DELFT UNIVERSITY OF TECHNOLOGY

MASTER THESIS

A Knowledge Discovery Framework for Understanding Energy Consumption Behavior using Social Data

Author:
Arkka DHIRATARA

Supervisor:
Dr. Ir. Alessandro BOZZON
Dr Achilleas PSYLLIDIS

A thesis submitted in fulfillment of the requirements for the degree of Master of Computer Science

in the

Web Information Systems
Department of Software Technology

September 14, 2017

Declaration of Authorship

I, Arkka DHIRATARA, declare that this thesis titled, "A Knowledge Discovery Framework for Understanding Energy Consumption Behavior using Social Data" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

Delft University of Technology

Abstract

Faculty Electrical Engineering, Mathematics and Computer Science
Department of Software Technology

Master of Computer Science

A Knowledge Discovery Framework for Understanding Energy Consumption Behavior using Social Data

by Arkka DHIRATARA

Understanding energy consumption behavior provide an insightful knowledge to improve energy efficiency, promote energy conservation, and importantly sustain the human life. However, currently energy consumption data are being gathered by (smart) energy meters at the household level or through surveys. While gathering data using smart meter is highly reliable, it lacks semantic information about how energy is consumed (e.g. using appliance). On the other hand, survey allow to gather semantically rich data, but the acquisition of the data is labor-intensive.

In this context, social media data data (e.g. twitter, instagram) which are semantically rich and publicly available can be used as an alternative source of data about energy consumption behavior. However, due to the noisy and ambiguous nature of social media data, the extraction of energy related information from micro posts is very challenging. The aim of this thesis is to introduce a general framework to discover knowledge about energy consumption behaviors from social media data. The framework explores the suitable of social media data as an alternative data source for capturing energy consumption behaviors, and thus to be used to complement conventional data sources. Using the state-of-the-art methods and approaches in social media data analytics field, we compose the framework which structured into three main stages: data collection, data enrichment & processing, and data analysis & visualization.

To study the performance of our framework, we set up an experiment aiming at identifying energy consumption behavior patterns in two different world cities: Jakarta (Indonesia) and Amsterdam (The Netherlands). On data collection stage, we collected 1,306,336 tweets from both cities. Next, on data enrichment & processing stage, we pre-processed the collected tweets and conduct dictionary-based annotation using our 8,329 energy consumption related terms. As a result, we identified 509,471 tweets (39%) of the corpus as energy consumption related tweets, which categorize into four different energy consumption behaviors: food, dwelling, mobility and leisure. Using the annotated streams as noisy datasets, we implement distant supervision machine learning technique using binomial classifier to identify energy consumption related tweets. Following this approach, we are able to achieve good classifier's performance on identifying energy consumption related tweets. Finally, on data analysis & visualization stage, we conduct statistical analysis and found strong positive correlation ($r = 0.73$) between energy consumption data extracted from social media and actual electricity load. Following this result, we show that social media data has the potential to be used as supplementary source of information for energy consumption studies.

Contents

Declaration of Authorship **iii**

Abstract **v**

Contents **vii**

List of Figures **xi**

List of Tables **xiii**

Preface **xv**

1 Introduction **1**
1.1 Energy Consumption . . . 2
1.2 Research Questions . . . 3
1.3 Scope . . . 4
1.4 Methods . . . 4
1.5 Contributions . . . 6
1.6 Outline . . . 6

2 Related Work **9**
2.1 Measure Energy Consumption . . . 9
2.2 Behavior Pattern in Social Data . . . 10
2.3 Discover Energy Consumption using Social Media . . . 12

3 Knowledge Discovery Framework **15**
3.1 Data Collection . . . 15
3.2 Data Enrichment and Processing . . . 16
3.2.1 Dictionary Generation and Processing . . . 16
3.2.2 Social Media Stream Processing . . . 18
3.2.3 Dictionary-based Stream Annotation . . . 18
3.2.4 Feature Extraction . . . 19
3.2.5 Machine-Learning-based Topic Detection . . . 20
3.3 Data Analysis and Visualization . . . 20
3.3.1 Text Analysis . . . 20
3.3.2 Heatmaps Visualization . . . 21
3.3.3 Displacement distribution . . . 21
3.3.4 Radius of gyration . . . 22
3.3.5 Temporal Analysis . . . 22

4 Implementation **25**
4.1 Cluster-computing Platform: Apache Spark . . . 26
4.2 System Architecture: SocioKnowledge . . . 27
4.2.1 Study Module . . . 27

4.2.2 Dictionary Module . . . 28
4.2.3 Stream Module . . . 29
4.2.4 Dataset Module . . . 30

5 Experiments and Results: Energy Consumption Behavior 31
5.1 Dictionary: Energy Consumption . . . 31
5.1.1 Dictionary Expansion using ConceptNet . . . 32
5.1.2 Dictionary Quality Assessment . . . 33
5.2 Stream: Twitter . . . 34
5.3 Matching Stream with Energy Consumption Dictionary . . . 35
5.4 Feature Extraction: TF-IDF with Hashing Trick and Word2Vec . . . 36
5.5 Distant Supervision Machine Learning . . . 37
5.5.1 Dataset Preparation . . . 38
5.5.2 Classifiers: Setup . . . 39
Dataset sample size . . . 39
Features: TF-IDF vs Word2Vec . . . 40
Classification algorithm: Logistic Regression vs SVMs . . . 40
Threshold tuning . . . 40
5.5.3 Classifiers: Energy Consumption in Amsterdam and Jakarta . . . 41
5.5.4 Classifiers vs Dictionary-based Annotation . . . 42
5.6 Analysis . . . 42
5.6.1 Temporal Analysis: Electricity Load Correlation . . . 44
5.6.2 Spatial Analysis and Visualization . . . 46
5.6.3 Heat Maps . . . 46
User Displacement distribution . . . 48
Radius of Gyration . . . 48

6 Discussions and Conclusions 53
6.1 Discussions . . . 53
6.1.1 Discover Energy Consumption Behaviors from Social Data . . . 53
6.1.2 Dictionary Extensiveness . . . 53
6.2 Threat of Validity . . . 54
6.2.1 Social Data as Data Source . . . 54
6.2.2 Ambiguity of Energy Consumption Behaviors in Social Data . . . 54
6.2.3 Geo-tagged Information in Social Data . . . 55
6.2.4 Limited Classifier Algorithms . . . 55
6.2.5 Limited Baseline Data for Evaluation . . . 55
6.3 Conclusion . . . 55
6.3.1 Research Questions . . . 56
RQ1: What is the state-of-the art in social media data analytics on discovering energy consumption in urban contexts? 56
RQ2: How could we design, develop, and implement a knowledge discovery framework to identify energy-related consumption behavior of people in cities using social data? . . . 56
RQ3: What is the performance of the proposed framework in various energy-related use cases? . . . 56
6.4 Outlook . . . 56
6.4.1 Specific Energy Consumption Activity Type . . . 57
6.4.2 Broadening the classifier algorithm scope . . . 57

A Dictionary: Energy Consumption **59**
A.1 Valid ConceptNet Relationships 59
A.2 Hashing Trick Algorithm 60

Bibliography **61**

List of Figures

1.1 CODALoop a Feedback-Loop model 3
1.3 Research outline 8

3.1 The Knowledge Discovery Framework 15
3.2 Dictionary Generation and Processing 17
3.3 Stream Processing 18
3.4 Dictionary-based Matching Workflow 19

4.1 Apache Spark Architecture 26
4.2 SocioKnowledge Architecture 27
4.3 Dictionary Module Classes 28
4.4 Stream Module Classes 29

5.2 Collected Twitter Data by City 35
5.3 Match vs Not Match Energy Consumption Dictionary 36
5.4 Dictionary Matched Streams by Class 37
5.6 Classifier: Dataset Sample Size 39
5.7 ROC Curves: TF-IDF vs Word2Vec Features 40
5.8 ROC Curves: Logistic Regression vs SVMs 41
5.9 EC Classifiers ROC Curve 42
5.10 EC-LEISURE Classifiers ROC Curve 43
5.11 EC-MOBILITY Classifiers ROC Curve 43
5.12 Tennet: The Netherlands Actual Electricity Load 45
5.14 Leisure Energy Consumption in Amsterdam 47
5.15 Leisure Energy Consumption in Jakarta 47
5.16 Mobility Energy Consumption in Amsterdam 47
5.17 Mobility Energy Consumption in Jakarta 48
5.18 User Displacement Distribution in Amsterdam 49
5.19 User Displacement Distribution in Jakarta 49
5.20 Radius of Gyration Distribution in Amsterdam 50
5.21 Radius of Gyration Distribution in Jakarta 50

List of Tables

2.1 Related Work on Behavior Pattern in Social Data 11
2.2 Related Work on Discover Energy Consumption using Social Data . . 13

3.1 Different Methods Discover Energy Consumption Behaviors using Socia Data 21

5.1 Energy Consumption Seed Dictionary by Source 32
5.2 Expanded Energy Consumption Dictionary by Source 34
5.3 Matching Algorithms Performance 36
5.4 Datasets 39
5.5 Binomial Classifier Threshold Tuning 41
5.6 Classifier Performances 44
5.7 Dictionary-based Annotation Performances 45
5.8 Descriptive Statistics: Actual Electricity Load vs Energy Consumption Streams 46

A.1 Valid ConceptNet Relationships 59

Preface

This document is my Master thesis that concludes my two years as a Computer Science student at Delft University of Technology. Throughout working on this project for almost a year, I have a great experience that allowed me to grow both scientifically as well as personally. Working on this thesis is definitely more challenging and time-consuming than my previous bachelor final project back in 2010. Looking back at the past month, I am very proud that I can finish this thesis. However, this work would be never have been possible without the valuable advice and support of the many people that I would like to thank in this preface.

First, I am deeply grateful to my scholarship providers, Lembaga Pengelola Dana Pendidikan (LPDP) for giving me the opportunity to study abroad, broaden my horizon, and pursue my dreams.

I would like to thank my supervisor, Dr. Ir. Alessandro Bozzon and Dr. Achilleas Psyllidis, for the opportunity to conduct this project. Throughout the project, you have always challenged me with critical comments, provide helpful suggestions, and most importantly for giving me the freedom to explore the new topics and ideas.

I would like to express my deepest gratitude to Sepideh Mesbah for the numerous discussion and exchange of ideas. Your critical thoughts have certainly helped me grow and made my analyses more sound.

My sincere gratitude to the members of my master thesis committee, Prof. Dr. Ir. Geert-Jan Houben and Dr. M.M. (Mathijs) de Weerdt for their willingness and availability to review this thesis, as well as for honoring me by serving on my thesis committee.

This last paragraph I have reserved for my family, my wife, Saras, my mother, July, my brother, Dhika, and my Sister, Nadhya whom always be part of my support system. Thank you for all of your intangible support that made what I am today. This master thesis is dedicated to you. Also, I would like to dedicate this thesis in memory of my father, Suryo, whose always inspire me to pursue study in computer science and studying higher education abroad. I believe that you must be very proud of my achievement and that I have fulfilled my promise.

Arkka Dhiratara
Delft, The Netherlands
September 14, 2017

Chapter 1

Introduction

Energy consumption is one of human activity that directly impact the sustainability of our planet. However, many people are not aware of earth depleting supply of non-renewable energy resources being used for their daily activities. End-user energy consumption behavior is an important factor that shape a sustainable energy system (Alrowaily and Kavakli, 2015). Understanding energy consumption behavior provide an insightful knowledge to improve energy efficiency, promote energy conservation, and importantly sustain the human life. Existing conventional energy consumption data are being gathered by (smart) energy meters at the household level or surveys. Gathering data using smart meter is highly reliable, but lacks semantic information about how energy is consumed. On the other hand, survey allow to gather semantically rich data, but the acquisition of the data is labor-intensive and expensive.

Social media has become foundational component of modern human interactions. Social media enables us to easily create and share information to our friends, colleagues or any other social circles. We can share our thoughts on Twitter, share music or movie that we enjoy on Path, share our daily commute struggle on Waze, share our workout activity on Nike+ and even share a photo of our lunch menu on Instagram. These social media platforms generate abundance amount of social media data, which contains rich information about human activities.

Studying human activity enable us to have a better understanding about the collective characteristics of people. Human activity is a broad topic; which comprise of all kind of activities that human do or cause to happen. This is where social media data can provide significance contribution by providing rich information about human activities. Following the potential of studying human activity using social media, there are already several research attempts has been conducted on this topic. For instance, understanding human mobility behavior to identify potential traffic congestion Pan et al., 2013, studying social connectivity to have a better marketing strategy Bampo et al., 2008, or even studying human interaction to detect a disease outbreak Chew and Eysenbach, 2010. These examples show that social media is able to deliver an insightful knowledge about collective human activity characteristic for different fields. Despite considerable studies have been conducted on specific kind of human activities, there is few specific research on studying energy consumption behavior using social media (Abbar, Mejova, and Weber, 2015; Alrowaily and Kavakli, 2015; Soytas, Sari, and Ewing, 2007; Bodnar et al., 2017). Understanding energy consumption could provide useful insight about how people consume the energy and also the underlying influences of that behavior.

This research aims to explore the potential usefulness of social media data as supplementary source of information about energy consumption. We propose a framework that incorporates steps and methodologies necessary to make use social media data for energy consumption studies. Furthermore, this research also

provides technical implementation of the framework as a data pipeline. The experiment conducted using collected twitter data from two word cities, which are Jakarta (Indonesia) and Amsterdam (The Netherlands).

1.1 Energy Consumption

Energy played a pivotal role in the development and evolution of human society. Today's modern civilization is heavily dependent on different forms of energy to perform our daily human activities, for instance, turn on the light with electricity, moving a car with fuel, and above all sustain human live with food. Energy is generated from different energy resources, such as fossil fuels, nuclear fuels or renewable energy. Unfortunately, nowadays most energy is generated from a non-renewable resource, whereas oil, coal, gas and nuclear are 90.43% of total world energy resources consumption (World Energy Council, 2016). Based on this fact, it is clear that our current energy consumption is not sustainable for generations to come.

Preserving energy can be done by consuming energy more efficiently and more prudently. Revilla, 2016 proposed a concept of sustainable energy lifestyle that consists of energy-consuming practices that allow people to meet their personal needs and aspirations, with energy sustainability into account. These energy-consuming practices might be seen as trivial for energy consumption efficiency, but if everyone is doing it the impact will be significant. For example, using public transport instead of cars, choosing nearby destinations, exercising more outdoors and using less energy intensive gadgets. These practices show that people are able to fulfill their personal needs, but using less energy in the process and less ecological footprint.

Transforming existing energy consumption lifestyle into more sustainable is not an easy task because providing people with the knowledge about their energy consumption in order to use energy responsibly is difficult. Following this challenges, since 2016 The Community Data-Loops for energy-efficient urban lifestyles (CODALoop)[1] project is initiated. This project aims at enabling behavior change in energy use through a feedback-loop model. CODALoop consider four energy consumption behaviors:

- **Dwelling**, the energy necessary to satisfy the demand of activities within the home and consequently the appliances used to perform those activities.
- **Mobility**, the energy necessary to bring one person from one place of activity to another place of activity.
- **Food Consumption**, the energy necessary to bring food to our plate, taking into account the energy consumed at all the stages of the food chain (production, processing, distribution, consumption and waste) but focusing on the consumption phase since we are analyzing lifestyles at the city scale.
- **Leisure**, the energy necessary to perform the activities that provide people with the enjoyment needed to have a satisfying life.

As a data-driven social learning model and behavior adaptation feedback loops, CODALoop implementation's performance depends heavily on the availability of energy consumption data. As depicted in Figure 1.1, we could see that energy-related data is present in all feedback loops. However, as discussed before, acquisition for energy consumption data using conventional data sources is challenging. As

[1] http://jpi-urbaneurope.eu/project/codaloop/

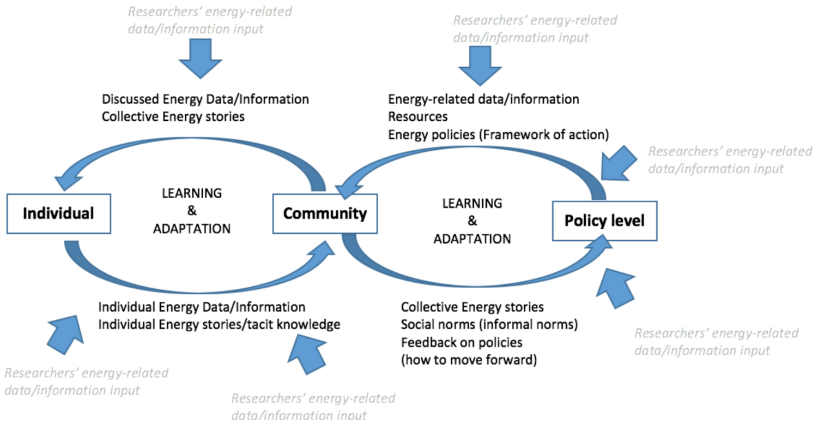


FIGURE 1.1: CODALoop as Feedback-Loop model.

a consequence, an alternative data source for energy consumption is required. This is one of domain specific problem that social media data has potential for usage.

This thesis explore the suitable of social media data as a supplementary data source for energy consumption data.

1.2 Research Questions

In relation to challenging addressed by this research, we have to define the main research question and the research sub-question as follows:

- **MRQ.** *How to discover behavioral patterns of people regarding energy consumption in cities, using social media data?*

  Social media provide an alternative source of data for studying human activity. However, due to to the unstructured nature of social media data compared to conventional data source (i.e. smart meters, surveys, sensors), different approaches and methodologies are required to collect, process, and analyze such knowledge.

- **RQ1.** *What is the state-of-the-art in social media data analytics on discover energy consumption in urban contexts?*

  As the first step in designing knowledge discovery framework for understanding energy consumption behavior using social media, we are required to conduct literature studies on several related previous studies that use social media data to study human activity patterns that require energy consumption. Based on this study, we are able to identify common approaches and methods being used and use these as the basis of our framework.

- **RQ2.** *How could we design, develop, and implement a knowledge discovery framework to identify energy-related consumption behavior of people in cities using social data?*

  We compose approaches and methods required to effectively identify energy related consumption behavior from social data.

- **RQ3.** *What is the performance of the proposed framework in various energy-related use cases?*

Following previous research question, we are required to implement the framework and test the performance of the framework in various energy consumption domains, which are mobility and leisure.

1.3 Scope

This thesis propose a general framework to discover knowledge using social data. As a scope, the framework specifically explore the extent which social data can be used as supplementary sources of information about energy consumption behavior patterns. There are different kind of energy consumption behaviors. In this thesis, we choose two different energy consumption sectors: leisure and mobility. To our knowledge, there have been few studies done specifically on energy consumption in leisure sectors. Moreover, we choose mobility sectors because based on Eurostat[2], transportation accounts for 33% of the total energy consumption, the highest among other sectors. By using this framework, we aims to deliver an alternative dataset for energy consumption studies, which can be used for further analysis.

As part of the experimentation, we implement the framework using social data gathered from Twitter within two weeks period (12-26 February 2017) in two world cities: Amsterdam (The Netherlands) and Jakarta (Indonesia).

1.4 Methods

In order to discover knowledge about energy consumption behavior patterns from social data, we require to identify energy consumption related posts from social data streams. As classification problem, we turn to machine learning to build an automated classifier. Building a classifier require both training and test dataset using a labeled dataset, consists of energy consumption and non-energy consumption related posts. However, energy consumption information is not explicitly available as an attribute in social data. Hence, further preprocessing of social data stream is required to generate such labeled dataset.

Social data is collected from the social media platform Twitter. Twitter is chosen due its popularity in both cities, Amsterdam and Jakarta. Moreover, their public API also provide capability to listen the entire posts (tweets) stream in their platform. Tweets are collected using Stream API using geographic bounding box within two cities and without any keyword filtering. Based on this collected tweets, we conduct a dictionary-based annotation to identify energy consumption related tweets. The dictionary that being used is initially generated manually by identifying energy consumption related terms, and then by including several external sources of dictionary. In order to enrich our dictionary, we further expand it using terms from ConceptNet. Dictionary expansion is conducted by using existing dictionary as a seed to discover any semantic related terms.

Before conducting dictionary-based annotation, we require to pre-process both dictionary and streams. This preprocessing is important to improve dictionary-based annotation performance, specifically on matching social data content's text (tweets) with the dictionary. We experiment with several pre-processing techniques, such as tokenization, remove stopwords, stemming and w-shingle. Dictionary-based annotation is conducted using two matching algorithm, exact matching and partial

[2]http://ec.europa.eu/

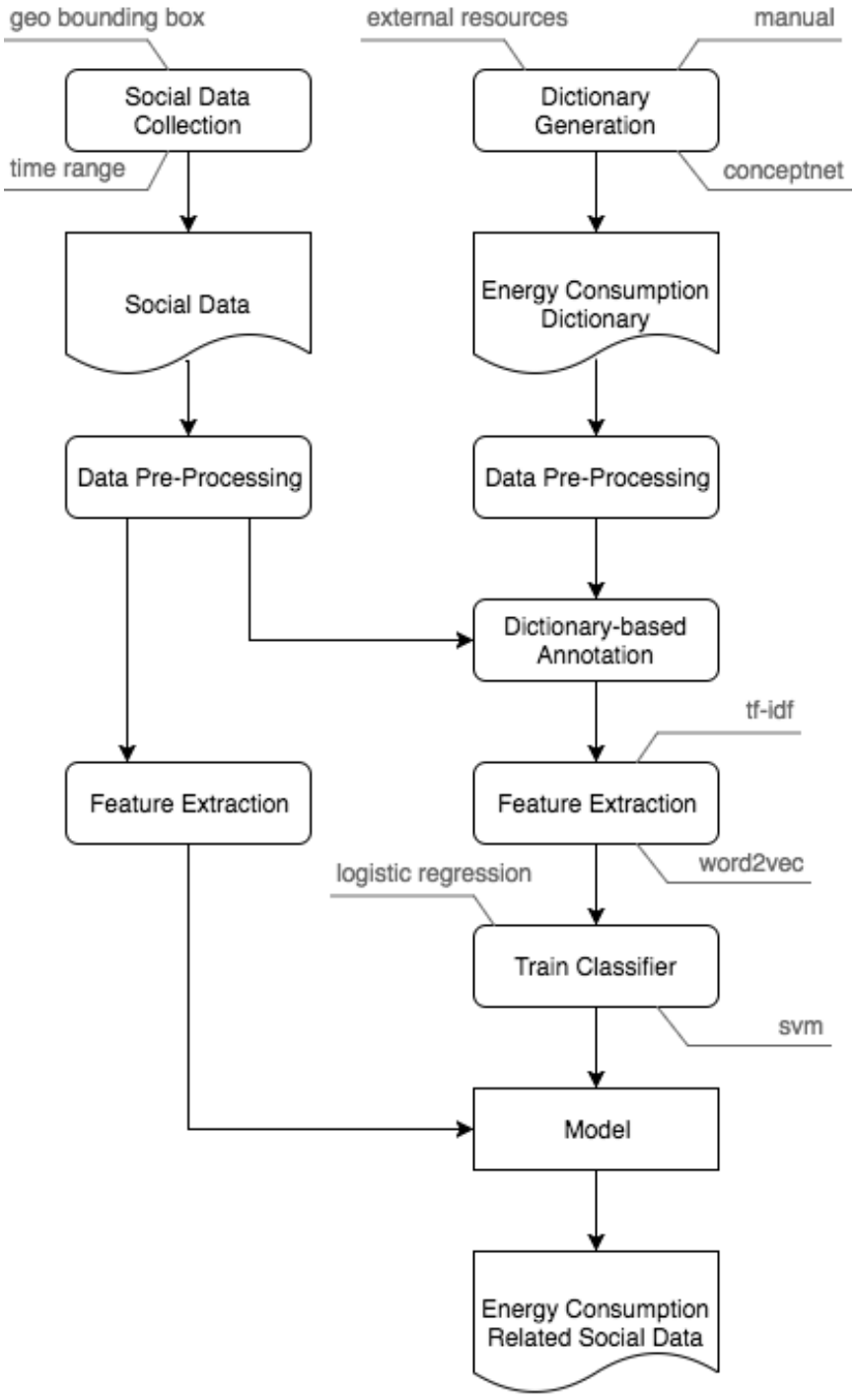


FIGURE 1.2: Methods

matching. Exact match is simply find an exact match between text's token and dictionary's term. On the other hand, partial match using Jaccard coefficient index to determine similarity between text's token and dictionary's term. The tweets containing any related terms of the dictionary (i.e. using the exact or partial match) are labeled as a energy consumption related tweet. In addition, we also conduct quality assessment for both dictionary and streams to ensure the reliability of the generated dataset. The assessment is conducted by observing the generated dataset, identify any irrelevant labeled tweets, and made further adjustment.

Using the labeled dataset, we train a binomial classifier to identify energy consumption related tweets. Based on this labeled tweets, we are able to generate both training and test dataset for the classifier. Because the dataset is noisy (or weakly labeled), due to heuristics nature of the specified rules, this approach is distant supervision machine learning. As additional features, we extract Term Frequency and Inverse Document Frequency (TF-IDF) for each term in the dataset. We explore different classifier algorithms and configurations (e.g. number of samples or features) to maximize the classifier performance. Finally, using this classifier, we are able to effectively identify energy consumption related tweets and provide an alternative dataset for energy consumption studies. As summary, methods being used in this thesis are visualized in Figure 1.2.

1.5 Contributions

The contributions of this thesis are listed as follow:

1. Literature survey on social media data analytics and energy consumption topic. We identify and synthesize several methodologies and approaches on studying human activity using social media in general, and specifically on energy consumption. We highlight the state-of-the-art advancement within the use of social media data to infer energy consumption. Literature survey is covered in Chapter 2.

2. Knowledge Discovery Framework that provide sets of procedures and steps required to generate an alternative dataset for studying energy consumption from social data. The framework is covered in Chapter 3.

3. A software library, which is a suite of implementation code of our framework that can be used for future studies on this topic. The software library is written in Python programming language and licensed as open source project, which enables further development, customization and extension by the research community. The software library is part of implementation in Chapter 4.

4. Energy consumption behavior patterns in two world cities. As part of the experiment, we implement the framework in Amsterdam (The Netherlands) and Jakarta (Indonesia) using social data gathered from Twitter. The experimentation is covered in Chapter 5.

1.6 Outline

In order to answer the research question and providing concrete contributions, this thesis is structured as follows:

- **Chapter 1**: Provides an introduction about research problem that we wanted to solve, followed by defining research questions and scope of the research.
- **Chapter 2**: Provides general overview of related works on extracting knowledge from social media, specifically on energy consumption.
- **Chapter 3**: Defines the framework with detail description for each underlying components, which consists of several approaches and methodologies to extract knowledge from social media.
- **Chapter 4**: Describes the implementation of the framework, with detail technical guidelines and the system architecture.
- **Chapter 5**: Evaluates the implementation of the framework through experiment in two different cities.
- **Chapter 6**: Discusses the result of the experiment and the framework in general, followed by overview conclusion of this thesis.

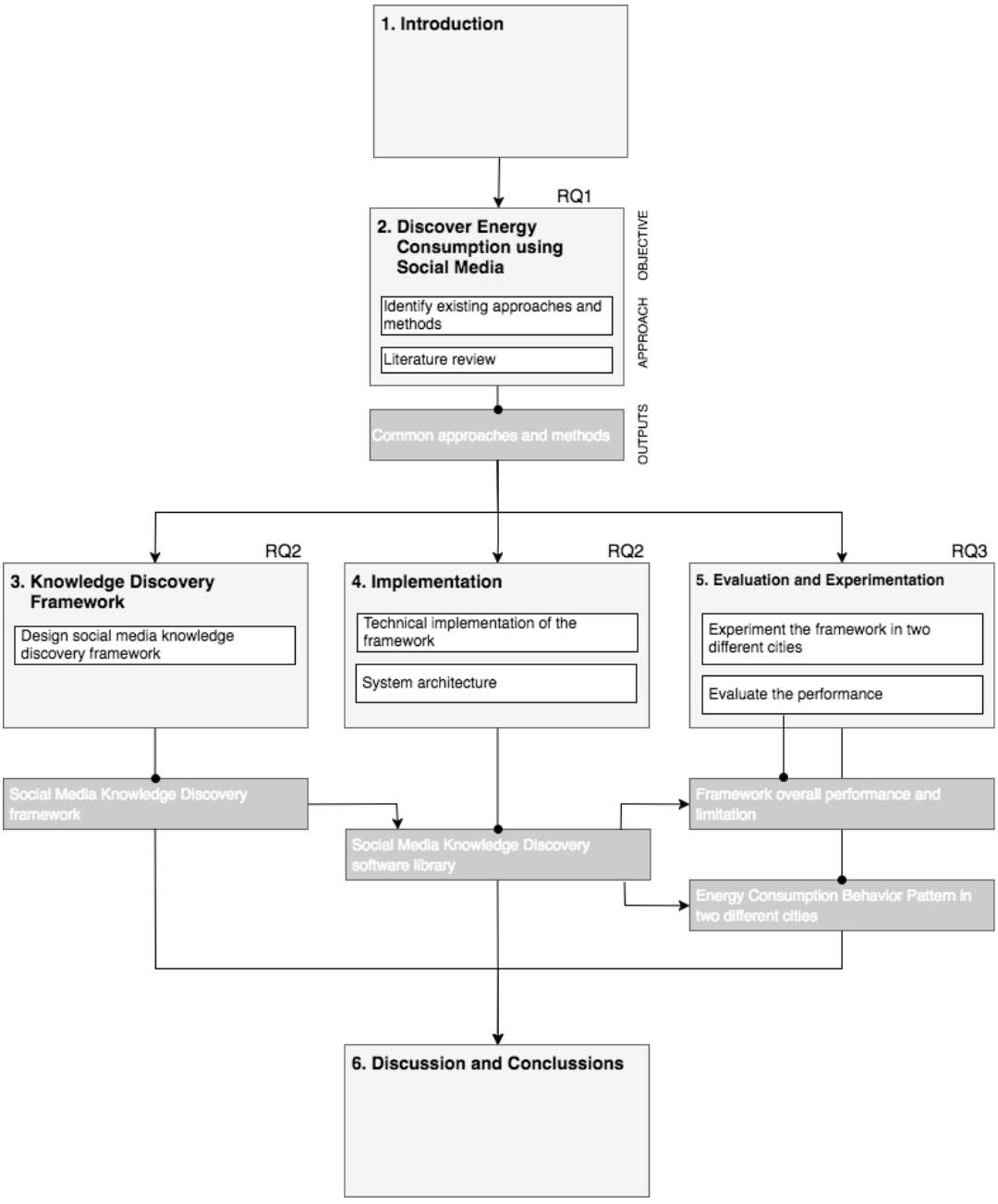


FIGURE 1.3: Research outline.

Chapter 2

Related Work

This chapter surveys previous work in discover energy consumption behaviour using social media data, and discuss in detail how our framework advances the state of the art of this topic. This chapter is part of our effort to answer `RQ1`.

- First, we review current conventional approaches and applications of energy consumption data.
- Second, we survey the advancement of social media data studies and the potential of inferring people's behaviour pattern in the urban environment.
- Finally, we discuss similar works in other domain specific problem which bear interesting relevancy with our objective, estimating energy consumption behaviour.

2.1 Measure Energy Consumption

Conventionally, measuring energy consumption is conducted by energy-related institutions (e.g. IEA[1], OECD[2], EEA[4]), governments (e.g. EIA[5], NBS[6], BPS[7]) or energy-related companies (e.g. Shell[8], BP[9]). These institutions basically measure energy consumption as part of their business needs. For example, an electric company should have transaction data of electricity usage within an area, or an oil and gas company who own gas station knows exactly, how many litter gasoline consumed based on their sales. However, most of these data are aggregated on city or country level; which unable to show energy consumption on more granular, such as household or individual.

Following the increasing need of more granular energy consumption data, some of these institutions, especially governments, conduct a comprehensive survey to gather energy consumption data. For instance, United States' Residential Energy Consumption Survey (RECS)[10] and China's Residential Energy Consumption Survey (CRECS)[11]. These surveys aim to identify energy consumption characteristics at the household level, which produce a high-spatial dataset for energy consumption. However, as a nature of the survey, this approach is obviously labor-extensive,

[1] https://www.iea.org/
[2] https://data.oecd.org/, World Bank[3]
[4] https://www.eea.europa.eu/
[5] https://www.eia.gov
[6] http://www.stats.gov.cn
[7] https://www.bps.go.id
[8] http://www.shell.com
[9] http://www.bp.com
[10] https://www.eia.gov/consumption/residential/about.php
[11] Zheng et al., 2014.

and require a lengthy process to be completed; US's RECS can only be conducted every four years periodically. As a result, there is a trade-off occurred, where survey provides fine spatial granularity, but low temporal granularity.

Although these conventional data sources have limitations, these data is still the most reliable data available for quantifying and estimating energy consumption. Hence, there are several notable works conducted using these data. Using historical data that span for decades, these data suitable for predictive analysis of future trends. Kraft and Kraft (1978) pioneered the causality study between the relationship of energy and Gross National Product (GNP), which later followed by several studies within the same economic sectors(e.g. Soytas and Sari, 2003; Shiu and Lam, 2004; Lee, 2005). Next, the usage of energy consumption data goes beyond economic sectors, Soytas, Sari, and Ewing (2007) investigate Granger causality relationship between energy with national income and carbon emission; York (2007) investigate the relationship between energy consumption with population demographic trends. These studies show that energy consumption data has the potential to predict trends in different fields of study.

2.2 Behavior Pattern in Social Data

As an alternative source, social media data provide rich human interactions data could be used to represent energy consumption behaviour pattern. Kwak et al. (2010) unearth the potential of Twitter[12] as a great wealth of information that publicly available. Moreover, they also demonstrate the ability of twitter to detect emerging trends within our social space, and provide comprehensive spati-temporal analysis. Social media data also has rich information on human activity that we can use to infer specific human activity. Pan, Ochi, and Matsuo (2013) successfully infer daily life behaviour pattern of a person (i.e. sleep time or work pressure/stress) using their communication activity data from Gmail[13], Facebook[14], and Twitter. By using collective data in a one-year period, this study has successfully predicted individual's daily performance based on the defined indicators. However, this study only conducted on a relatively small sample of a population, which only consists of 50 people. Therefore, a much larger individual behaviour pattern is required to have a better understanding of general human activity behaviour within an urban area.

Next, Zhu et al. (2013) pioneered human activity classifier using social media data. They categorize human activity into 10 categories, which are *Socializing, Relaxing, & Leisure; Eating & Drinking; Sports, Exercise, & Recreation; Consumer Purchases; Work-Related; Education; Traveling; Professional Services; Household Activities; Personal Care*. Using a machine learning technique, they train human activity recognition classifier using labelled dataset, which effectively infers human daily activities with an overall accuracy of 83.9%. However, it is important to note that the dataset is generated or labelled through crowd-source task on CrowdFlower[15], which requires taking into account financial constraint to determine the size of the dataset.

Using the high spatiotemporal dataset provided by social media, we can effectively identify collective activity pattern within the urban environment. Cranshaw et al., 2012 develop a model to extract distinct regions of the city that reflect current collective activity patterns, named as 'livelihoods'. This study shows the advantage

[12]https://twitter.com
[13]http://gmail.com
[14]http://facebook.com
[15]http://wwww.crowdflower.com

	Behavior Pattern in Social Data			
Reference	**Objective**	**Challenges**	**Methods**	**Key Findings**
Kwak et al., 2010	Study the topological characteristics of Twitter and its power as a new medium of information sharing	Extract information from text using machine learning approach	Using classifier that combines text analysis, anomaly detection, and social network analysis	Pioneered quantitative study of twitter and unearth the potential use of social data
Pan, Ochi, and Matsuo, 2013	Extract personal behavior patterns from social data	Identify people's behavior based on their social data activity	Infer daily life behavior pattern of a person based on their communication activity in Gmail, Facebook and Twitter.	Showed the possibility of discovering behavior patterns from social data
Zhu et al., 2013	Develop a human activity recognition using crowd-generated self-report data	Lack of conventional sensor data (as ground-truth) to verify findings inferred by social data	Using machine learning approach to infers human daily activity using Twitter data	Pioneered human activity classifier using social media data
Cranshaw et al., 2012	Extract distinct regions of the city that reflect current collective activity pattern	Present spatial clustering that reflect current social dynamics of a city	Using spectral clustering of geospatial check-in data from Facebook and Twitter	Clustering model for mapping a city based on the collective behaviors of its residents

TABLE 2.1: Related Work on Behavior Pattern in Social Data

of social media data compared to conventional data sources, where we are able to cluster or aggregate collective activity pattern beyond conventional boundaries of municipal units or neighborhoods.

However, with all the advantageous of social media can offer, there are also some deficiencies that need to be considered. As a caveat, we need to understand that although the amount of social media users is considerable, it doesn't mean any findings can be generalized to the entire population; because there are also people who not actively present in social media, may result in underrepresentation (observation bias). Furthermore, as the unstructured nature of social media data, there are also possibility of contextual bias on inferring the information that need to be considered. In order to overcome this limitations, we should positioned social media data as an emerging alternative data sources that need to be embraced as a complementary source of information for conventional data sources. Key highlights of related work on behavior pattern in social data is summarized in Table 2.1.

2.3 Discover Energy Consumption using Social Media

Bodnar et al. (2014) shows that social media users could be used as sensors of their activity pattern because their micro posts contain rich information about the human activity. Therefore, using this sensor, we are able to infer human activity in general, and specifically, energy consumption. Moreover, Cheng et al., 2011 describe detailed quantitative analysis and modeling of location sharing service users. The study provided several methods to process embedded geographic location from social media data and extract several spatial features, such as user's home location, user displacement, the radius of gyration, returning the probability of users for each location of interest and also identify factors that influence user's mobility pattern. Using such mobility inference, we could also quantify amount of energy consumed for that particular user.

As a subset of human activity, we can use social media data to study energy consumption. Abbar, Mejova, and Weber, 2015 identify food consumption activity from tweets. On this study, they define food-related keywords as a dictionary and apply these to find exact matches on the given tweets. Therefore, this study did not consider partial matches that effective to identify slang words or typos that commonly found on tweets or social media text in general. Bodnar et al., 2017 propose a system to identify the relationship between topics in social media data and electricity usage pattern. This study able to approximate electricity utilization patterns by conducting Granger causal relationship to find a relationship between discovered events and energy consumption through statistical analysis. This study provides huge methodological contribution on our framework to approximate quantify energy consumption from social media data.

One rationale for our approach in this research come from Marchetti-Bowick and Chambers (2012), where they using distant supervision learning to assist topic detection in tweets. The methodology starts by defining a list of keywords for each particular topic of interest. By using these keywords, they search any relevant tweets simply from Twitter's search and generate a topic-labeled dataset from this tweets. Next, they use this dataset to train a binomial classifiers for each topic, based on stream that consists related keywords (positive) or not relevant (negative). Surprisingly, this approach successfully outperforms lexical-driven keyword approach that commonly used for topic identification. However, it is important to mention that the scope of their study is topic identification for political forecasting, which only uses a relatively small set of dictionary keywords for each topic. On the contrary, topic identification for energy consumption should require a more extensive set of keywords to effectively identify a different kind of energy consumption related activities. Mesbah et al. (2017) provide another example use of distant supervision learning to identify underlying topic or class within a text. This study also uses a binomial classifier to identify specific topics, which is Logistic regression classifier trained from extracted TF-IDF features. Using this approach, this study has successfully extract semantic representation metadata from scientific publications with high precision and recall for all classes.

In conclusion, these related works show that extracting knowledge about energy consumption behavior patterns from social media is possible. However, due to unstructured nature of social media data, extensive pre-processing is required. Although social media data generate a noisy dataset, there are already several successful attempts on using distance supervision machine learning approach with good performance. Key highlights of related work on discover energy cOnsumption using social data is summarized in Table 2.2.

	Discover Energy Consumption using Social Data			
Reference	**Objective**	**Challenges**	**Methods**	**Key Findings**
Bodnar et al., 2014	Develop a novel system for social-media based disease detection at the individual level	Extract information from text using machine learning approach	Using classifier that combines text analysis, anomaly detection, and social network analysis	Social media users could be used as sensors for real-world event or phenomenon
Cheng et al., 2011	Quantitative assessment of human mobility patterns using social media user's location footprints	Identify temporal characteristics of how people use location sharing services and to model patterns of human mobility	Analyze social media user's location footprints using spatial, temporal, social and textual aspects	Extensive quantitative analysis and modeling to study human mobility patterns using social media
Abbar, Mejova, and Weber, 2015	Examine the potential of Twitter to provide insight about food consumption	Provide detailed and accurate data on the cultural and individual behaviors that lead to unhealthy dietary habits	Using dictionary-based annotation with food-related keywords	Foods mentioned in the daily tweets of users are predictive of the national obesity and diabetes statistics
Bodnar et al., 2017	Modeling real-time energy utilization patterns using social media data	Ubiquitous sensing systems that provide comparable level of information and knowledge with conventional sensor systems	Social media network-driven model that utilizes large-scale textual and geospatial data	Semi-supervised knowledge discovery that infers events from topics generated from social media network data
Marchetti-Bowick and Chambers, 2012	Political forecasting using Twitter data	Identify political alignment based on social media data	Using dictionary-based annotation and distant supervision learning to assist topic detection in tweets	Distant supervision for topic identification outperform sentiment lexicon
Mesbah et al., 2017	Semantic annotation in scientific publications	Classify sentences according to the nature of the contained information	Using dictionary-based annotation and distant supervision learning to identify underlying topic within a text in scientific publications	Distant supervision for topic identification achieve acceptable performance

TABLE 2.2: Related Work on Discover Energy Consumption using Social Data

Chapter 3

Knowledge Discovery Framework

In this chapter, we introduce a general knowledge discovery framework to extract knowledge about energy consumption behaviour from social media; as part of our answer on the second research question RQ2. The main challenge on extracting such knowledge from social media compared with conventional data sources lies in the unstructured nature of the considering data. The framework implements a data pipeline to effectively extract such knowledge. In brief, the framework consists of three main stages, which are Data Collection, Data Processing & Enrichment and Data Analysis, as depicted in Figure 3.1.

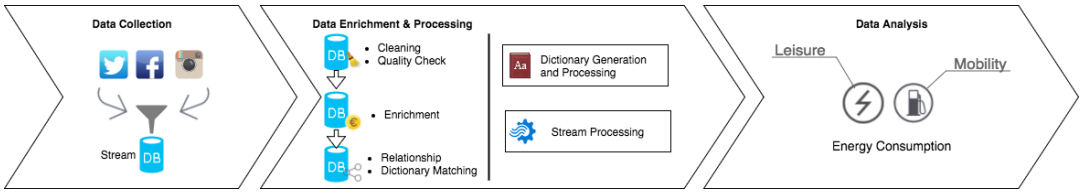


FIGURE 3.1: The Knowledge Discovery Framework

3.1 Data Collection

The initial and common stage on studying social media data is data collection. Data collection should ensure that the collected data is aligned with the objective and scope of the study, in terms of temporal periods, geographical boundaries and also relevant social media platforms. Temporal periods and geographical boundaries heavily depend on the scope of the experiment that we choose to perform. Following many social media platforms that are available today, there are some criteria that need to be considered, which are:

1. The platform is publicly publishing its data. A platform that publicly publishes its data using accessible methods, such as API, significantly eases the whole data collection process compared than crawling the web pages.

2. The platform provides necessary information that are relevant to the objective of the study. There are much social platform available today with their own unique use case or purpose. For example, micro-blogging (i.e. Twitter, Tumblr[1]), social network (i.e. Facebook, Google Plus[2]) or media sharing (i.e.

[1]https://www.tumblr.com
[2]https://plus.google.com

YouTube[3], Instagram). Therefore, it is recommended to choose social media platform that aligns with the study objective.

3. The platform is popular within the geographical area or targeted community of the study. There are several social media platforms that provide similar features to users. In some cases, the popularity of such social media platform is different for each region or country. For instance, Twitter is the leading microblogging social media platform globally, but Weibo [4] is the most popular platform in China.

4. The platform is following the asymmetric network model. Asymmetric network model enables us to collect social media data from any user without requiring mutual friendship with that specific user. For instance, on Twitter, we can follow or read tweets from any celebrities, public figures or users which not even in our actual social circle. As a result, we are able to collect more data from social media platform with asymmetric network model.

By following these criteria, one can improve the relevance of the study objective and the collected dataset. Moreover, these criteria also maximize the number of collected social data.

3.2 Data Enrichment and Processing

Our objective in this stage is to pre-process the collected social media data and enrich them in such degree that suitable to extract knowledge about energy consumption behaviour. Our approach is inspired by Marchetti-Bowick and Chambers (2012) study, where they using distance supervision to extract specific knowledge, political alignment from social media data. We can use the same approach to extract energy consumption behavior. However, studying energy consumption require more extensive dictionary compared to political alignment (see chapter 2). Therefore, as part of this stage, we explore several dictionary generation strategies to generate energy consumption dictionaries.

We divide this stage into three main sub-stages, which are enrichment and processing for energy consumption dictionary, social media data stream and dictionary-based stream annotation.

3.2.1 Dictionary Generation and Processing

The initial step to generate energy consumption dictionary (`Dictionary`) begins by creating a seed dictionary (see Figure 3.2). Seed dictionary can be compiled by manually identifying keywords or terms related to energy consumption behaviors. In addition, it is also possible to use any existing relevant dictionary available (i.e. Oxford References[5]) or extract terms directly from the Internet (i.e. culinary website[6], e-commerce website [7]). Next, terms should be categorized into different energy consumption behavior sectors. Based on Revilla (2016) study, we classify terms into four different classes, which are dwelling, mobility, food consumption and leisure. The

[3]https://www.youtube.com
[4]http://www.weibo.com
[5]http://www.oxfordreference.com
[6]http://allrecipes.com/
[7]https://www.mediamarkt.nl

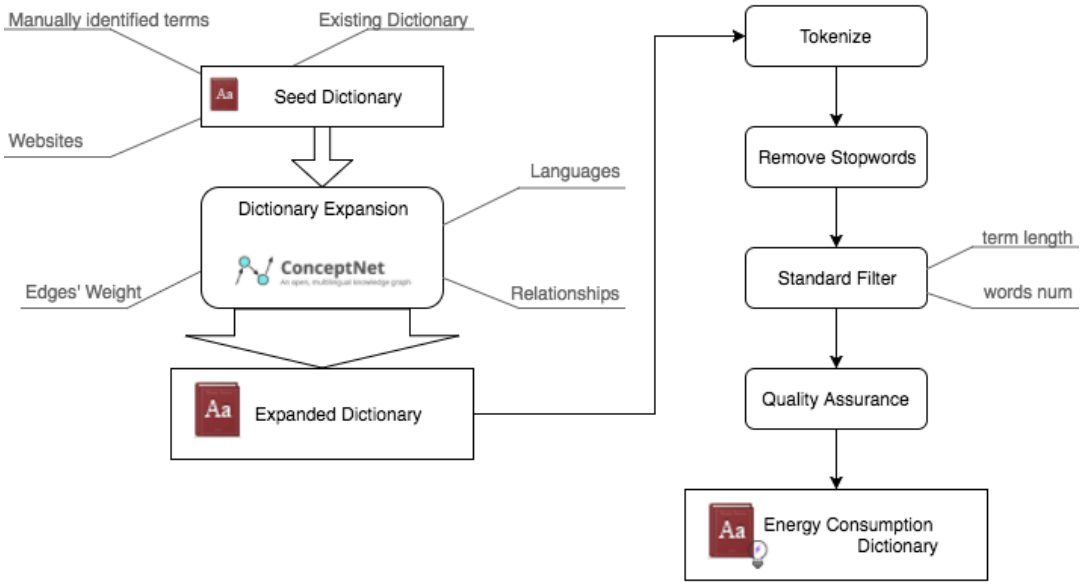


FIGURE 3.2: Dictionary Generation and Processing

classification is multi-label, where multiple energy consumption behavior labels can be assigned to each term.

Although by following the above approach we already yield related terms for energy consumption, it is possible to expand the dictionary further. We can apply several methods to discover related terms from the existing terms in the dictionary. The related terms could be discovered by simply looking into their translation in other languages, finding the synonym of the term or any other semantic relationships, such as WordNet[8] or ConceptNet[9]. This step ensures the generated dictionary consists of extensive terms that effectively and accurately annotate the collected social media data.

Next, the expanded dictionary should be processed further to improve the overall quality of the dictionary. The sequence processing for this purpose is conducted as follows:

1. Tokenize the dictionary terms. By simply chopping the terms into pieces based on the whitespace into tokens. There are also several tokenization strategies other than using whitespace (e.g. split the terms based on any non-alphanumeric characters). The selection of this strategy is language-specific, that each language has distinctive signature patterns.

2. Remove stop words from the dictionary terms. Stop words are set of commonly used words in any language. By removing these very commonly used words, we can focus on the important words instead. For instance, in the English language, we can find words like "the", "a" or "an" within the corpus. As a result, we are able to generate more representative tokens.

3. Words stemming. For the grammatical reason, terms use different forms of words, such as eat, eats and eating. Stemming aims to reduce inflectional or

[8]https://wordnet.princeton.edu
[9]http://conceptnet.io

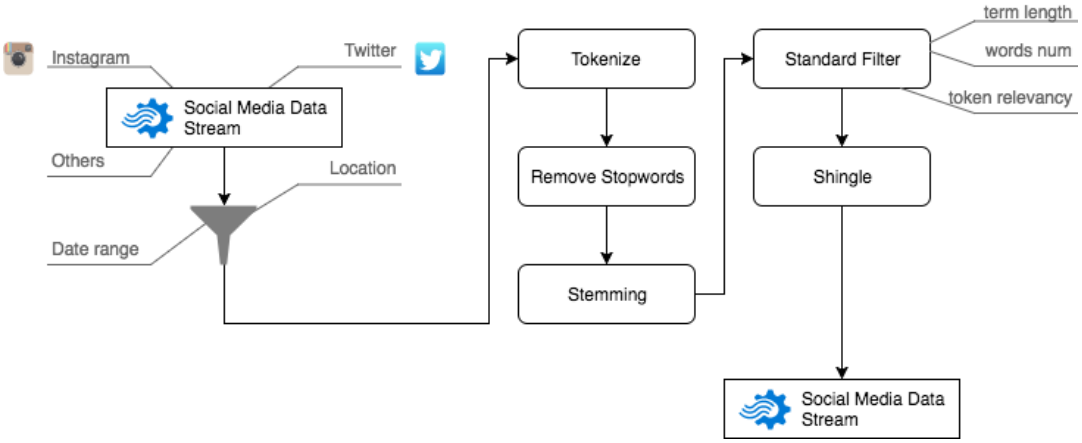


FIGURE 3.3: Stream Processing

derivational forms of a word to their base or root form. For instance, in the English language, we can stem words like "drives" or "drive" into "drive".

4. Standard filter of valid terms. Removing undesirable terms from the dictionary, such as minimum length of the terms, the maximum number of words and also desired selected languages of the terms. This filter effectively affects a number of terms within the dictionary and configured based on the scope of the study.

5. As quality assurance, we suggest also to manually check the generated dictionary to ensure the quality and conduct necessary configuration fine-tuning for better results.

3.2.2 Social Media Stream Processing

Stream processing begins by acquiring social media data from different platforms. Next, it is also necessary to filter social media data stream (`Stream`) based on the selected date range and location of the experiment, because specific time period or location may have their own energy consumption behavior patterns. Similar with `Dictionary` processing, we also require to process `Stream` further to improve the overall quality and increase the dictionary-based annotation performance on the next steps.

In addition, we implement w-shingling processing on the content text. This method generates a set of unique shingles from each adjacent tokens, as an addition to existing tokens to improve dictionary matching performance. For instance, if we have a text "Cooking classic spaghetti bolognese" and we apply 2-shingling the output would be "Cooking classic", "classic spaghetti" and "spaghetti bolognese". However, it is also important to note that higher number of shingling also increase the complexity of matching performance and directly impact on the processing time.

3.2.3 Dictionary-based Stream Annotation

Following two prior processing for both dictionary and stream, we are able to conduct dictionary-based annotation on the stream to produce a labelled dataset for our distance supervised learning. The matching algorithm on this stage can be divided

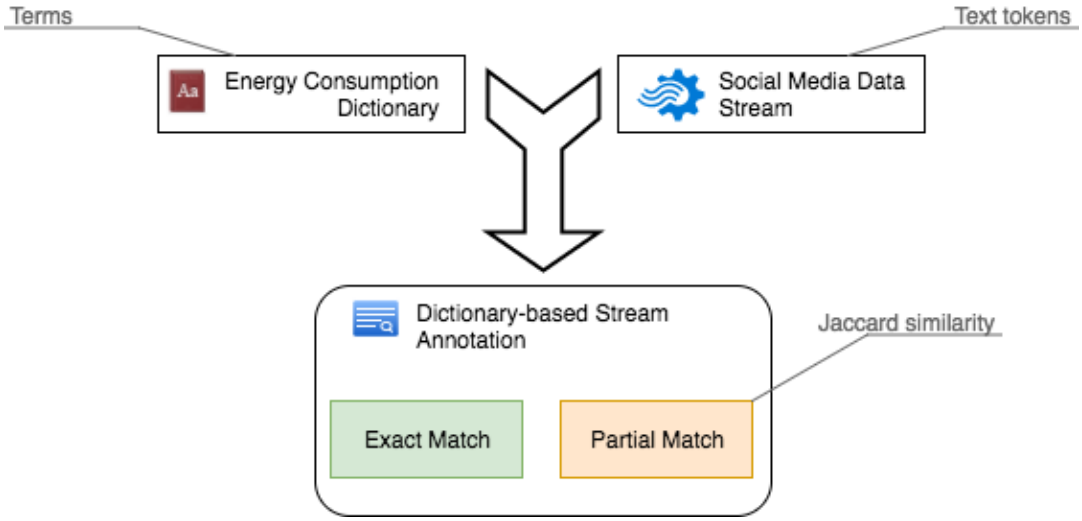


FIGURE 3.4: Dictionary-based Matching Workflow

into two types, which are exact matching and partial matching as depicted in Figure 3.4.

Exact matching is conducted by simply matching dictionary terms with the stream's tokens. On the other hand, partial matching requires implementing similarity algorithm to determine whether the given dictionary terms match the stream's tokens. For this purpose, we can use Jaccard Index, which is a similarity coefficient that is defined as the magnitude of the intersection of the two terms divided by the magnitude of the union of them both (see Equation 3.1).

$$jaccard(A, B) = |A \cap B| \cdot |A \cup B| \tag{3.1}$$

Partial matching is effective to match a text that consists of typos, slang words or informal words that very common occurred on social media data. By combining both matching approaches, we are able to determine dictionary class for each of text content and generate a labelled dataset for our study.

3.2.4 Feature Extraction

Using the labelled dataset generated from previous steps, we are able to extract features or derive values from the given text content. Features extraction is of paramount importance in any machine learning task and determines the overall performance of the classifier. Currently, there is several word representation feature extraction method that we can use, such as TF-IDF, Word2Vec or any other sophisticated features.

Term frequency-inverse document frequency (TF-IDF) is a feature vectorization of a document that provides the importance of each word to a document in the corpus. The method basically consists of two parts. Term frequency $TF(t, d)$, denotes the count of occurrences of a given term t in the given document d. Document Frequency $DF(t, D)$ is the number of documents that contains term t. DF is important to refrain over-emphasize terms that appear repeatedly across documents but

carry insignificant information about the document. Hence, Inverse document frequency is computed as a numerical measure to determine the importance of a term (see 3.2). Finally, TF-IDF is simply the product of TF and IDF.

$$IDF(t, D) = log\frac{|D| + 1}{DF(t, D) + 1'} \tag{3.2}$$

$$TFIDF = (t, d, D) = TF(t, d) \cdot IDF(t, D) \tag{3.3}$$

Word2Vec is a feature vectorization of a text that provides distributed vector representations of words. Using this method, words that share common contexts in the corpus are located close in the vector space, which makes the generalization to novel patterns easier and model estimation more robust. Word2Vec construct a vocabulary from the training text data and then learns vector representation of the words.

3.2.5 Machine-Learning-based Topic Detection

The main objective of this step is to detect topic within social media data. Based on the previous dataset, we have generated a labelled dataset with their respected features. Based on Marchetti-Bowick and Chambers (2012) study, distant supervision approach is more effective to infer topic from social media data compared to the pure keyword search. Distant supervision uses a weakly labeled training set or noisy signals to identify positive examples of a topic within test dataset.

Next, we are able to train a simple binary Logistic regression classifier as a proof of concept of our approach. Subsequently, we can also train more sophisticated classification algorithms, such as SVM; and conduct performance comparison analysis between these algorithms. As a result, we are able to determine which classification algorithm that works best for the study, because every machine learning algorithm works best under a given set of conditions.

Logistic Regression is a simple classification algorithm for predictive analysis in the binary class problem, such as a binary outcome between `0` and `1`, or `True` and `False` given a set of independent variables. The algorithm uses the logistic function to find a (`logit`) model that fits with the data points.

Support Vector Machines (SVMs) is discriminative classification algorithm that constructs optimal hyperplanes in high-dimensional space that aims largest distance (maximum margin) between the nearest training data point of any class and form a good separation between them.

3.3 Data Analysis and Visualization

The main objective of this entire framework is to discover knowledge about energy consumption behavior patterns from social media. The first two stages prepare social media data into a state where further analysis can be conducted. In this section, we identify specific analysis methods that are suitable to infer each energy consumption behaviors as shown on Table 3.1.

3.3.1 Text Analysis

Text analysis is the most common analysis approach to discover insight from social media data due to their rich text information. Using content's text from social

	Dwelling	Mobility	Food	Leisure
Text	text analysis, temporal analysis	text analysis, temporal analysis	text analysis, temporal analysis	text analysis, temporal analysis
PoI		heatmap, displacement distribution		heatmap
Trajectory		radius of gyration, trajectory segmentation		radius of gyration, trajectory segmentation

TABLE 3.1: Different Methods Discover Energy Consumption Behaviors using Socia Data

data, we are able to implement several methods to extract features that can be used to identify energy consumption. These are several text analysis methods that are commonly used on social media data:

- **Word frequency**. Lists of words and their frequencies.
- **Collocation**. Identify words that commonly appearing near each other.
- **Concordance**. Discover the contexts of a given word or sets of words (sentences).
- **N-grams**. Generate pair of word phrases.
- **Entity recognition**. Identify entity within text. (i.e. names, places, time periods, etc.)
- **Dictionary-based annotation**. Find a specific set of terms in the texts.

Text analysis approaches are suitable for every single energy consumption behaviors, because most of social media data provide text-based content.

3.3.2 Heatmaps Visualization

Heatmaps is a two-dimensional graphical representation of data where the individual values are represented by colors. This visualization is the one of the simplest approach for spatial analysis. This visualization provides an immediate visual summary of information, by using different intensified colors to present the density on the particular location. In order to generate such data, a clustering algorithm is required to effectively cluster the data based on their location. For instance, we can use (Ester et al., 1996) density–based clustering, Density-based Spatial Clustering of Applications with Noise (DBSCAN) algorithm, which is today's de facto standard algorithm for density-based clustering. DBSCAN group together spatial points that are closely packed together and mark the outliers points as low-density regions.

3.3.3 Displacement distribution

Using the embedded geographic location attribute from social media data, we characterise the mobility patterns of individuals based on their posts sequences. Based on these sequences, we can determine whether individual user is moving, and the distance between their two consecutive reported locations.

3.3.4 Radius of gyration

Another use of embedded geographic location attribute is to compute radius of gyration. This metric measures the standard deviation of distances between the user's reported location and the user's center of mass. As a result, we are able to identify both how frequently and how far a user moves. For instance, we can use average radius of gyration of users to visualize a heatmap that shows the dynamics of cities. The radius of gyration formalized on equation 3.4.

$$r_q = \sqrt{\frac{1}{n}\sum_{n}^{i=1}(r_i - r_{cm})^2} \tag{3.4}$$

3.3.5 Temporal Analysis

As a general analysis approach for all energy consumption behaviors, we can use temporal analysis to discover insight from the dataset. Social media data provide high-temporal dataset that potential for temporal analysis. Moreover, we are able to combine social media dataset with existing dataset from conventional data sources. However, there are some key characteristics that need to be considered when using social media data with those conventional dataset. First, social media tends to provide high-temporal dataset with time granularity in seconds. However, it is rare for real-world events to be reported at such temporal resolution. Second, it is obvious that a single social media data might be not significant to represent an actual real-world event.

Bodnar et al. (2017) suggest to overcome this discrepancies by normalizing social media dataset into conventional dataset's time scale. They pair social media data with real-world event, by defining a document d_j as the aggregation of social media data v_j, which is derived from several micro-posts m_j that occur between real-world event's time scale, x_q and x_{q+1} respectively (see equation 3.5). As a result, by combining both of social media dataset and conventional dataset, we are able to conduct both causation or correlation analysis.

$$d_j = \{v_j | time(x_q \leq time(m_j) < time(x_{q+1})\} \tag{3.5}$$

Granger causality relationship determine whether two different time series data are related to each other. The general idea of this analysis is based on the fact that time does not run backward, where if event `X` happens before event `Y`, then we can consider that event `X` is might causing event `Y` to happen. However, it is not possible for the opposite. Hypothesis for causal effect of X on Y shown on equation 3.6, where $\mathbb{P}$ is the probability, A is an arbitrary non-empty dataset, and both $\mathcal{I}(t)$ and $\mathcal{I}_{-X}(t)$ denote the information within time t in the entire universe, and in the modified universe where X is excluded. We can use this analysis to identify whether social media's event can be used to predict a particular upcoming event in a real world.

$$\mathbb{P}[Y(t+1) \in A \mid \mathcal{I}(t)] \neq \mathbb{P}[Y(t+1) \in A \mid \mathcal{I}_{-X}(t)] \tag{3.6}$$

Pearson correlation coefficient measure the correlation strength between two different variables. The coefficient has a value between $+1$ and -1, where $+1$ is total positive linear correlation, 0 is no linear correlation, and 1 is total negative correlation. The formula is simply a covariance cov of the two variables divided by the product of their standard deviations σ, as shown on equation 3.7. This correlation

analysis is useful to determine whether social data's trend can be used to approximate actual real world's trend.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{3.7}$$

Chapter 4

Implementation

In this chapter, we describe the implementation of the proposed framework from the previous chapter as a software library. In general, we envision a framework that implements a data pipeline that provide generic infrastructure for different kind of study of knowledge discovery using social media. For the implementation, we have identified several points to be considered, as follows:

- First, the implementation should be generic and flexible. The proposed framework acts as main guidelines to discover knowledge using social media. However, on every stages, it is also possible to rearrange or reconfigure based on the study requirement. Therefore, the implementation should provide data pipeline architecture that help organize different kind of study objective, with variety of extensible modules.

- Second, the implementation should be scalable, which enable to process large amount of social media data with an immense real-time streams. Hence, the implementation leverages cluster computing system architecture that provide extensive computing power necessary.

- Third, the implementation should provide general ETL (Extract, Transform and Load) process, which support data extraction from homogeneous or heterogeneous data sources (i.e. database, documents, CSVs or media); data transformation with high-level operators; and finally, data loading to different kind of data format.

A common design pattern that is being used for studying social media heavily depends on file dumps for each step of the data pipelines, typically in a format like CSV or compressed CSV. For each step, there are dedicated or parametrized scripts to process the data as required, such as unpacking the files, transforming into optimal query format and computing necessary processing tasks. However, these approach is not scalable when file dumps are large, where some simple steps can significantly slow down the entire data pipeline. Moreover, this approach is basically not a data pipeline, because there is no dependency resolution or work-flow management exists to ensure the reliability of the data pipeline. As a result, this approach is not suitable for building complex pipelines of batch jobs and also difficult to maintain.

Based on these considerations, we suggest to use exiting big data processing framework that are currently available, such as Apache Spark[1] as the data pipeline infrastructure for the framework.

[1] https://spark.apache.org

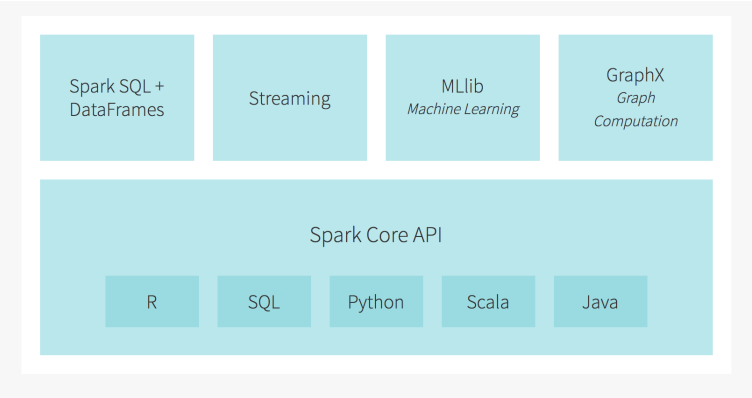


FIGURE 4.1: Apache Spark Architecture

4.1 Cluster-computing Platform: Apache Spark

Apache Spark is a cluster computing platform that provides a faster and more general data processing platform. Spark claim that is able to run programs up to 100x faster in memory, or 10x faster on disk, than Hadoop. The main advantages of using Spark is their rich APIs for operating on large datasets, which natively build for distributed architecture. The APIs consists of a collection of over 100 operators for transforming data and familiar data frame APIs for manipulating semi-structured data. Spark efficiently use fast in memory computing approach, which keeps track of the data that each task produces and enable applications to reliably store this data in memory. Thus, significantly improve Spark's performance by avoiding applications on costly disk accesses.

Apache Spark is packaged with higher-level libraries that support SQL queries, streaming data, graph processing, and the most important library for our framework, machine learning (see Figure 4.1. As a result, this swiss-army tools for big data processing should significantly help to develop complex work-flows required by the framework. Additional key features that Spark has to offers, such as:

- The APIs are available in different programming languages, which are Scala, Java, Python and R. These capability is significantly eases the learning curve for the new adopters.
- Integrates very well with the Hadoop ecosystem and different data sources (HDFS, Amazon S3, Hive, HBase, Cassandra)
- Able to run program on clusters managed by Hadoop YARN or Apache Mesos, and also standalone.

Using Apache Spark as the data pipeline infrastructure for our framework, delivers a truly unified approach to social media data analytics at scale. This approach, enable us to process a simple or computation-extensive task on either standalone or clusters of thousands of nodes with the same code. As a result, the implementation code of the framework is more maintainable, reusable and easily extensible by the community.

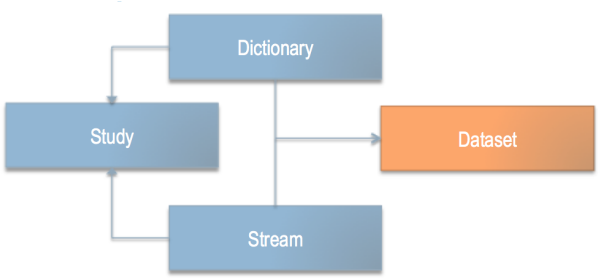


FIGURE 4.2: SocioKnowledge Architecture

4.2 System Architecture: SocioKnowledge

Apache Spark provides general programming model that eases developers to develop an application by composing arbitrary operators, such as mappers, reducers, joins, group bys and filters. This feature simplifies a wide array of computations, machine learning, streaming, complex query and batch, executed solely within Spark platform.

We proposed an application library, named SocioKnowledge as the implementation of our proposed framework for studying social media data. SocioKnowledge is a library that acts as an abstraction of Apache Spark platform which implements this framework approach. The library is written in Python programming language and designed as a package module, which is easier to distribute, implement and extended by the community. As depicted in Figure 4.2, SocioKnowledge mainly consists of four main modules, which are `Study`, `Dictionary`, `Stream` and `Dataset`.

4.2.1 Study Module

`Study` act as the main entry of the library. Using object of this class, we define an isolated environment for different study by initializing dependent components, such as Apache Spark Context, Data Storage, and Database connection. For instance, we can simply construct an object with study name "energy-consumption" using the given parameter as shown on listing 4.1.

```
1 study = Study('energy-consumption',bucket_url="data/")
```

LISTING 4.1: Study Module using local data storage

Apache Spark Context is the main entry point for Spark functionality. This context represents the connection to a Spark cluster, where we could easily configure whether we want to run Apache Spark as a standalone or as clusters with detail resource management configuration (define number of executors and their cpu & memory configuration). All of these setup preferences is configurable directly from system variables [2]. Next, data storage is a bucket to store documents or files required to be ingest as an input to our data pipeline, and also to store the result output. Our library has support two different bucket strategy, which are the ordinary local data storage that create separated folder for each study, and cloud storage service hosted in Amazon S3[3] by easily use the bucket URL as the parameter. Lastly, this module

[2]https://spark.apache.org/docs/latest/configuration.html
[3]https://aws.amazon.com/s3

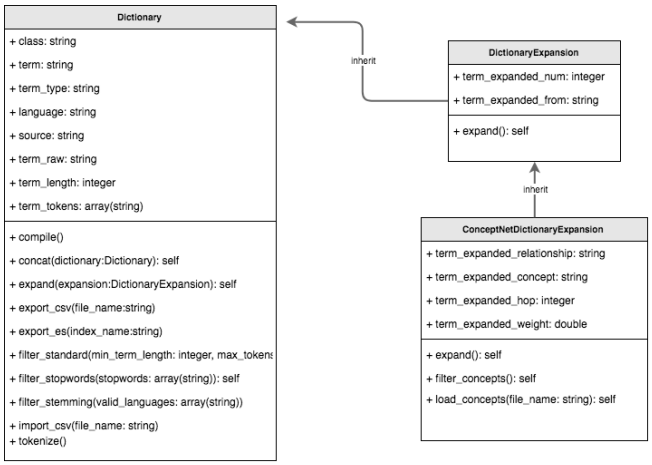


FIGURE 4.3: Dictionary Module Classes

also establish several database connections for the data pipeline, which currently support for MySQL, MongoDB and Elasticsearch. Similar with Spark configuration, connection configuration for these databases is configurable directly from system variables. The main reason that most of the component configurable directly using system variables is to enable containerization using Docker[4]; and deliver this whole study application as a lightweight, stand-alone and executable docker images.

4.2.2 Dictionary Module

`Dictionary` handles the dictionary generation pipeline, which includes providing a generic schema for the dictionary that streamlined different sources of the dictionary into a unified dictionary data model and several processing operations as depicted in Figure 4.3. This module acts as the higher-level abstraction of Spark's Data Frame and provide predefined methods based on the framework workflow.

```
1 expansion = ConceptNetDictionaryExpansion(\
2     dictionary=dictionary,\
3     conceptnet_file="conceptnet.csv", \
4     input_col="term",\
5     valid_min_weight=1.0,\
6     valid_languages=['en','nl','id','ms']
7 )
8 dictionary.expand(expansion)
```

LISTING 4.2: ConceptNetDictionaryExpansion Class

The main highlight on this module is on the `DictionaryExpansion` class, which responsible to further expand the `Dictionary`. Dictionary expansion is conducted by iterating through seed dictionary terms and discover related terms using particular approach. In this thesis, we explore the use of ConceptNet to expand our dictionary. We extend `DictionaryExpansion` class to develop and implement the expansion algorithm, as described the details on section 3.2.1. In addition to the expansion process, we also take a note for each expanded term, such as total number

[4]http://docker.com

FIGURE 4.4: Stream Module Classes

of expanded on particular term, ConceptNet's URI, total number of hop conducted, and edge's weight to reach this particular term. This expansion's meta data is really helpful for further analysis and improvements. Code snippet 4.2 shows how to expand existing dictionary object using ConceptNet with some given configurations; this approach can be replicated on any other dictionary expansion methods.

4.2.3 Stream Module

`Stream` handles social media data stream pipeline. Similarly to the Dictionary module, this module also offers a general unified data schema that enables us to store data stream from different social media platforms and conduct several processing operations, as depicted in Figure 4.4.

4.2.4 Dataset Module

`Dataset` is a simplified form of `Stream`, that serves as labeled dataset for machine-learning related tasks. This class provide several feature extraction methods, such as TF-IDF and Word2Vec, which already packaged as `mllib` natively on Apache Spark. For each of these feature extractions, we are able to configure different parameter configuration as part of our experiment. Moreover, this module also provide dimensionality reduction using PCA, transform multi-label dataset into binomial dataset, and several classifier algorithms.

Chapter 5

Experiments and Results: Energy Consumption Behavior

In this chapter, we conduct experiments that aim at showing how energy consumption behavior could be generated from social media data. As the objective of this experiment, we implement and evaluate the performance of our proposed framework to identify different energy consumption behaviour pattern. The study initially starts by generating energy consumption dictionary, and then followed up by identifying crawled data from Twitter in two different cities, which are Amsterdam (The Netherlands) and Jakarta (Indonesia) using the dictionary.

5.1 Dictionary: Energy Consumption

In the following, we describe the dictionary generation process for energy consumption topics in accordance with the four energy lifestyle sectors: dwelling, mobility, food consumption and leisure. First, we initially define keywords or terms that relevant for each sector. Based on the scope of the experiment, we identify three different languages of terms that need to be defined, which are English, Dutch, and Bahasa Indonesia.

A manually defined dictionary is obviously not extensive enough to identify broad topic of energy consumption. Therefore, we also extract dictionary terms from several different resources, which are:

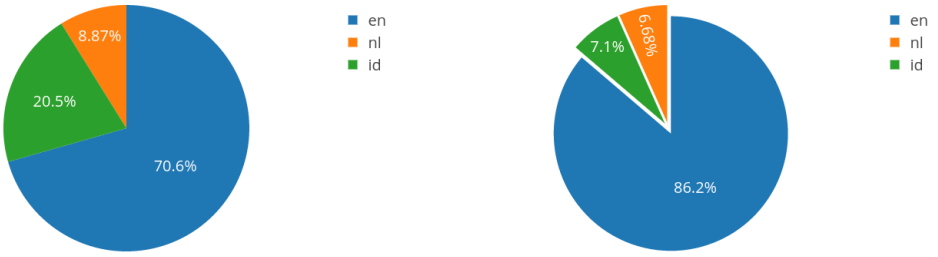
- **General Dictionary Reference** for food related terms from Oxford Food Reference [1].
- **Wikipedia** for general dictionary terms, especially for mobility and leisure related terms.
- **E-Commerce Websites** for dictionary related to dwelling. List of home appliances product category collected from Amazon [2] (English), MediaMarkt (Dutch) [3] and Lazada [4] (Indonesian).

The result as shown on Table 5.1, includes 1,262 terms as our energy consumption seed dictionary. Related terms for food consumption is dominating with nearly half of the entire dictionary, 48.7%, which most of them are contributed from Oxford Food References.

[1] http://www.oxfordreference.com/
[2] http://www.amazon.com
[3] http://www.mediamarkt.nl/
[4] http://www.lazada.co.id/

Class	Seeds Dictionary				
	Manual	*E-Commerce Websites*	*Oxford References*	*Wikipedia*	TOTAL
Dwelling	122	146			268
Food	171		444		615
Leisure	93			82	175
Mobility	87			117	204
TOTAL	**473**	**146**	**444**	**199**	**1,262**

TABLE 5.1: Energy Consumption Seed Dictionary by Source



(A) Seed Dictionary (B) Expanded Dictionary

FIGURE 5.1: Dictionary by languages: English (en), Dutch (nl) and Bahasa Indonesia (id)

On another perspective, we compare term's language distribution for both the seed and the expanded dictionary in Figure 5.1a. In both dictionaries, the English language is consistently dominant over other languages, 70.6& and 86.2% respectively. This condition is expected because we label common terms as the English language. For example, terms like "burger", "yogurt" or "pub" are registered as English terms despite those terms might be used in other languages.

In order to discover more terms for each class, it is required to expand the dictionary further. Using the initial dictionary as the seed dictionary, we leverage ConceptNet to expand our dictionary.

5.1.1 Dictionary Expansion using ConceptNet

ConceptNet is a freely-available semantic network, designed to help computers understand the meanings of words that people use. By using ConceptNet, we are able to discover semantically related terms from our existing dictionary. ConceptNet provides an API endpoint that freely accessible at http://api.conceptnet.io to query the semantic network. However, there is rate limit for a total number of request each hour that being permitted. Due to extensive-computation required for the expansion algorithm, as consequence, we lookup the related terms directly from ConceptNet's raw tab-separated file[5], which is a pre-built list of all the edges (assertions). The raw database of ConceptNet consists of five fields, which are:

- The URI of the whole edge

[5]https://github.com/commonsense/conceptnet5/wiki/Downloads

- The relation expressed by the edge
- The node at the start of the edge
- The node at the end of the edge
- A JSON structure of additional information about the edge, such as its weight

An edge, or assertion, is a unit of knowledge in ConceptNet which identify the relationship between natural-language terms, according to a particular source. ConceptNet provide several semantic relationships[6] that we can explore. Based on our iterative observation of the related terms, we identified 23 ConceptNet relationships that relevant within the context of the study (see Appendix A.1).

Not every edge provided by ConceptNet is required for this experiment. For instance, we only filter languages which are relevant to our experiment: English, Dutch, and Bahasa Indonesia. Currently, ConceptNet has total 304 languages available. Language for each concept is identifiable from the second part of the URI of the concept; for instance, term 'car' is registered with URL $/c/en/car$, where we can extract en and identify as English based on a two- or three-letter code that's been standardized by IANA. Following the scope of the experiment, we filter the ConceptNet dataset using three different languages, which are English (en), Dutch (nl), and Bahasa Indonesia/Malaysian (ms). Important to note that on ConceptNet, Bahasa Indonesia and Bahasa Malaysia both appear under the language code `ms`[7]. Moreover, it is also important to consider the edge's `weight` that can be extracted from the JSON meta data field. This attribute is provided by ConceptNet to defines the strength of the edge expresses the assertion. A typical edge's weight is one, but weight can be higher or lower. Hence, we only choose edges that have a weight greater or equal than one to ensure only the strength edges that need to be considered. The expansion approach is conducted by simply iterate through the dictionary terms and lookup any related terms based on the edges.

After dictionary expansion, we conduct several pre-processing routines. First, we apply simple whitespace tokenization for each term, because all of the three languages is basically using the Latin alphabet. Second, we remove any stop words found on the dictionary terms using the common stop words for that three languages[8]. As the final touch, we remove any terms which have more than two words and remove any duplication occurred in each class.

As a result shown on table 5.2, this method yields 7,317 new energy consumption related terms or expanded seven times from the initial seed dictionary. The expanded dictionary now consists of more than eight thousand terms related to energy. In addition, as depicted in Figure 5.1b English still dominate the entire dictionary, but we have successfully balance the number of terms for both Bahasa Indonesia and Dutch.

5.1.2 Dictionary Quality Assessment

The generated energy consumption dictionary plays a pivotal role for this research. The quality of terms enlisted in the dictionary effect the performance of our dictionary-based stream annotation process. As a standard approach, we filter out any undesirable or irrelevant terms using several defined rules and followed by a manual assessment. The rules that we followed are:

[6] https://github.com/commonsense/conceptnet5/wiki/Relations
[7] https://github.com/commonsense/conceptnet5/wiki/Languages
[8] https://github.com/stopwords-iso

Class	ConceptNet Expanded Dictionary		
	Seed	*ConceptNet*	TOTAL
Dwelling	207	1,795	2,002
Food	493	2,684	3,177
Leisure	145	1,418	1,563
Mobility	167	1,420	1,587
TOTAL	**1012**	**7,317**	**8,329**

TABLE 5.2: Expanded Energy Consumption Dictionary by Source

- Term must have minimum three characters
- Term must have minimum one word
- Term must have maximum two words
- Term must not contain any stop words
- Term must be registered in word stem (base or root form)
- Term must be registered in valid languages: English (en), Bahasa Indonesia (id) and Dutch (nl)
- Term must be unique across energy consumption domain (or class)
- Term must be relevant with the respective energy consumption domain (or class)

5.2 Stream: Twitter

The experiments in this paper use forty days of tweets from Twitter collected between January 23, 2017 and February 26, 2017, within the bounding box area of Amsterdam and Jakarta. The corpus contains over 1,306,336 tweets, collected through Twitter Stream API without any keyword filtering.

Based on Masih and Masih (1996) study, population size and age structure have clear effects on different energy consumption behaviors. In our experiment, two world cities from different countries and continents, Jakarta and Amsterdam clearly have different energy consumption behaviors. Taking this consideration, we decided to conduct the experiment for each city separately. Therefore, we split the collected stream dataset based on their city of origin, Amsterdam, and Jakarta with 219,436 and 1,086,900 tweets respectively. As depicted in Figure 5.2, Jakarta significantly generates more data compared to Amsterdam with 83.5% of the entire collected tweets. This finding is aligned with the fact that Jakarta once crowned as the most active twitter city in the world [9].

Similar to the dictionary processing, stream processing initially starts with whitespace tokenization to generate tokens from the text. However, as an unstructured nature of social media, it is mandatory to perform further data cleansing. We identify several entities exist in the text and conduct text processing, such as strip any non-alphanumeric characters, remove overly short tokens (lower than 4 characters), remove any user mentions, and remove any URL address found. Consequently, we

[9] https://www.forbes.com/sites/victorlipman/2012/12/30/the-worlds-most-active-twitter-city-you-wont-guess-it

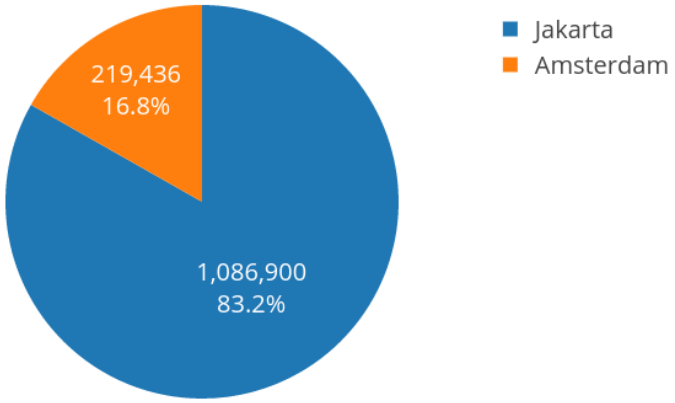


FIGURE 5.2: Collected Twitter Data by City

also remove any stop words found in the text and reduce any inflectional word for each token using word stemming. Finally, following our decision to limit dictionary terms into two words, we generate `two`-shingles for each document as additional tokens.

5.3 Matching Stream with Energy Consumption Dictionary

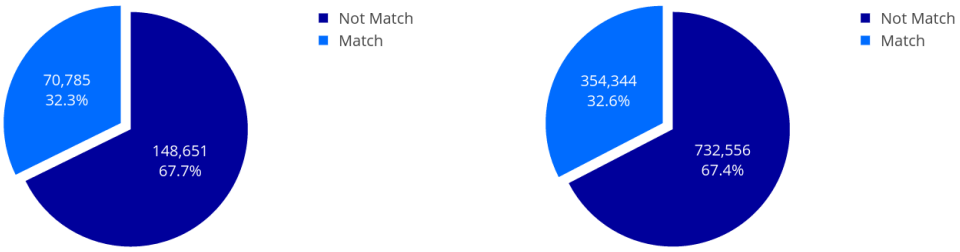
Using the processed dictionary and stream from previous steps, we are able to conduct dictionary-based stream annotation. This process goal is to generate an energy consumption labeled stream that will be used as the dataset for our distant supervision learning approach that we have described in subsection 3.2.3.

This process is considered as one of the steps that require extensive computation due to the complex iterative algorithm. For each document in the corpus, we need to match document's tokens with terms in our dictionary and annotate relevant energy consumption sectors. Therefore, iteration for both document tokens and dictionary terms is required. Moreover, we also implement two different matching algorithm, which is exact matching and also partial matching. Although exact matching is considered as a simple algorithm, partially matching obviously increases the complexity of the entire algorithm. In partial matching, we require computing Jaccard index for each document tokens with our entire dictionary terms; followed by identifying a positive partial match for pairs that have a minimum threshold of similarity score. For this threshold, we use 0.75 as the minimum similarity score to accommodate partial matching for typos or slang words.

In this experiment, we train a logistic classifier using two scenarios, which are dictionary-based annotation with (i) only exact match, and (ii) both matching algorithms (i.e. exact and partial). As shown in Table 5.3, the simple exact matching algorithm is consistently outperform partial matching setup in all metrics. Although partial match is successfully match typos or slang words, this capability also falsely annotated tweets. There are cases where words have similar Jaccard coefficient index

Metric		Matching Algorithms	
		Exact	*Exact + Partial*
ROC		0.80	0.79
Accuracy		0.80	0.79
Class 0 Non-EC	Precision	0.87	0.72
	Recall	0.81	0.93
	F1	0.84	0.82
Class 1 EC	Precision	0.70	0.64
	Recall	0.78	0.85
	F1	0.78	0.74

TABLE 5.3: Matching Algorithms Performance



(A) Amsterdam (B) Jakarta

FIGURE 5.3: Match vs Not Match Energy Consumption Dictionary

but contextually not relevant. For instance, a term like "cat" is partially matched with term "car". Based on these findings and also taking into consideration of exhaustive-computation required by the partial matching algorithm, we proceed the experiment using only with the exact matching algorithm.

As a result depicted din Figure 5.3a, we successfully matched 70,785 tweets (32.3%) from Amsterdam dataset and 354,344 matched tweets (32.6%) from Jakarta dataset.

5.4 Feature Extraction: TF-IDF with Hashing Trick and Word2Vec

In this section, we extract quantified and measured features from our stream dataset. These features will be used for our distant supervision machine learning approach and train energy consumption classifier. For this experiment, we choose to extract TF-IDF and Word2Vec.

In addition to our explanation on TF-IDF from previous subsection 3.2.4, this approach could produce a huge amount of features which can be expensive for a large corpus, which is not scalable. Therefore, we use hashing trick to apply a hash function to the term and calculate term frequencies from this hashed terms. The hashing procedure start by generating raw feature that mapped into a term by computing a hash function. Consequently, the term frequencies not be calculated directly from

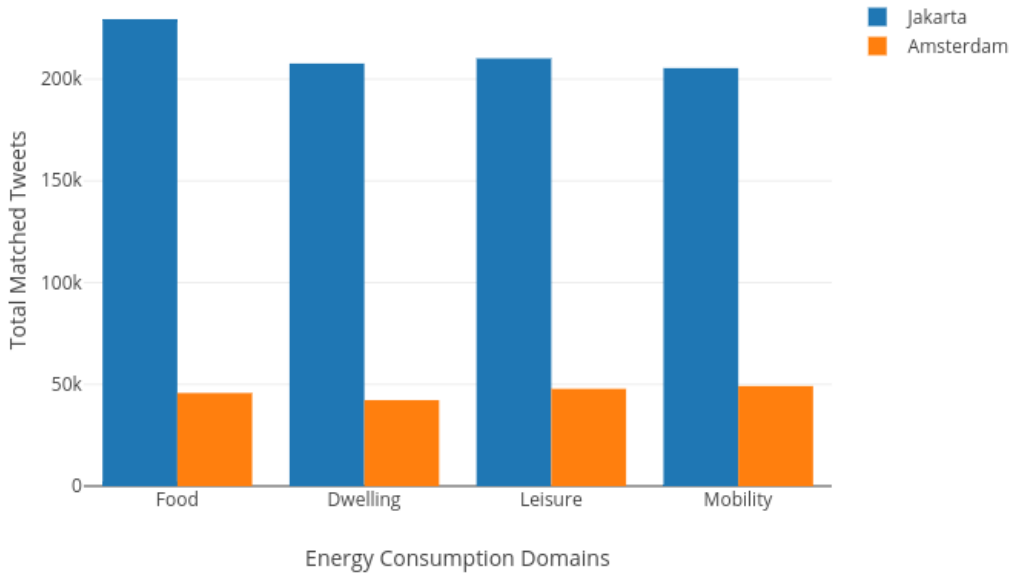


FIGURE 5.4: Dictionary Matched Streams by Class

the tokens, but based on the mapped terms (see Appendix A.2). By using this approach, we can define an arbitrary number of target feature dimensions (buckets of the hash table) that we desired.

Specifically, on TF-IDF features with hashing trick, we are able to use an arbitrary number of features required. However, a different number of features have a direct import on the classifier's performance. Therefore, we conduct an experiment using simple logistic regression with a different number of TF-IDF features to determine the optimal classifier's performance. For this experiment, we use the area under the curve of Receiver Operating Characteristic (ROC) as the performance metric. The curve is computed by ranking classifier's test result in decreasing order. Each positive result raises the lineup, and negative result moves the line right. Both variables are the fractions of the positive and the negative result of the test.

As depicted in Figure 5.5, we could see that using a small number of TF-IDF features, such as 100 only produce a small area under the curve of Receiver Operating Characteristic (ROC), which is 66%. Next, we increase gradually the number of features to determine the optimal number of features. As a result, the increasing number of features tends to also improve AUC. However, starting from 10,000 features there is only insignificant improvement occurred for each increment. Based on this findings, we choose 15,000 hash features as the most optimal number of features.

5.5 Distant Supervision Machine Learning

In this section, we explore different classification algorithms and conduct performance comparison to discover the most suitable classifier for the experiment. We

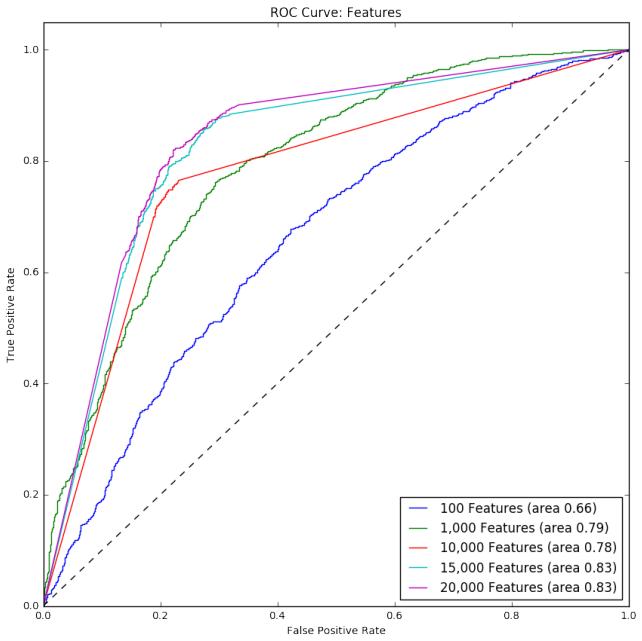


FIGURE 5.5: ROC Curve: Number of TF-IDF Features.

denote our classifier to identify energy consumption as `EC`, and followed by specific classifier for each sector, which are `EC-LEISURE` and `EC-MOBILITY`

5.5.1 Dataset Preparation

In order to train our classifiers, there are several dataset preparation steps that need to be conducted.

First, extract dataset from the dictionary-based annotated stream. Each document is multi-labeled with related energy consumption sectors, which are `dwelling`, `mobility`, `food`, `leisure` and `none`. We require exploding the multi-labeled column to ensure that we have an independent set of the document for each energy consumption sector.

Second, re-assign exploded energy consumption sector dataset and assemble binary dataset. For general energy consumption dataset `EC`, we simply assign any energy consumption sector (i.e. `dwelling`, `mobility`, `food` and `leisure`) as positive, and `none` as negative. On the other hand, for each specific energy consumption sector dataset (i.e. `EC-LEISURE` and `EC-MOBILITY`) we use that particular sector's label as positive, and `none` as negative.

Third, as a result shown on Table 5.4, we can see that the proportion between positive (+) and negative (-) label on every dataset is not balanced. A balanced dataset is favorable to avoid over-fitting when training the classifier. Therefore, we attempt to balance the dataset by randomly sampling each city's dataset. For Amsterdam, we consider 101,000 documents for each class, which consists of 100,000 documents as the training set and 1,000 as the test set. Accordingly, for Jakarta, we consider 11,000 documents for each class, which consists of 10,000 documents as training set and 1,000 as the test set. Both test sets are manually checked and validated to ensure the reliability of classification performance evaluation.

Dataset	EC		EC-LEISURE		EC-MOBILITY	
	+	-	+	-	+	-
Amsterdam	765,115	71,324	40639	71,324	49,194	71,324
Jakarta	765,115	399,048	188,386	399,048	183,664	399,048

TABLE 5.4: Datasets

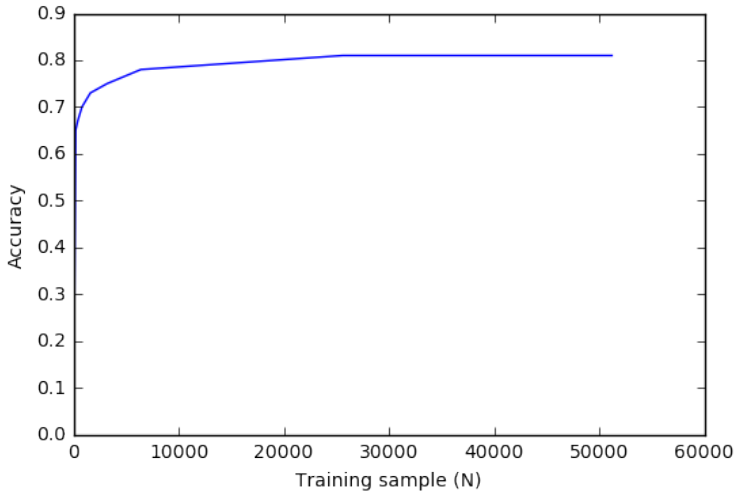


FIGURE 5.6: Classifier: Dataset Sample Size

5.5.2 Classifiers: Setup

Using the balanced dataset from the previous step, we are able to train classifiers and evaluate the performance. In our endeavor to achieve the best classifier for identifying energy consumption behavior from social data, we explore different classification algorithms and several experiment configurations. We evaluate the performance of each classifier by measuring Area Under the Curve (AUC) of Receiver Operating Characteristic (ROC). ROC curve shows how well a classifier can separate positive and negative examples, and identify the best threshold for separating them. The overall performance of the classifier is identified by how big the area under the curve, where 0.5 or half the plot means the classifier have a moderate (or random) performance, and 1.0 for the perfect classifier. Besides ROC, we also provide more detailed metrics, such as precision, recall, and accuracy.

Dataset sample size

Dataset sample size directly impacts on the overall performance of the classifier. A larger sample can yield more accurate results, but excessive samples can be expensive or unnecessary. In this step, we explore different sample size using logistic regression with TF-IDF features. Using random sampling from EC dataset of both cities, we gradually increase the number of sample and benchmark classifier's accuracy. As depicted in Figure 5.6, starting from 5,000 samples there is no significant increase of accuracy, where become stagnant beyond 25,000 samples. Following this result, we proceed the experiment with minimum 25,000 samples for each dataset.

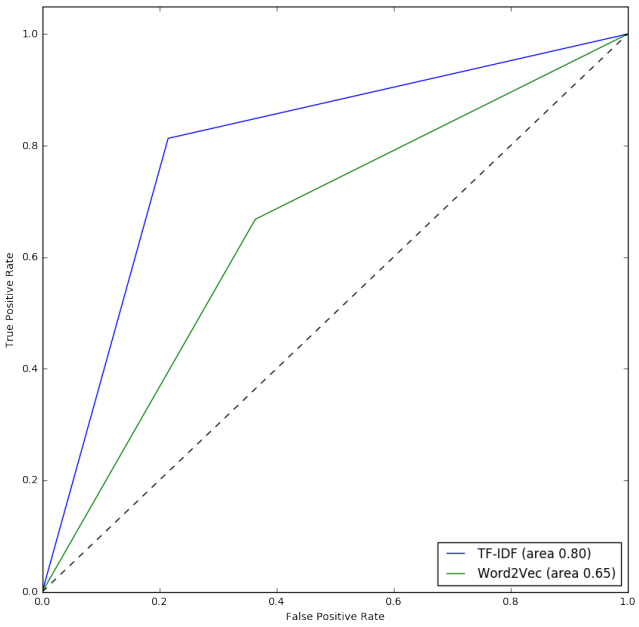


FIGURE 5.7: ROC Curves: TF-IDF vs Word2Vec Features

Features: TF-IDF vs Word2Vec

Classifier trained using features extracted from social data. These features aim to efficiently represent distinct characteristics of data as a compact feature vector. In this step, we compare the performance of classifier based on extracted features from two different methods: TF-IDF and Word2Vec. We use ROC curve to determine the performance of both feature extraction method. ROC curve effectively shows the degree of separation between classes. Based on Figure 5.7, we can see that classifier using TF-IDF features significantly outperform Word2Vec. This result shows that TF-IDF features are more effective to achieve high performing classifier.

Classification algorithm: Logistic Regression vs SVMs

We explore two different classification algorithms for our experiment. Logistic regression is the common algorithms for binomial classification problem. However, Support Vector Machines (SVMs) is one of the popular classification algorithms, which is widely used in industry. Using ROC curve as performance metric depicted in Figure 5.8, logistic regression is slightly performed better than SVMs. Moreover, logistic regression is rather a simple algorithm and inexpensive compared to SVMs. Based on this findings, we choose logistic regression as our classification algorithm.

Threshold tuning

Threshold tuning is important for binary classification problem, because the trained model may output a probability for each class. Thus, there are some cases that model need to be tuned in order to predict a class when the probability is very high. Based on our research objective, we require identifying energy consumption related social data precisely. As a result, a trade-off between precision and recall need to be tuned for this purpose. Initially, as in previous steps, we are using the default 0.5

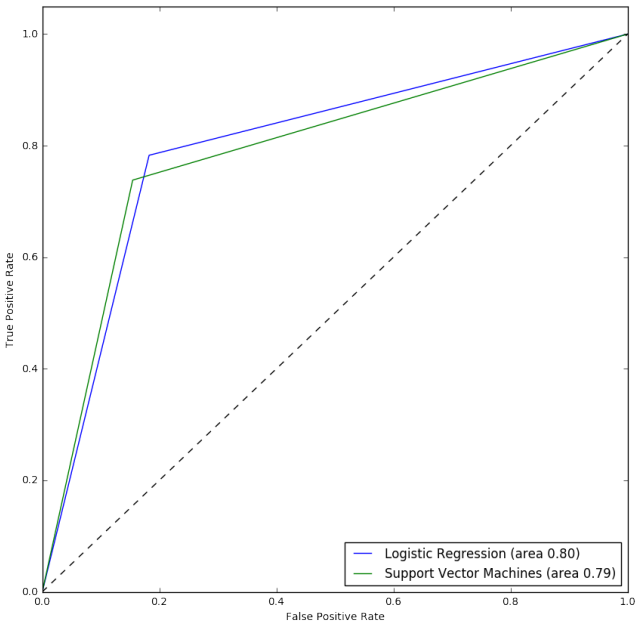


FIGURE 5.8: ROC Curves: Logistic Regression vs SVMs

Metric		Threshold			
		0.5	*0.6*	*0.7*	*0.8*
ROC		0.80	0.80	0.75	0.67
Accuracy		0.78	0.80	0.79	0.75
Class 0 Non-EC	Precision	0.91	0.87	0.80	0.74
	Recall	0.72	0.81	0.88	0.93
	F1	0.80	0.84	0.84	0.82
Class 1 EC	Precision	0.64	0.70	0.75	0.74
	Recall	0.88	0.78	0.62	0.41
	F1	0.74	0.74	0.68	0.54

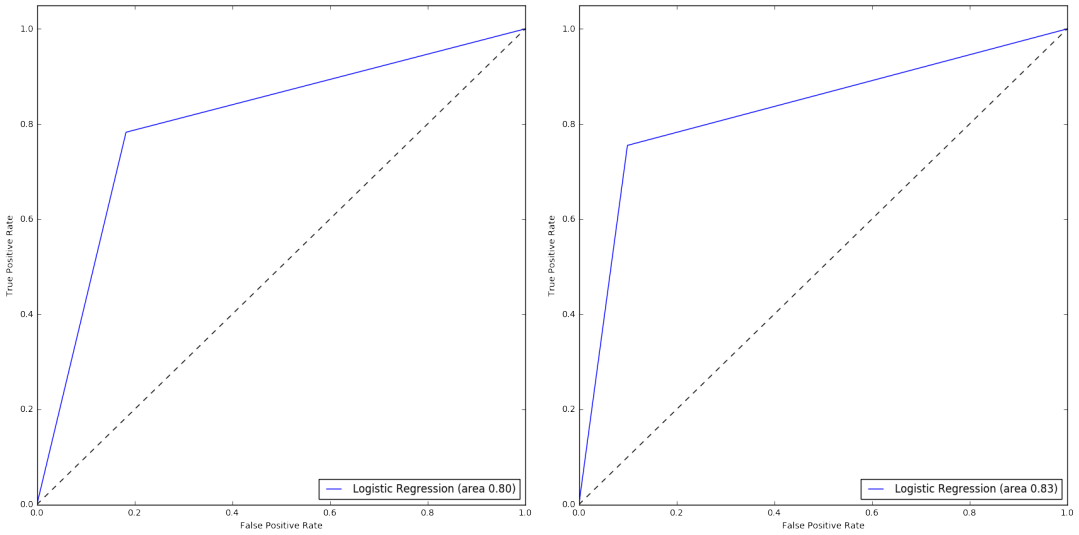
TABLE 5.5: Binomial Classifier Threshold Tuning

as the threshold to decide whether a social data (i.e. micro post) is related to energy consumption or not. In this experiment, we explore different (bigger) threshold configurations for our classifier to achieve better performance.

Based on Table 5.5, increasing threshold have a significant impact on improving the precision of the (EC) class. However, as a trade off, recall is deteriorating if we increase the threshold. We consider to have more precise result with low false positive, but also with a moderate recall. As a result, we choose 0.6 as the threshold of our classifier.

5.5.3 Classifiers: Energy Consumption in Amsterdam and Jakarta

Based on classifier's setup described in the previous section, we begin our experiment using the separated dataset from Jakarta and Amsterdam. EC classifiers performed very well in both cities. As depicted in Figure 5.9, Logistic Regression algorithm achieve AUC ROC 80% for Amsterdam, and 83% for Jakarta. Furthermore, specific energy consumption sector classifiers (e.g. EC-LEISURE and EC-MOBILITY)



(A) Amsterdam (B) Jakarta

FIGURE 5.9: EC Classifiers ROC Curve

also have slightly similar AUC ROC for both cities, as depicted in Figure 5.10 and Figure 5.11 respectively.

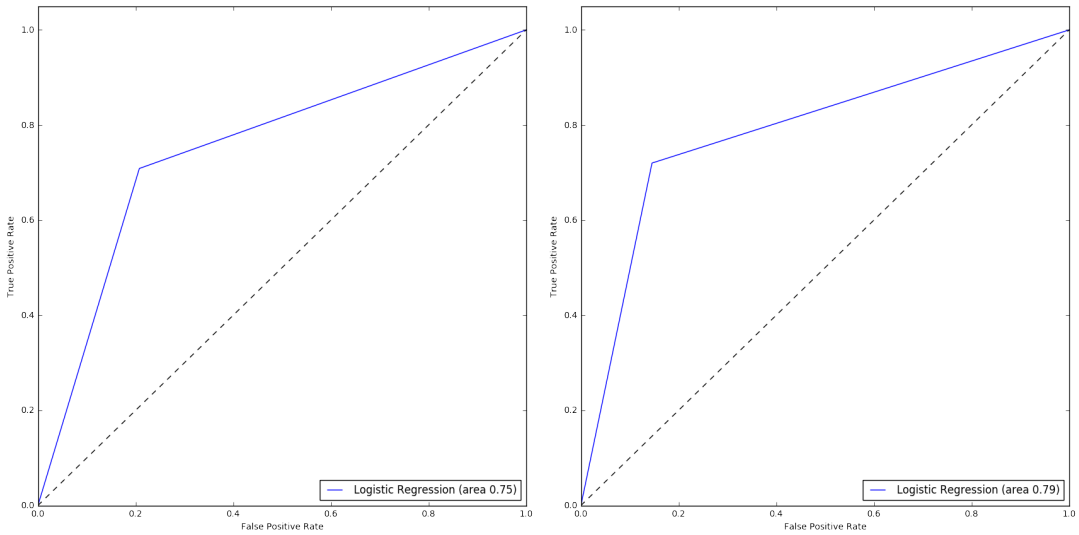
Although we are able to achieve high AUC ROC for all classifiers (i.e. above 75%), there are performance differences on predicting energy consumption related tweets or positive class (Class 1 EC) on all datasets. As shown in Table 5.6, classifiers trained with Jakarta dataset achieve better precision compared to Amsterdam dataset. Next, general energy consumption classifiers (`EC`) outperform specific energy consumption sector classifiers (`EC-LEISURE` and `EC-MOBILITY`) for both cities. Furthermore, we found that `EC-MOBILITY` have low precision on predicting positive class or energy consumption related in mobility sector for both Amsterdam and Jakarta dataset, 56% and 65% respectively.

5.5.4 Classifiers vs Dictionary-based Annotation

In this subsection, we compare classifiers performance with dictionary-based annotation using their own respective test dataset. By comparing evaluation metrics for dictionary-based annotation performance shown in Table 5.7, and classifier performance in Table 5.6, we found that dictionary-annotation generally outperform classifiers. For example, on Jakarta dataset `EC` dictionary-based annotation achieve 89% accuracy, and the classifier is only 80% accuracy. Although dictionary-based annotation performs better, the performance gap (i.e. accuracy) is not beyond 10 percent. Furthermore, take into consideration excessive computing power required to conduct dictionary-based annotation, especially on a large dictionary, the classifier is more applicable and scalable.

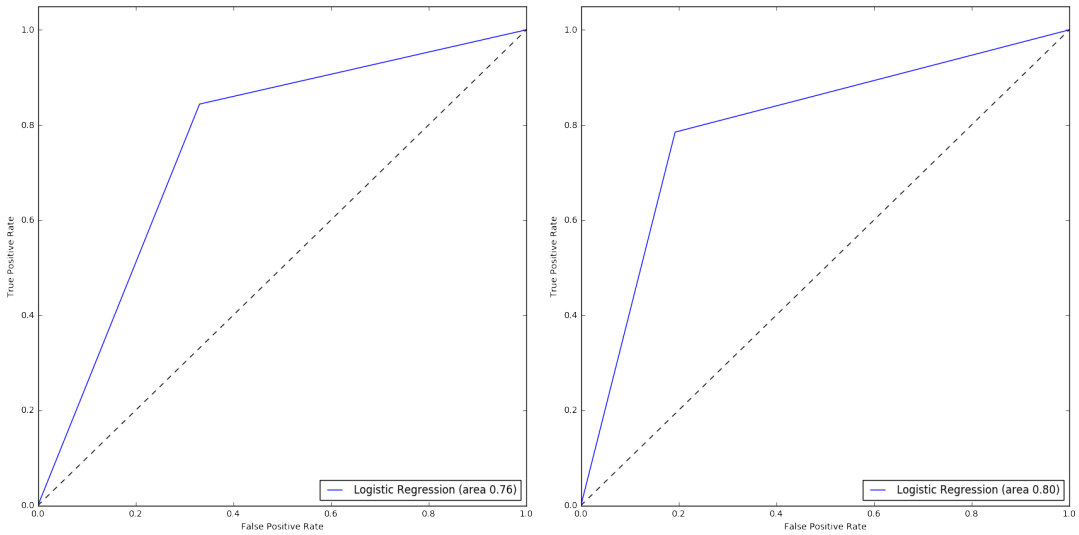
5.6 Analysis

In this experiment, we successfully implemented the framework by generating energy consumption dictionary, conducting dictionary-based stream annotation and implementing distant supervision classifiers. As a result, using the classifiers, we



(A) Amsterdam (B) Jakarta

FIGURE 5.10: EC-LEISURE Classifiers ROC Curve



(A) Amsterdam (B) Jakarta

FIGURE 5.11: EC-MOBILITY Classifiers ROC Curve

Metric		*EC*	*EC-LEISURE*	*EC-MOBILITY*
Amsterdam Dataset				
ROC		0.80	0.75	0.80
Accuracy		0.80	0.76	0.80
Class 0 Non-EC	Precision	0.87	0.84	0.86
	Recall	0.81	0.79	0.80
	F1	0.84	0.81	0.83
Class 1 EC	Precision	0.70	0.62	0.56
	Recall	0.78	0.70	0.74
	F1	0.74	0.66	0.65
Jakarta Dataset				
ROC		0.80	0.75	0.80
Accuracy		0.80	0.76	0.80
Class 0 Non-EC	Precision	0.84	0.83	0.88
	Recall	0.90	0.85	0.80
	F1	0.87	0.84	0.84
Class 1 EC	Precision	0.83	0.74	0.65
	Recall	0.75	0.72	0.78
	F1	0.79	0.73	0.71

TABLE 5.6: Classifier Performances

are able to discover energy consumption behavior using social media data as alternative sources of energy consumption data.

In this section, we present several analysis approaches that can be used to study energy consumption behavior patterns.

5.6.1 Temporal Analysis: Electricity Load Correlation

This analysis explores the capability of social media data to approximate real world trends or events. As the scope of this thesis, we would like to discover whether there is a correlation between actual electricity load and a total number of posts related to energy consumption.

The analysis begins by collecting related data for the analysis, which is actual electricity load (consumption). We found that in the Netherlands, there is a company named Tennet[10], a state-owned company, that operates the entire electricity transmission system of the Netherlands. On their website, they provide such actual electricity load dashboard, which is very relevant to our needs. Although the website did not provide raw data (i.e. CSV or spreadsheets) to be download, we can crawl the dashboard to collect the data. As a result, we have an actual electricity load in megawatt hour (MW) for every 15 minutes (see figure 5.12).

However, we are unable to retrieve an actual electricity load specifically in the city of Amsterdam, because the data is already aggregated within the Netherlands. On the other hand, we are unable to find any similar data for Jakarta. Due to data limitation, in this section we only cover electricity load estimation for Amsterdam. Based on Tennet's data, we are using the ratio of 2016 population data for both The Netherlands (16,979,729) and city of Amsterdam (851,373), which is 5% of actual electricity load.

Using the actual electricity load data, we are able to identify the correlation with our energy-consumption-labeled streams. It is important to note that these two time-series data have metric discrepancies, where Tennet's actual electricity load is

[10]https://www.tennet.eu/

Metric		*EC*	*EC-LEISURE*	*EC-MOBILITY*
Amsterdam Dataset				
Accuracy		0.89	0.85	0.81
Class 0 Non-EC	Precision	0.95	0.92	0.93
	Recall	0.86	0.84	0.85
	F1	0.89	0.83	0.89
Class 1 EC	Precision	0.84	0.80	0.77
	Recall	0.90	0.83	0.81
	F1	0.87	0.81	0.79
Jakarta Dataset				
Accuracy		0.89	0.87	0.88
Class 0 Non-EC	Precision	0.97	0.96	0.97
	Recall	0.90	0.88	0.89
	F1	0.93	0.92	0.93
Class 1 EC	Precision	0.95	0.83	0.72
	Recall	0.82	0.79	0.84
	F1	0.88	0.81	0.78

TABLE 5.7: Dictionary-based Annotation Performances

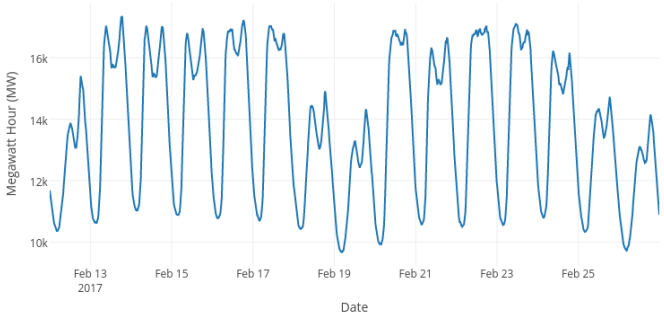


FIGURE 5.12: The Netherlands Actual Electricity Load

reported every 15 minutes and social media data in seconds. Following this discrepancies, we require to aggregate existing streams also into 15 minutes interval. Furthermore, it is also obvious that these both time series have two different unit values, which are Megawatt hour and a total number of posts. In order to overcome this, we require normalizing the value by rescaling into the same scale. As depicted in Figure 5.13, we can see clearly that there is such correlation between these two data, where peaks and slopes occurred in relatively at the same time.

In detail, we conduct quantitative analysis to measure the correlation between these two time-series data using Pearson correlation coefficient. First, we are using descriptive statistics to identify the characteristic of each time-series data (see table 5.8). Actual Electricity load has a very wide standard deviation (SD = 2,999.47), which reflect electricity load's various huge difference within different time. Pearson Correlation between Actual Electricity Load in the Netherlands and Energy Consumption Streams in Amsterdam is high at 0.73.

Following this strong and positive correlation, we are able to use social media data to approximate electricity load. Finally, we can approximate the actual electricity load in Amsterdam by rescaling stream data based on Tennet's data using

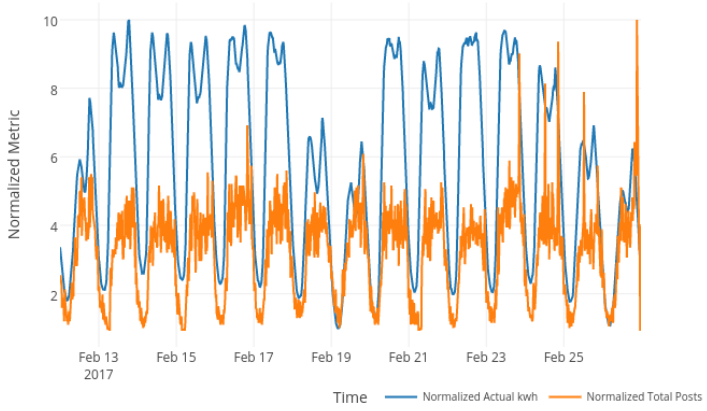


FIGURE 5.13: Normalized Actual Electricity Load vs Normalized Total Stream Posts

	Mean	*Min*	*Max*	*Std. Deviation*	*N*
Actual Electricity Load (NL)	260,039	9,674.56	17,356.64	2,299.47	1,440
Energy Consumption Streams (AMS)	46.57	0	185	26.52	1,440

TABLE 5.8: Descriptive Statistics: Actual Electricity Load vs Energy Consumption Streams

equation 5.1.

$$\frac{max_{kwh} - min_{kwh}}{max_{post} - min_{post}} \cdot (v - max_{post}) + max_{kwh} \tag{5.1}$$

5.6.2 Spatial Analysis and Visualization

Using Twitter stream dataset that we have described in Section 5.2, we found that only 20% or 258,872 tweets have geographic location information. Based on that data, most of the tweets are originated from Jakarta (236,635 tweets) and only small portion originated from Amsterdam (22,237 tweets).

5.6.3 Heat Maps

We plot a clustered tweets of energy consumption for both leisure and mobility sector as a heat map to visualize the spatial dynamics of the city.

Heat maps clearly visualize the point of interests or attractions for leisure in Amsterdam and Jakarta, as depicted in Figure 5.14 and Figure 5.15 respectively. For Amsterdam, the heat map is dominated in the old center area and the canal ring of Amsterdam. These areas are known as the epicenter of tourism in Amsterdam that offers several leisure attractions, such as museums, theaters or casinos. On the other hand, Jakarta's energy consumption for leisure is focused on several malls or shopping arcades, which are located in several Jakarta's arterial roads.

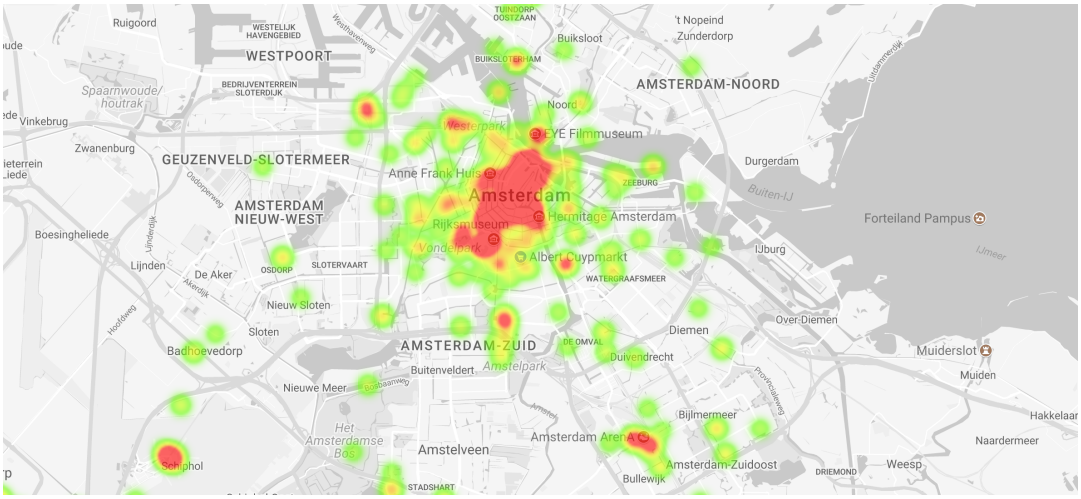


FIGURE 5.14: Leisure Energy Consumption in Amsterdam

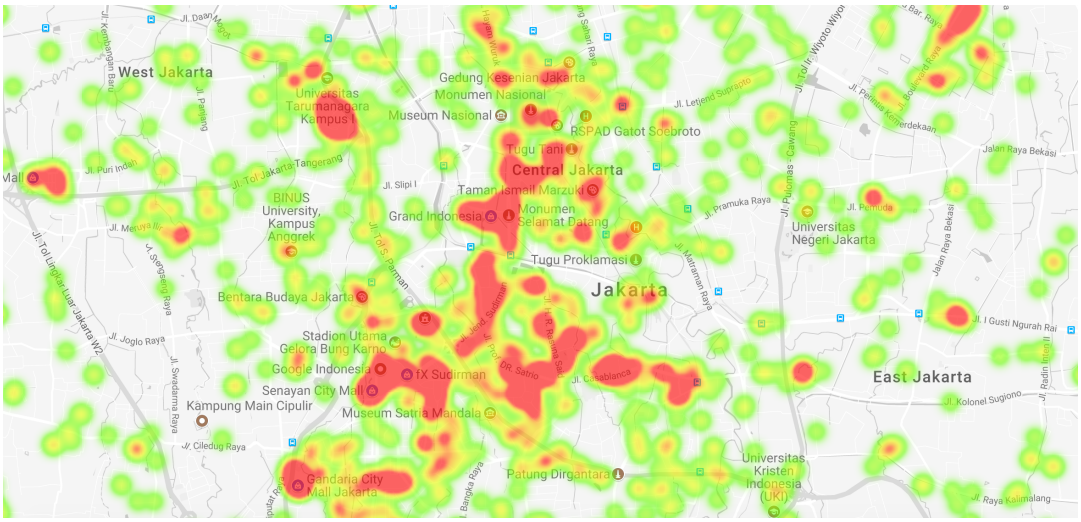


FIGURE 5.15: Leisure Energy Consumption in Jakarta

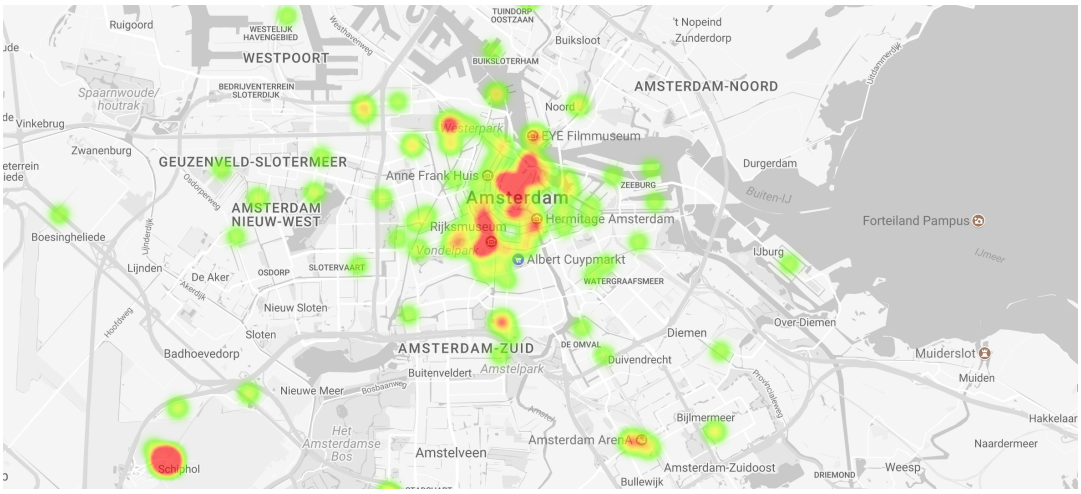


FIGURE 5.16: Mobility Energy Consumption in Amsterdam

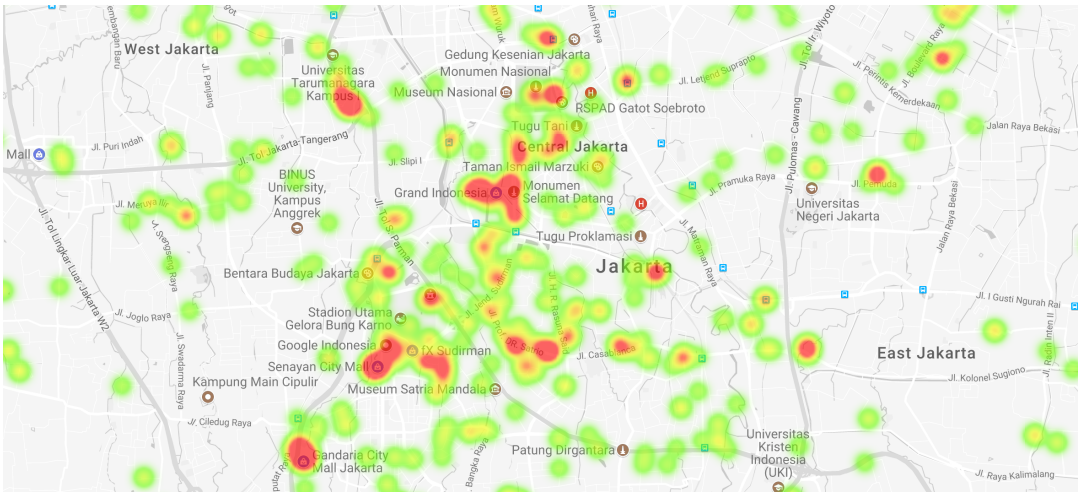


FIGURE 5.17: Mobility Energy Consumption in Jakarta

Next, we visualize heat map for energy consumption for mobility sector in both cities. Both heat maps depicted in Figure 5.16 and Figure 5.17 have accurately identified major transportation hubs, such as airport and train station. However, there are several questionable points of interests that highlighted as energy consumption in mobility sector. For example, we can observe a similar pattern that we found also on leisure's energy consumption heat maps. However, this phenomenon can also be correct, because such places (e.g. malls, shopping arcades, or museums) attracted people to go.

User Displacement distribution

In addition, to have a better understanding on energy consumption in mobility sector, we explore distance-based displacement of consecutive check-ins made by users. Based on the dataset, we discovered unique users of Amsterdam and Jakarta are 5,121 and and 63.817 respectively. Next, we identified there are 9,217 consecutive check-ins (i.e. trip between two points) in Amsterdam, and 138,832 consecutive check-ins for Jakarta.

As depicted in Figure 5.18 and Figure 5.19, users in both cities tend to travel in a short distance or up to 1 KM frequently, and less frequent beyond than 1 KM. However, in Amsterdam's user displacement distribution, frequency along the x-axis (distance displacement) decreases gradually. On the other hand, there is significant decrease in Jakarta's user displacement distribution between 1 KM and 2 KM, which later also decreased gradually along the x-axis. The findings confirmed that both cities are indeed an urban area, which most people did not require to have long distance trip frequently to satisfy their needs.

Radius of Gyration

Using the same consecutive check-ins from the previous section, we consider the radius of gyration of each user to identify standard deviation of distance between the user's check-ins and the user's center of mass. This metric determines whether a user travels mainly locally or long distance.

As depicted in Figure 5.20 and Figure 5.21, most user only travel within 1 KM from their center of mass. Therefore, we can conclude that user within the city of

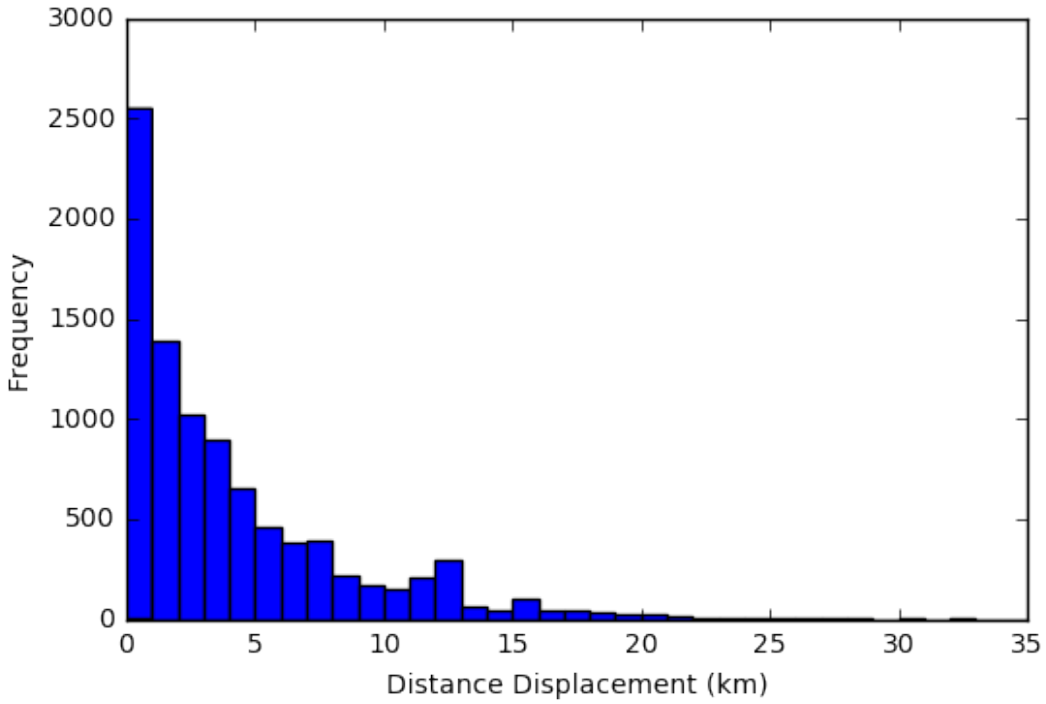


FIGURE 5.18: User Displacement Distribution in Amsterdam

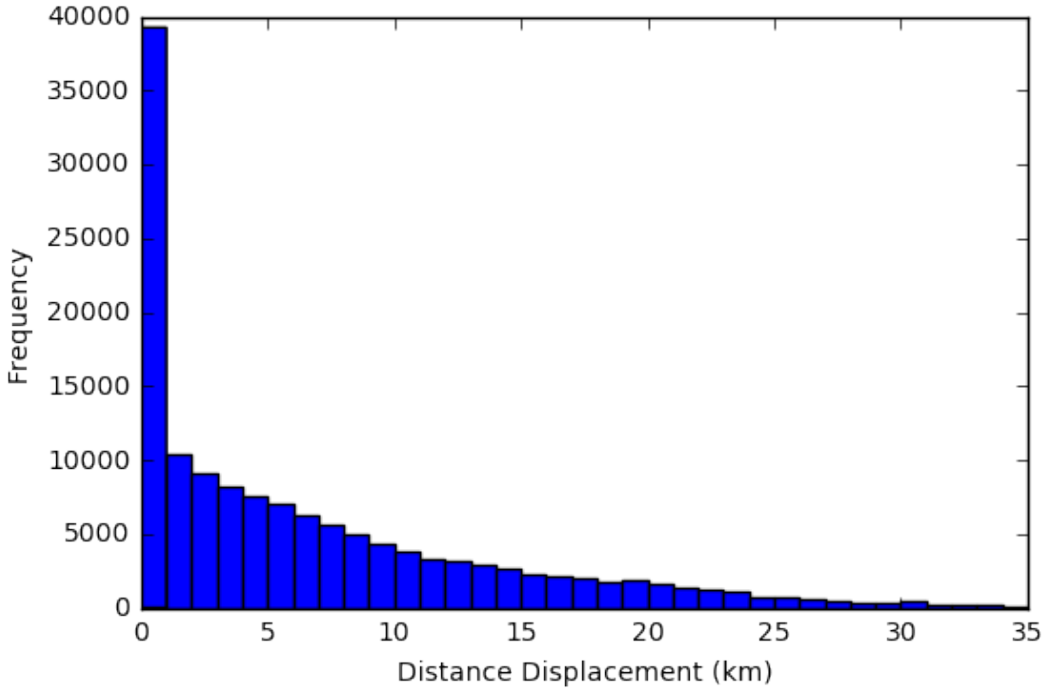


FIGURE 5.19: User Displacement Distribution in Jakarta

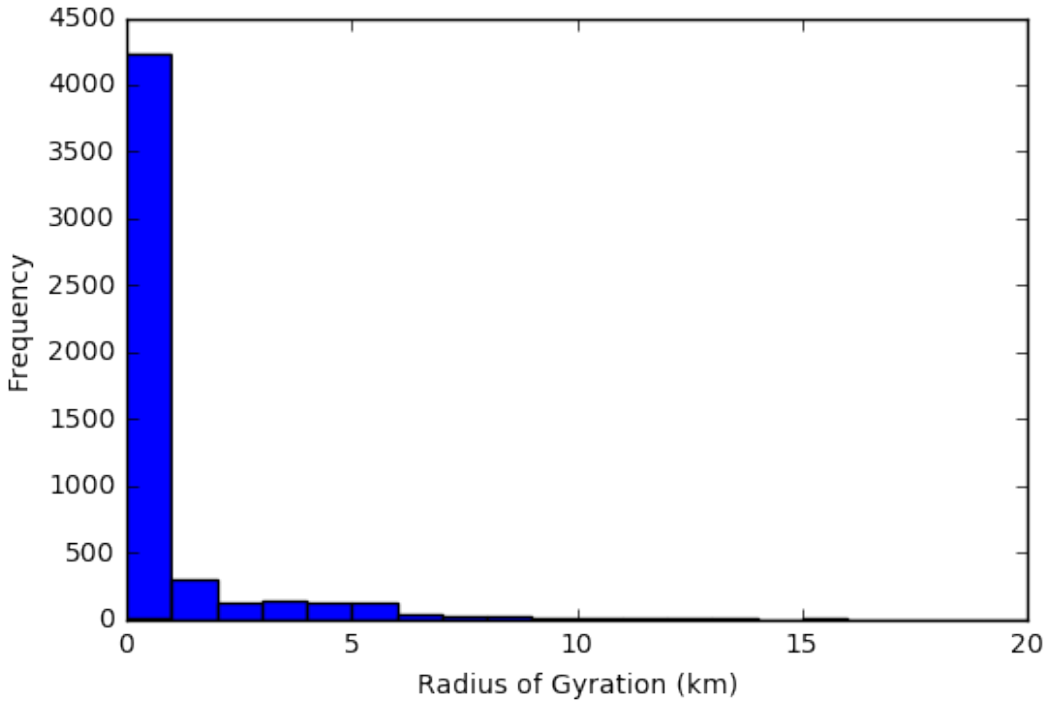


FIGURE 5.20: Radius of Gyration Distribution in Amsterdam

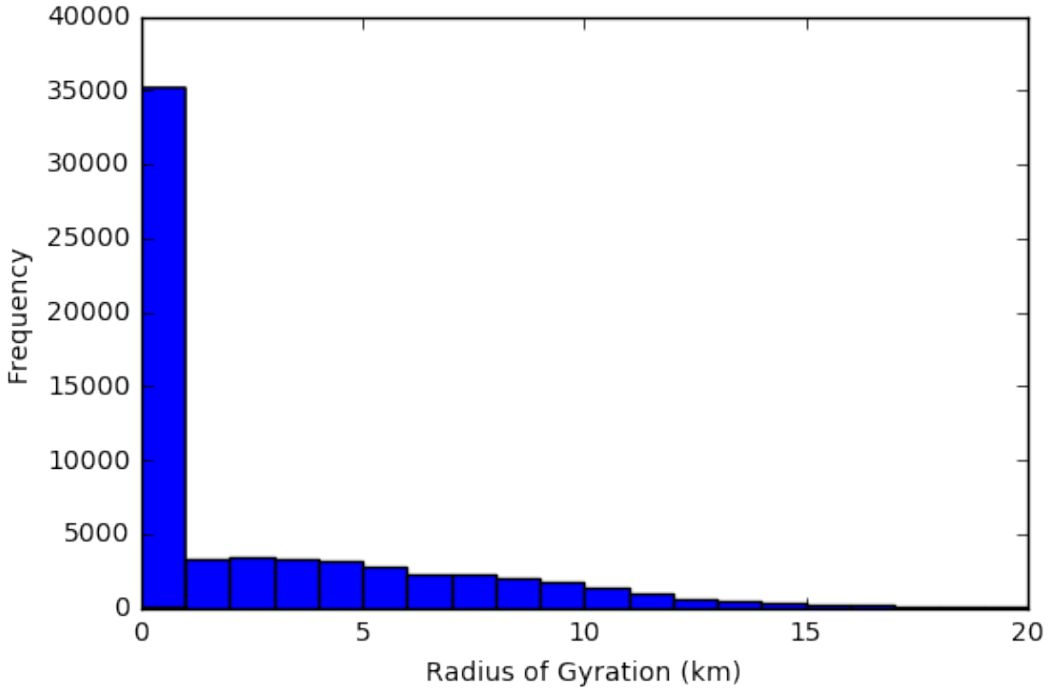


FIGURE 5.21: Radius of Gyration Distribution in Jakarta

Amsterdam and Jakarta tends to travel locally or within the city area. However, it is important to note that radius of gyration in Jakarta has a long tail, which stretches more than 15 KM. This finding shows that there is also a significant number of people that require having a long distance trip.

Chapter 6

Discussions and Conclusions

6.1 Discussions

In this chapter, we will reflect on the design and implementation of the framework from previous chapters.

6.1.1 Discover Energy Consumption Behaviors from Social Data

Social data has been increasingly used as a sensor to discover human activity and infer real-world event or phenomenon (Bodnar et al., 2014). In an effort to clearly demonstrate the usefulness of social data to discover energy consumption behaviors, we conduct correlation study between annotated energy consumption tweets and actual energy consumption data. On this study, we explore four different energy consumption sectors, which are dwelling, mobility, leisure, and food. Unfortunately, it is clear that there are no specific real-world data that able to quantify energy consumption on the specific particular sector. For example on mobility sector, people are able to move using a different mode of transport. Each mode of transport consumed different energy resources, such as a conventional car using petrol, a bus using gas and train using electricity. Although energy consumption data for each energy source is available, we can not generalize that these energy sources are being used exclusively for mobility sector. On the other hand, the same mode of transport can also use different energy source, such as an electric car that using electricity.

As a consequence, there is no definitive dataset to evaluate our system. On the contrary, Marchetti-Bowick and Chambers, 2012 study on political forecasting using Twitter have a definitive dataset (i.e. Gallup daily polling) that can be used as a ground-truth to evaluate their system performance. In this thesis, we used one particular energy resource dataset to evaluate our system, which is electricity. Despite several biases that can be occurred as discussed earlier, electricity consumption undoubtedly occurred in all energy consumption sectors. As depicted in Figure 5.13, we found a high degree of correlation between energy consumption related tweets with actual electricity load. Based on this result, we conclude that social data is suitable as a supplementary data source for energy consumption data.

6.1.2 Dictionary Extensiveness

Dictionary plays a pivotal role in the framework. We use the dictionary as part of dictionary-based annotation process to identify energy consumption related tweets that we used to train energy consumption classifiers. Energy consumption is clearly a broad domain, which needs to be represented by an extensive dictionary to achieve high recall of energy consumption related tweets. However, an extensive dictionary does not imply a high precision performance due to the existence of term biases.

In order to overcome these challenges, we ensure the extensiveness of the dictionary by expanding the seed dictionary using ConceptNet and manually assess the quality of the terms to suppress term biases. Our experiment have successfully identified 8,329 relevant terms in four different energy consumption lifestyle sectors as our dictionary (see Table 5.2).

As shown in Table 5.6, `EC` classifier outperform sector specific classifiers, `EC-LEISURE` and `EC-MOBILITY`. In brief, `EC` consider the entire terms in the dictionary. On the other hand, sector specific classifiers only consider terms that relevant (or labeled) with the respective sector. This result shows that dictionary extensiveness clearly has direct effect on classifier's performance. However, an extensive dictionary does not always require a high number of terms. Dictionary extensiveness depends on the dictionary scope of the domain itself. For example, Marchetti-Bowick and Chambers, 2012 successfully achieve high performing classifiers by using only six terms in average for each (political) dictionary.

6.2 Threat of Validity

As our work was confined to its time frame, so was the scope of this thesis. We have carefully to make our analysis as sound as possible, by following state-of-the-art methodologies and approaches from recent studies in the field. However, there are some threats to its validity because of the exploratory nature of this study.

6.2.1 Social Data as Data Source

Social media users are not a representative sample of the actual population. It is important to note that social data is generated by over-representation of affluent and tech-savvy demographic groups, which can result in sampling bias. Moreover, as our experiment data source, we collect social data from Twitter Stream API that only provide 1% of the actual Twitter Stream. In addition, there are also content pollutions or noises in social data generated by bots or malicious users. Despite these problems, our experiment shows a strong correlation between energy consumption related tweets and actual energy consumption (electricity) within the same period of time.

6.2.2 Ambiguity of Energy Consumption Behaviors in Social Data

The unstructured nature of social data is problematic to correctly identify energy consumption behaviors. Using dictionary-based annotation approach, we simply identify any occurrences of dictionary's terms in the given tweets. However, a single term can have different meanings which heavily depend on the sentence's context. For example, if we identify the term "*car*" as energy consumption within mobility sector, this term can mistakenly annotate sentence that consists "*car toys*" as mobility. Next, there is also time ambiguity within sentence's context where energy consumption is not currently happening at the given time. For example, if we identify term "*concert*" as energy consumption within leisure sector, this term can mistakenly annotate sentence "*going to #coldplay concert next month!*" as leisure, although the concert itself is not currently happening at the given time of the tweet. Users basically can share everything on Twitter, and not exclusively share an activity that consumes energy. For example, a tweet like "I'm selling my car VW Golf 2015. DM!" consists term "*car*", but it is clearly not an energy consumption for mobility sector.

6.2.3 Geo-tagged Information in Social Data

In our endeavor to discover energy consumption in mobility sector, we make use of the geographic information embedded in social data. Following this approach, we can expect another sampling biases due to the limited number of GPS-enabled devices are used by users to generate a geo-tagged social data. Based on our dataset, there are only 20% of the dataset or 258,635 tweets that have embedded with geographic information.

6.2.4 Limited Classifier Algorithms

Due to time constraints, we limited our experiment with only two different classifier algorithms, which are Logistic Regression and Support Vector Machines (SVMs). In chapter 5, we explore the performance of both classifiers and proceed the experiment with the best-performed classifier, which is Logistic Regression. However, there are several other classifier algorithms available that can be used in this study and may provide different results.

6.2.5 Limited Baseline Data for Evaluation

In order to evaluate the performance of our framework, a baseline data is required. However, as discussed in Section 6.1, there is no definitive dataset to evaluate our system. Therefore, we used any other existing relevant energy consumption data, such as energy source consumptions.

Unfortunately, a baseline data with high granularity is difficult to acquire. In order to conduct a comprehensive evaluation, at least we require a baseline data that have a similar level of granularity with social data. In our experiment, we have successfully acquired actual electricity load of the Netherlands for every 15 minutes interval. As the baseline data is at the country level, we estimate actual electricity load for the city of Amsterdam based on population ratio. This approach could introduce biases because there are many other factors that can affect electricity consumption besides population. On the other hand, we are unable to gather actual electricity load data with such standard from Jakarta, or even Indonesia. Therefore, we recommend using any other relevant energy consumption data in future studies to confirm our findings.

6.3 Conclusion

In this thesis, we have designed and presented a general framework to discover knowledge about energy consumption behavior patterns. In our implementation, we have used Twitter as a data source, and conduct experiment with more than 1 million tweets originated from two different cities, Amsterdam (The Netherlands) and Jakarta (Indonesia).

We tested our implementation using real world energy consumption data, which is actual electricity load. This analysis showed promising results on identifying energy consumption related tweets using our classifiers. Overall, we can conclude that social data can be used as an alternative data source for studying energy consumption behaviors.

6.3.1 Research Questions

Our main research question *"How to discover behavioral patterns of people regarding energy consumption in cities, using social media data?"* was divided into three separate research sub-questions.

RQ1: What is the state-of-the art in social media data analytics on discovering energy consumption in urban contexts?

In Chapter 2, we presented related works on social media data analytics in general, and specifically in understanding energy consumption behavior using social media data. We discovered that existing energy consumption studies are heavily depended on conventional data collection methodology, such as a labor-extensive survey. Despite the limitation, these studies able to unlock the potential use of energy consumption data for different applications. Next, we have identified previous study efforts on discovering human activity from social media data. Social data provide a great wealth of information, but the unstructured nature of social data require us to apply different approaches or methodologies to extract specific information that we want to discover, which is energy consumption.

RQ2: How could we design, develop, and implement a knowledge discovery framework to identify energy-related consumption behavior of people in cities using social data?

Taking into consideration of previous related works in Chapter 2, we designed a general framework for studying energy consumption using social data. In Chapter 3, we introduced a knowledge discovery framework which presented as a data pipeline that consists of three main stages, which are Data Collection, Data Processing & Enrichment, and Data Analysis. Next, we implement the framework as a software library that leverages big data processing framework from Apache Spark as described in Chapter 4. The software library is published as an open source project[1], which can be used for future studies.

RQ3: What is the performance of the proposed framework in various energy-related use cases?

In Chapter 5, we explore the performance of the framework in two energy consumption behavior sectors, which are leisure and mobility. Despite training with a noisy dataset, our distant-supervision learning approach yield a good performance on both precision and recall. Moreover, as discussed in Section 5.5.4, our classifiers also achieve a considerably similar performance compared to the computation-extensive dictionary-based annotation approach for both energy consumption behavior sectors. Finally, energy consumption related tweets identified by our classifiers have a positive strong correlation with the evaluation data (actual electricity load).

6.4 Outlook

In this section we laid out possible future research that can extend the scope of our research and reduce some of the threats to validity.

[1] https://github.com/arkka/socioknowledge

6.4.1 Specific Energy Consumption Activity Type

As indicated in Section 6.1.2, dictionary plays a pivotal role that determine framework's overall performance. Our domain problem, energy consumption is considerably as high level and very broad. Therefore, achieving an extensive dictionary that effective for dictionary-based annotation is challenging.

In this thesis, we only categorize energy consumption behavior based on lifestyle sectors (e.g. leisure and mobility). We suggest a more specific domain problem for energy consumption, which is based on energy consumption activity type. For instance, instead energy consumption in dwelling sector for the dictionary domain, we can focus on the specific energy consumption activity, such as cooking or gardening. As a result, with a more focus domain problem, we can find more suitable baseline data to evaluate the performance.

6.4.2 Broadening the classifier algorithm scope

In this thesis, we limited our work to only two different classifier algorithms. We can implement several other classifier algorithms in this framework, and enrich the analysis that we have described in Section 5.5. These additional classifier algorithms should be taken into account to confirm our findings.

Appendix A

Dictionary: Energy Consumption

A.1 Valid ConceptNet Relationships

Relation URI	Description	Example
/r/IsA	A is a subtype or a specific instance of B; every A is a B. This can include specific instances; the distinction between subtypes and instances is often blurry in language. This is the hyponym relation in WordNet.	car -> vehicle; Chicago -> city
/r/PartOf	A is a part of B. This is the part meronym relation in WordNet.	gearshift -> car
/r/HasA	B belongs to A, either as an inherent part or due to a social construct of possession. HasA is often the reverse of PartOf.	bird -> wing; pen ->w ink
/r/UsedFor	A is used for B; the purpose of A is B.	bridge -> cross water
/r/AtLocation	A is a typical location for B, or A is the inherent location of B. Some instances of this would be considered meronyms in WordNet.	butter -> refrigerator; Boston -> Massachusetts
/r/Causes	A and B are events, and it is typical for A to cause B.	exercise -> sweat
/r/HasSubEvent	A and B are events, and B happens as a subevent of A.	eating -> chewing
/r/MotivatedByGoal	Someone does A because they want result B; A is a step toward accomplishing the goal B.	compete -> win
/r/Desires	A is a conscious entity that typically wants B. Many assertions of this type use the appropriate language's word for "person" as A.	person -> love
/r/CreatedBy	B is a process or agent that creates A.	cake -> bake
/r/Synonym	A and B have very similar meanings. They may be translations of each other in different languages. This is the synonym relation in WordNet as well. Symmetric.	sunlight <-> sunshine
/r/DerivedFrom	A is a word or phrase that appears within B and contributes to B's meaning.	pocketbook -> book
/r/Entails	If A is happening, B is also happening. (This may be merged with HasPrerequisite in a later version.)	run -> move
/r/MannerOf	A is a specific way to do B. Similar to "IsA", but for verbs.	auction -> sale
/r/LocatedNear	A and B are typically found near each other. Symmetric.	chair <-> table
/r/EtymologicallyRelatedTo	Etymologically relationship within different culture context	
*/r/dbpedia/**	Relationship that provisionally imported from DBpedia	

TABLE A.1: Valid ConceptNet Relationships

A.2 Hashing Trick Algorithm

```
function hashing_vectorizer(
    features : array of string,
    N : integer
):
  x := new vector[N]
  for f in features:
        h := hash(f)
      idx := h mod N
      if (f) == 1:
          x[idx] += 1
      else:
          x[idx] -= 1
  return x
```

LISTING A.1: Psudocode for Hashing Trick

Bibliography

Abbar, Sofiane, Yelena Mejova, and Ingmar Weber (2015). "You Tweet What You Eat: Studying Food Consumption Through Twitter". In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. CHI '15. Seoul, Republic of Korea: ACM, pp. 3197–3206. ISBN: 978-1-4503-3145-6. DOI: 10.1145/2702123.2702153. URL: http://doi.acm.org/10.1145/2702123.2702153.

Alrowaily, Majed Abdullah and Manolya Kavakli (2015). "The Use of Smart Meters and Social Media in Promoting Conservation Behaviour". In: *Proceedings of the 2015 8th International Conference on U- and e- Service, Science and Technology (UNESST)*. UNESST '15. Washington, DC, USA: IEEE Computer Society, pp. 50–56. ISBN: 978-1-4673-9852-7. DOI: 10.1109/UNESST.2015.24. URL: http://dx.doi.org/10.1109/UNESST.2015.24.

Bampo, Mauro et al. (2008). "The effects of the social structure of digital networks on viral marketing performance". In: *Information systems research* 19.3, pp. 273–290.

Bodnar, T. et al. (2017). "Using Large-Scale Social Media Networks as a Scalable Sensing System for Modeling Real-Time Energy Utilization Patterns". In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* PP.99, pp. 1–14. ISSN: 2168-2216. DOI: 10.1109/TSMC.2016.2618860.

Bodnar, Todd et al. (2014). "On the Ground Validation of Online Diagnosis with Twitter and Medical Records". In: *Proceedings of the 23rd International Conference on World Wide Web*. WWW '14 Companion. Seoul, Korea: ACM, pp. 651–656. ISBN: 978-1-4503-2745-9. DOI: 10.1145/2567948.2579272. URL: http://doi.acm.org/10.1145/2567948.2579272.

Cheng, Zhiyuan et al. (2011). "Exploring millions of footprints in location sharing services." In: *ICWSM* 2011, pp. 81–88.

Chew, Cynthia and Gunther Eysenbach (2010). "Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak". In: *PloS one* 5.11, e14118.

Cranshaw, Justin et al. (2012). "The livehoods project: Utilizing social media to understand the dynamics of a city". In:

Ester, Martin et al. (1996). "Density-based spatial clustering of applications with noise". In: *Int. Conf. Knowledge Discovery and Data Mining*. Vol. 240.

Kraft, John and Arthur Kraft (1978). "On the relationship between energy and GNP". In: *The Journal of Energy and Development*, pp. 401–403.

Kwak, Haewoon et al. (2010). "What is Twitter, a Social Network or a News Media?" In: *Proceedings of the 19th International Conference on World Wide Web*. WWW '10. New York, NY, USA: ACM, pp. 591–600. ISBN: 978-1-60558-799-8. DOI: 10.1145/1772690.1772751. URL: http://doi.acm.org/10.1145/1772690.1772751.

Lee, Chien-Chiang (2005). "Energy consumption and GDP in developing countries: a cointegrated panel analysis". In: *Energy economics* 27.3, pp. 415–427.

Marchetti-Bowick, Micol and Nathanael Chambers (2012). "Learning for Microblogs with Distant Supervision: Political Forecasting with Twitter". In: *Proceedings of*

the 13th Conference of the European Chapter of the Association for Computational Linguistics. EACL '12. Avignon, France: Association for Computational Linguistics, pp. 603–612. ISBN: 978-1-937284-19-0. URL: http://dl.acm.org/citation.cfm?id=2380816.2380890.

Masih, Abul M.M. and Rumi Masih (1996). "Energy consumption, real income and temporal causality: results from a multi-country study based on cointegration and error-correction modelling techniques". In: *Energy Economics* 18.3, pp. 165 – 183. ISSN: 0140-9883. DOI: http://dx.doi.org/10.1016/0140-9883(96)00009-6. URL: http://www.sciencedirect.com/science/article/pii/0140988396000096.

Mesbah, Sepideh et al. (2017). "Semantic Annotation of Data Processing Pipelines in Scientific Publications". In: *European Semantic Web Conference*. Springer, pp. 321–336.

Pan, Bei et al. (2013). "Crowd Sensing of Traffic Anomalies Based on Human Mobility and Social Media". In: *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. SIGSPATIAL'13. New York, NY, USA: ACM, pp. 344–353. ISBN: 978-1-4503-2521-9. DOI: 10.1145/2525314.2525343. URL: http://doi.acm.org/10.1145/2525314.2525343.

Pan, Rui, Masanao Ochi, and Yutaka Matsuo (2013). "Discovering Behavior Patterns from Social Data for Managing Personal Life". In: URL: https://www.aaai.org/ocs/index.php/SSS/SSS13/paper/view/5819.

Revilla, Beatriz Pineda (2016). "Changing Energy Needs: Pursuing 'energy conscious lifestyles' through data-driven social learning and individual behaviour adaptation feedback loops." In:

Shiu, Alice and Pun-Lee Lam (2004). "Electricity consumption and economic growth in China". In: *Energy policy* 32.1, pp. 47–54.

Soytas, Ugur and Ramazan Sari (2003). "Energy consumption and GDP: causality relationship in G-7 countries and emerging markets". In: *Energy economics* 25.1, pp. 33–37.

Soytas, Ugur, Ramazan Sari, and Bradley T Ewing (2007). "Energy consumption, income, and carbon emissions in the United States". In: *Ecological Economics* 62.3, pp. 482–489.

World Energy Council (2016). "World Energy Resources 2016". In:

York, Richard (2007). "Demographic trends and energy consumption in European Union Nations, 1960–2025". In: *Social science research* 36.3, pp. 855–872.

Zheng, Xinye et al. (2014). "Characteristics of residential energy consumption in China: Findings from a household survey". In: *Energy Policy* 75, pp. 126–135.

Zhu, Zack et al. (2013). "Human Activity Recognition Using Social Media Data". In: *Proceedings of the 12th International Conference on Mobile and Ubiquitous Multimedia*. MUM '13. New York, NY, USA: ACM, 21:1–21:10. ISBN: 978-1-4503-2648-3. DOI: 10.1145/2541831.2541852. URL: http://doi.acm.org/10.1145/2541831.2541852.