LAPPEENRANTA-LAHTI UNIVERSITY OF TECHNOLOGY LUT
School of Engineering Science
Software Engineering
Petri Rämö

TWITTER TOPIC MODELING AND SENTIMENT ANALYSIS ON SMART CITIES

Examiners: Assistant Professor Antti Knutas

TIIVISTELMÄ

Lappeenrannan-Lahden teknillinen yliopisto LUT
School of Engineering Science
Tietotekniikan koulutusohjelma
Petri Rämö

Twitter topic modeling and sentiment analysis on Smart cities
Diplomityö 2022

44 sivua, 11 kuva, 6 taulukko, 1 liite
Työn tarkastajat: Apulaisprofessori Antti Knutas
Hakusanat: Aihemallinnus, Tunneanalyysi, Älykaupunki
Keywords: Topic modeling, Sentiment analysis, Smart city

Älykaupungit ovat kehittyneitä kaupunkeja, jotka käyttävät infrastruktuurissa olevaa teknologiaa keräämään dataa ja analysoimaan sitä. Tämä opinnäytetyö tutkii älykaupunkeja Twitterin kautta käyttäen aihemallinnusta ja tunneanalyysiä. Aihemallinnus käyttää LDA metodia ja tunneanalyysi antaa tulokset neutraalina, positiivina tai negatiivina. Twitter on valtava sosiaalisen median sivu joka sisältää lyhyitä viestejä joita kutsutaan twiiteiksi. Twitter data, jota tässä työssä käytettiin, oli Archive Teams Twitter Grab. Tämä sisältää monia miljoonia twiittejä. Tämä data oli jaettu neljään vuosineljännes osaan vuodesta 2020. Tulokset eivät olleet tyydyttäviä, koska data seteistä löytyi niin vähän twiittejä liittyen älykaupunkeihin. Jotkut löydetyt twiitit eivät myöskään olleet järkeviä, koska ne sisälsivät hashtageja, ihmisten nimimerkkejä ja/tai muita kieliä mitkä eivät olleet latinalaisilla aakkosilla. Tämä aiheutti sen, että aihe mallinnus ja tunne analyysi eivät olleet niin mielekkäitä, koska data niiden tekemiseen oli niin vähä ja se ei ollut ideaalia.

ABSTRACT

Lappeenranta-Lahti University of Technology LUT
School of Engineering Science
Software Engineering
Petri Rämö

Twitter topic modeling and sentiment analysis on Smart cities
Master's Thesis 2022

44 pages, 11 figures, 6 tables, 1 appendix
Examiners: Assistant Professor Antti Knutas
Keywords: Topic modeling, Sentiment Analysis, Smart city

Smart cities are developed cities that utilize technology that is in infrastructure to collect data and then analyze it. This thesis studies Smart cities using topic modeling and sentiment analysis to analyze the topics and opinions on Twitter on the subject. Topic modeling utilizes the LDA method and sentiment analysis gives sentiments in neutral, positive, and negative. Twitter is a huge social media site that contains small messages called tweets. Twitter data that was used was from Archive Teams Twitter Grab. This contained multiple millions of tweets. This data has been divided into four quarters of the year 2020. The results were not satisfying as there were so few tweets in these data sets that contained Smart cities. Some of the tweets also didn't make any sense because there are tweets that contain hashtags, people's nicknames, and/or other languages that didn't have Latin alphabets. This made the topic models and sentiment analyses not so meaningful as the data for doing them were so little and not ideal.

ACKNOWLEDGEMENTS

Huge thanks go to Antti Knutas who helped me during the thesis work even though it was quite slow.

TABLE OF CONTENTS

LIST OF SYMBOLS AND ABBREVIATIONS

IaaS	Infrastructure-as-a-Service
IoT	Internet of Things
IT	Information Technology
KDD	Knowledge Discovery from Data
LDA	Latent Dirichlet Allocation
NLP	Natural Language Processing
OT	Operational Technology
PaaS	Platform-as-a-Service
SaaS	Software-as-a-Service

1 INTRODUCTION

This is a master's thesis in computer science made at the Lappeenranta-Lahti University of Technology LUT. This thesis studies smart cities through social media to find out the public topics and opinions. This is being done through topic modeling and sentiment analysis. This section introduces the context and structure of the thesis.

1.1 Background

Smart cities are cities where structures for the city are monitored by different kinds of technologies to design, construct or maintain them. This way computers for example make a solution from the data that is being gathered by this monitoring. These structures can be for example roads, airports, water, and power [1]. Nam and Pardo [2] listed many cities on all continents that were awarded as Smart21 communities where the cities had earned high enough scores in five factors that are part of smart cities. It is expected according to Frost and Sullivan [3] that there will be over 26 smart cities among the global cities and 50% of these cities will be in North America and Europe.

Twitter is a social network where you can send tweets and receive them. You can get these tweets from singular people, businesses, and organizations [4]. Archive team is a collection of people who tries to save our digital heritage [5]. One of these is tweets that they collect on their Twitter stream [6]. With this data from each month, I can do a topic modeling and sentiment analysis to survey topics and sentiments that arise from the data. For this I am making a tool that does this for me utilizing R code language and its Shiny package.

1.2 Goals and delimitations

This thesis has two goals. One is to determine what kind of topics rises from the Twitter data using topic modeling and compare them to other similar kinds of research. The other is to determine what kind of sentiments rises from the Twitter data using sentiment analysis.

For topic modeling, the delimitations are, that we are only using Latent Dirichlet allocation (LDA) to do the topic modeling. The number of topics that are shown is determined by a tool. For sentiment analysis, the delimitations are, that the tool gives three kinds of sentiments: positive, negative, and neutral. For the Twitter data, we have delimited it to contain one year's worth of data for the Archive Team. This one year has been separated into four quarters where every quarter has three months' worth of data.

This thesis has two research questions:

1. What kind of topics rises from the twitter data relating to smart cities?
2. What kind of sentiments do these tweets contain relating to smart cities?

1.3 Structure of the thesis

Section 2 contains a literature review that consists of the Internet of Things, Cloud computing, Smart city, Social media and microblogging, Twitter, Data mining, and Text mining. Section 3 contains methods that were used in this research. These were Quantitative research, topic modeling, and programming language. Section 4 contains a description of the tool that was used in this research. Section 5 shows and analyses the results that this tool gave. Section 7 is about follow-up research that could be done about the same topic. Section 8 contains the summary of the research. After this comes references and an appendix.

2 LITERATURE REVIEW

This literature review was done to get more knowledge of the research problem and its context. This starts by explaining the smart cities and research that has been done before on this topic. After that, it introduces the social media, Twitter, which is being used in this research and research before about this. Last it explains what text mining is, and what is being used to conduct this research about the smart cities from Twitter.

There were multiple sources to get the information for this literature review. The main sources were LUT Primo, IEEE Xplore, and Google Scholar. The search terms, their plurals, and combinations that were used in this literature review: smart city, twitter, sentiment analysis, topic modeling, and text mining. Through these articles, more articles were discovered from their references.

2.1 Internet of Things

Internet of Things (IoT) was a phrase from a 1999 presentation by Kevin Ashton to link radio-frequency identification to Procter & Gamble supply chain [7]. Nowadays IoT can mean devices that are connected to the internet around the clock every day of the week. These devices can be anything nowadays like alarm clocks or televisions in your home or bus schedule board in the bus station. What IoT does to things is give them sensing, actuating, computing, and communication capabilities which makes these things smart. [8]

IoT devices consist of two domains: Information Technology (IT) and Operational Technology (OT). OT domain contains sensors and devices that can be connected to machines or other devices. These devices collect data that can be sent somewhere else. IT domain contains things like servers, databases, and applications. These things handle connectivity between things and compute the data that is given by the OT things. These domains can also be seen in Figure 1. [8]

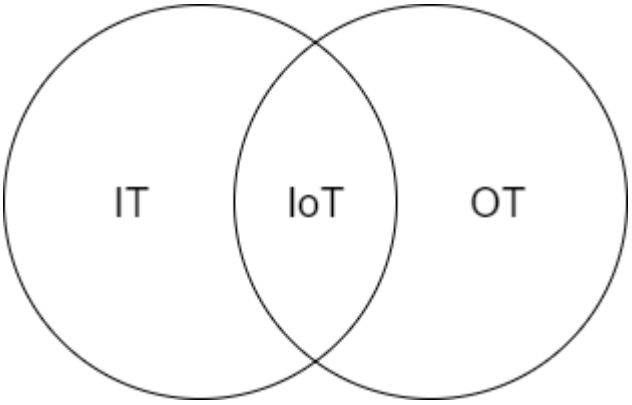


Figure 1 Venn diagram of IoT [8]

For an object to be smart it needs sensors, actuators, memory, communication device, power source, and processing unit. Sensors sense things and make it to data. Actuators are triggered by the data and behave in a way that they are meant to in the physical world. Memory is where the data is stored. Processing unit is for processing and analyzing the data. Communication unit is for communicating with other smart units through a wired or wireless connection. Power source is for giving the object electricity so that it can operate. There are different kinds of trends when it comes to smart objects: decrease in size, increase in processing power, decrease in power consumption, and improved communication capabilities. Decreasing the size makes it easier to include these smart objects in our everyday life. Increased processing power allows making the smart object more complex and connected. A decrease in power consumption makes smart objects last longer without the need of external power or a change of batteries. Improvement in communication capabilities makes the sending data faster and over a wider area with wireless technologies. [8]

2.2 Cloud Computing

Cloud computing is accessing computing resources through a network. This way you can access for example other networks, servers, storages, applications, and services. This is visualized in Figure 2. These resources are made possible through virtualization which has two types: application virtualization and server virtualization. Application virtualization means that an application is hosted on a virtual machine that the users can access through

the network. This way the user doesn't need the computing power, or the data used in an application on his personal device to use this application. In server virtualization, you access a server that can have multiple virtual machines. These virtual machines can have different operating systems and applications that the users have installed. This way many physical machines can be put into one physical machine as virtual machines. [9]

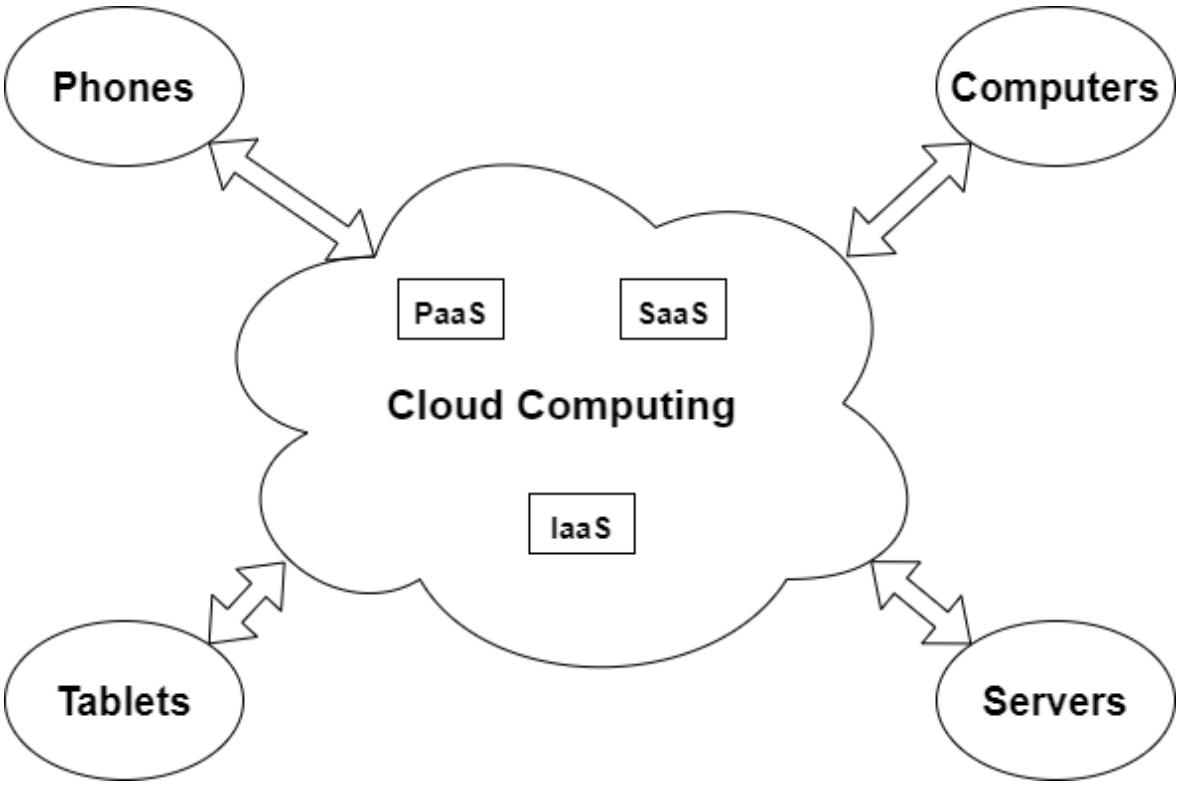


Figure 2 Cloud computing

As Marinescu [10] said in his book Cloud Computing: Theory and Practice: "There are three cloud delivery models: Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS), and Infrastructure-as-a-Service (IaaS)". SaaS offers a cloud application to use by the service provider. This application can be accessed using different clients that utilize the network. In this, the infrastructure of the cloud is defined by the service provider and can't be changed by the user. PaaS offers a cloud infrastructure to host applications. In this, the service provider has defined the programming languages and tools that it supports. Also as in the SaaS, you can't change the cloud infrastructure that the provider offers. IaaS offers a part of computing resources from a physical machine. This means for example processing, storage,

and networks. In this, you also can't change the cloud infrastructure but you can manage the operating system, storage, deployed applications, and some network components. [10]

2.3 Smart city

Smart cities are meant to "function as a device". The technology, that computes and gathers data from day-to-day life, is hidden in the infrastructure. This is possible nowadays through IoT and cloud computing, where the sensors that gather data are inserted in almost every possible thing and then send to the cloud. This way the data can be real-time and accessed almost everywhere. [11]

There are also other definitions of smart cities. McCord et al. [12] defines them as an "urban development project that design computation systems and sensory technology to monitor activity and regulate energy consumption and resource distribution". Nam et al. [2] have many definitions for it and every one of them has one thing in common: to make the city better. Hall et al. [1] say that this can be done by monitoring and integrating the city's critical infrastructures. These are for example transportation, communication, water, and power.

In Figure 3 we can see application areas that smart cities automate or facilitate according to Lim and Maglio [13]. In Table 1 these areas are opened to 5 categories: connection, collection, computation, communications, and co-creation. Connections mean the connection between things and people. Collection means data that has been gathered through things. Computation means using computational processes for decision-making using an algorithm and expert knowledge. Communications means both machine-to-machine communications and machine-to-human communications. Co-creation means value between customers and the providers in the system through activities. [13]

Smart service system	**Connection**	**Collection**	**Computation**	**Communication**	**Co-creation value**
Smart energy	Between customers, providers, other stakeholders, and things	Energy operations data	Optimize energy usage.	Between machines, facilities, etc.	Activities of energy customers, energy providers, and other stakeholders
Smart transportation	Between vehicles, roads, and other related infrastructure	Vehicle operations and health data	Safety and efficiency	Between vehicles, people, etc.	Activities of drivers, riders, and other stakeholders
Smart logistics	Between facilities, vehicles, and goods	Production and logistics data	Optimal operations management	Between facilities, vehicles, people, etc.	Activities of manufacturers, distributors, and other stakeholders
Smart health	Between people, devices, and the health care environment	Health-related data	Diagnosis and prognosis	Within or through technology-equipped people, living, and care environment	Activities of patients, healthy people, healthcare providers, and other stakeholders
Smart farming	Between living properties and farming equipment	Condition and environment data	Optimal health management	Within or through a technology-equipped farm	Activities of farmers, agriculture companies, and other stakeholders
Smart building	Between customers, providers, other	Work-related and building operations data	Comfort and performance optimization	Within or through a technology-equipped building	Activities of building occupants, managers,

	stakeholders, and things				and other stakeholders
Smart home	In-home and home-around connections	Living-related data	Context-awareness	Wireless within or through a technology-equipped house	Activities of residents and related stakeholders
Smart security	Between customers and providers	Property condition and environment data	Real-time surveillance	Real-time communications between stakeholders	Activities of property owners and protectors
Smart hospitality	Between people and the service environment	Stay-related data	Context-awareness	Within or through a technology-equipped hospitality environment	Activities of guests and service providers
Smart education	Between people, devices, and education environments	Study-related data	Maximal learning and satisfaction	Within or Through technology-equipped education devices and environment	Activities of students, teachers, and other stakeholders
Smart city	Between people and organizations	Public purpose data	Optimal administration and living conditions of citizens	Between stakeholders	Activities of citizens, public infrastructures, government agencies, and other stakeholders

Table 1 Characteristics of Smart Service Systems [13]

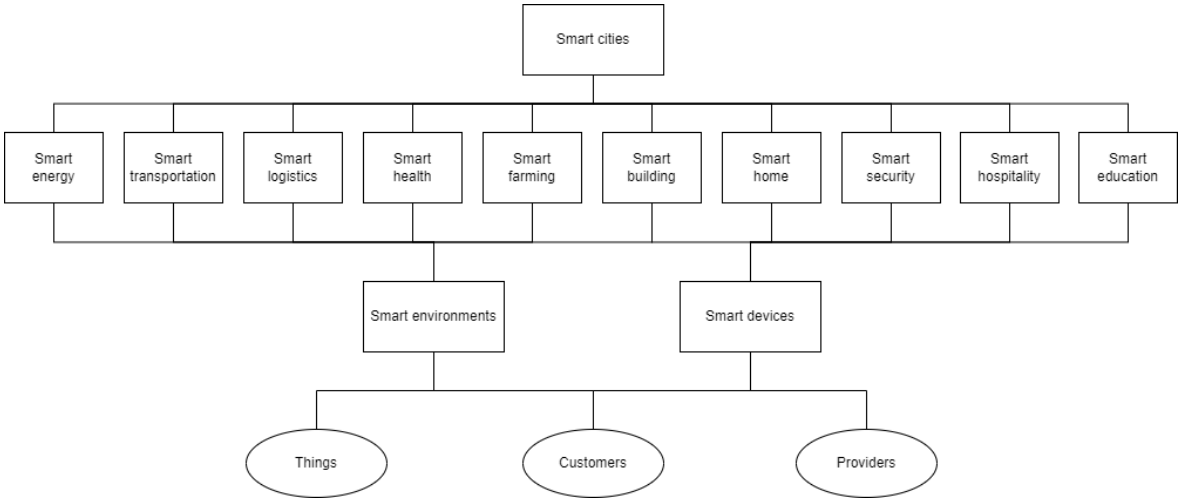


Figure 3 Application areas related to smart cities [13]

Lim et al. [14] researched smart cities using topic modeling in 2018. For data, they used a total of 2856 articles to find keywords from them. They used Latent Dirichlet Allocation for this. keywords that came up were data, system, service, network, urban, technology, sensor, environment, citizen, public, social, mobility, sustainable, life, open, knowledge, policy, integration, decision, and local.

Because a smart city consists of many different aspects of a city, there have been many different types of research on it. Lim et al. [13] address the problem that arises from gathering the big data that is gathered from different infrastructures of the city. These areas can be seen in Figure 3. That research gives reference models to help with it and challenges that arise from transforming that data into something useful information. Badii et al. [15] have a similar kind of problem where they present a web and a mobile app to show this data that has been gathered in an informative way. McCord et al. [12] on the other hand present and review one smart city project, Sidewalk Toronto, which is quoted in the research as to be "district wide energy system". Pradeep et al. [16] present on a more general level what kind of challenges are in developing technology to utilize smart city and offers a solution for this. It also addresses the security and privacy aspects that the data that is being gathered can contain.

2.4 Social Media and Microblogging

Social media contains websites and applications that use social networking. These websites and applications range from text format to video format [17] and allow interaction, networking, and collaboration among users. For brands and individual people, social media can offer different things. For brands it offers platforms to market, connect with customers, and have feedback for example. For individual people, it offers to place to socialize, express yourself, do research, and search for entertainment for example. Nowadays social media users can influence other people on decisions and products. This can be a problem as Alkawaz et al. [18] have studied social media has also fake news. [19]

Microblogging is a form of social media [20]. Microblogs offer interaction with your friends and showing your own opinions, and have increased the number of users using them over the years [21]. In microblogs, you can send a small amount of content [22]. This content can contain for example text, images, and video links [22]. Some of the more popular microblogging sites are Facebook, Sina Weibo, and Twitter [20].

2.5 Twitter

Twitter is a microblogging service where people can share short messages with everybody. Its first publication was on 16th July 2006 and has since grown in popularity. Normally these microblog messages are written by people individually to their profiles but there are also profiles for example companies and political parties whose messages are updated by their teams of communication. Depending on the user, these messages follow period updates. Some update once a day, some by every hour. The topics of these blogs are determined by the users and can vary a lot. Reader count also varies depending on the user and the follower amount that this microblog has. According to Martínez-Cámara et al. Twitter was eight most popular website in 2014. [23]

On Twitter, you can send 280 characters (originally 140 characters) messages, called tweets [24]. These messages can contain links to other websites, images, or videos [23]. These

messages are shown to your followers and others that search your Twitter user. You can also decide who you are sending your messages to if you don't want the world to see them. This can be done on devices that have access to the Internet. Each of these messages has its own URL which means that these messages are their own web pages. In Figure 4 you can see the template of a tweet. [4]

Figure 4 Tweet template [25]

There have also been many types of research that utilize Twitter to get data. Doshi et al. [26] present an application that shows trending hashtags and active users by using real-time Twitter data from Twitter API. This application also utilizes users' coordinates to visualize where the tweets are from on a world map. Kaur et al. [27] use Twitter API to extract tweets from Twitter and get additional information about these tweets. This information could be useful for example marketing companies. Zhang et al. [28] studied over 69 million tweets to show if there is Twitter trend manipulation. This is being done by determining what reason some topics become trending.

2.6 Data Mining

Nowadays different organizations and institutions collect and store data for future use. This data can be in any format, be very large, and have difficult data structures, which makes it hard to handle and use. Data mining is discovering knowledge from data through an iterative

process. It offers a solution to find new information from a large amount of data that can't be handled by humans but by a computer. There is predictive data mining and descriptive data mining. In predictive you use variables and fields in the data to predict the outcome. In descriptive you find patterns from the data that can be understood by humans. [29]

Data mining is a part of knowledge discovery from data (KDD). All the steps that are in KDD are data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, and knowledge presentation. In these steps the data is first cleaned, then combined, if there are multiple data sources, then relevant data is selected and transformed to the correct format, then intelligent methods are applied which means mining, and last, the patterns are evaluated and transformed so that the knowledge can be presented. These steps are visualized in Figure 5. This data that can be mined can be any kind of data. It can be database data, data warehouse data, transactional data, or some other form of data like spatial data or text data. [30]

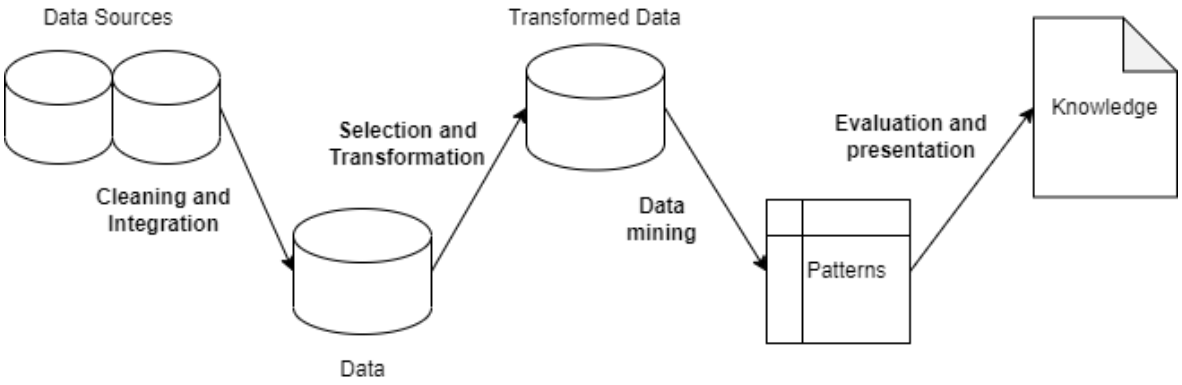


Figure 5 Knowledge discovery from data [30]

2.7 Text Mining

Text mining is a special type of data mining. As said in the previous chapter, data mining means the process of getting knowledge from any kind of data. In this case, because it is text mining, it means that the data is text. This text is usually unstructured data that contains strings that have words in it. This text is also written in natural language and excludes

artificial languages such as source code or mathematical equations. Typical text mining tasks are classification, clustering, and association. [31]

Text mining started in the 1950s when text classification and text clustering came up as a part of pattern recognition. The classification and clustering at that time was book and information classification and clustering which was done based on the topics and contents of texts. New applications were needed when Internet technology started to popularize in the 1980s and 1990s as the text mining field developed and grew. In 1987 first Message Understanding Conference took place where the performance of text mining techniques was evaluated. Nowadays text mining has been done on social media. [32]

Text is a set of words that is written in natural language. These words consist of multiple characters that define the meaning of the word. These words are grouped in sentences and then grouped into paragraphs. These sentences and paragraphs have meaning based on the words and word order that they have. The sentences have rules which are called grammar. Sentences start usually with a capital letter, have white spaces between words, and end with punctuation marks. These punctuation marks are also how the sentences are divided into paragraphs. [31]

In classification, one or more categories are assigned where the data is then categorized. There are two kinds of approaches to this: one where human has made the category rules and one where the machine has made the rules. In the human approach, you make the rules yourself, and then the data is categorized by them. In the machine approach, rules are made by the machine using sample data. In clustering data is processed from one group into smaller subgroups that have similarities among the subgroup items. The most favorable outcome is where the similarities have been maximized among the items in the subgroup and similarities have been minimized between the subgroups. In association rules are defined from the data sets using the if-then form. This means that if something happens then something else happens. This has been used for example to find purchasing trends where if something is bought what else is bought with that item also. [31]

The researches on text mining vary very much. Matsumoto et al. [33] studied data where the data contains numerical and text data. For this, they made a mining framework that would analyze this text and numerical data. Zhong et al. [34] studied how to mine patterns from text documents. For this, they presented an innovative and effective pattern discovery technique. This technique consists of processes of pattern deploying and pattern evolving. Tekiner et al. [35] studied text mining with high-performance computing. This was chosen because there is exponential growth in text data from gigabytes to terabytes.

3 METHODS

This methods section explains the methods that were used to create this research. First, it explains the research method that was used to conduct this research. After that, it explains topic modeling and sentiment analysis that are done on the smart city about the Twitter data. Lastly, it introduces the programming languages that were used to make the tool that then does this topic modeling and sentiment analysis.

This section was made similar way as the literature review to get reliable information about the methods. Same sources were used but with the correct search term, combinations of them, plurals of them, for the methods. These search terms were Descriptive analysis, Quantitative research, Topic modeling, Latent Dirichlet Allocation (LDA), Sentiment analysis, Bing, R, and R shiny.

3.1 Quantitative research

Quantitative research is a research method where you collect and analyze data to test objective theories [36]. This data is structured and can be represented numerically [37]. Through this, it produces accurate and reliable measurements that can be used in statistical analysis [37].

Quantitative research has a hypothesis or research question/s that is being studied through data. This hypothesis or research question/s should be clearly stated and ideally narrowed down to a one-sentence statement of the problem. Quantitative research also has a design on what way the data is being studied. These are Descriptive, Correlational, Experimental, and Quasi-experimental. The design is chosen depending on the aim of the research. [36]

Descriptive research is the design that is used in this study. Descriptive research is restricted to factual registration and it doesn't try to explain why things are this way. There is no need to form a hypothesis or development of theory for descriptive research. Descriptive research tries to be objective or neutral, and it just tries to show how the reality is. The approach to

the descriptive research depends on the study width and study depth. Study width means the data that is being processed and depth means cases that are being studied through this data. [38]

3.2 Topic modeling

Topic modeling is a technique used in text mining and Natural Language Processing (NLP) [39]. Its purpose is to find a collection of words from textual data that is given [40]. It tries to identify patterns and relationships from this data [41]. This group of words is called topic. This way you can find recurring patterns of words from textual data. There are multiple topic modeling methods. These are for example Vector space model, Latent semantic indexing, Probabilistic latent semantic indexing, and Latent Dirichlet allocation (LDA). In Figure 6 we can see the framework of topic modeling. [42]

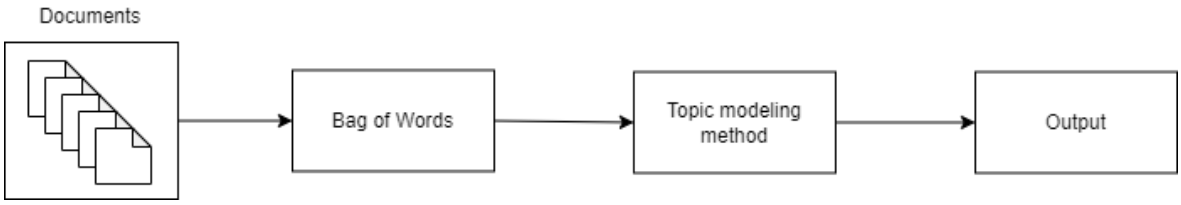


Figure 6 Framework of topic modeling [42]

LDA is an unsupervised topic modeling method [39], [41]. In LDA topics are modeled as a probability distribution over the words that are in the corpus[40], [43]. LDA contains an algorithm that looks for words that have co-occurrences in the documents in the corpus [40]. It also assumes that these words that appear in the same documents belong also to the same topic [40]. For LDA you need to choose the number of topics before you can run it [40]. The word that has the highest probability in the topic usually gives a hint of what the topic is [39].

There are earlier studies where topic modeling was utilized from Twitter's tweets. Dahal et al. [40] try to evaluate public opinion on global climate change through topic modeling and sentiment analysis. They also have geotagged the tweets so they can see how the opinion

changes over the world. Hidayatullah et al. [41] utilize topic modeling for Indonesian Twitter accounts that tweet about football. They then make a word cloud of the topics where you can see what word is most frequent in the topic based on their size. Yu et al. [44] present a new topic model on Twitter data which is called Twitter hierarchical latent Dirichlet allocation. it should mine hierarchical dimensions of tweets' topics which would help in their online analytical processing.

3.3 Sentiment analysis

Sentiment analysis is a category in NLP and text mining [45]. Sentiment analysis can be used to identify the emotional state or opinion from textual data [40]. This textual data is processed to present some kind of topic so that the sentiment analysis gives an opinion on a specific topic [45]. Figure 7 shows the framework for sentiment analysis. There are three approaching types for text in sentiment analysis according to Nann et al. [45]: using a machine learning-based classifier, using relevant n-grams of the text and document with an unsupervised semantic orientated scheme, or using publicly available lexicon dictionaries which provide positive, negative or neutral scores for each word.

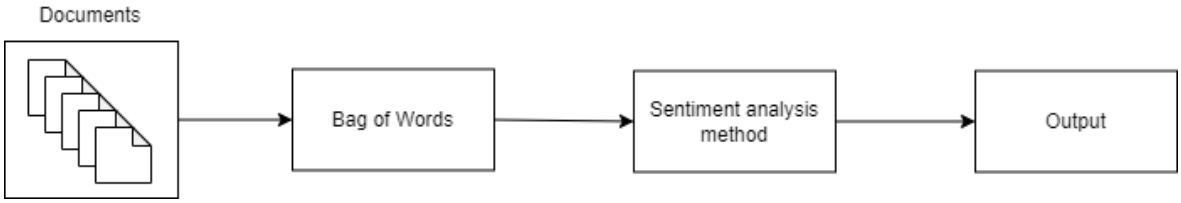


Figure 7 Framework of sentiment analysis

A lexicon dictionary is a set of words. These words are given a sentiment numeric value which shows words' polarity and intensity. More advanced lexicon dictionary methods have also grammar rules to boost the quality of sentiment analysis. Also, this lexicon approach can be used in combination with machine learning methods. An advantage that these lexicon approaches have is that they can be used without any additional information on the sentiment analysis. One limitation that these lexicons have is that they are troublesome to make since the lexicons are usually made through manual labor. [46]

There are studies on sentiment analysis from Twitter data. Rahman et al. [47] present a model that does sentiment analysis from Twitter data. This model utilizes in combination supervised and unsupervised machine learning algorithms. The subjects of this study were McDonald's and KFC to see which is more popular. Wagh and Punde [48] research in their study different types and techniques used in sentiment analysis. For this Twitter was used as data. It shows what different levels of analysis there are in sentiment analysis and the process of sentiment analysis. Prakruthi et al. [49] present an application that does sentiment analysis on live Twitter data. The application searches tweets with user-given hashtags and the number of recent tweets and shows the results as pie charts and histograms.

3.4 Programming language

R is a language and environment available as Free Software and Open Source. R provides an environment where you can make various statistics and graphics from data. It is also very extendable as you can make yourself your extensions. These extensions are called packages. As said on the R documentation about page, it has an effective data handling and storage facility, a suite of operators for calculation on arrays, a large integrated collection of intermediate tools for data analysis, graphical facilities for data analysis, and display either on-screen or on hardcopy, and a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input, and output facilities. [50]

Shiny is a framework that helps you create web applications using R code. It requires zero ability to code HTML, CSS, or JavaScript to make Shiny apps. In Shiny you have general building blocks that build your web app as you want to. Shiny is designed to feel easy so you can concentrate on processing data. [51]

4 ANALYSIS TOOL AND PROCESS

The tool for making topic modeling and sentiment analysis is called Smamta (Sosiaalisen Median Aihe Mallinnus ja Tunne Analyysi translated Social Media Topic Modeling and Sentiment Analysis). This section explains the tool, how it was made, what it can do, and how to utilize it. This section also goes into detail about how the code that was made works and how the results can be read. This section also explains what you need to make this tool work.

4.1 Prerequisite

For this tool to work you need these tweets in the correct format. This is being done using a parser [52] that I modified to parse these folders containing tweets into a .csv file. This parser is made using Python Jupyter Notebook and it utilizes Python 2. Parser goes into every subfolder in the Twitter stream folder and collects every tweet it can from these files and puts them in their own row to the .csv file. The parser also counts how many tweets it has extracted from the stream and how many it has failed. When this is done the tweets from the corresponding folder are in a file that's size is way smaller and easier to handle by the Smamta.

4.2 User interface

The user interface is pretty simple and easy to understand. It consists of input and output areas that utilize RShiny elements. The input area is made with a sidebar panel element and consists of text input element, select input element, and action button element. The output area consists of two text output elements and two plot output elements. Outside these areas is the title panel element which has the tool's name. Figure 8 shows the user interface. In the grey box on the left is the input field and on the right in white is the output field.

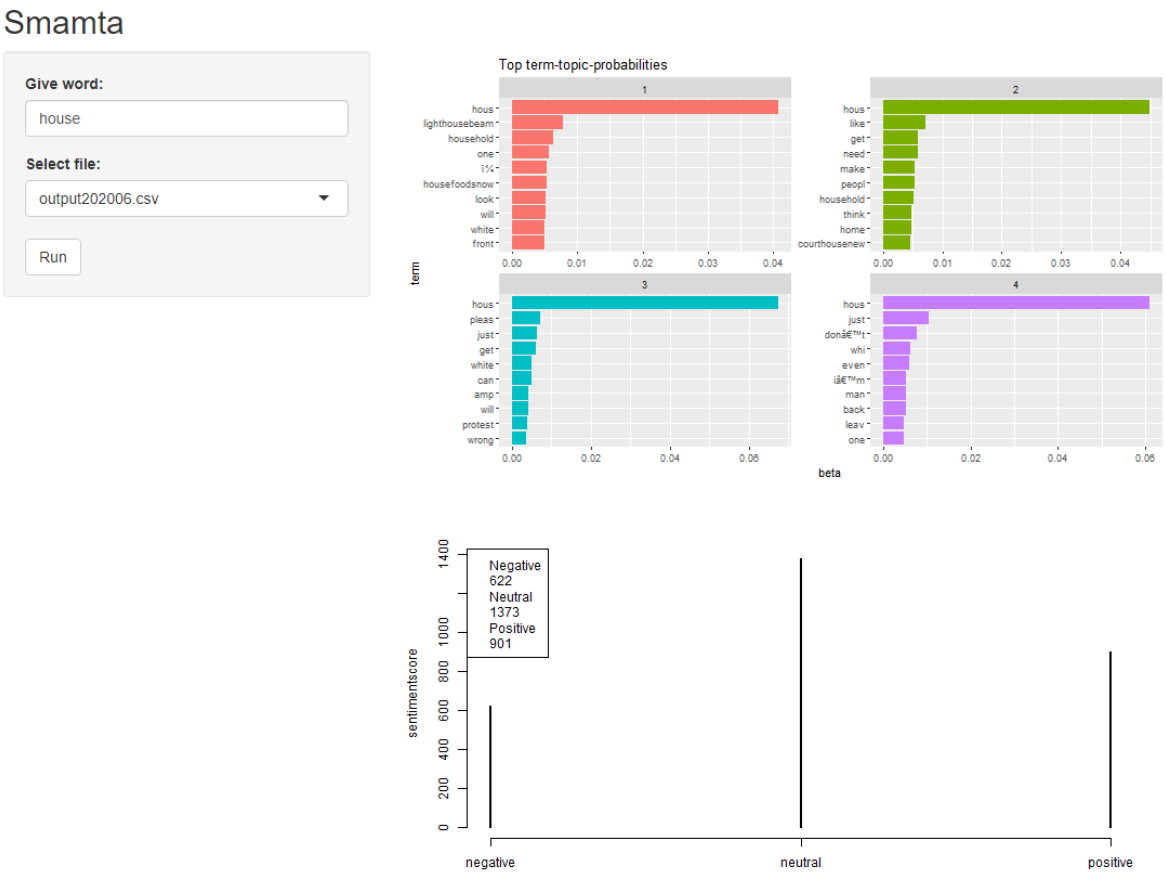


Figure 8 User interface example

Input areas text input element is for inputting a word that user wants to utilize to define the topic modeling and sentiment analysis. The select input element is for you to choose which dataset you want to utilize. It gives the dataset as a slide menu where you can choose the correct one you want. The action button element is for running the topic modeling and sentiment analysis.

Output areas text output elements are for giving the title to plots so users will know which one is topic modeling and which one is sentiment analysis. Plot output elements are for the outputs that the tool gives as it is run. One output is for the topic modeling and the other is for sentiment analysis so that they both can be shown at the same time.

4.3 Workflow

In Figure 9 you can see the workflow of Smamta in activity diagram form. Work starts when you have the preprocessed file that was described in section 4.1 so that you can use the tool. Now you have a user interface in front of you and you can put inputs and run the tool. After this tool checks if the inputs are correct. If they are not correct, the tool gives you an error message and you can put new inputs. After correct inputs user, the work that is being done is not shown to the user until the final output. What's happening in the background is that first the preprocessed file is being parsed, using the input that the user has put, into a corpus. After that, the corpus is cleaned from spaces, punctuations, numbers, stopwords (articles and conjunctions for example), transformed into lowercase, and stemmed. After this, the document is transformed into a document term matrix which is needed for topic modeling and sentiment analysis. In the topic modeling side, variable K is first searched. This variable determines how many topics there are in the corpus. After this topic modeling is done using this K value and LDA. Basically at the same time sentiment analysis is being done using an existing dictionary to determine if the words are negative, positive, or neutral. After these, the output is shown to the user where, in topic modeling, there are as many topics as the K was determined, and the 10 most common words in that topic are shown, and in sentiment analysis, the amount of each sentiment is put in a bar diagram and is shown in numeric value. These results are also put into a .csv files so that they can be used outside this tool.

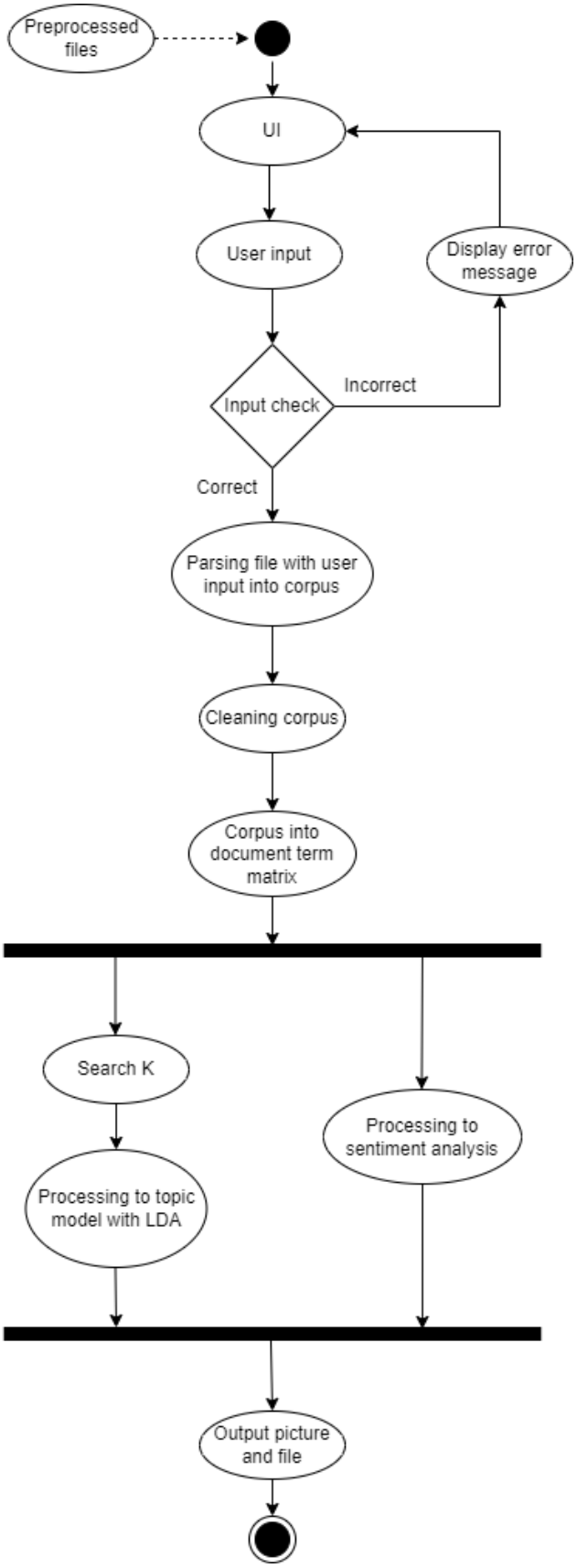


Figure 9 Smamta activity diagram

5 RESEARCH RESULTS

This section contains research results that the tool gave from the Twitter data of 2020. The search words that were used were different forms of singular and plural of the word Smart City. These words were smart city, smart cities, smartcity, smartcities, Smart city, Smart cities, Smartcity, Smartcities, Smart City, Smart Cities, SmartCity, SmartCities, smart City, smart Cities, smartCity and smartCities. The data were distributed over 4 quarters. Each quarter has 3 months of worth of data from the year 2020, which means that quarter 1 has January, February, and March, quarter 2 has April, May, and June, quarter 3 has July, August, and September, and quarter 4 has October, November, and December.

Topic models have been shown in as a table, where each word of the topic is shown, and at the bottom is the topic that these words represent as can be seen in Table 2, 3, 4, and 5. Sentiment analysis is shown as a clustered column chart where the results are shown in quantity and percent form. The results were not exactly what was expected as there were so few tweets that were found with these search words. As there were very few tweets, I also checked the tweets to help determine the topics. The unedited results can be seen in Appendix 1.

5.1 Quarter 1

In Table 2 you can see the first quarter of the year 2020 where the tool found 166 tweets as can be seen in Table 6. There are many unknown topics as there were tweets that contained only hashtags and/or people's names that the topic model took as a topic. Topic words 3 had the first topic that contained a smart tree. Searching from the tweets I found that the smart tree was a tree that could have Wi-Fi and other smart features. Topic words 5 contained development and budget as topics and topic words 8 transportation and environment. In Figure 3 topic words 3 would go to smart environment category, topic words 5 would go to smart city category, and topic words 8 would go to smart transportation category and smart environment category.

Topic words 1	Topic words 2	Topic words 3	Topic words 4	Topic words 5	Topic words 6	Topic words 7	Topic words 8	Topic words 9
citi	veri	enricomolinari	smartciti	smart	hmoindia	anilvijminist	smartcitylab	authackeray
smart	amrutcitygzb	smartciti	ces	citi	httpstcofwrxtlmrsp	ashook	climat	kamlesh
bbmpcomm	cmofficeup	amp	long	develop	ltgovdelhi	bjpharyana	ðŸš¶â€• â™€ï¸□ðŸš´ðŸ›µðŸ›´ðŸšŒðŸš‹ðŸš—	mieknathshind
bbmpmayor	myogiadityanath	charg	aussi	smartciti	mohuaindia	girish	infograph	narendramodi
blrsmartciti	myogioffic	dubai	ciudad	solut	pleas	httpstcoquswqok	plan	thanecitypolic
cmofkarnataka	pradeep	easeofliv	dâ€™eyfranceâ€¦	amp	pmoindia	mcfaridabadâ€¦	smartciti	tmcatweetaway
httpstcogmpgbbugr	smartcitieshua	finserv	digital	challeng	reâ€¦	mcfbd	smartcityexpo	tmcsmartciti
karnatakavartheâ€¦	swachhbharat	jmcjabalpur	doit	urbanmobl	smartcitieshua	mcffaridabad	â€œachiev	uddhavthackeray
mnreddiip	swachhbharatgov	jodhpur	dorotheebell	leader	smartcitiesind	smartcityfbd	ahok	citi
myadhunandan	upgovt	mohuaindia	effectu	ampâ€¦	swachsurvekshan	citi	aiâ€•	smart
vkuchhangi		mygovindia	ell	best	tweetndmc		aia	
		new	gestion	bridg			aid	
		renew	hubciti	budget			apwa	
		smart	lâ€™heur	deliv			cio	
		smartcitytumkur	las	design			citi	
		smartmangaluru	mue	essenti			commit	
		smartphon	pour	forumâ€™			commut	
		tree	publiqu	gap			cto	
		via	technolog	help			govtech	
		wef		hitachi			httpâ€¦	
		wifi		httpstcomztbllxkcf			httpstcovnwbrpek	
				httpstcozsezafi			lateshift	
				iingwen			nyc	
				innov			pilot	
				learn			smart	
				map			smartcitiesd	
				oper			smartcityc	
				peopl			via	
				polici			worker	
				rohanrgupta				
				taiwan				
				technolog				
				tool				
				zero				
Unknow	Unknow	Smart tree	Unknown	Development and budget	Unknown	Unknown	Transportation and environment	Unknown

Table 2 Quarter 1 topics

There were mostly neutral tweets as many of the tweets contained only hashtags and/or people names which don’t have any sentiment value. Other than that, the tweets that got a value were mostly positive regarding smart cities.

5.2 Quarter 2

There were fewer tweets about smart cities in quarter 2 as can be seen in Table 6. This also affected the topics as there were 5 topics whereas the first quarter had 9 as you can see from Tables 2 and 3. Some topics didn’t make sense, same as in quarter 1. Topic words 2 had a topic of health and topic words 3 had a topic of IoT. In Figure 3 topic words 2 would go into smart health category and topic words 3 would go to any of the categories as usually every one of them has IoT.

Same as in quarter 2, there were mostly neutral tweets for the same reasons as in quarter 1. Some tweets had some writing that was not in the Latin alphabet which can be seen in topic words 1 as there are just some meaningless character strings. Other than that there were more positive tweets than negative but compared to the first quarter positive tweets had a bigger decrease than negative tweets.

Topic words 1	Topic words 2	Topic words 3	Topic words 4	Topic words 5
smartcityrajkot	smart	smartciti	ajuntamentvlc	agrapolic
collectorrjt	citi	live	cevcv	chiefsecyup
bouddhk	dholerasmartc	infograph	confecomerccv	cmofficeup
cmoguj	aisitaisidemo	internet	fehvhosteleria	hardeepspuri
jayeshssd	class	thing	smartcityvlci	mohuaindia
àª•àª®à«‡àª¨à«□àªŸ	doctor	citi	taxivah	officeofdmagra
àª•àª°àªµà«€	hpareshaan	â€˜â€™smart	valenciactiva	pibiâ€¦
àªŸà«□àªµàª¿àªŸ	httpstcojtvmpihuh	â€œinternet	vitemprend	pmoindia
àª®àª¾¼àª,	roflgandhi	alykhansatchu	vlcshoppinâ€¦	prabhun
digic	world	borisjohnson	vlctechciti	sakshisur
		incubam		smartcityagra
		join		
		model		
		pullul		
		sensor		
		spoke		
		thinâ€¦		
		togeth		
		will		
Unknown	Health	IoT	Unknown	Unknown

Table 3 Quarter 2 topics

5.3 Quarter 3

Quarter 3 also had fewer tweets than quarters 1 and 2, as can be seen in Table 6, but had the same quantity of topics that quarter 2 had. Topics words 1 had IoT as a topic, topic words 2 had infrastructure as a topic, topic words 3 had transportation as a topic, and topic words 4 had disturbance as a topic. There was also one topic that didn't make sense as you can see from Table 4. In Figure 3 topic words 1 could go to any of the categories, topic words 2 would go to smart cities, smart environment, smart transportation, and smart building category, topic words 3 would go to smart transportation, and topic words 4 could go to any of the categories as the topic hint towards what the smart cities can cause.

Topic words 1	Topic words 2	Topic words 3	Topic words 4	Topic words 5
citi	bike	citi	citi	à«§à««à«¦à«®à«¨à«¦
smart	csclоffici	smart	smart	cmoguj
ndtv	dedic	technolog	miâ€¦	cmohri
affirm	httpstcosrcnbwao	antgrasso	httweet	lrd
aux	import	around	jayashreenandi	lrdmale
banayeg	itdpindia	better	just	mayurdesai
banwaya	lane	chang	photo	mcffaridabad
data	mohuaindia	citizen	post	mix
dholera	roadsâ€¦	creat	smartcityoman	rahulkmaurya
dont	smartcitieshua	edg	ðŸ˜·ðŸ™Šð	sewer
driftkhi	timesofindia	efficienâ€¦	acharyasarmila	smartcityfbad
ellemÃªm		environ	airdrop	smartcityrajkot
enabl		even	disruptor	soil
face		experâ€¦	great	vijayrupanibjp
filmciti		help	httpstcofetgyrry	àª›à«‡àª¤à«‹
foriegn		increas	httpstcoirffbjht	àª¥àª¾àª¯
god		ioten	httpstcolckrqvjux	àªàà«• àª°àª¾
havâ€¦		lidar	httpstcovobmutsj	àªàª°àª¤à«€
httpstcohdnlpnlm		live	iotâ€¦	àª®àª¨à«‡
les		need	miss	àª°à«‹àªœ
linnov		roadsafeti	next	àªµàª°à«• àª·
lonu		smartciti	pcr	àªµàª¿àª¶à«• àªµàª¾àª‚â€¦
modi		smarter	saharanpur	
particuliÃ¨r		speed	smartciti	
partner		techhq	srivatsayb	
power		technâ€¦	à¤…à¤°à¥□à¤•à¥‹	
problÃ¨meâ€¦		transport	à¤–à¤°à¥□à¤š	
que		use	à¤¦à¥‡à¤–à¤¿à¤¨à¥□à¤›	
readi		way	à¤¨à¤¾à¤—à¤°à¤¿à¤•	
sensor		world	à¤¨à¤¿à¤¶à¥□à¤šà¤¿à¤¤à¤¿	
smartciti			à¤¬à¤šà¤¾à¤• à¤°	

stakehold			à¤¬à¤¨à¤¾à¤‰à¤¨à¤¨à¥‡	
sustain			à¤¬à¤¾à¤Ÿ	
technolog			à¤¬à¥• à¤¯à¤¾à¤ªà¤¾à¤°à¥€à¤•	
thank			à¤®à¤¹à¤²	
trust			à¤¹à¤¾à¤¡à¤–à¥‹à¤°	
will				
yogi				
à¸• à¸²à¸£à¸—à¸µà¹†à¸ªà¸²à¸¡à¸²à¸£à¸–à¹€à¸”à¸´à¸™à¸—à¸²à¸‡à¹€à¸¥à¸·à¸à¸¢à¸™à¹„à¸Ÿà¸Ÿà¹‰à¸²à¸«à¸£à¸·à¸à¸£à¸–à¸¢à¸™à¸•à¹Œà¸ªà¹ˆà¸§à¸™à¸•à¸±à¸§				
à¸□à¸²à¸£à¸¢à¹€à¸›à¹‡à¸™				
à¹ƒà¸”à¹†				
à¹□à¸•à¹^à¹€à¸¡à¸·à¸‡à¸‡à¸²à¸™à¸-à¸±à¸^à¸‰à¸£à¸´à¸¢à¸°â€‹à¸—à¸µà¹^à¸”à¸µà¸„à¸·à¸à¸„à¸§à¸²à¸¡à¸„à¸´à¸”à¹€à¸«à¹‡à¸™				
à¹„à¸¡à¹^à¸^à¸³à¹€à¸›à¹‡à¸™à¹€à¸™à¸•à¹‰à¸-à¸‡à¸¡à¸£à¸°à¸šà¸š				
à¸«à¸£à¸·à¸•à¸±à¸§à¸§à¸±à¸§à¸š				
IoT	Infrastructure	Transportation	Disturbance	Unknown

Table 4 Quarter 3 topics

The distribution of positive, neutral, and negative tweets changed a little bit from the first two quarters. This quarter had fewer neutral tweets even though there were some same issues as in the first 2 quarters. There were also mostly positive tweets as there were no negative tweets.

5.4 Quarter 4

Quarter 4 has the least quantity of tweets compared to other quarters as can be seen in Table 6. It was also the first quarter that didn't have any unknown topics as can be seen in Table 5. Topic words 1 had drones and government, topic words 2 had transportation and smart city industry, topic words 3 had sustainability and intelligent transportation, and topic words 4 had training cyber professionals as a topic. In Figure 3 topic words 1 government would go to smart city category and drones could go to multiples like smart logistics, smart building, and smart farming as it could be used in all of those. Topic words 2 transportation would go to smart transportation and smart city industry would go to smart energy, smart

logistics, and smart farming. Topic words 3 sustainability would go to smart environment, and intelligent transportation to smart transportation. Topic words 4 would go to smart education.

Topic words 1	Topic words 2	Topic words 3	Topic words 4
citi	citi	smart	smartciti
smart	smart	citi	citi
just	can	latest	smart
post	will	sustain	cyber
aargelich	adam	system	daili
approach	amp	across	forb
award	analys	banayeng	gen
bareilli	architectur	citiesâ€¦	httpstcoomvrqzrlz
build	area	citiesðŸ“¶ðŸš¦	httpstcouonsawsxk
categori	arrog	countri	httpstcoxnytkjzen
cdwgwagov	asseco	dailyâ€¦	latest
coâ€¦	bullock	deliv	lifestyl
code	camâ€¦	develop	millenni
complet	car	dikhan	much
contributor	chang	especializado	securearm
cool	cityâ€¦	film	tecmundo
dataop	companii	futuro	thank
daysofcod	connect	good	train
devop	control	govern	via
discuss	data	httpstcodfptmj	will
drone	dell	httpstcoppkqoppz	worker
egovern	dumitrascuedi	httpstcosjoqclvmon	
emerg	electr	hybrid	
finalist	emiss	idstch	
homorodean	fenc	integr	
httpstcokztcmomkc	green	intellig	
httpstcoxyhjdrdakr	hull	itn	
httpstcozfvycsd	humil	japantim	
ignor	idl	juml	
import	industria	kyo	
join	interes	las	
live	jennison	madhya	
manag	join	mission	
movement	lead	nasirkha	
mrdaniel	leverag	need	
part	manag	newindia	

photo	microsoft	one	
singl	must	part	
smartcâ€¦	natur	pradesh	
softwar	nod	present	
strateg	optim	realis	
tech	orang	rescu	
technolog	park	sapna	
temsictexpo	precum	sitio	
testb	princip	smartcityâ€¦	
use	realiti	smarter	
video	rob	tecnologÃa	
visakhapatnam	sau	thehackersmeetup	
vital	shanti	thmcommun	
world	smâ€¦	thmindia	
à¨¸à©• à¨²à¨¤à¨¾à¨¨à¨ªà©• à¨°	smartciti	thmnov	
à¨²à©‹à¨§à©€	stpiindia	thmwebinar	
	sunt	transport	
	surveil	trend	
	sustainablefutur	uasugv	
	sustaingram	unman	
	telekom	urbana	
	toward	use	
	town	vision	
	veri	wale	
	within	world	
Drones and government	Transportation and smart city industry	Sustainability and intelligent transportation	Training cyber professionals

Table 5 Quarter 4 topics

The distribution of tweets sentiments were similar to in quarter 3. There were fewer neutral tweets as this quarter had the most meaningful tweets. It also had more positive tweets as there were no negative tweets.

5.5 Overview of the results

Overall there were not many tweets found from the data sets. There are reasons for this that I didn't see in the beginning when I tested these data sets. I tested these data sets with more

common words and not so specific as a smart city. These more common words game me more tweets and more meaningful results. For example, I used the word dog which gave results of different kinds of dog-related topics and even political topics because there is a Turkish politician named Erdogan who has a dog in his name. Other testing words that I used were house and horse which also gave more meaningful answers. Because there were so few tweets the topic models looked a little bit different than in testing. In testing, there were usually about 10 most common words of the topic, but as you can see in Table 2, 3, 4, and 5 there are topics that have more than 10 most common words. This happens because there are so many same amounts of words that the tool doesn't choose but shows all of them. This is because there is about one tweet or retweets of that same tweet that the same words come up in the same amount.

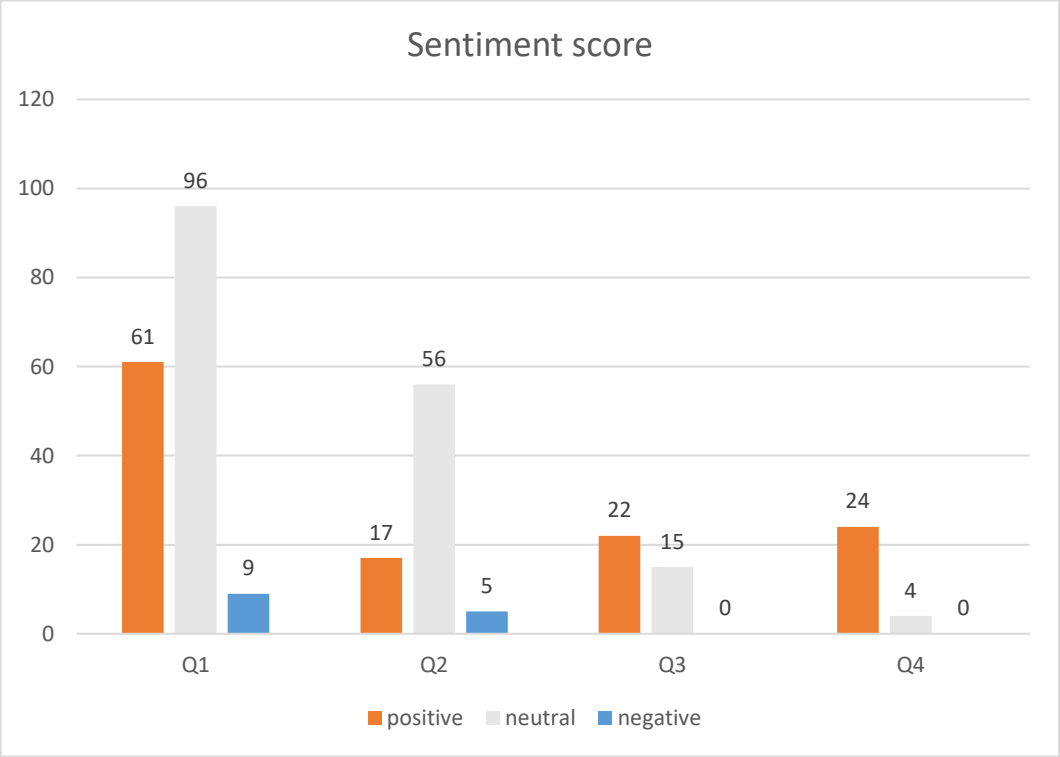


Figure 10 2020 Sentiment score

Another reason that there weren't so many answers could be that these data sets were "Spritzer" version of Twitter grabs [6]. This is said to be the most light and shallow version [6]. This can be seen in Table 6 where there are about five to six million tweets every quarter. Gabriel Stricker [53] says that there are over 500 million tweets sent each day. If I have five

to six million tweets from three months, it seems so little compared to how many are sent in a day.

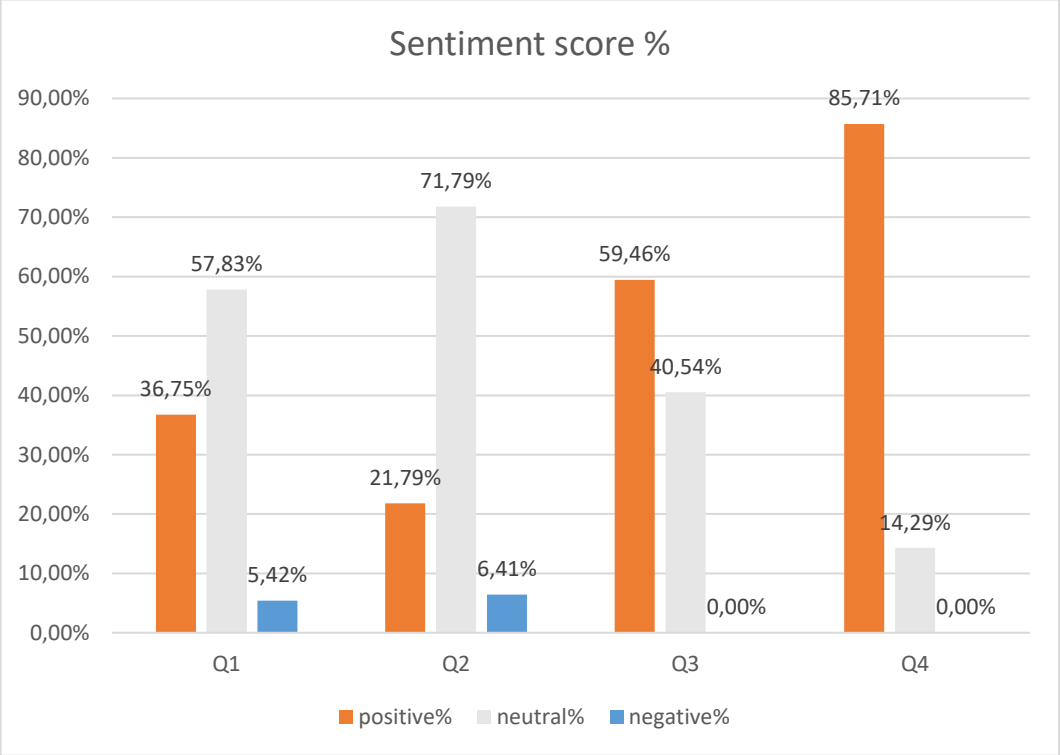


Figure 11 2020 Sentiment score as a percent

One reason that the results were not so meaningful was also the language. I only used English words and the Latin alphabet. On Twitter, you can use any language that you want and any alphabet you want. This means that these data sets can have various languages and alphabets. This was seen in some results as there was a used the word smart city in a tweet but the rest of the tweet was in a different language. Also for example in Table 4 topic words 4 there are some strings that are characters that mean nothing. This happens if the word is not in the Latin alphabet or has some emojis as my parser could not translate them to meaningful Latin alphabet words.

Tweets/Quarters	Q1	Q2	Q3	Q4
Total Tweets	5309485	5544064	5772039	5360555
Found Tweets	166	78	37	28

Table 6 Quantity of total tweets and found tweets

The last reason that some of the topics didn't have a meaning was the tweets themself. Some tweets only contained hashtags and/or people's Twitter nicknames. These results became because these hashtags and/or people's Twitter nicknames had the words that were searched in them. This means that someone has just tagged some hashtags and nicknames to a tweet that itself doesn't maybe have a meaning in text form. These tweets can have meaning as they can contain links, pictures, and/or videos that are not seen in text format.

Sentiment analysis gave a different kind of insight on found tweets. As I didn't take any other results than the end results, I can't see what sentiment weights it gave to each tweet. But I speculate that most of the neutral tweets didn't have a meaning in the first place which result in a neutral weight to the tweet. These tweets would contain non-Latin alphabets or hashtags only for example. As for the tweets that had a positive or negative weight there could be also these tweets that didn't have a meaning but had a word that has a positive or negative weight that would push it outside of the neutral. Although this could be the case it can be seen from Figure 10 and 11 that there is mostly a positive or neutral view of smart cities in these data sets.

6 DISCUSSION

From the previous section we can see that we got some topics out of the data sets that would go to the categories. Categories that rose from the datasets were smart environment, smart city, smart transportation, smart health, smart building, smart logistics, smart farming, smart energy, and smart education. In each quarter there were more positive tweets than negative tweets. There were also many neutral tweets in quarter 1, 2, and 3.

Comparing these results to earlier research results, I used Lim et al. [14] results that were presented in section 2.3. Some of the results in topic modeling were the same as what they got from the articles in smart city literature. The same topics were sustainable and environment. Some that were close to Lim et al. [14] keywords were IoT as this contains sensors and transportation as this contains mobility.

The results were not what I expected. There were so few tweets in each quarter. This made that the results in Tables 2, 3, 4 and 5 didn't always show 10 most used words in topics as there were multiple words that had the same quantity. Some quarter also had many neutral tweets which could be because there were tweets that had nicknames, hashtags, and/or not Latin alphabets that cause tweets to have neutral sentiment value.

Data sets affect the generalization of the results. As the data set is light and shallow, it doesn't give a full view of Twitter topics and sentiments from these quarters. In this research we got around five to six millions tweets per each quarter when there is over 500 million tweet per day as said in section 5.5.

For follow-up research that could be done is to use better data sets for the research that give more meaningful data. This could be done using Twitter API. With this, you could acquire tweets on Smart Cities straight and would not need to parse the data set in the tool. The basic access to this would give you 500000 tweets per month with essential access which would be more than the data sets that I used gave [54]. If you could get academic research access you could acquire 10 million tweets a month [54].

Another one is to use a different source than Twitter. This could mean a similar kind of platforms like Facebook or a completely different one. This could be done because the tool only requires that the data is in .csv format. If a different source is used, a different parser would need to be used to parse the data into .csv format if needed. This would mean that this tool could be used for other topics than Smart City.

The last one would be that you could enhance the tool. Right now the interactive part of the tool is quite bare and doesn't have many options for the user. This could be enhanced like that you could give a different kinds of parameters if you would like to specify the data. Also, the output could be enhanced, for example, to look a little bit more pretty as the sentiment analysis is quite rough. Also, I needed to modify the output in Microsoft Excel to make them look like they know do, so if the tool could give the output like that in the end, it would save a lot of time.

7 CONCLUSION

The research questions for this were what kind of topics and sentiments rise from the data sets. Some topics didn't make sense but some topics made sense. Some of the topics are made into one category like health and others into many like IoT. The most common category for these topics was transportation as four topics contained this. Most of the tweets in each quarter were mostly positive or neutral. The number of negative tweets in the data sets about smart cities in every quarter was always the lowest. However, these results may not be so reliable, as there were so few tweets that were found from the data sets.

For professional data analysts, this would mean not using the data sets the way that I used. This means that these data sets are not good if you are searching answers for this specific topic. These data sets could be used for more common topics like what were used when testing these data sets for the first time. These results also showed that the tool could be evolved to show a more refined output.

As said in section 6 there were some similarities with the results that Lim et al. [14] got but there were also differences. Others that didn't feature in Lim et al. [14] research were smart tree, development, budget, health, infrastructure, disturbance, drones, government, smart city industry, and training cyber professionals. Biggest difference between Lim et al. [14] research and mine was that their articles for data set and I used Twitter.

This work could be continued in many ways. You could acquire different data sets that could show more ideal data. You can also develop the tool further to give a different kinds of outputs. Lastly, you could use a different source or even a different topic as a base to search topics and sentiments of it.

REFERENCES

[1] R. Hall, B. Bowerman, J. Braverman, J. Taylor, H. Todosow, and U. Wimmersperg, "The vision of a smart city," *2nd Int. Life* ., Jan. 2000.

[2] T. Nam and T. Pardo, "Conceptualizing smart city with dimensions of technology, people, and institutions," Jun. 2011, pp. 282–291. doi: 10.1145/2037556.2037602.

[3] F. & Sullivan, "Frost & Sullivan: Global Smart Cities market to reach US$1.56 trillion by 2020." https://www.prnewswire.com/news-releases/frost--sullivan-global-smart-cities-market-to-reach-us156-trillion-by-2020-300001531.html (accessed Mar. 11, 2021).

[4] L. Fitton, M. Gruen, and L. Poston, *Twitter For Dummies*, 1. Aufl. Hoboken: For Dummies, Wiley, John Wiley & Sons, Incorporated, Wiley Pub, 2009.

[5] "Archive Team," Dec. 03, 2020. https://wiki.archiveteam.org/ (accessed Mar. 10, 2021).

[6] "Archive Team: The Twitter Stream Grab." https://archive.org/details/twitterstream (accessed Mar. 10, 2021).

[7] K. Ashton, "That 'Internet of Things' Thing | RFID JOURNAL," *That "Internet of Things" Thing*, Jun. 22, 2009. https://www.rfidjournal.com/that-internet-of-things-thing (accessed Feb. 01, 2022).

[8] M. Alam, K. A. Shakil, and S. Khan, Eds., *Internet of Things (IoT): Concepts and Applications*. Cham: Springer International Publishing, 2020. doi: 10.1007/978-3-030-37468-6.

[9] Nayan B. Ruparelia, *Cloud Computing*. Cambridge, Massachusetts: The MIT Press, 2016. Accessed: Feb. 16, 2022. [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=e000xww&AN=1238001&site=ehost-live

[10] Dan C. Marinescu, *Cloud Computing : Theory and Practice*, vol. 1st ed. Boston: Morgan Kaufmann, 2013. Accessed: Feb. 16, 2022. [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=e000xww&AN=486501&site=ehost-live

[11] R. Mehmood, S. See, I. Katib, and I. Chlamtac, *Smart Infrastructure and Applications: Foundations for Smarter Cities and Societies*. Cham: Springer International Publishing AG, Springer, 2019.

[12] C. Mccord and C. Becker, *Sidewalk and Toronto: Critical Systems Heuristics and the Smart City*. 2019. [Online]. Available: https://explore.openaire.eu/search/publication?articleId=od_____18::ce4729010d6af33d5749a7babec2a0e2

[13] C. Lim and P. P. Maglio, "Data-Driven Understanding of Smart Service Systems Through Text Mining," *Service Science*, vol. 10, no. 2, pp. 154–180, Jun. 2018, doi: 10.1287/serv.2018.0208.

[14] C. Lim, K.-J. Kim, and P. P. Maglio, "Smart cities with big data: Reference models, challenges, and considerations," *Cities*, vol. 82, pp. 86–99, 2018, doi: https://doi.org/10.1016/j.cities.2018.04.011.

[15] C. Badii, P. Bellini, P. Nesi, and M. Paolucci, "A smart city development kit for designing Web and mobile Apps," Aug. 2017, pp. 1–8. doi: 10.1109/UIC-ATC.2017.8397569.

[16] P. P. T and S. K. L, "Smart City Services - Challenges and Approach," Feb. 2019, pp. 553–558. doi: 10.1109/COMITCon.2019.8862243.

[17] G. F. Hurlburt, "Web 2.0 Social Media: A Commercialization Conundrum," *IT Professional*, vol. 14, no. 6, pp. 6–8, Nov. 2012, doi: 10.1109/MITP.2012.115.

[18] M. H. Alkawaz, S. A. Khan, and M. I. Abdullah, "Plight of Social Media Users: The Problem of Fake News on Social Media," in *2021 IEEE 11th IEEE Symposium on Computer Applications Industrial Electronics (ISCAIE)*, Apr. 2021, pp. 289–293. doi: 10.1109/ISCAIE51753.2021.9431841.

[19] D. E. Vlad, *Concepts of Quality Connected to Social Media and Emotions*. Wiesbaden: Springer Fachmedien Wiesbaden, 2020. doi: 10.1007/978-3-658-28867-9.

[20] B. Liu *et al.*, "Data Acquisition, Hot Issues and System of Microblog Mining," in *2015 International Conference on Network and Information Systems for Computers*, Jan. 2015, pp. 116–119. doi: 10.1109/ICNISC.2015.89.

[21] Y. Lu and J. Chen, "Public Opinion Analysis of Microblog Content," in *2014 International Conference on Information Science Applications (ICISA)*, May 2014, pp. 1–5. doi: 10.1109/ICISA.2014.6847451.

[22] A. Kaplan and M. Haenlein, “The early bird catches the news: Nine things you should know about micro-blogging,” *Business Horizons*, vol. 54, pp. 105–113, maaliskuu 2011, doi: 10.1016/j.bushor.2010.09.004.

[23] E. MARTÍNEZ-CÁMARA, M. T. MARTÍN-VALDIVIA, L. A. UREÑA-LÓPEZ, and A. R. MONTEJO-RÁEZ, “Sentiment analysis in Twitter,” *Natural language engineering*, vol. 20, no. 1, pp. 1–28, 2014, doi: 10.1017/S1351324912000332.

[24] “Counting characters.” https://developer.twitter.com/en/docs/counting-characters (accessed Apr. 21, 2021).

[25] “default-tweet-3.png (1200×501).” https://www.tweetgen.com/c/default-tweet-3.png (accessed Feb. 10, 2022).

[26] Z. Doshi, S. Nadkarni, K. Ajmera, and N. Shah, “TweerAnalyzer: Twitter Trend Detection and Visualization,” in *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, Aug. 2017, pp. 1–6. doi: 10.1109/ICCUBEA.2017.8463951.

[27] H. Kaur, M. Talluri, and J. S. He, “Get Twitter information: A collaborative Android application for big data analysis,” in *2015 International Conference on Collaboration Technologies and Systems (CTS)*, Jun. 2015, pp. 483–484. doi: 10.1109/CTS.2015.7210475.

[28] Y. Zhang, X. Ruan, H. Wang, H. Wang, and S. He, “Twitter Trends Manipulation: A First Look Inside the Security of Twitter Trending,” *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 1, pp. 144–156, Jan. 2017, doi: 10.1109/TIFS.2016.2604226.

[29] M. Kantardzic, *Data mining: concepts, models, methods, and algorithms*. Hoboken, NJ: Wiley-Interscience, 2003.

[30] Jiawei Han, Jian Pei, and Micheline Kamber, *Data Mining: Concepts and Techniques*, vol. 3rd ed. Burlington, MA: Morgan Kaufmann, 2011. Accessed: Feb. 03, 2022. [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=e000xww&AN=377411&site=ehost-live

[31] T. Jo, “Text Mining Concepts, Implementation, and Big Data Challenge.” Springer International Publishing, Cham, 2019. doi: 10.1007/978-3-319-91815-0.

[32] C. Zong, R. Xia, and J. Zhang, *Text Data Mining*. Singapore: Springer Singapore, 2021. doi: 10.1007/978-981-16-0100-2.

[33] T. Matsumoto, W. Sunayama, Y. Hatanaka, and K. Ogohara, "Data Analysis Support by Combining Data Mining and Text Mining," in *2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, Jul. 2017, pp. 313–318. doi: 10.1109/IIAI-AAI.2017.165.

[34] N. Zhong, Y. Li, and S.-T. Wu, "Effective Pattern Discovery for Text Mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 1, pp. 30–44, Jan. 2012, doi: 10.1109/TKDE.2010.211.

[35] F. Tekiner, Y. Tsuruoka, J. Tsujii, and S. Ananiadou, "Highly scalable Text Mining - parallel tagging application," in *2009 Fifth International Conference on Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control*, Sep. 2009, pp. 1–4. doi: 10.1109/ICSCCW.2009.5379432.

[36] R. Ingham-Broomfield, "A nurses' guide to Quantitative research," *Australian journal of advanced nursing*, vol. 32, no. 2, pp. 32–38, 2014.

[37] M. J. Goertzen, "Applying Quantitative Methods to E-book Collections," *Library technology reports*, vol. 53, no. 4, pp. 1-, 2017.

[38] T. M. de Jong and D. J. M. van der Voordt, "Ways to study and research urban, architectural, and technical design." DUP Science, Delft, The Netherlands, 2002.

[39] H. Jelodar *et al.*, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimedia tools and applications*, vol. 78, no. 11, pp. 15169–15211, 2019, doi: 10.1007/s11042-018-6894-4.

[40] B. Dahal, S. A. P. Kumar, and Z. Li, "Topic modeling and sentiment analysis of global climate change tweets," *Social network analysis and mining*, vol. 9, no. 1, pp. 1–20, 2019, doi: 10.1007/s13278-019-0568-8.

[41] A. F. Hidayatullah, E. C. Pembrani, W. Kurniawan, G. Akbar, and R. Pranata, "Twitter Topic Modeling on Football News," Apr. 2018, pp. 467–471. doi: 10.1109/CCOMS.2018.8463231.

[42] B. V. Barde and A. M. Bainwad, "An overview of topic modeling methods and tools," Jun. 2017, pp. 745–750. doi: 10.1109/ICCONS.2017.8250563.

[43] L. He, Y. Jia, W. Han, and Z. Ding, *Mining user interest in microblogs with a user-topic model*, vol. 11. 2014. doi: 10.1109/CC.2014.6911095.

[44] D. Yu, D. Xu, D. Wang, and Z. Ni, “Hierarchical Topic Modeling of Twitter Data for Online Analytical Processing,” *IEEE access*, vol. 7, pp. 12373–12385, 2019, doi: 10.1109/access.2019.2891902.

[45] N. H. Khun and H. A. Thant, “Visualization of Twitter Sentiment during the Period of US Banned Huawei,” Nov. 2019, pp. 274–279. doi: 10.1109/AITC.2019.8921014.

[46] F. Viegas, M. S. Alvim, S. Canuto, T. Rosa, M. A. Gonçalves, and L. Rocha, “Exploiting semantic relationships for unsupervised expansion of sentiment lexicons,” *Information systems (Oxford)*, vol. 94, pp. 101606-, 2020, doi: 10.1016/j.is.2020.101606.

[47] S. A. E. Rahman, F. A. AlOtaibi, and W. A. AlShehri, “Sentiment Analysis of Twitter Data,” Apr. 2019, pp. 1–4. doi: 10.1109/ICCISci.2019.8716464.

[48] R. Wagh and P. Punde, “Survey on Sentiment Analysis using Twitter Dataset,” Mar. 2018, pp. 208–211. doi: 10.1109/ICECA.2018.8474783.

[49] V. Prakruthi, D. Sindhu, and D. S. A. Kumar, “Real Time Sentiment Analysis Of Twitter Posts,” Dec. 2018, pp. 29–34. doi: 10.1109/CSITSS.2018.8768774.

[50] “R: What is R?” https://www.r-project.org/about.html (accessed Jun. 16, 2021).

[51] “Welcome | Mastering Shiny.” https://mastering-shiny.org/ (accessed Jun. 17, 2021).

[52] joshua_been, “Archive Team: The Twitter Stream Grab (Historic Twitter Content),” Nov. 02, 2018. https://blogs.baylor.edu/digitalscholarship/2018/11/02/archive-team-the-twitter-stream-grab-historic-twitter-content/ (accessed Aug. 19, 2021).

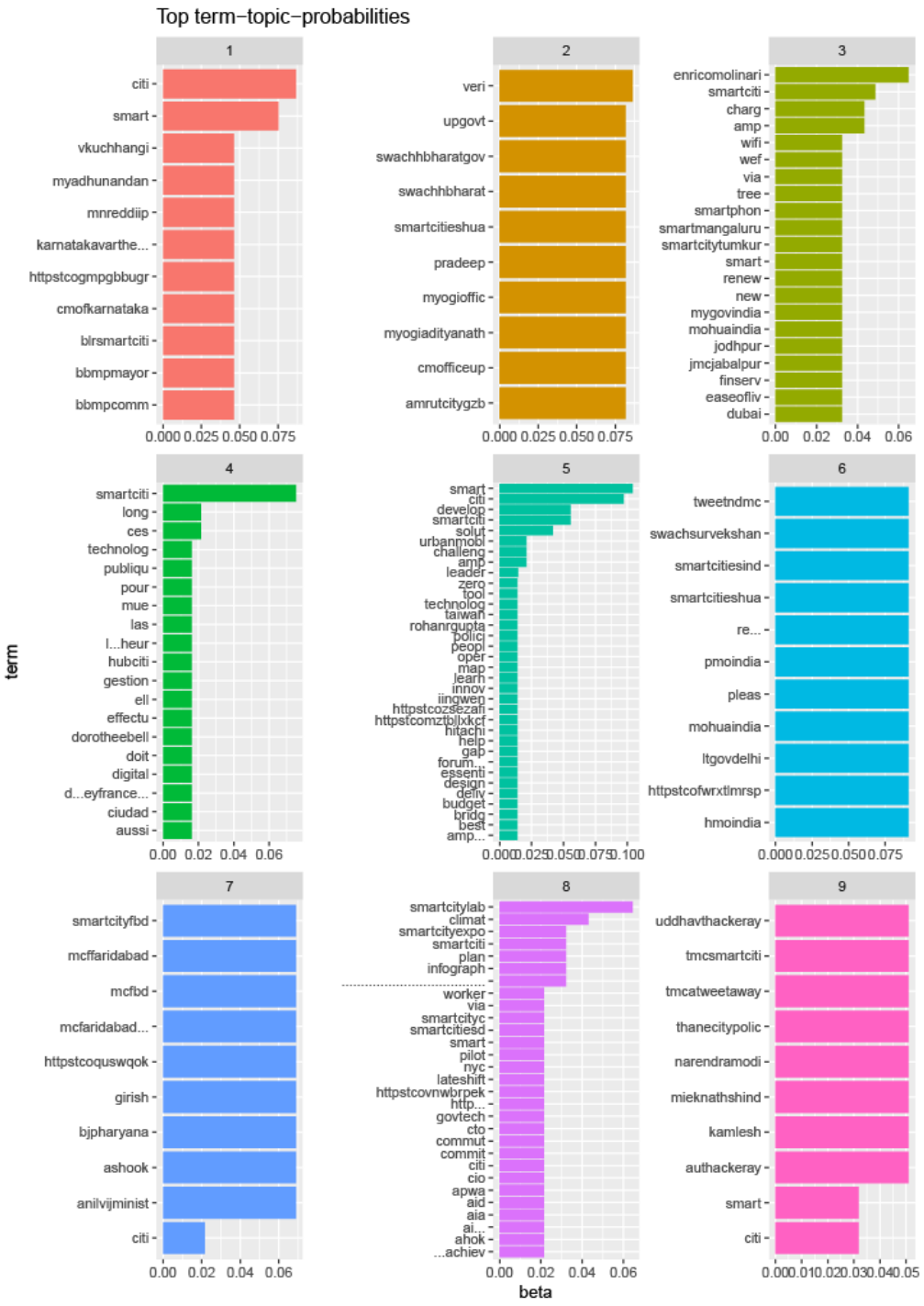
[53] G. Stricker, “The 2014 #YearOnTwitter.” https://blog.twitter.com/en_us/a/2014/the-2014-yearontwitter (accessed Apr. 12, 2022).

[54] “Twitter API Documentation.” https://developer.twitter.com/en/docs/twitter-api (accessed Apr. 13, 2022).

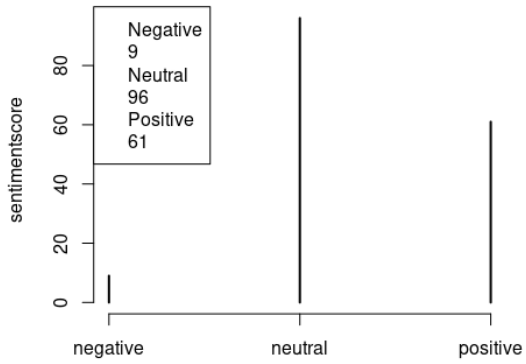
APPENDIX 1. Results from the tool

This appendix contains topic models and sentiment analysis as they came from the tool. Every quarter consist of 3 months of data transformed into topic model and sentiment analysis.

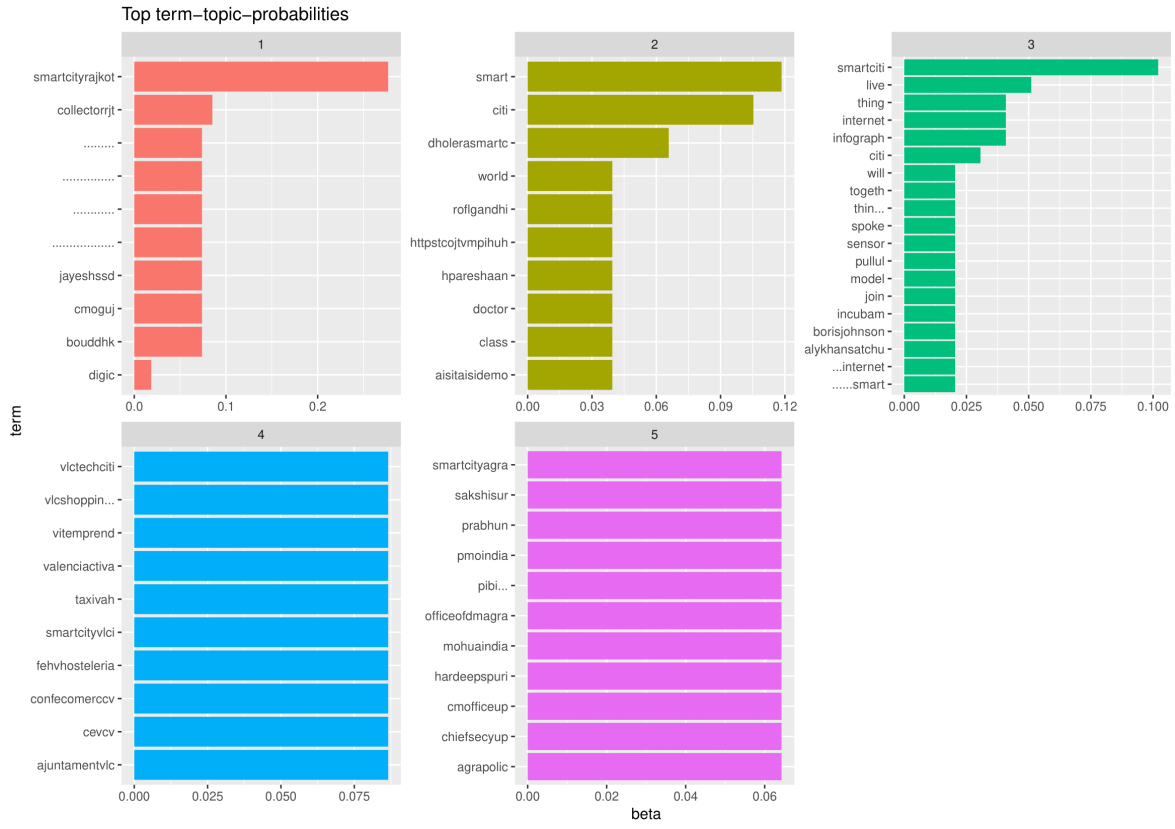
A 1.1 Quarter 1



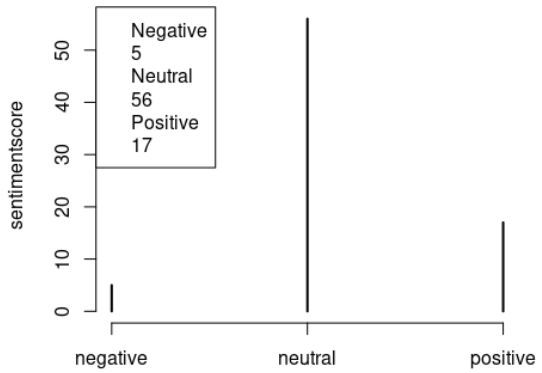
APPENDIX 1. (continues)



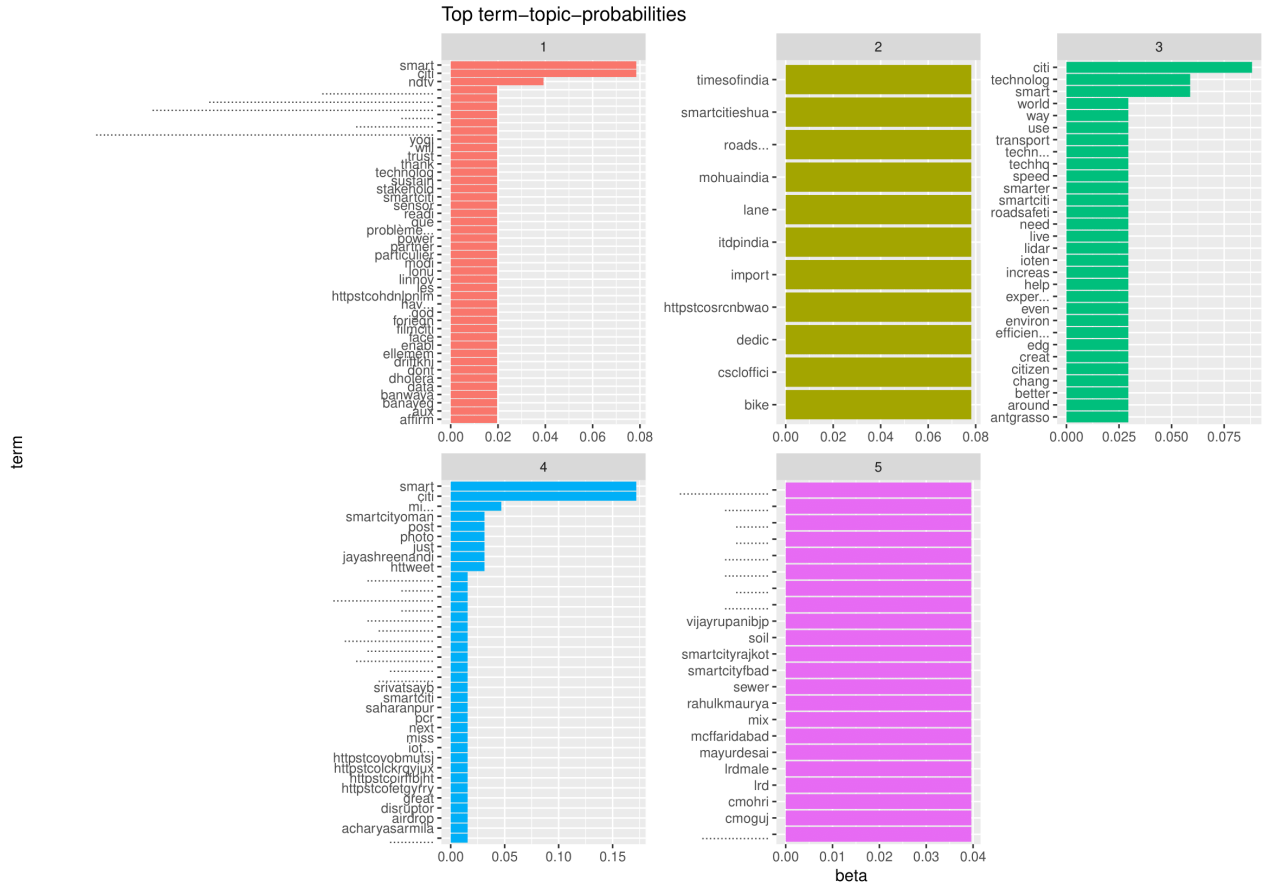
A 1.2 Quarter 2



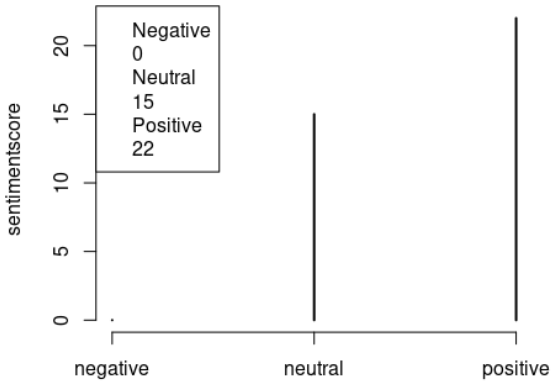
APPENDIX 1. (continues)



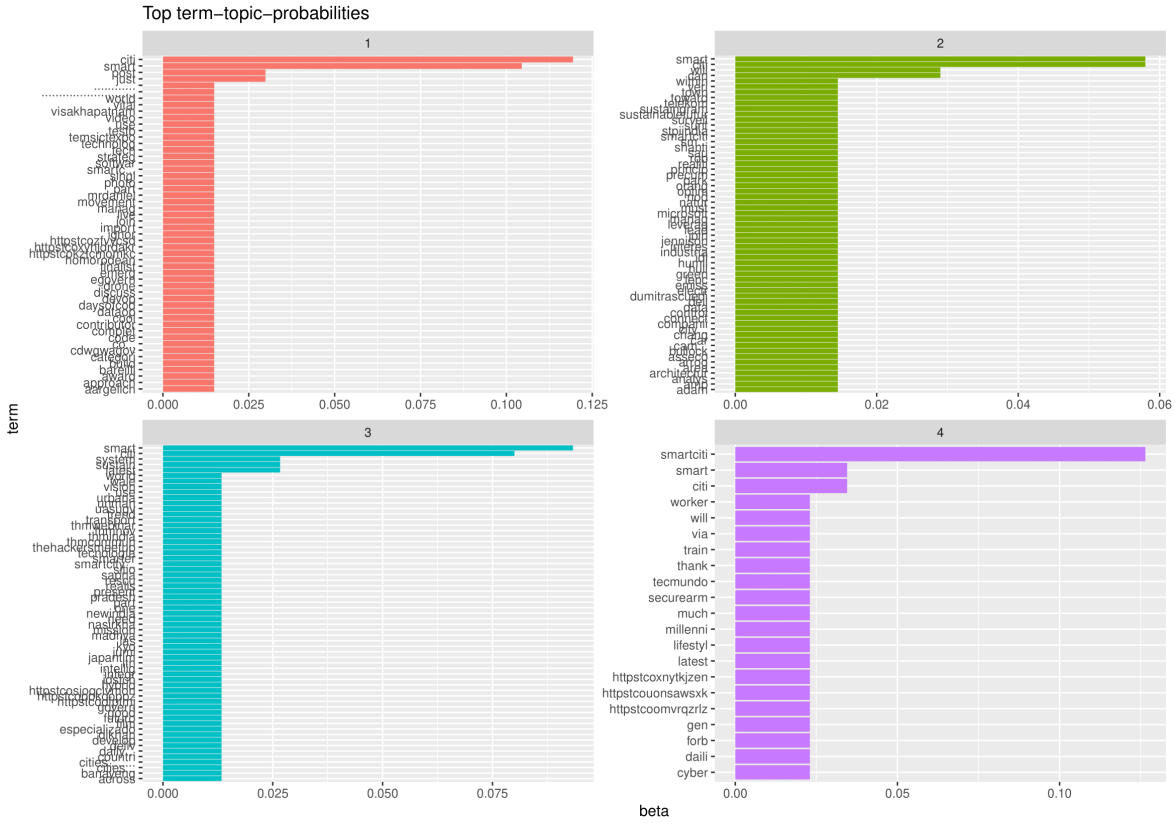
A 1.3 Quarter 3



APPENDIX 1. (continues)



A 1.4 Quarter 4



(continues)

APPENDIX 1. (continues)