UNIVERSITETET I AGDER

Machine Learning Techniques to Predict Pandemic from Social Media

VIDA NEJATIMOGHADAM

SUPERVISOR

Dr. Jaziar Radianti

University of Agder, 2018
Faculty of Engineering and Science
Department of Information and Communication Technology

UNIVERSITY OF AGDER

IKT-590 MASTER THESIS

Machine Learning Techniques to Predict Pandemic from Social Media

Author:
Vida Nejatimoghadam

Supervisor:
Dr. Jaziar Radianti

This Thesis is submitted in Partial Fulfillment of the Requirements for the degree of Master of Science in Information and Communication in the

Faculty of Engineering and Science
Department of Information and Communication Technology

June 3, 2018

Abstract

In recent years, there has been a particular focus on improving public health through the means of prediction and preparedness of pandemic diseases. Early detection, prediction, and analysis of disease outbreaks allow the authority agencies to mitigate the side effects of Pandemic and immune the people. Nowadays, social media such as Twitter or Facebook play a vital role in the crisis situation. By means of social media, people from all over the world can be aware of the recent pandemic outbreaks. In fact, the mainstream adoption of social media in people daily life has caused a paradigm shift in how people communicate, create, cooperate, and use information during a crisis. Moreover, by analyzing data, which broad-casted during a crisis, the relevant health organizations or agents can discover much useful information.

On the other side, the volume and velocity of messages or tweets during crises today tend to be extremely high and make it hard for discerning and taking an actions. Therefore, machine-learning techniques can be used to help analyzing this big flow of messages. They are the useful way to discovering the knowledge from big data. Various methods and machine learning algorithms have been proposed and applied in various cases.

In this work, we adopted data analysis and predictive techniques. Several features from tweets have been extracted and data is modeled by binary classification for pandemic prediction. Three different predictive models (Support Vector Machine, Decision Tree, and Naive Bayes) have been conducted in order to pandemic prediction. Our experimental results illustrate that SVM technique outperforms other techniques. However, there is no global best predictive model and it depends on various parameters such as dataset, configuration, etc.

Acknowledgements

This report serves as a Master Thesis to fulfill the requirements of the Master Program in Information and Communication Technology (ICT) at the Faculty of Engineering and Science, University of Agder (UiA) in Grimstad, Norway. It is followed the work performed in the context of IKT590 Master thesis, which counts 30 ECTS credits towards the degree of MSc in ICT. The thesis work started on 15th of January 2018 and ended on 25th of May 2018.
I am very grateful to my supervisor Dr. Jaziar Radianti. Her constant support, supervision, and guidance helped me a lot and leads my endeavors to reach ideal outcomes. Last, but certainly not the least, I would like to offer my sincerest gratitude to my family, especially my Mom and Dad for supporting me during my life.

Finally, I would like to express my deepest gratitude to my kind husband, Meisam, who was a source of encouragement for me. The completion of this thesis would not be possible without his infinite love and support.

Vida Nejatimoghadam
Grimstad, Norway
Spring 2018

Contents

List of Figures

List of Tables

List of Abbreviations

AI	**A**rtificial **I**ntelligence
BoW	**B**ag **O**f **W**ord
CART	**C**lassification **A**nd **R**egression **T**rees
CL	**C**onfidence **L**evel
DT	**D**ecision **T**ree
IT	**I**nformation **T**echnology
ILI	**I**nfluenza **L**ike **I**lness
MPQA	**M**ulti **P**erspective **Q**uestion **A**nswering
MERS	**M**iddle **E**ast **R**espiratory **S**yndrome
NLP	**N**atural **L**anguage **P**rocessing
Naive Bayes	**N**aive **B**ayes
ONLP	**O**pen **N**atural **L**anguage **P**rocessing
ROC	**R**eceiver **O**perating **C**haracteristic
SNA	**S**ocial **N**etwork **A**nalysis
SVM	**S**upport **V**ector **M**achine
SARS	**S**evere **A**cute **R**espiratory **S**yndrome
URL	**U**niform **R**esource **L**ocator
WHO	**W**orld **H**ealth **O**rganization

Chapter 1

Introduction

1.1 Background

In the recent years, social media have experienced massive growth in their user base. There are more than one billion members belonging to various social Medias such as Facebook and Twitter. There are a huge number of various social media platforms or applications, which can be categorized as web-logs, micro-blogs, discussion forums, and so on. In general, social media refers to a conversational, distributed model of the content generation, dissemination, and communication between communities. The mainstream adoption of social media applications has effected a paradigm shift in how people collaborate, communicate, create, and use information in various situations. Many scholars such as computer scientists, economist, mathematicians, bio-informaticists, and others are demanding access to the huge quantities of information created by people, and their interactions during events via social media. Therefore, social media can be known as a type of living lab, which allows academics to collect large amounts of data generated in a real-world environment [1].

Diseases with slow rates of transmission or less rates of symptomatic are classified as pandemics. A pandemic is an epidemic disease that has diffused across populations over wide areas such as continents or globally. A number of viruses include pandemic potential. For example, the corona virus responsible for the severe acute respiratory syndrome (SARS) appeared in southern China for the first time in November 2002. It caused 8096 cases and

774 deaths cross 26 countries [2]. Thus, a pandemic is not to be confused with an endemic. An endemic disease can be controlled with regard to how many people may infect. On the other hand, pandemic have most fatal threats to human throughout history. Two most well-known pandemics are tuberculosis, which is an airborne bacterial infection and smallpox, which is a viral infection that has influenced humankind for thousands of years. The other recent examples of pandemics are human immunodeficiency virus (HIV) and the H1N1 (Swine Flu) pandemic of 2009.

According to the World Health Organization (WHO), the H1N1 influenza pandemic, which is known as the Swine flu was the first influenza pandemic in the 21st century. Pandemic influenza can be catastrophic, the spread of this virus was impressive from 1918 to 1919 is estimated to have infected 500 million persons globally and have killed 50 to 100 million persons [3]. The government of Mexico had reported cases of Influenza like-Illnesses (ILI) in various regions of the country in March 2009. There were over 854 cases of Pneumonia and 59 deaths from the capital on 23rd of March [4]. Cases were also reported in Southern California (USA). The scientists from the center for disease control and prevention (USA) were detected the emergence of the novel influenza virus 2009. The director general of the WHO under the advisement of the Emergency Committee called under the laws of the International health regulations and announced the ongoing event as a Public Health Emergency of International Concern [5].

The chain of following this announcement happened so fast. The WHO director-general raised the influenza pandemic alert level from 3 to 4 on 27th of April utter that the probability of a pandemic has increased [6]. Two days later the alert level was raised further to level 5 because of the H1N1 influenza virus capacity to spread rapidly in all over the world. Moreover, on the 11th of June 2009, nearly 30,000 cases in over 74 different countries were reported and the WHO raised the pandemic alert level to 6 . Influenza pandemic spread globally and caused nearly 17,000 deaths by the beginning of 2010 [7]. The H1N1 2009 pandemic represented a public health emergency of unknown scope, effect, and duration. The WHO utilized six-stage classification systems to explain how this virus changes from a disease that infects many people to one that has become pandemic. In addition, the WHO has considered an initiative plans, instructions, and recommendations on how to manage an outbreak of pandemic influenza.

There is no proper model that can estimate the impact of this virus in the societies yet. Models are based on data from industrialized countries, which may underestimate the real impact of a pandemic in developing countries [8]. However, forecasting is a fundamental component in a decision support system for preparing and responding to pandemic outbreaks. Different types of simulation models have been proposed in order to predict the result of pandemic outbreaks in the presence of different interferences.

Every government has estimated and dedicated the high percentage of the budget for Information Technology (IT) based health care support services, especially in crisis situations. The adoption of micro blogging platforms such as Facebook, Twitter, and on-line newspapers during crises, emergencies, and critical events has increased recently. Easy access to social networks leads to produce and retrieve information in various forms, such as textual message, videos, and images [9]. Access to the critical information becomes more important especially in the first hours when no other information sources such as news channels or traditional ways are available. In addition, quick access to vital data can help emergency responders, health services and decision makers acquire situational awareness, detection, sooner decision making, and more efforts accordingly.

Since Twitter launch in 2006, it has not only attracted over 500 million registered users but also turned into a popular subject of scholarly research. It has more than 1400 scientific publications on Scopus, 470 on the web of science, and 10,000 more available via Google Scholar. These publications emerge from various scholarly fields and disciplines and address different research questions [10]. In general, Twitter used by fewer people than Facebook, but more studies have been done on Twitter. One possible reason might be that it is relatively easy to get access to Twitter data in comparison with other social media. Twitter research is based on valuable data sets, which can be retrieved via the Twitter API and other special tools. In some cases, this is known as working with big data, which it means the number of collected tweets or monitored users rather than to the capacity of a device or the size of storage required for the respective files. While there are, also some challenges for computational infrastructure to manage big Twitter data for analysis, developing theories, methods and models for particular research are also more important challenges in order to make sense of thousands and millions of tweets.

1.2 Motivation

At the point of mass emergencies, a phenomenon recognized as the collective behavior becomes manifest [11]. It includes of socio-behaviors, which consist of information contagion and amplified information search [12]. Disasters and emergencies bring unsure situations. People involved in such situations search for prompt answers to their quick queries. Access to critical information becomes more vital when no other resources such as traditional media are available, especially in the first few hours. In these situations, people desire to know where exactly their families and friends are as not being able to reach them or might be able to contact them, can have frightening instants during these crisis situations. Availability of immediate information can save lives during emergencies. People share information about approaching threats, where to go for help, donations and so on. Thus, it is necessary to keep abreast of the latest developments. However, this is hard since the produced information by users under crisis situations is scattered with different qualities. Social media is utilized for social interaction. They are enabled by communication technologies such as the smart phones, web, and they turn communication into an interactive dialogue. Interactions on social media are highly distributed, decentralized, and occur in real time.

Since social media propose a uniquely powerful and quick way to disseminate information, accurate or inaccurate equally similar to incorrect information, which can spread like fire. However, social media tend to spread valid information over rumors [13].

Twitter and Facebook are popular examples of using social media in crisis situations. Twitter is a micro-blogging service with a lightweight chat permitting users to post and exchange messages called as tweets. Although most tweets are chatter, they are also used to share news and relevant information. It becomes a valuable tool in disaster situations. In emergencies, tweets either process first-person observations or find relevant knowledge from external sources [14]. Information from different sources such as official sources, reputable sources, and others is propagated. Other users then elaborate and combine this pool of information to generate derived interpretations.

During the Mumbai terror attacks of 2008, a group of online users created a Twitter page voluntarily to share and update situational information on the attacks. A study acquired that 52.67 % of the tweeted H1N1 material in 2009 to be related to news and information on swine flu [15]. During the Haiti earthquake, Twitter was employed to create awareness about the disaster and mobilize people to help [16].

Social network Analysis (SNA) is a sociological method for analyzing patterns of interactions and relationships between social actors. In this way, the underlying social structure such as central nodes that act as hubs, leaders, and patterns of interactions between groups are discovered [17]. What has been discovered is that the social media is solely valuable to the statistical study of IT, social behavior [18] within quantitative attributed to e-learning process and simulation design for further data mining.

In this study, we propose to use SNA to study the community of Twitter users disseminating information during the related pandemic tweets in order to determine interesting patterns and features within this on-line community. The automatic learning based approach is applied to predict the people fears in the content of the tweet messages. Automatic prediction with the power of machine has much lower costs in comparison with the human-based work. In addition, this automatic approaches can measure very large datasets, which would be impossible to handle with human-based classifications.

Thus, we collected various tweets from different users refer to pandemic via Twitter between February to May 2018 by using the Knime Analytics platform. The collected data has been analyzed, and machine-learning technique has been applied for pandemic prediction.

1.3 Problem Statement

All crisis events represent a new set of difficulties that need preparedness and quick responses to be taken by a plenty of organizations and governmental institutions, different nations and cultures, and all with different world views [19]. When an influenza pandemic appears, all countries worldwide will inevitably be affected. However, the impact may vary within countries. During the Spanish flu pandemic from 1918 to 1920 the estimated deaths among different countries indicates that mortality rates in North America and Europe were significantly lower than those in Asia, Latin America, and Sub-Saharan Africa [20]. A recent study that estimated the global impact of the Spanish flu pandemic shows that a considerable difference in mortality rates was observed between high and low-income countries. Why the pandemic caused such high mortality rates in developing countries is an open question yet. Several factors may have been involved, such as weak public health infrastructures, lack of access to adequate medical care, social factors (housing conditions and population density), and nutritional status. Unfortunately, it is so difficult, if not impossible, to predict when and which infections disease will affect the world. Another potential factor likely to happen in a future pandemic is the high HIV/AIDS outbreak in some developing countries. A study concluded that 96 % of the estimated 63 million deaths in a future pandemic would happen in developing countries [21]. The impact of high mortality rates clearly requires to be taken into account when creating pandemic preparedness plans for developing countries.

Surveillance systems and realizing the dynamics of infectious disease transmission have improved considerably since the 1968 influenza pandemic. These improvements can be utilized to support policymakers in current pandemic management. Understanding the problem for pandemic management starts with understanding the various societies. For example, there may be a considerable difference in the spread of a disease in the rural area with a low level of commuting compared to a main urban area with the wide public transportation system. There may also be diversity in the age distribution of the population and social structure. Therefore, local decision makers have to make a decision on what interventions to use or activate. Some determined interventions may not be available or applicable to the specific community. Others may have less or no impact on the spread of the disease. All these issues must take into account in the preparedness and response process.

One of the biggest challenges to be addressed by computer science can be the management of large volumes of data in crisis situations. The wide range of data acquisition sources available at the instant of crisis creates a need for data integration, aggregation, and visualization. These techniques assist crisis management officials to optimize the decision-making method, which the quality of these decisions depends on the quality of the available data.

It is stochastic and unpredictability nature of pandemic disease emergence, which is the largest challenge. This is somewhat comparable to earthquakes or tsunamis [22]. There is a quick need to come up with "early warning signals", which can proper predict when and which pathogens might emerge.

Larsson states that it is important to look back and learn from past experiences, successes, and failures in crisis management [23]. Moreover, the major part of crisis management is the exchange of information and communication in the crises. Larsoon also emphasizes data analysis and how media reports on crisis events is very important.

In this work, the collected twitter data about pandemic is divided in three categories based on how tweets can be relevant or not relevant to pandemic outbreaks. Therefore the groups named as Pandemic_ Related, Communicable Disease, and Not relevant. Then, by extracting different features from collected tweets, data will analysis precisely. In addition, by applying machine learning predictive models, we will compare the performance of these models and find features to improve their performance. However, we will perform analysis methods on each data feature, or different classifiers, to gain insight into the contributions and limitations of each feature.

1.4 Purpose and research questions

To fulfill the problem statement, the main research questions in this project are:

- How to detect pandemic issues from Twitter messages?
- How to analyze people sentiments derived from their tweets about the pandemics? Are people always concerning about the pandemic or the concern changes frequently?
- Which tweets or users had more influences on the network?
- Which tweets by which users have been re-tweeted more?
- Can we locate pandemic issues from different countries based on user location?
- Can machine learning techniques be used to predict some pandemic outbreaks?
- Which machine learning techniques can perform more successfully?

1.5 Structure of this thesis

This thesis is organized as follows:

This chapter introduces the purpose and research questions of the thesis and provides an overview of the scope of the study. In chapter two, the related theoretical background and review are introduced. The data collection and method of analysis are explained in chapter three. In chapter four, the analysis and the obtained results are illustrated. Finally, we draw some conclusion and summarize this project.

Chapter 2

Literature Review

In line with the research area of the thesis, the theoretical background of the fields of crisis management and social media will be outlined. The purpose of this section is providing an overview and understanding of the theories applied in the thesis. In addition, the crisis definition, crisis types, and emergency management stages will be discussed.

2.1 Introduction of Crisis Management

2.1.1 Crisis Definition and Types

The word "Crisis" can be hard to describe, and no general definition is agreed by everyone. Therefore, when trying to explain and categorize crises, this can vary between different organizations. The literature on crisis management offers different definitions on the word crisis. A crisis is the realization of an unpredictable occasion that threatens main expectancies of stakeholders and can influence an organization's efficiency with the negative outcome [24].

In addition, Hermann [25] states a crisis is something that 1) threatens high-priority values of the organization, 2) present a restricted amount of time in which a response can be made, and 3) is unexpected or unanticipated by the organization. Based on the Cambridge advanced learner's dictionary, the crisis is defined as a situation that has reached a very difficult or dangerous point, a time of great disagreement, uncertainty or suffering.

Nearly all of the definitions have similar and common things, the poverty of time, high level of uncertainty, and threats to values. What distinguishes them is their focus. Researchers have tried to categorize different dimensions of crises. Mitroff and Pauchant [26] divided crises into technical, social and human crisis, and economic. Meanwhile, Fearn-banks defined types of crises as external economic attacks, mega-damage, external information attacks, psychological and internal crises [27].

Falkheimer and Heide define “societal crisis” and “organizational crisis”, regarding work made by Sundeluis, Stern, and Bynander. A societal crisis is based on a situation where the central operators experience that considerable values are threatened, with only limited time and unpredictable. An organizational crisis is defined by the equivalent characteristics, considerable values, unpredictability, and urgency [28]. By media technologies development, the probability of crisis in the organization is gradually increased.

Shaluf and Agmadun Said [22], crisis types can be divided into four groups. Community crisis, Non-community crisis, conflict crisis, and Non-conflict crisis. The first types contain a crisis, which occurs due to natural disaster, industrial or non-industrial crisis that happens due to the political or non-conflict crisis. The second types, which do not affect the community itself, are mostly transportation accidents. The third types contain a crisis, which happens because of any inter-humanitarian conflict or struggle. From the community point-of-view, these types can be an external crisis, such as way, threat or terrorism or internal crisis, such as religious conflicts or dictatorships. The last group can contain any kind of economic crisis or social crisis.

In the crisis situation always is required decisions be made quickly due to damage avoidance. In order to control damages and make decision crisis management is determined. It is the designed process of strategies in order to support an organization faced with a sudden negative or unpredictable event.

In general, crisis management consists of methods used to respond to both reality and conception of crises. The response should contain action in the area of crisis prevention, crisis assessment, crisis handling, and crisis termination. For this reason, a crisis management planning is definitely recommended.

2.1.2 Emergency Crisis Management Stages

It is essential to examine the specific stages of crises. Generally, researchers have categorized five stages of a crisis. Fearn-Banks illustrated five stages of a crisis as detection, prevention/preparation, containment, recovery, and learning [27]. Fink S proposes that a crisis can be divided into four separate stages [29]: prodromal crisis stage, acute crisis stage, chronic crisis stage, and crisis resolution stage. According to [24], crisis management has three stages: pre-crisis (prevention and preparation), crisis response.

The literature on crisis management typically determines four to eight phases of the emergency management process. Nearly all classifications contain four basic phases. Mitigation, preparedness, response, and recovery are counted as emergency management phases.

Mitigation refers to a pre-disaster operation to identify risk, diminish, and decrease negative effects of the identified type of disaster event on personal property and human life. The aim of mitigation is preventing future emergencies or minimizing their effects.

Preparedness refers the actions taken former to a possible disaster that enables the emergency managers and people to be able to respond adequately when a disaster occurs. Preparedness also includes having adequate information systems up and running and practicing with them so that they can be used for command and control to coordinate emergency personnel and locate resources and keep track of the location of evacuees, for instance. Generally, it is preparing to handle an emergency.

The **response** phase contains operation immediately prior to a foretold event, as well as during and after the disaster event, that helps to diminish human and property losses. Examples of such actions include placing emergency supplies and personnel, searching for, rescuing, and treating victims, and housing them in a temporary, relatively safe place.

The **recovery** phase is sometimes never completed, its aim is to enable the population affected to return to their normal social and economic activities. Therefore, for example, recovery would include replacing a destroyed bridge or other missing infrastructure, as well as rebuilding permanent housing that was lost in the disaster. The models and maps contained in geographical information systems are the main aids in the planning and management of a recovery process. For any emergency operation, there must be a combination of communications and information technology and a resource management system to support those involved in any phase of the emergency.

Since Second World War emergency management has centralized primarily on preparedness. Mostly it is involved readiness for an enemy attack. Community preparedness for all disasters needs expertise and identify resources in advance, then planning how to apply them in a disaster. Although, preparedness is one of the emergency management phases. Current thinking determines more phases of emergency management.

2.2 Pandemic and Crisis Management

Pandemics are regionalized infectious disease outbreaks that deserve more consideration for crisis management. It more or less similar to the response and decision making of other natural disasters, such as earthquakes and hurricanes. Although these kind of infectious disease outbreaks are solely a gentle threat when occurring in conjunction with these natural disasters [30], they can be a much higher threat when happening separately.

Response to epidemic outbreaks uniquely benefits from local system attentions that raise early analysis of an epidemic, assist treatment, manage public concern, and limit the spread and influence of the disease. In the 21st-century influenza outbreaks were responsible for more than 43 million deaths. These globally spread epidemics or pandemics were the serious from 1918 to 1919. The Spanish flu pandemic also killed more than 40 million people, and two moderate pandemics, the 1957-1958 Asian flu, and 1968-1969 Hong Kong flu killed people 2 million and 1 million respectively [31].

The WHO has taken initially to plans, instructions, and recommendations on how to manage to manipulate an outbreak of pandemic diseases. On the national level, governments also have developed rules, policies, and plans regarding WHO's recommendations, for managing spreads and have adopted legislation to deal with threats and risks. Local authorities also are responsible for planning and interventions implementation in case of the real outbreak with considering both national and local policy. Decision makers are expected to take action to decrease the consequences of the pandemic; however, they meet serious problems that may prevent an effective response [32].

A systematic preparedness and response process has a cyclic nature consist of Evaluating, planning, exercises, and outbreak response. By evaluating, the situation to determine risks, needs, vulnerabilities, and resource is the natural first step, which makes the input to the planning steps. Combination of evaluation and planning can be conducted before an outbreak, however, some aspects of the diseases and its effects may be unknown, so the resulting plan must be considered imperfect and hypothetical.

Exercises provide opportunities to test plans and train key personnel in a controllable and safe environment. The final test of preparedness is an actual outbreak. Decision makers must analyze the situation, explore alternative courses of necessary action, taking resources and information into account, and adapt the plan respectively. The experience aggregated from the response can provide valuable feedback for preparing the next outbreak.

2.3 Social Media in the Analysis Technique

2.3.1 Sentiment Analysis

In recent years, social media has provided information sharing in social networks. Last researches have defined many factors that drive information propagation such as content-related features (e.g., hashtag inclusion, topics, and URL (uniform resource locater). In addition, user and network characteristics such as popularity need to take into account. However, some research has drawn attention to emotions as another potential driver of information propagation in a social media setting, especially regarding the user's data sharing behavior.

Social media content quotes information about the author's emotional case, his or her judgment or evaluation of a certain person or topic, or the defined emotional communication (i.e., the emotional impact the sender wishes to have on the receiver), which is termed as "sentiment".

Results from different studies on social media, discussion forums, on-line news portals or other context shown that the effective dimensions of messages both negative and positive sentiment could trigger more cognitive involvement in terms of attention and influence on feedback and social sharing behavior [33].

Sentiment analysis allows us to manually or automatically classify tweets based on their emotionality such as positive or negative. For on-line communication, [34] processes evidence of sentiment diffusion, which is shown how messages including positive or negative emotions and words are mostly received verbal responses. Studies have also represented that emotionally-charged content is more likely to be shared by on-line users. There are many tools, which might be used to automatically categorize tweets with respect to their sentiments.

In this work, the Knime Analytics Platform is utilized for sentiment analysis and results will be discussed in result chapter.

2.3.2 Tracking

Social media tracking or monitoring is the process of reading, watching or listening to the content of media resources continuously. Then, identifying, saving, and analyzing content that includes on certain keyword or topics. While the news is obviously the core content to monitor, most organizations now realize the growing importance of tracking social media. For example, what is published on blogs and sites such as Twitter or YouTube have gained significant influence. In addition, many organizations are tracked work of mouth mentions about their jobs in social media such as Facebook, Twitter, and forums. There are many tools for tracking and monitoring, which organizations applied to their business in order to get the better view about their futures. By applying this method, they also will be able to do many data analysis regarding their jobs or topics, which they selected.

2.3.3 Predicting

All decision makers are expected to take necessary action to reduce the consequences of the pandemic, but they face some difficulties that may prevent an effective response or action. Limited supplies, incomplete information about the virus transmission and effectiveness of various medicines force decision makers to be considered an alternative strategy during the pandemics. Population distribution age, school structure, and infrastructure, social and economic conditions, affect how the disease will spread in the society or local community.

Forecasting is a vital component in a decision support system in order to prepare and respond to pandemic outbreaks. Various simulation models have been developed in order to predict the outcome of pandemic outbreaks in the presence of different interventions [35]. However, these simulations also have limitations with respect to the decision maker's demands. A decision support system can assist the decision maker overcoming these problems in the planning and response process. In this project, we present three different types of predictive models for pandemic outbreaks.

2.3.4 Application in Crisis Management

In last decades, social media has experienced wonderful growth in user base and have an impact on the public communication and discussion in society. For example, in 2012, twitter counted more than 200 million accounts [36]. Social media services such as Facebook or Twitter are increasingly utilized for communication during crisis situations by the members of the affected public, individuals, organizations, and professional media [37].

Twitter is an important channel for public communication and emerged as a widely used social reporting tool to spread information on the social crisis [38]. In Twitter, the users are restricted to writing messages of no more than 140 characters these are turned into short messages. These short messages make important information dissemination in the network and become Twitter as a successful social network for content dissemination.

Recent research indicated that Twitter is a rapid information diffusion tool under large-scale crisis such as terror attacks, natural disasters, and social movements. Twitter turned out to be much rapid in tweeting situational reports to the on-line community, due to its short texting service interface on cell phones and easy access [12]. Therefore, allows the first responders to collectively cope with the crisis situations.

Vieweg analyzed the Twitter logs for Red River Floods and the Oklahoma Grass-fires (March and April 2009) from the aspect of situational awareness content [14]. An automated framework to increase situational awareness during emergency situations was developed. They Geo-location and location-reference information extracted from user's tweet, which helped in increasing situational awareness during emergency events [14].

Agrawal and Raghav [39] did another related work, they analyzed twitter stream during the 2008 Mumbai terrorist attacks. Their analysis indicated how information available on on-line social media during the attacks helped the terrorists in their decision-making by increasing their social awareness. One main conclusion obtained was that during emergency situations, users utilize a certain and specific vocabulary to quote tactical information on Twitter.

Moreover, by using brief keywords or abbreviation, prefixed by the symbol '#' which called hashtag, makes tweets more easily search-able among all Twitter message traffic. It enables users to follow real-time feeds of all messages including hashtag. Social media technologies mediate human communications in social crisis situations and represent different patterns of crisis communication. In the Twitter, users not only consume the incoming information from their own network, but also spread information to their own network. Twitter is considered an efficient communication tool in different situations, especially in crisis situations.

2.4 Predicting Pandemic by Applying Machine Learning Techniques

2.4.1 Theories on Used Technique

Machine learning is a subset of artificial intelligence (AI), which is concerned with the development of different algorithms permit the computer to learn. Machine learning theory, also known as computational learning theory, aims to understand the basic principles of learnings as a computational process. The concept of these techniques is to learn the theory automatically from the data, via a process of inference, learning from examples or model fitting. A model is determined with some parameters and learning is the performance of a computer program using training data to optimize the execution of the model. The model also can make predictions in testing dataset applying the knowledge achieved through learning from examples [40].

The goals of this theory are both to understand basic issues in the learning process and help to design of better-automated learning methods. This theory plans elements from both the theory of computation and statistics and contains tasks such as:

- Create mathematical models that capture key aspects of machine learning, in which one can analyze the difficulty of various types of learning problems.
- Mathematically analysis general issues, such as: “When can be sure about predictions made from limited data?” “What kinds of techniques can apply in the presence of distracting information?”
- Proving guarantees for algorithms and developing proper machine learning algorithms that meet required criteria.

Machine learning has divided into several subfields dealing with various kind of learning tasks. The supervised and unsupervised, Active and Passive Learners, and online and Batch Learning Protocol are all in the taxonomy of learning paradigms. The supervised learning will be only focused because it has been utilized in this project. Supervised learning is a type of system in which both input and output are provided. The input and output data are labeled for classification in order to further data processing. Supervised machine learning systems process the learning algorithms with known quantities to support future judgments. They are mostly associated with retrieval-based AI. However, they may be capable of using a generative learning model. The training data for supervised learning contains a set of samples with paired input subjects and output [41].

All pandemics control needs two steps: early detection of new pandemic and development of vaccines. It required a system for quick detection of pandemic strains, which involves. Different numbers of machine learning techniques are applied to recognize relationship or aggregation in biological data, to group same genetic elements, and to analyze and predict diseases [42]. Machine learning is concerned with the automatic acquisition of models from data and applying them for automatic inference and prediction. Since addressing pandemic viral analysis is concerned with pandemic viral strains classification, the role of certain positions and modeling for future prediction, machine-learning techniques offers much in this regards.

2.4.2 Existing Use of the Technique

Most machine learning algorithms or models view learning as a standalone process, focusing on prediction accuracy as the measure of performance. However, the systematic ways to utilize machine learning is limited. For example, one would like the more powerful algorithms that optimize multiple objectives. One would like models that capture the process of deciding what to learn and how to learn. There has been some theoretical work on these problems, but there is much more have to do.

Machine learning approaches are applied in various applications such as search engines, natural language processing, pattern recognition etc. Current research domains where machine learning used are multiple sequence alignment, molecular clustering and classification, and regression analysis. It has achieved superior success in many applications [41].

Various studies have been done on the possibility of social media data to be a predictor of public health, with research into Twitter analysis such as the Swine Flu pandemic [43], influenza [44], and Dengue outbreaks [45].

Machine learning classifiers require different features for learning so different researchers from time to time have selected various features for comparing results. Agarwal [46] and Pak [47] selected different features as bigrams, pos tagging, hashtags, ngrams etc. and found mixed classification results. Hassan khan approach contains rigorous data pre-processing followed by supervised machine learning. They gathered labelled data collections of different sources so that machine learning will not be limited to a particular domain [48].

Agarwal state that in order to have better results by using machine learning approaches and finding correct features are challenging tasks [46]. They used various features for their classification tasks. Different approaches and classifiers such as Naïve Bayes (NB), Support vector machine, lexicon based etc, have been applied time to time with different parameters for evaluation.

In this project, we tried to apply three different predictive machine learning techniques in the domain of pandemic outbreaks sequence analysis. We have conducted decision tree classification, Naïve Bayes and support vector machine techniques. Most importantly, we compare the performance of utilized techniques, which the results will demonstrate in the fourth chapter.

Chapter 3

Methods in Data Collection and Data Analysis

This chapter is organized as follows. The data source is described in section 3.1. Section 3.2 introduces the tools and techniques. In section 3.3 data analysis procedures are discussed. Then, section 3.4 presents the application of machine learning algorithms. Figure 3.1 illustrates the data collection and methods overview, which will be discussed in this chapter.

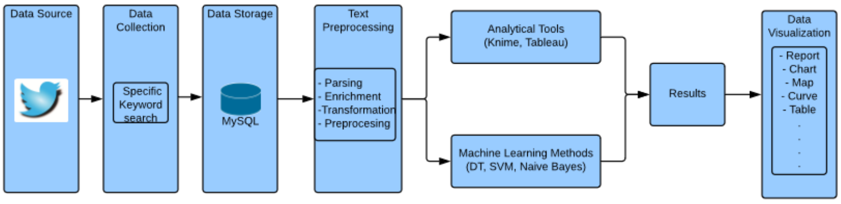


FIGURE 3.1: Data and methods overview.

3.1 Data Source

In Twitter, users are restricted to writing messages of no more than 140 characters likes a short message. These small messages create important information dissemination in the network and lead Twitter to become a successful social network for content dissemination. There are different ways to access Twitter data; in this project, the data is collected via the Twitter Search API. It was found in REST architecture, which refers to a collection of network design rules that determine resources and ways to address and access data. By allowing third-party developers have access to its API. The REST API lets access to the nouns and verbs of Twitter such as User Profile, Timeline, Tweets, Followers, User language, user location and so on. In order to get the Twitter search API, four keys are necessary: the customer secret, consumer key, access token and access token secret. In this project by setting up these parameters in the Knime can be connected and collected relevant project tweets.

One of the issues faced in gathering the tweets is the rate-limiting restriction imposed by Twitter i.e. 350 requests per hour for registered users and 150 for anonymous users. Therefore, I collect a set of tweets and since Twitter only makes the most current tweets available, it is not guarantee that the sample is the perfect representation of the population of tweets for this research. In addition, another issue is some gaps between the data collection period since Twitter only makes the last 6-day tweets available to the public. Therefore, back in time and capture them is not possible.

3.2 Tools and Techniques

A wide range of tools both open source and commercial is now available to get the first impression of a special social media channel. Tools are developed for any of the popular social media such as Twitter, Facebook, Google, and YouTube to provide an overview of the mentioned social media channels. These tools are designed as websites or web application components and serve their interface to a cloud-based application that collects the data. Knime Analytics platform and Tableau are the tools we use for analyzing data, visualization, and prediction of the pandemics.

Knime Analytics Platform is an open source software for data-driven innovation and helping find the hidden potential in data. It has more than 2000 modules, hundreds of examples and widest choice of advanced algorithms. It is a perfect toolbox for data analytics, reporting, and integration. In this project, Knime has been adopted due to the variety of features for data analysis, text processing, and machine learning. For example, data preprocessing has three main steps: extraction, transformation and data loading, which Knime does all three. It provides a user with a graphical interface to let them for assembly of nodes for data processing and easy to add plugins. In Knime each unit of computation is named node. Each node can be interconnected with other nodes and formed a workflow. The initial requirement is a source of input data, which can be real data acquired through a test or simulated data to validate the logic of node. Nodes can drag from the node repository and put them on workflow. In order to indicate the flow of data, draw an arrow from an output port of a node and import node of the next node in the flow [49].

Tableau is an interactive data visualization and powerful business intelligence tool. It enables non-technical users to interactive and apt visualization in form of worksheets or dashboard to obtain insight data insights for better development. In this project, we demonstrated Geo-location of tweets by using Tableau. In addition, sentiment analysis, text classification, and three different machine-learning techniques have been applied for data analysis and pandemic prediction [50].

3.3 Data Analysis Procedures

The Knime Text processing plugin was developed to parse and process textual and transform data into numerical data. It is a combination of Natural Language Processing (NLP), information retrieval and text mining. This plugin allows for the parsing of texts available in different formats (e.g. Microsoft Word, PDF, Xml, and Database).

It is feasible to determine and label various types of named entities and enriching the documents in a semantic way. In addition, documents can be filtered (e.g. RegEx filter, stop word, column filter), stemmed by stemmers for different languages and preprocessed in different ways. Frequencies of words can be computed, keyword or hashtags can be extracted, and documents can be visualized such as tag clouds. Moreover, the processed text data can be transformed into vector spaces and represent numerical data.

Finally, the data is ready for analysis tasks and applying machine learning techniques.

In order to reach more consistent and better results more disciplined needs for handling of data. Generally in this project, the process for making data ready for analysis and machine learning algorithm can be summarized in the below steps:

1. **Data Parsing**:

   The data is parsed by using Database connector and reader nodes. These nodes allow us to reach the collected data with the relevant keyword search ("pandemic" and "#pandemic"), which has been stored in MySQL database. In addition, the outputs can save in different format by writer nodes such as Excel Writer and CSV Writer. Moreover, by utilizing the reader nodes can read and load the data for other workflows. 3.2 shows the simple workflow for parsing the data.

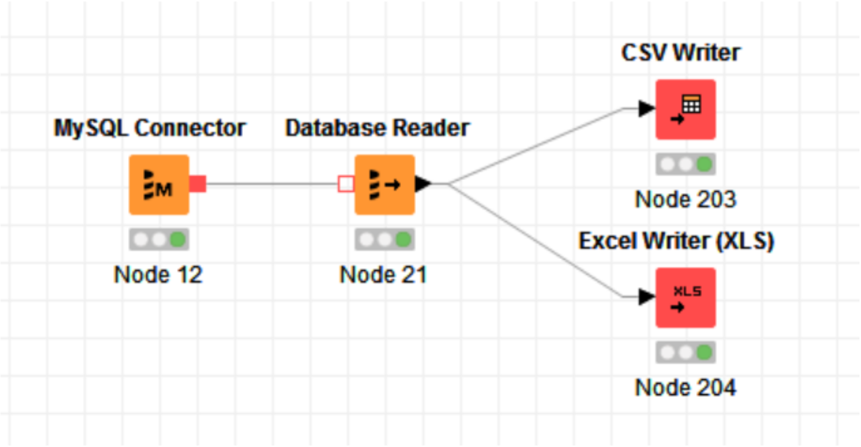


FIGURE 3.2: Data Parsing Simple Workflow.

2. **Enrichment and Labeling**:

   All enrichment nodes need a data table including exactly on the column of document cells and return data table with exactly one column of document cells. In this step, each node starts to analyses the document in a way that it mixes multiple words to terms, and then utilizes specific tags to these terms. A created document after this tagging process includes more information than a document before. Due to the further analysis, the collected data is required to specify with a label.

TABLE 3.1: Categories of labeled data

Category	Keyword
Pandemic_Related	epidemic, fever, infection,...
Communicable Disease	Flu, HIV, AIDS, Influenza, H1N1,...
Not relevant	Others except for two first categories

TABLE 3.2: Data label result.

Row	Tweet	Category
1	Scientists uncover new aspect of the flu virus and how it interacts with antibodies in the lungs flu influenza	Communicable Disease
2	Researchers discovered more than 200 previously unknown viruses in a category whose members cause illnesses such as flu	Communicable Disease
3	We need a universal flu vaccine before the next pandemic strikes	Pandemic_related
4	WHO knows this because they will start it World health chief warns a pandemic that will kill millions star	Pandemic_related
5	PS4Trophies Does anyone have any recommendations on a good cooperative board game that plays well with 2 people We've been playing	Not relevant
6	Geo-Engineering,Manufacturing Global Warming Weather War Fare and 20 reduction of,sunlight,Impacts Vitamin D	Not relevant

Labeled data is a set of samples that have been tagged with one or more labels. Table 3.1 represents the labels, which are applied in this project. The Pandemic_Related category is included all words, which are related to project keyword search such as infection, epidemic, outbreak, etc. The communicable diseases are caused by microorganisms such as bacteria, parasites, and viruses, which can burst directly or indirectly from one person to another [51]. This category is contained HIV, AIDS, Influenza, etc. The rest tweets, which were not part of two first categories counted as Not relevant. Table 3.2 demonstrates the some samples of labeled data.

This labeling process has been handled by Java snippet node in order to create a new document for further analysis. Java snippets can be applied to implement methods that are not available as native Knime nodes. It lets an interesting and powerful new method to work with extensions, which are compatible with the new functionality. Therefore, the Java snippet is programmed due to specify the categories.

In addition, in the text processing part, the OpenNLP (Natural Language processing) tokenization has been applied to words. This method breaks up the sentences and words into smaller parts [52].

3. **Transform Data**:

   The next step after enriched documents is preprocessing data, i.e. filtered, punctuation erases, etc., however, data transformation is needed in advance. In this step, the structure of the data has to be transformed into a Bag of Words (BoW). This model is a way of presenting text data when modeling text with machine learning algorithms. It has great success in issues such as language modeling and document classification [53]. Therefore, this node has to be used. The BoW node needs a list of documents as input data, including exactly one column of document cells. The input documents and its terms are transformed into a bag of words structure. The output file includes two columns, the first is the term column and the second one is the document column. The terms tags are listed in brackets after the terms. In fact, it is used the tokenized words for each observation and find out the frequency of each token.

4. **Data Preprocessing**:

   A technique applied to convert the raw data into clean dataset is called data preprocessing. Thus, specific steps are executed to convert the data into the clean dataset. Once a bag of words is created preprocessing nodes such as filtering, regular expression (RegEx) filter, and column filter, etc. can be applied. A RegEX is a string of characters used to make a search pattern. It also provides a various range of special characters that can be assembled into filters. In this project, RegEX filter has been applied to the tweets column to remove all the punctuations and URLs such as "?!/()=#:;”.

Moreover, the non-English languages such as Japanese and Korean, which were shown like question marks in the tweet column, were discarded. In addition, by creating Wild-card tagger node, creating new BoW, and new term frequency node the hashtags have been extracted from the tweets. BoW model used in NLP and is aimed to categorize documents. The idea is to analyze and classify various "bags of words", and by matching the different categories, can identify which "bag" a certain block of text comes from [53]. In this project, different filters simply filter rows according to the terms and the filter criteria in preprocessing Meta node. For example, the “Stop word filter” node filters terms that are stop words.

The current set of filters allows to filter non-relevant terms and reduce the size of the bag of word. In addition, the used preprocessing nodes in this project can modify terms, such as “Porter stemmer” that reduce the words to their stems by the Porter algorithm. This is a process for deleting the commoner morphological and inflexional endings from words in English. To stem all terms included in the input document, the kuhlen stemming algorithm or Porter stemmer can be used. The Kuhlen stemmer can be used on English documents only [54]. Moreover, the “Snowball stemmer” node has been applied due to other tweet languages. It is a small string processing language modeled for making stemming algorithms using in Information Retrieval. The snowball compiler translates a relevant script into another language [55].

Figure 3.3 presents a part of the preprocessing workflow, which is located in a Meta node at the beginning of the other used workflows.

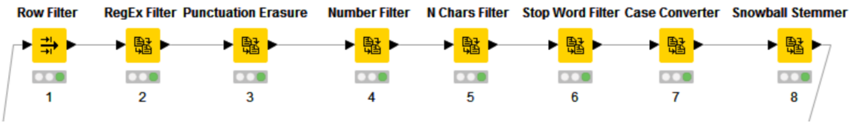


FIGURE 3.3: Sample of used data preprocessing workflow.

Functionalities of nodes have been described below:

1. Row Filter: This node excluded the non-English rows by determining the languages in the tab of "use pattern matching".

2. RegEX filter: This node removed the special characters and URLs such as "~_#():;+" from all tweets. This can be set by regular expression section in node.

3. Punctuation Erasure: This node deleted all the punctuation such as "?!." and append a column named "Preprocessed document". In fact, the output of this step is the cleaned tweets.

4. Number filter: All the terms included in the input documents, which consist of digit, decimal separation is set in this node.

5. N chars filter: It refers to filtering all terms included in the input documents with less than the specified number N characters, which is define 3 in this work.

6. Stop word filter: It is a general strategy for defining a stop list is to sort the terms by collection frequency. It takes the most frequent terms as a stop list, the members of which are then discarded during the indexing.

7. Case converter: It converts all terms of input documents to lower or upper case, which lower case is selected in this work.

8. Snowball stemmer: It changes tweet languages. It is a small string processing language modeled for making stemming algorithms using in information retrieval. the Porter stemmer is selected in relevant nodes.

3.4 Application of Machine Learning Algorithms

Machine learning is one of the most important techniques gaining the interest of researchers because of its accuracy and adaptability. However, it depends on how data ingestion is feed the model. It denotes a broad class of computational methods, which help to extract a model of a system from the simulation of this system. The purpose of such models is predicting the behavior of such system in some unobserved situations or aim to understand its previously observed behavior. In machine learning technique, the key to the accuracy of a classifier is the selection of the proper features. There are two groups of learning supervised and unsupervised, which in this project the collected data is supervised. Supervised learning refers to the subcategory of machine learning models, which proceed models in the form of input-output relationships. The purpose of supervised learning is to recognize a mapping from some input variables to some output variables, which are called attributes or features. In this project, three different predictive algorithms have been employed, which are presented as follows:

- **Decision Tree algorithm**

  Decision Tree (DT)is a simple but powerful machine learning algorithms, which has been used for classification issues successfully. The decision tree technique utilizes a supervised approach for classification, where the leaves on the tree show classifications and branches represent conjunctions of features that lead to classification.

  In simplest form, a decision tree merges several binary tests in a tree structure. Moreover, it is built by using training data set to decrease the average depth of each path from the root to the leaf node. This technique of classification has been used for a wide range of applications in bio-informatics [56]. There exist many different of decision tree learning methods. The main idea behind tree algorithm is to find a simple tree, which has good predictive performance on the learning sample.

The enumeration of all feasible trees is intractable, thus most tree algorithms are based on the heuristic. One of the most common heuristics is a greedy top-down recursive partitioning approach. It begins with a single node tree regarding the finished learning sample and recognizes a way to split this node by choosing a test among a set of candidate tests. Then, the algorithm proceeds recursively to split the substitute for this node. The entire process results in a partition of the learning sample into smaller and smaller subsets. The expansion of a branch is stopped when some stop-splitting rule applies. Finally, each terminal node of the tree is labeled with a prediction (vector of class probabilities and class name), which is computed based on the subset of objects that reach this node.

To split a node, a score measure is defined and the test between the complexes of the candidate, which realized the best score, is selected. The score evaluates the ability of the test to decrease the impurity of the class within the local learning sample. Most of the score measures are like the following form [57] Eq.(1):

$$Score(S,T) = I(S) - \sum_{i=1}^{P} \frac{N_i}{N} I(S_i) \tag{3.1}$$

Where T denotes the candidate test, S is the local learning sample of size N which is split by T into p subsets Si of size N_i (i = 1, . . . , p), and I(S) measures the impurity of the output in the sample S.

In the context of a classification issue with m classes, two popular measures of impurity are Shannon's (or logarithmic) entropy:

$$I(S) - \sum_{c=1}^{m} \frac{N_c}{N} log \frac{N_c}{N} \tag{3.2}$$

, and Gini's (or quadratic) entropy:

$$I(S) = \sum_{c=1}^{m} \frac{N_c}{N} 1 - \frac{N_c}{N} \tag{3.3}$$

Where N is the size of the local learning sample S and N_c is the number of objects of output class c in S. These score measures give very similar results in practice and have been widely used in the context of decision tree or normalized in different ways [58].

- **Support Vector Machine Algorithm**

  Support Vector Machine (SVM) is a discriminative based classifier, which has been applied to many problems such as image processing and face recognition, text classification, spam detection and many more issues in social media [59]. In fact, SVM is set of related supervised learning models used for classification and regression.

  It is an alignment-free method, which utilizes vectors to classify objects. It does not depend on multiple alignments, therefore, it can avoid errors, if any, in multiple alignment files. This can be counted as one of the SVM advantages. Another advantage of this method is that it can classify sequences with low similarity or even with very short lengths.

  The method described the idea of splitting points of various classes in clouds of points with a line, which is optimally distanced from the classes. A special property of SVM is that minimizes the empirical classification error and maximize the geometric margin simultaneously. Therefore, SVM called maximum margin classifiers. If the SVM is used for classification then finding the proper function is a kind of issue. However, SVM supports different functions (HyperTangent, Polynomial, and RBF).

  However, computing between each class will increase the runtime. In the real world application, finding the perfect class for more than thousands of training data set is time-consuming. Therefore, regularization parameter is needed. The selection of the appropriate kernel function has been made first in order to find the proper model configuration. Among three offered kernels in the Knime Analytic platform (Hyper-Tangent, Polynomial, and RBF), the Polynomial kernel performed best. The model presentation function $f(X)$ with polynomial kernel is formulated as follows [60] Eq. (3-6).

$$f(x) = \sum_{i=0}^{m} \omega_i K((X_i, Y_i)^p + b. \tag{3.4}$$

Optimization formula is as follows:

$$\min\left(\frac{1}{2}||\omega||^2 + C\sum_{t=0}^{t}(\xi_i + \xi_j^*)\right), \tag{3.5}$$

subject to:

$$\begin{cases} 1(y_j) - < \omega, x_j > -b \leq \xi + \xi_j, \\ \langle \omega, x_j \rangle + b - y_j \geq \xi + \xi_j^*, \\ \xi + \xi_j^* \leq 0. \end{cases} \tag{3.6}$$

K denotes kernel function, exponent p the order of polynomial, x_i and y_i input and output variables, index m is the number of attributes, ξ_i, ξ_i^* demonstrate the slack variables used to cope with possible constraints in the optimization issue. Another parameter is gamma and Bias, which are defined 1.0 and show the influence of a single training sample with low values meaning "far" and high values meaning "close".

- **Naïve Bayes Algorithm**

  A naïve Bayes is a classification algorithm for binary (two-class) or multi-class classification problems. A Naïve Bayes classifier is a simple generative classifier according to the application of the Bayes' theorem with the powerful hypothesis that the features are highly independent.

This technique is easiest to realize when explained using categorical or binary input values. It is called Naïve Bayes or idiot Bayes due to the calculation of the probabilities for each assumption are simplified to make their calculation tractable. Contrary to the Naïve design and this simplified assumption, Naïve Bayes' classifiers have worked great in many complex real situations such as text classification [61], sentiment classification [62], and spam detection [63].

This model works very well in the problems, which the features are independent. It calculates the probability of an object belonging to each of the classes. Given a class label C for a data class, which is shown by a feature vector x (x_1,...x_f). According to the Bayes rule we can calculate class posterior probability P(c|X) as follows [64] Eq. (7-10):

$$P(c|X) = P(c|X_1, ..., X_f) = \frac{P(C)P(X_1, ..., X_f)|c}{P(X_1, ..., X_f)} \tag{3.7}$$

Then, by using the naive independence assumption, we can write:

$$P(x_i|c, x_1, ..., x_i - 1, x_i + 1, ...x_f) = P(x_i|c) \tag{3.8}$$

For all i=1...f using this equation the class posterior can be written as:

$$P(c|X_1, ..., X_f) = \frac{P(C)\prod_{i=1}^{n} P(x_i|c)}{P(X_1, ..., X_f)} \tag{3.9}$$

Since P(x_1,...,x_f) is constant for all classes, we can apply the following classification:

$$\hat{c} = \arg\max_{i} P(c) \prod_{i=1}^{n} P(x_i|c) \tag{3.10}$$

The class with highest posterior probability would be selected as the class label for a given sample.

The mentioned machine learning techniques are aimed to identify relationships or associations in biological data, find the similar group of elements, analyze, and predict pandemic outbreak. The main idea behind these methods is to learn the theory automatically from the data, through a process of inference, model fitting or learning from samples. Most of the research domains are applied machine learning techniques in multiple sequence alignment, prediction, classification and expression analysis. The machine learning techniques are used for prediction and analysis and the proper technique is suggested in the next chapter.

Chapter 4

Results

This chapter presents in details how the dataset was collected and demonstrated the experiment results. The pandemic data has been collected from Twitter and various experiments have been done. Then, it explains how to configure the machine learning models that are considered for this project and explain the analytics task on data. This chapter is organized as follows: Section 4.1 describes the datasets that we used for this work and their collection methods. We further explain how data migrated into a relational database. In the section, 4.2 the machine learning techniques configuration is addressed. Then in the section 4.3, the prediction experiment evaluation and comparison of models with each other are discussed.

4.1 Dataset Collection

As mentioned before in chapter 3, there are different ways to access Twitter data. In this project, we collect data via Twitter Search API. The Knime Twitter nodes permit users to search for Tweets on Twitter, retrieve information about users, and post tweets via the Knime. In order to use the twitter nodes having a Twitter account, creating a Twitter App and set up Twitter API connector node is needed, which are summarized as following steps:

1. Log in to Twitter account in https://apps.twitter.com.

2. Select "Create New App".

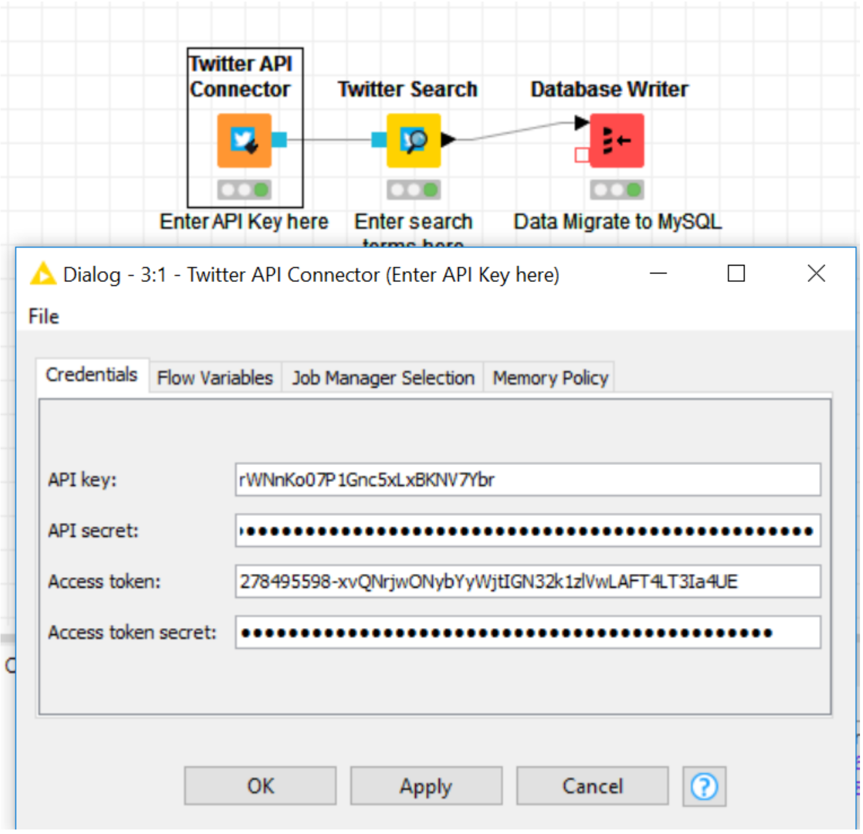
3. Fill out the form but it is optional.

4. Click "API Keys".

5. Click "Create my access token".

6. Set up the Twitter API connector node by API key, API secret, Access token and Access token secret.

In the next step, by applying the Twitter search node and defining the search key, tweets field, and user field all necessary data is collected. Then, the collected data is migrated to MySQL database by database writer node. Figure 4.1 shows the Twitter search API configuration and data collection workflow:



API key: rWNnKo07P1Gnc5xLxBKNV7Ybr

API secret: ••

Access token: 278495598-xvQNrjwONybYyWjtIGN32k1zlVwLAFT4LT3Ia4UE

Access token secret: ••

FIGURE 4.1: Twitter API configuration in Knime.

For our empirical analysis, we collected tweets with the keyword of the "Pandemic" or "#pandemic" from February to May 2018. Figure 4.2 demonstrates the collected data that is saved in a table on database, and named to Twitter.

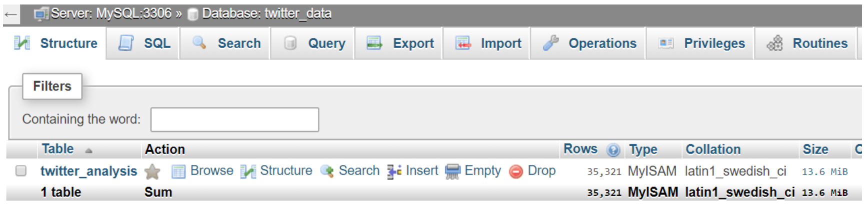
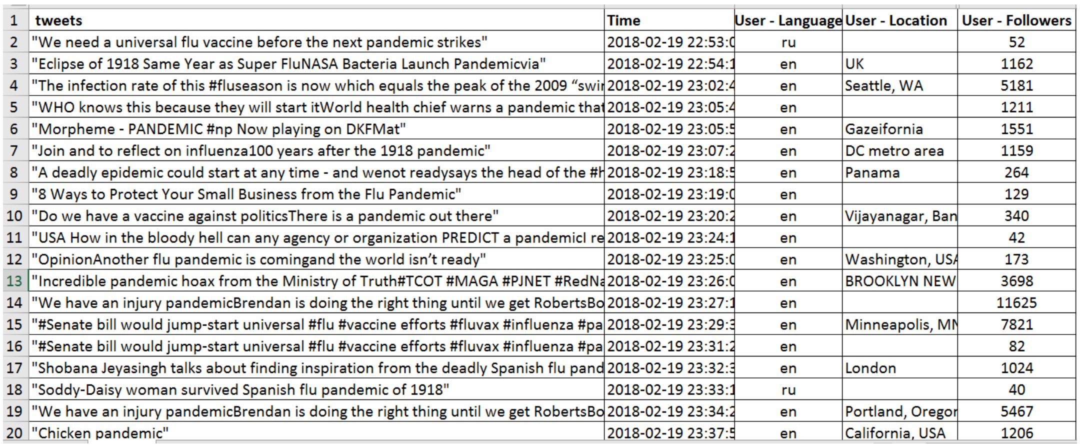


	Table	Action	Rows	Type	Collation	Size	C
□	twitter_analysis	Browse Structure Search Insert Empty Drop	35,321	MyISAM	latin1_swedish_ci	13.6 MiB	
	1 table	Sum	35,321	MyISAM	latin1_swedish_ci	13.6 MiB	

FIGURE 4.2: MySQL Database Table.

The MySQL table contains 35,321 data that includes pandemic or #pandemic Tweets. We extracted all the tweets and user's information such as user name, user id, user location, tweet id, time, country, URLs, etc but some filtering will be applied based on the project requirements. Figure 4.3 presents part of the collected data from the Twitter:



1	tweets	Time	User - Language	User - Location	User - Followers
2	"We need a universal flu vaccine before the next pandemic strikes"	2018-02-19 22:53:0	ru		52
3	"Eclipse of 1918 Same Year as Super FluNASA Bacteria Launch Pandemicvia"	2018-02-19 22:54:1	en	UK	1162
4	"The infection rate of this #fluseason is now which equals the peak of the 2009 “swir	2018-02-19 23:02:4	en	Seattle, WA	5181
5	"WHO knows this because they will start itWorld health chief warns a pandemic that	2018-02-19 23:05:4	en		1211
6	"Morpheme - PANDEMIC #np Now playing on DKFMat"	2018-02-19 23:05:5	en	Gazeifornia	1551
7	"Join and to reflect on influenza100 years after the 1918 pandemic"	2018-02-19 23:07:2	en	DC metro area	1159
8	"A deadly epidemic could start at any time - and wenot readysays the head of the #h	2018-02-19 23:18:5	en	Panama	264
9	"8 Ways to Protect Your Small Business from the Flu Pandemic"	2018-02-19 23:19:0	en		129
10	"Do we have a vaccine against politicsThere is a pandemic out there"	2018-02-19 23:20:2	en	Vijayanagar, Ban	340
11	"USA How in the bloody hell can any agency or organization PREDICT a pandemicI re	2018-02-19 23:24:1	en		42
12	"OpinionAnother flu pandemic is comingand the world isn't ready"	2018-02-19 23:25:0	en	Washington, USA	173
13	"Incredible pandemic hoax from the Ministry of Truth#TCOT #MAGA #PJNET #RedNa	2018-02-19 23:26:0	en	BROOKLYN NEW	3698
14	"We have an injury pandemicBrendan is doing the right thing until we get RobertsBo	2018-02-19 23:27:1	en		11625
15	"#Senate bill would jump-start universal #flu #vaccine efforts #fluvax #influenza #pa	2018-02-19 23:29:3	en	Minneapolis, MN	7821
16	"#Senate bill would jump-start universal #flu #vaccine efforts #fluvax #influenza #pa	2018-02-19 23:31:2	en		82
17	"Shobana Jeyasingh talks about finding inspiration from the deadly Spanish flu pand	2018-02-19 23:32:3	en	London	1024
18	"Soddy-Daisy woman survived Spanish flu pandemic of 1918"	2018-02-19 23:33:1	ru		40
19	"We have an injury pandemicBrendan is doing the right thing until we get RobertsBo	2018-02-19 23:34:2	en	Portland, Oregon	5467
20	"Chicken pandemic"	2018-02-19 23:37:5	en	California, USA	1206

FIGURE 4.3: Sample of Collected Twitter data.

4.2 Twitter data Analysis

Obtain results in this section have been achieved based on the described procedures in section 3.3. One of the primary visualization tasks in this project is extracting the Hashtags from the tweets. Using Hashtags make it easier for viewers to find and follow the discussions regarding the events. Tweets with the Hashtags can double the times of more engagement rather than tweets without. Moreover, Hashtags let social media sites and users to categorize the content of their Tweets, which provides a clear picture of what the Tweets are about. Therefore, in data preprocessing workflow (See Appendix A) a meta node is applied. It contains a set of filters, BoW creator, Wildcard tagger nodes, and document data extractor. This Meta node named as Extract Hashtags. Figure 4.4 represents the Hashtags extracted result as a cloud in order to get a clearer overview of Tweets in this project.

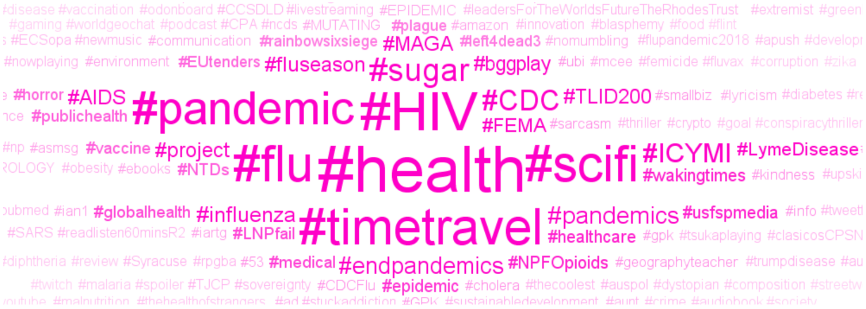


FIGURE 4.4: Extracted Hashtags from user Tweets.

With respect to the research questions, this part of the project aimed to visualize the most influential users' network, Retweet network, and top 20 Tweeters. In this regards, the analyzing Twitter data workflow is created, which includes Text processing, Network Analysis, and summary statistics. I explore further the data by analyzing the network of the influential users. They are important to the analyzing and managing of propagation in tweeter social networks. If a user increases his or her usage and the people connected increase their usage as well, it can be proposed that this identifies this user as influential.

With this data and notion of influence, the influential users whose had the most significant impact on the network, have been identified. Due to the huge number of users in the network, the influential users have been restricted. They selected based on their number followers and number of retweets. However, having many friends or followers does not make users influential. Figure 4.5 shows the top 50 influential users network. The result indicates the power of the users to propagating a news via social media such as Twitter. Moreover, It should be considered that user image profile is protected and shown as hidden due to the General Data Protection Regulation (GDPR).

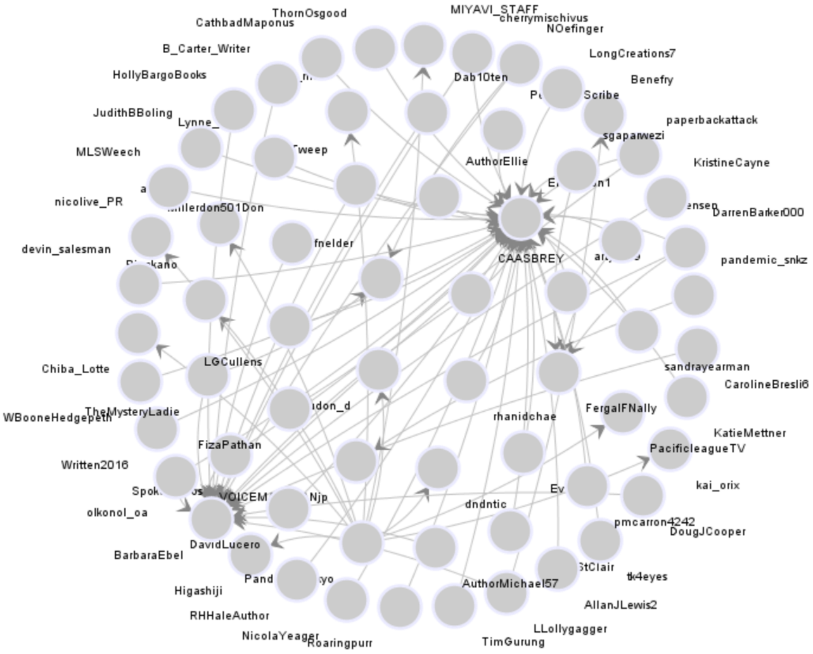


FIGURE 4.5: Influential User's network.

TABLE 4.1: Top 20 Tweeters.

Row	User	Count(Tweet)	Sum(Retweeted)
1	Pandemic_ukyo	22	6299
2	paperbackattack	16	899
3	AllanJLewis2	9	676
4	DavidLucero	7	594
5	Spokenamos	19	1023
6	Millerdon501Don	22	858
7	FizaPathan	7	652
8	EvaFarohi	7	594
9	DarrenBarker000	19	941
10	TudorTweep	17	835
11	WBooneHedgepeth	8	652
12	JudithBBoling	8	594
13	CAASBREY	16	899
14	pmcarron4242	12	782
15	cherrymischivus	13	617
16	Lynne_Jean	7	594
17	olkonol_oa	16	899
18	rhanidchae	9	740
19	AuthorEllie	7	594
20	Roaringpurr	7	594

One of the important functionality offered by Twitter is the Retweet. It lets users reposted a received tweet to their followers and networks. In the next step, the 20 top tweeters have been recognized with respect to their number of Tweets and sum of their Retweeted. Table 4.1 represents the 20 top tweeters. This result in finding the tweets of users who are not being directly followed. It represents the diffusion of information to users, which are not targeted (i.e, followers of their followers). The relevant workflow is attached in Appendix A.

In order to find the people sentiments forward the pandemic diseases, a sentiment analysis workflow has been deployed by using the Knime analytics platform. Data labeling is one of the requirements for data analysis. In this step, the data is labeled by using a Java snippet program. Then, it is categorized into 3 groups:

1. Pandemic_Related: This group consists of all words, which are related to project keyword search such as infection, epidemic, outbreak,etc.
2. Communicable Disease: This groups includes diseases, which are happened by microorganisms such as bacteria, viruses, HIV, AIDS, Influenza, etc.
3. Not relevant: This group covers the tweets, which are not part of two first groups.

Figure 4.6 illustrates the proportion of these three categories.

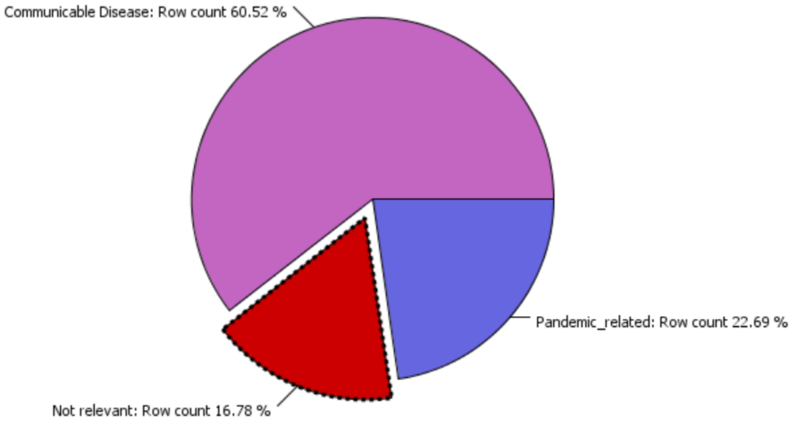


FIGURE 4.6: Data Category.

The pie chart result indicates that the 22.69 percentage of people tweets are related to the pandemic diseases such as infection, 60.52 percentage of tweets referred to Communicable diseases such as HIV, AIDS, and etc, and 16.78 percentage of tweets are not relevant to our research area. Therefore, the "Not relevant" group has been ignored from the data analysis and prediction.

By using the Multi Perspective Question Answering (MPQA) Corpus node reader sentiment tagging is applied to the data. The MPQA Corpus has included news articles from a wide range of news sources manually commentated for positive or negative opinions (i.e., beliefs, sentiments, emotions, speculations, and so on). This node reads the positive and negative words from relevant files and matches them with the relevant sentiments [65]. This procedure is named sentiment tagging. Afterward, preprocessing and transformation steps are performed and the positive and negative words have been extracted from the Tweets. Figure 4.7 presents the sentiment analysis of tweets. As can be seen, the positive words printed with green color and negative ones with red color. They are categorized based on the tweets sentiments, which are accordance with MPQA Corpus files. It can be concluded that majority of the people have worry regarding the pandemic outbreak. The relevant workflow is attached in Appendix A.

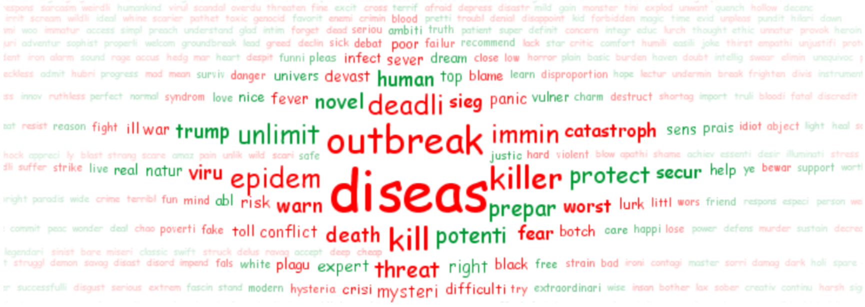

respons sarcasm weirdli humankind virul scandal overdu threaten fine excit cross terrif afraid depress disastr mild gain monster tini explod unwant quench hollow
irrit scream wildli ideal whine scarier pathet toxic genocid favorit enemi crimin blood pretti troubl denial disappoint kid forbidden magic time evid unpleas pundit hilari down
woo immatur access simpl preach understand glad intim forget dead seriou ambiti truth patient super definit concern integr educ lurch thought ethic unnatur provok heroin
adventur sophist properli welcom groundbreak lead greed declin sick debat poor failur recommend lack star critic comfort humili easili joke thirst empathi unjustifi
iron alarm sound rage accus hedg mar heart despit funni pleas infect sever dream close low horror plain basic burden haven doubt intellig swear elimin unequivoc
admit hubri progress mad mean surviv danger univers devast human top blame learn disproportion hope lectur undermin break frighten divis
innov ruthless perfect normal syndrom love nice fever novel deadli sieg panic vulner charm destruct shortag import truli bloodi fatal discredit
resist reason fight ill war trump unlimit outbreak immin catastroph sens prais idiot object light heal
hock appreci ly blast strong scare amaz pain unlik wild scari safe justic hard violent blow apathi shame achiev essenti desir illuminati stress
suffer strike live real natur viru epidem diseas killer protect secur help ye bewar support
paradis wide crime terribl fun mind abl risk warn prepar worst lurk littl wors friend respons especi person
commit peac wonder deal chao poverti fake toll conflict death kill potenti fear botch care happi lose power defens murder sustain
legendari sinist bare miseri classic swift struck delus ravag accept deep cheap
struggl demon savag disast disord impend fals white plagu expert threat right black free strain bad ironi contagi master sorri damag dark holi spare
successfulli disgust serious extrem fascin stand modern hysteria crisi mysteri difficulti try extraordinari wise insan bother lax sober creativ continu harsh

FIGURE 4.7: User Sentiment Analysis from their Tweets .

4.3 Data Analysis

With respect to our research question, we are aimed to locate the geographical origin of the Twitter messages, which can be analyzed from number of Tweets and Retweets expressing concern on disease in different part of the world. Figure 4.8 illustrates the user distribution in all over the world during the project data collection.

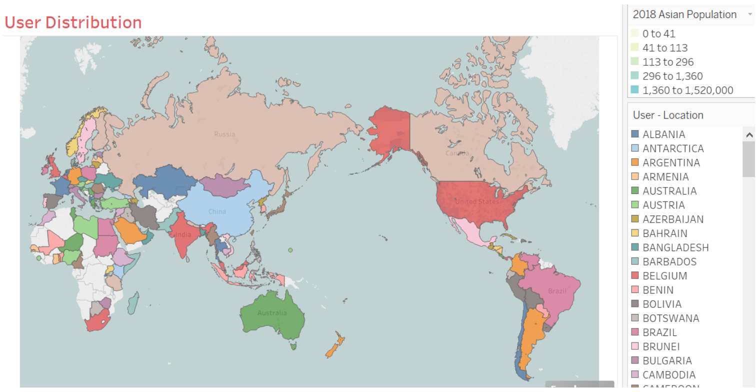


FIGURE 4.8: Twitter user distribution.

Figure 4.8 shows that most of the users were involved from various countries. It should be mentioned that some user locations were null or unknown, which they are filtered on our visualization.

With respect to our research questions, we are aimed to realize that "Are the pandemics topic that continuously feared by people or related to certain time? ". Therefore, the number of tweets is visualized per hour. Figure 4.9 presents the continuously tweets related to our data categories.

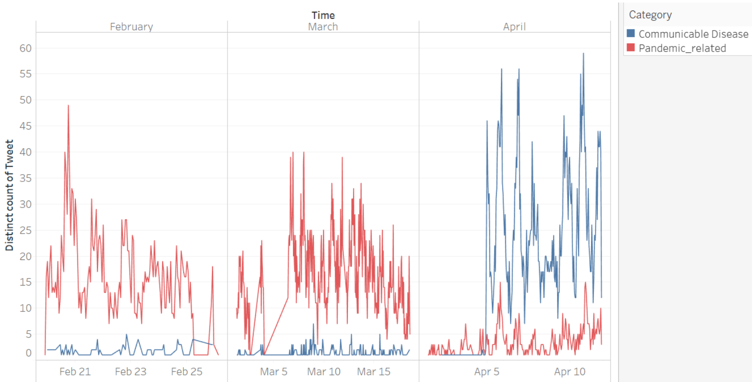


FIGURE 4.9: Number of Tweets Per Hour.

It can be concluded that people were more worried about pandemic issues during the months of February and March 2018. In addition, the communicable diseases were the hot topic in April 2018. In order to find more details of the people worries, the website of world health organization has been checked. According to the [66] and [67] total of 2189 laboratory confirmed cases of Middle East respiratory syndrome (MERS), containing 782 deaths were reported globally. The majority of these cases were happened in Saudi Arabia during February and March 2018. Moreover, during April 2018 people were worried about the communicable diseases such as Influenza in china, Yellow fever in Brazil, and Dengue in France [68].

Afterward, the number of Tweets and Retweets have located on the maps due to find which users from which countries had more concern about the pandemic and communicable diseases. The countries with highest number of Tweets and Retweets are demonstrated in the Figures 4.10 and 4.11 respectively.

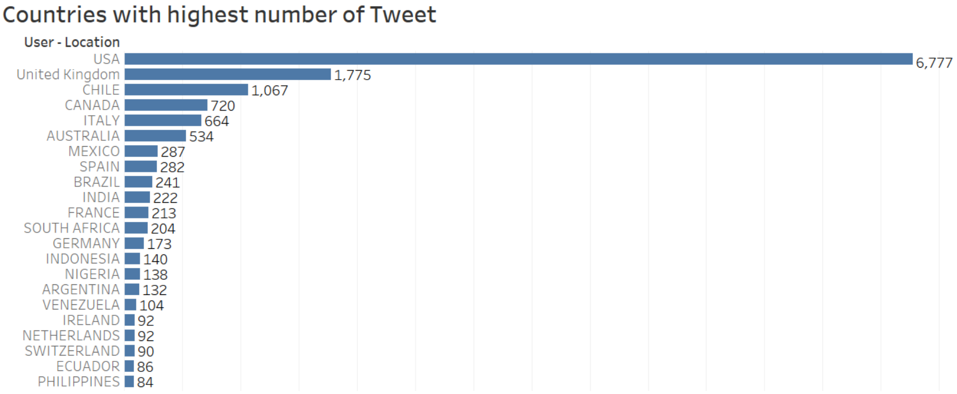


FIGURE 4.10: Countries with highest number of Tweet.

Figures 4.12 and 4.13 shows the diffusion maps of Tweets and Retweets per country.

It can be concluded that most of the developing societies have concern about the pandemic outbreaks. Also, Figure 4.14 shows the dashboard visualization of Twitter information based on frequency of tweets per country. By clicking Dashboard (*https://public.tableau.com/views/ Book2_21041/Dashboard1?:embed=y& display_count=yes*) the findings can be observed.

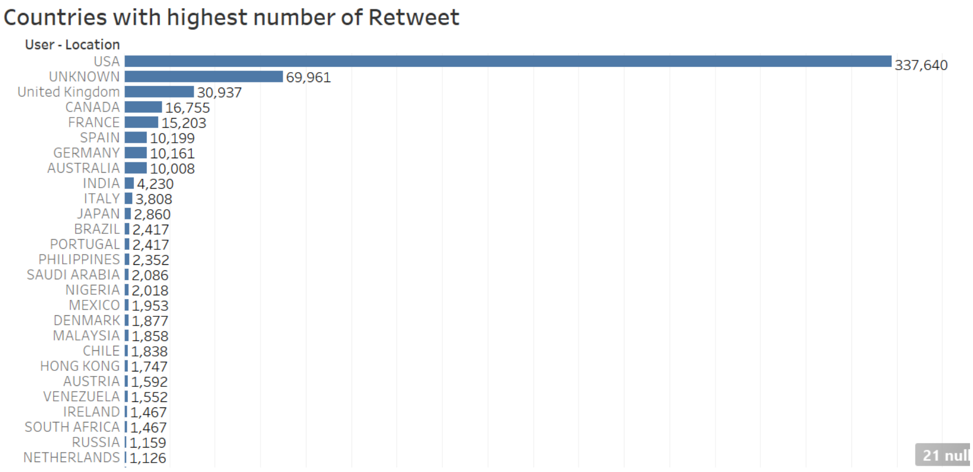


FIGURE 4.11: Countries with highest number of Retweet.

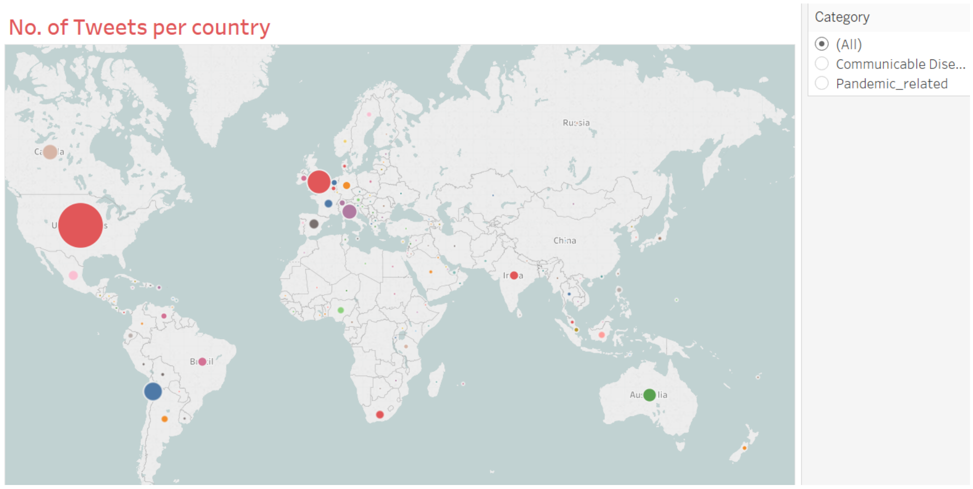


FIGURE 4.12: Number of Tweets Per Country.

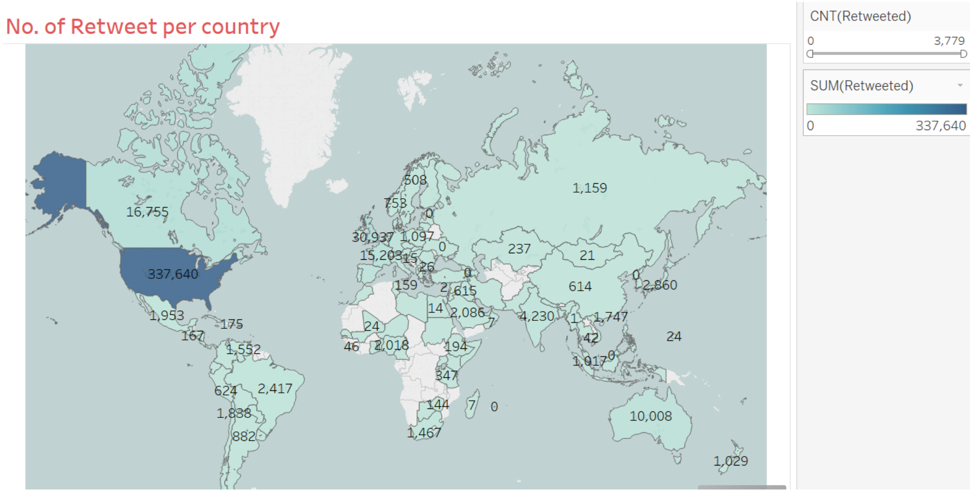


FIGURE 4.13: Number of Retweets Per Country.

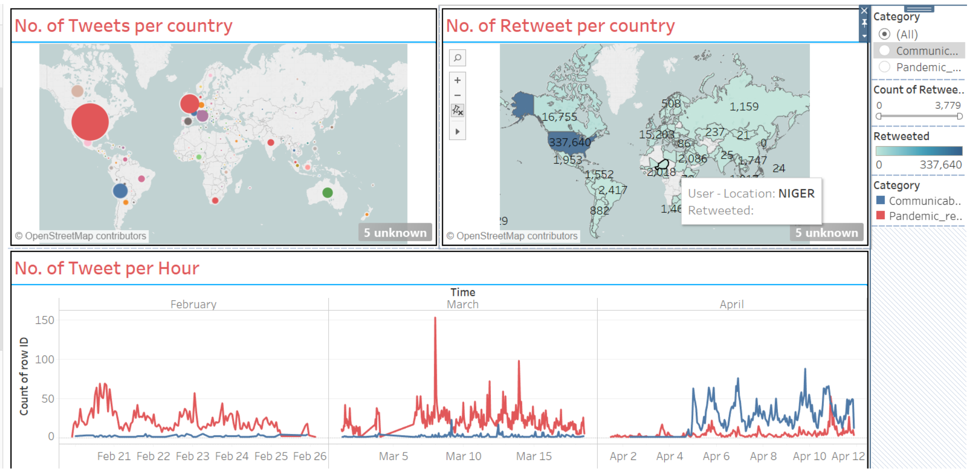


FIGURE 4.14: Data Dashboard.

4.4 Experiments

Our experimental design is summarized in Figure 4.15. In brief, the comparison analysis is conducted with three predictive models, Decision tree, Naïve Bayes, and SVM. These three models have been described precisely in chapter 3, however, the setup and implementation are discussed as follow:

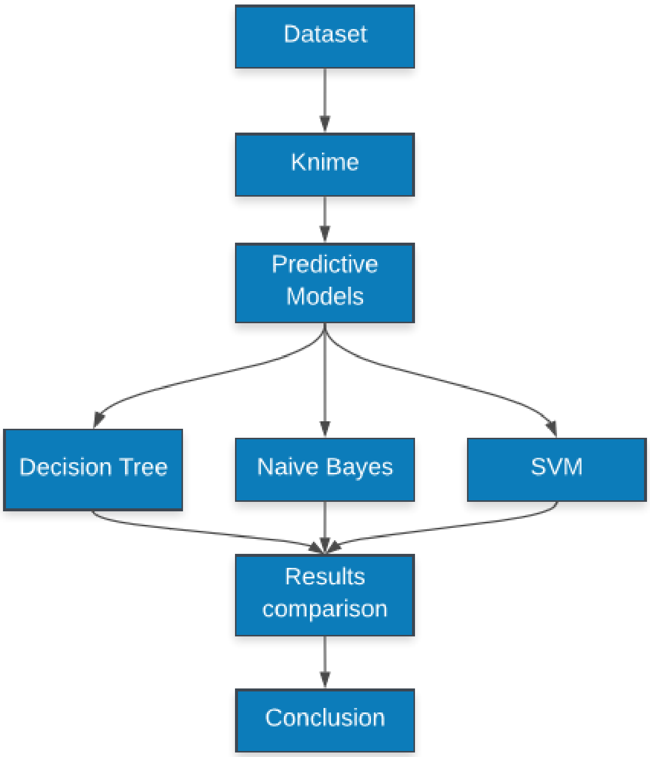


FIGURE 4.15: Overview of Prediction approach.

4.4.1 Implementation of techniques

The Knime Analytics platform is equipped with the decision tree predictive nodes in order to make a workflow for pandemic prediction. The methodology of the decision tree classification (i.e. Data collection, Data preprocessing, and transformation) has been applied and data is ready for training.

Figure 4.16 shows an overview of the decision tree workflow in the Knime. It consists of 5 steps. The data is parsed via the database connector and reader nodes. In the second step, document is transformed by making label for data and other necessary nodes such as row filter. The preprocessing is the third step, which data is cleaned by using different nodes. The fourth step is the data partitioning. In training set, the data is divided into 2 partitions with the ratio of 70% for the first partition and 30% for the second partition. The training sets are selected randomly. Then, the score is measured in the last step. It should be noted that different algorithms apply different metrics for measuring best and many alternative measures for the information gain [69].

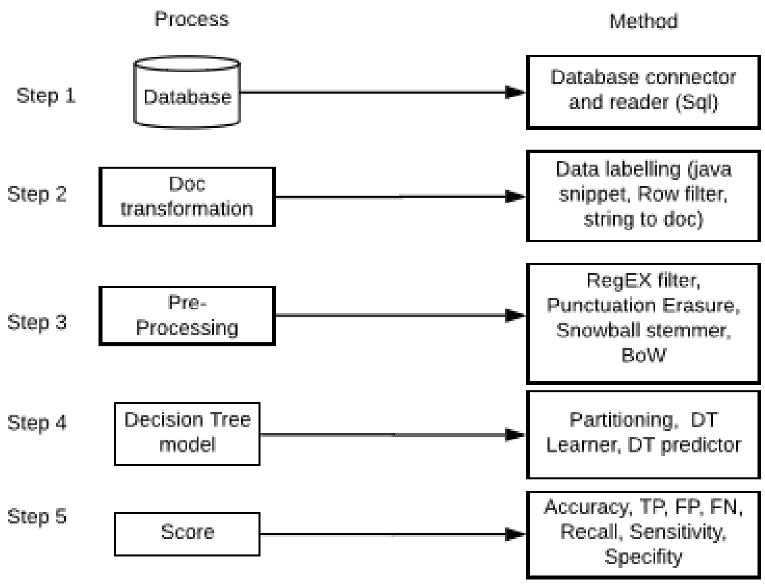

Process
Method
Step 1
Database
Database connector and reader (Sql)
Step 2
Doc transformation
Data labelling (java snippet, Row filter, string to doc)
Step 3
Pre-Processing
RegEX filter, Punctuation Erasure, Snowball stemmer, BoW
Step 4
Decision Tree model
Partitioning, DT Learner, DT predictor
Step 5
Score
Accuracy, TP, FP, FN, Recall, Sensitivity, Specifity

FIGURE 4.16: Decision Tree overview diagram.

The most popular one is Gini index, which is used in the CART (classification and Regression Trees) and it is also applied in this project. Moreover, there is a post pruning method to decrease the tree size and increase prediction accuracy. The most of the techniques, which are used in decision tree implementation, can be found in C4.5 program for machine learning [70], which has been described in chapter 3. The relevant workflow and part of the created decision tree are attached in appendix A.

SVM is another predictive model, which is used for pandemic outbreak prediction. The Knime Analytic platform is equipped with the SVM learner and predictor nodes. It also consists 5 steps and the first 3 steps are applied similar as decision tree method. In the step 4, the SVM learner node trains a support vector machine on the input data. The training data sets are splitted the same as decision tree method. This model supports different kernels (HyperTangent, Polynomial, and RBF) and SVM learner supports multiple class problems [71]. However, computing between each class will increase the runtime. In the real world application, finding the perfect class for more than thousands of training data set is time-consuming. Therefore, regularization parameter is applied.

In this step, the kernel and parameters are configured respectively. The overlapping penalty is adjusted on 1.0 and the polynomial parameter is worked based on the described formula in chapter 3. It is calculated separation line in the higher dimension, which is called kernel trick. The next parameters are gamma and Bias, which are defined 1.0 and shows the influence of a single training sample with low values meaning "far" and high values meaning "close" [72]. In the last step, the scores such as accuracy, TP, FP, recall, and sensitivity are computed, which the results will be demonstrated in this chapter. Figure 4.17 demonstrates an overview of SVM workflow. The Knime workflow is attached in Appendix A.

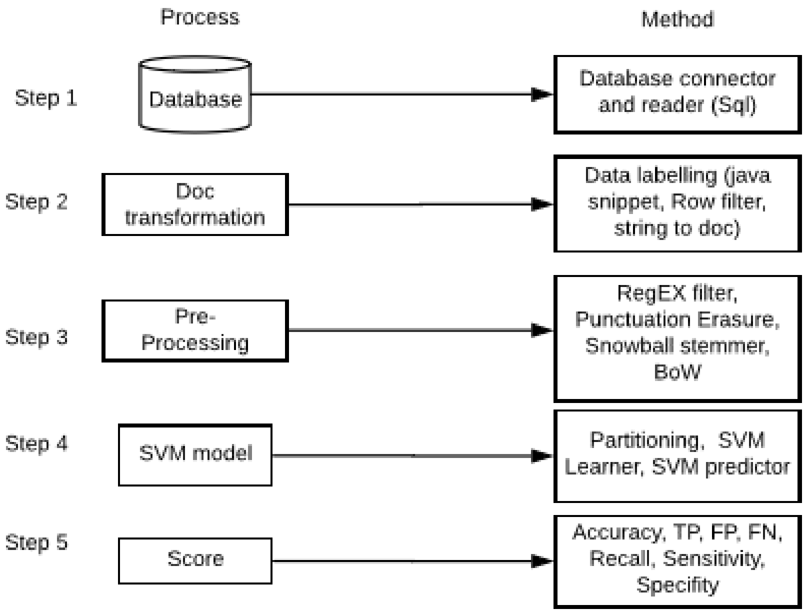

Process
Method
Step 1
Database
Database connector and reader (Sql)
Step 2
Doc transformation
Data labelling (java snippet, Row filter, string to doc)
Step 3
Pre-Processing
RegEX filter, Punctuation Erasure, Snowball stemmer, BoW
Step 4
SVM model
Partitioning, SVM Learner, SVM predictor
Step 5
Score
Accuracy, TP, FP, FN, Recall, Sensitivity, Specifity

FIGURE 4.17: SVM overview diagram.

The last predictive model is Naive Bayes model. The same as the previous models, the Knime Analytic platform is equipped with the Naive Bayes learner and predictor nodes. This workflow consists 5 steps. The three first steps are the same as the previous predictive models. The fourth step is data partitioning. Naive Bayes learner node creates a Bayesian model based on the training data sets, which is splitted the same as the other approaches. It calculates the number of rows per attribute value per class for nominal attributes and it is like a Gaussian distribution for numerical attributes [73].

The maximum number of unique values per attribute is set to 20 and the missing values are ignored. In addition, the class posterior probability is set to 0.001 and calculation process follows the formula presented in chapter 3. Then, the created node needs to connect with the Naive Bayes predictors, which predicts the class per row according to the learned model. The score is computed in the fifth step based on the mentioned formula. Figure 4.18 illustrates a predictive Naive Bayes overview. In addition, the Knime workflow is attached in Appendix A.

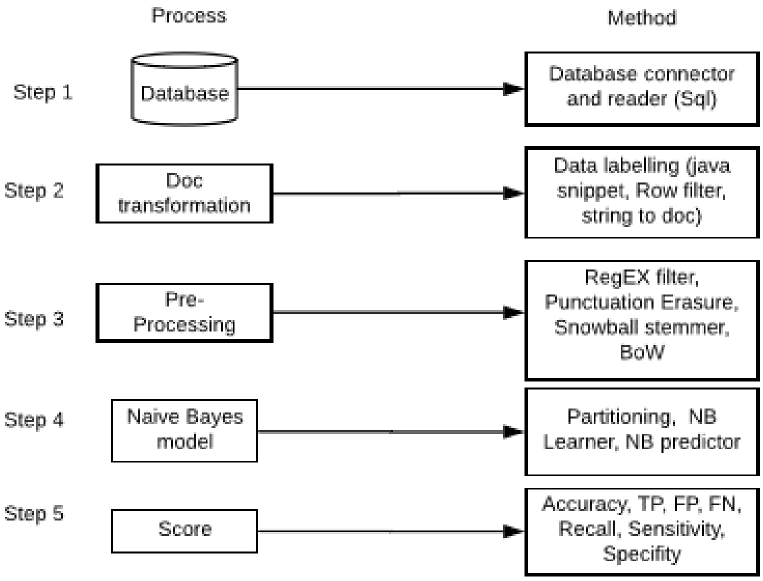

Process
Method
Step 1
Database
Database connector and reader (Sql)
Step 2
Doc transformation
Data labelling (java snippet, Row filter, string to doc)
Step 3
Pre-Processing
RegEX filter, Punctuation Erasure, Snowball stemmer, BoW
Step 4
Naive Bayes model
Partitioning, NB Learner, NB predictor
Step 5
Score
Accuracy, TP, FP, FN, Recall, Sensitivity, Specifity

FIGURE 4.18: Naive Bayes overview diagram.

4.5 Evaluation Metrics

Regarding the project research questions, our goal in this step is to predict the pandemic outbreaks. It is essential to explain what we mean by good prediction and how we can measure such a prediction. The evaluation ways adapted for classification performance play a vital role in design and classifiers selection, especially when we are faced with an imbalanced dataset. Imbalanced dataset means having at least one class in minority relative to others, which would be a challenge in the real world of machine learning usage.

In order to evaluate the performance of the classifiers, various metrics can be used. The different choices for evaluation metric can be made based on the goal of the problem. We first explain the common evaluation metrics that have been used for binary classification tasks. Examples of these measurements are the Error rate, Recall, Sensitivity, Precision, Specificity, and Accuracy. Below the definitions of the common evaluation metrics are described, which they are used for our problem [74] Eq. (1.5):

The Positive Rate (TPR) or Recall, determines to what extent all the samples, which needed to be classified, can be considered as positive. If a positive sample is classified as positive, it is counted as a true positive.

$$\text{true positive rate} = \frac{TP}{TP + FN} \text{= Sensitivity} \tag{4.1}$$

True Negative Rate (TNR), is the percentage of negative samples correctly classified within negative class. If the class label of a sample is negative and it is classified as negative, then it is counted as a true negative.

$$\text{true negative rate} = \frac{TN}{TN + FP} \text{ = Specificity} \tag{4.2}$$

False Positive Rate (FPR), is the percentage of negative samples wrongly classified as belonging to the positive class.

$$\text{false positive rate} = \frac{FP}{FP + TN =} \text{ = 1 - Specificity} \tag{4.3}$$

Precision is the proportion of positive examples, which are positive and illustrating how accurate is the learning method. However, in imbalance classes since FPR would be greater than TPR, then this would affect precision, which is not useful.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.4}$$

Accuracy can be determined as the number of correct predictions. It can be calculated as follow:

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{4.5}$$

By using the Knime scoring node in all prediction workflows the mentioned evaluation metrics have been calculated automatically. Tables 4.2, 4.3, and 4.4 show the predictive models scores with one time run, respectively.

TABLE 4.2: Decision Tree evaluation scores

	Results for the credibility of Decision Tree classification					
Category	Error Rate	Recall	Prec.	Specificity	Sensitivity	Accuracy
Pandemic_Related	1.96	0.983	0.988	0.975	0.983	0.983
Communicable Disease		0.975	0.965	0.983	0.975	

TABLE 4.3: SVM evaluation scores

	Results for the credibility of SVM classification					
Category	Error Rate	Recall	Prec.	Specificity	Sensitivity	Accuracy
Pandemic_Related	1.95	0.982	0.988	0.977	0.982	0.984
Communicable Disease		0.977	0.965	0.982	0.977	

TABLE 4.4: Naive Bayes evaluation scores

	Results for the credibility of Naive Bayes classification					
Category	Error Rate	Recall	Prec.	Specificity	Sensitivity	Accuracy
Pandemic_Related	6.17	0.913	0.995	0.99	0.913	0.936
Communicable Disease		0.99	0.847	0.913	0.99	

4.5.1 Comparison of different classifiers

A Number of learning schemes such as decision tree, SVM, and Naive Bayes models are tried in this step. The supervised classifiers are trained to predict credibility levels on Twitter posts. This part of the project aimed to precisely find the correct samples and high accuracy. This data can be followed by the labels of "corrected" and "wrong" classified. Therefore, the experiments are done for performance comparison of three different predictive classifiers. Tables 4.5, 4.6, and 4.7 present the confusion matrix of 3 predictive models, respectively.

TABLE 4.5: Credibility summary of decision tree classification.

Corrected classified	7221
Wrong classified	145
Accuracy	98.03%
Error rate	1.96%
Kappa statistic	0.955

TABLE 4.6: Credibility summary of SVM classification.

Corrected classified	7222
Wrong classified	144
Accuracy	98.04%
Error rate	1.95%
Kappa statistic	0.956

TABLE 4.7: Credibility summary of Naive Bayes classification.

Corrected classified	6767
Wrong classified	445
Accuracy	93.83%
Error rate	6.17%
Kappa statistic	0.866

Afterward, the confidence level is computed. The confidence level (CL) is the frequency of possible confidence intervals, which include the true value of the corresponding parameter. A particular confidence interval gives a range of plausible values for the parameter of interest. The confidence level is computed as follow [75] (Eq.6):

$$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right). \tag{4.6}$$

Where P denotes the probability of distribution, $\bar{X}$ is the sample mean, and σ is standard deviation. In addition, $\sqrt{n}$ denotes the number of terms and μ is the unknown parameter between the stochastic endpoints.

TABLE 4.8: Calculation of important parameters

Standard Deviation	2.680303543
Standard Error	0.03156139
Margin Error	0.06312278

Table 4.8 shows the calculation of necessary parameters for finding the confidence level. In the first step, the mean values have been computed by adding up the accuracy scores for each method divide by the total number of samples. The standard deviations have been calculated in the second step. In the third step, the standard errors have been computed by dividing the standard deviation over the square root of the sample size. The margin errors have been defined by multiplying the standard error in the fourth step. Finally, in the last step, the confidence levels have been computed by adding the margin of error to the mean from Step 1 and then subtracting the margin of error from the mean [76].

In order to calculate the confidence interval, needs to set it as 90%, 95%, or 99%. In this work, the confidence level is 95%, which used most commonly in research area. It presents a particular interval within which the data is 95% sure or certain for a specific outcome.

TABLE 4.9: Confidence level results

Methods	Mean of accuracy	CL	lower endpoint	upper endpoint
SVM	97.79	95.11	93.15	97.07
DT	97.47	94.78	92.82	96.74
NB	93	90.31	88.35	92.27

Table 4.9 illustrates the obtained results of the mean of accuracy from three different methods after running four times. The performance of all

models has not been improved after four times due to the small number of data samples. Therefore, the performance has been measured based on the four times running of predictive models. In this project, the SVM method performs better in comparison with the decision tree and Naive Bayes. It can be concluded that the pandemic outbreak increased gradually based on the SVM result. The confidence interval is computed as 95.11%, which is higher than the other methods. The lower and upper endpoints represent the upper and lower limit by subtracting or adding the value of 1.96. Figure 4.19 represents the comparison result of the confidence level for the relevant predictive techniques.

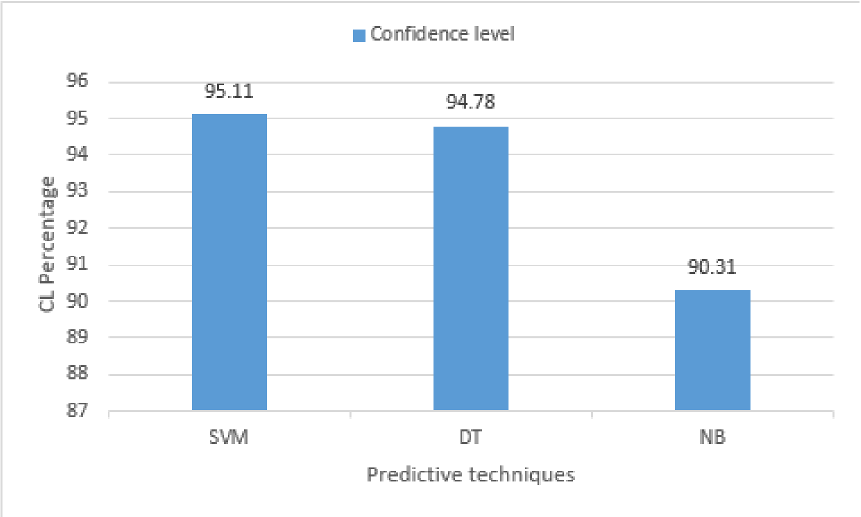


FIGURE 4.19: Comparison of used predictive techniques.

In addition, the Receiver Operating Characteristic (ROC) curves are plotted for the three used techniques. ROC graph is a useful technique for visualizing the techniques performance. This type of graph is mostly used in medical decision making, and recently have been increasingly adopted in the data mining and machine learning research communities. It plots the curve between TPR and FPR of an algorithm [77].

Figure 4.20 illustrates the comparison of the ROC curve of the three used predictive techniques in this work. Each point on the ROC curve shows a sensitivity/specificity pair regarding a particular decision threshold. A test, with no overlapping in two classes, has a ROC curve, which passes through

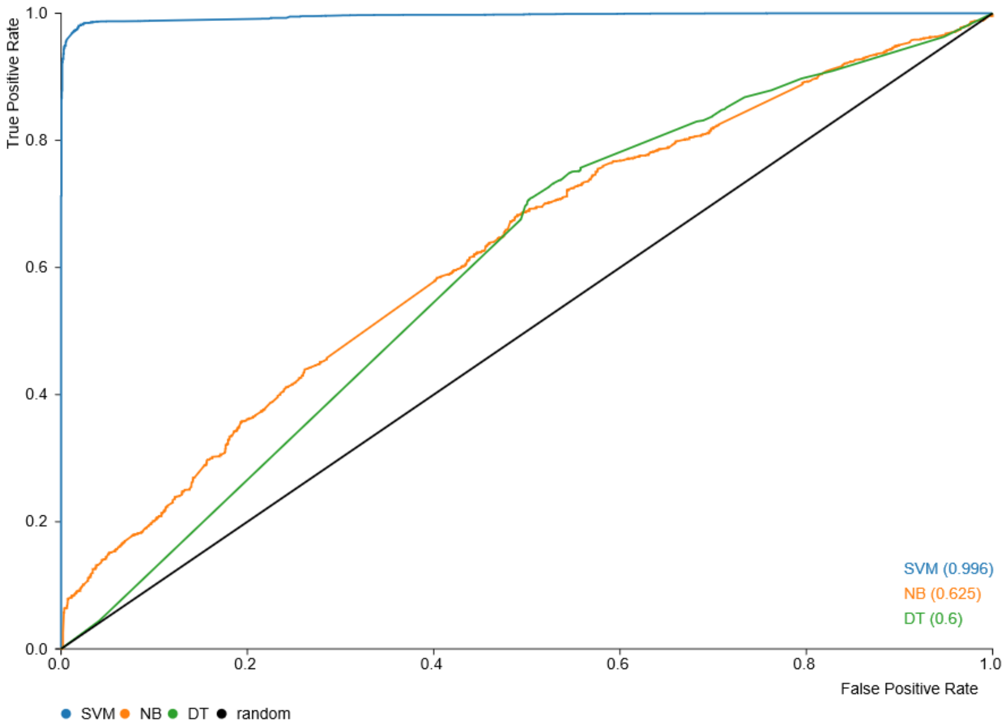


FIGURE 4.20: Comparison of the ROC curve of the three predictive techniques.

the upper left corner. Therefore, the closer ROC curve to the upper left corner is the higher overall accuracy of the test. In this work, the SVM technique outperforms other machine learning techniques. The DT technique has a better performance than the NB technique with respect to the pandemic outbreak prediction.

To conclude, the dataset collection, data analysis and machine learning experiments have been described in this chapter. The performance of adopted methods has been compared and our results show that while a particular classifier may predict and perform well, another classifier might perform better prediction.

Chapter 5

Conclusion and Future work

This chapter conclude the thesis by providing a brief summary of the progress made towards achieving the goals of the thesis. The future works also suggest a new research direction that leads to improving the results of this study.

5.1 Conclusion and Future Work

In this project Twitter used as SNA to study interactions between users during February to March 2018. Predicting the pandemic outbreaks from the content in social media has attracted various research activities in the last decades. In addition, affected people and humanitarian organizations have a prompt information needs during the time of disasters and emergency situations.

In the case of the Twitter social network, predicting the popularity of tweets has a vital role in several applications such as crisis management, popular news detection, and data analysis. Current studies represent the social media content posted during a crisis to satisfy peoples information demands. However, getting relevant information from social media platform such as Twitter is not a trivial task. Despite advances machine learning techniques, quick access to relevant information is still a challenging task.

Tweets can be used to analysis opinions. Several user-based and tweet-based features have been extracted from the body of tweets and the users who posted the tweets. By applying various analysis techniques the influential users, top 20 tweets and retweets, user's sentiment, and Twitter user distribution are identified based on the project research questions. The results represent the power of users to propagating a news or event via social media such as Twitter. In addition, people from most of the countries having a concern regarding the pandemic or communicable diseases. Moreover, three different machine learning techniques (SVM, DT, and NB) are utilized to predict pandemic outbreaks. Our experimental results illustrate that SVM technique outperforms other techniques.

One of the weaknesses of the research is the lack of enough number of collected tweets. It is referred to the period limitation of this project. This is also difficult task as Twitter limits the number of tweets, which can be downloaded. Thus, a first future work is adding more data sources in order to improve the tweet collection rate and obtain much larger samples.

Another weakness can be relevant collected tweets where irrelevant tweets can be also gathered by searching a certain keyword. For example when searching a keyword "virus", it can be collected both computer virus and disease related virus. Therefore, the other task for future work is determining an accurate keyword search and reducing the number of irrelevant tweets.

Another extension for future work would be to specify diseases either communicable disease or pandemic ones in order to see if more efficient and accurate classifies can be trained. Furthermore, more advanced techniques and deep learning can be employed for this problem to improve the performance of classification and efficient results.

Appendix A

Appendix A: The Knime Workflows

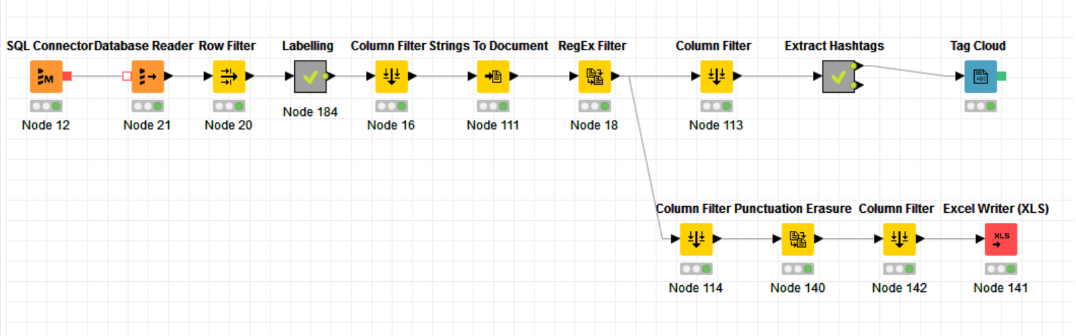


FIGURE A.1: Preprocessing & Extracted Hashtag.

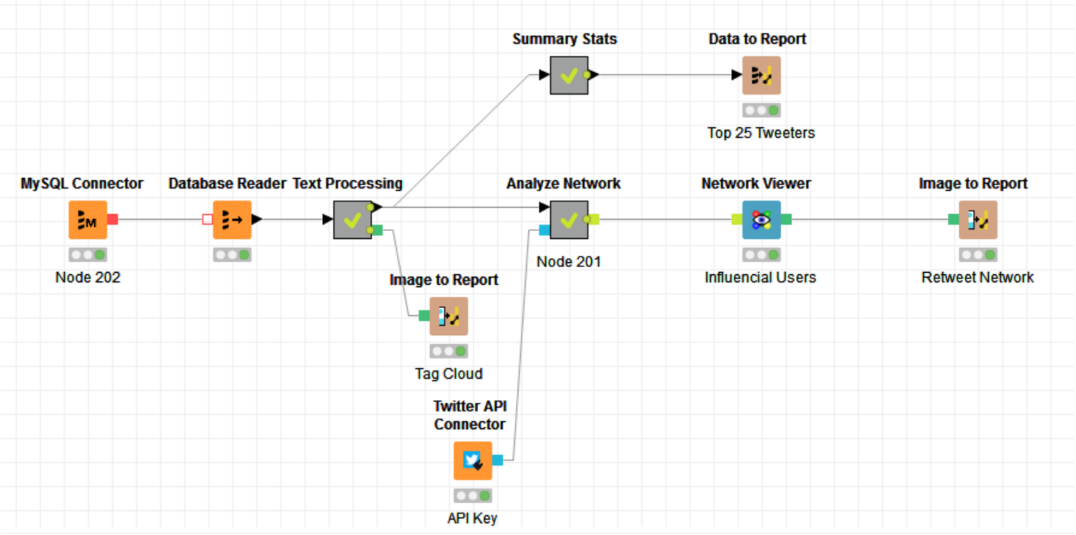


FIGURE A.2: The top Tweeters, influential users, and Retweet network Workflow.

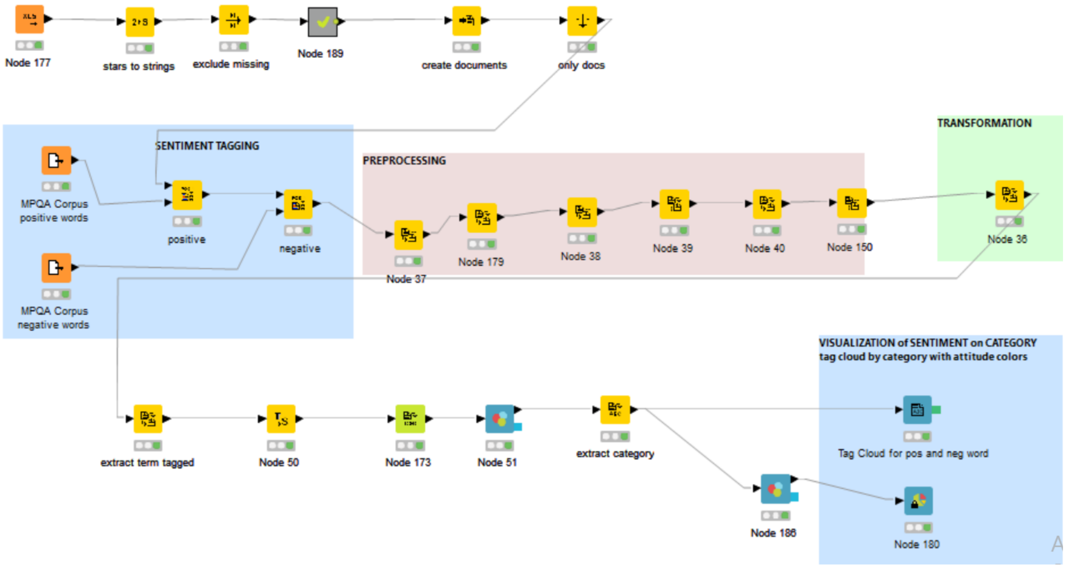
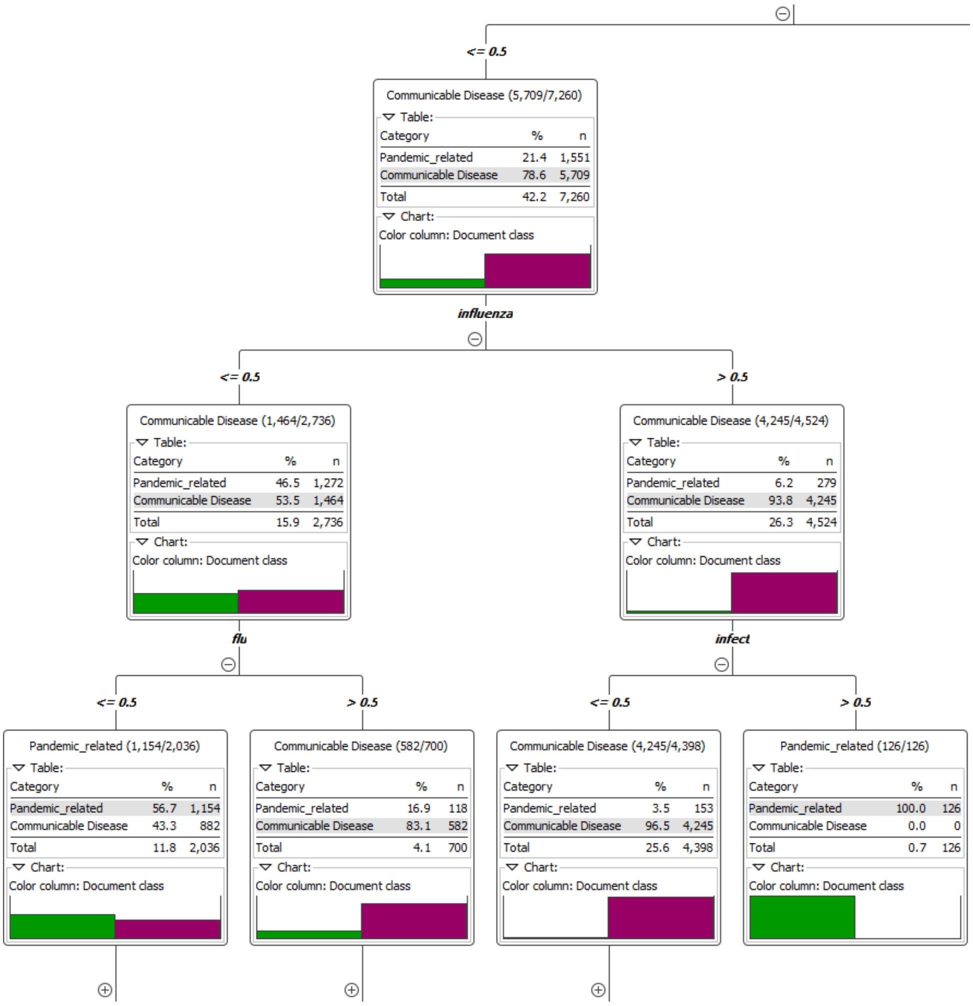


FIGURE A.3: Sentiment analysis workflow.



Communicable Disease (5,709/7,260)

Table:

Category	%	n
Pandemic_related	21.4	1,551
Communicable Disease	78.6	5,709
Total	42.2	7,260

Chart:
Color column: Document class

Communicable Disease (1,464/2,736)

Table:

Category	%	n
Pandemic_related	46.5	1,272
Communicable Disease	53.5	1,464
Total	15.9	2,736

Chart:
Color column: Document class

Communicable Disease (4,245/4,524)

Table:

Category	%	n
Pandemic_related	6.2	279
Communicable Disease	93.8	4,245
Total	26.3	4,524

Chart:
Color column: Document class

Pandemic_related (1,154/2,036)

Table:

Category	%	n
Pandemic_related	56.7	1,154
Communicable Disease	43.3	882
Total	11.8	2,036

Chart:
Color column: Document class

Communicable Disease (582/700)

Table:

Category	%	n
Pandemic_related	16.9	118
Communicable Disease	83.1	582
Total	4.1	700

Chart:
Color column: Document class

Communicable Disease (4,245/4,398)

Table:

Category	%	n
Pandemic_related	3.5	153
Communicable Disease	96.5	4,245
Total	25.6	4,398

Chart:
Color column: Document class

Pandemic_related (126/126)

Table:

Category	%	n
Pandemic_related	100.0	126
Communicable Disease	0.0	0
Total	0.7	126

Chart:
Color column: Document class

FIGURE A.4: Part of a created decision tree.

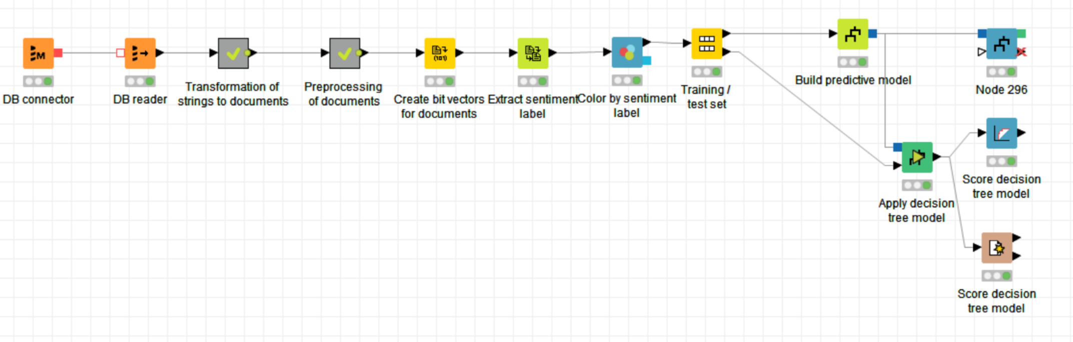

DB connector
DB reader
Transformation of strings to documents
Preprocessing of documents
Create bit vectors for documents
Extract sentiment label
Color by sentiment label
Training / test set
Build predictive model
Node 296
Apply decision tree model
Score decision tree model
Score decision tree model

FIGURE A.5: Decision Tree workflow.

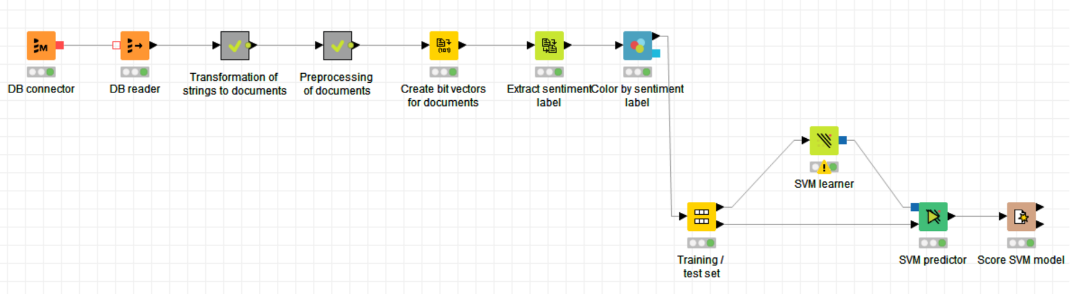


FIGURE A.6: SVM predictive workflow.

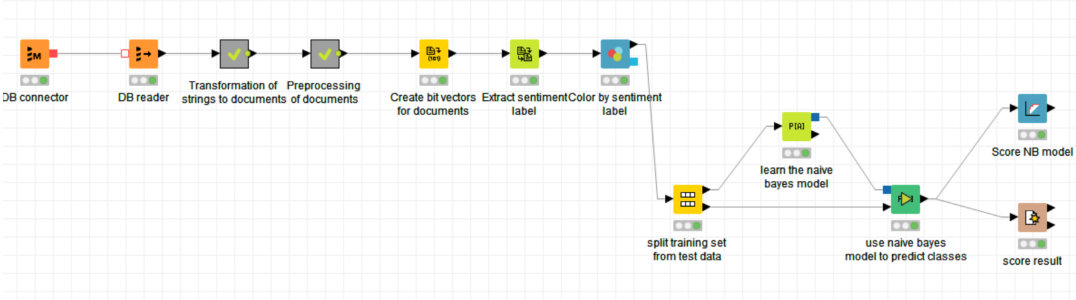

DB connector
DB reader
Transformation of strings to documents
Preprocessing of documents
Create bit vectors for documents
Extract sentiment label
Color by sentiment label
learn the naive bayes model
Score NB model
split training set from test data
use naive bayes model to predict classes
score result

FIGURE A.7: Naive Bayes predictive workflow.

Bibliography

[1] C. K. boyd d, “Critical questions for big data. information, communication and society.”, *Routledge Publisher*, vol. 15, no. 662-679, 2012.

[2] WHO., 2009. [Online]. Available: http://whqlibdoc.who.int/publications/2010/9789241599924_eng.pdf.

[3] W. health org., http://whqlibdoc.who.int/publications/2010/9789241599924_eng.pdf, [WHO: Evolution of a Pandemic A (H1N1) 2009, 2nd Edition], 2009.

[4] WHO., http://www.who.int/csr/don/2009_04_24/en/index.html, 2009.

[5] W. health org., http://www.who.int/csr/don/2009_04_26/en/index.html, 2009.

[6] WHO., http://www.who.int/mediacentre/news/statements/2009/h1n1_20090427/en/index.html, 2009.

[7] W. health org., http://www.who.int/mediacentre/news/statements/2009/h1n1_pandemic_phase6_20090611/en/index.html, 2009.

[8] K. F. MI.M̃eltzer Cox NJ, “The economic impact of pandemic influenza in the united states: Priorities for intervention.”, *Emerging Infectious Diseases*, vol. 5, no. 5, 1999.

[9] I. M.L. J. Timothy C. Germann Kai Kadau and C. A. Macken, “Mitigation strategies for pandemic influenza in the united states.”, *The National Academy of Sciences of the USA*, 2006.

[10] R. C. Rohan D.W; S. Anand ; Subbalakshmi, “Twitter analytics: Architecture, tools and analysis.”, *MILITARY COMMUNICATIONS CONFERENCE, 2010 - MILCOM 2010*, 2010.

[11] D. R. R. Dynes Quarantelli and E. Quarantelli, "Group behavior under stress: A required convergence of organizational and collective behavior perspectives.", *Sociology and Social Research*, no. 52, 1968.

[12] P. L.H.A. L. Starbird K. and S. Vieweg, "Chatter on the red: What hazards threat reveals about the social life of micro blogged information.", *ACM Conference on Computer Supported Cooperative Work, CSCW*, no. 52, pp. 241–250, 2010.

[13] M. M. Castillo C. and B. Poblete, "Information credibility on twitter.", *Hyderabad, India*, 2011.

[14] S. Vieweg, "Microblogged contributions to the emergency arena: Discovery, interpretation and implications.", *Computer Supported Collaborative Work*, 2010.

[15] E. G. Chew C. and M. Sampson, "Pandemics in the age of twitter: Content analysis of tweets during the 2009 h1n1 outbreak.", *Published online 2010 Nov 29. doi: 10.1371/journal.pone.0014118*, vol. 11, no. 5, pp. 1–13, 2010.

[16] D. Yates and S Paquette, "Emergency knowledge management and social media technologies: A case study of the 2010 haitian earthquake.", *International Journal of Information Management*, vol. 31, no. 1, pp. 6–13, 2011.

[17] S. Wasserman and K. Faust, *Social network analysis: Methods and applications.* New York, U.S.A, 1994. [Online]. Available: `http://nptel.ac.in/courses/105105110/pdf/m4l05.pdf`.

[18] L. A.S.A.L.B.D.B.N.C.N.C.J.F.M.G.e. a. David Lazer Alex Sandy Pentland, "Life in the network: The coming age of computational social science.", *AAAS*, vol. 323, 2009.

[19] B. A.K. A. Ansell C., "Managing transboundary crises: Identifying the building blocks of an effective response system.", *Contingencies and Crisis Management*, vol. 18, no. 4, pp. 195–207, 2010.

[20] P. G. Patterson KD, "The geography and mortality of the 1918 influenza pandemic.", *American Public Health Association*, vol. 65, pp. 4–21, 1991.

[21] C. B.F.D.H. K. Murray CJ Lopez AD, "Estimation of potential global pandemic influenza mortality on the basis of vital registry data from the 1918–20 pandemic: A quantitative analysis.", *Lancet*, p. 368, 2007.

[22] F. A. M. Shaluf I and A M. Said, "A review of disaster and crisis," disaster prevention and management.", vol. 12, no. 1, pp. 24–32, 2009.

[23] L. Larsson, "Kris och lärdom - kriskommunikation från tjernobyl till tsunamin.", 2008.

[24] W. T. Coombs, *Ongoing crisis communication: Planning, managing, and responding.* CA: Sage, 2012.

[25] C. Hermann, *Some consequences of crisis which limit the viability of organizations.* 2015.

[26] T. C. Pauchant and I. I. Mitroff, *Transforming the crisis-prone organization: Preventing individual, organizational, and environmental tragedies.* San Francisco, Ca: Jossey-Bass, 1992.

[27] K. Fearn-Banks, *Crisis communications: A casebook approach.* New York., 2011.

[28] J Falkheimer and M. Heide, "Multicultural crisis communication: Towards a social constructionist perspective.", *Journal of Contingencies and Crisis Management*, vol. 14, no. 4, pp. 180–189, 2006.

[29] S. Fink., *Crisis management planning for the inevitable.* 1986.

[30] C. M. A. Gayer M. and J. T. Watson, "Epidemics after natural disasters. emerging infectious diseases.", *World Health Organization, Geneva, Switzerland*, vol. 13, pp. 1–5, 2007.

[31] WHO., "Who handbook for journalists: Influenza pandemic.", *World Health Organization, Geneva, Switzerland*, 2005. [Online]. Available: http://www.who.int/csr/don/Handbook_influenza_pandemic_dec05.pdf.

[32] M. S. Meade and E. J. Earickson, *Medical geography.* New York, 2005.

[33] H. B. F. Cha M.; Haddadi and G. K., “Measuring user influence on twitter: The million follower fallacy.”, *AAAI Conference on Weblogs and Social Media*, no. pp. 10-17, 2010.

[34] D. Huffaker, “Dimensions of leadership and social influence in on line communities.”, *Human Communication Research*, no. pp. 593-617, 2010.

[35] N. A. Longini I. M., U. K.H.W.C.D.A. T. Xu S., and E. M. Halloran, “Containing pandemic influenza at the source.”, *Science*, no. 1083–1087, 2005.

[36] huffpost tech., “Twitter user statistics revealed.”, 2010.

[37] A. K.M.G.M.J.S.D.P.M.G. D. Palen L., *A vision for technology-mediated support for public participation and assistance in mass emergencies and disasters.* 2010.

[38] L. B.A. L. Jin Y., “Examining the role of social media in effective crisis management: The effects of crisis origin, information form, and source on publics’ crisis responses.”, *SAGE*, vol. 41, no. 1, 2014.

[39] M. A. Onook Oh and H. R. Rao., “Information control and terrorism: Tracking the mumbai terrorist attack through twitter.”, *Information Systems Frontiers*, vol. 13, no. 1, 2011.

[40] N. G.R. D. G. Lu K. Buyyani and Z. Chen., “Influenza a virus informatics: Genotype-centered database and genotype annotation.”, *IEEE International Multi-Symposiums on Computer and Computational Sciences (IMSCCS’07)*, 2007.

[41] S. S. Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms.* 2014.

[42] A. Sami and M. Takahashi., “Decision tree construction for genetic applications based on association rules.”, *IEEE TENCON, Melbourne, Australia*, 2005.

[43] E. d. Q. Martin Szomszor Patty Kostkova, “Twitter predicts swine flu outbreak.”, *International Conference on Electronic Healthcare*, no. 18-29, 2009.

[44] A. Culotta, “Detecting influenza outbreaks by analyzing twitter messages.”, *International Conference on Electronic Healthcare*, 2010.

[45] E. L.M.H.R.B. F. Gomide, “Evolving fuzzy modeling using participatory learning.”, 2010.

[46] I. V.O. R. A Agarwal B Xie, “Proceedings of the workshop on languages in social media.”, 2011.

[47] A. Pak and P. Paroubek., “Twitter as a corpus for sentiment analysis and opinion mining.”, vol. 10, 2010.

[48] U. Q. Khan Farhan Hassan and S. Bashir., “A semi-supervised approach to sentiment analysis using revised sentiment strength based on sentiwordnet.”, no. 1-22, 2016.

[49] knime., https://www.knime.com/knime, 2018.

[50] Tableau., https://www.tableau.com/, 2018.

[51] C. Disease, http://www.euro.who.int/en/health-topics/communicable-diseases, 2018.

[52] A. text processing, http://kt.ijs.si/theses/phd_matic_perovsek.pdf, 2016.

[53] A. H. S. F. Sayeedunnissa and M. Hameed, *Supervised opinion mining of social network data using a bag of words approach on the cloud.* 202. 2013.

[54] I. Research, http://www.informationr.net/ir/19-1/paper605.html#.Wv7NvEiFM2w, 2014.

[55] M. Porter, http://snowball.tartarus.org/texts/introduction.html, 2001.

[56] R. F. Firouzi, “A decision tree-based approach for determining low bone mineral density in inflammatory bowel disease using weka software.”, *Eur J Gastroenterol Hepatol*, 2007.

[57] L. W. Pierre Geurts Alexandre Irrthum, “Supervised learning with decision tree-based methods in computational and systems biology .”, *NCBI*, 2009.

[58] R. O. L. Breiman J. Friedman and C. Stone., “Classification and regression trees.”, *Wadsworth International*, 1984.

[59] R. Jafari and H. R. Arabnia., “A survey of face recognition techniques.”, 2009.

[60] W. B.Z. D. Bell B., “Forecasting river run off through support vector machines.”, *International Conference on Cognitive Informatics and Cognitive Computing(ICCICC)*, 2012.

[61] E. Frank and R. R. Bouckaert., *Naive bayes for text classification with unbalanced classes.* 2006.

[62] I. A. Vivek Narayanan and A. Bhatia., “Fast and accurate sentiment classification using an enhanced naive bayes model.”, *In Intelligent Data Engineering and Automated Learning IDEAL*, 2013.

[63] D. M. Freeman., “Using naive bayes to detect spammy names in social networks.”, *ACM workshop on Artificial intelligence and security*, 2013.

[64] E. Frank and R. R. Bouckaert, *Naive bayes for text classification with unbalanced classes.* 2006.

[65] MPQA, http://mpqa.cs.pitt.edu/, 2018.

[66] W. health org, http://www.emro.who.int/pandemic-epidemic-diseases/mers-cov/mers-situation-update-february-2018.html, 2018.

[67] WHO, http://www.emro.who.int/pandemic-epidemic-diseases/mers-cov/mers-situation-update-march-2018.html, 2018.

[68] Cdtr, https://ecdc.europa.eu/sites/portal/files/documents/Communicable-disease-threats-report-7-apr-2018.pdf, 2018.

[69] J. R. Quinlan, *Programs for machine learning.* 3. 1993, vol. 16.

[70] M. M. John C. Shafer Rakesh Agrawal, “Sprint: A scalable parallel classifier for data mining.”, *Proceedings of the 22th International Conference on Very Large Data Bases*, no. 544-555, 1996.

[71] S. Keerthi, S. Shevade, C. Bhattacharyya, and K. Murthy, "Improvements to platt's smo algorithm for svm classifier design", CSA, Banglore, India, Tech. Rep., 1999.

[72] J. C. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines", ADVANCES IN KERNEL METHODS - SUPPORT VECTOR LEARNING, Tech. Rep., 1998.

[73] M. S.S. S. Tina R. Patil, "Performance analysis of naive bayes and j48 classification algorithm for data classification.", *International Journal of Computer Science and Applications*, vol. 6, 2013.

[74] T. Fawcett., "An introduction to roc analysis.", *Science Direct*, 2005.

[75] Confidence, `https://en.wikipedia.org/wiki/Confidence_interval`, 2018.

[76] Online, `https://measuringu.com/ci-five-steps/`, 2018.

[77] T. Fawcett, "Using rule sets to maximize roc performance.", *IEEE Internat. Conf. on Data Mining (ICDM-2001)*, 2001.