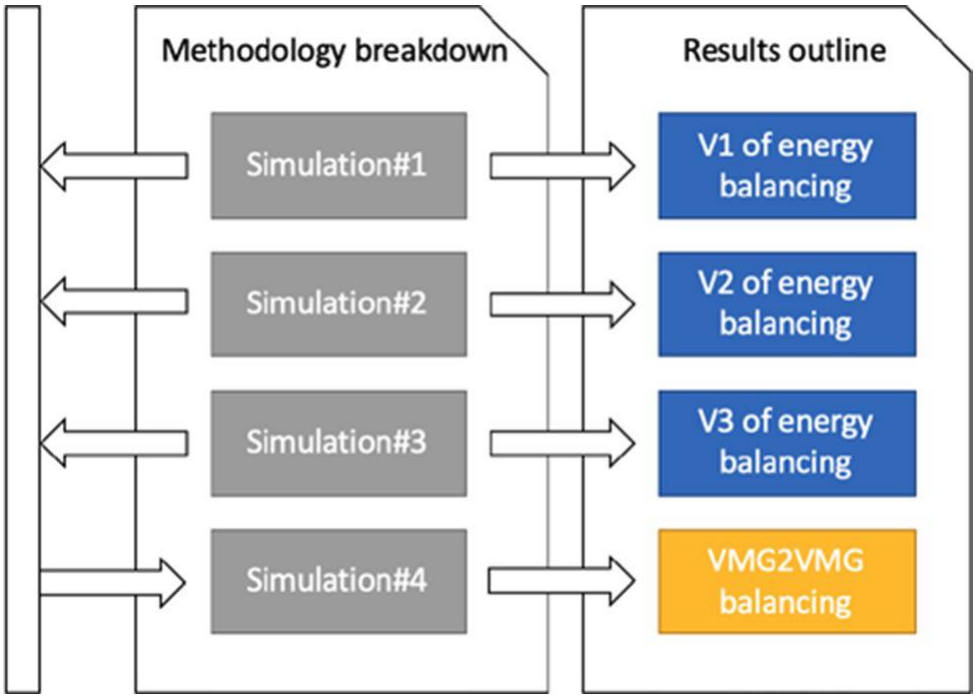

Fig. 44. EBA research summary and outcomes.

A step-by-step approach is followed, incorporating experimentation and elaboration to produce results that aid organizing next research steps described in Sect. 10.2.2. Each of the simulations, results in a version of energy balancing that can act as stand-alone component, able to feed the final simulation that deals with VMG2VMG balancing, one of the research aims of this thesis.

Simulation#1 enforces the rules defined (Sect. 7.3.2) for balancing at the P2P level. Simulation#2 utilizes a quantile-based discretization for appointing prosumers to bins. Simulation#3 transposes prosumer energy values from continuous variables to categorical, appointing the prosumer entries to equally sized bins. Finally, Simulation# 4 performs VMG2VMG balancing by utilizing EBA, an exhaustive search and balancing algorithm that is activated by the results from either of the previous simulations (#1–3).

Appendices J–N provide captions for a detailed presentation of the experimentation process. These aim at further clarifying the proposed approach by presenting the process steps analytically.

9.1.2 Introducing a Novel Approach in One Step Ahead Energy Load Forecasting

Appendices O-R contain the results of each forecasting method combined with the three ensemble methods that are tested in a visualized way (presentation using plots) as well as a summarized forecast accuracy table, acting as a point of reference for enhanced clarification of the results presented in this section.

By observing the overview of results (Appendix R) it is recognizable that most of the times the metric values fail to be consistent. To that end, it seemed more appropriate to highlight one accuracy metric for presenting the overall achievements in this study in a more clear and understandable way.

Therefore, to present the results from all three ensemble methods, highlighting the accuracy performance, the SMAPE metric is chosen. This is done because SMAPE can be considered one of the most reliable accuracy measures in percentiles. Also, according to Sect. 7.3.4 it is already clarified that when values of that metric reside between 0 and 10% are at its optimal indications.

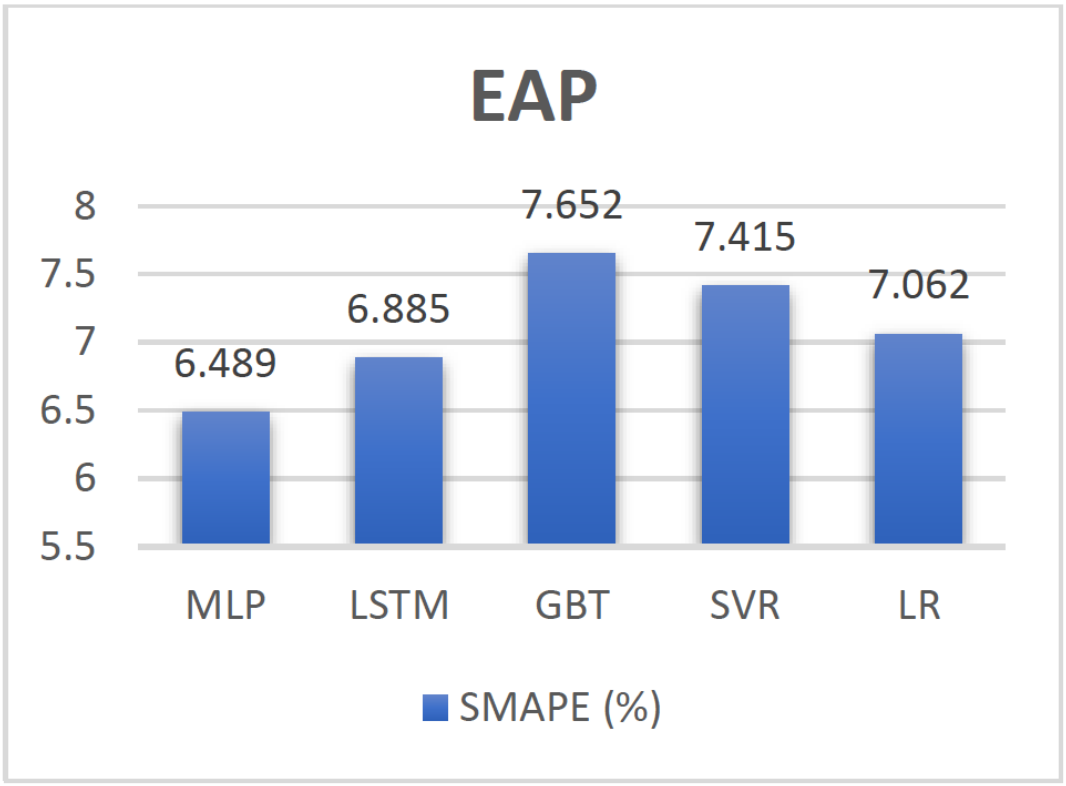

Fig. 45. SMAPE accuracy for EAP.

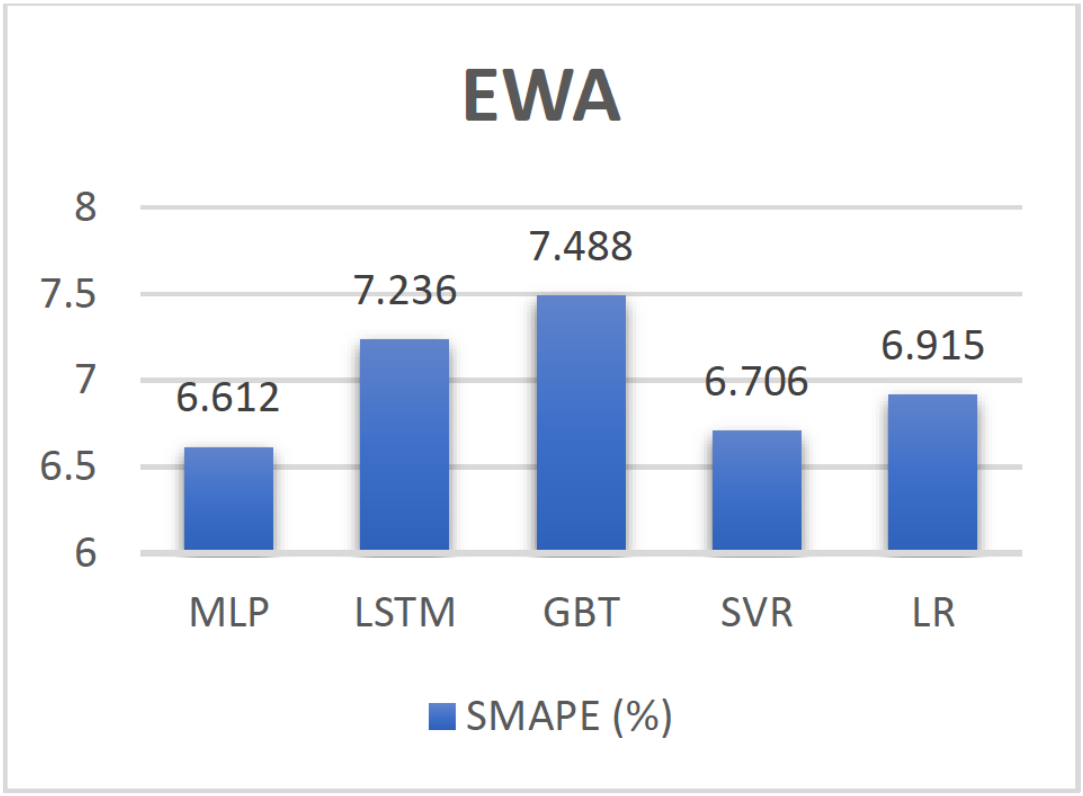

Fig. 46. SMAPE accuracy for EWA.

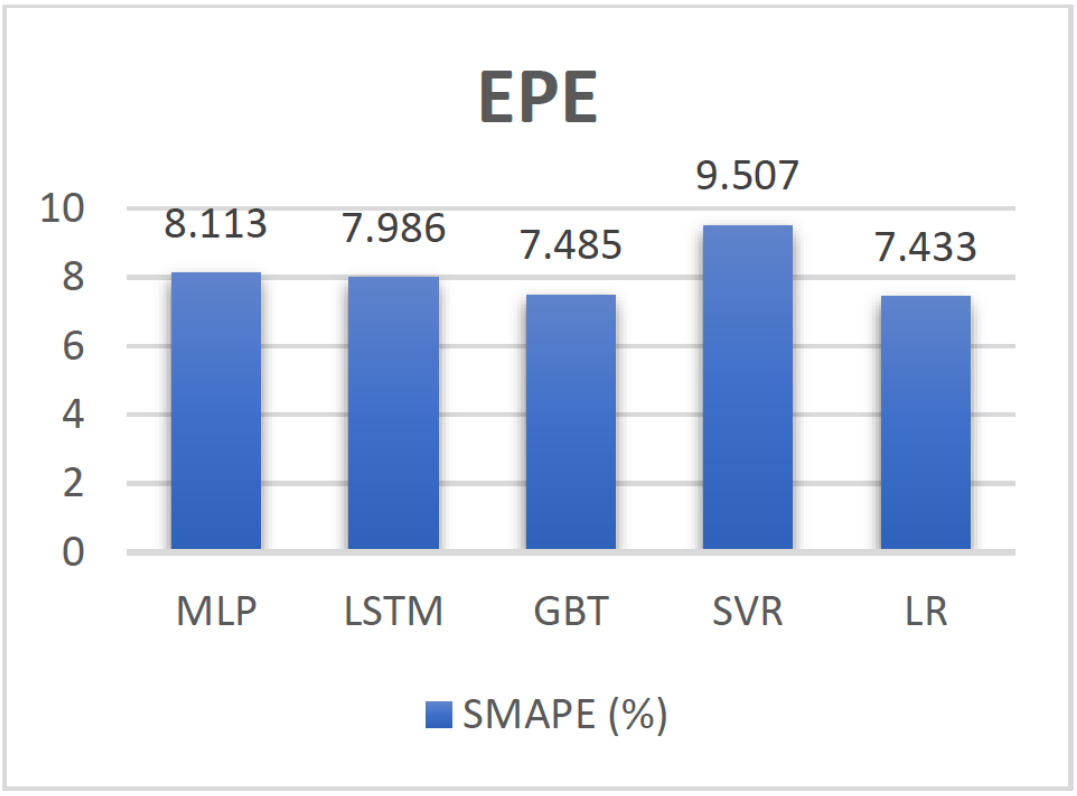

Fig. 47. SMAPE accuracy for EPE.

Finally, according to Fig. 45, Fig. 46 and Fig. 47 it is expected to have more consistent results, since the accuracy results from just one metric is presented. That would allow to perform a better comparison on the ensemble level (EAP, EWA and EPE). For example, the SVR prediction model outputs the greater deviations with EAP_SVR value 7.415%, EWA_SVR value 6.706% and EPE_SVR value 9.507%. Similarly, the MLP prediction model while for EAP_MLP and EWA_MLP outputs relatively consistent values, 6.489% and 6.612% respectively, with EPE_MLP outputs 8.113%. All that mentioned, this performance can be attributed to the way that intra-model specific processes, for observing errors.

One more observation is related with the ability to accurately forecast the peak load timesteps. It becomes more noticeable on plots supplied in Appendices O-Q, where most of the methods fail to accurately forecast, with MLP and LR having the most optimistic results.

9.1.3 A Tri-layer Optimization Framework for Day-Ahead Energy Scheduling based on Cost and Discomfort minimization

This study envisions a DSS framework that optimizes the day-ahead energy scheduling for aggregators and consumers. It aims to produce material for reference for both academic and industry, focusing on the concepts of multi- objective optimization, applied to Demand-Side Management material, such as, DR, flexibility, energy load forecasting and discomfort. For each actor an optimization architectural layer describes and solves an individualized problem while their interaction is managed by a third architectural DSS layer.

The importance of effective energy management is highlighted as well as the benefits it yields while envisioning a cooperative energy management between the aggregator and consumer. Any improvements in this type of cooperation, generates new prospects for efficient energy monitoring and management for modern energy systems. For presenting and validating findings two distinct scenarios are considered, one referring to consumer optimization and one to aggregator portfolio optimization. The first scenario models cost minimization offering acceptable options without violating the comfort of the consumers and addresses the problem with a large-scale nonlinear solver. The second scenario enables portfolio cost minimization for the aggregators and solves the problem using mixed integer linear programming.

Both problem solutions utilize a common framework for communicating the results envisioning a DR scheme that improves prospects for their collaboration. An inclusive collaborative schema has been implemented enabling both the aggregator and the consumer to engage in DR events, yet retaining their autonomy, especially the consumer, following a human-centric approach towards the end-user, that is the consumer him/herself. The results show significant gains in cost savings for both the aggregator and the consumer.

Chapter 10: Conclusions

This chapter discusses conclusions of this thesis for both PART I and PART II and elaborates on recommendations for future work. Sect. 10.1 summarizes accomplishments, limitations and implications narrating on employed research tasks showcasing the final theory of the overall thesis research design. Sect. 10.2 organizes future directions according to each domain of studies, Social Media and Energy.

10.1 DISCUSSION

This section offers a high-level discussion regarding accomplishments of this thesis. That is, an interdisciplinary approach for data analytics, exposing various data mining and machine learning capabilities for enhanced knowledge acquisition. Novel research tasks deal with Social Media and Energy domains, forming two distinct theoretical frameworks (one for each domain) that utilize mutual options regarding incorporated methods/algorithms. The presented novel research tasks deal with the topics of SM Types, SM Topic Extraction, SM Sentiment Analysis and Energy Balancing, Energy Load Forecasting, Energy Optimal Day-Ahead Scheduling, respectively.

This thesis assesses the evolution of SM, envisions, and evaluates a novel hypothesis-based data driven methodology for analysing SMPs and proposing a new taxonomy on SMTs (Koukaras, Tjortjis and Rousidis, 2020). It establishes the hypothesis that the number of SMTs is smaller than what literature suggests while making observations on SM usage and features for experimenting on association rules and clustering algorithms. Features were grouped under utilities forcing an adjective comprehension. The utilized SM were chosen based on user penetration and available features. Since some SM implement fewer features than others, such an analysis might be impaired by this disparity.

Frequent itemset extraction produced generalized rules for forming new SMTs, yet with relatively high confidence, but rather low support. A simple grouping was implemented introducing ambiguity to the results, due to the subjectivity of the process of feature grouping. Resulting SMTs were further reduced by removing dominant utilities. As a trial and error extra analytical step, it was assumed that by

removing one by one the three most frequent utilities, the results might be improved. When this process ended, three SMTs were proposed, namely Social, Entertainment and Profiling networks, envisioned to capture emerging SMP services. Defining SMTs effectively can aid researchers, SM users and professionals by informing about SM selection according to offered functionalities, new SM trends and collaborations and acquisitions. There are many features and characteristics for each SM, with even more feature overlaps, due to similarities. Users may be confused. By considering personalized content and experiences, such a SM selection can enhance activities like public communication and social connectivity. A novel SMTs' taxonomy can facilitate the identification of new trends/features in the future, by incorporating more functional, well-structured and up-to-date SM that marketers and researchers could use. Finally, SM ecosystem is characterized by continuous collaborations and acquisitions between SM. This work classifies and suggests new hybrid categories, considering that complementary features can be merged or collaborate. Similarly, a popular SM can acquire a distinct feature from another SM.

This thesis utilizes data from Twitter to perform ARM extracting knowledge about public attitudes. It exploits the COVID-19 pandemic as a use case. The proposed methodology integrates topic extraction and visualization techniques, such as wordclouds, to form clusters or themes of opinions. It identifies frequent wordsets and generates rules that infer to user attitudes. Only strong wordsets are stored after discarding trivia ones. Frequent wordset identification is also employed attempting to reduce the number of extracted topics. Data were retrieved just from one SM (Twitter). From 27/2/2020 until 28/8/2020, 2.146.243 unique tweets were collected in a worldwide scale. Inevitably this dataset (at some point) generates limitations for a more robust analysis. Although multiple data preprocessing steps were implemented, it is almost impossible to completely clean the text. It is not possible to ascertain synonyms of strong words that are used in slang or jargon or they are just misspellings.

Also, there are always missing data that may cause faulty representation of results. The most frequent wordsets were extracted attempting to form new topics, yet with the higher possible leverage and support. Ideally, the strongest rules should be identified (high confidence and support) while supplementary considering leverage and lift. For that reason, the resulting grouping of wordsets to topics may be ambiguous, due to the general subjectivity that this methodology introduces. Findings

showcase that 50 initially retrieved topics are narrowed down to just a few ones, using methods such as LDA along with ARM. This thesis exposes implications by enhancing the extraction of insights regarding SM user opinions, attitudes, and discussions. These insights could be implemented in a plethora of other topics (not just a pandemic) due to the lack of a predefined ontology or vocabulary.

The proposed methodology could be expanded to additional SMPs rather than Twitter, generating new possibilities for a holistic compare and contrast analysis. During the COVID-19 pandemic, unstructured and unfiltered messages flushed SM, presenting opinions and ideas often resulting in negative outcomes or actions. A mechanism being able to identify, review and categorize these data streams under generic topics (yielding a more precise overview of discussions and topics) may be very useful for a variety of stakeholders (governments, health experts etc.). On the other hand, SM may also provide valuable medical relevant content. Involved parties can take advantage of a more accurate topic extraction method and engage with the SM public opinion. Then if and when deemed necessary an intervention can take place publishing content/information/guidelines evaluated by experts.

Researchers and practitioners attempt to extract knowledge by evaluating public sentiment in response to global events focusing on microblog data. Tracking and alerting the public in case of a pandemic can exploit valuable insights for reducing the negative impact in terms of human lives, societies, economies etc.

This study aims to evaluate public attitudes towards the COVID-19 spread, by performing sentiment analysis on tweets in English. The tweets were globally gathered for a period of seven months (February to August 2020). No geolocation restrictions were applied when crawling the data, forming a dataset that contains tweets from all over the world. Findings are subject to limited generalization since data from only one SMP (Twitter). It was not possible to retrieve all the available data related with COVID-19 available for the examined period. There were minor technical issues related with the server uptime that facilitated the crawling script. Although, the dataset accumulated around 14k English tweets per investigated day. The sentiment polarization process was not enhanced; but the overall preprocessing was improved, rendering the output of the sentiment polarization, subject to disputes. The hypotheses generated, and their validation process (Pearson correlations and p-values) contain certain arbitrary concepts that introduce subjectivity to the validation of results. This

thesis investigates whether there is a correlation between public sentiment and the number of cases and deaths attributed to COVID-19. Findings highlight a correlation of sentiment polarity and deaths, starting 41 days and expanding up to three days prior to the count. A strong correlation is also detected, between COVID-19 twitter conversation polarity and reported cases, but a weak correlation between polarity and reported deaths.

Knowledge extraction from online posts may be utilized for discovering correlations between attitudes of masses in times of need. Sentiment analysis calculates how positive or negative is the online text content. In case of a pandemic (COVID-19), tools that enable timely tracking and alerting of the public may become very useful for tackling issues (economic, social etc.). This thesis envisions such functionalities exposing valuable insights for predicting disease outbreaks through monitoring and evaluation of multivariable correlations (COVID-19 posts' sentiment polarity /cases/deaths). Exploitation parties involve healthcare/medical professionals, research community or governments may retrieve useful indicators for psychological correlations resulting from SM. Also, SM posts offer new capabilities for retrieving or publishing information with health context to the communities, enhancing online presence of public health policy makers (e.g. governments). Therefore, implications of this part of this thesis, envision a system that makes decisions about a worldwide healthcare/medical crisis utilizing online content. Numerous extra parameters could be retrieved such as population characteristics (e.g. age, gender), indexes (e.g. economic, regional), vaccination programs, government policies (e.g. lockdowns) and more, along with posts' sentiment analysis from multiple SM.

A way to improve energy management is to perform balancing both at the P2P level (prosumer to prosumer) and then at the VMG2VMG level, while considering the intermittency of available RES and especially PV. This thesis proposes an interdisciplinary analytics-based approach for the formation of VMGs addressing energy balancing (Koukaras, Tjortjis, *et al.*, 2021). Such an approach incorporates Computer Science and ML methods to address an Energy sector problem, utilizing data preprocessing techniques. Such an approach offers options for generating VMGs of prosumers through clustering and binning, prior to utilizing EBA that performs VMG2VMG balancing. Limitations can be attributed to the fact that, this research task, poses as a Computer Science approach attempting to investigate an Energy sector

problem setting the requirements for an interdisciplinary approach. Also, certain levels of domain details are disregarded e.g. power transmission, heating or PV production losses. By observing the outputs after running the VMG balancing process multiple times, every attempt remains biased by the nature of RES (PV). A high production is expected during peak hours (sunlight), and zero production after (nightfall).

Implications involve a few identified use cases in case this tool can be integrated as an external standalone component/toolkit. i) Functional cost minimization or profit maximization when transferring energy from VMGs handled by aggregators to the virtual portfolio of the DSO enhancing DER load distribution. These grids are conceptualized as clusters. ii) New possibilities that arise from VMG2VMG transfer of energy, when large scale RES utilization is feasible. VMGs enable P2P energy transfer, offering optimal energy distribution in close proximity, maximizing savings from main grid tariffs and reducing green gas emissions. The impartiality and stability of such an energy management system may be enforced through local ecosystem constraints. iii) VMGs allow energy trading to cover DR needs. DR energy transfer (uni/bilaterally) on P2P level enables prosumers to participate to energy-sharing schemes or DR programs.

This thesis presents an accurate approach for ELF (Koukaras, Bezas, *et al.*, 2021), investigating improvements in load forecasting capabilities for residential house energy requirements. It utilizes historical data from a state-of-the-art nZEB smart home, performing multiple tests for improved ELF. It focuses on the aspects of one step ahead ELF, while aiming at presenting an approach that can be utilized regardless of the data time resolution (15-min, 30-min, one-hour etc.) yielding results with high accuracy.

The proposed methodology integrates multiple separate models, nullifying any chances for dependent modelling between consequent prediction steps while focusing on STLF. Therefore, for predicting the hourly energy load for the next day, 24 independent models were created for each hour. The utilized metrics are RMSE, MAPE and SMAPE based on their high usability by the academic and practitioner community. It is acknowledged that there are numerous other metrics that can be incorporated, possibly mitigating any unidentified inconsistencies, or tampering on the final evaluation of results. Also, there are many other ensemble methods, prediction algorithms and strategies for timeseries forecasting. Although, the ones incorporated

in this study are considered to be characteristic and educative enough options as well as exemplar timeseries predictors. More specifically, experimentation implements five forecasting models (MLP, LSTM, XGBOOST, SVR and LR) combined with three fine-tuned ensemble forecasting models (EAP, EWA and EPE) combined into one proposed single step forecasting strategy. The evaluation of forecasting results is performed utilizing popular accuracy metrics (MAPE, SMAPE and RMSE) and an Execution Time (ET) metric. The experimentation is executed under the same characteristics regarding forecasting horizons, model configurations, metrics, and objectives. The findings utilizing pilot data and the methodology's exploitation, envision to constitute to a work that can be utilized by researchers and practitioners, as a point of reference regarding ELF with multi-domain applications. For example, Energy sector stakeholders (e.g., DSO and aggregators) can take advantage of better forecasting methods yielding improved ELF precision that may enable enhanced planning of DR management strategies.

This thesis proposes a framework for a multi-objective analysis, acting as a novel tool that evaluates new possibilities for optimal energy management integrating a DSS (Koukaras, Gkaidatzis, *et al.*, 2021). Two distinct optimization problems are considered one for consumers and one for aggregators. Each solution may completely or partly interact with the other in the form of DR signal exchange. The overall optimization is formulated by a bi-objective optimization problem for the consumer yielding cost minimization and discomfort reduction; and a single objective optimization problem for the aggregator yielding portfolio cost minimization. Three architectural layers are conceived, namely the consumer, aggregator and DSS, forming a tri-layer optimization framework.

During the experimentation process constraints such as power transmission loss, heating losses, inventors, distances etc. were not considered. That is because data were retrieved from API pilots (part of Terni Distribution System portfolio and ITI/CERTH nZEB) reporting on final values. For incorporating the proposed DR scheme only two system stakeholders were considered, the consumer and the aggregator. For a more holistic approach, DSO should also be included, expanding experimentation on even more complex interactions. In that case, this framework will be able to identify, test and evaluate a wider range of requirements for a modern energy distribution network. An analytical comparative analysis on possible solvers was not performed (e.g.

compare with metaheuristics and genetic algorithms). Instead, large-scale nonlinear optimization and a mixed integer programming solver were used for outputting results for the two actors (consumer and aggregator respectively). The main goal is to achieve optimal management of energy resources considering both actors' preferences and goals. The integration of RES, storage and more related parameters may offer enhanced energy grid elasticity, since there are more options for handling issues related with peak loads, broken power grid links etc. Solving each problem in a stand-alone manner allows interactions between consumers (P2P level) while allowing the DSO to validate DR programs for the aggregator.

Nowadays, aggregators have to manage very large portfolios showcasing the importance for more efficient energy load scheduling. A modern and efficient framework should also consider consumer preferences enabling dynamic asset management through a reliable DR signal confirmation scheme. The DSO should be able to retrieve that information from the aggregator and approve the DR portfolio schedule while enforcing and validating a reliable grid functionality. DSO should be able to identify power distribution network malfunctions and notify the aggregator in order to adjust day-ahead energy load and flexibility requests. This part of the thesis considers multiple energy assets and actors that may comprise energy ecosystems. The proposed approach yields prospects for assisting or becoming a point of reference for both the industry and the academia, when referring to DERs, P2P energy transfer models and energy multi-objective optimization and scheduling.

The research accomplishments of this thesis are communicated through the conception of two theoretical frameworks reported in Chapter 5: i) "A Multi-Functional Framework for defining Social Media Types, extracting Topics and Inferences, and discovering Correlations based on Public Sentiment" and ii) "A Novel Framework for P2P and VMG2VMG Energy Balancing, Incorporating One Step Ahead Load Forecasting and Optimization for Day-Ahead Energy Scheduling". These theoretical frameworks, integrate common data mining methods/algorithms envisioning research methodologies for interdisciplinary approaches enabling enhanced knowledge acquisition in data analytics.

10.2 FUTURE DIRECTIONS

This section outlines future research directions of this thesis based on the research achievements per domain of studies, Social Media and Energy.

10.2.1 Social Media

There are future plans to elaborate more on SMTs, by continuing to monitor their evolution. It is likely to observe more aggressive merges of SMPs soon, forcing updates on the proposed taxonomy. The next step is to improve the methodology to better handle biases (Sect. 3.4.1) in order to improve the quality of the research output by performing an empirical study on the understanding the usage of each SM from the user perspective.

Furthermore, the methodology could be automated in a way that even when new SM become popular, new features are added or biased data entries persist, SM allocation on a SMT should be effectively adjusted. This way it should become feasible to track future changes in SM when new features are added. As mentioned in (Engelbrecht, Gerlach and Widjaja, 2016), SM are under a rapid evolution, growth and metamorphosis. Scientists around the world have started using online tools and various technologies dedicated to SM, but the adoption and acceptance is still poor across the wider research community. This work could help academics and practitioners to keep track of the evolution on SMTs by having a point of reference regarding the essence of SM usage. For example, which list of SM should refer to, when it is required to research on market trends, which one for people's discussions, which one for entertainment purposes, and so on.

In the future, the conducted research aims at further mitigating the biases and limitations regarding SMTE (Sect. 3.4.2). In addition:

i) This thesis also envisions the developed methodology as a part of a complete Decision Support System (DSS). This will engage in predictive and prescriptive analytics (Koukaras and Tjortjis, 2019) utilizing SM historical data to forecast public attitudes/sentiment regarding healthcare issues. Implications of such a DSS involve policy makers when taking actions for mitigating issues arising during a worldwide crisis.

ii) Expand on algorithmic improvements regarding ARM options (Ghafari and Tjortjis, 2019). One of the basic concerns of the methodology refers to decisions

made regarding the appropriate values of minimum of support and minimum confidence for finding the most frequent item sets and extracting the rules, respectively. Future investigation should elaborate on possible options that effectively perform this process in an automated way, without manually setting the support and confidence levels (Ghafari and Tjortjis, 2016).

iii) Create an algorithm that automatically classifies input tweets to the generated topics resulting from this research. That way topic extraction will have a dynamic extraction feature, enabling real-time monitoring and classification of SM data input streams.

Based on findings from task SMSA, a software or tool that offers text sentiment analysis functionalities that validates the proposed hypotheses related with a health crisis, might be of great help for the communities. For example, healthcare/medical professionals, research community or governments could extract indicators about existent psychological correlations resulting from SM data. SM offer numerous opportunities for extracting or communicating public health information; therefore, such proactive opportunities should be exploited for improving public health policy makers' online presence.

According to these points, this task should be extended as follows:

i) Perform Association Rule Mining (ARM) on tweets for discovering most inferenced topics/discussions and visualize results.

ii) Compare the outputs of the current sentiment analysis approach with contradicting outputs from other sentiment analysers, such as an approach utilizing Neo4j/graphs and SentiStrength.

iii) Conduct forecasting analysis using ensemble or a multivariate approach utilizing the SM dataset for predicting the course of the virus spread in relation with tweets polarization.

Expand on implications, findings and i), ii) and iii) while envisioning the generation of a decision support system. This system takes into consideration SM data to decide about Healthcare/Medical related problems (Koukaras, Rousidis and Tjortjis, 2020), such as epidemics. It combines more than 30 different parameters like population characteristics (gender, life expectancy, age distribution and more), indexes (economic, medical and more), healthcare dynamics like ICUs, medical staff per

capita, vaccination programs. Also, it considers various government policies applied (social distancing, lockdown and more), interactions like tourism and air-travel connections, isolation from other countries and sentiment analysis on COVID-19 data retrieved from Twitter, but also from other SM platforms.

10.2.2 Energy

Conducted research on Energy balancing can be extended by addressing the following points:

i) Continue monitoring of the evolution of research novelties on RES energy distribution balancing.

ii) Improve the methodology by tackling biases presented in Sect. 7.4.1 and enhancing the perception of balancing in a more energy-centric manner.

iii) Extend ongoing work on distinct energy profile generation and update the dataset entries to form lists of P2P energy transfer, aiming to generate closed communities within VMGs. This approach would aid further improvement of energy balancing and allow further simulations with more datasets in order to calibrate the approach.

iv) Improve the model by further automating the proposed methodology to act as a stand-alone component or toolkit. An extended reference to such tools was made in Sect. 6.1.2 and specifically to attempts incorporating Blockchain technology, as in (Cioara *et al.*, 2018; Pop *et al.*, 2018; Droriano *et al.*, 2019). Blockchain can offer the infrastructure for a secure (Tsolakis *et al.*, 2018) and distributed energy balancing ecosystem (Pop *et al.*, 2020) through smart contracts. This architectural approach can thus be complemented at the application level (Pop *et al.*, 2019), whilst exposing Business Intelligence opportunities (Mendling *et al.*, 2017).

v) Regarding DR programs, this study conceptualizes that the proposed approach can be expanded to use the Time of Use (ToU) category, or any other DR program described in Sect. 6.1.1. It was initially considered utilizing ToU, since currently the approach solely relies on PV energy production for VMG energy output. This seems more appropriate as the dataset contains data that represent prosumer energy generation originated from RES, such as PV. Considering these points, a minimization on operating costs and pollution emissions in an individualized way (per VMG), or in a super-cluster of multiple VMGs can be achieved. Having presented future prospects of this work, it is envisioned assisting academics and

practitioners alike when referring to DERs, VMG2VMG, P2P energy transfer models and energy balancing utilizing clustering, and exhaustive balancing techniques merged with distributed energy transactions using technologies such as Blockchain i.e. (Conejo *et al.*, 2005; Stecchi *et al.*, 2019).

In addition, reference is made to the identified possible improvements and future work of the proposed OSA-ELF approach for STLF. Thus, the following points are enumerated:

i) It is in future plans to mitigate or tackle the biases as they were presented in the corresponding section (Sect. 7.4.2), aiming at further improving STLF accuracy.

ii) Expand testing of the proposed approach, by implementing a broader evaluation involving bigger datasets with energy load data entries from multiple distinct sources (e.g., a cluster of industrial or residential building).

iii) Implement the proposed OSA-ELF approach as a standalone toolkit enabling the integration with other tools for improving the DRM strategies in Microgrids while envisioning accuracy improvements in power management systems.

iv) Expand the proposed single-step experimentation domains by modifying its current conception for testing it for other ELF horizons, such as VSTLF, MTLF or LTLF.

v) The proposed approach predicts a numerical feature: the energy load in hourly resolution. Discretizing this feature to produce class labels lends itself to an interesting classification problem.

The proposed framework of task EODS conceptualizes a DSS tool for energy stakeholders based on an optimization engine. It envisions multi-level, multi-factor and multi-objective problem modelling for solving practical problems in the energy sector. This study could be improved according to the following points.

i) Continue monitoring the evolution of research novelties on the multi-objective optimization on the energy sector. Enforce improvements in proposed methodology by tackling limitations as explained in Sect. 7.4.3 and by enhancing the perception on the objective functions.

ii) Improve the proposed framework by adding more automation to the methodology in a way that it can act as a stand-alone software, which can be utilized given just the appropriate input of datasets.

iii) Add one more architectural layer, that is the DSO layer, leading to a tri-level optimization approach. The DSO layer is to be conceived as tri-objective optimization problem optimizing considering three objective functions simultaneously. Minimization of grid energy losses, voltage profile improvement and cost reduction of environmental emissions.

iv) Implement the proposed DR scheme in a more sophisticated way, including technologies such as Blockchain (Lu and Hong, 2019). This would enhance the business perspective of the proposed framework while addressing more practical applications of this study.

Bibliography

Abadi, M. *et al.* (2016) 'Tensorflow: A system for large-scale machine learning', in *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pp. 265–283.

Abd-Alrazaq, A. *et al.* (2020) 'Top concerns of tweeters during the COVID-19 pandemic: A surveillance study', *Journal of Medical Internet Research*, 22(4), p. e19016. doi: 10.2196/19016.

Adamopoulou, A. A., Tryferidis, A. M. and Tzovaras, D. K. (2016) 'A context-aware method for building occupancy prediction', *Energy and Buildings*, 110, pp. 229–244. doi: 10.1016/j.enbuild.2015.10.003.

Aggarwal, C. C. and Zhai, C. X. (2012) 'An introduction to text mining', in *Mining text data*. Springer, pp. 1–10. doi: 10.1007/978-1-4614-3223-4_1.

Agrawal, R. and Srikant, R. (2013) 'Fast Algorithms For Mining Association Rules In Datamining', in *International Journal of Scientific & Technology Research*. Citeseer, pp. 13–24.

Aiello, L. M. *et al.* (2013) 'Sensing trending topics in twitter', *IEEE Transactions on Multimedia*, 15(6), pp. 1268–1282. doi: 10.1109/TMM.2013.2265080.

Akbary, P. *et al.* (2019) 'Extracting Appropriate Nodal Marginal Prices for All Types of Committed Reserve', *Computational Economics*, 53(1), pp. 1–26. doi: 10.1007/s10614-017-9716-2.

Alvarez-Benitez, J. E., Everson, R. M. and Fieldsend, J. E. (2005) 'A MOPSO algorithm based exclusively on pareto dominance concepts', in *Lecture Notes in Computer Science*, pp. 459–473. doi: 10.1007/978-3-540-31880-4_32.

Alvarez-Melis, D. and Saveski, M. (2016) 'Topic modeling in Twitter: Aggregating tweets by conversations', in *Proceedings of the 10th International Conference on Web and Social Media, ICWSM 2016*, pp. 519–522.

Amarasinghe, K., Marino, D. L. and Manic, M. (2017) 'Deep neural networks for energy load forecasting', in *IEEE International Symposium on Industrial Electronics*. IEEE, pp. 1483–1488. doi: 10.1109/ISIE.2017.8001465.

Amjady, N. and Daraeepour, A. (2011) 'Midterm demand prediction of electrical power systems using a new hybrid forecast technique', *IEEE Transactions on Power Systems*, 26(2),

pp. 755–765. doi: 10.1109/TPWRS.2010.2055902.

Anoh, K. *et al.* (2020) 'Energy Peer-to-Peer Trading in Virtual Microgrids in Smart Grids: A Game-Theoretic Approach', *IEEE Transactions on Smart Grid*, 11(2), pp. 1264–1275. doi: 10.1109/TSG.2019.2934830.

Arthur, D. and Vassilvitskii, S. (2007) *K-means++: The advantages of careful seeding*, *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms*. Stanford.

Asmus, P. (2010) 'Microgrids, Virtual Power Plants and Our Distributed Energy Future', *Electricity Journal*, 23(10), pp. 72–82. doi: 10.1016/j.tej.2010.11.001.

Asur, S. *et al.* (2012) 'Trends in Social Media: Persistence and Decay', in *SSRN Electronic Journal*. doi: 10.2139/ssrn.1755748.

Awad, M. and Khanna, R. (2015) 'Support vector regression', in *Efficient learning machines*. Springer, pp. 67–80.

Badami, M. *et al.* (2019) 'A decision support system tool to manage the flexibility in renewable energy-based power systems', *Energies*, 13(1), p. 153. doi: 10.3390/en13010153.

Balijepalli, V. S. K. M. *et al.* (2011) 'Review of demand response under smart grid paradigm', in *2011 IEEE PES International Conference on Innovative Smart Grid Technologies-India, ISGT India 2011*. IEEE, pp. 236–243. doi: 10.1109/ISET-India.2011.6145388.

Barkur, G. and Vibha, G. B. K. (2020) 'Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from India', *Asian journal of psychiatry*, 51, p. 102089.

Beleveslis, D. *et al.* (2019) 'A hybrid method for sentiment analysis of election related tweets', in *2019 4th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM)*. IEEE, pp. 1–6.

Bendaoud, N. M. M. and Farah, N. (2020) 'Using deep learning for short-term load forecasting', *Neural Computing and Applications*, 32(18), pp. 15029–15041. doi: 10.1007/s00521-020-04856-0.

Berrios, R., Totterdell, P. and Kellett, S. (2015) 'Eliciting mixed emotions: A meta-analysis comparing models, types and measures.', *Frontiers in Psychology*, 6(MAR). doi: 10.3389/fpsyg.2015.00428.

Bintoudi, A. D. *et al.* (2018) 'Novel hybrid design for microgrid control', in *Asia-Pacific Power and Energy Engineering Conference, APPEEC*. IEEE, pp. 1–6. doi: 10.1109/APPEEC.2017.8308958.

Bishop, C. M. (2007) 'Pattern recognition and machine learning', *Choice Reviews Online*, 44(09), pp. 44-5091-44–5091. doi: 10.5860/choice.44-5091.

Blake, S. T. and O'Sullivan, D. T. J. (2018) 'Optimization of Distributed Energy Resources in an Industrial Microgrid', *Procedia CIRP*, 67, pp. 104–109. doi: 10.1016/j.procir.2017.12.184.

Blei, D. M., Ng, A. Y. and Jordan, M. T. (2002) 'Latent dirichlet allocation', *Advances in Neural Information Processing Systems*, 3, pp. 993–1022.

Botchkarev, A. (2018) 'Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology', *arXiv preprint arXiv:1809.03006*. Available at: http://arxiv.org/abs/1809.03006.

Boyd, D. M. and Ellison, N. B. (2007) 'Social network sites: Definition, history, and scholarship', *Journal of Computer-Mediated Communication*, 13(1), pp. 210–230. doi: 10.1111/j.1083-6101.2007.00393.x.

Bragatto, T. *et al.* (2019) 'A real-life experience on 2nd life batteries services for Distribution System Operator', in *Proceedings - 2019 IEEE International Conference on Environment and Electrical Engineering and 2019 IEEE Industrial and Commercial Power Systems Europe, EEEIC/I and CPS Europe 2019*. IEEE, pp. 1–6. doi: 10.1109/EEEIC.2019.8783963.

Brin, S. *et al.* (1997) 'Dynamic Itemset Counting and Implication Rules for Market Basket Data', in *SIGMOD Record (ACM Special Interest Group on Management of Data)*, pp. 255–264. doi: 10.1145/253262.253325.

Bunnoon, P., Chalermyanont, K. and Limsakul, C. (2009) 'Mid term load forecasting of the country using statistical methodology: Case study in thailand', in *2009 International Conference on Signal Processing Systems, ICSPS 2009*. IEEE, pp. 924–928. doi: 10.1109/ICSPS.2009.174.

Cagnano, A., De Tuglie, E. and Mancarella, P. (2020) 'Microgrids: Overview and guidelines for practical implementations and operation', *Applied Energy*, 258, p. 114039. doi: 10.1016/j.apenergy.2019.114039.

Cardoso, G. *et al.* (2018) 'Battery aging in multi-energy microgrid design using mixed integer linear programming', *Applied Energy*, 231, pp. 1059–1069. doi: 10.1016/j.apenergy.2018.09.185.

Casado-Mansilla, Di. *et al.* (2018) 'A Human-Centric Context-Aware IoT Framework for Enhancing Energy Efficiency in Buildings of Public Use', *IEEE Access*, 6, pp. 31444–31456. doi: 10.1109/ACCESS.2018.2837141.

Cataldi, M., Di Caro, L. and Schifanella, C. (2010) 'Emerging topic detection on Twitter based on temporal and social terms evaluation', in *Proceedings of the 10th International Workshop on Multimedia Data Mining, MDMKDD '10*, pp. 1–10. doi: 10.1145/1814245.1814249.

Chaffey, D. (2021) *Global social media statistics research summary [updated 2021]*, *Smart Insights*. Available at: https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/ (Accessed: 31 May 2021).

Charalambous, A. *et al.* (2019) 'Phase balancing and reactive power support services for microgrids', *Applied Sciences (Switzerland)*, 9(23), p. 5067. doi: 10.3390/app9235067.

Chen, T. and Guestrin, C. (2016) 'XGBoost: A scalable tree boosting system', in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. doi: 10.1145/2939672.2939785.

Choi, H. J. and Park, C. H. (2019) 'Emerging topic detection in twitter stream based on high utility pattern mining', *Expert Systems with Applications*, 115, pp. 27–36. doi: 10.1016/j.eswa.2018.07.051.

Choi, S. S., Cha, S. H. and Tappert, C. C. (2009) 'A survey of binary similarity and distance measures', *WMSCI 2009 - The 13th World Multi-Conference on Systemics, Cybernetics and Informatics, Jointly with the 15th International Conference on Information Systems Analysis and Synthesis, ISAS 2009 - Proc.*, 3(1), pp. 80–85.

Chollet, F. (2015) 'others.(2015). Keras. GitHub'.

Ciftci, O. *et al.* (2019) 'Chance-constrained microgrid energy management with flexibility constraints provided by battery storage', in *2019 IEEE Texas Power and Energy Conference, TPEC 2019*, pp. 1–6. doi: 10.1109/TPEC.2019.8662200.

Cinelli, M. *et al.* (2020) 'The covid-19 social media infodemic', *Scientific Reports*, 10(1), pp. 1–10.

Cioara, T. *et al.* (2018) 'Enabling New Technologies for Demand Response Decentralized Validation Using Blockchain', in *Proceedings - 2018 IEEE International Conference on Environment and Electrical Engineering and 2018 IEEE Industrial and Commercial Power Systems Europe, EEEIC/I and CPS Europe 2018*. IEEE, pp. 1–4. doi: 10.1109/EEEIC.2018.8493665.

Conejo, A. J. *et al.* (2005) 'Forecasting electricity prices for a day-ahead pool-based electric energy market', *International Journal of Forecasting*, 21(3), pp. 435–462. doi: 10.1016/j.ijforecast.2004.12.005.

Copp, E. (2018) *Top 5 Social Media Trends in 2019 (And How Brands Should Adapt)*, *Hootsuite*. Available at: https://blog.hootsuite.com/social-media-trends/ (Accessed: 4 December 2018).

Cordeiro, M. (2012) 'Twitter event detection: combining wavelet analysis and topic inference

summarization', in *the Doctoral Symposium on Informatics Engineering - DSIE'12*, pp. 11–16. Available at: http://paginas.fe.up.pt/~prodei/dsie12/papers/paper_14.pdf.

Cui, Y. *et al.* (2017) 'Review: Multi-objective optimization methods and application in energy saving', *Energy*, 125, pp. 681–704. doi: 10.1016/j.energy.2017.02.174.

Daneshi, H., Shahidehpour, M. and Choobbari, A. L. (2008) 'Long-term load forecasting in electricity market', in *2008 IEEE International Conference on Electro/Information Technology, IEEE EIT 2008 Conference*. IEEE, pp. 395–400. doi: 10.1109/EIT.2008.4554335.

Daneshvar, M. *et al.* (2020) 'Chance-constrained models for transactive energy management of interconnected microgrid clusters', *Journal of Cleaner Production*, 271, p. 122177.

Dikaiakos, M. D. *et al.* (2009) 'Cloud computing: Distributed internet computing for IT and scientific research', *IEEE Internet Computing*, 13(5), pp. 10–11. doi: 10.1109/MIC.2009.103.

Ding, Y. M., Hong, S. H. and Li, X. H. (2014) 'A demand response energy management scheme for industrial facilities in smart grid', *IEEE Transactions on Industrial Informatics*, 10(4), pp. 2257–2269. doi: 10.1109/TII.2014.2330995.

Droriano, L. *et al.* (2019) 'Decentralized blockchain flexibility system for Smart Grids: Requirements engineering and use cases', in *CANDO-EPE 2018 - Proceedings IEEE International Conference and Workshop in Obuda on Electrical and Power Engineering*. IEEE, pp. 39–44. doi: 10.1109/CANDO-EPE.2018.8601171.

Ellahi, M. *et al.* (2019) 'Recent approaches of forecasting and optimal economic dispatch to overcome intermittency of wind and photovoltaic (PV) systems: A review', *Energies*, 12(22), p. 4392. doi: 10.3390/en12224392.

Engelbrecht, A., Gerlach, J. P. and Widjaja, T. (2016) *Understanding the anatomy of data-driven business models - Towards an empirical taxonomy*, *24th European Conference on Information Systems, ECIS 2016*. Available at: http://aisel.aisnet.org/ecis2016_rphttp://aisel.aisnet.org/ecis2016_rp/128 (Accessed: 31 May 2021).

Ester, M. *et al.* (1996) 'A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise', in *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, pp. 226–231.

Evans, J. D. (1996) *Straightforward statistics for the behavioral sciences*. Thomson Brooks/Cole Publishing Co.

Farsi, B. *et al.* (2021) 'On Short-Term Load Forecasting Using Machine Learning Techniques

and a Novel Parallel Deep LSTM-CNN Approach', *IEEE Access*, 9, pp. 31191–31212. doi: 10.1109/ACCESS.2021.3060290.

Feng, S. *et al.* (2015) 'A word-emoticon mutual reinforcement ranking model for building sentiment lexicon from massive collection of microblogs', *World Wide Web*, 18(4), pp. 949–967.

Feng, S. *et al.* (2019) 'Attention based hierarchical LSTM network for context-aware microblog sentiment classification', *World Wide Web*, 22(1), pp. 59–81.

Forrest, C. (2014) *The top 7 acquisitions of all-time in social media, and why they matter - TechRepublic*. Available at: https://www.techrepublic.com/article/the-top-7-acquisitions-of-all-time-in-social-media-and-why-they-matter/ (Accessed: 1 June 2021).

Gebben, F., Bader, S. and Oelmann, B. (2015) 'Configuring artificial neural networks for the prediction of available energy in solar-powered sensor nodes', in *2015 IEEE SENSORS - Proceedings*. IEEE, pp. 1–4. doi: 10.1109/ICSENS.2015.7370253.

Ghafari, S. M. and Tjortjis, C. (2016) 'Association rules mining by improving the imperialism competitive algorithm (ARMICA)', in *IFIP Advances in Information and Communication Technology*. Springer, pp. 242–254. doi: 10.1007/978-3-319-44944-9_21.

Ghafari, S. M. and Tjortjis, C. (2019) 'A survey on association rules mining using heuristics', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4), p. e1307. doi: 10.1002/widm.1307.

Goodwin, D. (2020) *10 Important 2020 Social Media Trends You Need to Know*. Available at: https://www.searchenginejournal.com/2020-social-media-trends/342851/#close (Accessed: 1 June 2021).

Greene, D., O'Callaghan, D. and Cunningham, P. (2014) 'How many topics? Stability analysis for topic models', in *Joint European conf. on machine learning and knowledge discovery in databases*. Springer, pp. 498–513. doi: 10.1007/978-3-662-44848-9_32.

Greff, K. *et al.* (2017) 'LSTM: A Search Space Odyssey', *IEEE Transactions on Neural Networks and Learning Systems*, 28(10), pp. 2222–2232. doi: 10.1109/TNNLS.2016.2582924.

Grimes, D. *et al.* (2014) 'Analyzing the impact of electricity price forecasting on energy cost-aware scheduling', *Sustainable Computing: Informatics and Systems*, 4(4), pp. 276–291. doi: 10.1016/j.suscom.2014.08.009.

Grmanová, G. *et al.* (2016) 'Incremental ensemble learning for electricity load forecasting', *Acta Polytechnica Hungarica*, 13(2), pp. 97–117. doi: 10.12700/aph.13.2.2016.2.6.

Guan, C. *et al.* (2013) 'Very short-term load forecasting: Wavelet neural networks with data

pre-filtering', *IEEE Transactions on Power Systems*, 28(1), pp. 30–41. doi: 10.1109/TPWRS.2012.2197639.

Gundecha, P. and Liu, H. (2012) 'Mining Social Media: A Brief Introduction', *2012 TutORials in Operations Research*, pp. 1–17. doi: 10.1287/educ.1120.0105.

Haerder, T. and Reuter, A. (1983) 'Principles of transaction-oriented database recovery', *ACM Computing Surveys (CSUR)*, 15(4), pp. 287–317. doi: 10.1145/289.291.

Hahsler, M., Grün, B. and Hornik, K. (2005) 'Arules - A computational environment for mining association rules and frequent item sets', *Journal of Statistical Software*, 14(15), pp. 1–25. doi: 10.18637/jss.v014.i15.

Hamzah, F. B. *et al.* (2020) 'CoronaTracker: worldwide COVID-19 outbreak data analysis and prediction', *Bull World Health Organ*, 1(32).

Han, J., Kamber, M. and Pei, J. (2012) 'Data Mining: Concepts and Techniques', *Data Mining: Concepts and Techniques*, 5(4), pp. 83–124. doi: 10.1016/C2009-0-61819-5.

Han, J., Pei, J. and Yin, Y. (2000) 'Mining frequent patterns without candidate generation', *ACM sigmod record*, 29(2), pp. 1–12.

Hawkes, A. D. and Leach, M. A. (2009) 'Modelling high level system design and unit commitment for a microgrid', *Applied Energy*, 86(7–8), pp. 1253–1265. doi: 10.1016/j.apenergy.2008.09.006.

Hipp, J., Güntzer, U. and Nakhaeizadeh, G. (2000) 'Algorithms for association rule mining — a general survey and comparison', *ACM SIGKDD Explorations Newsletter*, 2(1), pp. 58–64. doi: 10.1145/360402.360421.

Hirsch, A., Parag, Y. and Guerrero, J. (2018) 'Microgrids: A review of technologies, key drivers, and outstanding issues', *Renewable and Sustainable Energy Reviews*, 90, pp. 402–411. doi: 10.1016/j.rser.2018.03.040.

Hochreiter, S. and Schmidhuber, J. (1997) 'Long short-term memory', *Neural computation*, 9(8), pp. 1735–1780.

Van Houdt, G., Mosquera, C. and Nápoles, G. (2020) 'A review on the long short-term memory model', *Artificial Intelligence Review*, 53(8), pp. 5929–5955. doi: 10.1007/s10462-020-09838-1.

How social media traffic is adapting to COVID-19 - Marx Layne (2020). Available at: https://marxlayne.com/social-traffic-covid-19/ (Accessed: 3 June 2021).

Hsiao, Y. H. (2015) 'Household electricity demand forecast based on context information and user daily schedule analysis from meter data', *IEEE Transactions on Industrial Informatics*,

11(1), pp. 33–43. doi: 10.1109/TII.2014.2363584.

Hsu, W.-Y., Hsu, H.-H. and Tseng, V. S. (2019) ‘Discovering negative comments by sentiment analysis on web forum’, *World Wide Web*, 22(3), pp. 1297–1311.

Huang, C. *et al.* (2020) ‘Mining the characteristics of COVID-19 patients in china: Analysis of social media posts’, *Journal of Medical Internet Research*, 22(5), p. e19087. doi: 10.2196/19087.

Ioannidis, J. P. A. (2018) ‘The proposal to lower P value thresholds to. 005’, *Jama*, 319(14), pp. 1429–1430.

Islam, M. S. *et al.* (2020) ‘COVID-19–related infodemic and its impact on public health: A global social media analysis’, *The American Journal of Tropical Medicine and Hygiene*, 103(4), p. 1621.

ISO 7730 (2005) *ISO 7730: Ergonomics of the thermal environment Analytical determination and interpretation of thermal comfort using calculation of the PMV and PPD indices and local thermal comfort criteria*, *Management*. ISO (International standards). Available at: http://linkinghub.elsevier.com/retrieve/pii/S0267726105000503.

Jaccard, P. (1912) ‘the Distribution of the Flora in the Alpine Zone.’, *New Phytologist*, 11(2), pp. 37–50. doi: 10.1111/j.1469-8137.1912.tb05611.x.

Jacob, M., Neves, C. and Vukadinović Greetham, D. (2020) *Forecasting and Assessing Risk of Individual Electricity Peaks*. Springer Nature.

Jahn, J. (1985) ‘Scalarization in Multi Objective Optimization’, in *Mathematics of Multi Objective Optimization*. Springer, pp. 45–88. doi: 10.1007/978-3-7091-2822-0_3.

Jain, A. K. (2010) ‘Data clustering: 50 years beyond K-means’, *Pattern Recognition Letters*, 31(8), pp. 651–666. doi: 10.1016/j.patrec.2009.09.011.

Javaid, S. and Javaid, N. (2020) ‘Comfort evaluation of seasonally and daily used residential load in smart buildings for hottest areas via predictive mean vote method’, *Sustainable Computing: Informatics and Systems*, 25, p. 100369. doi: 10.1016/j.suscom.2019.100369.

Kafle, Y. R. *et al.* (2016) ‘Towards an internet of energy’, in *2016 IEEE International Conference on Power System Technology, POWERCON 2016*. IEEE, pp. 1–6. doi: 10.1109/POWERCON.2016.7754036.

Kanellopoulos, Y. *et al.* (2011) ‘K-attractors: A partitional clustering algorithm for numeric data analysis’, *Applied Artificial Intelligence*, 25(2), pp. 97–115. doi: 10.1080/08839514.2011.534590.

Kaplan, A. M. and Haenlein, M. (2010) ‘Users of the world, unite! The challenges and

opportunities of Social Media', *Business Horizons*, 53(1), pp. 59–68. doi: 10.1016/j.bushor.2009.09.003.

Karmellos, M. and Mavrotas, G. (2019) 'Multi-objective optimization and comparison framework for the design of Distributed Energy Systems', *Energy Conversion and Management*, 180, pp. 473–495. doi: 10.1016/j.enconman.2018.10.083.

Kaufman, L. and Rousseeuw, P. (2002) 'Statistical Data Analysis Based on the L1-Norm and Related Methods', *Statistical Data Analysis Based on the L1-Norm and Related Methods*, pp. 405–416. doi: 10.1007/978-3-0348-8201-9.

Kennedy-Shaffer, L. (2019) 'Before p< 0.05 to beyond p< 0.05: using history to contextualize p-values and significance testing', *The American Statistician*, 73(sup1), pp. 82–90.

Ketchen, D. J. and Shook, C. L. (1996) 'The application of cluster analysis in strategic management research: An analysis and critique', *Strategic Management Journal*, 17(6), pp. 441–458. doi: 10.1002/(sici)1097-0266(199606)17:6<441::aid-smj819>3.0.co;2-g.

Khairalla, M. A. *et al.* (2018) 'Short-Term Forecasting for Energy Consumption through Stacking Heterogeneous Ensemble Learning Model', *Energies*, 11(6), p. 1605. doi: 10.3390/en11061605.

Khuntia, S. R., Rueda, J. L. and van der Meijden, M. A. M. M. (2016) 'Forecasting the load of electrical power systems in mid- and long-term horizons: A review', *IET Generation, Transmission and Distribution*, 10(16), pp. 3971–3977. doi: 10.1049/iet-gtd.2016.0340.

Kietzmann, J. H. *et al.* (2011) *Social media? Get serious! Understanding the functional building blocks of social media*, *Business Horizons*. doi: 10.1016/j.bushor.2011.01.005.

Kim, E. H. J. *et al.* (2016) 'Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news', *Journal of Information Science*, 42(6), pp. 763–781. doi: 10.1177/0165551515608733.

Kim, M. *et al.* (2004) 'SPEA2+: Improving the performance of the strength pareto evolutionary algorithm 2', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 742–751. doi: 10.1007/978-3-540-30217-9_75.

Kingma, D. P. and Ba, J. L. (2015) 'Adam: A method for stochastic optimization', *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.

Kolen, J. F. and Kremer, S. C. (2010) 'Gradient Flow in Recurrent Nets: The Difficulty of Learning LongTerm Dependencies', *A Field Guide to Dynamical Recurrent Networks*. A field

guide to dynamical recurrent neural networks. IEEE Press. doi: 10.1109/9780470544037.ch14.

Kong, X. *et al.* (2020) 'Optimal operation strategy for interconnected microgrids in market environment considering uncertainty', *Applied Energy*, 275, p. 115336. doi: 10.1016/j.apenergy.2020.115336.

Korenek, P. and Šimko, M. (2014) 'Sentiment analysis on microblog utilizing appraisal theory', *World Wide Web*, 17(4), pp. 847–867. doi: 10.1007/s11280-013-0247-z.

Koukaras, P., Gkaidatzis, P., *et al.* (2021) 'A Tri-Layer Optimization Framework for Day-Ahead Energy Scheduling Based on Cost and Discomfort Minimization', *Energies 2021, Vol. 14, Page 3599*, 14(12), p. 3599. doi: 10.3390/en14123599.

Koukaras, P., Tjortjis, C., *et al.* (2021) 'An interdisciplinary approach on efficient virtual microgrid to virtual microgrid energy balancing incorporating data preprocessing techniques', *Computing*, pp. 1–42. doi: 10.1007/s00607-021-00929-7.

Koukaras, P., Bezas, N., *et al.* (2021) 'Introducing a Novel Approach in One-step Ahead Energy Load Forecasting', *Sustainable Computing: Informatics and Systems*, 32, p. 100616. doi: 10.1016/j.suscom.2021.100616.

Koukaras, P., Rousidis, D. and Tjortjis, C. (2020) 'Forecasting and prevention mechanisms using social media in health care', in *Studies in Computational Intelligence*. Springer, pp. 121–137. doi: 10.1007/978-3-662-61114-2_8.

Koukaras, P. and Tjortjis, C. (2019) 'Social Media Analytics, Types and Methodology', in *Machine Learning Paradigms*. Springer, pp. 401–427. doi: 10.1007/978-3-030-15628-2_12.

Koukaras, P., Tjortjis, C. and Rousidis, D. (2020) 'Social Media Types: introducing a data driven taxonomy', *Computing*, 102(1), pp. 295–340. doi: 10.1007/s00607-019-00739-y.

Leiner, D. J. *et al.* (2018) 'Functional domains of social media platforms: Structuring the uses of Facebook to better understand its gratifications', *Computers in Human Behavior*, 83, pp. 194–203. doi: 10.1016/j.chb.2018.01.042.

Li, C. R. J. and Deng, Z. H. (2007) 'Mining frequent ordered patterns without candidate generation', *Proceedings - Fourth International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2007*, 1(2), pp. 402–406. doi: 10.1109/FSKD.2007.402.

Li, D., Chaudhary, H. and Zhang, Z. (2020) 'Modeling spatiotemporal pattern of depressive symptoms caused by COVID-19 using social media data mining', *International Journal of Environmental Research and Public Health*, 17(14), pp. 1–23. doi: 10.3390/ijerph17144988.

Li, L. *et al.* (2021) 'Short-term apartment-level load forecasting using a modified neural

network with selected auto-regressive features', *Applied Energy*, 287, p. 116509. doi: 10.1016/j.apenergy.2021.116509.

Li, T., Qian, Z. and He, T. (2020) 'Short-Term Load Forecasting with Improved CEEMDAN and GWO-Based Multiple Kernel ELM', *Complexity*, 2020. doi: 10.1155/2020/1209547.

Li, Y., Che, J. and Yang, Y. (2018) 'Subsampled support vector regression ensemble for short term electric load forecasting', *Energy*, 164, pp. 160–170. doi: 10.1016/j.energy.2018.08.169.

Liu, N. *et al.* (2017) 'Energy-Sharing Model with Price-Based Demand Response for Microgrids of Peer-to-Peer Prosumers', *IEEE Transactions on Power Systems*, 32(5), pp. 3569–3583. doi: 10.1109/TPWRS.2017.2649558.

Liu, W. *et al.* (2018) 'Game theoretic non-cooperative distributed coordination control for multi-microgrids', *IEEE Transactions on Smart Grid*, 9(6), pp. 6986–6997. doi: 10.1109/TSG.2018.2846732.

Livingston, D. *et al.* (2018) 'Applying blockchain technology to electric power systems', *Smart Energy International*. JSTOR. Available at: https://www.smart-energy.com/industry-sectors/business-finance-regulation/applying-blockchain-technology-electric-power-systems/.

Long, C. *et al.* (2018) 'Peer-to-peer energy sharing through a two-stage aggregated battery control in a community Microgrid', *Applied energy*, 226, pp. 261–276.

Lopez, C. E., Vasu, M. and Gallemore, C. (2020) 'Understanding the perception of COVID-19 policies by mining a multilanguage Twitter dataset', *arXiv:2003.10359*. Available at: http://arxiv.org/abs/2003.10359.

Lu, R. and Hong, S. H. (2019) 'Incentive-based demand response for smart grid with reinforcement learning and deep neural network', *Applied Energy*, 236, pp. 937–949. doi: 10.1016/j.apenergy.2018.12.061.

Lu, Y. *et al.* (2015) 'Renewable energy system optimization of low/zero energy buildings using single-objective and multi-objective optimization methods', *Energy and Buildings*, 89, pp. 61–75. doi: 10.1016/j.enbuild.2014.12.032.

Lua, A. (2019) *Top Social Media Sites to Consider for Your Brand -*, *Buffer*. Available at: https://buffer.com/library/social-media-sites/ (Accessed: 30 January 2019).

Luo, L. and Chen, Y. (2020) 'Carbon emission energy management analysis of LCA-Based fabricated building construction', *Sustainable Computing: Informatics and Systems*, 27, p. 100405. doi: 10.1016/j.suscom.2020.100405.

Malekizadeh, M. *et al.* (2020) 'Short-term load forecast using ensemble neuro-fuzzy model',

Energy, 196, p. 117127. doi: 10.1016/j.energy.2020.117127.

Marino, C. *et al.* (2018) 'A chance-constrained two-stage stochastic programming model for reliable microgrid operations under power demand uncertainty', *Sustainable Energy, Grids and Networks*, 13, pp. 66–77. doi: 10.1016/j.segan.2017.12.007.

Marino, D. L., Amarasinghe, K. and Manic, M. (2016) 'Building energy load forecasting using Deep Neural Networks', in *IECON Proceedings (Industrial Electronics Conference)*. IEEE, pp. 7046–7051. doi: 10.1109/IECON.2016.7793413.

Martín-Gómez, C., Vidaurre-Arbizu, M. and Eguaras-Martínez, M. (2014) 'Sensor Placement for Bpm Analysis of Buildingsin Use to Implement Energy Savings Through Building Performance Simulation', *Journal of Engineering and Architecture*, 2(2). doi: 10.15640/jea.v2n2a10.

McKinney, W. (2010) 'Data Structures for Statistical Computing in Python', in *Proceedings of the 9th Python in Science Conference*. Austin, TX, pp. 56–61. doi: 10.25080/majora-92bf1922-00a.

Mendling, J. *et al.* (2017) 'Blockchains for Business Process Management-Challenges and Opportunities 0:1 Blockchains for Business Process Management-Challenges and Opportunities', *ACM Transactions on Management Information Systems*, 9(0), pp. 1–16. Available at: http://intelledger.github.io/.

Meng, X. *et al.* (2012) 'Entity-centric topic-oriented opinion summarization in twitter', in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 379–387. doi: 10.1145/2339530.2339592.

Menniti, D. *et al.* (2013) 'Demand response program implementation in an energy district of domestic prosumers', in *2013 Africon*. IEEE, pp. 1–7.

Moura, P. S. and de Almeida, A. T. (2010) 'Multi-objective optimization of a mixed renewable system with demand-side management', *Renewable and Sustainable Energy Reviews*, 14(5), pp. 1461–1468. doi: 10.1016/j.rser.2010.01.004.

Mukherjee, A. *et al.* (2020) 'Lightweight sustainable intelligent load forecasting platform for smart grid applications', *Sustainable Computing: Informatics and Systems*, 25, p. 100356. doi: 10.1016/j.suscom.2019.100356.

Murtagh, F. (1991) 'Multilayer perceptrons for classification and regression', *Neurocomputing*, 2(5–6), pp. 183–197. doi: 10.1016/0925-2312(91)90023-5.

Nair, V. and Hinton, G. E. (2010) 'Rectified linear units improve Restricted Boltzmann machines', in *ICML 2010 - Proceedings, 27th International Conference on Machine Learning*,

pp. 807–814.

Nickerson, R. S. (2000) ‘Null hypothesis significance testing: a review of an old and continuing controversy.’, *Psychological methods*, 5(2), p. 241.

Nojavan, S. *et al.* (2017) ‘A cost-emission model for fuel cell/PV/battery hybrid energy system in the presence of demand response program: ε-constraint method and fuzzy satisfying approach’, *Energy Conversion and Management*, 138, pp. 383–392. doi: 10.1016/j.enconman.2017.02.003.

Noor, M. N. *et al.* (2014) *Filling missing data using interpolation methods: Study on the effect of fitting distribution*. Trans Tech Publ.

Noor, S. *et al.* (2020) ‘Analysis of public reactions to the novel Coronavirus (COVID-19) outbreak on Twitter’, *Kybernetes*. doi: 10.1108/K-05-2020-0258.

Nosratabadi, S. M., Hooshmand, R. A. and Gholipour, E. (2017) ‘A comprehensive review on microgrid and virtual power plant concepts employed for distributed energy resources scheduling in power systems’, *Renewable and Sustainable Energy Reviews*, 67, pp. 341–363. doi: 10.1016/j.rser.2016.09.025.

Obar, J. A. and Wildman, S. S. (2015) ‘Social Media Definition and the Governance Challenge: An Introduction to the Special Issue’, *SSRN Electronic Journal*. doi: 10.2139/ssrn.2647377.

Oikonomou, L. and Tjortjis, C. (2018) ‘A method for predicting the winner of the usa presidential elections using data extracted from twitter’, in *2018 South-Eastern European Design Automation, Computer Engineering, Computer Networks and Society Media Conference (SEEDA_CECNSM)*. IEEE, pp. 1–8.

Osakwe, Z. T. *et al.* (2021) ‘Identifying public concerns and reactions during the COVID-19 pandemic on Twitter: A text-mining analysis’, *Public Health Nursing*, 38(2), pp. 145–151. doi: 10.1111/phn.12843.

Otte, E. and Rousseau, R. (2002) ‘Social network analysis: A powerful strategy, also for the information sciences’, *Journal of Information Science*, 28(6), pp. 441–453. doi: 10.1177/016555150202800601.

Paiva, L. T. and Fontes, F. A. C. C. (2018) ‘Optimal electric power generation with underwater kite systems’, *Computing*, 100(11), pp. 1137–1153. doi: 10.1007/s00607-018-0643-4.

Palli, N. *et al.* (1998) ‘An interactive multistage e-inequality constraint method formultiple objectives decision making’, in *Proceedings of the ASME Design Engineering Technical Conference*, p. V002T02A006. doi: 10.1115/DETC98/DAC-5598.

Pallis, G., Zeinalipour-Yazti, D. and Dikaiakos, M. D. (2011) 'Online social networks: Status and trends', *Studies in Computational Intelligence*, 331, pp. 213–234. doi: 10.1007/978-3-642-17551-0_8.

Paltoglou, G. and Thelwall, M. (2010) 'A study of information retrieval weighting schemes for sentiment analysis', in *Proceedings of the 48th annual meeting of the association for computational linguistics*, pp. 1386–1395.

Pang, B. and Lee, L. (2008) 'Opinion mining and sentiment analysis', *Foundations and Trends in Information Retrieval*, 2(1–2), pp. 1–135. doi: 10.1561/1500000011.

Papazoglou, M. P. and Georgakopoulos, D. (2003) 'Introduction: Service-oriented computing', *Communications of the ACM*, 46(10), pp. 24–28.

Park, H. W., Park, S. and Chong, M. (2020) 'Conversations and medical news frames on twitter: Infodemiological study on covid-19 in south korea', *Journal of Medical Internet Research*, 22(5), p. e18897.

Park, M., Lee, J. and Won, D.-J. (2020) 'Demand Response Strategy of Energy Prosumer Based on Robust Optimization Through Aggregator', *IEEE Access*, 8, pp. 202969–202979.

Paterl, D. (2019) *10 Social Media Trends to Watch in 2019*. Available at: https://www.entrepreneur.com/article/324901 (Accessed: 6 June 2021).

Pedregosa, F. *et al.* (2015) 'Scikit-learn: Machine Learning in Python', *Journal of Machine Learning Research 12*, 19(1), pp. 29–33.

Perrin, A. (2015) *Social Media Usage: 2005-2015 | Pew Research Center*, *Pew Internet & American Life Project2*. Available at: http://www.pewinternet.org/2015/10/08/social-networking-usage-2005-2015/ (Accessed: 1 June 2021).

Petrican, T. *et al.* (2018) 'Evaluating forecasting techniques for integrating household energy prosumers into smart grids', in *Proceedings - 2018 IEEE 14th International Conference on Intelligent Computer Communication and Processing, ICCP 2018*, pp. 79–85. doi: 10.1109/ICCP.2018.8516617.

Piatetsky-Shapiro, G. (1991) 'Discovery, analysis, and presentation of strong rules', *Knowledge discovery in databases*, pp. 229–248. Available at: http://ci.nii.ac.jp/naid/10000000985/.

Pochampally, R. and Varma, V. (2011) 'User context as a source of topic retrieval in Twitter', in *... on Enriching Information Retrieval (...*. Citeseer, pp. 1–3. Available at: http://select.cs.cmu.edu/meetings/enir2011/papers/pochampally-varma.pdf.

Poli, R. and Koza, J. (2014) 'Genetic programming', in *Search Methodologies: Introductory*

Tutorials in Optimization and Decision Support Techniques, Second Edition. Springer, pp. 143–186. doi: 10.1007/978-1-4614-6940-7_6.

Pop, C. *et al.* (2018) 'Blockchain based decentralized management of demand response programs in smart energy grids', *Sensors (Switzerland)*, 18(1), p. 162. doi: 10.3390/s18010162.

Pop, C. *et al.* (2019) 'Blockchain-based scalable and tamper-evident solution for registering energy data', *Sensors (Switzerland)*, 19(14), p. 3033. doi: 10.3390/s19143033.

Pop, C. *et al.* (2020) 'Blockchain based Decentralized Applications: Technology Review and Development Guidelines', *arXiv preprint arXiv:2003.07131*. doi: 10.3390/fi13030062.

Quiggin, D. *et al.* (2012) 'A simulation and optimisation study: Towards a decentralised microgrid, using real world fluctuation data', *Energy*, 41(1), pp. 549–559. doi: 10.1016/j.energy.2012.02.007.

Rafea, A. and Mostafa, N. A. (2013) 'Topic extraction in social media', in *Proceedings of the 2013 International Conference on Collaboration Technologies and Systems, CTS 2013*. IEEE, pp. 94–98. doi: 10.1109/CTS.2013.6567212.

Ren, H. *et al.* (2010) 'Multi-objective optimization for the operation of distributed energy systems considering economic and environmental aspects', *Applied Energy*, 87(12), pp. 3642–3651. doi: 10.1016/j.apenergy.2010.06.013.

Ribeiro, G. T., Mariani, V. C. and Coelho, L. dos S. (2019) 'Enhanced ensemble structures using wavelet neural networks applied to short-term load forecasting', *Engineering Applications of Artificial Intelligence*, 82, pp. 272–281. doi: 10.1016/j.engappai.2019.03.012.

Richthammer, C. *et al.* (2013) *Taxonomy for social network data types from the viewpoint of privacy and user control, Proceedings - 2013 International Conference on Availability, Reliability and Security, ARES 2013*. doi: 10.1109/ARES.2013.18.

Rideout, V. (2015) 'The Common Sense Census: Media Use By Tweens and Teens', *Common Sense Media*, pp. 1–104.

Rittho, O. *et al.* (2001) 'Yale: Yet another learning environment', in *LLWA 01-Tagungsband der GI-Workshop-Woche Lernen-Lehren-Wissen-Adaptivität*. Citeseer, pp. 84–92.

Röder, M., Both, A. and Hinneburg, A. (2015) 'Exploring the space of topic coherence measures', in *WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, pp. 399–408. doi: 10.1145/2684822.2685324.

Rodrigues, F. and Trindade, A. (2018) 'Load forecasting through functional clustering and ensemble learning', *Knowledge and Information Systems*, 57(1), pp. 229–244. doi:

10.1007/s10115-018-1169-y.

Rosenberg, H., Syed, S. and Rezaie, S. (2020) 'The Twitter pandemic: The critical role of Twitter in the dissemination of medical information and misinformation during the COVID-19 pandemic', *Canadian Journal of Emergency Medicine*, 22(4), pp. 418–421. doi: 10.1017/cem.2020.361.

Rousidis, D., Koukaras, P. and Tjortjis, C. (2020) 'Social media prediction: a literature review', *Multimedia Tools and Applications*, 79(9–10), pp. 6279–6311. doi: 10.1007/s11042-019-08291-9.

Rousseeuw, P. J. (1987) 'Silhouettes: A graphical aid to the interpretation and validation of cluster analysis', *Journal of Computational and Applied Mathematics*, 20(C), pp. 53–65. doi: 10.1016/0377-0427(87)90125-7.

Samuel, J. *et al.* (2020) 'Covid-19 public sentiment insights and machine learning for tweets classification', *Information*, 11(6), p. 314.

Schubert, E. *et al.* (2017) 'DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN', *ACM Transactions on Database Systems*, 42(3), pp. 1–21. doi: 10.1145/3068335.

Scott, J. *et al.* (1996) 'Social Network Analysis: Methods and Applications', *The British Journal of Sociology*, 47(2), p. 375. doi: 10.2307/591741.

Sefidgar-Dezfouli, A., Joorabian, M. and Mashhour, E. (2019) 'A multiple chance-constrained model for optimal scheduling of microgrids considering normal and emergency operation', *International Journal of Electrical Power and Energy Systems*, 112, pp. 370–380. doi: 10.1016/j.ijepes.2019.05.026.

Sharma, N. *et al.* (2014) 'Leveraging weather forecasts in renewable energy systems', *Sustainable Computing: Informatics and Systems*, 4(3), pp. 160–171. doi: 10.1016/j.suscom.2014.07.005.

Sibuya, M. (1993) 'A random clustering process', *Annals of the Institute of Statistical Mathematics*, 45(3), pp. 459–465. doi: 10.1007/BF00773348.

Sideratos, G., Ikonomopoulos, A. and Hatziargyriou, N. D. (2020) 'A novel fuzzy-based ensemble model for load forecasting using hybrid deep neural networks', *Electric Power Systems Research*, 178, p. 106025. doi: 10.1016/j.epsr.2019.106025.

Singh, L. *et al.* (2020) 'A first look at COVID-19 information and misinformation sharing on Twitter', *arXiv preprint arXiv:2003.13907*.

Sioshansi, F. (2019) *Consumer, prosumer, prosumager: How service innovations will disrupt the utility business model*, *Consumer, Prosumer, Prosumager: How Service Innovations will*

Disrupt the Utility Business Model. Academic Press. doi: 10.1016/C2018-0-01192-5.

SmartHome / ITI (2020). Available at: https://smarthome.iti.gr/ (Accessed: 4 June 2021).

Statistics for Social Networks: Top 15 Most Popular Social Networking Sites, Active users [2017] - DreamGrow (2017). Available at: https://www.dreamgrow.com (Accessed: 12 October 2017).

Stecchi, U. *et al.* (2019) 'A multipurpose ICT platform for supporting energy transition: First results in flexibility profiling', in *SyNERGY MED 2019 - 1st International Conference on Energy Transition in the Mediterranean Area*. IEEE, pp. 1–6. doi: 10.1109/SyNERGY-MED.2019.8764146.

Tabatabaei, N. M., Kabalci, E. and Bizon, N. (2019) *Microgrid architectures, control and protection methods*. Springer.

Tankovska, H. (2021) • *Number of social media users 2025 | Statista, Jan 28,2021*. Available at: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/ (Accessed: 31 May 2021).

Thelwall, M. *et al.* (2010) 'Sentiment strength detection in short informal text', *Journal of the American society for information science and technology*, 61(12), pp. 2544–2558.

Thelwall, M. (2014) 'Sentiment analysis and time series with Twitter', *Twitter and society*, pp. 83–95.

Thelwall, M. (2018) 'Social media analytics for YouTube comments: Potential and limitations', *International Journal of Social Research Methodology*, 21(3), pp. 303–316.

Thelwall, M., Buckley, K. and Paltoglou, G. (2011) 'Sentiment in Twitter events', *Journal of the American Society for Information Science and Technology*, 62(2), pp. 406–418.

Thelwall, M., Buckley, K. and Paltoglou, G. (2012) 'Sentiment strength detection for the social web', *Journal of the American Society for Information Science and Technology*, 63(1), pp. 163–173.

Top 50 2019's Acquisitions in Social Media - Index (2019). Available at: https://index.co/top/market/social-media/acquisitions/2019 (Accessed: 6 June 2021).

Torkzadeh, R. *et al.* (2014) 'Medium term load forecasting in distribution systems based on multi linear regression & principal component analysis: A novel approach', in *2014 19th Conference on Electrical Power Distribution Networks, EPDC 2014*. IEEE, pp. 66–70. doi: 10.1109/EPDC.2014.6867500.

Tsiara, E. and Tjortjis, C. (2020) 'Using Twitter to Predict Chart Position for Songs', in *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer,

pp. 62–72.

Tsolakis, A. C. *et al.* (2018) ‘A Secured and Trusted Demand Response system based on Blockchain technologies’, in *2018 IEEE (SMC) International Conference on Innovations in Intelligent Systems and Applications, INISTA 2018*, pp. 1–6. doi: 10.1109/INISTA.2018.8466303.

Twitter Inc. (2021) *Coronavirus: Staying safe and informed on Twitter*, *Twitter*. Available at: https://blog.twitter.com/en_us/topics/company/2020/covid-19.html (Accessed: 3 June 2021).

Tzovaras *et al.* (2019) ‘DRIMPAC—Unified Demand Response Interoperability Framework Enabling Market Participation of Active Energy Consumers’, in *Proceedings*, p. 15. doi: 10.3390/proceedings2019020015.

Vafeiadis, T. *et al.* (2018) ‘Machine Learning Based Occupancy Detection via the Use of Smart Meters’, in *Proceedings - 2017 International Symposium on Computer Science and Intelligent Controls, ISCSIC 2017*. IEEE, pp. 6–12. doi: 10.1109/ISCSIC.2017.15.

Valentini, C. and Kruckeberg, D. (2012) ‘New Media Versus Social Media: A Conceptualization of Their Meanings, Uses, and Implications for Public Relations’, *New Media and Public Relations*, pp. 3–12.

Vayansky, I. and Kumar, S. A. P. (2020) ‘A review of topic modeling methods’, *Information Systems*, 94, p. 101582. doi: 10.1016/j.is.2020.101582.

Vergados, D. J. *et al.* (2016) ‘Prosumer clustering into virtual microgrids for cost reduction in renewable energy trading markets’, *Sustainable Energy, Grids and Networks*, 7, pp. 90–103. doi: 10.1016/j.segan.2016.06.002.

Ves, A. V. *et al.* (2019) ‘A Stacking Multi-Learning Ensemble Model for Predicting Near Real Time Energy Consumption Demand of Residential Buildings’, in *Proceedings - 2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing, ICCP 2019*, pp. 183–189. doi: 10.1109/ICCP48234.2019.8959572.

Vesa, A. V. *et al.* (2020) ‘Energy flexibility prediction for data center engagement in demand response programs’, *Sustainability (Switzerland)*, 12(4), p. 1417. doi: 10.3390/su12041417.

Wang, C., Paisley, J. and Blei, D. M. (2011) ‘Online variational inference for the hierarchical Dirichlet process’, in *Journal of Machine Learning Research*. JMLR Workshop and Conference Proceedings, pp. 752–760.

Wang, L. *et al.* (2020) ‘Ensemble learning for load forecasting’, *IEEE Transactions on Green Communications and Networking*, 4(2), pp. 616–628.

Wang, X., Gerber, M. S. and Brown, D. E. (2012) ‘Automatic crime prediction using events

extracted from twitter posts', in *International conference on social computing, behavioral-cultural modeling, and prediction*. Springer, pp. 231–238. doi: 10.1007/978-3-642-29047-3_28.

Wang, Z., Wang, Y. and Srinivasan, R. S. (2018) 'A novel ensemble learning approach to support building energy use prediction', *Energy and Buildings*, 159, pp. 109–122. doi: 10.1016/j.enbuild.2017.10.085.

Wetherill, G. B. and Seber, G. A. F. (1977) *Linear Regression Analysis.*, *Journal of the Royal Statistical Society. Series A (General)*. John Wiley & Sons. doi: 10.2307/2345290.

WikiPedia (2019) *List of mergers and acquisitions by Facebook*, *Wikipedia*. Available at: https://en.wikipedia.org/wiki/List_of_mergers_and_acquisitions_by_Facebook (Accessed: 1 June 2021).

Xia, C., Wang, J. and McMenemy, K. (2010) 'Short, medium and long term load forecasting model and virtual load forecaster based on radial basis function neural networks', *International Journal of Electrical Power and Energy Systems*, 32(7), pp. 743–750. doi: 10.1016/j.ijepes.2010.01.009.

Yakhchi, S. *et al.* (2017) 'ARMICA-Improved: A new approach for association rule mining', in *Int'l Conf. on Knowledge Science, Engineering and Management*. Springer, pp. 296–306. doi: 10.1007/978-3-319-63558-3_25.

Yang, R. and Wang, L. (2012) 'Multi-objective optimization for decision-making of energy and comfort management in building automation and control', *Sustainable Cities and Society*, 2(1), pp. 1–7. doi: 10.1016/j.scs.2011.09.001.

Yang, S. *et al.* (2020) 'A multi-objective stochastic optimization model for electricity retailers with energy storage system considering uncertainty and demand response', *Journal of Cleaner Production*, 277, p. 124017.

Yang, S. H. *et al.* (2014) 'Large-scale high-precision topic modeling on twitter', in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1907–1916. doi: 10.1145/2623330.2623336.

Yang, Y., Hong, W. and Li, S. (2019) 'Deep ensemble learning based probabilistic load forecasting in smart grids', *Energy*, 189, p. 116324. doi: 10.1016/j.energy.2019.116324.

Your 2019 Social Media Strategy: 4 Trends You Can't Ignore (2019). Available at: https://www.convinceandconvert.com/social-media-marketing/social-media-trends-2019/ (Accessed: 1 June 2021).

Yu, M. and Hong, S. H. (2017) 'Incentive-based demand response considering hierarchical

electricity market: A Stackelberg game approach', *Applied Energy*, 203, pp. 267–279. doi: 10.1016/j.apenergy.2017.06.010.

Yu, M., Lu, R. and Hong, S. H. (2016) 'A real-time decision model for industrial load management in a smart grid', *Applied Energy*, 183, pp. 1488–1497. doi: 10.1016/j.apenergy.2016.09.021.

Yuan, Y., Xu, H. and Wang, B. (2014) 'An improved NSGA-III procedure for evolutionary many-objective optimization', in *GECCO 2014 - Proceedings of the 2014 Genetic and Evolutionary Computation Conference*, pp. 661–668. doi: 10.1145/2576768.2598342.

Zafarani, R., Abbasi, M. A. and Liu, H. (2014) *Social media mining: An introduction*, *Social Media Mining: An Introduction*. Cambridge University Press. doi: 10.1017/CBO9781139088510.

Zare, M. *et al.* (2016) 'New Stochastic Bi-Objective Optimal Cost and Chance of Operation Management Approach for Smart Microgrid', *IEEE Transactions on Industrial Informatics*, 12(6), pp. 2031–2040. doi: 10.1109/TII.2016.2585379.

Zhang, C. *et al.* (2018) 'Peer-to-Peer energy trading in a Microgrid', *Applied Energy*, 220, pp. 1–12.

Zhang, D. *et al.* (2016) 'An Optimal and Learning-Based Demand Response and Home Energy Management System', *IEEE Transactions on Smart Grid*, 7(4), pp. 1790–1801. doi: 10.1109/TSG.2016.2552169.

Zhang, Y., Gatsis, N. and Giannakis, G. B. (2013) 'Robust energy management for microgrids with high-penetration renewables', *IEEE Transactions on Sustainable Energy*, 4(4), pp. 944–953. doi: 10.1109/TSTE.2013.2255135.

Zhao, W. X. *et al.* (2011) 'Topical keyphrase extraction from Twitter', in *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 379–388.

Zhou, Y. *et al.* (2017) 'Performance Evaluation of Peer-to-Peer Energy Sharing Models', *Energy Procedia*, 143, pp. 817–822. doi: 10.1016/j.egypro.2017.12.768.

Appendices

Appendix A

Table A1. The complete set of 112 SM sites.

SM sites			
Facebook	Gab	Cross.tv	Plurk
YouTube	Telegram	Flixster	LiveJournal
Instagram	Tagged	Gaia Online	Weibo
Twitter	Myspace	BlackPlanet	Qzone
Reddit	Badoo	MyMFB	QQ
Vine	Stumbleupon	Care2	Baidu
Pinterest	Foursquare	CaringBridge	Line
Ask.fm	MeetMe	GoFundMe	YY
Tumblr	Skyrock A192	Tinder	Sprybirds
Flickr	Pinboard	Crokes	Xing
Google+	Kiwibox	Goodreads	VampireFreaks
LinkedIn	Twoo	Internations	CafeMom

SM sites			
VK	Yelp	PlentyofFish	Ravelry
ClassMates	Snapfish	Minds	ASmallWorld
Meetup	Photobucket	Nexopia	ReverbNation
WhatsApp	Shutterfly	Glocals	SoundCloud
Messenger	500px	Academia.edu	Solaborate
Snapchat	DeviantArt	Busuu	eToro
Quora	Dronestagram	English, baby!	Xanga
GirlsAskGuys	Fotki	Italki.com	Ryze
Nextdoor	Fotolog	Untappd	Zynga
ProductHunt	Imgur	Doximity	Habbo
AngelList	Pixabay	Wayn	FunnyOrDie
Kickstarter	WeHeartIt	CouchSurfing	Tout
WeChat	43Things	TravBuddy	Classmates
Skype	Path	Tournac	MyHeritage
Viber	Uplike	Cellufun	MocoSpace
Viadeo	Last.fm	23andMe	Ancestry.com

Appendix B

Table B1. Mapping official features to utilities.

SM sites	Primary
Utility	Official Features
Connecting (Count=52)	Fans, Groups, Live Chat, Pokes, Gifts, Messaging, Explore, Instagram Direct, Direct Messaging, Discussion Website, Exploring, Profiles, Messaging to Blogs, Accounts, User Profiles, Circles, Communities, Collections, Emails, User Profile Network, Influencers, Synchronization with Other Social Networks, SMS Service, Members, Neighbors, Chatting, Drafts, Secret chats, Voice Calls, Bands, Dating, Mothers, Weaving, Christian, Talent, Muslims, Activists, Political, Authors, Expats, Follow, Teenagers, Celebrities, Relatives, User Groups, Messages, Group and Voice Chat, Video conferences, Conversations, Chat features
Multimedia (Count=29)	Photos, Videos, Text, Upload and download options for Photos, Playback Upload Quality and formats, Live Streaming, 3D Videos, 360o Videos, Images, Live Videos, Photographic Filters, Record Short Video Clips, Ability To "Revine" Videos on A Personal Stream, Stream, Photography, Voice, Image Filters, Short videos, Gab, Cloud-Based Messages, Audio, Files, Musicians, Crocheting, Photoblog, VideoBlog, AudioBlog, Pictures
Professional (Count=36)	Monetization, Licensing, Job Listings, Online Recruiting, For-Pay Research, Snapcash, Products, Startups, Investors, Funding, Channels, Enterprises, Purchases, Home Services, Drones, Knitting, Environmental, Treatments, Medical, Illness, Funding, Rewards, Academics, Papers, Teaching, Language, Health, Business, Promoting, Companies, Technology, Trading, Stock offering, Virtual Currency, Video Streaming for money, Video tutorials for money
Sharing (Count=23)	Post Text, Instagram Stories, Tweet, Retweet, Links, Hashtags, Sharing Content, Protected Posts, Pins, Boards, Send Questions, Queue, Tags, Questions, What's Hot, Post to And Read Community Boards, Post, Content Discovery, Location, Inspiration, Spinning, Sharing, Posting, Quoting
Entertainment (Count=17)	Games, Shopping, Gaming, Art, Music, Culture, Travel, Luxury, Movies, Animes, Books, Comedy, Online Social Gaming, Gamers, Concerts, Fashion, Sports

SM sites	Primary
Opinions (Count=15)	Polls, Answers, Suggest Edits, Feeds, Recommendations, Reviews, Advice, Recommendation, Discussions, Forums, Opinions, Reviews, Discussion forums,
Profile (Count=13)	Wall, Calendar, Embedded in Profile, Skills, Memories, Bookmarking, Goals, Career, Records, Professional Profiles, Profile, Journals, Diaries
Publishing (Count=11)	Dashboard (Blog Posts), Google+ Page, Locations, Google Local, Publishing Platform, Blog, Blogging, Weblog, Pulse, Blogs, Microblogging
Applications (Count=15)	Apps, Stand-alone Apps, Third-party Services, HTML editing, Interaction and compatibility, Filtering, Additional features, Deprecated Features, Applications, External, Third Party Applications, Mobile, SMS, Bots, third party development
Schedule (Count=8)	Organization, View Information About Upcoming Reunions, Organize Meetups, Events, Activities, Planning, Event, Event coordination
Privacy (Count=6)	Classified section, Access control, Identity Service, Privacy, Security and Technology, Enhanced Privacy
Voting (Count=7)	Likes, Web Content Rating, Voting, +1 Button, Like Buttons, Upvote/Downvote, Stickers
News (Count=7)	News Feed, Status, Follow People & Trending Topics, Social News Aggregation, Following, News, Tech News
Promoting (Count=4)	Fan Pages, Links, Advertising, Ad-Free

Appendix C

Table C1. Utility Occurrences on the SM dataset.

No	SM sites	Connecting	Multimedia	Professional	Sharing	Entertainment	Opinions	Profile	Publishing	Applications	Schedule	Privacy	Voting	News	Promoting
1	Facebook	7	4	-	-	1	1	1	-	1	-	1	1	2	2
2	YouTube	-	6	-	1	-	-	-	-	-	-	-	-	-	-
3	Instagram	2	3	1	1	-	-	-	-	2	-	-	-	-	-
4	Twitter	1	2	-	4	-	-	-	-	-	-	-	-	1	-
5	Reddit	1	1	-	3	-	-	-	-	-	-	-	2	1	-
6	Vine	-	2	-	1	-	-	-	-	-	-	-	-	-	-
7	Pinterest	1	2	-	2	-	-	-	-	-	-	-	-	1	
8	Ask.fm	1	-	-	1	-	-	-	-	-	-	-	-	-	1
9	Tumblr	1	-	-	4	-	-	-	2	1	-	-	-	-	
10	Flickr	1	2	1	-	-	-	-	-	2	1	1	-	-	-
11	Google+	5	2	-	1	-	-	1	3	2	-	2	1	-	-
12	LinkedIn	3	-	3	-	-	-	2	1	4	-	1	-	-	1
13	VK	4	-	-	-	-	-	-	-	-	-	1	1	1	-
14	ClassMates	2	-	-	1	-	-	-	-	-	1	1	-	-	-
15	Meetup	2	-	-	-	-	-	-	-	-	1	-	-	-	-
16	WhatsApp	1	3	-	-	-	-	-	-	-	-	-	-	-	-
17	Messenger	1	3	-	-	-	-	-	-	-	-	-	-	-	-
18	Snapchat	-	3	1	-	-	-	1	-	-	-	-	-	-	-
19	Quora	-	-	-	-	-	3	1	-	-	-	-	1	-	-
20	GirlsAskGuys	-	1	-	2	-	3	-	-	-	-	-	-	-	-
21	Nextdoor	1	-	-	-	-	-	-	-	-	2	-	-	-	-

No	SM sites	Connecting	Multimedia	Professional	Sharing	Entertainment	Opinions	Profile	Publishing	Applications	Schedule	Privacy	Voting	News	Promoting
22	ProductHunt	-	-	1	-	-	-	-	-	-	-	-	1	-	-
23	AngelList	-	-	2	-	-	-	-	-	-	-	-	-	-	-
24	Kickstarter	-	-	2	-	-	-	-	-	-	-	-	-	-	-
25	WeChat	2	-	-	-	1	-	-	-	-	-	-	-	-	-
26	Skype	1	3	-	-	-	-	-	-	-	-	-	-	-	-
27	Viber	1	3	-	-	-	-	-	-	-	-	-	-	-	-
28	Viadeo	-	-	3	-	-	-	-	-	-	-	-	-	-	-
29	Gab	1	1	-	-	-	-	-	-	-	-	-	-	-	1
30	Telegram	4	1	1	-	-	-	1	-	2	-	1	1	-	-
31	Tagged	1	-	-	-	-	-	-	-	-	-	-	-	-	
32	Myspace	1	1	-	-	-	-	-	-	-	-	-	-	-	-
33	Badoo	1	-	-	-	-	-	-	-	-	-	-	-	-	-
34	Stumbleupon	-	-	-	1	-	-	-	-	-	-	-	-	-	-
35	Foursquare	-	-	2	1		1	-	-	-	-	-	-	-	-
36	MeetMe	1	-	-	-	-	-	-	-	-	-	-	-	-	-
37	Skyrock A192	-	-	-	-	-	-	-	1	-	-	-	-	-	-
38	Pinboard	-	-	-	-	-	-	1	-	-	-	-	-	-	1
39	Kiwibox	-	1	-	-	1	-	-	1	-	-	-	-	-	-
40	Twoo	1	1	-	-	-	-	-	-	-	-	-	-	-	-
41	Yelp	-	1	1	-	-	2	-	-	-	1	-	-	-	-
42	Snapfish	-	1	-	-	-	-	-	-	-	-	-	-	-	-
43	Photobucket	-	2	-	-	-	-	-	-	-	-	-	-	-	-
44	Shutterfly	-	1	-	-	-	-	-	-	-	-	-	-	-	-
45	500px	-	1	-	-	-	-	-	-	-	-	-	-	-	-
46	DeviantArt	-	1	-	-	1	-	-	-	-	-	-	-	-	-
47	Dronestagram	-	1	1	-	-	-	-	-	-	-	-	-	-	-

No	SM sites	Connecting	Multimedia	Professional	Sharing	Entertainment	Opinions	Profile	Publishing	Applications	Schedule	Privacy	Voting	News	Promoting
48	Fotki	-	1	-	-	-	-	-	-	-	-	-	-	-	-
49	Fotolog	-	1	-	-	-	-	-	1	-	-	-	-	-	-
50	Imgur	-	1	-	-	-	-	-	-	-	-	-	1	-	-
51	Pixabay	-	2	-	-	-	-	-	-	-	-	-	-	-	-
52	WeHeartIt	-	1	-	-	-	-	-	-	-	-	-	-	-	-
53	43Things	-	-	-	1	-	1	1	-	-	-	-	-	-	-
54	Path	1	1	-	-	-	-	-	-	-	-	1	-	-	-
55	Uplike	-	1	-	1	-	-	-	-	-	-	-	-	-	-
56	Last.fm	-	-	-	-	1	1	-	-	-	-	-	-	-	-
57	Cross.tv	1	-	-	-	-	-	-	-	-	-	-	-	-	-
58	Flixster	-	-	-	-	1	-	-	-	-	-	-	-	-	-
59	Gaia Online	-	-	-	-	1	-	-	-	-	-	-	-	-	-
60	BlackPlanet	3	-	-	-	-	-	-	1	-	-	-	-	-	-
61	MyMFB	1	-	-	-	-	-	-	-	-	-	-	-	-	-
62	Care2	2	-	1	-	-	-	-	-	-	-	-	-	-	-
63	CaringBridge	-	-	3	-	-	-	-	-	-	-	-	-	-	-
64	GoFundMe	-	-	1	-	-	-	-	-	-	-	-	-	-	-
65	Tinder	1	-	-	1	-	-	-	-	-	-	-	-		
66	Crokes	2	-	-	-	-	-	-	-	-	-	-	-	-	-
67	Goodreads	-	-	-	-	1	1	-	-	-	-	-	-	-	-
68	Internations	1	-	-	-	-	-	-	-	-	-	-	-	-	-
69	PlentyofFish	1	-	-	-	-	-	-	-	-	-	-	-	-	-
70	Minds	-	-	2	-	-	-	-	-	-	-	1	-	-	-
71	Nexopia	-	-	-	-	-	2	-	-	-	-	-	-	-	-
72	Glocals	1	-	-	-	-	-	-	-	-	2	-	-	-	-
73	Academia.edu	1	-	2	-	-	-	-	-	-	-	-	-	-	-

No	SM sites	Connecting	Multimedia	Professional	Sharing	Entertainment	Opinions	Profile	Publishing	Applications	Schedule	Privacy	Voting	News	Promoting
74	Busuu	-	-	2	-	-	-	-	-	-	-	-	-	-	-
75	English, baby!	-	-	2	-	-	-	-	-	-	-	-	-	-	-
76	Italki.com	-	-	2	-	-	-	-	-	-	-	-	-	-	-
77	Untappd	-	1	-	-	-	2	-	-	-	-	-	-	-	-
78	Doximity	-	-	1	-	-	-	-	-	-	-	-	-	-	-
79	Wayn	-	-	-	-	1	-	-	-	-	-	-	-	-	-
80	CouchSurfing	-	-	-	-	1	-	-	-	-	1	-	-	-	-
81	TravBuddy	-	-	-	-	1	-	-	-	-	-	-	-	-	-
82	Tournac	-	-	-	1	1	-	-	-	-	-	-	-	-	-
83	Cellufun	-	-	-		1	-	-	-	-	-	-	-	-	-
84	23andMe	2	-	1	-	-	-	-	-	-	-	-	-	-	-
85	Plurk	1	-	-	1	-	-	-	1	-	-	-	-	-	-
86	LiveJournal	-	-	-	-	-	-	2	1	-	-	-	-	-	-
87	Weibo	-	-	-	2	-	-	-	1	-	-	-	-	-	-
88	Qzone	-	3	-	-	-	-	1	1	-	-	-	-	-	-
89	QQ	2	1	1	-	2	-	-	1	-	-	-	-	-	-
90	Baidu	-	2	-	3	-	1	-	-	-	-	-	-	-	-
91	Line	3	4	-	-	-	-	-	-	-	-	-	-	-	-
92	YY	1	-	3	-	4	-	-	-	-	-	-	-	-	-
93	Sprybirds	-	-	1	-	-	-	-	-	-	-	-	-	-	-
94	Xing	-	-	1	-	-	1	1	-	-	1	-	-	-	-
95	VampireFreaks	1	-	-	-	1	-	-	-	-	-	-	-	-	-
96	CafeMom	1	-	-	-	-	-	-	-	-	-	-	-	-	-
97	Ravelry	1	1	1	1	-	-	-	-	-	-	-	-	-	-
98	ASmallWorld	1	-	-	-	2	-	-	-	-	-	-	-	-	-
99	ReverbNation	-	-	-	-	1	-	1	-	-	-	-	-	-	-

No	SM sites	Connecting	Multimedia	Professional	Sharing	Entertainment	Opinions	Profile	Publishing	Applications	Schedule	Privacy	Voting	News	Promoting
100	SoundCloud	-	-	-	1	1	-	-	-	-	-	-	-	-	-
101	Solaborate	1	-	3	-	-	2	1	-	-	1	-	-	1	-
102	eToro	1	-	2	-	-	-	-	-	-	-	-	-	-	-
103	Xanga	1	3	-	-	-	-	1	2	-	-	1	-	-	-
104	Ryze	-	-	1	-	-	-	-	-	-	-	-	-	-	-
105	Zynga	-	-	-	-	1	-	-	-	-	-	-	-	-	-
106	Habbo	1	-	-	-	1	-	-	-	-	-	-	-	-	-
107	FunnyOrDie	1	1	-	-	1	-	-	-	-	-	-	-	-	-
108	Tout	-	-	1	-	-	-	-	-	-	-	-	-	-	-
109	Classmates	2	-	-	-	-	-	-	-	-	1	-	-	-	-
110	MyHeritage	-	1	-	-	-	-	1	-	-	-	-	-	-	-
111	MocoSpace	-	-	-	-	1	-	-	-	-	-	-	-	-	-
112	Ancestry.com	2	-	-	-	-	-	-	-	-	-	-	-	-	-

Appendix D

Table D1. SMP's Primary, Secondary and trivia utilities.

SM sites	Primary	Secondary	Trivia
Facebook	Connecting (7)	Multimedia (4)	Entertainment (1), Opinions (1), Profile (1), Applications (1), Privacy (1), Voting (1), News (2), Promoting (2)
YouTube	Multimedia (6)	Sharing (1)	-
Instagram	Multimedia (3)	Connecting (2)	Professional (1), Sharing (1), Applications (2)
Twitter	Sharing (4)	Multimedia (2)	Connecting (1), News (1)
Reddit	Sharing (3)	Voting (2)	Connecting (1), Multimedia (1), News (1)
Vine	Multimedia (2)	Sharing (1)	-
Pinterest	Multimedia (2), Sharing (2)	Connecting (1), News (1)	-
Ask.fm	Sharing (1), Connecting (1), Promoting (1)	-	-
Tumblr	Sharing (4)	Publishing (2)	Connecting (1), Applications (1)
Flickr	Multimedia (2), Applications (2)	Connecting (1), Professional (1), Schedule (1), Privacy (1)	-
Google+	Connecting (5)	Publishing (3)	Multimedia (2), Sharing (1), Profile (1), Applications (2), Privacy (2), Voting (1)
LinkedIn	Applications (4)	Connecting (3), Professional (3)	Profile (2), Publishing (1), Privacy (1)
VK	Connecting (4)	Privacy (1), Voting (1), News (1)	-
ClassMates	Connecting (2)	Sharing (1), Schedule (1), Privacy (1)	-
Meetup	Connecting (2)	Schedule (1)	-
WhatsApp	Multimedia (3)	Connecting (1)	-
Messenger	Multimedia (3)	Connecting (1)	-
Snapchat	Multimedia (3)	Professional (1), Profile (1)	-
Quora	Opinions (3)	Profile (1), Voting (1)	-

SM sites	Primary	Secondary	Trivia
GirlsAskGuys	Opinions (3)	Sharing (2)	Multimedia (1)
Nextdoor	Schedule (2)	Connecting (1)	-
ProductHunt	Professional (1), Voting (1)	-	-
AngelList	Professional (2)	-	-
Kickstarter	Professional (2)	-	-
WeChat	Connecting (2)	Entertainment (1)	-
Skype	Multimedia (3)	Connecting (1)	-
Viber	Multimedia (3)	Connecting (1)	-
Viadeo	Professional (3)	-	-
Gab	Connecting (1), Multimedia (1), Promoting (1)	-	-
Telegram	Connecting (4)	Applications (2)	Multimedia (1), Professional (1), Profile (1), Privacy (1), Voting (1)
Tagged	Connecting (1)	-	-
Myspace	Connecting (1), Multimedia (1)	-	-
Badoo	Connecting (1)	-	-
Stumbleupon	Sharing (1)	-	-
Foursquare	Professional (2)	Sharing (1), Opinions (1)	-
MeetMe	Connecting (1)	-	-
Skyrock A192	Publishing (1)	-	
Pinboard	Profile (1), Promoting (1)	-	-
Kiwibox	Multimedia (1), Entertainment (1), Publishing (1)	-	-
Twoo	Connecting (1), Multimedia (1)	-	-
Yelp	Opinions (2)	Multimedia (1), Professional (1), Schedule (1)	-

SM sites	Primary	Secondary	Trivia
Snapfish	Multimedia (1)	-	-
Photobucket	Multimedia (2)	-	-
Shutterfly	Multimedia (1)	-	-
500px	Multimedia (1)	-	-
DeviantArt	Multimedia (1), Entertainment (1)	-	-
Dronestagram	Multimedia (1), Professional (1)	-	-
Fotki	Multimedia (1)	-	-
Fotolog	Multimedia (1), Publishing (1)	-	-
Imgur	Multimedia (1), Voting (1)	-	-
Pixabay	Multimedia (2)	-	-
WeHeartIt	Multimedia (1)	-	-
43Things	Sharing (1), Opinions (1), Profiling (1)	-	-
Path	Connecting (1), Multimedia (1), Privacy (1)	-	-
Uplike	Multimedia (1), Sharing (1)	-	-
Last.fm	Entertainment (1), Opinions (1)	-	-
Cross.tv	Connecting (1)	-	-
Flixster	Entertainment (1)	-	-
Gaia Online	Entertainment (1)	-	-
BlackPlanet	Connecting (3)	Publishing (1)	-
MyMFB	Connecting (1)	-	-
Care2	Connecting (2) Professional (1)	-	-

SM sites	Primary	Secondary	Trivia
CaringBridge	Professional (3)	-	-
GoFundMe	Professional (1)	-	-
Tinder	Connecting (1), Sharing (1)	-	-
Crokes	Connecting (2)	-	-
Goodreads	Entertainment (1), Opinions (1)	-	-
Internations	Connecting (1)	-	-
PlentyofFish	Connecting (1)	-	-
Minds	Professional (2)	Privacy (1)	-
Nexopia	Opinions (2)	-	-
Glocals	Schedule (2)	Connecting (1)	-
Academia.edu	Professional (2)	Connecting (1)	-
Busuu	Professional (2)	-	-
English, baby!	Professional (2)	-	-
Italki.com	Professional (2)	-	-
Untappd	Opinions (2)	Multimedia (1)	-
Doximity	Professional (1)	-	-
Wayn	Entertainment (1)	-	-
CouchSurfing	Entertainment (1), Schedule (1)	-	-
TravBuddy	Entertainment (1)	-	-
Tournac	Sharing (1), Entertainment (1)	-	-
Cellufun	Entertainment (1)	-	-
23andMe	Connecting (2)	Professional (1)	-
Plurk	Connecting (1), Sharing (1), Publishing (1)	-	-
LiveJournal	Profile (2)	Publishing (1)	-

SM sites	Primary	Secondary	Trivia
Weibo	Sharing (2)	Publishing (1)	-
Qzone	Multimedia (3)	Profile (1), Publishing (1)	-
QQ	Connecting (2), Entertainment (2)	Multimedia (1), Professional (1), Publishing (1)	-
Baidu	Sharing (3)	Multimedia (2)	Opinions (1)
Line	Multimedia (4)	Connecting (3)	-
YY	Entertainment (4)	Professional (3)	Connecting (1)
Sprybirds	Professional (1)	-	-
Xing	Professional (1), Opinions (1), Profile (1), Schedule (1)	-	-
VampireFreaks	Connecting (1), Entertainment (1)	-	-
CafeMom	Connecting (1)	-	-
Ravelry	Connecting (1), Multimedia (1), Professional (1), Sharing (1)	-	-
ASmallWorld	Entertainment (2)	Connecting (1)	-
ReverbNation	Entertainment (1), Profile (1)	-	-
SoundCloud	Sharing (1), Entertainment (1)	-	-
Solaborate	Professional (3)	Opinions (2)	Connecting (1), Profile (1), Schedule (1), News (1)
eToro	Professional (2)	Connecting (1)	-
Xanga	Multimedia (3)	Publishing (2)	Connecting (1), Profile (1), Privacy (1)
Ryze	Professional (1)	-	-
Zynga	Entertainment (1)	-	-
Habbo	Connecting (1), Entertainment (1)	-	-

SM sites	Primary	Secondary	Trivia
FunnyOrDie	Connecting (1), Multimedia (1), Entertainment (1)	-	-
Tout	Professional (1)	-	-
Classmates	Connecting (2)	Schedule (1)	-
MyHeritage	Multimedia (1), Profile (1)	-	-
MocoSpace	Entertainment (1)	-	-
Ancestry.com	Connecting (2)	-	-

Appendix E

Table E1. Frequent itemsets (FP-Growth).

Size	Support	Item 1	Item 2	Item 3
1	0.473	Connecting	-	-
1	0.384	Multimedia	-	-
1	0.277	Professional	-	-
1	0.205	Entertainment	-	-
1	0.196	Sharing	-	-
1	0.134	Profile	-	-
1	0.116	Opinions	-	-
1	0.116	Publishing	-	-
1	0.089	Privacy	-	-
1	0.089	Schedule	-	-
1	0.071	Voting	-	-
1	0.062	Applications	-	-
1	0.054	News	-	-
1	0.045	Promoting	-	-
2	0.188	Connecting	Multimedia	-
2	0.107	Connecting	Professional	-
2	0.071	Connecting	Entertainment	-
2	0.098	Connecting	Sharing	-

Size	Support	Item 1	Item 2	Item 3
2	0.054	Connecting	Profile	-
2	0.062	Connecting	Publishing	-
2	0.080	Connecting	Privacy	-
2	0.062	Connecting	Schedule	-
2	0.045	Connecting	Voting	-
2	0.062	Connecting	Applications	-
2	0.054	Connecting	News	-
2	0.036	Connecting	Promoting	-
2	0.071	Multimedia	Professional	-
2	0.045	Multimedia	Entertainment	-
2	0.098	Multimedia	Sharing	-
2	0.062	Multimedia	Profile	-
2	0.045	Multimedia	Opinions	-
2	0.054	Multimedia	Publishing	-
2	0.054	Multimedia	Privacy	-
2	0.045	Multimedia	Voting	-
2	0.045	Multimedia	Applications	-
2	0.036	Multimedia	News	-
2	0.027	Professional	Sharing	-
2	0.045	Professional	Profile	-
2	0.036	Professional	Opinions	-

Size	Support	Item 1	Item 2	Item 3
2	0.036	Professional	Privacy	-
2	0.036	Professional	Schedule	-
2	0.036	Professional	Applications	-
2	0.027	Entertainment	Opinions	-
2	0.036	Sharing	Opinions	-
2	0.036	Sharing	Publishing	-
2	0.027	Sharing	Applications	-
2	0.027	Sharing	News	-
2	0.045	Profile	Opinions	-
2	0.045	Profile	Publishing	-
2	0.045	Profile	Privacy	-
2	0.036	Profile	Voting	-
2	0.036	Profile	Applications	-
2	0.027	Profile	Promoting	-
2	0.027	Opinions	Schedule	-
2	0.027	Publishing	Privacy	-
2	0.027	Publishing	Applications	-
2	0.036	Privacy	Voting	-
2	0.045	Privacy	Applications	-
2	0.027	Voting	Applications	
2	0.027	Voting	News	-

Size	Support	Item 1	Item 2	Item 3
3	0.045	Connecting	Multimedia	Professional
3	0.027	Connecting	Multimedia	Entertainment
3	0.054	Connecting	Multimedia	Sharing
3	0.036	Connecting	Multimedia	Profile
3	0.027	Connecting	Multimedia	Publishing
3	0.054	Connecting	Multimedia	Privacy
3	0.036	Connecting	Multimedia	Voting
3	0.045	Connecting	Multimedia	Applications
3	0.036	Connecting	Multimedia	News
3	0.027	Connecting	Professional	Profile
3	0.027	Connecting	Professional	Privacy
3	0.036	Connecting	Professional	Applications
3	0.027	Connecting	Sharing	Publishing
3	0.027	Connecting	Sharing	Applications
3	0.027	Connecting	Sharing	News
3	0.027	Connecting	Profile	Publishing
3	0.045	Connecting	Profile	Privacy
3	0.027	Connecting	Profile	Voting
3	0.036	Connecting	Profile	Applications
3	0.027	Connecting	Publishing	Privacy
3	0.027	Connecting	Publishing	Applications

Size	Support	Item 1	Item 2	Item 3
3	0.036	Connecting	Privacy	Voting
3	0.045	Connecting	Privacy	Applications
3	0.027	Connecting	Voting	Applications
3	0.027	Connecting	Voting	News
3	0.027	Multimedia	Professional	Applications
3	0.027	Multimedia	Sharing	News
3	0.027	Multimedia	Profile	Publishing
3	0.036	Multimedia	Profile	Privacy
3	0.027	Multimedia	Profile	Voting
3	0.027	Multimedia	Profile	Applications
3	0.027	Multimedia	Privacy	Voting
3	0.036	Multimedia	Privacy	Applications
3	0.027	Multimedia	Voting	Applications
3	0.027	Professional	Opinions	Schedule
3	0.027	Professional	Privacy	Applications
3	0.027	Profile	Publishing	Privacy
3	0.027	Profile	Privacy	Voting
3	0.036	Profile	Privacy	Applications
3	0.027	Profile	Voting	Applications
3	0.027	Privacy	Voting	Applications

Appendix F

Table F1. Results from clustering with DBSCAN.

DBSCAN							
id	Cluster 0	id	Cluster 1	id	Cluster 2	id	Cluster 3
1	Facebook	2	YouTube	16	WhatsApp	25	WeChat
3	Instagram	6	Vine	17	Messenger	46	DevianArt
4	Twitter	34	Stumbleupon	23	AngelList	58	Flixster
5	Reddit	55	Uplike	24	Kickstarter	59	Gaia Online
7	Pinterest	65	Tinder	26	Skype	79	Wayn
8	Ask.fm	97	Ravelry	27	Viber	81	TravBuddy
9	Tumblr	-	-	28	Viadeo	83	Cellufun
10	Flickr	-	-	31	Tagged	92	YY
11	Google+	-	-	32	Myspace	95	VampireFreaks
12	LinkedIn	-	-	33	Badoo	98	ASmallWorld
13	VK	-	-	36	MeetMe	105	Zynga
14	ClassMates	-	-	40	Twoo	106	Habbo
15	Meetup	-	-	42	Snapfish	107	FunnyOrDie
18	Snapchat	-	-	43	Photobucket	111	MocoSpace
19	Quora	-	-	44	Shutterfly	-	-
20	GirlsAskGuys	-	-	45	500px	-	-
21	Nextdoor	-	-	47	Dronestagram	-	-
22	ProductHunt	-	-	48	Fotki	-	-
29	Gab	-	-	51	Pixabay	-	-
30	Telegram	-	-	52	WeHeartIt	-	-
35	Foursquare	-	-	57	Cross.tv	-	-
37	Skyrock A192	-	-	61	MyMFB	-	-
38	Pinboard	-	-	62	Care2	-	-
39	Kiwibox	-	-	63	CaringBridge	-	-

DBSCAN							
41	Yelp	-	-	64	GoFundMe	-	-
49	Fotolog	-	-	66	Crokes	-	-
50	Imgur	-	-	68	Internations	-	-
53	43Things	-	-	69	PlentyofFish	-	-
54	Path	-	-	73	Academia.edu	-	-
56	Last.fm	-	-	74	Busuu	-	-
60	BlackPlanet	-	-	75	English, baby!	-	-
67	Goodreads	-	-	76	Italki.com	-	-
70	Minds	-	-	78	Doximity	-	-
71	Nexopia	-	-	84	23andMe	-	-
72	Glocals	-	-	91	Line	-	-
77	Untappd	-	-	93	Sprybirds	-	-
80	CouchSurfing	-	-	96	CafeMom	-	-
82	Tournac	-	-	102	eToro	-	-
85	Plurk	-	-	104	Ryze	-	-
86	LiveJournal	-	-	108	Tout	-	-
87	Weibo	-	-	112	Ancestry.com	-	-
88	Qzone	-	-	-	-	-	-
89	QQ	-	-	-	-	-	-
90	Baidu	-	-	-	-	-	-
94	Xing	-	-	-	-	-	-
99	ReverbNation	-	-	-	-	-	-
100	SoundCloud	-	-	-	-	-	-
101	Solaborate	-	-	-	-	-	-
103	Xanga	-	-	-	-	-	-
109	Classmates	-	-	-	-	-	-
110	MyHeritage	-	-	-	-	-	-

Table F2. Results from clustering with k-medoids.

k-medoids (k=4)							
id	Cluster 0	id	Cluster 1	id	Cluster 2	id	Cluster 3
1	Facebook	2	YouTube	10	Flickr	12	LinkedIn
25	WeChat	3	Instagram	13	VK	18	Snapchat
39	Kiwibox	4	Twitter	15	Meetup	19	Quora
46	DevianArt	5	Reddit	16	WhatsApp	30	Telegram
56	Last.fm	6	Vine	17	Messenger	38	Pinboard
58	Flixster	7	Pinterest	21	Nextdoor	86	LiveJournal
59	Gaia Online	8	Ask.fm	22	ProductHunt	88	Qzone
67	Goodreads	9	Tumblr	23	AngelList	94	Xing
79	Wayn	11	Google+	24	Kickstarter	101	Solaborate
80	CouchSurfing	14	ClassMates	26	Skype	103	Xanga
81	TravBuddy	20	GirlsAskGuys	27	Viber	110	MyHeritage
82	Tournac	34	Stumbleupon	28	Viadeo	-	-
83	Cellufun	35	Foursquare	29	Gab	-	-
89	QQ	53	43Things	31	Tagged	-	-
92	YY	55	Uplike	32	Myspace	-	-
95	VampireFreaks	65	Tinder	33	Badoo	-	-
98	ASmallWorld	85	Plurk	36	MeetMe	-	-
99	ReverbNation	87	Weibo	37	Skyrock A192	-	-
100	SoundCloud	90	Baidu	40	Twoo	-	-
105	Zynga	97	Ravelry	41	Yelp	-	-
106	Habbo	-	-	42	Snapfish	-	-
107	FunnyOrDie	-	-	43	Photobucket	-	-
111	MocoSpace	-	-	44	Shutterfly	-	-
-	-	-	-	45	500px	-	-
-	-	-	-	47	Dronestagram	-	-
-	-	-	-	48	Fotki	-	-

k-medoids (k=4)							
-	-	-	-	49	Fotolog	-	-
-	-	-	-	50	Imgur	-	-
-	-	-	-	51	Pixabay	-	-
-	-	-	-	52	WeHeartIt	-	-
-	-	-	-	54	Path	-	-
-	-	-	-	57	Cross.tv	-	-
-	-	-	-	60	BlackPlanet	-	-
-	-	-	-	61	MyMFB	-	-
-	-	-	-	62	Care2	-	-
-	-	-	-	63	CaringBridge	-	-
-	-	-	-	64	GoFundMe	-	-
-	-	-	-	66	Crokes	-	-
-	-	-	-	68	Internations	-	-
-	-	-	-	69	PlentyofFish	-	-
-	-	-	-	70	Minds	-	-
-	-	-	-	71	Nexopia	-	-
-	-	-	-	72	Glocals	-	-
-	-	-	-	73	Academia.edu	-	-
-	-	-	-	74	Bussu	-	-
-	-	-	-	75	English, baby!	-	-
-	-	-	-	76	Italki.com	-	-
-	-	-	-	77	Untappd	-	-
-	-	-	-	78	Doximity	-	-
-	-	-	-	84	23andMe	-	-
-	-	-	-	91	Line	-	-
-	-	-	-	93	Sprybirds	-	-
-	-	-	-	96	CafeMom	-	-
-	-	-	-	102	eToro	-	-
-	-	-	-	104	Ryze	-	-

k-medoids (k=4)							
-	-	-	-	108	Tout	-	-
-	-	-	-	109	Classmates	-	-
-	-	-	-	112	Ancestry.com	-	-

Appendix G

Table G1. 20 most common words per topic for LDA simulation_no=1.

Rank	Topic 0 words	Topic 0 weights	Topic 1 words	Topic 1 weights	Topic 2 words	Topic 2 weights	Topic 3 words	Topic 3 weights	Topic 4 words	Topic 4 weights	Topic 5 words	Topic 5 weights	Topic 6 words	Topic 6 weights	Topic 7 words	Topic 7 weights	Topic 8 words	Topic 8 weights	Topic 9 words	Topic 9 weights
1	covid	24884	coronavirus	97834	pandemic	29520	home	78711	like	53379	case	175921	need	82681	people	57381	virus	56481	close	35222
2	hand	24862	outbreak	23294	impact	20776	stay	70744	time	44102	new	97337	help	48615	die	42803	mask	37971	school	25535
3	corona	15035	read	17661	covid	20384	work	56798	know	34043	test	74308	fight	42479	days	30492	trump	32767	service	19689
4	wash	14589	pandemic	17564	business	19805	safe	36521	go	33540	deaths	66692	people	37625	infect	21544	spread	28767	measure	18411
5	vaccine	14339	say	15758	market	17994	time	20496	social	31223	report	57204	health	28402	patients	16445	china	26661	order	18382
6	ms	13098	covid	15148	new	16528	test	16538	look	31207	total	48146	world	25334	symptoms	15221	stop	24581	public	17740
7	dr	12896	question	14807	company	14765	get	14618	think	30573	number	43930	crisis	22265	kill	15074	people	22499	lockdown	17456
8	clean	9194	ill	14584	crisis	13455	family	13733	watch	30250	positive	41477	pandemic	19457	covid	13565	dont	22174	spread	16486
9	virus	8490	amid	14224	help	13174	quarantine	11941	good	29796	confirm	41127	time	18425	virus	12572	news	21813	open	16004
10	drug	8472	uk	13445	team	13073	individuals	11262	day	26768	coronavirus	38979	support	18286	disease	12537	corona	18651	pay	15023
11	est	6423	minister	12343	support	12964	people	11107	people	26198	update	35087	medical	18008	say	11888	check	16649	people	14286
12	pandemia	6368	news	12199	learn	12264	let	11001	live	24323	death	32060	thank	17460	hand	11739	say	15581	government	11818
13	go	5521	latest	12185	global	12144	help	10832	dont	22888	covid	27159	care	16722	person	11197	wear	15523	place	10877
14	unavailable	5362	travel	11788	economic	11671	im	10715	distance	22381	state	24057	workers	15606	risk	11140	chinese	15189	food	10729
15	youth	5134	countries	10952	die	11466	day	10506	come	20529	recover	20658	money	15234	care	11055	think	14114	state	10234
16	earn	5118	sign	10768	response	11267	advice	10039	life	19270	rise	19168	government	14289	spread	11045	f*ck	13611	shop	9481
17	cure	5053	update	10526	pm	11238	self	9910	im	17741	italy	17319	country	12959	doctor	10064	go	13426	store	8959
18	ecoins	4985	flight	10424	businesses	10757	dont	8797	happen	17246	toll	16794	doctor	12902	infection	9783	face	12967	shut	8889
19	gone	4890	answer	9525	work	10700	away	8742	right	16013	india	16669	paper	11132	cough	9066	know	12641	pour	8886
20	hate	4812	check	7983	plan	10549	tip	8567	get	15797	health	15437	protect	10485	second	8711	buy	12373	students	8880

Table G2. 20 most common words per topic for LDA simulation_no=2.

Rank	Topic 0 words	Topic 0 weights	Topic 1 words	Topic 1 weights	Topic 2 words	Topic 2 weights	Topic 3 words	Topic 3 weights	Topic 4 words	Topic 4 weights	Topic 5 words	Topic 5 weights	Topic 6 words	Topic 6 weights	Topic 7 words	Topic 7 weights	Topic 8 words	Topic 8 weights	Topic 9 words	Topic 9 weights
1	die	57325	good	38502	people	52825	coronavirus	79194	say	34340	time	35665	help	39187	home	78470	covid	21780	case	175922
2	people	38207	lockdown	35033	trump	40886	china	42523	people	32111	come	28982	crisis	31208	stay	70738	ill	16821	new	105304
3	im	17694	close	34359	dont	40688	news	39032	health	30884	need	28423	support	30851	work	43022	ms	13098	test	92928
4	man	17168	social	33502	think	36477	world	37090	live	25328	thank	28117	free	24199	mask	38539	cure	12093	deaths	66655
5	infect	14478	time	30177	know	34130	pandemic	35550	medical	23043	let	27850	need	24033	safe	38221	coronavirus	9839	report	52008
6	year	14433	school	25376	like	31250	outbreak	33467	doctor	20193	like	27339	pandemic	22122	face	24634	est	9836	total	47841
7	house	11887	distance	23947	want	22655	spread	26460	patients	18678	help	21934	business	20117	corona	21836	past	9642	number	43842
8	paper	11132	go	23560	go	22451	read	25658	pm	17420	right	21159	impact	19398	spread	21191	non	9015	positive	42096
9	toilet	9984	day	22481	question	19312	virus	22514	minister	17261	fight	20831	covid	17363	hand	20726	pour	8886	confirm	41127
10	old	9595	take	18858	say	19222	covid	19824	save	17249	world	20768	service	17259	protect	20374	check	7983	coronavirus	39431
11	wait	9462	weeks	17345	hand	19094	global	19527	care	17009	love	20743	new	14590	virus	17436	day	7266	update	37001
12	vote	8552	people	16470	watch	18367	market	19439	spread	16101	look	19429	businesses	14176	wear	15727	plus	6781	death	33540
13	self	8388	days	16114	kill	17747	countries	15223	fight	16030	know	19284	fund	13612	people	15155	pandemia	6369	covid	27211
14	men	8312	open	15536	get	17629	chinese	14796	virus	15391	im	18944	emergency	13590	dont	14167	go	5959	state	20900
15	white	7378	years	13369	real	16846	amid	14200	covid	15322	share	18498	company	13002	follow	13650	place	5815	recover	20705
16	additions	7024	week	11459	tell	15314	trump	13476	risk	15152	get	18237	offer	12959	wash	12921	im	5815	rise	17366
17	right	7017	today	11287	make	15092	warn	13259	government	14775	hope	18015	economy	12853	need	12370	sure	5375	march	16473
18	flu	6683	morning	11058	try	14817	vaccine	12924	india	14164	feel	16502	coronavirus	12663	buy	11217	others	5342	toll	15924
19	f*ck	6534	stop	10684	need	14639	say	12586	disease	12853	great	14693	people	12584	help	10342	inside	5323	rate	15317
20	virus	6526	start	10543	ask	14515	fear	11875	treat	10645	video	13068	response	12541	avoid	9758	youth	5134	record	13576

Table G3. 20 most common words per topic for LDA simulation_no=3.

Rank	Topic 0 words	Topic 0 weights	Topic 1 words	Topic 1 weights	Topic 2 words	Topic 2 weights	Topic 3 words	Topic 3 weights	Topic 4 words	Topic 4 weights	Topic 5 words	Topic 5 weights	Topic 6 words	Topic 6 weights	Topic 7 words	Topic 7 weights	Topic 8 words	Topic 8 weights	Topic 9 words	Topic 9 weights
1	people	78951	stay	70688	spread	56075	test	92899	close	28030	world	43763	virus	56842	time	78946	work	78334	case	175922
2	get	48693	home	55192	health	44307	say	32755	support	27973	china	32739	news	42270	need	57245	coronavirus	40953	new	100676
3	like	36208	hand	39831	social	33625	positive	28060	school	25478	new	19543	trump	42201	help	49637	die	23139	deaths	66693
4	go	34594	mask	38540	public	28807	pandemic	23738	open	19067	lockdown	16195	corona	30625	people	35385	covid	20147	report	57549
5	dont	31795	safe	37751	distance	25407	covid	22081	covid	18390	uk	14286	watch	27106	look	27713	home	15448	total	50897
6	im	27834	face	25968	measure	21049	medical	19930	business	17571	people	14117	live	23990	like	24544	ms	13098	number	44399
7	die	27074	wash	19902	people	19665	coronavirus	19383	fight	17198	hit	13550	know	21016	thank	24008	company	11605	coronavirus	41893
8	know	25957	buy	16169	travel	18803	patients	18972	ill	16821	government	13542	read	20736	share	23966	force	9976	confirm	41127
9	think	22968	wear	15727	minister	17723	workers	17620	pm	15975	time	13493	market	19619	love	19597	employees	9750	death	36727
10	kill	21765	dont	12816	outbreak	17704	need	17520	service	14953	come	13407	media	16693	think	18292	self	9382	update	35474
11	quarantine	20173	protect	10874	lockdown	17117	dr	16423	online	14600	country	13320	president	14301	hope	17513	person	8568	covid	33269
12	let	19823	use	10215	risk	16147	question	16051	latest	14512	end	12898	good	12699	know	17403	men	8312	recover	20719
13	right	17413	touch	10171	say	15324	care	15586	businesses	13719	pandemic	12712	video	12449	great	16774	cure	7450	rise	18755
14	way	16673	people	9540	coronavirus	15117	response	15147	update	13095	coronavirus	11510	lie	10707	life	16676	isolate	7161	toll	16795
15	live	16337	clean	9210	government	13943	crisis	14558	join	12910	paper	11125	pandemic	10642	feel	16027	additions	7024	italy	15762
16	save	15997	panic	9162	amid	13929	trump	12636	coronavirus	12760	war	10927	like	10524	good	15546	pandemia	6369	data	15053
17	happen	15974	avoid	8638	cancel	13643	health	12372	free	12675	state	10852	coronavirus	10502	day	15418	im	6118	rate	14997
18	thing	15904	follow	8492	take	13299	state	12064	help	12391	city	10589	impact	10101	learn	14158	meet	5447	state	14058
19	want	14963	water	8392	prevent	11906	result	11691	check	12289	toilet	9957	chinese	9911	things	13780	unavailable	5356	positive	14040
20	flu	13972	shop	8311	state	11698	ask	10844	march	11269	crisis	9187	want	9785	important	13599	others	5342	india	13839

Table G4. 20 most common words per topic for LDA simulation_no=4.

Rank	Topic 0 words	Topic 0 weights	Topic 1 words	Topic 1 weights	Topic 2 words	Topic 2 weights	Topic 3 words	Topic 3 weights	Topic 4 words	Topic 4 weights	Topic 5 words	Topic 5 weights	Topic 6 words	Topic 6 weights	Topic 7 words	Topic 7 weights	Topic 8 words	Topic 8 weights	Topic 9 words	Topic 9 weights
1	time	77313	help	50796	world	56633	home	78063	question	19783	test	100147	coronavirus	81292	dont	67133	covid	38634	case	175905
2	like	41220	work	49432	trump	54683	people	74285	cancel	18208	positive	37004	spread	55705	people	40222	health	32913	new	103876
3	look	30423	fight	45783	know	39104	stay	70741	get	17833	state	34796	lockdown	34045	say	34849	read	26924	deaths	66693
4	day	21163	hand	39343	right	32310	die	55660	week	16150	say	28178	social	33630	im	32842	risk	24755	report	55675
5	come	21049	need	33602	china	31424	safe	34103	days	14224	pm	25214	distance	25897	think	25155	check	22438	total	50897
6	good	20992	support	30249	live	29162	virus	22782	ask	13715	india	19797	outbreak	20404	know	25116	doctor	21419	number	44066
7	best	20268	thank	29924	watch	25672	work	19088	ms	13098	march	19642	measure	19722	want	19669	latest	20417	coronavirus	41656
8	market	19676	mask	27261	go	25616	infect	18986	answer	11005	crisis	18108	amid	18920	like	18818	information	19951	confirm	41127
9	today	18472	face	26288	people	22209	quarantine	18961	months	9594	minister	17683	public	17364	love	18577	update	18680	death	37177
10	people	18353	pandemic	25565	president	19348	corona	17665	air	9170	close	16707	follow	15771	year	18382	patients	18304	covid	35783
11	need	15603	crisis	20675	virus	19322	ill	16821	light	8711	health	15150	prevent	15608	get	17884	advice	17026	update	31880
12	buy	15533	wash	19889	way	17297	live	11737	flight	8151	april	14490	order	12346	virus	15161	impact	15837	recover	20718
13	great	15487	workers	19854	need	16838	save	10946	news	8032	government	14298	covid	12205	flu	14471	coronavirus	15071	italy	19487
14	stock	14953	people	19015	pandemic	16666	man	10617	tweet	7759	try	12617	avoid	11616	go	14316	care	14386	rise	19197
15	life	14556	pay	18857	happen	15558	stop	10426	postpone	7747	announce	12419	government	11203	tell	13915	medical	13519	toll	16799
16	hope	14530	join	16288	like	15314	f*ck	10094	wait	7435	plan	11427	pandemic	10805	worry	13791	pandemic	13091	news	15416
17	go	13123	protect	14866	let	15162	self	9964	track	7403	covid	11392	travel	10374	hear	12706	share	12249	rate	15004
18	feel	12517	food	14261	end	14944	non	9015	link	7074	government	10418	close	9815	kid	12685	article	12037	china	14124
19	online	11881	company	13018	lie	14843	healthy	8724	events	6639	result	10409	media	9742	individuals	12577	uk	11935	record	13704
20	help	11679	donate	12317	time	14799	person	8568	coronavirus	6528	school	9821	take	9416	house	11841	important	11898	south	12878

Table G5. 20 most common words per topic for LDA simulation_no=5.

Rank	Topic 0 words	Topic 0 weights	Topic 1 words	Topic 1 weights	Topic 2 words	Topic 2 weights	Topic 3 words	Topic 3 weights	Topic 4 words	Topic 4 weights	Topic 5 words	Topic 5 weights	Topic 6 words	Topic 6 weights	Topic 7 words	Topic 7 weights	Topic 8 words	Topic 8 weights	Topic 9 words	Topic 9 weights
1	coronavirus	97408	new	70554	work	52803	pandemic	74917	case	158748	people	79625	time	67393	close	36453	test	93419	stay	70601
2	covid	88860	day	32103	come	33210	crisis	37761	deaths	60990	virus	76865	like	49124	die	32494	positive	28183	home	53343
3	live	39322	state	27134	think	25471	health	27696	report	53223	trump	50604	im	42410	watch	30149	social	23842	hand	39831
4	update	31125	coronavirus	25766	people	25203	response	25725	total	50897	know	40046	get	42382	school	25570	support	22106	safe	38183
5	read	29171	china	24171	go	25086	world	23876	new	49314	corona	33760	people	29892	pm	23842	people	22016	spread	35473
6	latest	23191	italy	23218	need	24780	good	22929	confirm	41002	say	27434	pay	24708	lockdown	23502	service	19032	mask	35104
7	impact	20320	news	18954	thank	17485	change	20910	number	40367	dont	25802	feel	22075	video	22895	medical	18417	protect	27861
8	spread	17853	march	17243	great	17372	look	19450	death	30334	kill	22677	care	20906	government	21783	travel	18099	help	21617
9	economy	17819	case	17174	buy	17295	global	17987	recover	20720	world	22370	leave	20432	take	16815	distance	17377	face	21379
10	news	16364	try	16355	house	17050	listen	14562	health	20685	call	21310	go	17581	coronavirus	13919	provide	17053	wash	19902
11	help	15772	hit	15433	food	17047	need	13174	update	18478	right	20072	work	16628	vaccine	13491	government	16960	dont	17873
12	outbreak	15018	april	14376	time	16296	outbreak	13062	coronavirus	17540	go	20013	workers	16421	plan	13350	staff	16036	use	15023
13	save	14394	south	14038	job	15914	lead	12395	rise	16645	question	19612	dont	15932	measure	13001	fight	15923	prevent	14950
14	world	12417	record	13598	run	15189	link	12195	covid	16271	die	19163	sick	15487	man	11520	help	15467	stop	13858
15	daily	11463	days	12987	stock	14915	time	11531	positive	13916	want	17702	need	14210	year	10666	sign	14650	wear	13831
16	prepare	10849	countries	11422	company	14615	best	10804	toll	13236	think	17115	f*ck	13604	say	10321	government	14620	family	13386
17	urge	10088	lockdown	9992	open	14161	fight	10531	hours	12546	china	17060	look	12666	no	10063	patients	14389	ms	13098
18	information	10005	york	9813	weeks	14046	read	10298	india	12496	ill	16821	home	12498	est	9829	india	13747	avoid	12115
19	need	9959	uk	9573	years	13829	risk	10099	reach	12309	stop	15674	right	12471	play	9698	doctor	13593	love	11797
20	cause	9647	city	9366	happen	13647	current	9735	bring	12163	lie	15528	worry	12426	quarantine	8954	uk	12933	share	11539

Appendix H

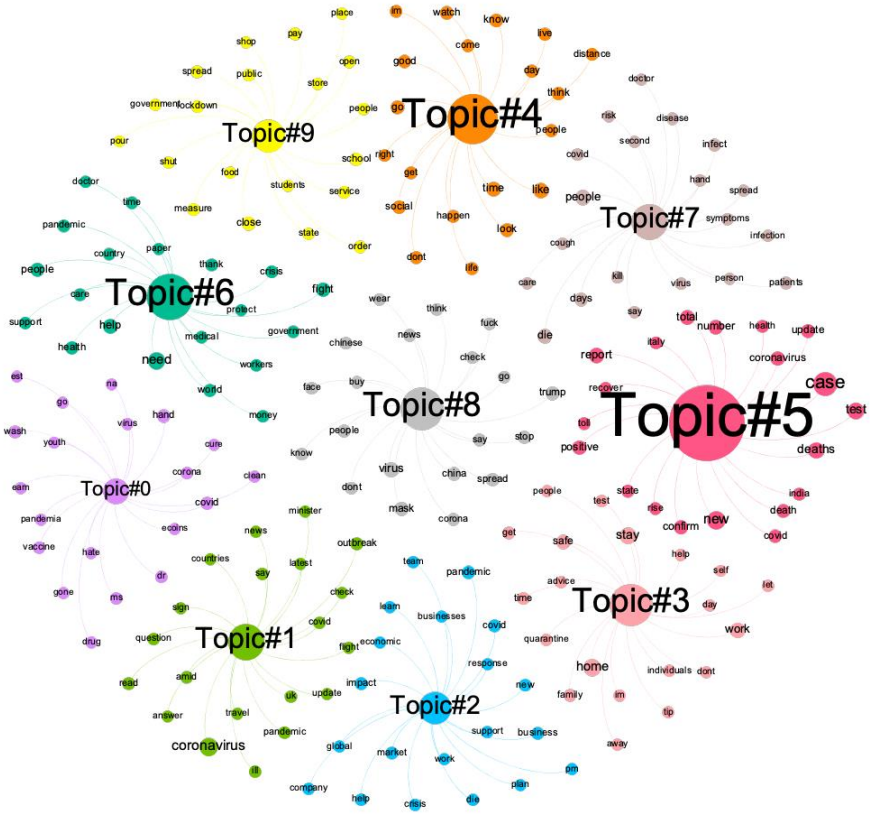

Fig. H1. LDA filtering with 2713 features.

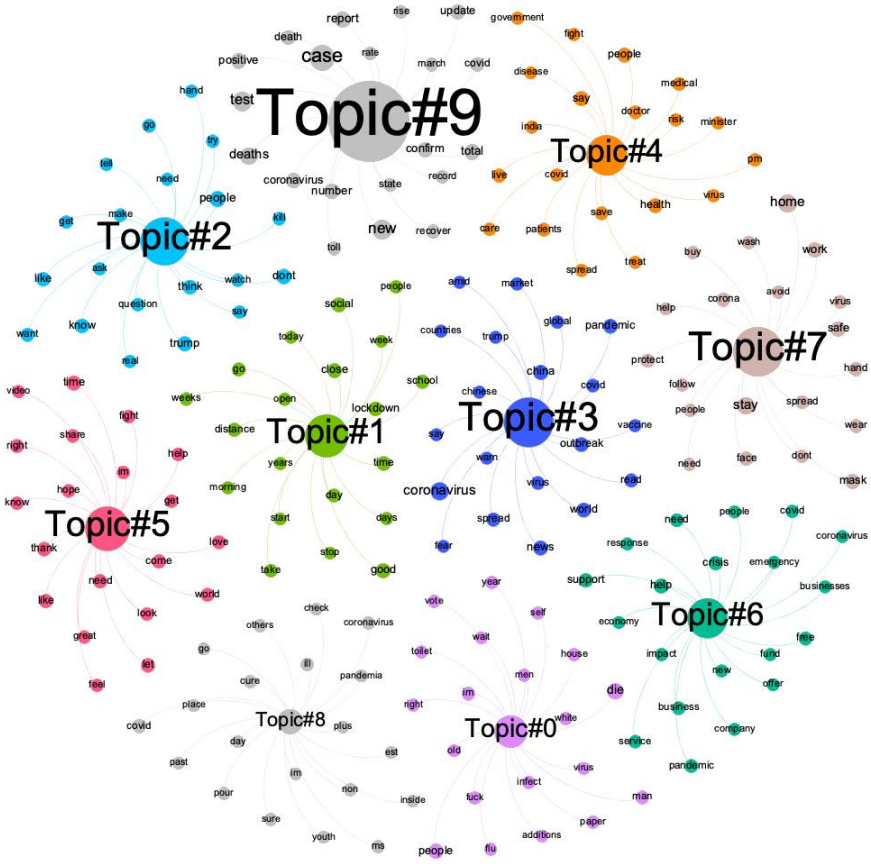

Fig. H2. LDA filtering with 1597 features.

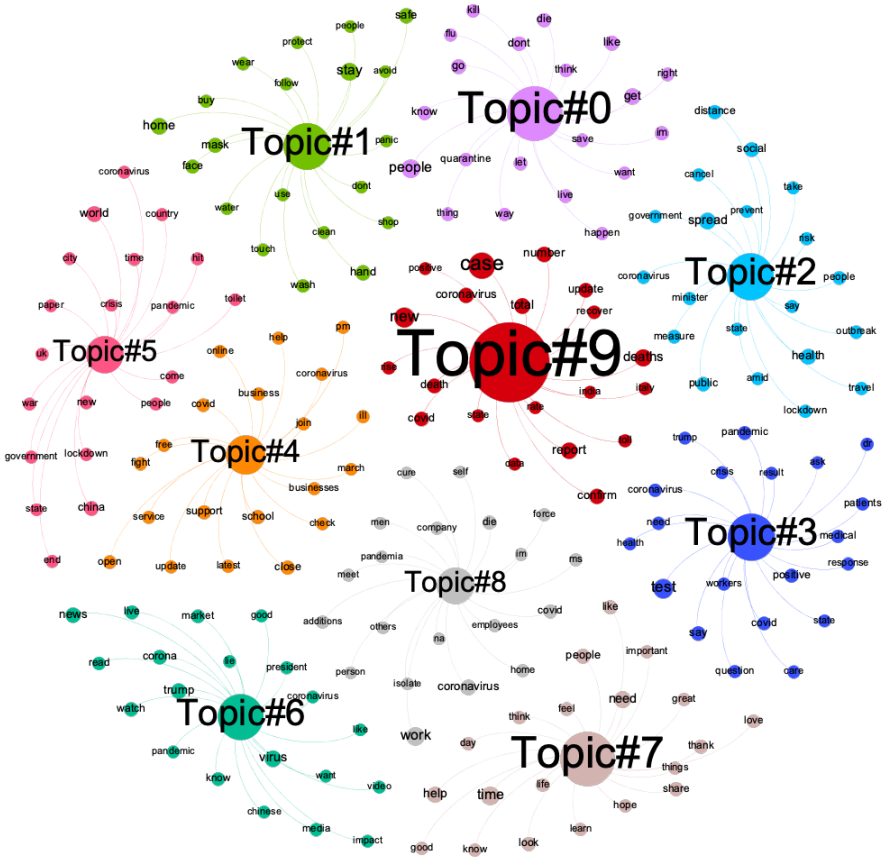

Fig H3. LDA filtering with 1153 features.

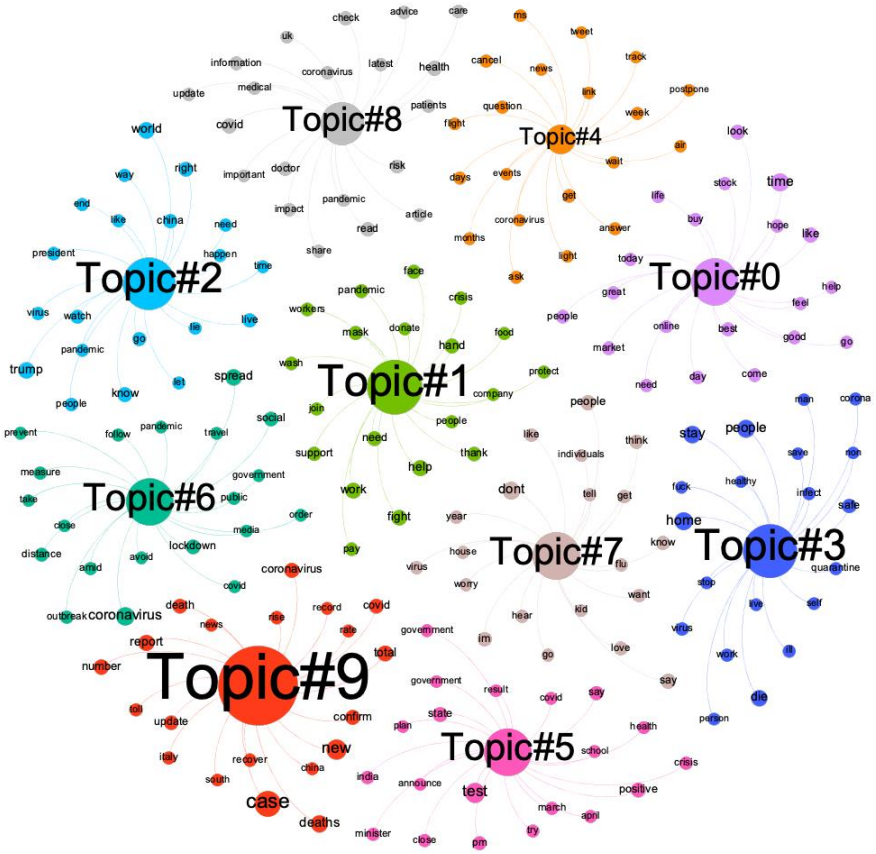

Fig. H4. LDA filtering with 886 features.

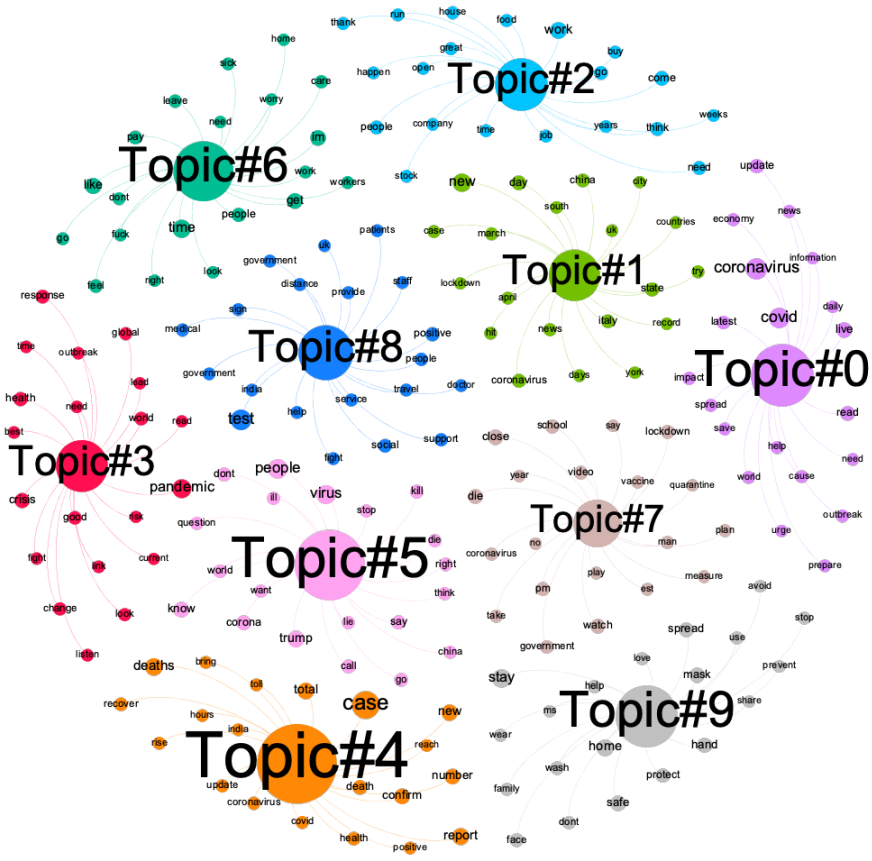

Fig. H5. LDA filtering with 713 features.

Appendix I

Table I1. Top 138 rules showing the strongest leverage, support, confidence and lift.

antecedents	consequents	antecedent supp	consequent supp	supp	conf	lift	leverage
coronavirus	covid	0,42	0,40	0,30	0,71	1,79	0,132
covid	coronavirus	0,40	0,42	0,30	0,75	1,79	0,132
covid, coronavirus	update	0,30	0,18	0,18	0,60	3,33	0,126
update	covid, coronavirus	0,18	0,30	0,18	1,00	3,33	0,126
know	like	0,20	0,20	0,16	0,80	4,00	0,120
like	know	0,20	0,20	0,16	0,80	4,00	0,120
know, people	think	0,16	0,16	0,14	0,88	5,47	0,114
think	know, people	0,16	0,16	0,14	0,88	5,47	0,114
covid	update	0,40	0,18	0,18	0,45	2,50	0,108
covid	update, coronavirus	0,40	0,18	0,18	0,45	2,50	0,108
know	think	0,20	0,16	0,14	0,70	4,38	0,108
know	think, people	0,20	0,16	0,14	0,70	4,38	0,108
think	know	0,16	0,20	0,14	0,88	4,38	0,108
think, people	know	0,16	0,20	0,14	0,88	4,38	0,108
update	covid	0,18	0,40	0,18	1,00	2,50	0,108
update, coronavirus	covid	0,18	0,40	0,18	1,00	2,50	0,108
go, people	think	0,22	0,16	0,14	0,64	3,98	0,105
think	go, people	0,16	0,22	0,14	0,88	3,98	0,105
coronavirus	update	0,42	0,18	0,18	0,43	2,38	0,104
coronavirus	update, covid	0,42	0,18	0,18	0,43	2,38	0,104
update	coronavirus	0,18	0,42	0,18	1,00	2,38	0,104
update, covid	coronavirus	0,18	0,42	0,18	1,00	2,38	0,104
know, go	think, dont	0,14	0,12	0,12	0,86	7,14	0,103
think, go	know, dont	0,14	0,12	0,12	0,86	7,14	0,103
think, go, people	know, dont	0,14	0,12	0,12	0,86	7,14	0,103
know, go, people	think, dont	0,14	0,12	0,12	0,86	7,14	0,103
think, go	know, dont, people	0,14	0,12	0,12	0,86	7,14	0,103
know, go	think, dont, people	0,14	0,12	0,12	0,86	7,14	0,103
know, dont	think, go	0,12	0,14	0,12	1,00	7,14	0,103
think, dont	know, go	0,12	0,14	0,12	1,00	7,14	0,103
think, dont, people	know, go	0,12	0,14	0,12	1,00	7,14	0,103
know, dont, people	think, go	0,12	0,14	0,12	1,00	7,14	0,103
think, dont	know, go, people	0,12	0,14	0,12	1,00	7,14	0,103
know, dont	think, go, people	0,12	0,14	0,12	1,00	7,14	0,103
think	know, dont	0,16	0,12	0,12	0,75	6,25	0,101
think	know, go, dont	0,16	0,12	0,12	0,75	6,25	0,101
know, people	think, dont	0,16	0,12	0,12	0,75	6,25	0,101
think, people	know, dont	0,16	0,12	0,12	0,75	6,25	0,101
think	know, dont, people	0,16	0,12	0,12	0,75	6,25	0,101
think, people	know, go, dont	0,16	0,12	0,12	0,75	6,25	0,101
know, people	think, go, dont	0,16	0,12	0,12	0,75	6,25	0,101
think	know, go, dont, people	0,16	0,12	0,12	0,75	6,25	0,101
new, coronavirus	case	0,16	0,12	0,12	0,75	6,25	0,101
know, dont	think	0,12	0,16	0,12	1,00	6,25	0,101
know, go, dont	think	0,12	0,16	0,12	1,00	6,25	0,101
know, dont, people	think	0,12	0,16	0,12	1,00	6,25	0,101
know, dont	think, people	0,12	0,16	0,12	1,00	6,25	0,101
think, dont	know, people	0,12	0,16	0,12	1,00	6,25	0,101
know, go, dont, people	think	0,12	0,16	0,12	1,00	6,25	0,101
think, go, dont	know, people	0,12	0,16	0,12	1,00	6,25	0,101
know, go, dont	think, people	0,12	0,16	0,12	1,00	6,25	0,101
case	new, coronavirus	0,12	0,16	0,12	1,00	6,25	0,101
know, think	go, dont	0,14	0,14	0,12	0,86	6,12	0,100
go, dont	know, think	0,14	0,14	0,12	0,86	6,12	0,100
go, dont, people	know, think	0,14	0,14	0,12	0,86	6,12	0,100

know, think, people	go, dont	0,14	0,14	0,12	0,86	6,12	0,100
go, dont	know, think, people	0,14	0,14	0,12	0,86	6,12	0,100
know, think	go, dont, people	0,14	0,14	0,12	0,86	6,12	0,100
go	like, people	0,26	0,16	0,14	0,54	3,37	0,098
go	know, people	0,26	0,16	0,14	0,54	3,37	0,098
go	think	0,26	0,16	0,14	0,54	3,37	0,098
go	think, people	0,26	0,16	0,14	0,54	3,37	0,098
pandemic	crisis	0,26	0,16	0,14	0,54	3,37	0,098
like, people	go	0,16	0,26	0,14	0,88	3,37	0,098
know, people	go	0,16	0,26	0,14	0,88	3,37	0,098
think	go	0,16	0,26	0,14	0,88	3,37	0,098
think, people	go	0,16	0,26	0,14	0,88	3,37	0,098
crisis	pandemic	0,16	0,26	0,14	0,88	3,37	0,098
new	case	0,18	0,12	0,12	0,67	5,56	0,098
new	case, coronavirus	0,18	0,12	0,12	0,67	5,56	0,098
case	new	0,12	0,18	0,12	1,00	5,56	0,098
case, coronavirus	new	0,12	0,18	0,12	1,00	5,56	0,098
know, people	go, dont	0,16	0,14	0,12	0,75	5,36	0,098
think	know, go	0,16	0,14	0,12	0,75	5,36	0,098
know, people	think, go	0,16	0,14	0,12	0,75	5,36	0,098
think, people	know, go	0,16	0,14	0,12	0,75	5,36	0,098
think	know, go, people	0,16	0,14	0,12	0,75	5,36	0,098
think	go, dont	0,16	0,14	0,12	0,75	5,36	0,098
think, people	go, dont	0,16	0,14	0,12	0,75	5,36	0,098
think	go, dont, people	0,16	0,14	0,12	0,75	5,36	0,098
go, dont	know, people	0,14	0,16	0,12	0,86	5,36	0,098
know, go	think	0,14	0,16	0,12	0,86	5,36	0,098
know, go, people	think	0,14	0,16	0,12	0,86	5,36	0,098
know, go	think, people	0,14	0,16	0,12	0,86	5,36	0,098
think, go	know, people	0,14	0,16	0,12	0,86	5,36	0,098
go, dont	think	0,14	0,16	0,12	0,86	5,36	0,098
go, dont, people	think	0,14	0,16	0,12	0,86	5,36	0,098
go, dont	think, people	0,14	0,16	0,12	0,86	5,36	0,098
know	think, dont	0,20	0,12	0,12	0,60	5,00	0,096
know	think, go, dont	0,20	0,12	0,12	0,60	5,00	0,096
know	think, dont, people	0,20	0,12	0,12	0,60	5,00	0,096
dont, people	know, think, go	0,20	0,12	0,12	0,60	5,00	0,096
know	think, go, dont, people	0,20	0,12	0,12	0,60	5,00	0,096
go, people	like	0,22	0,20	0,14	0,64	3,18	0,096
go, people	know	0,22	0,20	0,14	0,64	3,18	0,096
like	go, people	0,20	0,22	0,14	0,70	3,18	0,096
know	go, people	0,20	0,22	0,14	0,70	3,18	0,096
think, dont	know	0,12	0,20	0,12	1,00	5,00	0,096
think, go, dont	know	0,12	0,20	0,12	1,00	5,00	0,096
think, dont, people	know	0,12	0,20	0,12	1,00	5,00	0,096
think, go, dont, people	know	0,12	0,20	0,12	1,00	5,00	0,096
know, think, go	dont, people	0,12	0,20	0,12	1,00	5,00	0,096
go, people	know, dont	0,22	0,12	0,12	0,55	4,55	0,094
dont	get, people	0,22	0,12	0,12	0,55	4,55	0,094
dont	know, think, go	0,22	0,12	0,12	0,55	4,55	0,094
go, people	think, dont	0,22	0,12	0,12	0,55	4,55	0,094
go, people	know, think, dont	0,22	0,12	0,12	0,55	4,55	0,094
dont	know, think, go, people	0,22	0,12	0,12	0,55	4,55	0,094
know, dont	go, people	0,12	0,22	0,12	1,00	4,55	0,094
get, people	dont	0,12	0,22	0,12	1,00	4,55	0,094
know, think, go	dont	0,12	0,22	0,12	1,00	4,55	0,094
think, dont	go, people	0,12	0,22	0,12	1,00	4,55	0,094
know, think, go, people	dont	0,12	0,22	0,12	1,00	4,55	0,094
know, think, dont	go, people	0,12	0,22	0,12	1,00	4,55	0,094
need	time	0,28	0,24	0,16	0,57	2,38	0,093

time	need	0,24	0,28	0,16	0,67	2,38	0,093
know	go, dont	0,20	0,14	0,12	0,60	4,29	0,092
dont, people	know, go	0,20	0,14	0,12	0,60	4,29	0,092
know	go, dont, people	0,20	0,14	0,12	0,60	4,29	0,092
know	think, go	0,20	0,14	0,12	0,60	4,29	0,092
know	think, go, people	0,20	0,14	0,12	0,60	4,29	0,092
dont, people	think, go	0,20	0,14	0,12	0,60	4,29	0,092
dont, people	know, think	0,20	0,14	0,12	0,60	4,29	0,092
go, dont	know	0,14	0,20	0,12	0,86	4,29	0,092
go, dont, people	know	0,14	0,20	0,12	0,86	4,29	0,092
know, go	dont, people	0,14	0,20	0,12	0,86	4,29	0,092
think, go	know	0,14	0,20	0,12	0,86	4,29	0,092
think, go, people	know	0,14	0,20	0,12	0,86	4,29	0,092
think, go	dont, people	0,14	0,20	0,12	0,86	4,29	0,092
know, think	dont, people	0,14	0,20	0,12	0,86	4,29	0,092
go, people	dont	0,22	0,22	0,14	0,64	2,89	0,092
dont	go, people	0,22	0,22	0,14	0,64	2,89	0,092
time	look	0,24	0,12	0,12	0,50	4,17	0,091
virus	corona	0,24	0,12	0,12	0,50	4,17	0,091
im	get	0,18	0,16	0,12	0,67	4,17	0,091
get	im	0,16	0,18	0,12	0,75	4,17	0,091
look	time	0,12	0,24	0,12	1,00	4,17	0,091
corona	virus	0,12	0,24	0,12	1,00	4,17	0,091

Appendix J

The following figures depict k-means clustering on raw data for various timestamps. Different colouring is applied for marked entries that were positioned to different clusters.

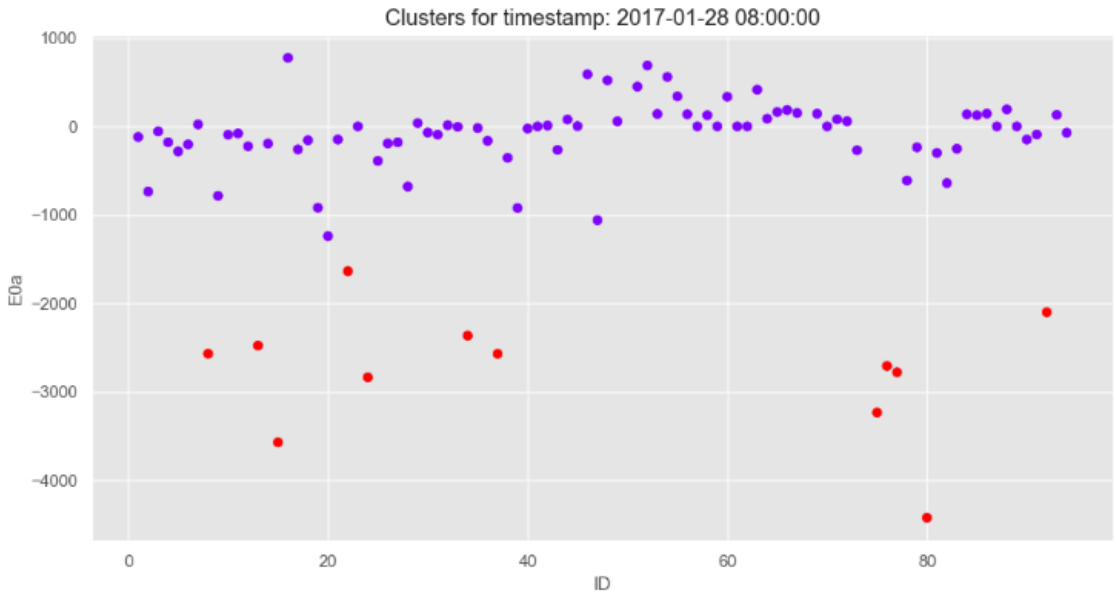

Fig. J1. k-means clustering for ‘2017-01-28 08:00:00’.

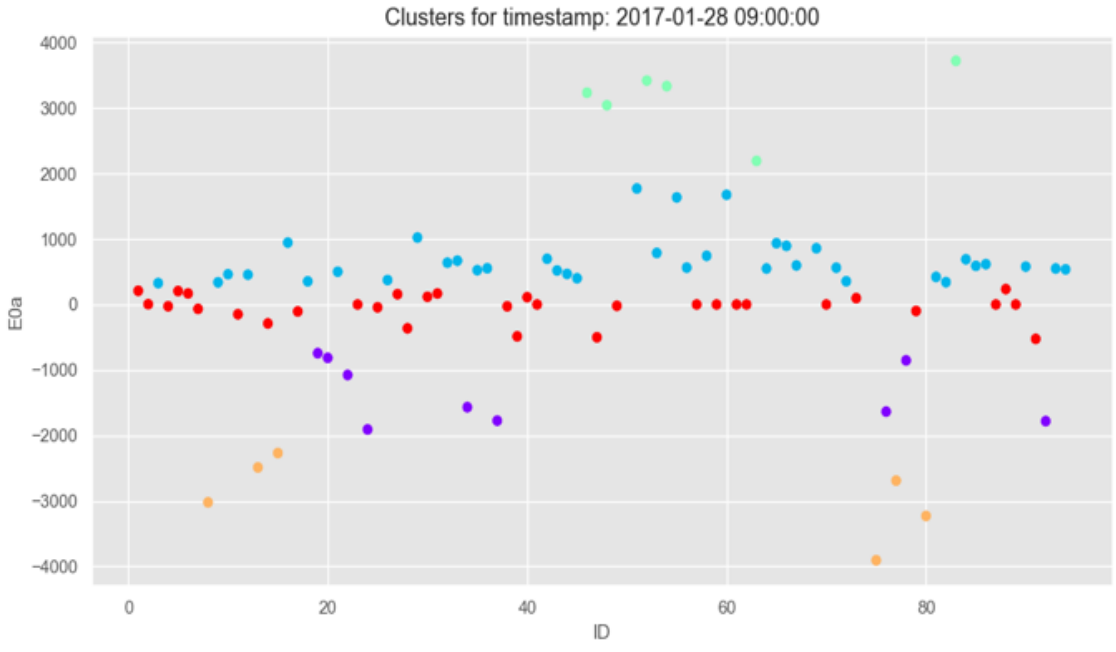

Fig. J2. k-means clustering for ‘2017-01-28 09:00:00’.

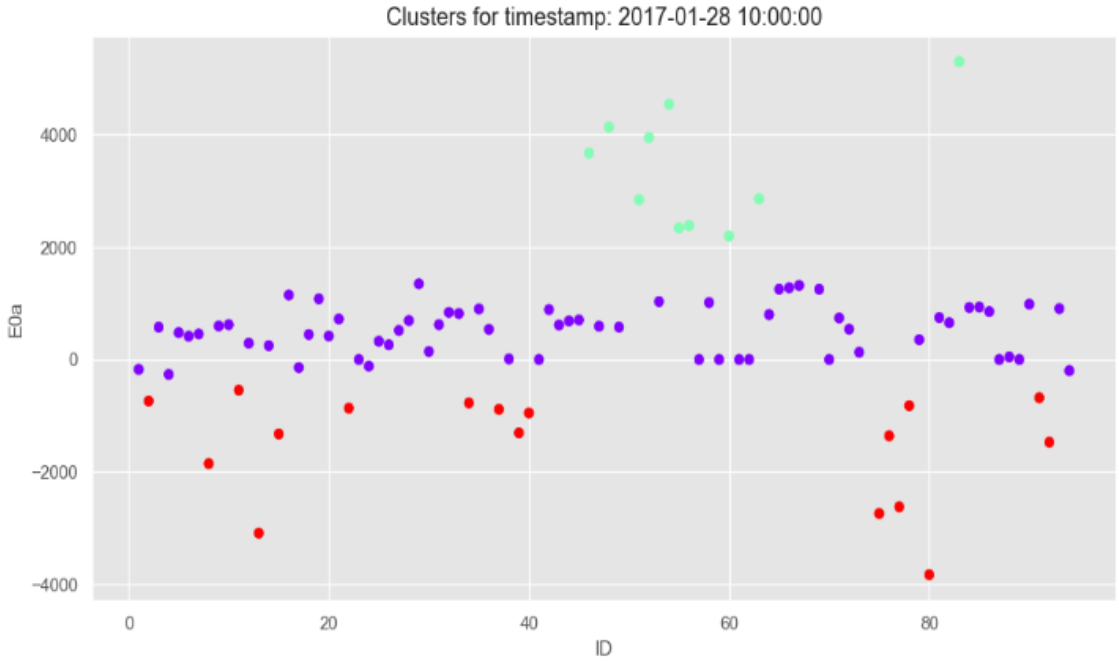

Fig. J3. k-means clustering for ‘2017-01-28 10:00:00’.

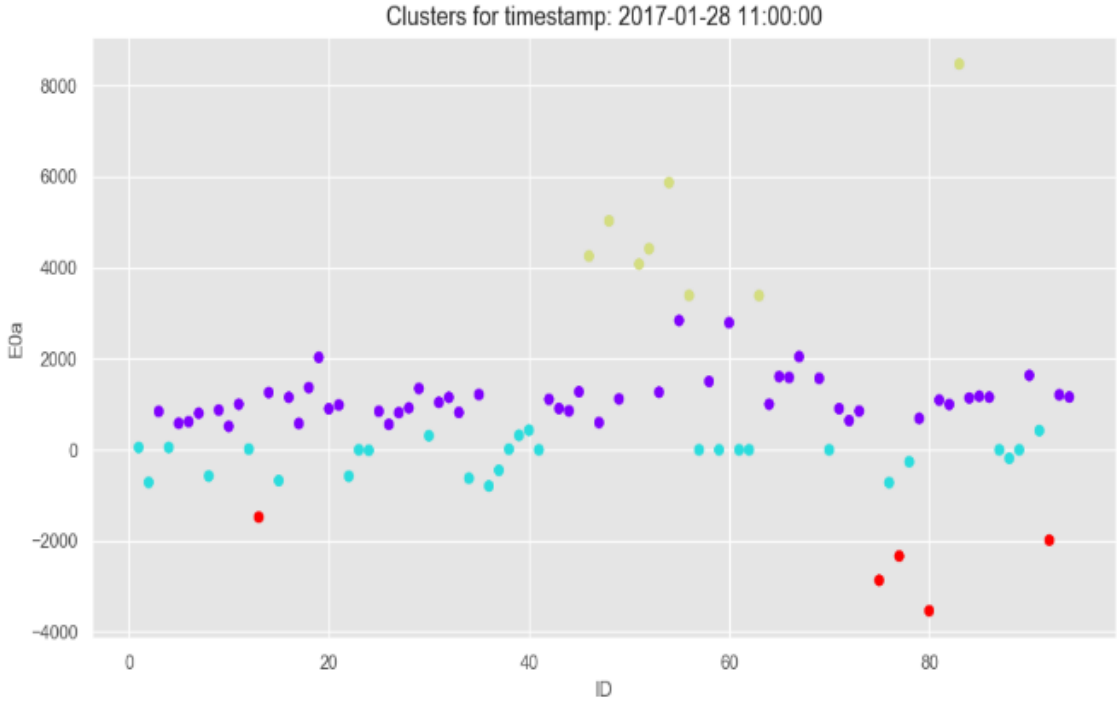

Fig. J4. k-means clustering for ‘2017-01-28 11:00:00’.

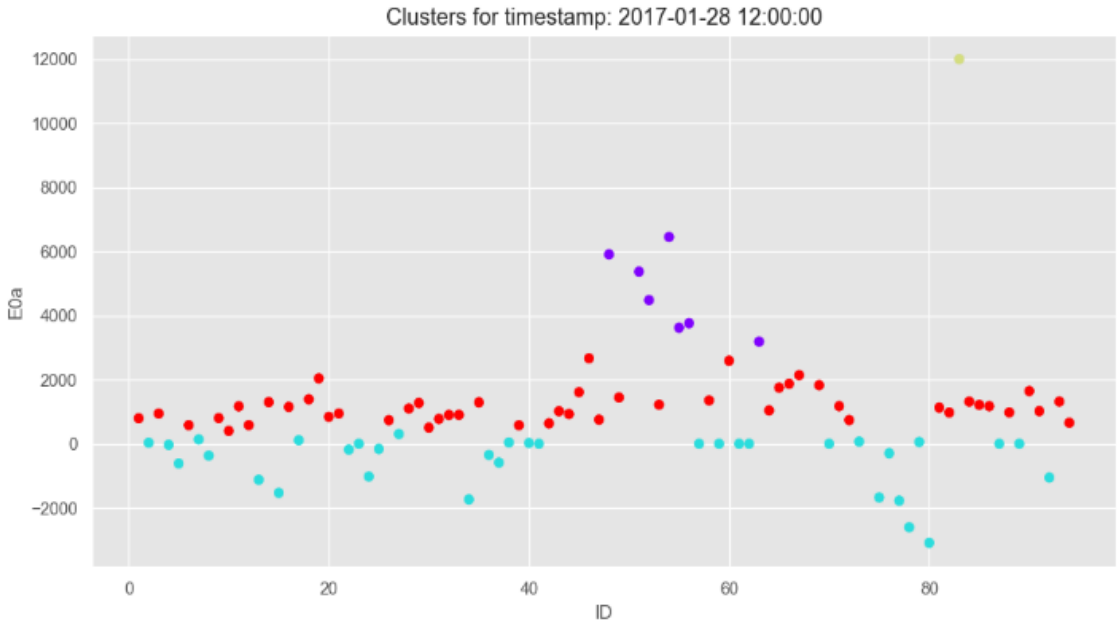

Fig. J5. k-means clustering for ‘2017-01-28 12:00:00’.

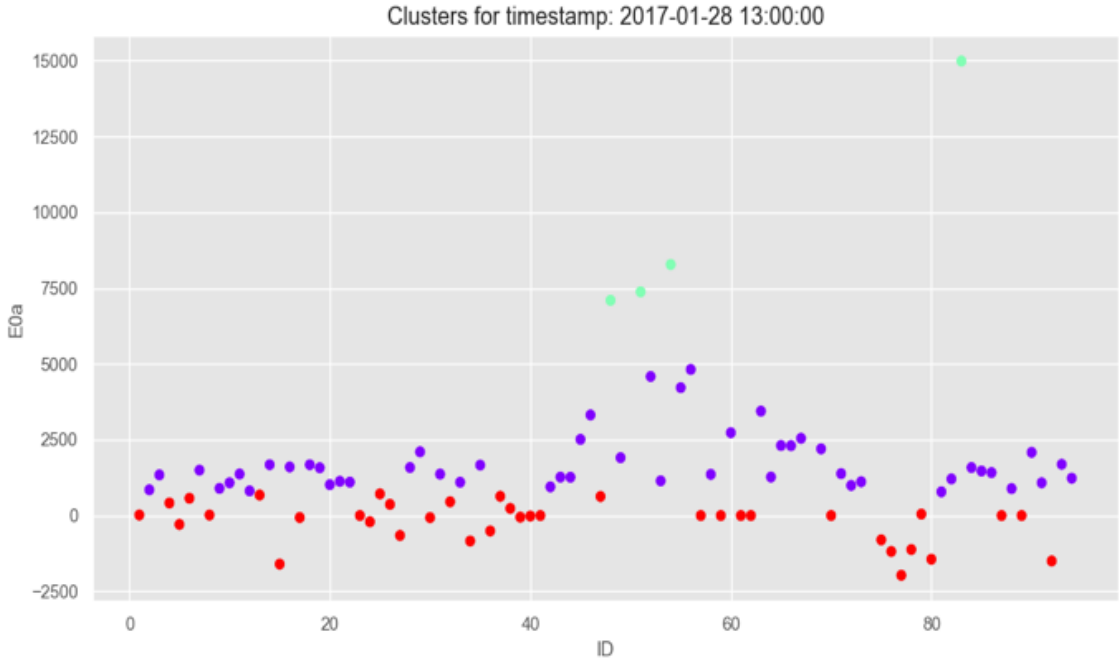

Fig. J6. k-means clustering for ‘2017-01-28 13:00:00’.

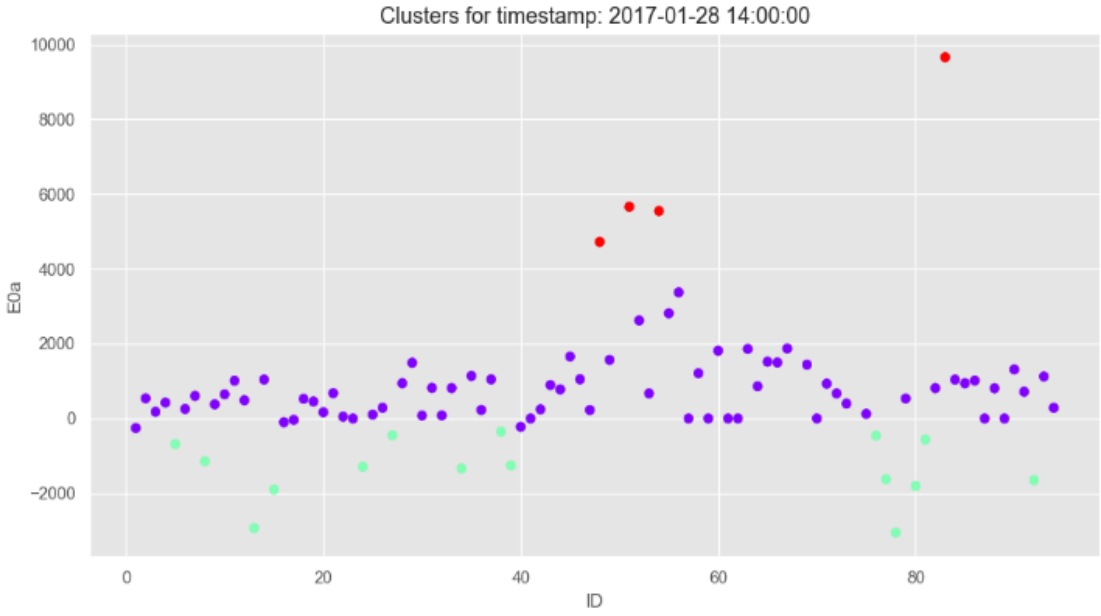

Fig. J7. k-means clustering for ‘2017-01-28 14:00:00’.

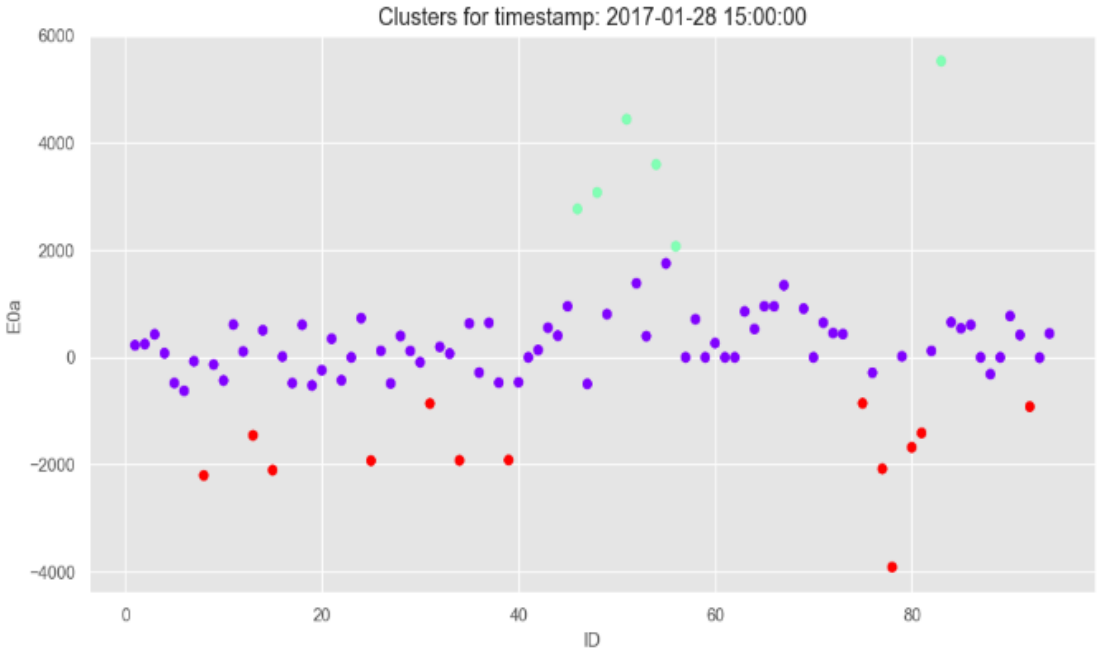

Fig. J8. k-means clustering for ‘2017-01-28 15:00:00’.

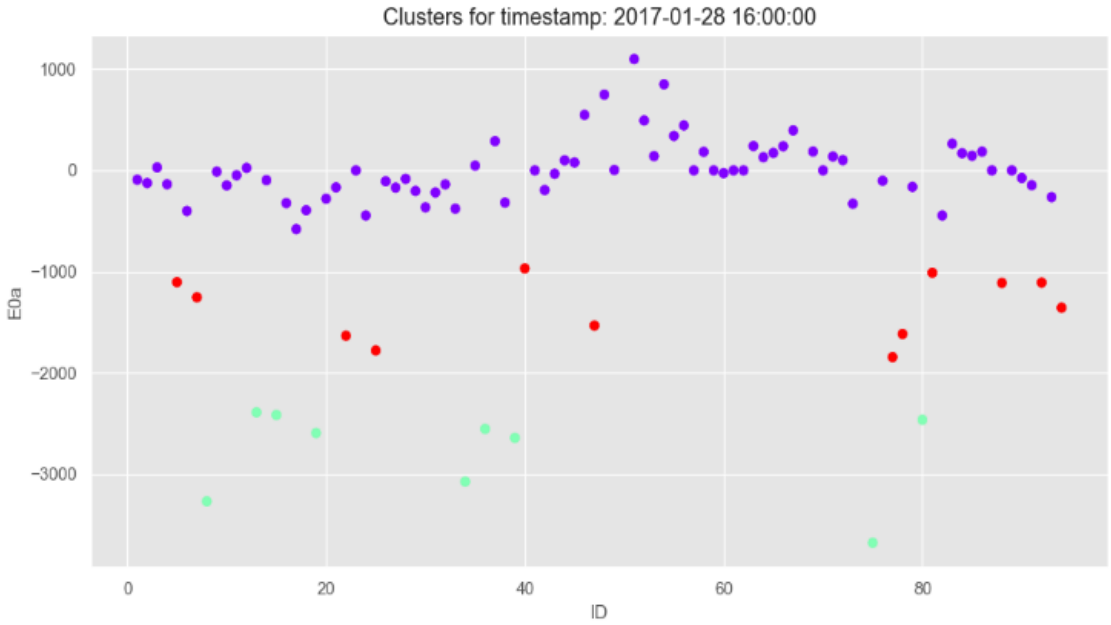

Fig. J9. k-means clustering for ‘2017-01-28 16:00:00’.

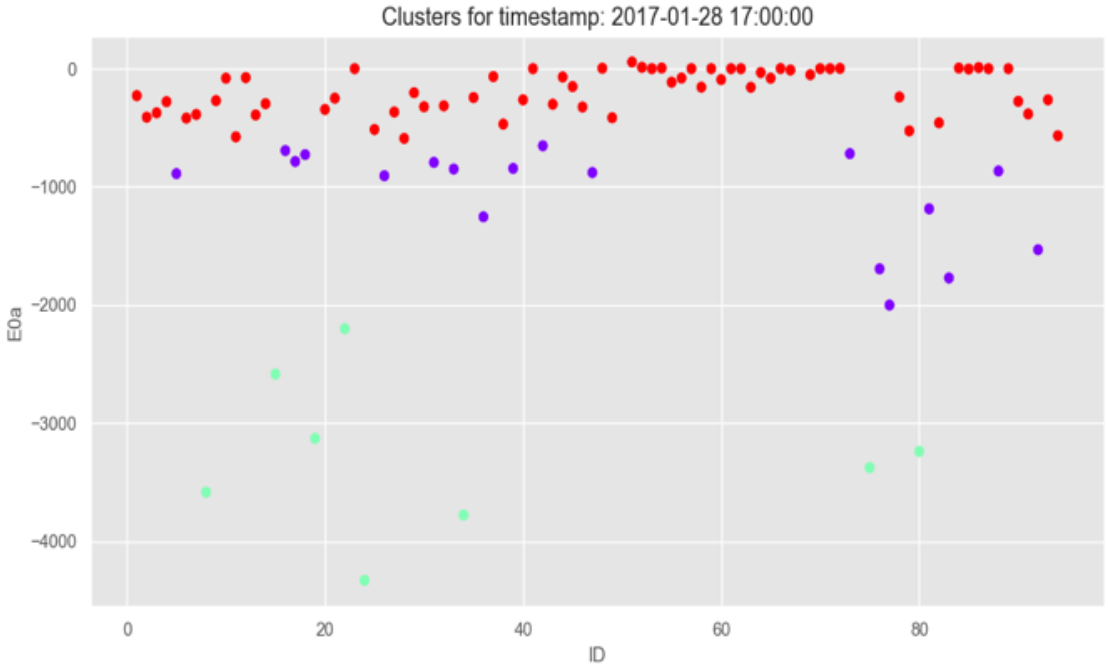

Fig. J10. k-means clustering for ‘2017-01-28 17:00:00’.

Appendix K

This Appendix contains an example of clustering and balancing on a P2P level. Each table refers to the following Fig., while different colouring is applied for marked entries that were positioned to different clusters.

Table K1. Clustering and balancing on a P2P level for '2017-01-28 08:00:00'.

Index	IDclass3	IDclass1	E0aclass3	E0aclass1	BalancedE0a
0	80	16	-4,424	774	-3,650
1	15	52	-3,572	688	-2,884
2	75	46	-3,235	588	-2,647
3	24	54	-2,837	559	-2,278
4	77	48	-2,780	521	-2,259
5	76	51	-2,780	449	-2,260
6	37	63	-2,709	414	-2,157
7	8	55	-2,571	339	-2,231
8	13	60	-2,570	335	-2,143
9	34	88	-2,478	192	-2,174
10	92	66	-2,366	185	-1,916
11	22	65	-2,101	162	-1,475
12	20	67	-1,637	151	-1,090
13	47	86	-1,060	144	-916
14	39	69	-923	143	-780
15	19	53	-920	140	-780
16	9	84	-785	137	-648
17	2	56	-737	137	-600
18	28	93	-681	130	-51
19	82	58	-641	126	-515
20	78	85	-612	126	-486
21	25	64	-390	87	-303
22	38	71	-356	79	-277
23	81	44	-300	77	-223
24	5	72	-284	58	-226
25	73	49	-270	57	-213
26	43	29	-266	37	-229
27	17	7	-260	22	-238
28	83	32	-253	13	-240
29	79	42	-237	9	-228
30	12	45	-225	3	-222

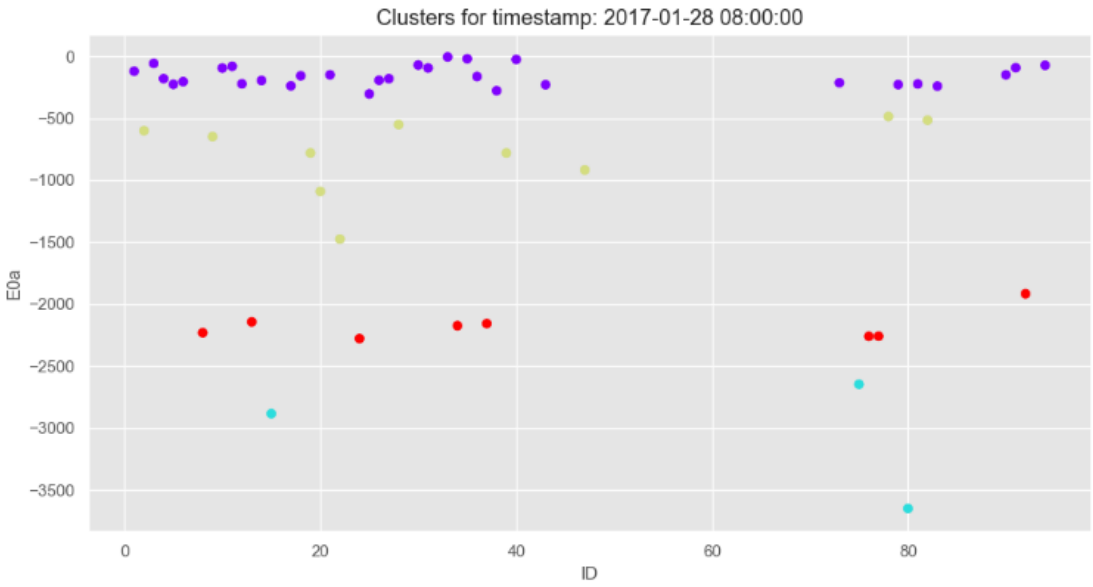

Fig. K1. Clustering and balancing on a P2P level for ‘2017-01-28 08:00:00’.

Table K2. Clustering and balancing on a P2P level for ‘2017-01-28 09:00:00’.

Index	IDclass3	IDclass1	E0aclass3	E0aclass1	BalancedE0a
0	83	75	3,722	-3,907	-185
1	52	80	3,421	-3,229	192
2	54	8	3,335	-3,022	313
3	46	77	3,235	-2,690	545
4	48	13	3,043	-2,489	554
5	63	15	2,193	-2,268	-75
6	51	24	1,772	-1,909	-137
7	60	92	1,676	-1,782	-106
8	55	37	1,636	-1,774	-138
9	29	76	1,022	-1,636	-614
10	16	34	945	-1,569	-624
11	65	22	936	-1,076	-140
12	66	78	895	-854	41
13	69	20	859	-815	44
14	53	19	788	-744	44
15	58	91	744	-525	219
16	42	47	698	-500	198
17	84	39	690	-487	203
18	33	28	670	-364	306
19	32	14	639	-290	349
20	86	11	616	-149	467
21	67	17	597	-106	491
22	85	79	590	-98	492
23	90	7	579	-67	512
24	56	25	563	-43	520
25	71	38	562	-28	534
26	36	4	553	-27	526
27	93	49	549	-20	529
28	64	nan	549	0	549

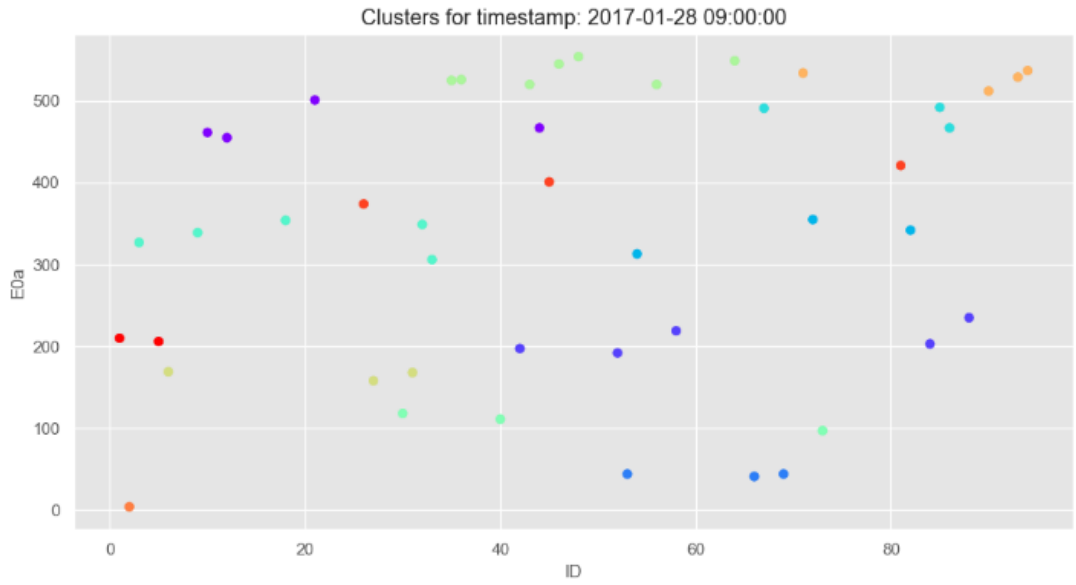

Fig. K2. Clustering and balancing on a P2P level for ‘2017-01-28 09:00:00’.

Table K3. Clustering and balancing on a P2P level for ‘2017-01-28 10:00:00’.

Index	IDclass3	IDclass1	E0aclass3	E0aclass1	BalancedE0a
0	83	80	5,302	-3,829	1,473
1	54	13	4,545	-3,088	1,457
2	48	75	4,139	-2,740	1,399
3	52	77	3,951	-2,620	1,331
4	46	8	3,675	-1,852	1,823
5	63	92	2,861	-1,471	1,390
6	51	76	2,844	-1,357	1,487
7	56	15	2,385	-1,324	1,061
8	55	39	2,341	-1,305	1,036
9	60	40	2,199	-950	1,249
10	29	37	1,348	-886	462
11	67	22	1,320	-865	455
12	66	78	1,277	-821	456
13	65	34	1,252	-772	480
14	69	2	1,251	-738	513
15	16	91	1,149	-680	469
16	19	11	1,080	-544	536
17	53	4	1,032	-264	768
18	58	94	1,015	-198	817
19	90	1	985	-175	810
20	85	17	935	-142	793
21	84	24	927	-117	810
22	93	nan	907	0	907

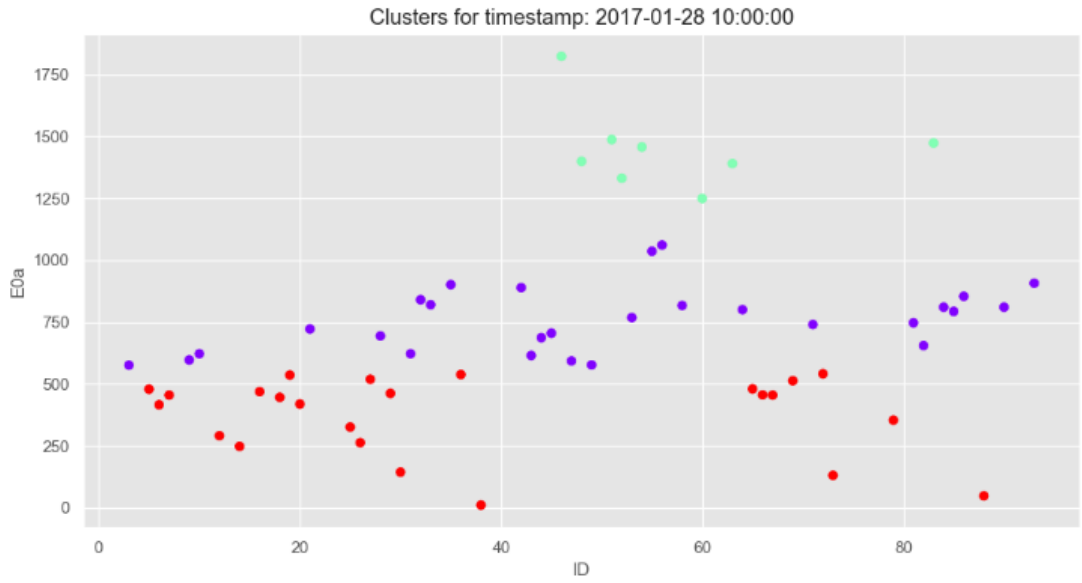

Fig. K3. Clustering and balancing on a P2P level for '2017-01-28 10:00:00'.

Table K4. Clustering and balancing on a P2P level for '2017-01-28 11:00:00'.

Index	IDclass3	IDclass1	E0aclass3	E0aclass1	BalancedE0a
0	83	80	8,473	-3,535	4,938
1	54	75	5,872	-2,865	3,007
2	48	77	5,030	-2,334	2,696
3	52	92	4,419	-1,986	2,433
4	46	13	4,255	-1,479	2,776
5	51	36	4,078	-795	3,283
6	56	76	3,390	-720	2,670
7	63	2	3,386	-715	2,671
8	55	15	2,842	-676	2,166
9	60	34	2,792	-625	2,167
10	67	22	2,045	-578	1,467
11	19	8	2,030	-574	1,456
12	90	37	1,636	-449	1,187
13	65	78	1,610	-263	1,347
14	66	88	1,590	-187	1,403
15	69	24	1,569	-11	1,558
16	58	nan	1,502	0	1,502

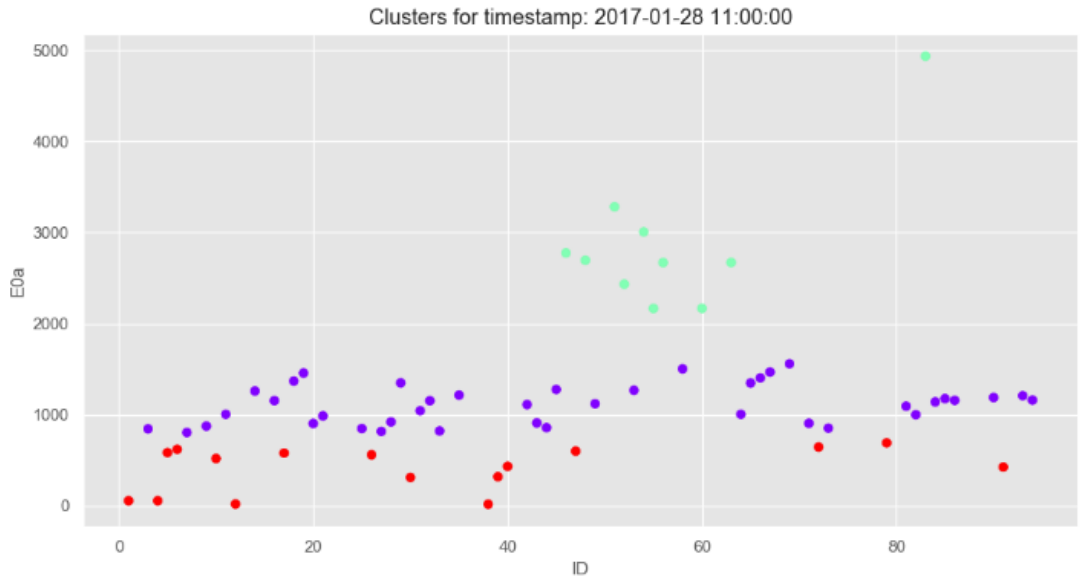

Fig. K4. Clustering and balancing on a P2P level for '2017-01-28 11:00:00'.

Table K5. Clustering and balancing on a P2P level for '2017-01-28 12:00:00'.

Index	IDclass3	IDclass1	E0aclass3	E0aclass1	BalancedE0a
0	83	80	11,998	-3,089	8,909
1	54	78	6,450	-2,605	3,845
2	48	77	5,906	-1,774	4,132
3	51	34	5,369	-1,737	3,632
4	52	75	4,481	-1,674	2,807
5	56	15	3,761	-1,530	2,231
6	55	13	3,621	-1,122	2,499
7	63	92	3,187	-1,055	2,132
8	46	24	2,666	-1,022	1,644
9	60	5	2,591	-615	1,976
10	67	37	2,141	-587	1,554
11	19	8	2,041	-370	1,671
12	66	36	1,871	-347	1,524
13	69	76	1,826	-296	1,530
14	65	22	1,752	-181	1,571
15	90	25	1,645	-161	1,484
16	45	4	1,610	-32	1,578
17	49	nan	1,443	0	1,443

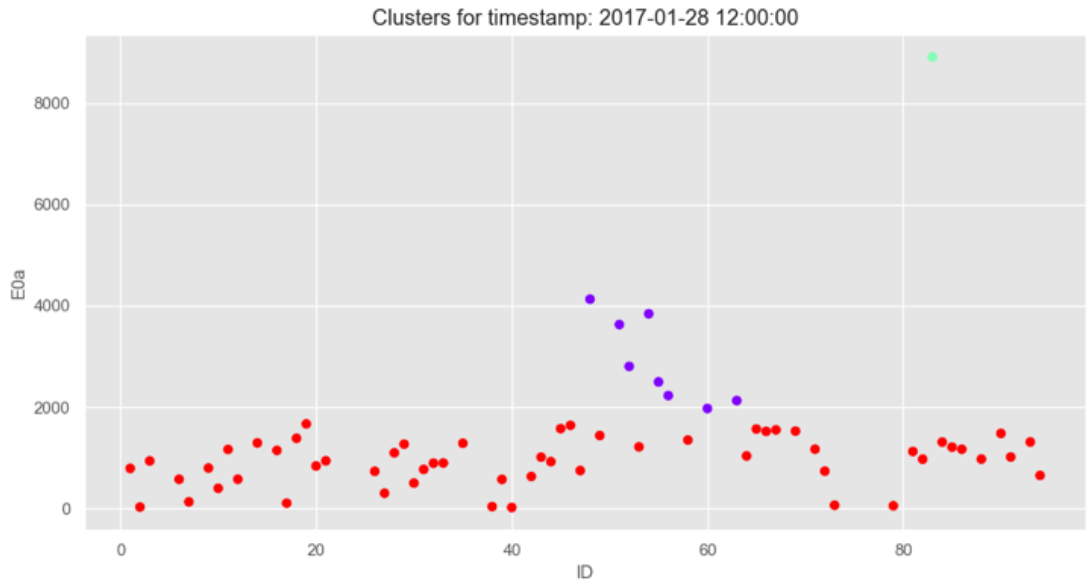

Fig. K5. Clustering and balancing on a P2P level for '2017-01-28 12:00:00'.

Table K6. Clustering and balancing on a P2P level for '2017-01-28 13:00:00'.

Index	IDclass3	IDclass1	E0aclass3	E0aclass1	BalancedE0a
0	83	77	14,987	-1,969	13,018
1	54	15	8,279	-1,600	6,679
2	51	92	7,379	-1,499	5,880
3	48	80	7,097	-1,440	5,657
4	56	76	4,817	-1,184	3,633
5	52	78	4,586	-1,120	3,466
6	55	34	4,218	-837	3,381
7	63	75	3,444	-800	2,644
8	46	27	3,317	-656	2,661
9	60	36	2,732	-506	2,226
10	67	5	2,546	-293	2,253
11	45	24	2,512	-206	2,306
12	65	30	2,310	-68	2,242
13	66	17	2,304	-65	2,239
14	69	39	2,201	-54	2,147
15	29	40	2,101	-16	2,085
16	90	nan	2,084	0	2,084

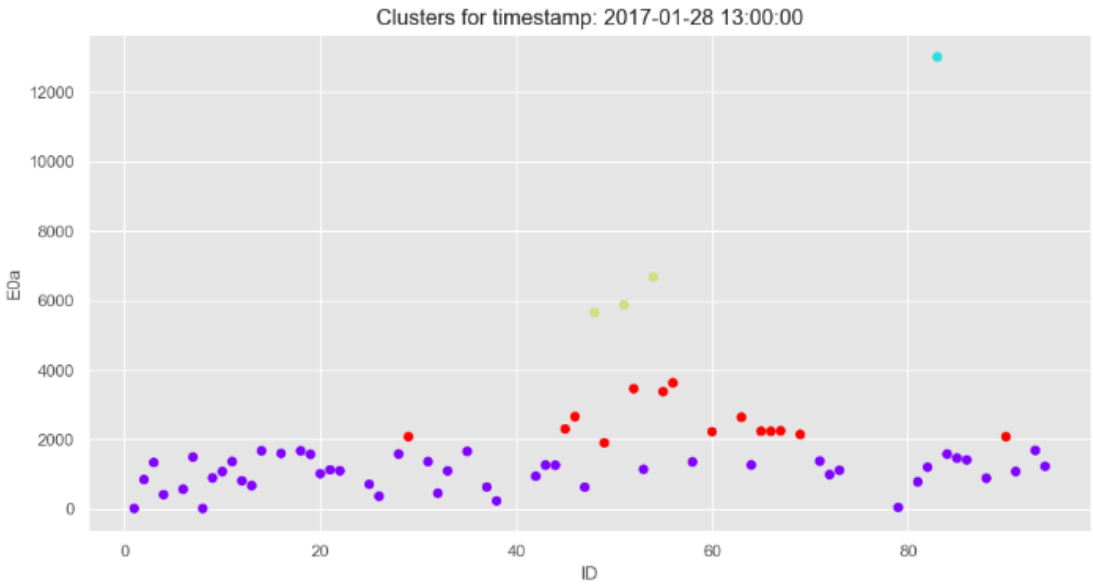

Fig. K6. Clustering and balancing on a P2P level for '2017-01-28 13:00:00'.

Table K7. Clustering and balancing on a P2P level for '2017-01-28 14:00:00'.

Index	IDclass3	IDclass1	E0aclass3	E0aclass1	BalancedE0a
0	83	78	9,663	-3,052	6,611
1	51	13	5,662	-2,932	2,730
2	54	15	5,549	-1,901	3,648
3	48	80	4,725	-1,805	2,920
4	56	92	3,374	-1,649	1,725
5	55	77	2,814	-1,628	1,186
6	52	34	2,622	-1,338	1,284
7	67	24	1,871	-1,293	578
8	63	39	1,859	-1,256	603
9	60	8	1,813	-1,147	666
10	45	5	1,657	-684	973
11	49	81	1,568	-561	1,007
12	65	76	1,517	-457	1,060
13	66	27	1,497	-448	1,049
14	29	38	1,490	-352	1,138
15	69	1	1,440	-256	1,184
16	90	40	1,311	-222	1,089
17	58	16	1,210	-100	1,110
18	35	17	1,139	-38	1,101

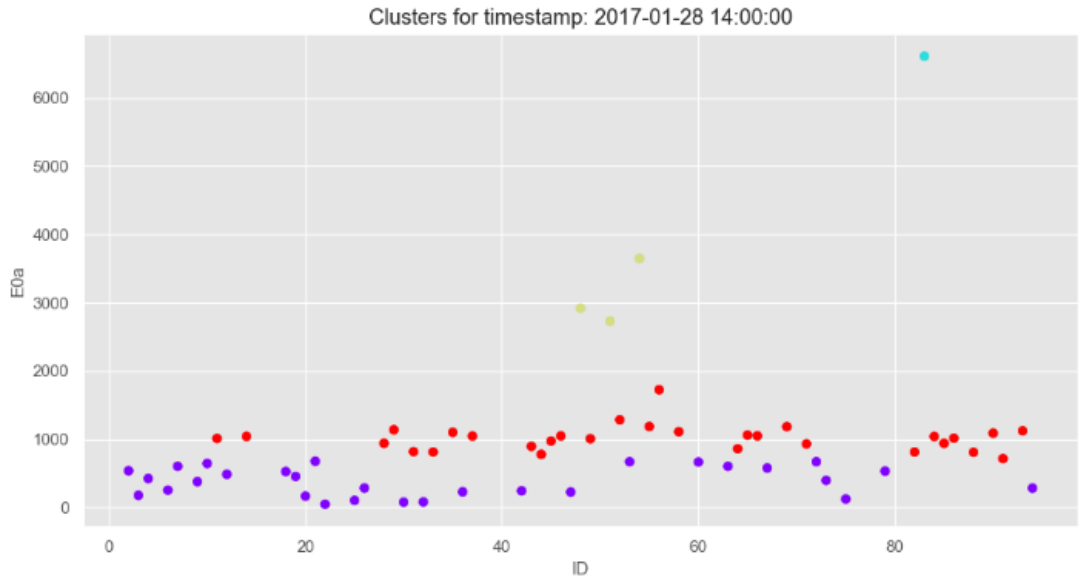

Fig. K7. Clustering and balancing on a P2P level for '2017-01-28 14:00:00'.

Table K8. Clustering and balancing on a P2P level for '2017-01-28 15:00:00'.

Index	IDclass3	IDclass1	E0aclass3	E0aclass1	BalancedE0a
0	83	78	5,529	-3,914	1,615
1	51	8	4,443	-2,202	2,241
2	54	15	3,600	-2,103	1,497
3	48	77	3,079	-2,079	1,000
4	46	25	2,769	-1,927	842
5	56	34	2,073	-1,922	151
6	55	39	1,754	-1,916	-162
7	52	80	1,384	-1,678	-294
8	67	13	1,345	-1,453	-108
9	65	81	954	-1,411	-457
10	45	92	953	-919	34
11	66	31	953	-860	93
12	69	75	910	-856	54
13	63	6	856	-623	233
14	49	19	807	-522	285
15	90	47	771	-495	276
16	24	27	733	-487	246
17	58	17	713	-480	233
18	84	5	657	-479	178
19	71	38	649	-472	177
20	37	40	646	-464	182
21	35	10	633	-431	202
22	11	22	611	-427	184
23	18	88	607	-314	293
24	86	76	605	-286	319
25	43	36	555	-282	273
26	85	20	543	-240	303
27	64	9	529	-134	395
28	14	30	506	-91	415
29	72	7	453	-73	380
30	94	93	445	-5	440

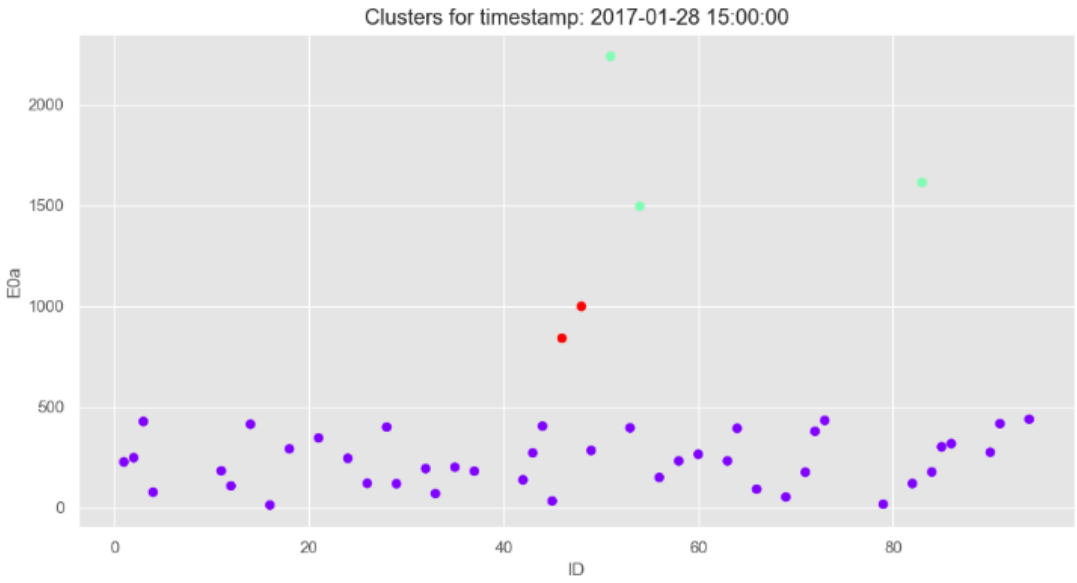

Fig. K8. Clustering and balancing on a P2P level for '2017-01-28 15:00:00'.

Table K9. Clustering and balancing on a P2P level for '2017-01-28 16:00:00'.

Index	IDclass3	IDclass1	E0aclass3	E0aclass1	BalancedE0a
0	75	51	-3,682	1,100	-2,582
1	8	54	-3,272	851	-2,421
2	34	48	-3,077	749	-2,328
3	39	46	-2,644	548	-2,096
4	19	52	-2,596	493	-2,103
5	36	56	-2,556	444	-2,112
6	80	67	-2,466	395	-2,071
7	15	55	-2,418	341	-2,077
8	13	37	-2,392	289	-2,103
9	77	83	-1,846	263	-1,583
10	25	63	-1,781	241	-1,540
11	22	66	-1,635	239	-1,396
12	78	69	-1,617	186	-1,431
13	47	86	-1,534	185	-1,349
14	94	58	-1,358	183	-1,175
15	7	65	-1,255	172	-1,083
16	88	84	-1,112	169	-943
17	92	85	-1,109	145	-964
18	5	53	-1,105	141	-964
19	81	71	-1,012	137	-875
20	40	64	-969	129	-840
21	17	72	-581	102	-479
22	24	44	-446	100	-346
23	82	45	-446	78	-368
24	6	35	-401	47	-354
25	18	3	-394	29	-365
26	33	12	-379	24	-355
27	30	49	-365	4	-361
28	73	nan	-330	0	-330

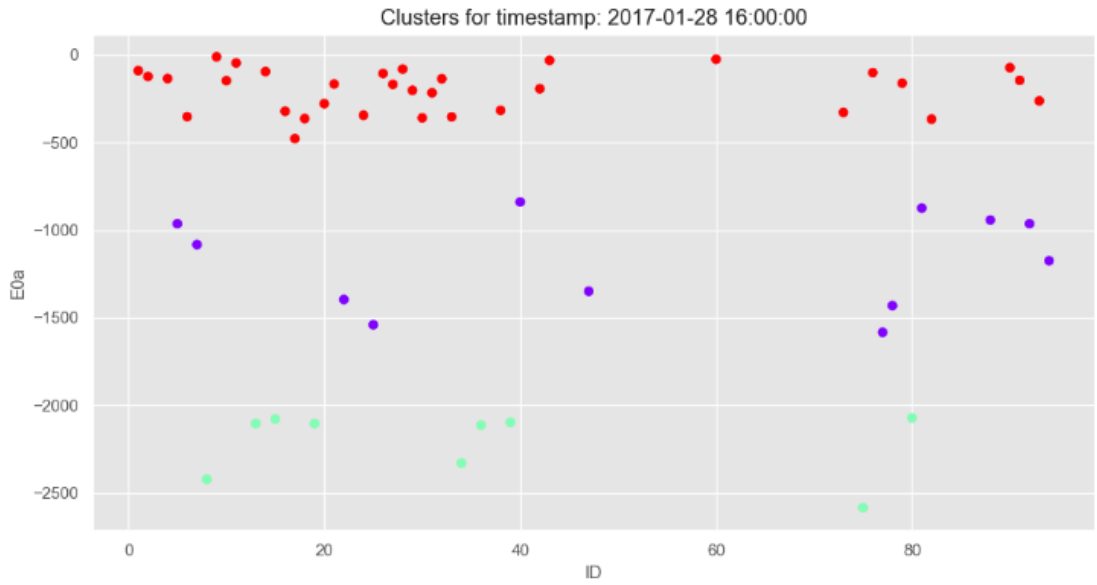

Fig. K9. Clustering and balancing on a P2P level for '2017-01-28 16:00:00'.

Table K10. Clustering and balancing on a P2P level for '2017-01-28 17:00:00'.

Index	IDclass3	IDclass1	E0aclass3	E0aclass1	BalancedE0a
0	24	51	-4,333	56	-4,277
1	34	52	-3,780	11	-3,769
2	8	86	-3,586	9	-3,577
3	75	84	-3,378	5	-3,373
4	80	54	-3,241	5	-3,236
5	19	48	-3,131	3	-3,128
6	15	72	-2,586	2	-2,584
7	22	nan	-2,202	0	-2,202

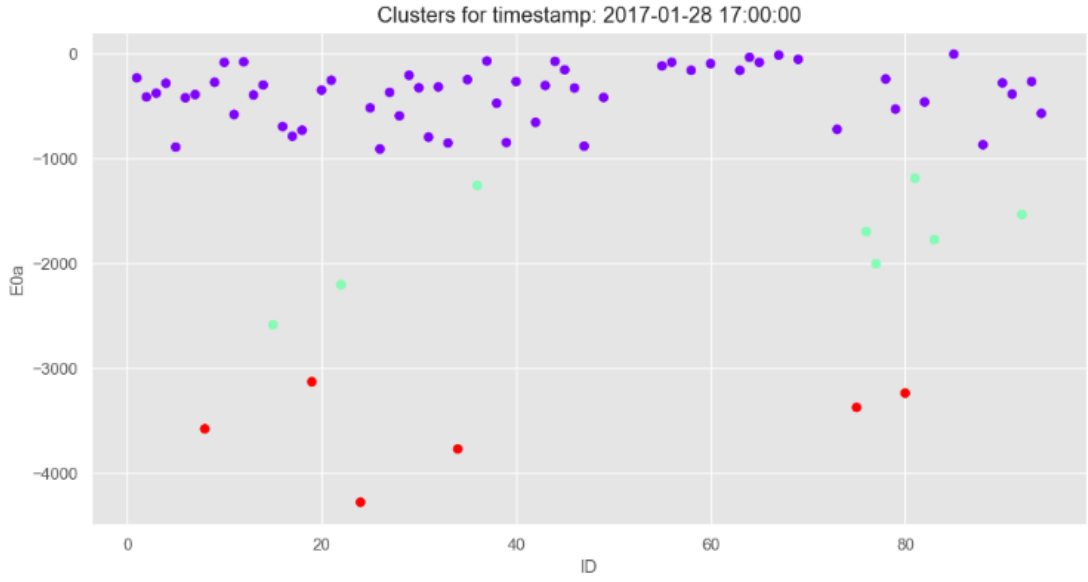

Fig. K10. Clustering and balancing on a P2P level for '2017-01-28 17:00:00'.

Appendix L

This Appendix contains a binning overview on raw data on a specific timestamp.

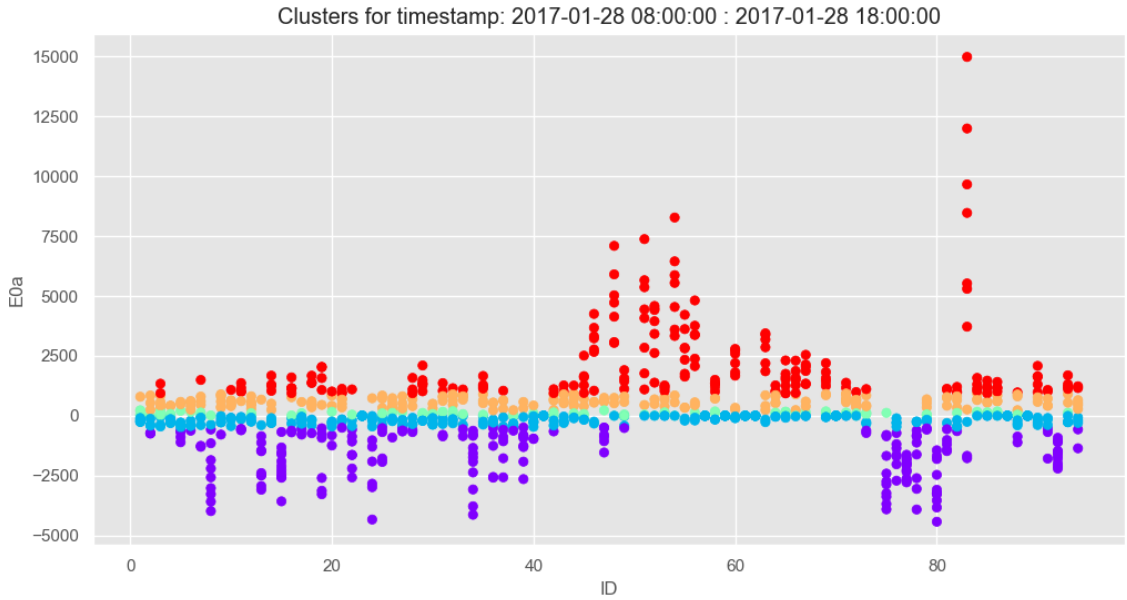

Fig. L1. CUT: #create Bins by pandas.cut (binning) (McKinney, 2010) specifically define the bin value edges.

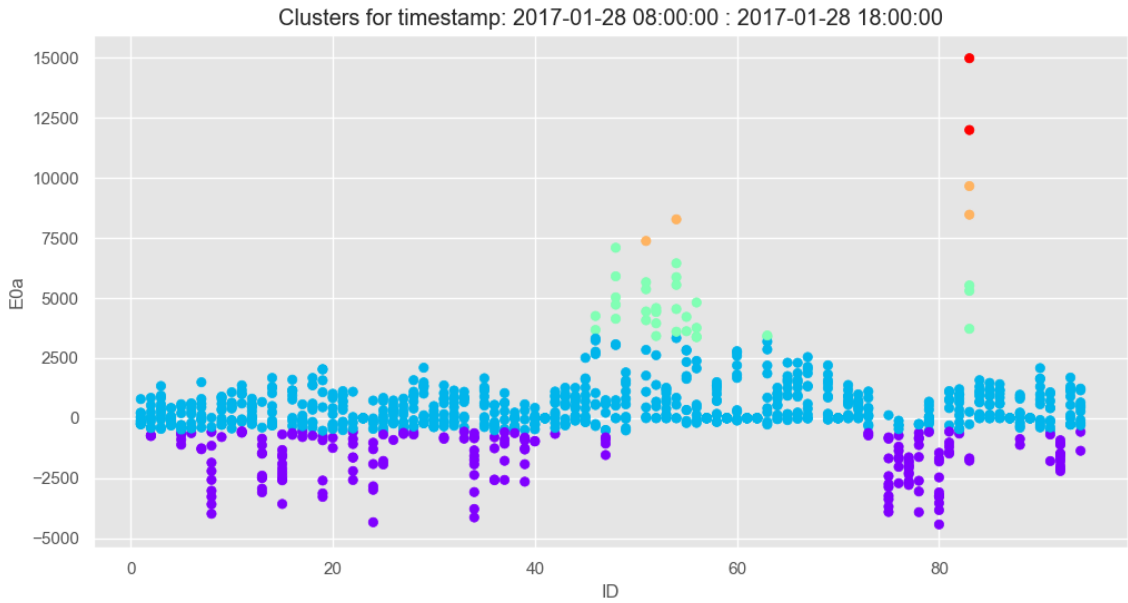

Fig. L2. QCUT: create Bins by pandas.qcut (binning) (McKinney, 2010) "Quantile-based discretization function" equally sized bins.

Appendix M

QCUT method results for “sunlight” timestamp examples. Each table is followed by a Fig. depicting coloured points for each bin (labelled as cluster).

Table M1. QCUT binning for ‘2017-01-28 08:00:00’.

BinLabel	E0a count	Sum (Watt)
1	19	-39,627
2	18	-5,209
3	23	-1,183
4	13	824
5	18	6,058

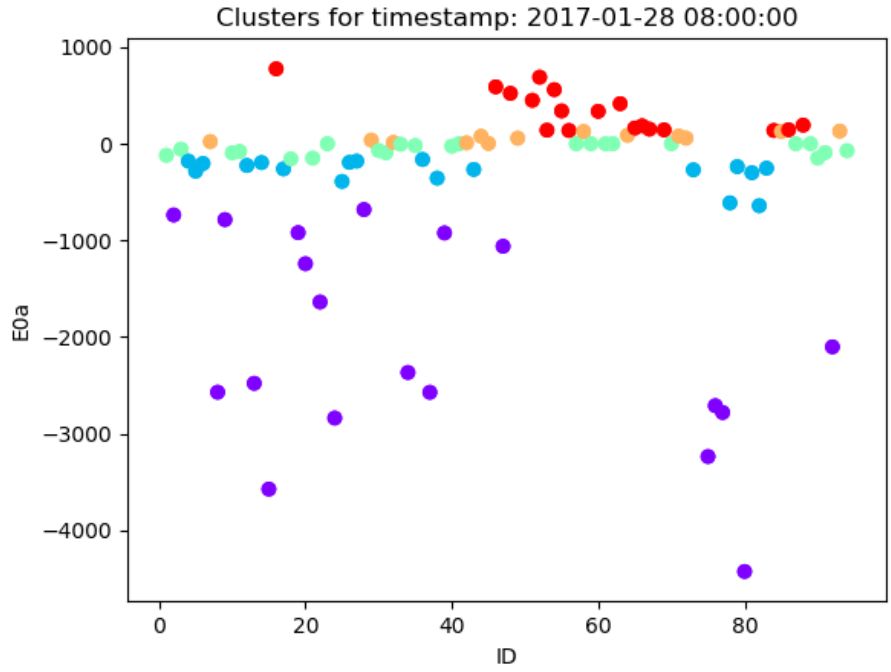

Fig. M1. QCUT binning for ‘2017-01-28 08:00:00’.

Table M2. QCUT binning for ‘2017-01-28 09:00:00’.

BinLabel	E0a count	Sum (Watt)
1	19	-31,640
2	18	-828
3	18	4,389
4	18	9,933
5	18	31,610

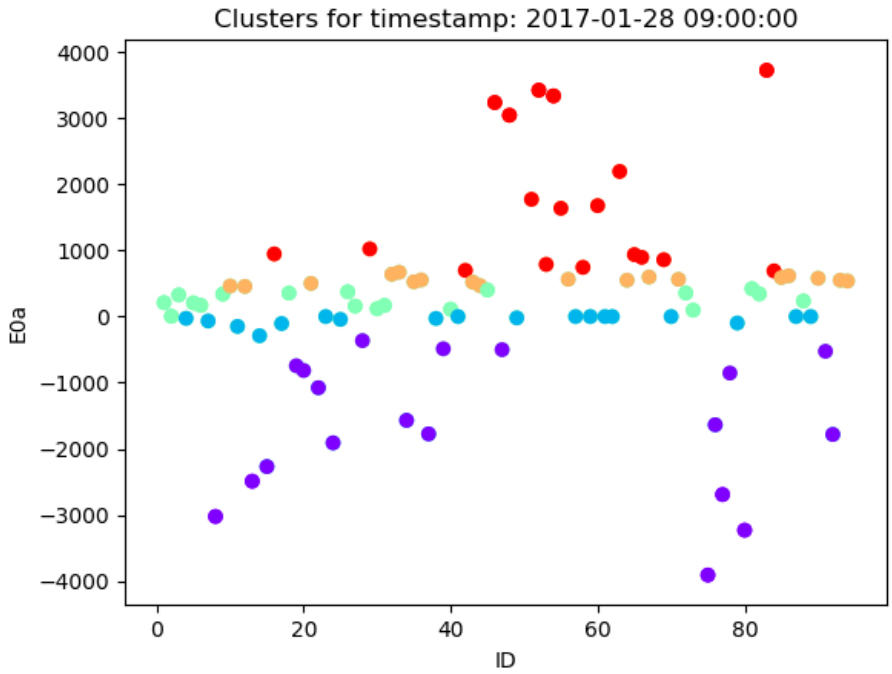

Fig. M2. QCUT binning for '2017-01-28 09:00:00'.

Table M3. QCUT binning for '2017-01-28 10:00:00'.

BinLabel	E0a count	Sum (Watt)
1	19	-26,304
2	18	411
3	18	8,986
4	18	14,823
5	18	43,951

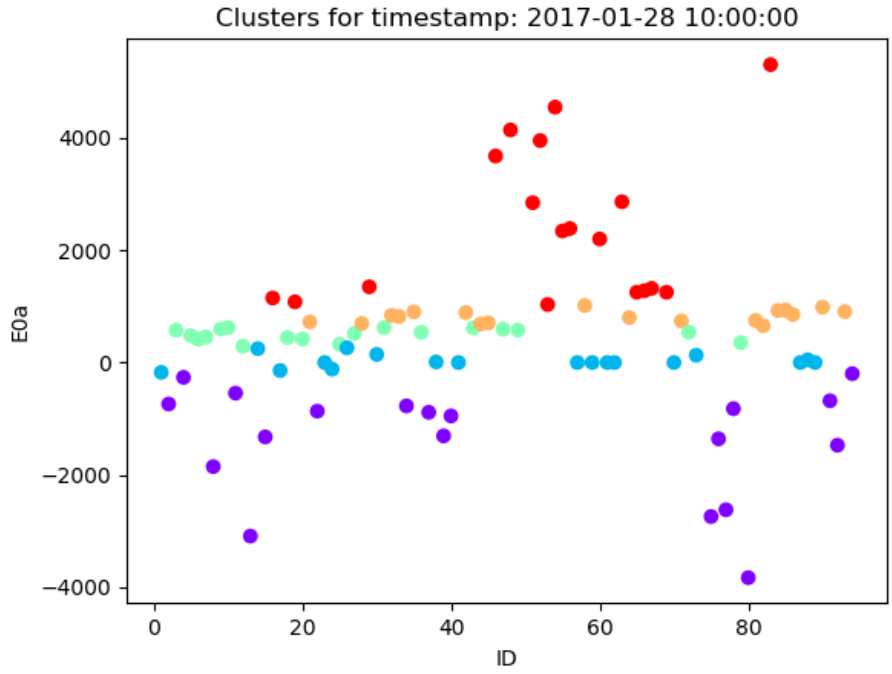

Fig. M3. QCUT binning for ‘2017-01-28 10:00:00’.

Table M4. QCUT binning for ‘2017-01-28 11:00:00’.

BinLabel	E0a count	Sum (Watt)
1	25	-17,792
2	12	3,852
3	18	14,871
4	18	20,874
5	18	57,887

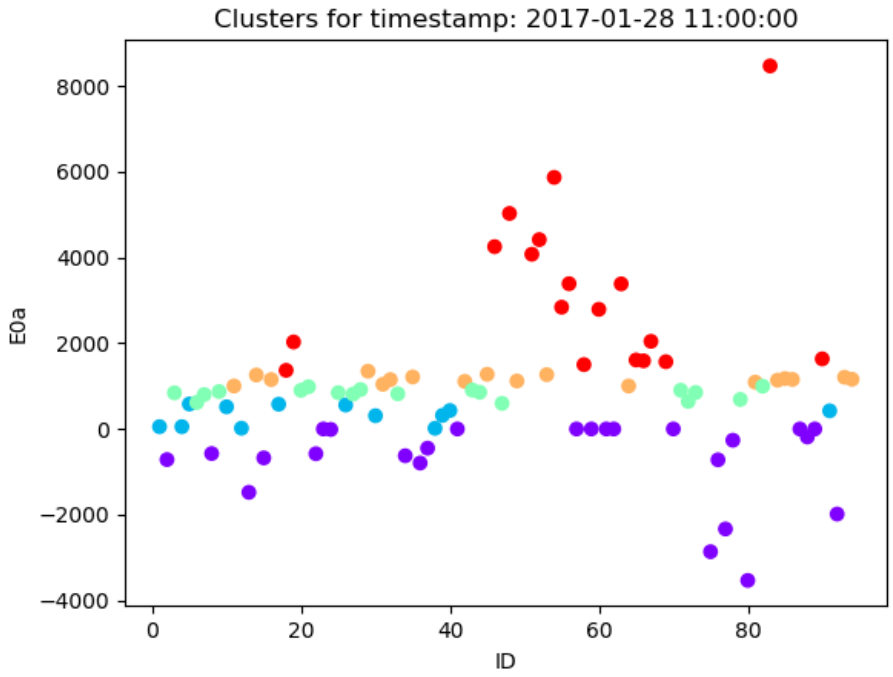

Fig. M4. QCUT binning for ‘2017-01-28 11:00:00’.

Table M5. QCUT binning for ‘2017-01-28 12:00:00’.

BinLabel	E0a count	Sum (Watt)
1	26	-18,197
2	11	2,255
3	18	14,458
4	18	21,635
5	18	64,359

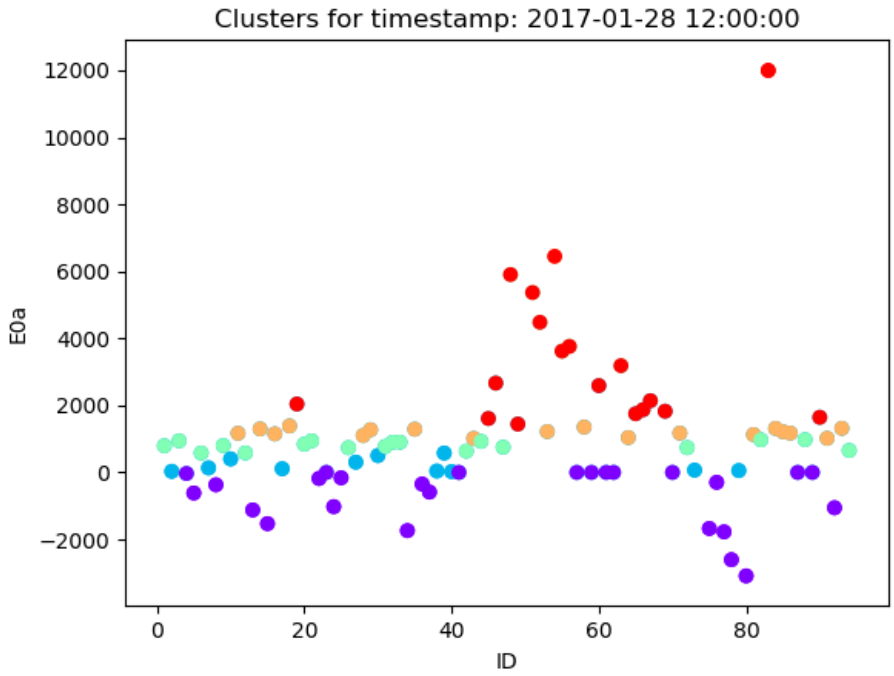

Fig. M5. QCUT binning for '2017-01-28 12:00:00'.

Table M6. QCUT binning for '2017-01-28 13:00:00'.

BinLabel	E0a count	Sum (Watt)
1	25	-12,313
2	12	4,784
3	18	18,641
4	18	26,781
5	18	78,822

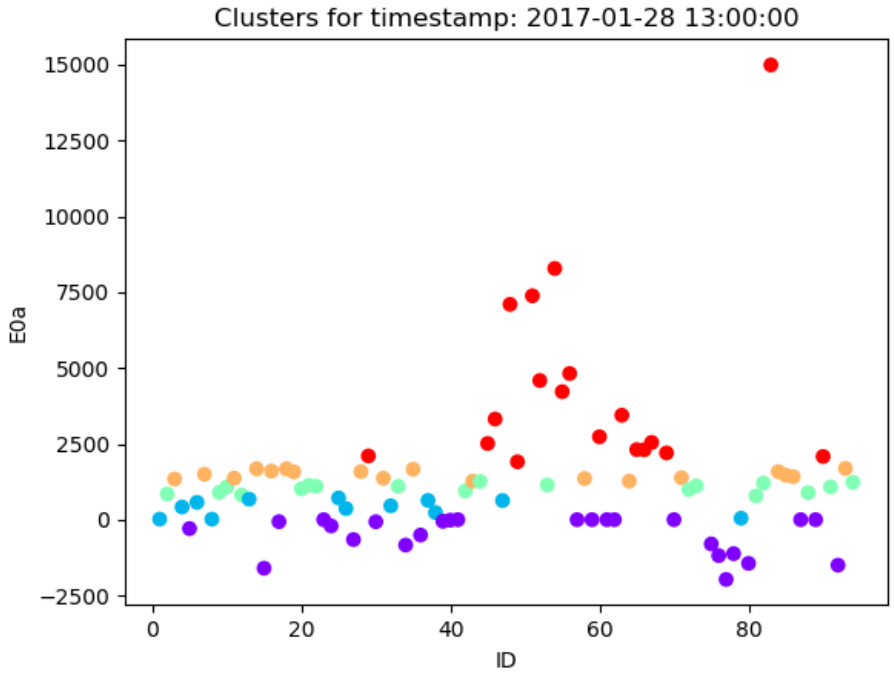

Fig. M6. QCUT binning for ‘2017-01-28 13:00:00’.

Table M7. QCUT binning for ‘2017-01-28 14:00:00’.

BinLabel	E0a count	Sum (Watt)
1	19	-21,119
2	18	1,234
3	18	8,785
4	18	17,061
5	18	51,642

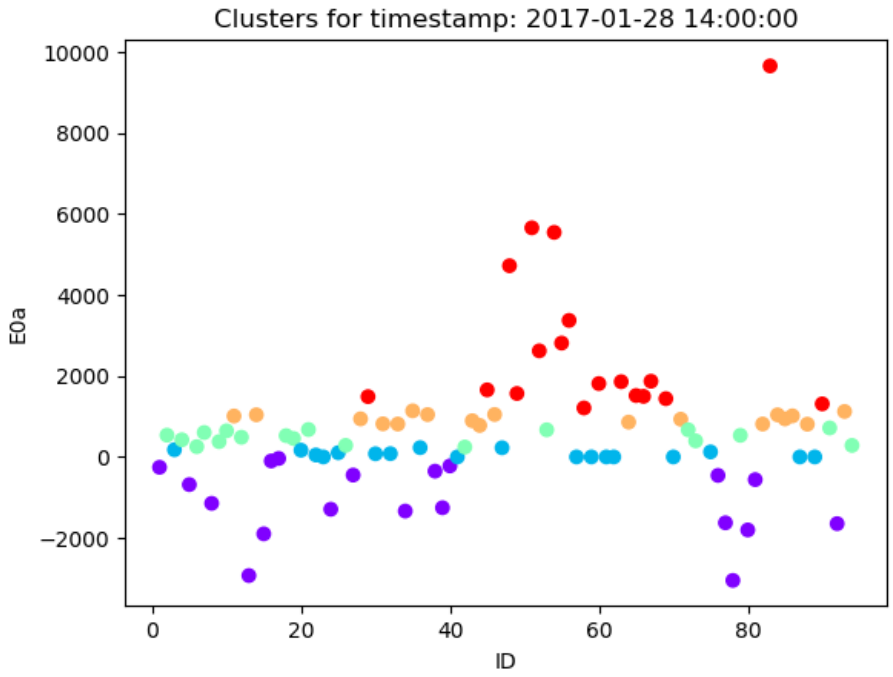

Fig. M7. QCUT binning for ‘2017-01-28 14:00:00’.

Table M8. QCUT binning for ‘2017-01-28 15:00:00’.

BinLabel	E0a count	Sum (Watt)
1	19	-26,326
2	21	-3,219
3	15	2,474
4	18	9,527
5	18	33,626

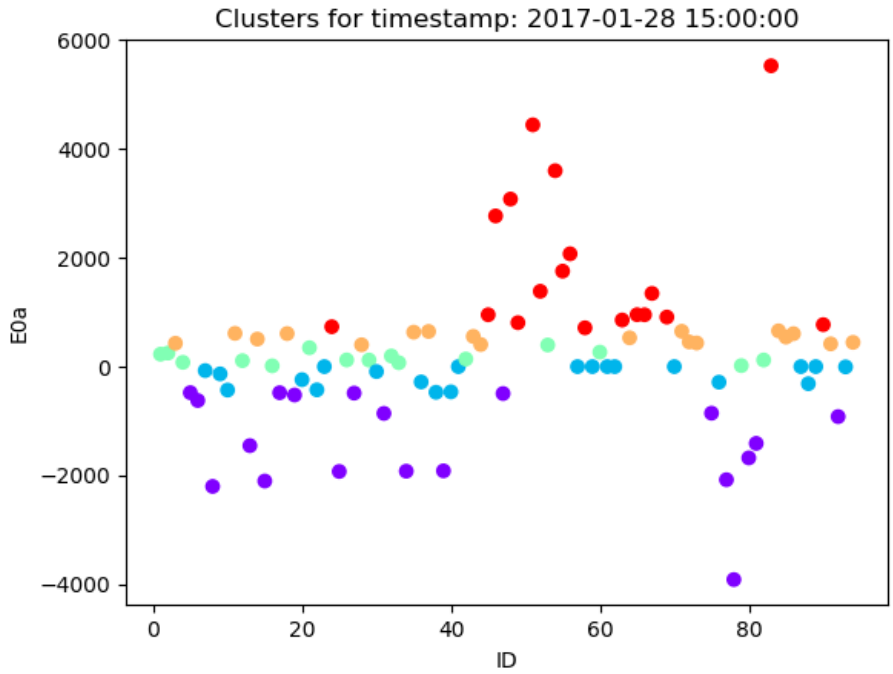

Fig. M8. QCUT binning for '2017-01-28 15:00:00'.

Table M9. QCUT binning for '2017-01-28 16:00:00'.

BinLabel	E0a count	Sum (Watt)
1	19	-39,455
2	18	-7,295
3	26	-1,706
4	10	791
5	18	6,993

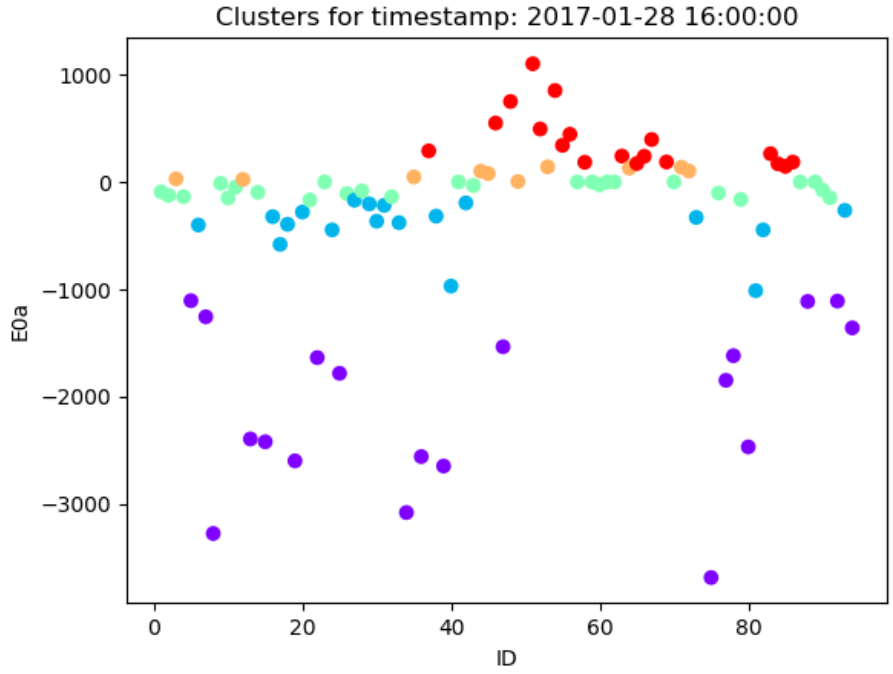

Fig. M9. QCUT binning for ‘2017-01-28 16:00:00’.

Table M10. QCUT binning for ‘2017-01-28 17:00:00’.

BinLabel	E0a count	Sum (Watt)
1	19	-40,074
2	18	-10,565
3	18	-5,514
4	28	-1,664
5	8	92

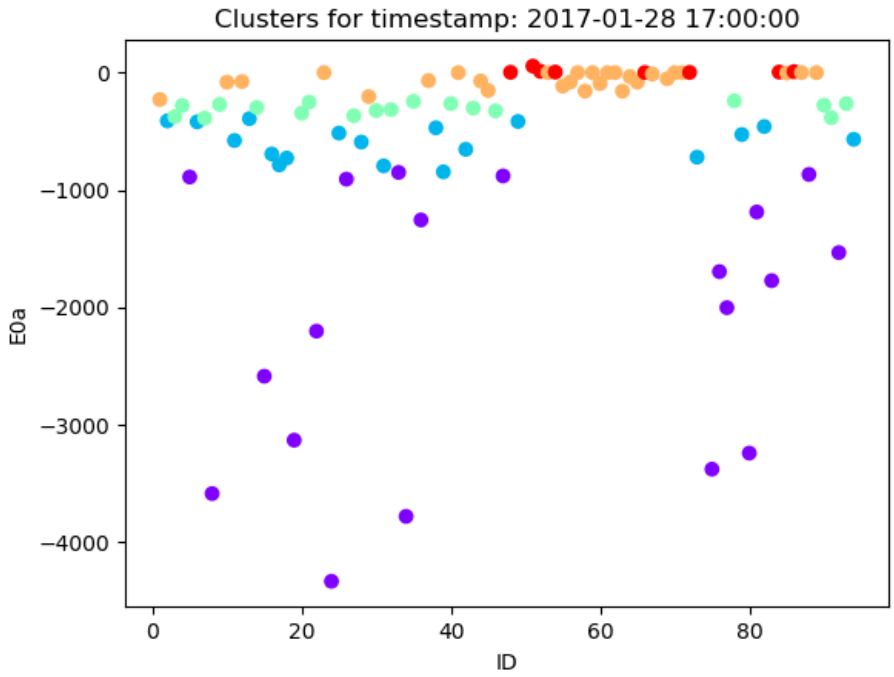

Fig. M10. QCUT binning for ‘2017-01-28 17:00:00’.

Appendix N

CUT method results for “sunlight” timestamp examples. Each table is followed by a Fig. depicting coloured points for each bin (labelled as cluster).

Table N1. CUT binning for ‘2017-01-28 08:00:00’.

BinLabel	E0a count	Sum (Watt)
1	2	-7,996
2	8	-21,546
3	2	-3,738
4	15	-9,466
5	64	3,609

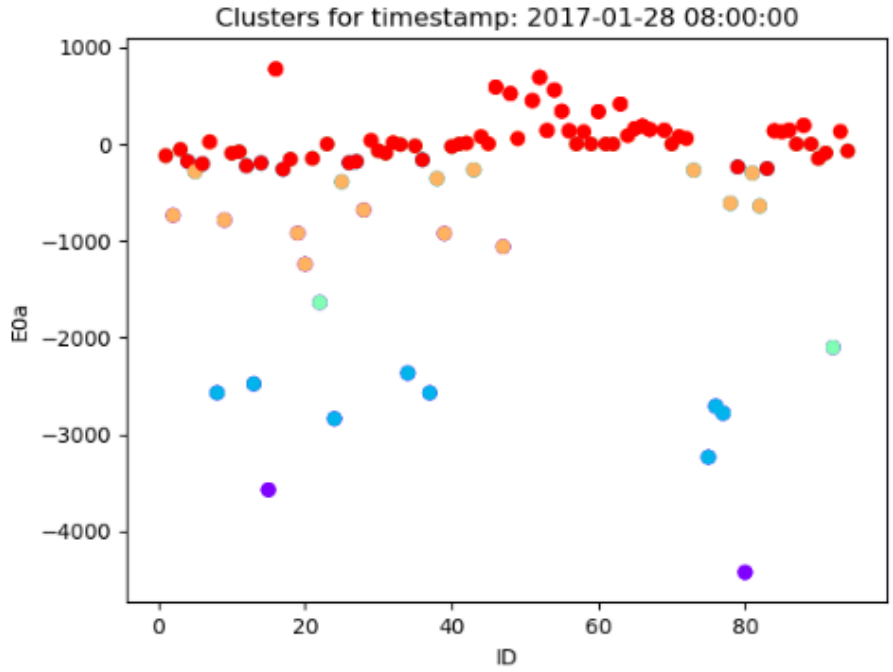

Fig. N1. CUT binning for ‘2017-01-28 08:00:00’.

Table N2. CUT binning for ‘2017-01-28 09:00:00’.

BinLabel	E0a count	Sum (Watt)
1	5	-15,337
2	7	-12,014
3	61	9,204
4	13	14,854
5	5	16,756

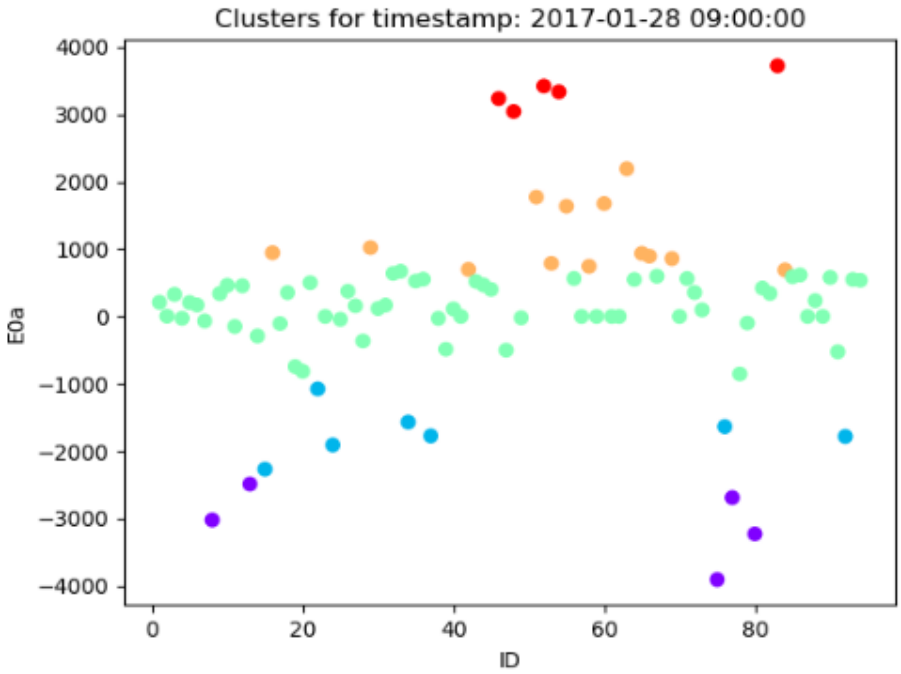

Fig. N2. CUT binning for '2017-01-28 09:00:00'.

Table N3. CUT binning for '2017-01-28 10:00:00'.

BinLabel	E0a count	Sum (Watt)
1	4	-12,277
2	15	-14,027
3	62	33,929
4	5	12,630
5	5	21,612

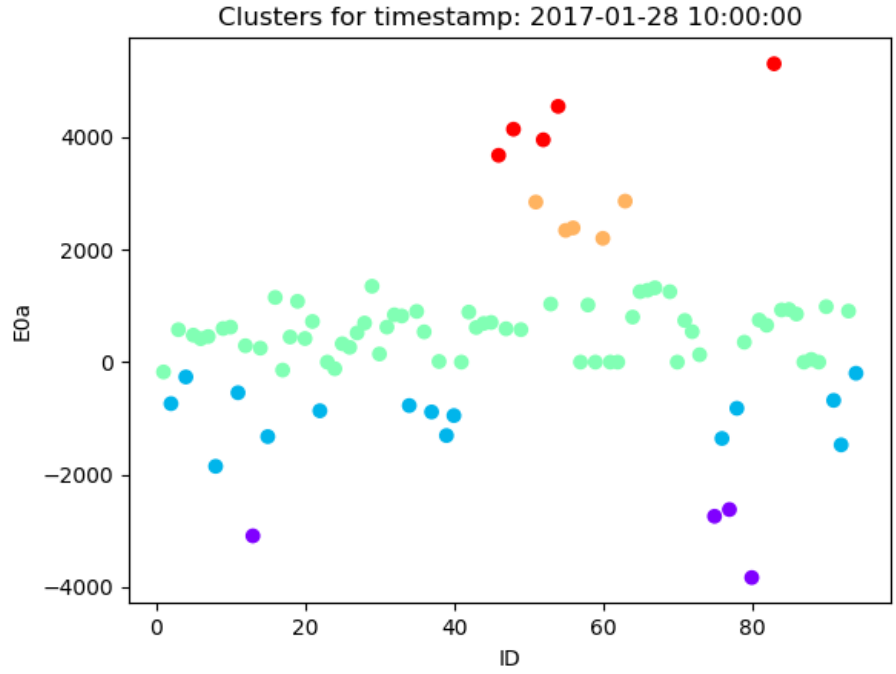

Fig. N3. CUT binning for ‘2017-01-28 10:00:00’.

Table N4. CUT binning for ‘2017-01-28 11:00:00’.

BinLabel	E0a count	Sum (Watt)
1	5	-12,199
2	66	31,380
3	14	28,384
4	5	23,654
5	1	8,473

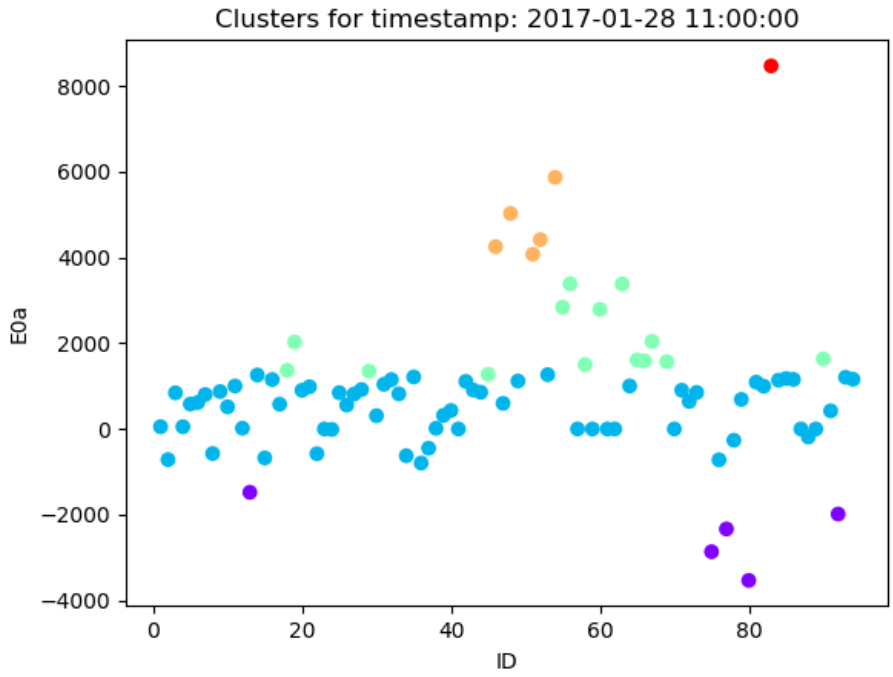

Fig. N4. CUT binning for ‘2017-01-28 11:00:00’.

Table N5. CUT binning for ‘2017-01-28 12:00:00’.

BinLabel	E0a count	Sum (Watt)
1	16	-18,165
2	67	57,902
3	6	26,325
4	1	6,450
5	1	11,998

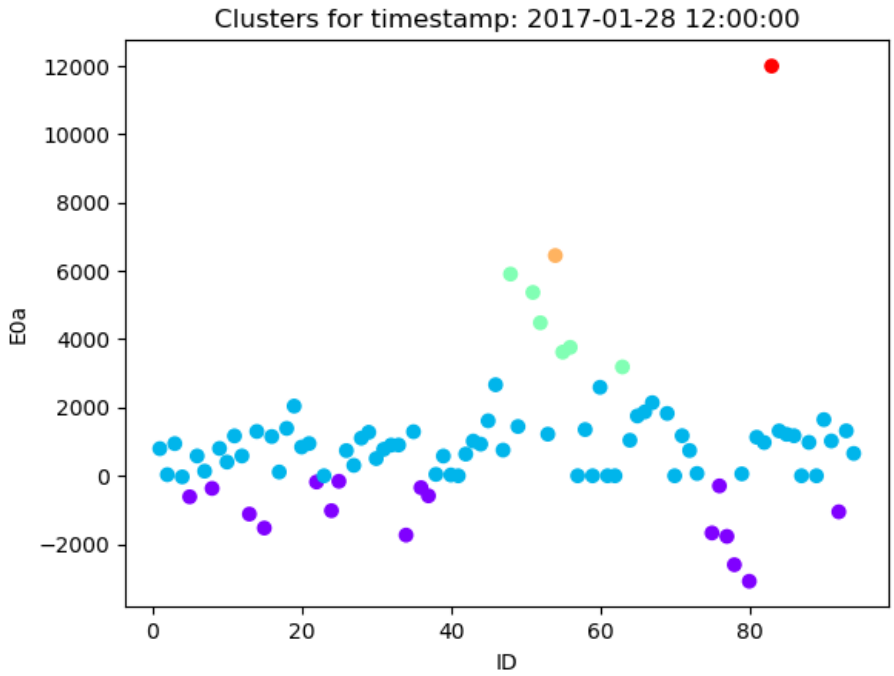

Fig. N5. CUT binning for '2017-01-28 12:00:00'.

Table N6. CUT binning for '2017-01-28 13:00:00'.

BinLabel	E0a count	Sum (Watt)
1	63	21,876
2	23	52,280
3	3	19,293
4	1	8,279
5	1	14,987

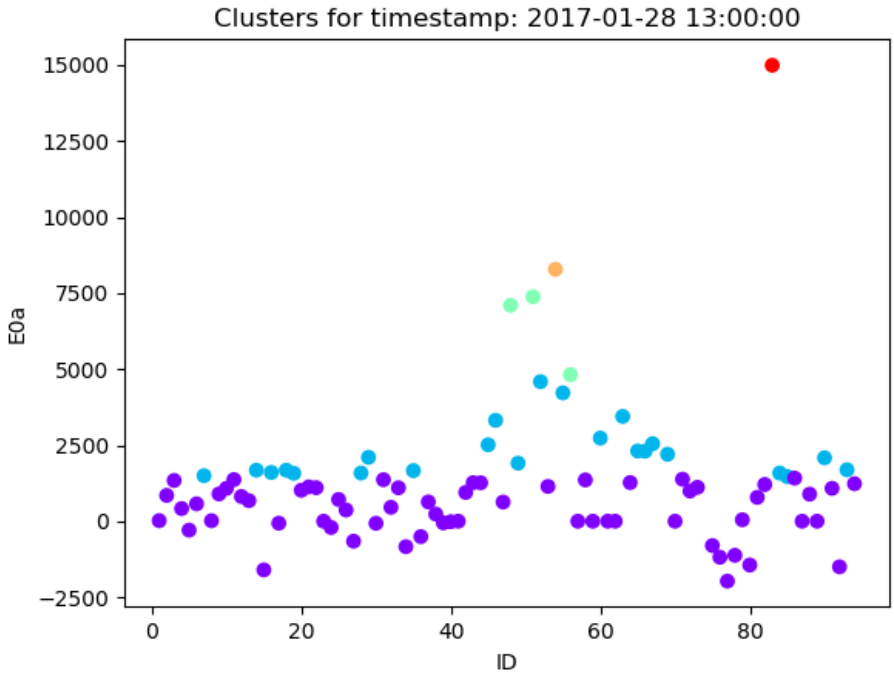

Fig. N6. CUT binning for ‘2017-01-28 13:00:00’.

Table N7. CUT binning for ‘2017-01-28 14:00:00’.

BinLabel	E0a count	Sum (Watt)
1	12	-19,246
2	72	42,440
3	3	8,810
4	3	15,936
5	1	9,663

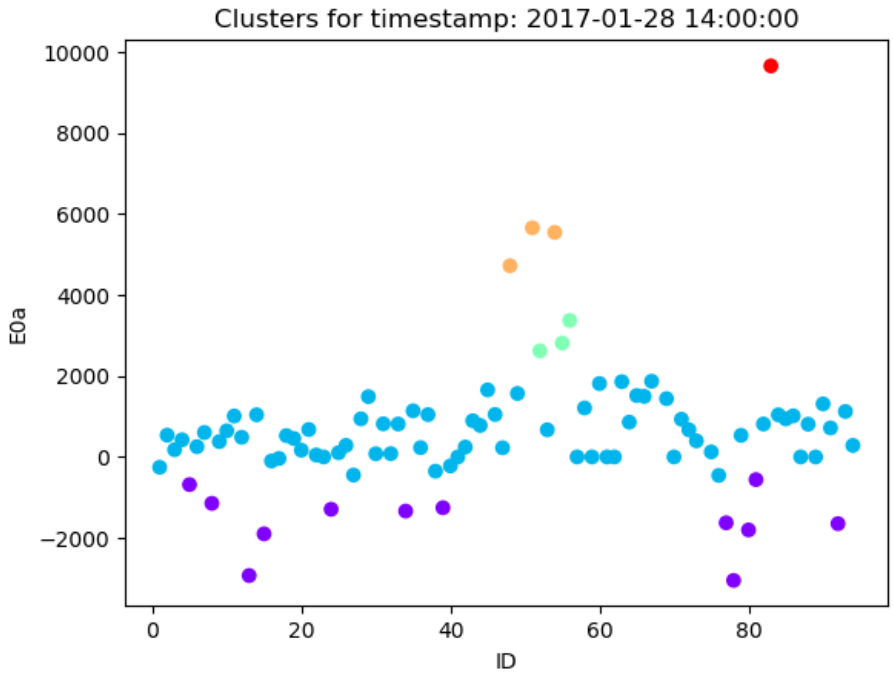

Fig. N7. CUT binning for ‘2017-01-28 14:00:00’.

Table N8. CUT binning for ‘2017-01-28 15:00:00’.

BinLabel	E0a count	Sum (Watt)
1	4	-10,298
2	23	-18,944
3	57	22,077
4	5	13,275
5	2	9,972

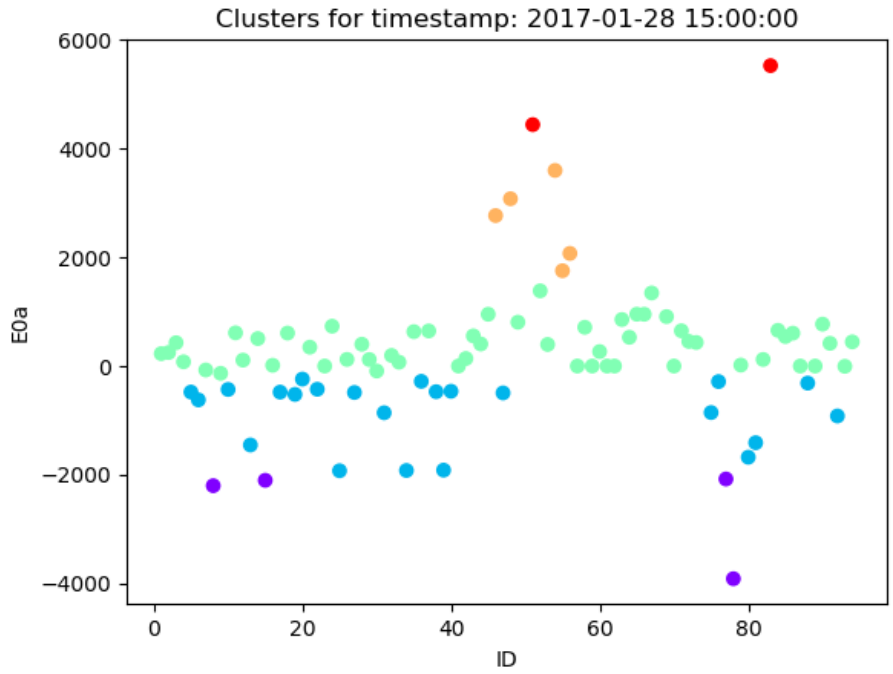

Fig. N8. CUT binning for ‘2017-01-28 15:00:00’.

Table N9. CUT binning for ‘2017-01-28 16:00:00’.

BinLabel	E0a count	Sum (Watt)
1	3	-10,031
2	8	-18,699
3	10	-12,706
4	52	-6,229
5	18	6,993

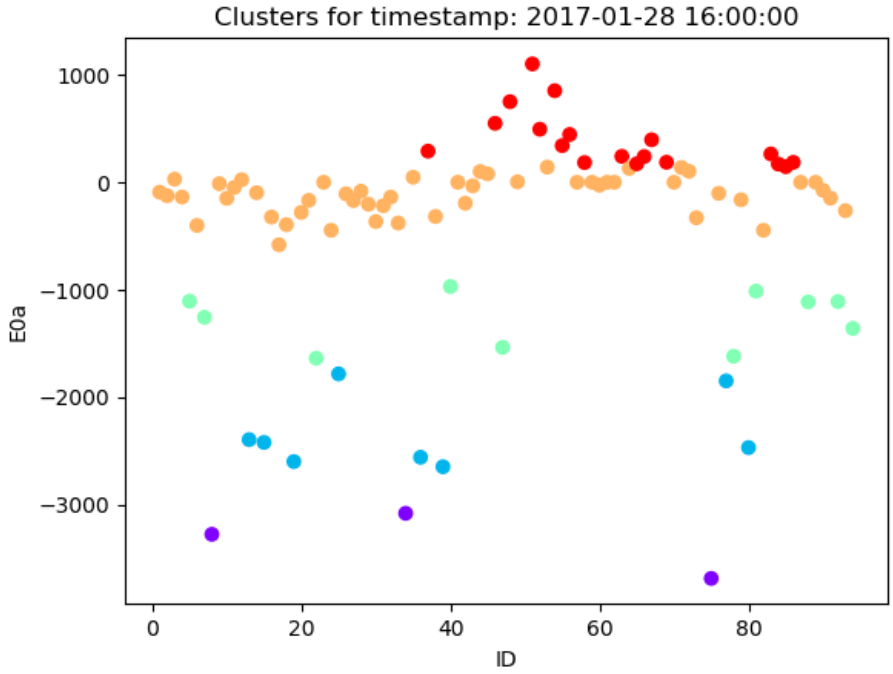

Fig. N9. CUT binning for ‘2017-01-28 16:00:00’.

Table N10. CUT binning for ‘2017-01-28 17:00:00’.

BinLabel	E0a count	Sum (Watt)
1	3	-11,699
2	4	-12,336
3	3	-5,976
4	10	-10,908
5	71	-16,806

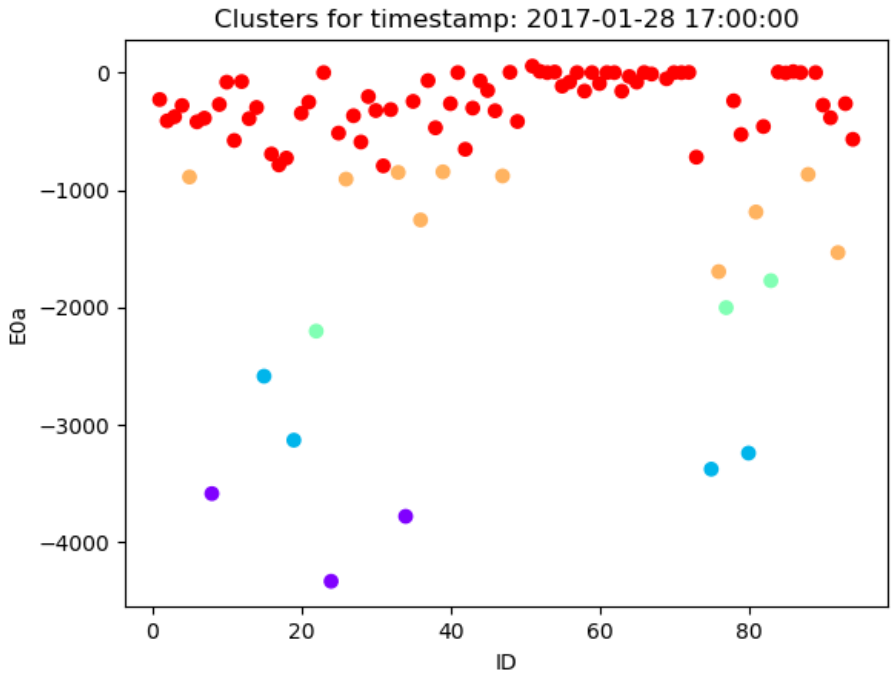

Fig. N10. CUT binning for ‘2017-01-28 17:00:00’.

Appendix O

This Appendix contains results from Ensemble utilizing Averaged Predictions (EAP).

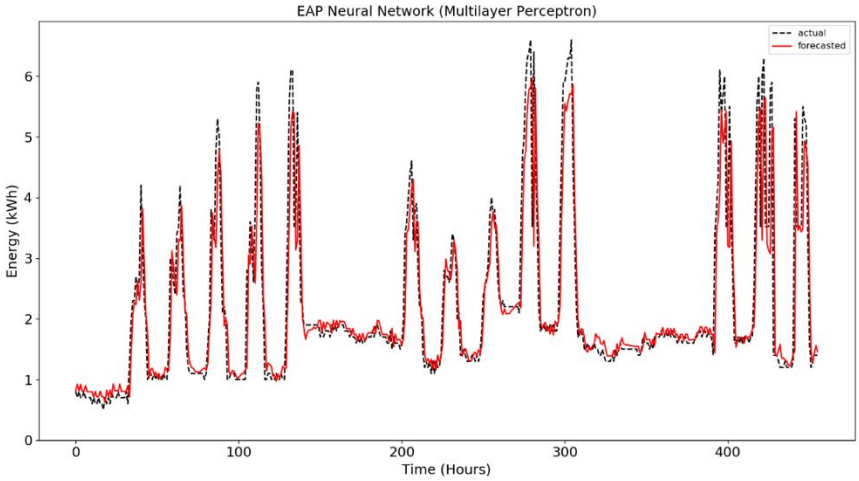

Fig. O1. EAP Neural Network (Multilayer Perceptron).

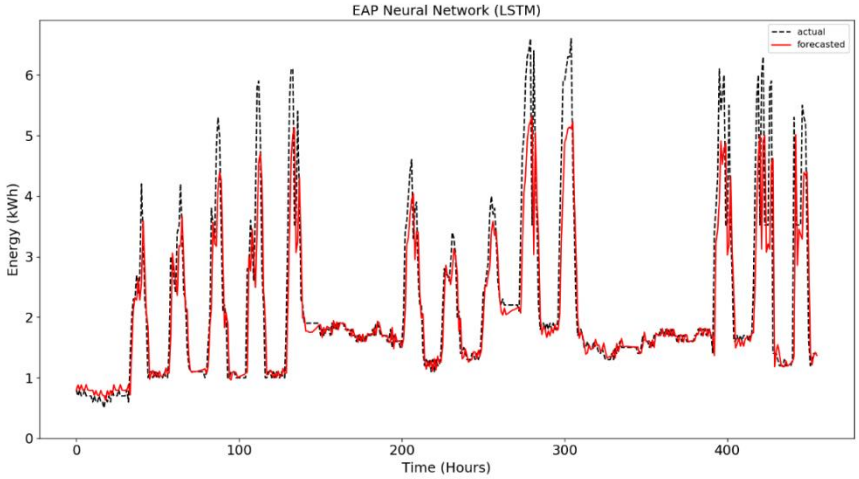

Fig. O2. EAP Neural Network (LSTM).

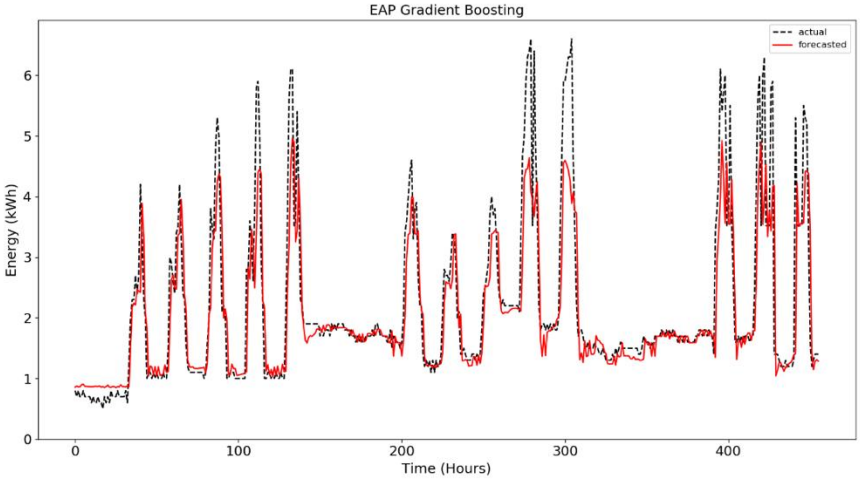

Fig. O3. EAP Gradient Boosting.

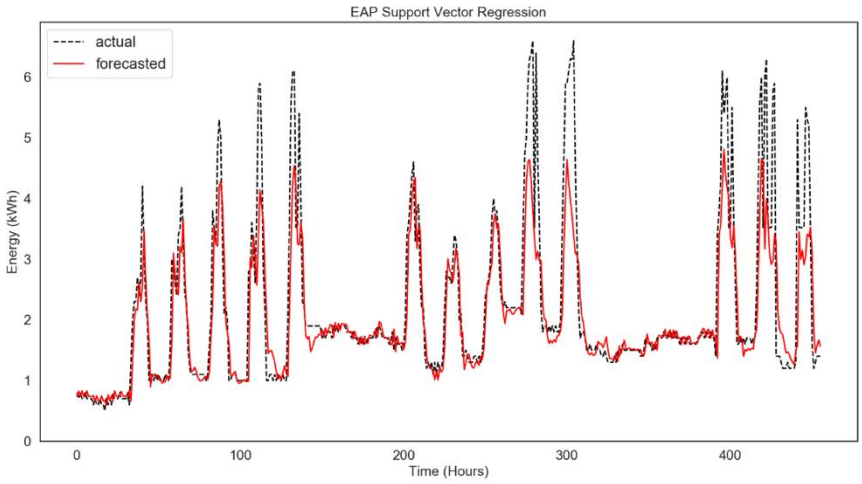

Fig. O4. EAP Support Vector Regression.

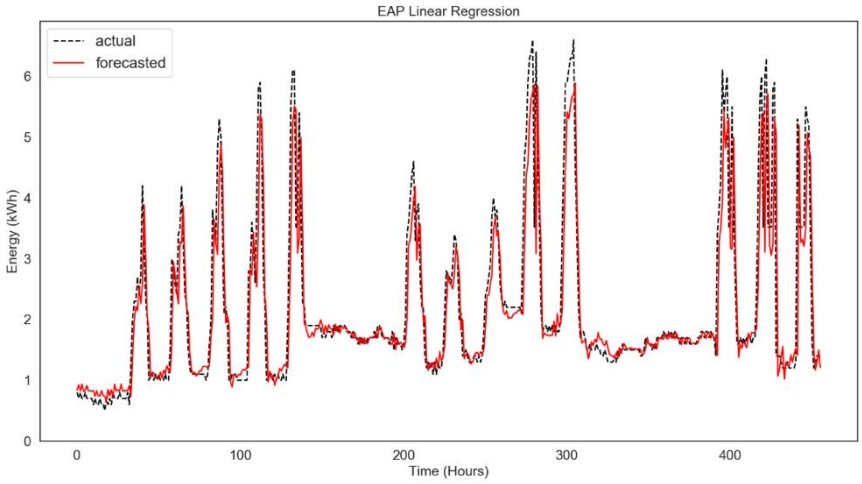

Fig. O5. EAP Linear Regression.

Appendix P

This Appendix contains results from Ensemble utilizing Weighted Averages (EWA).

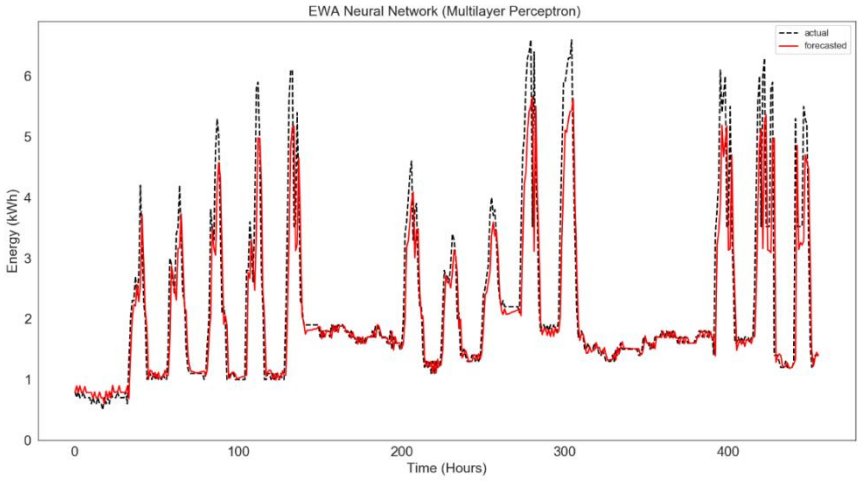

Fig. P1. EWA Neural Network (Multilayer Perceptron).

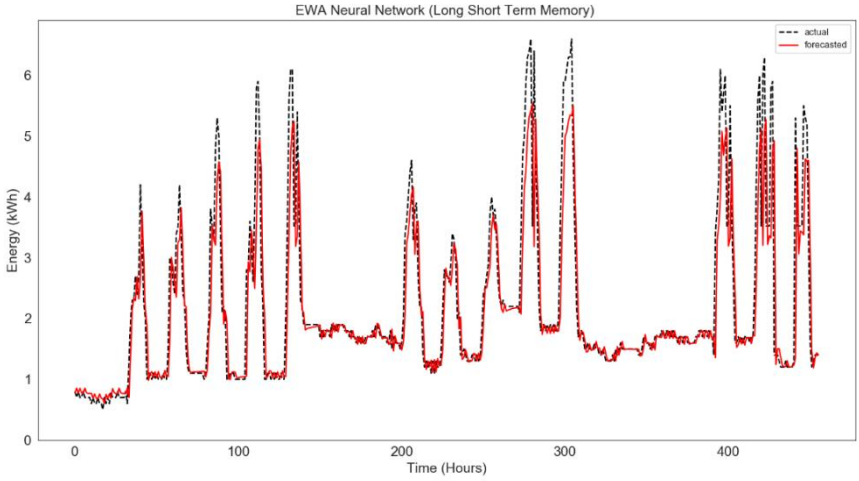

Fig. P2. EWA Neural Network (LSTM).

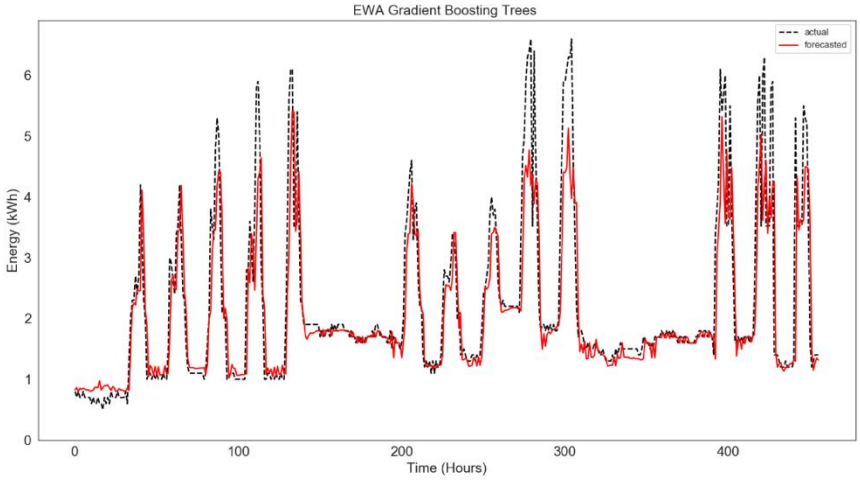

Fig. P3. EWA Gradient Boosting.

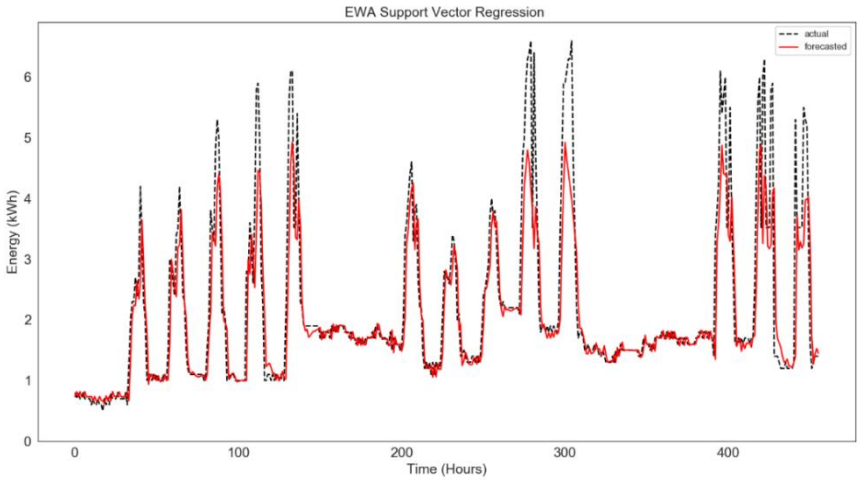

Fig. P4. EWA Support Vector Regression.

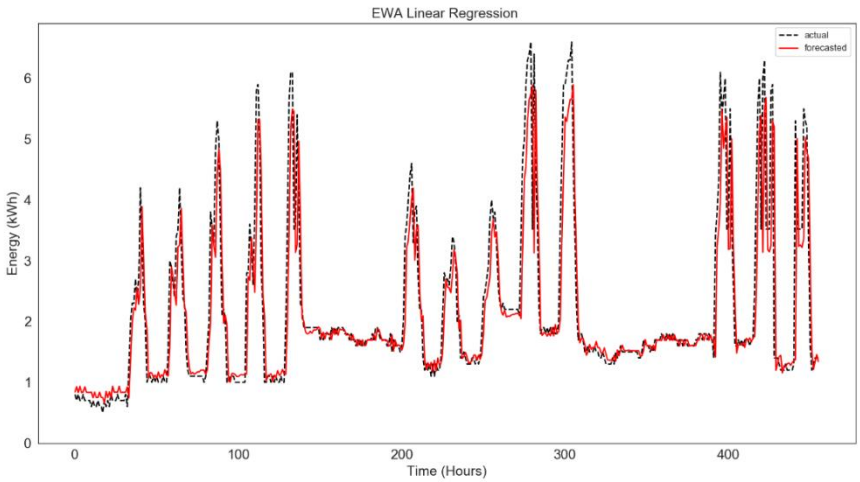

Fig. P5. EWA Linear Regression.

Appendix Q

This Appendix contains results from Ensemble utilizing Polynomial Exhibitor (EPE).

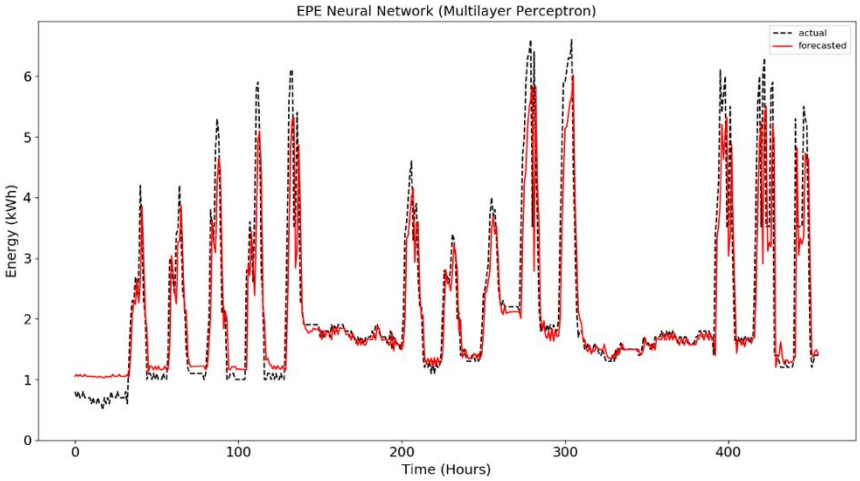

Fig. Q1. EPE Neural Network (Multilayer Perceptron).

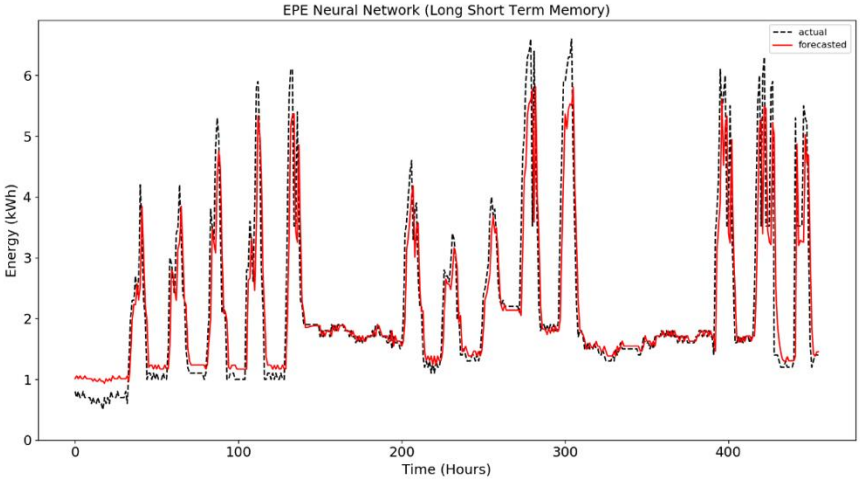

Fig. Q2. EPE Neural Network (LSTM).

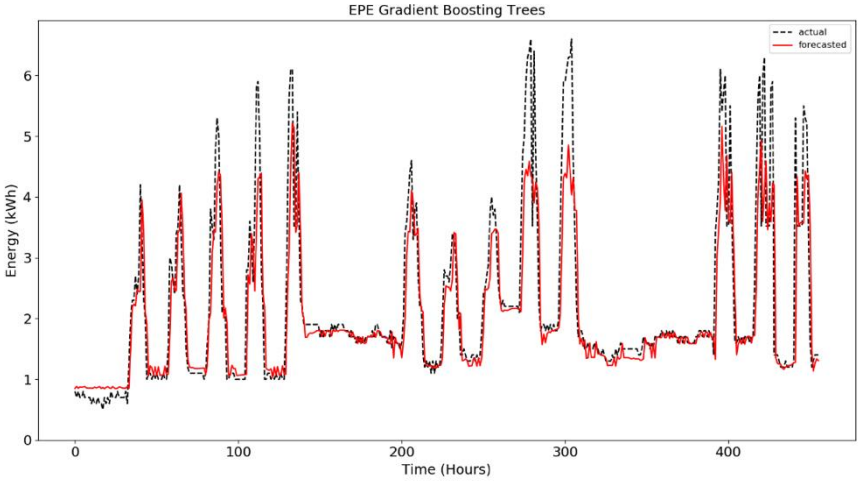

Fig. Q3. EPE Gradient Boosting.

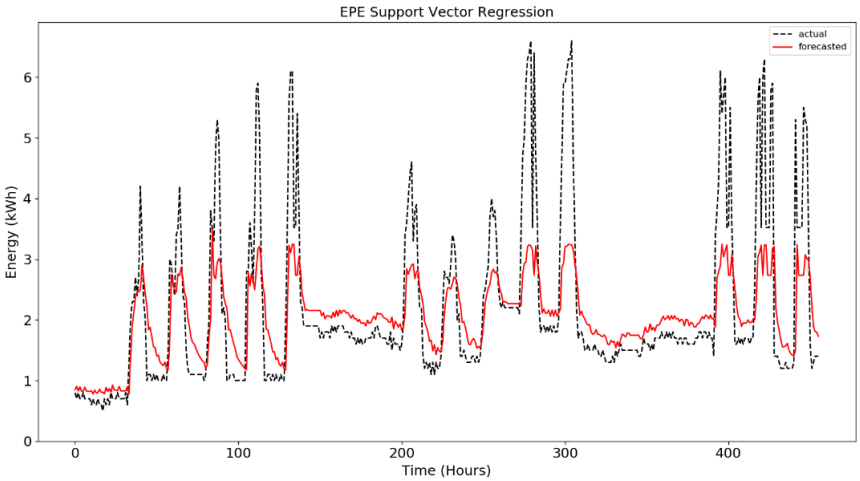

Fig. Q4. EPE Support Vector Regression.

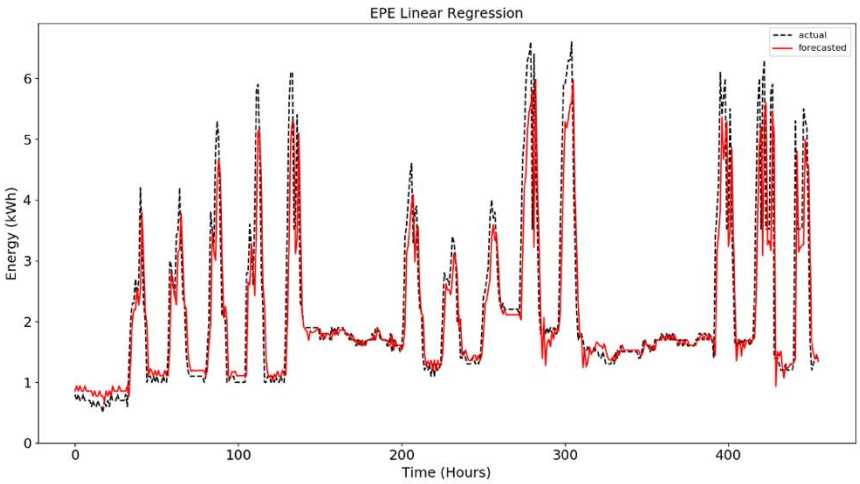

Fig. Q5. EPE Linear Regression.

Appendix R

This Appendix contains the summarized results for forecast accuracy.

Table R1. Summary of forecast accuracy.

Prediction Model	MAPE (%)	SMAPE (%)	RMSE (kWh)	ET (seconds)
Ensemble utilizing Averaged Predictions (EAP)				
MLP	**12.962**	**6.489**	**0.651**	238
LSTM	14.093	6.885	0.668	825.5
GBT	15.373	7.652	0.699	72.6
SVR	14.354	7.415	0.723	2.05
LR	14.506	7.062	0.664	**0.37**
Ensemble utilizing Weighted Averages (EWA)				
MLP	13.234	6.612	0.668	232
LSTM	15.191	7.236	0.663	960
GBT	15.116	7.488	0.698	72

SVR	13.201	6.706	0.684	2.13
LR	14.239	6.915	0.674	0.42
Ensemble utilizing Polynomial Exhibitor (EPE)				
MLP	17.298	8.113	0.686	227.9
LSTM	17.138	7.986	0.679	858.73
GBT	15.113	7.485	0.699	83.64
SVR	17.323	9.507	0.981	6.88
LR	15.293	7.433	0.686	2.6

**Best values for MAPE, SMAPE, RMSE and ET are considered the minimum values column-wise. These values are highlighted with bold font.

Appendix S

Table S1. Cost minimization and discomfort options for consumer.

t	No Opt	ND	ND Chg(%)	SD	SD Chg(%)	AD	AD Chg(%)	MD	MD Chg(%)	HD	HD Chg(%)
0	6.418	5.143	-19.866	4.493	-29.994	4.493	-29.994	4.493	-29.994	4.493	-29.994
1	6.347	5.183	-18.339	4.443	-29.998	4.443	-29.998	4.443	-29.998	4.443	-29.998
2	6.356	5.38	-15.356	4.449	-30.003	4.449	-30.003	4.449	-30.003	4.449	-30.003
3	6.386	5.432	-14.939	4.47	-30.003	4.47	-30.003	4.47	-30.003	4.47	-30.003
4	6.309	5.298	-16.025	4.416	-30.005	4.416	-30.005	4.416	-30.005	4.416	-30.005
5	6.276	5.012	-20.140	4.393	-30.003	4.393	-30.003	4.393	-30.003	4.393	-30.003
6	6.409	4.486	-30.005	4.486	-30.005	4.486	-30.005	4.486	-30.005	4.486	-30.005
7	6.951	4.866	-29.996	4.866	-29.996	4.866	-29.996	4.866	-29.996	4.866	-29.996
8	9.763	6.834	-30.001	6.834	-30.001	6.834	-30.001	6.834	-30.001	6.834	-30.001
9	30.334	27.767	-8.462	25.262	-16.721	22.851	-24.669	21.234	-29.999	21.234	-29.999
10	37.325	35.455	-5.010	33.18	-11.105	30.99	-16.973	28.778	-22.899	26.127	-30.001
11	37.855	36.18	-4.425	33.97	-10.263	31.843	-15.882	29.694	-21.559	26.499	-29.999
12	38.13	36.895	-3.239	34.831	-8.652	32.844	-13.863	30.837	-19.127	26.96	-29.295
13	36.906	35.675	-3.336	33.611	-8.928	31.626	-14.307	29.619	-19.745	25.834	-30.001
14	36.11	34.5	-4.459	32.312	-10.518	30.205	-16.353	28.077	-22.246	25.277	-30.000
15	36.691	34.821	-5.097	32.546	-11.297	30.356	-17.266	28.144	-23.295	25.684	-29.999
16	37.139	34.398	-7.380	31.834	-14.284	29.367	-20.927	26.875	-27.637	25.997	-30.001
17	37.803	34.242	-9.420	31.408	-16.917	28.68	-24.133	26.462	-30.000	26.462	-30.000

t	No Opt	ND	ND Chg(%)	SD	SD Chg(%)	AD	AD Chg(%)	MD	MD Chg(%)	HD	HD Chg(%)
18	37.065	33.172	-10.503	30.227	-18.449	27.394	-26.092	25.945	-30.001	25.945	-30.001
19	34.999	31.495	-10.012	28.68	-18.055	25.97	-25.798	24.499	-30.001	24.499	-30.001
20	13.621	11.252	-17.392	9.535	-29.998	9.535	-29.998	9.535	-29.998	9.535	-29.998
21	6.819	4.933	-27.658	4.773	-30.004	4.773	-30.004	4.773	-30.004	4.773	-30.004
22	6.673	5.05	-24.322	4.671	-30.001	4.671	-30.001	4.671	-30.001	4.671	-30.001
23	6.593	5.324	-19.248	4.615	-30.002	4.615	-30.002	4.615	-30.002	4.615	-30.002
Total	**495.28**	**448.79**	**-9.386**	**414.31**	**-16.35**	**388.57**	**-21.55**	**366.61**	**-25.98**	**346.96**	**-29.95**

Appendix T

Table T1. Partial day-ahead portfolio scheduling (10 consumers) with lower (l) and upper (u) bounds of flexibility contribution in kWh.

t	C1 (l,u)	C2 (l,u)	C3 (l,u)	C4 (l,u)	C5 (l,u)	C6 (l,u)	C7 (l,u)	C8 (l,u)	C9 (l,u)	C10 (l,u)
0	(0, 0.46)	(0, 7.86)	(0, 0.32)	(0, 1.79)	(0, 0.57)	(0, 2.04)	(0, 0.49)	(0, 0.37)	(0, 0.76)	(0, 0.62)
1	(0, 0.77)	(0, 4.72)	(0, 0.33)	(0, 0.84)	(0, 0.54)	(0, 1.66)	(0, 0.88)	(0, 0.76)	(0, 0.63)	(0, 0.69)
2	(0, 0.5)	(0, 4.75)	(0, 0.35)	(0, 1.45)	(0, 0.53)	(0, 1.91)	(0, 0.87)	(0, 0.65)	(0, 0.26)	(0, 0.42)
3	(0, 0.58)	(0, 4.15)	(0, 0.43)	(0, 0.82)	(0, 0.42)	(0, 1.05)	(0, 0.49)	(0, 0.4)	(0, 0.3)	(0, 0.72)
4	(0, 0.69)	(0, 4.96)	(0, 0.52)	(0, 1.38)	(0, 0.56)	(0, 0.62)	(0, 1.16)	(0, 0.23)	(0, 0.85)	(0, 0.56)
5	(0, 0.3)	(0, 4.1)	(0, 0.34)	(0, 1.55)	(0, 0.57)	(0, 1.12)	(0, 1.2)	(0, 0.39)	(0, 0.78)	(0, 0.84)
6	(-0.75, 0)	(-4.47, 0)	(-0.33, 0)	(-2.18, 0)	(-0.9, 0)	(-2.53, 0)	(-1.16, 0)	(-0.61, 0)	(-0.61, 0)	(-0.92, 0)
7	(-1.41, 0)	(-4.5, 0)	(-0.37, 0)	(-1.92, 0)	(-1.1, 0)	(-2.26, 0)	(-1.43, 0)	(-0.47, 0)	(-0.58, 0)	(-0.96, 0)
8	(-5.3, 0)	(-5.1, 0)	(-0.34, 0)	(-2.24, 0)	(-1.03, 0)	(-1.75, 0)	(-1.7, 0)	(-1.39, 0)	(-0.85, 0)	(-1.23, 0)
9	(-3.6, 0)	(-4, 0)	(-0.43, 0)	(-1.34, 0)	(-1.42, 0)	(-1.69, 0)	(-1.99, 0)	(-1.24, 0)	(-0.79, 0)	(-0.79, 0)
10	(-4.83, 0)	(-3.16, 0)	(-0.52, 0)	(-6.59, 0)	(-1.1, 0)	(-2.27, 0)	(-1.46, 0)	(-1.56, 0)	(-0.83, 0)	(-1.1, 0)
11	(0, 1.82)	(0, 3.44)	(0, 0.33)	(0, 2.18)	(0, 1.23)	(0, 2.09)	(0, 0.73)	(0, 0.44)	(0, 1)	(0, 0.87)
12	(0, 3.29)	(0, 7.53)	(0, 0.33)	(0, 4.43)	(0, 2.32)	(0, 4.58)	(0, 1.5)	(0, 1.38)	(0, 1.99)	(0, 0.83)
13	(0, 3.71)	(0, 7.43)	(0, 0.39)	(0, 3.71)	(0, 1.85)	(0, 4.8)	(0, 1.19)	(0, 1.78)	(0, 2.5)	(0, 2.24)
14	(0, 5.45)	(0, 9.83)	(0, 0.22)	(0, 5.32)	(0, 1.41)	(0, 4.81)	(0, 1.02)	(0, 0.87)	(0, 2.8)	(0, 0.98)
15	(-2.89, 0)	(-5.95, 0)	(-0.07, 0)	(-4.1, 0)	(-1.14, 0)	(-2.48, 0)	(-0.67, 0)	(-1.41, 0)	(-1.24, 0)	(-1.88, 0)
16	(-3.01, 0)	(-5.87, 0)	(-0.23, 0)	(-2.74, 0)	(-1.38, 0)	(-2.64, 0)	(-0.56, 0)	(-0.96, 0)	(-1.86, 0)	(-1.8, 0)
17	(-3.16, 0)	(-6.04, 0)	(-0.3, 0)	(-3.74, 0)	(-1.06, 0)	(-3.35, 0)	(-0.81, 0)	(-0.58, 0)	(-2.28, 0)	(-1.96, 0)

t	C1 (l,u)	C2 (l,u)	C3 (l,u)	C4 (l,u)	C5 (l,u)	C6 (l,u)	C7 (l,u)	C8 (l,u)	C9 (l,u)	C10 (l,u)
18	(-3.74, 0)	(-6.1, 0)	(-0.28, 0)	(-3.41, 0)	(-0.53, 0)	(-2.83, 0)	(-0.6, 0)	(-0.7, 0)	(-2.77, 0)	(-1.43, 0)
19	(-0.47, 0)	(-6.66, 0)	(-0.25, 0)	(-2.22, 0)	(-0.41, 0)	(-4.21, 0)	(-0.76, 0)	(-1.51, 0)	(-2.23, 0)	(-0.89, 0)
20	(-0.58, 0)	(-8.46, 0)	(-0.28, 0)	(-2.03, 0)	(-0.64, 0)	(-3.9, 0)	(-0.62, 0)	(-1.56, 0)	(-1.57, 0)	(-0.63, 0)
21	(-0.5, 0)	(-9.41, 0)	(-0.21, 0)	(-3.56, 0)	(-0.43, 0)	(-3.79, 0)	(-0.34, 0)	(-0.9, 0)	(-1.57, 0)	(-0.41, 0)
22	(0, 0.88)	(0, 10.64)	(0, 0.3)	(0, 4.36)	(0, 0.5)	(0, 3)	(0, 0.95)	(0, 1.31)	(0, 2.2)	(0, 0.4)
23	(0, 0.88)	(0, 7.79)	(0, 0.43)	(0, 4.89)	(0, 0.46)	(0, 4.11)	(0, 0.98)	(0, 0.47)	(0, 1.58)	(0, 0.54)