S	
E1	
E2	
For Office Use Only	

Masters Project Final Report (MCS) 2019

Project Title	Sentiment based analysis of social media data using fuzzy-rough set classifier for the prediction of the presidential election
Student Name	W.A.Suresha Nilmini Perera
Registration No. & Index No.	2016/MCS/081 16440815
Supervisor's Name	Dr.Kasun Karunanayaka

For Office Use ONLY

Sentiment based analysis of social media data using fuzzy-rough set classifier for the prediction of the presidential election

A dissertation submitted for the Degree of Master of Computer Science

W.A.S.N.Perera
University of Colombo School of Computing
2019

Declaration

The thesis is my original work and has not been submitted previously for a degree at this or any other university/institute.

To the best of my knowledge it does not contain any material published or written by another person, except as acknowledged in the text.

Student Name: W.A.Suresha Nilmini Perera

Registration Number: 2016/MCS/081

Index Number: 16440815

Suresha

Signature: Date: 20/11/2020

This is to certify that this thesis is based on the work of

Ms. W.A.Suresha Nilmini Perera

under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by: Dr. Kasun Karunanayaka

Supervisor Name: Dr. Kasun Karunanayaka

Kasun

Signature: Date: 20/11/2020

Abstract

As an interdisciplinary research field, sentiment analysis is one of the momentous applications in Natural Language Processing, for quantifying the emotional value in vast data in the form of text available in social media networks to gain an understanding of the attitudes, opinions, and emotions expressed. There is a great deal of literature on the various ways to address sentiment analysis with social media and this project focuses on Machine Learning techniques with twitter data analysis. Special attention is drawn towards the classifiers based on the Fuzzy Set and Rough Set approach which are two powerful mathematical components of the computational intelligence with its new dimension involved in the field of sentiment analysis. However, there is a minimal number of review papers discussed about rough-fuzzy classifier involvement in sentiment analysis and there is a plethora of work that must be done with text mining in natural language processing. Sentiment Analysis, also called opinion mining, which is the field of computational study that analyzes people's opinions, attitudes and emotions toward an entity.

Decision-making behavior in an online community is predominant because many individuals now use community-based web services thus the opinions of others are readily available in online environments where people often rely on other individuals' decisions for social validation to make their own. The power of Social Media Websites is rampant and growing a plethora of information and data convoluted with varying interests, opinions, and emotions with human generated baselines. The widespread use of the above media encourages positive and negative attitudes about people, organizations, places, events, and ideas. Sentiment Analysis, also called opinion mining, which is the field of computational study that analyzes people's opinions, attitudes and emotions toward an entity.

The mission of this project is to develop a sentiment based classifier using machine learning and fuzzy-rough set theory. Further, it carries automatic sentiment classification with twitter corpus collected during a certain time period with regard to the case study for the prediction of results at the presidential election 2019, Sri Lanka.

The fuzzy rough classifier is developed using the Fuzzy Rough Nearest Neighbor algorithm. The performance of the classifier was evaluated with precision, recall, accuracy and F1 score against the Multinomial Naive Bayes and Support Vector Mechanism. The accuracy of the fuzzy rough set based classifier is high compared to other classifiers. The actual results of the presidential election of 2019 are tally with the predicted results of the classifier. Therefore, the current state of art for the prediction of political sentiment with microblogging which is probable with the social media data as witnessed with this case study and this can be used in the other cases as well.

Acknowledgements

I would like to acknowledge my indebtedness and render my warmest thanks to my supervisor, Dr.Kasun Karunanayaka, who made this work possible. His friendly guidance, constant supervision and expert advice have been invaluable throughout all stages of the work.

Also, I would like to convey my utmost gratitude to all the other academic members of University of Colombo School of Computing - UCSC for the knowledge they passed throughout all four years.

This thesis has been written during my stay at the Department of Computer Science, The Open University of Sri Lanka. I would also wish to express my gratitude to Mr. Duminda de Silva, The Head of the Department of Computer Science, The Open University of Sri Lanka, for providing me with time to engage with this project and the continuous encouragement provided. I take this opportunity to thank him.

The members of my family are important to me in the pursuit of this project. I wish to thank my loving and supportive husband, Mr.Nuwan Hettiarachi, for his continuous support and understand. My two wonderful children, Thisari Sandevmee and Oshadha Gethyan, provided unending inspiration. My thanks are extended to my parents for their constant encouragement whose love and guidance are with me in whatever I pursue.

Finally I would like to thank my colleagues specially Ms.Shalini Rajasingham who gave me valuable support and encouragement given endlessly, to make this project success.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

ABBREVIATION

Sentiment Analysis	SA
Machine Learning	ML
Natural Language processing	NLP
Gotabaya Rajapaksa	GR
Sajith Premadasa	SP
Natural Language Toolkit	NLTK
Term Frequency and Inverse Document Frequency	TF-IDF
Fuzzy Rough Nearest Neighbor	FRNN

PUBLICATIONS

1. The paper titled "Sentiment Classification of Social Media Data with Supervised Machine Learning Approaches: Common Framework, Challenges, and New Dimensions" written based on this project has been accepted and will be presented in the SLAAI-International Conference on Artificial Intelligence – 2020 (SLAAI-ICAI -2020).

Chapter 1

Introduction

1.1 Introduction

With the rapid technological advancements, the numbers of internet users are growing fast where many of them use online review sites, forum discussions, social networks and personal blogs to express their opinions and ideas by interacting, sharing, posting and manipulating content. The power of social media is rampant and growing with information and data convoluted with varying interests, opinions, and emotions. There is a prolific expansion with the above media and those encourage people to get the consent of others towards people, products, places etc. Those platforms become new benchmarks for sharing the opinionated text on various products, services, experiences etc. Further, opinionated text can be further utilized to identify the attitudes towards the general public which is considered as really helpful in making various decisions.

Microblogging platforms on social media permits the users to communicate among themselves and this has become more popular with its free and easy accessibility. Twitter and Facebook are the most prominent platforms whilst providing a user friendly platform for their subscribers. As a result of that, millions of messages are broadcasting via those platforms resulting in big data for the researchers. The amount of data carried with microblogging sites are huge compared to other resources available. Further, this is identified as a media consist of plethora of sentiment inbuild with the text messages.

Text mining permits the process of examining large collections of unstructured and inconsistent raw data to discover new information, find the patterns and correlations between huge data sets to predict outcomes which are hard to extract. Those information are further converted into structured form which is possible to further analyzed or classified. There is a prolific increase in the area of text mining including government funded projects, research in higher education, political context, education, medicine, microbiology, businesses etc. Therefore, the underlying idea of Sentiment Analysis (SA) is to identify and extract the subjective information hidden in the text (this can be either a comment for someone’s opinion, judgement etc.) by using an automated process in order to classify it as positive, negative or neutral with polarity detection.

It is very common that people make decisions based on others' opinions. Nowadays, many industries use this to evaluate the success of their campaigns or new products in almost all areas. This becomes more practical in the context of politics as individuals assess others' opinion before making a voting decision in a particular election. It is the nature of a human being to evaluate the close friends and family members on how they perceive the world before making an opinion. The same concept can be seen everywhere including organizations where it is a need that they are practicing even before the technical advancements boom up in this area and simply follow surveys, polls, interviews and focus groups to access the public opinion. There is a massive increase in the number of papers devoted to SA during recent years.

There are three different levels of SA :

1. Document Level Analysis - Classify whether a whole opinion document expresses a positive, negative or neutral sentiment [1, 2].

2. Sentence Level Analysis - The polarity is calculated for each sentence and determines the sentiment.

3. Entity/Aspect Level Analysis - This discovers sentiments on entities and/or their aspects.

An election provides the opportunity for all citizens to choose their political leader as the representative in the government. In a democracy, everybody has the right to vote for their candidate. It is very common to conduct a survey or polls to predict the outcome as mentioned above. Data with social media can be effectively used for the prediction of the outcome of an election since this is a powerful communication platform resulting in hundreds and millions messages posted which can be utilized to analyze the sentiment. Those messages include the opinion of people, interests, personal beliefs etc. Those data can be analyzed to predict the world and find interesting conclusions in certain domains. As a result of that social media platforms become vital in the political domain to analyze the events. A vast collection of researches contributed to the political domain with respect to the social media analysis however there exists many unsolved problems. This research is being looked for an efficient way of classifying the sentiment of people towards the candidates before the presidential election of Sri Lanka 2019, according to the collected twitter data by using Machine Learning (ML) and data mining techniques. The results of this study help explain the victory of Gotabhaya Rajapaksha (GR) over Sajith Premadasa (SP).

1.2 Problem domain

This study is correlated with few major areas in modern computer science. ML (ML) based classification, fuzzy-rough sets, SA, social media and clustering algorithms are some of them. Under this background section it is explained in brief what these fields are in order to compose the background of the quest which have been carried out.

1.2.1 Sentiment analysis

SA, also referred to as opinion mining, is listed under Natural Language processing (NLP) to detect, extract or characterize subjective information. For instance, finding the opinion hidden in a given text can be considered. It is very common to analyze the opinion of the people with the aid of comments in social media websites, blog posts, product review websites etc. Positive, neutral and negative are the polarity of opinions identified by the SA. This is directly affecting humans and their activities/behaviors. There exist many business organizations offering the service of SA as their one of the top most services such as: BRAND24, Twitratr, Social Mention. There is a massive increase in the number of papers devoted to SA during recent years.

1.2.2 Social Media

Social media is turning out to be increasingly well known and turning into a significant theme for research in numerous fields. Individuals can speak with their friends so they can share their own emotions. Monitoring and tracking of social media has become more popular Web-based social networking observing or following is the most significant subject in the present situation. Therefore, it requires an effective and convenient way for the analysis of enormous volumes of information in an effective and convenient way since this process is considered as important in the context of Social Science in modeling public opinion. People also can express their opinions and views freely with the help of those social media platforms which has become more popular during the last two decades.

Twitter is one of the most popular platforms where people communicate and mention their view and comments via SMS-based service. However, they have given only 160 characters per message. This becomes more less when it comes to twitter as it requires 20 characters for the username whilst the message is restricted to 140 characters allowing the users to have brief explanations of ideas. The access of the twitter can be done either via web site or application dedicated for smartphones. Twitter uses a special function called “following” to subscribe to the tweets posted by other users. It consists with features such as:

- Each user has a profile
- Each user can add a photo and information about themselves
- Users can *follow* each other
- Users can *tweet*
- Users can interact with a Tweet via *comments (replies)*, *likes*, and *shares (retweets)*
- Users can interact with other users via *direct messaging*
- Users can create a *thread:* A series of connected tweets
- Users use *hashtags* (e.g., #mannheim) in order to associate their tweets with certain topics and to make them easier to find
- Users can search for keywords/hashtags in order to find relevant tweets and users

All the tweets by users can be seen in the public timeline of twitter. Apart from the basic concept of personal status updates in microblogging, it provides many other information such as: political aspects, business related information, personal beliefs and thoughts. Therefore, this platform contains all the information around the world. The business organizations get the use of this platform to analyze the public opinion about the company and their products as well as the customer satisfaction on products as new research areas and they invest money. Those analyses influence executives of the companies to change the organization's future plans. Further, they can make decisions on future designs as well. Furthermore, this has become more popular among the politicians, where they get the use of the twitter data analysis to understand their social images. This research project comes under the branch of political domain where it analyses tweets of the general public towards a particular political leader, ultimately the outcome which is the sentiment represents voters opinion before the election commencing.

1.2.3 Fuzzy sets

Fuzzy set and rough set theory are two diverse approaches for the vagueness represented in two different ways. However, this cannot be used as a cure for the difficulties with the classical set theory. Fuzzy set hypothesis addresses gradualness of information, communicated by the fuzzy membership while rough set hypothesis addresses granularity of information, communicated by the indiscernibility relation.

The fuzzy alludes to things which are not clear or are not clear (vague). Fuzzy logic centers on computing based on "degrees of truth" instead of the boolean rationale of true or false. When this comes to NLP, it is not easy to categorize it into 1 or 0. In this manner, in fuzzy rationale take and 1 as extraordinary cases of truth whereas counting different states of truth in between as in figure 1 and works closer to the way the human brains work by giving a really important adaptability for thinking.

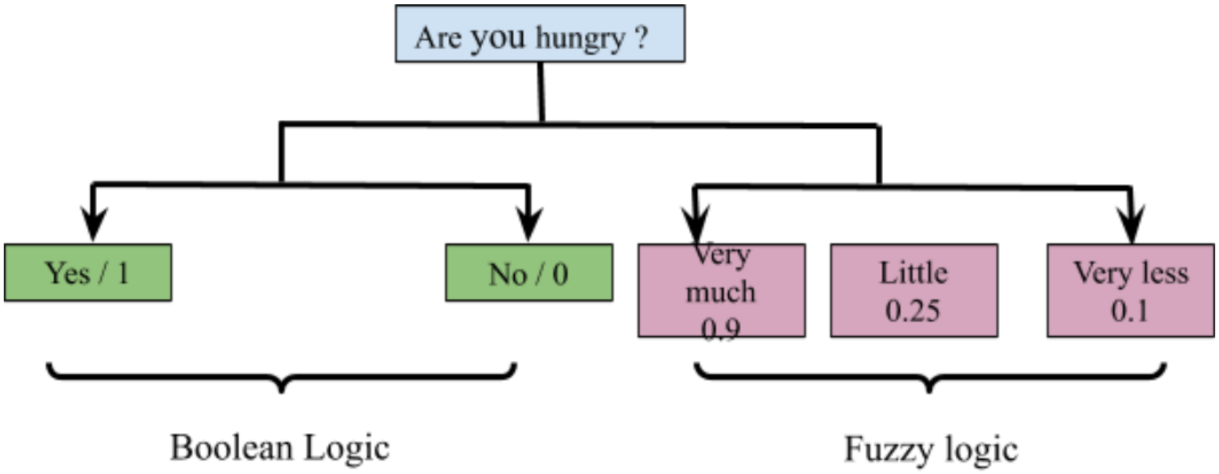


Figure 1: The difference between fuzzy and boolean logic

Fuzzy logic architecture consists of four parts as in Figure 2.

- RULE BASE consists of if-then rules for the purpose of decision making.
- FUZZIFICATION: the numerical values(crisp) converted to fuzzy values with the use of a knowledge base in the system.
- INFERENCE ENGINE: The inference engine is same as the processor of the computer and draws a substantial result by analyzing and concluding all the data that it gets from the fuzzification module.
- DEFUZZIFICATION: The information received from the inference engine is not in a user-readable format and defuzzification converts those values into a crisp value.

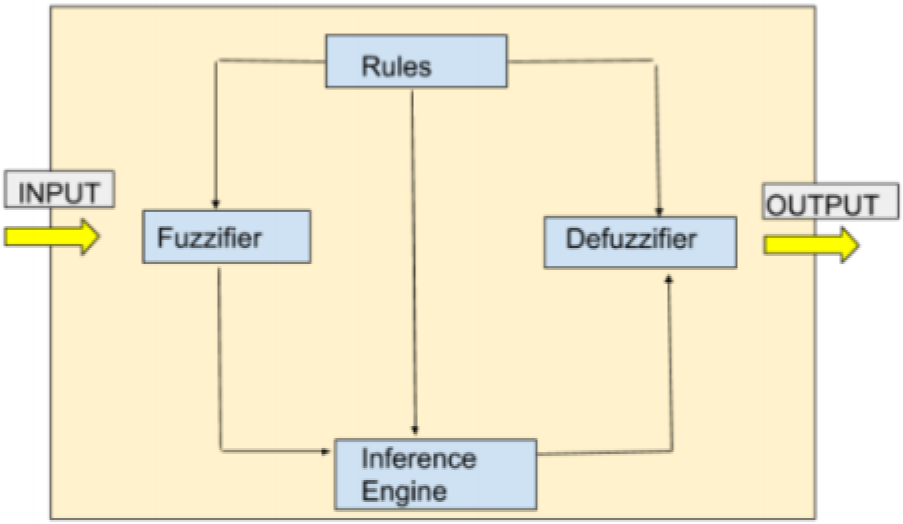


Figure 2 Fuzzy logic architecture

Fuzzy logic is derived from the fuzzy set theory. In the classical set theory:

$f_A(x)$- characteristic function of set A;

where: X = universe of discourse), x = elements of X, A = crisp set.

$f\ A(x): X \rightarrow 0,1$ where $f_A(x) = \begin{cases} 1, if\ x \in A \\ 0, if\ x \notin A \end{cases}$

The membership function of the fuzzy set is $\mu_A(x)$ and the members are varying with different values with in the interval of [0,1].

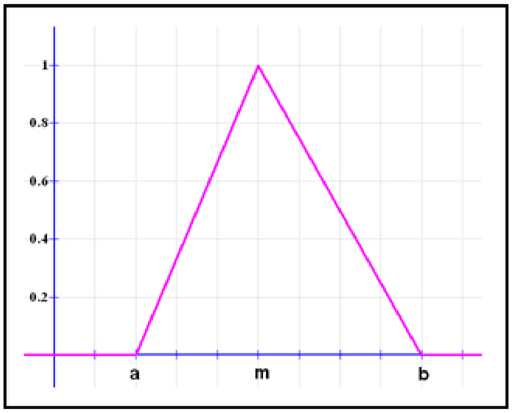
$f\ A(x): X \rightarrow [0,1]$ where $\mu_A(x) = \begin{cases} 1,\ if\ x\ is\ totally\ in \\ 0,\ if\ x\ is\ not\ A; \\ U\ (0<u<1),\ if\ x\ i \end{cases}$

A can be written as a collection of ordered pairs as follows.

$A = \{(X, \mu_i(X)), x \in X\}$

The membership function consist with different shapes as follows:

1. Triangular membership function as in figure 3

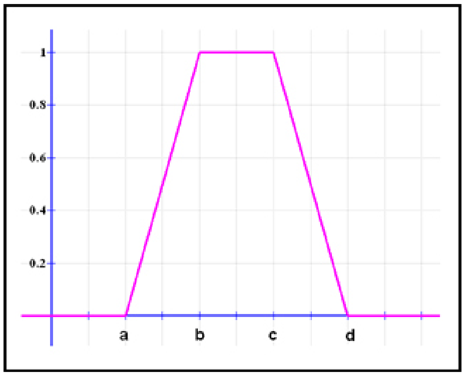


$$\mu_A(x) = \begin{cases} 0, & x \le a \\ \dfrac{x-a}{m-a}, & a < x \le m \\ \dfrac{b-x}{b-m}, & m < x < b \\ 0, & x \ge b \end{cases}$$

Figure 3 Triangular membership function

where; a = lower limit, b = upper limit, m = is a value , where a < m < b

2. Trapezoidal as in figure 4.

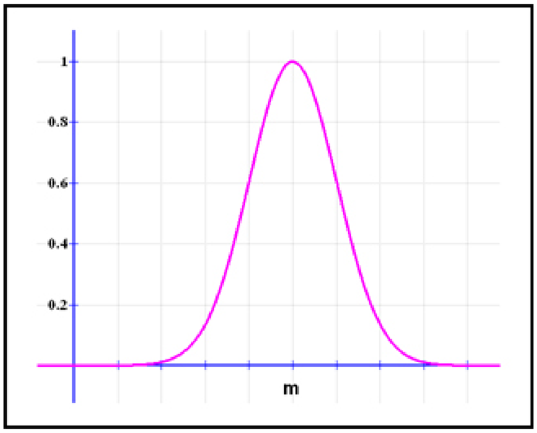


$$\mu_A(x) = \begin{cases} 0, & (x < a)\ or\ (x > d) \\ \dfrac{x-a}{b-a}, & a \le x \le b \\ 1, & b \le x \le c \\ \dfrac{d-x}{d-c}, & c \le x \le d \end{cases}$$

Figure 4 Trapezoidal membership function

where: a = lower limit, d = upper limit , b = a lower support limit, and c = upper support limit, where a < b < c < d

3. Gaussian as in figure 5.



$$\mu_A(X) = e^{-\frac{(x-m)^2}{2k^2}}$$

Figure 5 Gaussian membership function.

Where; m = central value, a standard deviation k > 0

Some operations associated with fuzzy set:

Complement set $\underline{A}$: $\mu_{\underline{A}}(x) = 1 - \mu_A(x)$

Union $A \cup B$: $\mu_{A \cup B}(x) = MAX[\mu_A(x), \mu_B(x)]$

Intersection$A \cap B$: $\mu_{A \cap B}(x) = MIN[\mu_A(x), \mu_B(x)]$

1.2.4 Rough sets

With the recent development of Artificial intelligence, the problem of imperfect knowledge has become a problematic for many scientists. To understand and manipulate imperfect knowledge a successful approach was proposed by Pawlak in 1982 [3] as Rough set theory. It is capable of handling vagueness. This concept has been used in many research areas such as ML, image processing etc. The specialty behind this rough set theory is that it is not required for prior knowledge with respect to its data.

In Rough sets, it is important to understand about the information system, tuple and column.

An information system is a pair of (U,A).

- $IS = (U, A)$, where
- U is a non-empty finite set of objects called the universe.
  U= $\{x_1, x_2, x_3, \ldots\ldots, x_n\}$
- A is a non-empty finite set of attributes such that for every $a\epsilon A$.
  $a: U \rightarrow V_a$
- V_a is called the value set of a (or set of values of attribute).

The upper approximation is defined as the objects which are possibly belong to the target set while lower approximation is defined as the objects that positively belong to the target set. There are three disjoint regions available to represent the universe. The positive region contains objects are definitely within the set X. Boundary may or may not contain objects while negative region contains the objects which are definitely can confirm ad the members of X as shown in the figure 6.

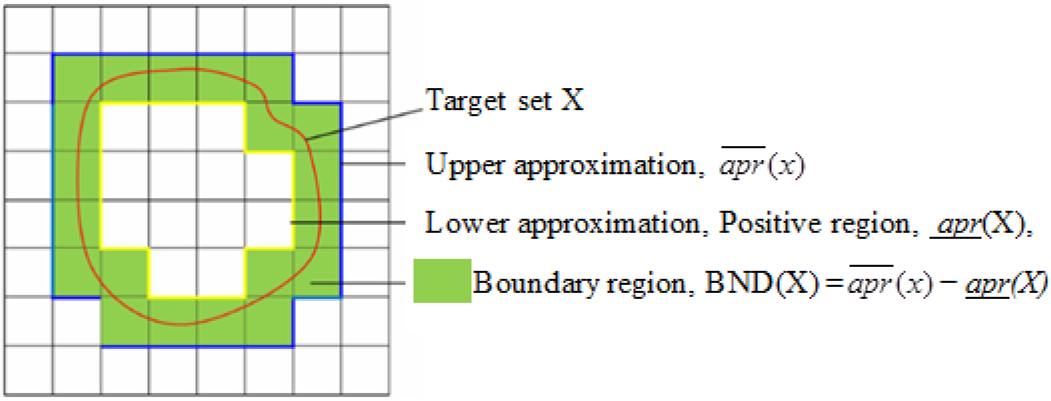

Target set X
Upper approximation, $\overline{apr}(x)$
Lower approximation, Positive region, $\underline{apr}$(X),
🟩 Boundary region, BND(X) = $\overline{apr}(x) - \underline{apr}(X)$

Figure 6 represents the rough set and its three disjoint regions.

1.2.5 Fuzzy rough sets

Fuzzy set and Rough set theory complement each other. Dubois and Prade in 1990 [4] introduced the hybrid concept for both rough and fuzzy sets. It composes the fuzzy-rough set which concerns related however distinct concepts of vagueness and indiscernibility [5]. This approach gives tremendous benefits to handle knowledge acquisition in information systems with fuzzy decisions. With the equivalence relation, fuzzy sets turn to rough and Fuzzy-rough sets then allocate the lower and upper approximation. Further, the above equivalence relation can be converted to an equal fuzzy relation. It is very clear that the rough sets are based on crisp equivalence classes while fuzzy-rough sets are based on fuzzy equivalence classes [5].

In the field of granular computing this theory becomes more popular since each other can be benefited by working together. Therefore, this can provide flexible and solid solutions in data imperfection whilst providing a successful solution in the field of data analysis.

In fuzzy-rough sets the equivalence class is fuzzy. In addition, fuzziness is introduced in the Output classes too.

The reflexivity, symmetry, and transitivity associated with Fuzzy equivalence classes. Here the s is considered as a fuzzy similarity relation.

- reflexivity : $\mu_S(x,x) = 1$
- symmetry : $\mu_S(x,y) = \mu_S(y,x)$
- Transitivity : $\mu_S(x,z) \geq \mu_S(x,y) \wedge \mu_S(y,z)$

Fuzzy-rough sets

The lower and upper approximation of the fuzzy rough set is really important and many research papers have highlighted those concepts in detail [5, 6, 7].

$$\mu_{\underline{P}X}(F_i) = inf_x max\{1 - \mu_{F_i}(x)\,, \mu_X(x)\} \quad \forall i$$

$$\mu_{\overline{P}X}(F_i) = sup_x min\{\mu_{F_i}(x)\,, \mu_X(x)\} \quad \forall i$$

The fuzzy lower and upper approximations can be defined as:

$$\mu_{\underline{P}X}(x) = \sup_{F \in \mathbb{U}/P} min\left(\mu_F(x)\,, \inf_{y \in \mathbb{U}} max\{1 - \mu_F(y)\,, \mu_X(y)\}\right)$$

$$\mu_{\overline{P}X}(x) = \sup_{F \in \mathbb{U}/P} min\left(\mu_F(x)\,, \sup_{y \in \mathbb{U}} min\{\mu_F(y)\,, \mu_X(y)\}\right)$$

The tuple <$\underline{P}X$,$\underline{P}X$> is called a fuzzy-rough set [8].

1.2.6 Attribute selection

The attribute selection task plays a vital role due to the generation of huge amounts of real-valued data in social media. This is defined as “the process of finding a best subset of features, from the original set of features in a given data set, optimal according to the defined goal and criterion of feature selection (a feature goodness criterion)” [9]. The selected subset of originally available attributes which is more relevant and less redundant can be subsequently used in the model. A littler subset of attributes progresses the performance as the search space is little. This also decreases the hazard of overfitting. Additionally, this method is associated with creating a lesser number of off-base choices and meeting with deluding with this approach. Less attribute set is simpler to examine and more productive and more secure to apply. Rough set based approach [3, 10, 11] is one of the foremost vital methods used for attribute determination that obtains information from the dataset which does not require outside data for attribute since it handles the vagueness within the data framework. However, this strategy associated with many limitations associated with uncertainty and proposed approach of fuzzy rough set by Dubois & Prade settle both uncertainty and vagueness associated within the dataset [12].

1.3 Problem

For a democratic country, elections contributed in many ways such that citizens chose a person to hold public office. There are various types of elections such as: presidential election, parliamentary election, provincial council election, local authorities' election etc. in Sri Lanka. As election is very famous the same way election prediction is again not a new keyword. Still there is a challenging task to predict accurate results.

To enforce the feasibility and reliability of a system that will help in political decision accurately has emerged. It is common to utilize online social substance for analyzing political occasions. Tweets can be taken as a representative of the citizens' opinion to test the society which carries a positive, negative or neutral estimation towards the election and the candidate.

Many researchers have utilized tweets and other social media information to analyze and identify trends in politics. Foreseeing the election result is a significant political milestone which helps the politicians to effectively utilize their campaign budget in a fruitful way. Election-related social media information can be very complex and deluding which cannot be effectively decided from their online action. Most of the existing literature needs an orderly treatment of the issues concerning social media information. Researchers have recommended that stakeholders of social media may be less vulnerable to social desirability than participants in a survey when they are examining their political inclinations. Be that as it may, opinions of the twitter client base may not represent the people who have the voting power but it is still beneficial to mine conclusion from Twitter since it represents a certain segment of the population where their voices play a vital role by making influences for the opinion of the wider range. The previous literature needs a dependable information gathering strategy as the information mined from social media isn't examined consistently, and consequently may not precisely speak to the entire segmentation of online clients. A few of the algorithms utilized by the past researchers like NB classifier have numerous areas recognized with moving forward and due to its unreasonable assumption of independence numerous people have suggested various improvements with distinctive aspects.

Therefore, this proposes an approach on developing a novel model to analyze the social network feeds and sentiment information mined using ML models to predict election outcomes by overcoming the limitations of social media data and other identified drawbacks with existing approaches. The case study taken for the research is the presidential election of Sri Lanka 2019, GR and SP.

1.4 Motivation

SA of micro-blogging may be a generally modern research theme and have plenty of areas to be improved by doing various researches by considering the prior contributions by previous researchers. Decent amount of researchers have been conducted in the area of SA with document and blog article reviews, user reviews, etc. The finest outcomes achieved with supervised classification strategies such as NB and SVM. However, this requires the manual labeling which is exceptionally costly. A few work has been done on unsupervised and semi-supervised approaches and identified many improvements to be done in those fields. Different findings have been tested in this area and compare their results to base-line performance. There's a requirement of appropriate and formal comparisons between these results to choose the best classifier for specific applications.

There's an increased request for automated investigation of data which could be challenging. Many people propose supervised ML procedures prepared on bag-of-words concepts to classify the sentiment. Subsequently, above approaches highly depend on the domain and require huge annotated corpora. Frameworks on microblogging services like Twitter and other social communication stages are still very immature, in spite of the fact that numerous applications use the utilize of social media and particularly Twitter. After the victory of Barack Obamain 2008 at the U.S. presidential election, Twitter has ended up an authentic communication channel within the political arena with a new chapter. Hence, it can be anticipated that the SA has become a part of a political campaign today. This becomes more essential when it is needed to focus mostly with online campaigns in situations like COVID 19 pandemic.

Social media has the ability to influence political communication since they are apparatuses that can be utilized to illuminate and mobilize clients in better approaches. Users have the chance to associate politicians directly and campaign managers and involve with political activities in better and novel approaches. Maybe one of the foremost self-evident recent examples of online political campaigns is Donald Trump. In spite of the fact that it was certainly questionable, no one can deny that the U.S. President's campaign was greatly successful. Numerous parties claim that it helps him to win the election. This has become a hot topic among every party and discusses the utilization of social media as a new marvel in communications studies. As a result of that, political campaigns presently utilize social media to set up the candidate's political character, to teach and pull in voters, and to spread their political information. The predominance of social media in politics makes candidates more reachable to voters that are really important. The capacity to distribute information and broadcast it to millions of individuals immediately permits campaigns to be more successful. The candidates have the chance to carefully build their images based on powerful analytics in real time with zero cost.

In order to have a successful digital campaign, understanding the sentiment of people towards the candidate is really important. It helps to take a general opinion about the position of the candidate among people. Therefore, the effective classification technique is essential for the SA of peoples' opinion towards the candidates of the Sri Lankan political context.

1.5 Research Contribution

1.5.1 Goal

The goal of this project is to develop a sentiment based classifier using ML and fuzzy-rough set theory expecting accurate results. Further, it carries automatic sentiment classification with twitter corpus collected during a certain time period with regard to the case study for the prediction of results at the presidential election 2019, Sri Lanka.

1.5.2 Objectives of the study

To achieve the above mentioned goal, the following objectives have been identified.

- Develop a model for pre-processing of harvested un-structured noisy data from Social networks.

- Analyze ML algorithms used in other classification models and evaluate their suitability for the problem of classifying data for SA.

- Analyze the performance and accuracy of Rough Sets and Fuzzy Rough sets in classification models.

- Develop classifiers using the algorithms identified above and extract features that will allow them to classify sentiments into the positive, negative or neutral.

- Analyze the performance of new model with existing models.

1.6 Scope

The data collection is based on political tweets distributed on Twitter's public message board some time recently the presidential election 2019 in Sri Lanka. Thereafter, uses the hashtags for crawling of tweets to gather all related information. Tweets comprise with emoticons and they are utilized for opinion mining. Only the tweets in English language are considered for the study but this can be expanded for the other languages as Twitter API permits to indicate the dialect of the recovered posts. Raw tweets are full of noise as well as an obscure dataset. Hence, the data must be normalized to make a dataset which can be effectively used with different classifiers. Therefore, data preprocessing and cleaning steps needed to be used to standardize and optimize the dataset. The preprocessing steps includes removal of punctuations and special characters, tokenization, stemming and many other steps that are discussed under the methodology chapter.

The simplest method of SA is the finding of the presence of single words or tokens in the corpus. However this classification is further improved with the unigrams, bigrams and N-grams features along with the comparison among each other. Fuzzy-rough set based classifiers will be used to classify tweets as the ML algorithm. The same data set will be classified with the use of traditional NB and SVM algorithms in order to compare the results with the newly identified classifier.

Evaluation metrics will be used to measure the performance in order to identify the accuracy of the novel classifier with other specified ML algorithms against.

1.7 Organization of the thesis

This dissertation consists of five and used figures, tables and charts to depict the information in a very concise way. Provides a comprehensive description about the problem domain along with the scope. The rest of the chapters of this dissertation consists of a plethora of descriptions of the study much deeper as given below.

In chapter 2, literature review will be presented with SA, classification methods, algorithms used, political SA techniques, twitter SA with political context will be discussed together with the challenges of existing research.

Chapter 3 consists of the methodology which explains the steps taken to complete and achieve the targets of the case study. The identification of data collection platforms, data collection methods, pre-processing and data cleaning methods, classification techniques will be discussed with this chapter. The used twitter APIs and the collected corpus about the presidential election 2019 of Sri Lanka will be presented along with testing and training dataset.

Chapter 4, discusses experiment results and the evaluation of the case study. Several ML algorithms will be discussed including rough-fuzzy classifiers, NB, SVM. Performance evaluation and challenges arise in this chapter. Finally, this chapter evaluates the results, improvement and the accuracy compared to the existing classifiers.

Finally, in chapter 5, analysis and conclusions are provided with final comments and thoughts about the study.

Chapter 2

Literature review

2.1 Introduction

There are significant research contributions on opinion mining with social media data, review sites, blogs, opinion polls, surveys and interviews. Most of the approaches are associated with ML approaches and the unlimited data would be facilitated with the help of above resources. This literature review chapter focuses on the techniques used for SA, challenges and future directions for the researchers.

SA has been conducted over a period of time with different topics by many researchers for instance: stock market [13], movie reviews [14], product reviews [15], news and blogs and election reviews with political data [16]. Here discusses some of the SA concepts with different approaches.

There are many approaches available for the analysis of social media data in order to obtain the hidden sentiment such as: lexicon based approach, ML based approaches and Hybrid based approach [17] as shown in figure 7. Lexicon based approaches identify a sentiment lexicon and use it to describe the polarity as positive, negative or neutral. It should be noted that the active engagement of a human being is required at the analysis phase. This approach can be further divided into two categories as: Dictionary based approach and corpus based approach. Further, machine learning approach can be divided into supervised, unsupervised and semi-supervised learning and it requires a large data set to be effective which is the main drawback. This provides more accuracy compared to the lexicon based approach where many researchers focused on NB classification and SVM under the ML approach. Hybrid approach combines both ML and lexicon-based methods and has plenty of research areas that no one has touched yet. Supervised ML techniques have shown relatively better performance than the unsupervised lexicon based methods [18].

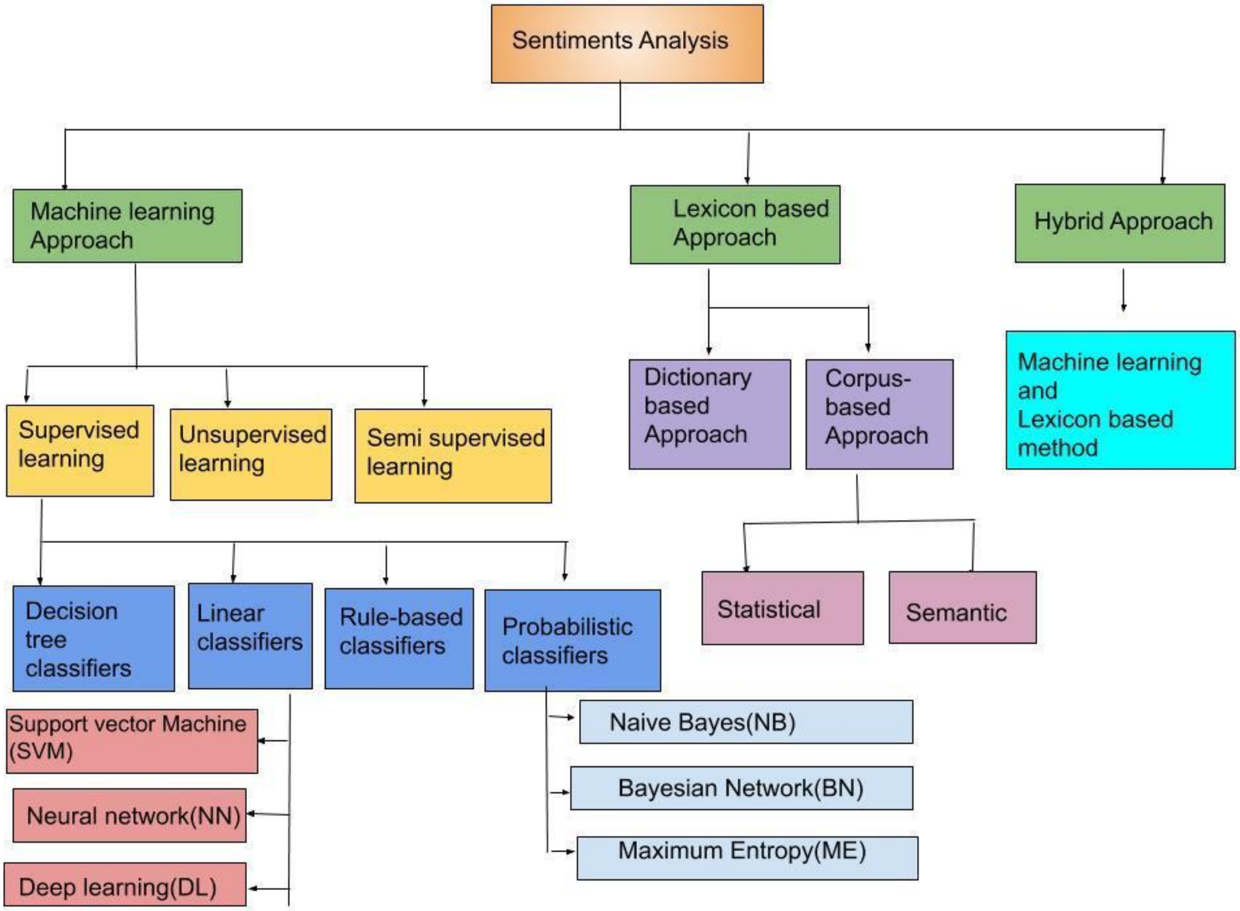


Figure 7 SA Methods

Srinivas *et al.* [19] have exploited a rough-fuzzy classifier that combines rough set theory with the fuzzy set to predict heart diseases. The rule generation has been conducted with the use of rough set theory and prediction has been done using fuzzy classifiers. Some of the features associated with the development of the classifier have been automated and rule generation is one of them. The attribute identification is also carried out automatically. Further, it uses the rough sets and rules applied for the fuzzy classifier in order to predict heart disease. Their developed classifier is based on rough sets and fuzzy sets whilst outperformed with 80% of accuracy with the Switzerland datasets. Keerthika et al. [20] attempted on developing a Fuzzy Temporal Rule Based Classifier with the use of fuzzy rough set along with temporal logic. This classifier is developed to identify the temporal patterns in the medical domain databases. The classifier was compared with the FNN (Fuzzy Neural Network) and it can be concluded that the accuracy of TFRBC is 88 % when compared with the FNN which is 74%. As a future direction, the rule based classifier can be improved with effective decisions by taking upper approximations while reducing the number of rule sets when compared with the lower approximations.

Srividya and Sowjanya [21] discussed a methodology to analyze collected reviews over a period of time in Facebook using NB classifier into relevant and non-relevant to determine public opinion on the popularity of android based and identified the most preferable versions of OS for android phones and found Android KitKat is more preferable than others. They have shown that there are many areas to be improved with the NB classifier in different areas.

Tweets have been used by Pritee and Sachin [22] to predict the outcome of the election results. The data collection was conducted by taking the tweets that are matching to certain keywords and by collecting all the tweets with the help of twitter streaming API. In order to do the sentimental analysis, the collected data were pre-proceed with lower case conversion, punctuation and number removal, stemming and striping white spaces. NB, a supervised ML approach for classification used on the above training data set including emotions for the sentimental analysis. The dictionary based approach with eleven lexicons used for the classification of identified variables. The case study of US and Gujarat Rajya Sabha elections were analyzed based on the above approach. They have found that the use of twitter data for the political SA can be done successfully in order to make predictions in an election. Further, it can be observed the likelihood of the voters and how they would influence others who have the voting power. Finally, it can be concluded that the above findings can be utilized to analyze the sentiment of the final results of the election before commencing.

Election forecasting has a political side which helps in forecasting the results. In order to forecast an election there are opinion polls as well as used and many scientifically proven statistical models [23]. However sometimes polls also fail in predicting the results of the election even in the developed countries. According to [24], it listed several failed polls results such as in the 1992 British General Elections, French presidential etc.

Electoral analysis of Twitter data is straightforward and optimistic even though it is a challenge for the research community in today's world. Wong et al [25] used Twitter streaming API to collect tweets which contain the identified keyword relevant to the events identified before the election and used lexicon-based SA package, to extract the sentiment of tweets as a ternary (positive, negative, neutral) classification. They used the consistency relationship between tweeting and retweeting behavior. For each tweet they set the value either 1,-1 or 0 based on the relevance and did not consider intermediate values. Vadivukarassi et al [26] proposed a model to analyze Twitter data where the data streaming is done using both streaming and search API to access the real-time feeds and archived data for analytics. The above extracted raw data are preprocessed using Natural Language Toolkit techniques. The word scores of the features are tested based on Chi-square method and key words were scored sentiments during the analysis of data. NB classifier is used for training and testing the features and also evaluating the sentimental polarity hence this helps to obtain the summarized report about the opinion from Twitter.

Pak and Paroubek [27] developed a classifier to do linguistic analysis on twitter corpus with English language only. Further they have mentioned that this approach can be used for the other languages as well. However, the developed classifier is capable of predicting the sentiment of the tweets as positive, negative or neutral. For the collection of corpus they queried twitter for both happy and sad emoticons and trained the data set with positive and negative classifiers to identify the positive and negative sentiments. For the analysis of the corpus, they used a plot of word frequencies with Zipf's law to understand how terms are distributed across collected corpus and TreeTagger (Schmid, 1994) [28]. This is basically associated with the part-of speech and lemma information. Under the training of the classifier they followed the method of filtering, tokenization, removal of stop words, n-grams feature with tweets. As the classifier NB was selected and it outperformed with best results. The accuracy was obtained by calculating the entropy (Shannon and Weaver, 1963) [29] and calculating "salience" for all n-grams. During the process, they trained the classifier and determined the sentiment of twitter corpus using multinomial NB.

Prabhu et al. [30] proposed a methodology to distribute political party's tickets during the election with twitter data. Upon receiving all the necessary data related to a candidate, NB Algorithm used to predict the most deserving candidate. Their final conclusion was that this classifier is suitable for any type of election in India and elsewhere.

Nowadays, this area has become popular in broad applications in various fields. The idea of this is to categorize the texts with the use of NLP algorithms as positive, negative or neutral together with some emotional feelings such as love, anger, kindness etc. *Rao et al.* [31] identified the importance of identifying the general sentiment polarity of a news article before publishing. They use a ML approach based on feature based classification model on Twitter data to classify them as popular/unpopular classes. The model extracts several features from the collected twitter data and uses NB algorithm and SVMs to train the classifier. Then the model used to predict the popularity of the news. The accuracy ensured with precision and recall together with the unigram, bigram and hybrid features. Hybrid features with SVM classifier outperform among other classifiers which use various features.

Wei and Gulla [32] present an analysis technique based on a tree of feelings of ontological features. The product's attributes labeling is handled with the novel HL-SOT approach. Hierarchical Learning (HL) process used for analyzing their associated sentiments in product reviews. Further they used a defined Sentiment Ontology Tree (SOT) with the above process. The HL-flat approach ignores the hierarchical relationships among labels when training each classifier and this is a "flat" version of HL-SOT. The H-RLS algorithm only uses identical threshold values for each classifier in the classification process where HL-SOT enables the threshold values to be learned separately for each classifier in the training process. The research found that the HL-SOT approach outperforms two baselines: the HL-flat and the H-RLS approach. The classification approach used is based on hierarchical classification algorithms. The classification approach used is based on hierarchical classification algorithms. Neviarouskaya *et al.* [33] article describes a method "SentiFul" the segmentation of the text into certain levels.

Pang et al. [1,34] investigated the performance of NB, Maximum Entropy and SVM in SA on different features. They used n-gram features, part-of-speech and position information by considering adjectives with ML approaches. They have identified a number of outcomes based on the study. The presence of the features in the corpus is more important than its frequency. The bigram features decrease the accuracy while the part-of-speech increases. The use of position information also increased the accuracy compared with only use of adjectives. Further they have observed that NB performs well with a lesser number of feature whilst SVM perform better when the feature space is increased.

Part of Speech tagging (POS) is important in SA analysis as words may carry different meanings with different parts-of-speech. Turney [2] gets the use of this approach with trigram features. He allocated a point based system to cater the semantic orientation built with phrases.

Hao et al. [35] proposed a crowdsourcing platform to develop the SA system during the presidential election US, 2012by using the Amazon Mechanical Turk (AMT) as the baseline dataset. The developed system consists of sentiment labeling together with sarcasm, humor. The model performed well with the category of SA analysis.

Mostafa and Mohamed, [36] used an expert-predefined lexicon to analyze the sentiment of the consumer brand. They used the most famous microblog Twitter to collect data for the study. They used various techniques to guarantee the representativeness by randomly selecting the corpus. They focused on the lexicon-based method for sentiment orientation. They use the R software package for their analysis. Finally, used the StreamGraph in order to visualize the trend associated with tweets over a period of time.

WordNet was used by Kamps et al. [37] to observe degree of association with the emotional content with a word based on various dimensions. For the SA, Xia et al. [38] used an ensemble framework with the use of features and classification techniques. POS tagging and Wordrelations were used as feature sets. NB, Maximum Entropy and SVM used to construct the classifiers. Further, they used the weighted, fixed and meta classifier combination as ensemble methods.

It has been observed that, at a certain place the accuracy of the classifiers remains as it is even though adding a huge training data set. Barbosa et al [39] used a labeled data set for training the classifier. They have observed that the accuracy is higher if the training labels can achieve more than 50% accuracy. Further they have examined that accuracy of the classifier is increasing with the number of labels increases. When the data set becomes larger and larger the noisy associated with the labels also increases. To overcome that, they have used a large number of tweets for this experiment.

A study has been conducted [40] on Pang Corpus which is a movie review database and opinions collected from the website Epinions.com named as Taboada Corpus to train a sentiment classifier with SVM together with n-grams and different weighting schemes: Term Frequency Inverse Document Frequency (TFIDF), Binary Occurrence (BO) and Term Occurrence (TO). Further, it uses chi-square weight features to select informative features. It is evident that the chi-square feature selection improves the accuracy of the classification. Further, it shows that the unigrams outperform the other n-gram models for both datasets.

An optimized classifier proposed by Bhumika and Vaghelawith [41] evident that SVM outperformed than the other existing systems. The study was based on movie review, twitter and gold dataset. Researchers also found that the performance of SVM depends upon the dataset as well as the ration on training and testing data set [42].

A case study conducted to carry qualitative analysis on social media networking websites related to political leaders to identify different sentiments [43]. The results show that Tuning Multinomial NB performs better than NB. Another study of twitter data [44] which focuses on presidential elections in Egypt 2012 was conducted and find out that NB scores the highest accuracy and the lowest error rate for Arabic text classification. Ringsquandl and Petkovic [45] conducted a study to analyze the presidential candidates of the Republican Party in the USA based on their campaign topics. It found out that data retrieval methods should be smoothed, pre-processing steps are really important, and Natural Language Toolkit (NLTK)'s functionalities are not sufficient. As future work they propose to collect information other domain-specific opinion. Kassraie et al. [46] conducted a research on predicting the US 2016 elections results and found that there are people who tweet do not have the voting power. Further they found that the social media isn't always reliable.

Raghuwanshi and Pawar [47] performed a comparative study on NB, SVM, and Logistic regression with crowd source information that compares linear and probabilistic approach. The results revealed that SVM turns out to be best among all and can work with linear or non-linear data. Processing and extracting exact emotions are two major areas to work in this field, need more efficient ML, deep learning algorithms for better classifiers, a lot of ways to deal with spam posts/tweets and use better mining techniques to deal with NLP more efficiently.

A web crawling framework to facilitate the quick discovery of sentimental contents of movie and hotel reviews conducted with two supervised ML algorithms: K-Nearest Neighbour (K-NN) and NB [48]. For movies review, NB is better than K-NN. However for the hotel reviews it acts negatively. The researchers suggested testing the results with random forest, SVM etc. Further they suggested combining and building a new one using the benefits of both.

One of the challenges with SA is the large feature space. Chena et al. [49] proposed a way for drug reviews using fuzzy rough feature reduction. They used Random Forest algorithm with fuzzy-rough QuickReduct feature selection. They found that the Fuzzy Rough Feature Selection (FRFS) showed good results. As future development they want to guarantee about more accuracy with NLP.

A fuzzy rough set-based feature selection algorithm has been used for hierarchical feature selection with sibling strategy and revealed that it is more efficient and more versatile [50]. Further it shows that the classifier resulted with higher performance with establishing the efficiency and effectiveness. It opens new research trends by combining fuzzy rough sets with hierarchical feature selection problems.

Many researchers now combine fuzzy rough set theory with other technologies to improve classifiers. Genetic Search Fuzzy Rough (GSFR) is one of the feature selection algorithm used by [51] using the evolutionary sequential genetic search technique with fuzzy rough set theory to early identification of cancer. This is an extended approach of Fuzzy-rough nearest neighbor (FRNN) classifier. This classifier outperforms with number of features, accuracy, and precision, recall, F-measure and computation time compared to other classifiers. This research opens new doors to hybridize fuzzy rough sets with both Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) in order to improve further.

A new approach based on FRNN proposed by R. Jensen and Cornelis [52] and tested with nine data sets. FRNN-FRS uses the traditional operations, a t-norm and an implicator. FRNN-VQRS is the fuzzy quantifier-based approach used. It resulted that, FRNN outperforms both FRNN-O, as well as the traditional Fuzzy Nearest Neighbour (FNN) algorithm. This research opens new areas by providing explanations of the impact of the choice towards the fuzzy relations, connectives and quantifiers. The accuracy of the classifier upon the feature selection and preprocessing is another new research area arising from this.

The above-discussed methods used the concept of the fuzzy rough set to generate the minimal reduct set.

2.2 Conclusion and Novelty

As per the above literature review it has been explored that the SA has used to find out the expressed opinion of a document or a text is positive, negative or neutral. With that, the need of a system for interpreting the public sentiment towards various entities has emerged. Researchers have used various approaches with different algorithms such as SVM, NB, Maximum Entropy, NBSVM etc. in SA. To cater the above identified requirements, many scholars conducted researches by contributing a lot in the domain of SA as mentioned in the above literature. However, there exist many gaps to be filled in this area. Consequently, SA has become a popular research area among the academics as well as along the business organizations.

SA is a challenging research area with many improvements and investigations should be done. Some of them related to twitter have been listed in this paragraph. One major problem associated with text is, one place it plays as subjective but some other places as objective in the same text. The meaning of the sentences can vary with respect to the domain it is associated with. Sarcasm detection is another problematic area where many researchers currently actively contribute. People can pass their negative opinion by using positive words towards a target group. Thwarted expression [53, 54] is another challenge with natural languages where a small part of the document/text will carry the overall opinion of the entire text/document. With the bag-of-words approach this will give the opposite opinion of the polarity as that approach is basically aligned with the number of words appeared in the document. Negation handling is again difficult in both implicit and explicit cases. The order dependence with the words in a sentence also makes wrong opinions about the corpus. Many of the researchers focused their experiments on the English language only. There is a room for all the researchers for the experiment of SA where the text involves emotional status and attitudes of people and expressed well in their native language. The bag of words model also carries some problems. The total number of dimensions with the bag of word model is equal to the size of vocabulary in the corpus. To cater this, researchers can use dimensionality reduction methods effectively. Further, it does not consider the neighbour words where that is useful with semantic relation of nearby words.

Most of the algorithms used in ML approach belong to supervised classification and provide higher accuracy and performance [55, 56]. Therefore, Research results show algorithms such as SVM, NB, Maximum Entropy have the highest accuracy. In some cases, lexicon-based methods perform well however it needs humans to be involved in the process. The supervised ML method needs the training data set which is labeled by humans manually. It is a tedious task associated with supervised learning methods as well as expensive. With considering the above mentioned facts, there is a lot of room for improvement in the context of supervised ML as well. As mentioned in the literature review, the more accuracy can be obtained with more cleaner data. Therefore, the optimizations of the data pre-processing steps are always important to improve the accuracy of the entire classifier. There is another area where researchers can focus on is the use of various features associated with the dataset. The use of ngram models still needs more improvements as it acts differently in different places. Some researchers found that the bigram model performed well compared to others as observed in the literature survey. Furthermore, it requires a common framework to analyse the results of different classifiers in order to select the best one. Hence, new researchers can focus on the study of combining ML methods into opinion lexicon methods to obtain better accuracy. Next chapter explains the steps followed in identifying and applying best methods to address this study.

Chapter 3

Methodology

3.1 Introduction

The methodology speaks the way that the research question has been taken care of and attended to with the information picked up in the literature review. This project spread over numerous computer-enabled scientific areas. As referenced before, the extent of this research project spread over various cutting edge technologies to achieve the goal of the project. Therefore, this chapter is to present the comprehensive overview on the steps followed in order to successfully achieve the goal. Therefore, this chapter describes the developed classified data collection process, data pre-processing steps and all the other steps involved in developing the classifier.

3.1.1 Representation of the problem

The aim of the study is to develop a sentiment based classifier using fuzzy-rough set theory and train it with labeled twitter data to retrieve accurate results with unknown tweets. This process is to develop an automatic sentiment classification method and the twitter data will be collected by taking the case study for the prediction of results at the presidential election 2019, Sri Lanka. On 16th of November 2019, is a special day for all the Sri Lankans as they walk towards the polls to elect a new president among 35 candidates resulting in a big competition. However the fight is with two famous candidates: The United National Party's (UNP) SP and GR from the Sri Lanka People's Front (SLPP) party. According to the statistics of the Election Commission, the participation of voters during the election is high. The voters mark their first and second choice in their ballots. The second choice is for in case a candidate did not receive 50% of the total votes then the second preference will be taken into consideration. The election day is 16th November where the voters cast their vote for their favorable candidate and 17th November 2020 will release the results by announcing the winner.

There is a big trend for finding methods on classifying huge twitter datasets using SA and ML techniques. SA is a difficult task since it totally depends on the domain and the culture it relates to. Therefore, the prediction of the winning party before the election by using twitter data is a tedious task. However, if can predict that earlier, that can bring tremendous benefits for each political party to adjust their campaign strategy. The solution for this is composed with two major strategies. As the first step, the classifier will be trained with known labeled data and classify them into a pre-known number of classes. At the second step, classifier will take unlabeled twitter data sets and classify them into relevant previously identified classes. Now the classifier is capable of predicting the sentiment towards the political candidate as negative, positive or neutral. The comparison of this has to be carried out for the GR and SP. In this case all the unclassified tweets will be classified into one of pre-known class according to the likelihood it gains from user given, pre-known tweet set. A comparison will be carried out towards the two politicians SP and GR.

3.2 The process of the framework

Figure 8 depicts the flow chart of the general pipeline of the framework proposed to develop the sentiment classifier. The right hand side describes the learning phase and how the sentiment classification model is developed and trained with collected twitter data using the sample training data set. Then a data preprocessing will be carried to clean the dataset. Then the text features will be extracted with TFIDF vectorization where the numerical values will be assigned for all the features. Thereafter, the feature reduction will be processed to get the minimum number of features aligned with the feature reduction step in order to achieve higher performance. Once trained, the same used features during the training process will be taken at the prediction step as well. The testing phase and the prediction phase is simpler than the training phase. The left hand side is representing the steps in the prediction process of the unlabeled tweets.

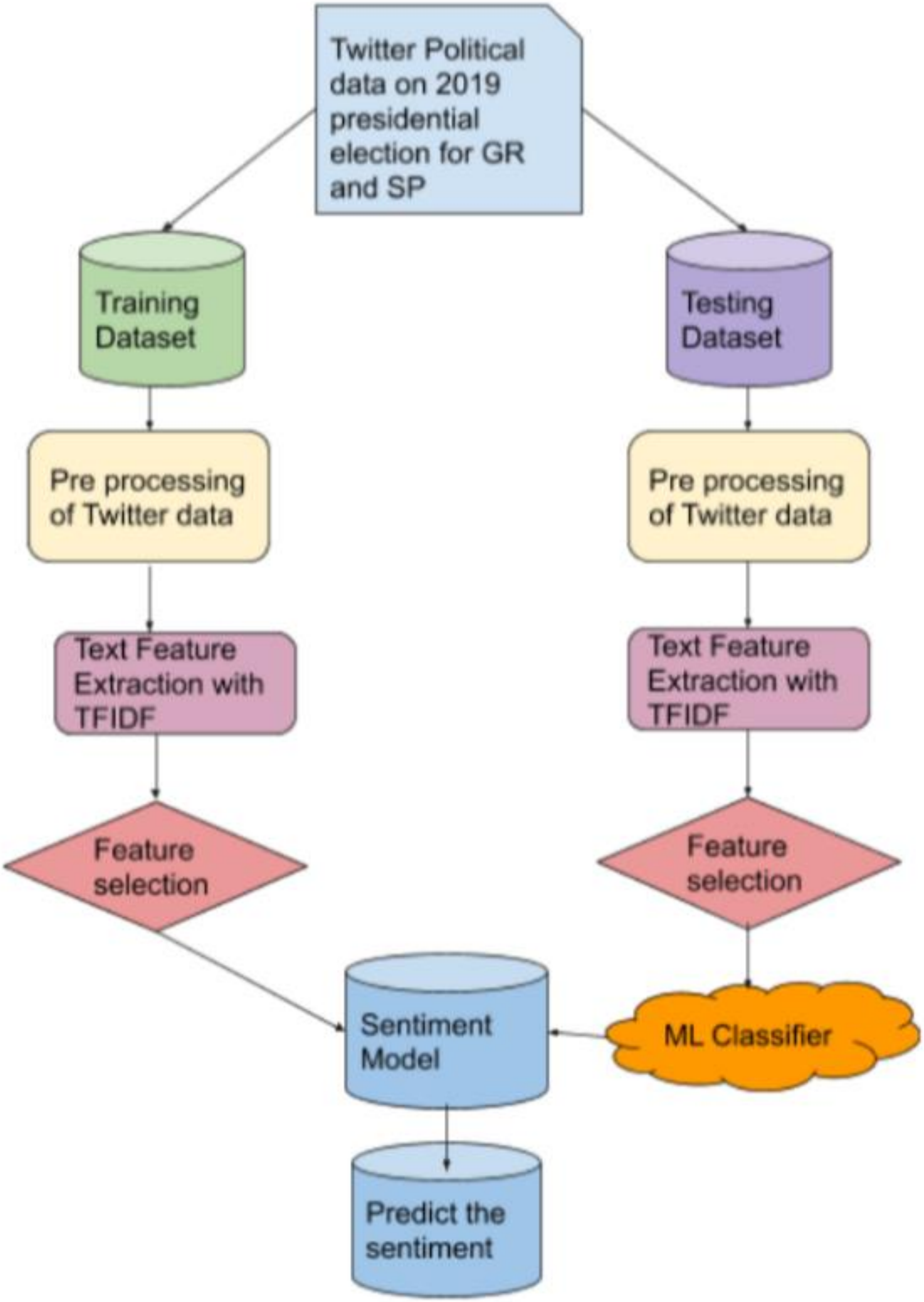

Twitter Political data on 2019 presidential election for GR and SP
Training Dataset
Testing Dataset
Pre processing of Twitter data
Pre processing of Twitter data
Text Feature Extraction with TFIDF
Text Feature Extraction with TFIDF
Feature selection
Feature selection
Sentiment Model
ML Classifier
Predict the sentiment

Figure 8 the general pipeline of the framework

3.3 Develop the classifiers

All the steps identified above in the above diagrams show the steps to develop the classifier and discuss them in detail under this section.

3.3.1 Raw data collection

At the data collection stage, tweets were selected as the training and testing data to train the classifier and predict the results of the outcome of the case study. For the collection of older and live tweets uses two twitter APIs respectively Streaming and Search APIs. The collected tweets were manually annotated as positive, negative or neutral towards the candidate.

The twitter data set is collected with APIs mentioned above between 01st September and the 15th of November 2019 (two months before the election) based on the main political candidates GR and SP with the use of hash tags.

It requires API key, API secret, Access token and Access token secret in order to download tweets using twitter APIs. The following steps were followed to obtain them.

- Browse the twitter developer web site https://developer.twitter.com/
- Login to the twitter account using the previously created login credentials.
- Create a twitter developer account using “Apply for a developer account”
- After filling all the requested information and agreeing upon the agreement the account will be granted after the review process.
- Once approved, the "API key" and "API secret" can be obtained under the "API keys" tab.
- Then click "Create my access token" to get the "Access token" and "Access token secret".

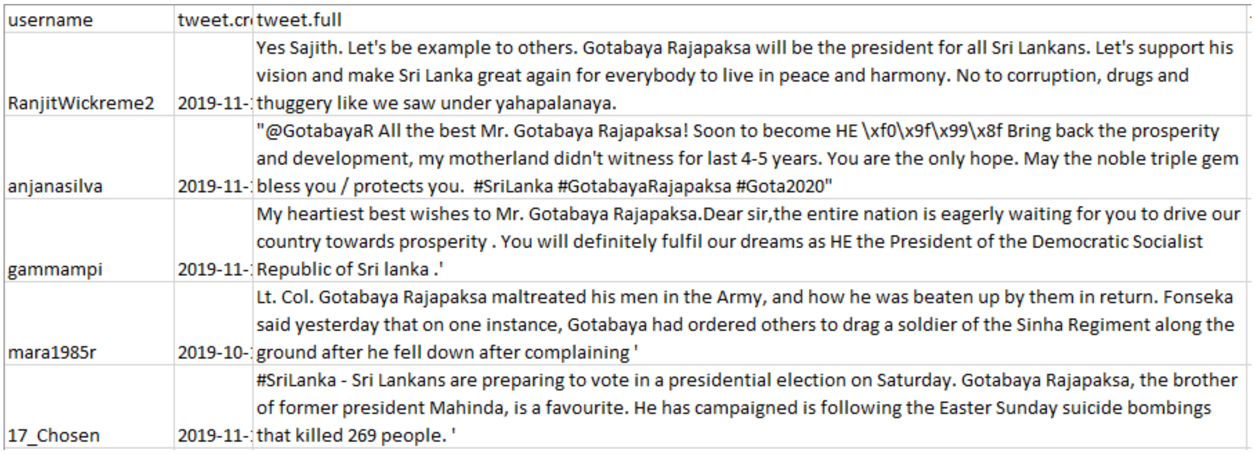
Then used the Python library called Tweepy to connect Twitter Streaming API to obtain the data set. Further, it uses the command line utility of GetOldTweets3 Python 3 library to access old tweets. For the experiment purpose we use only the tweets in English and all the other non-English tweets were discarded.

Dataset was obtained from twitter from 01st October 2020 to 10th November 2020 for both GR and SP using twitter API. The collected data set consists of 40000 tweets for GR and 40000 for SP. From the above data set 1000 tweets were randomly selected for SP and another 1000 for GR to use for training and testing dataset. Another 1000 from each data set was selected to be classified using the developed classifier. Table 1 shows the number of tweets obtained, training data set, testing data set and predicting data set in relation to GR and SP.

Table 1 Famous residential candidates, Sri Lanka, 2019 and the number of corresponding tweets obtained during the period of 01st October 2020 to 10th November 2020

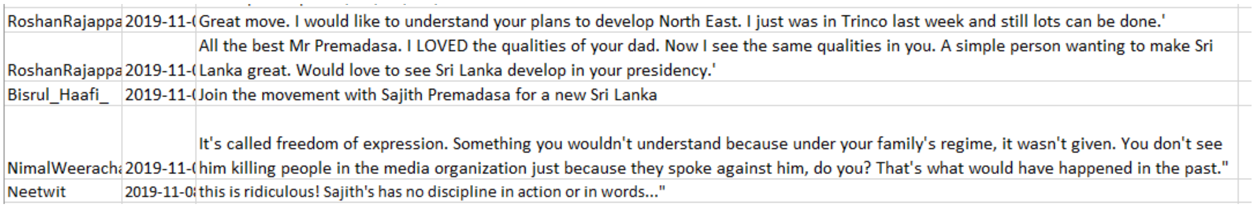
Description	**No of Tweets**	
	GR	**SP**
Extracted tweets	40000	40000
Training dataset	700	700
Testing dataset	300	300
Data set to be classified	100	100

The dataset is taken to the CSV file format as it is the input format for the classifier. Some of the tweets obtained from twitter data with respect to the GR is shown in the figure 9 and SP in figure 10. Those tweets must be preprocessed before using for the classification.



username	tweet.cr	tweet.full
RanjitWickreme2	2019-11-	Yes Sajith. Let's be example to others. Gotabaya Rajapaksa will be the president for all Sri Lankans. Let's support his vision and make Sri Lanka great again for everybody to live in peace and harmony. No to corruption, drugs and thuggery like we saw under yahapalanaya.
anjanasilva	2019-11-	"@GotabayaR All the best Mr. Gotabaya Rajapaksa! Soon to become HE \xf0\x9f\x99\x8f Bring back the prosperity and development, my motherland didn't witness for last 4-5 years. You are the only hope. May the noble triple gem bless you / protects you. #SriLanka #GotabayaRajapaksa #Gota2020"
gammampi	2019-11-	My heartiest best wishes to Mr. Gotabaya Rajapaksa.Dear sir,the entire nation is eagerly waiting for you to drive our country towards prosperity . You will definitely fulfil our dreams as HE the President of the Democratic Socialist Republic of Sri lanka .'
mara1985r	2019-10-	Lt. Col. Gotabaya Rajapaksa maltreated his men in the Army, and how he was beaten up by them in return. Fonseka said yesterday that on one instance, Gotabaya had ordered others to drag a soldier of the Sinha Regiment along the ground after he fell down after complaining '
17_Chosen	2019-11-	#SriLanka - Sri Lankans are preparing to vote in a presidential election on Saturday. Gotabaya Rajapaksa, the brother of former president Mahinda, is a favourite. He has campaigned is following the Easter Sunday suicide bombings that killed 269 people. '

Figure 9 Sample of tweets collected for GR according to the given hash tags

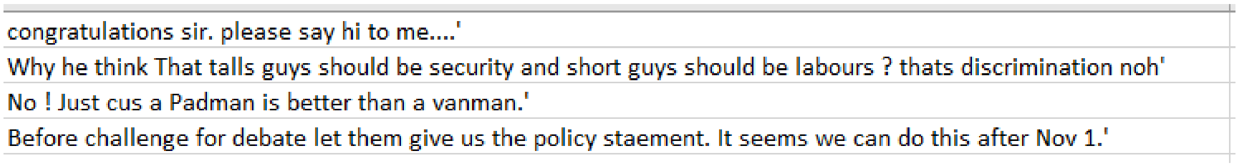


RoshanRajappa	2019-11-(	Great move. I would like to understand your plans to develop North East. I just was in Trinco last week and still lots can be done.'
RoshanRajappa	2019-11-(	All the best Mr Premadasa. I LOVED the qualities of your dad. Now I see the same qualities in you. A simple person wanting to make Sri Lanka great. Would love to see Sri Lanka develop in your presidency.'
Bisrul_Haafi_	2019-11-(	Join the movement with Sajith Premadasa for a new Sri Lanka
NimalWeeracha	2019-11-(	It's called freedom of expression. Something you wouldn't understand because under your family's regime, it wasn't given. You don't see him killing people in the media organization just because they spoke against him, do you? That's what would have happened in the past."
Neetwit	2019-11-0	this is ridiculous! Sajith's has no discipline in action or in words..."

Figure 10 Sample of tweets collected for GR according to the given hash tags

3.3.2 Lexicon development and Data preprocessing

In most of the social media media texts including tweets, people use very informal language to communicate with each other. Therefore, they consist of noisy data. Some of the texts are incomplete where the meaning cannot be extracted very clearly. Since all those are related to NLP the SA process has become more challenging. In the context of this political domain, people have used many slangs, misspelling, punctuations, words that are not in the dictionary etc. Therefore, it is very important to handle those impurities in the collected data set. Some of the examples from the collected dataset with above impurities is shown in the figure 11.



congratulations sir. please say hi to me....'

Why he think That talls guys should be security and short guys should be labours ? thats discrimination noh'

No ! Just cus a Padman is better than a vanman.'

Before challenge for debate let them give us the policy staement. It seems we can do this after Nov 1.'

Figure 11 Some of the collected tweets with impurities which need to be preprocessed.

Zhang et al.[57] used Stanford Parser Tools1 for POS tagging and parsing while the Natural Language Toolkit2 was used for removing stop words and lemmatization.

The preprocessing is done to prepare these datasets for experiment by removing noisy, inconsistent, and incomplete data by following the steps shown in the figure 12.

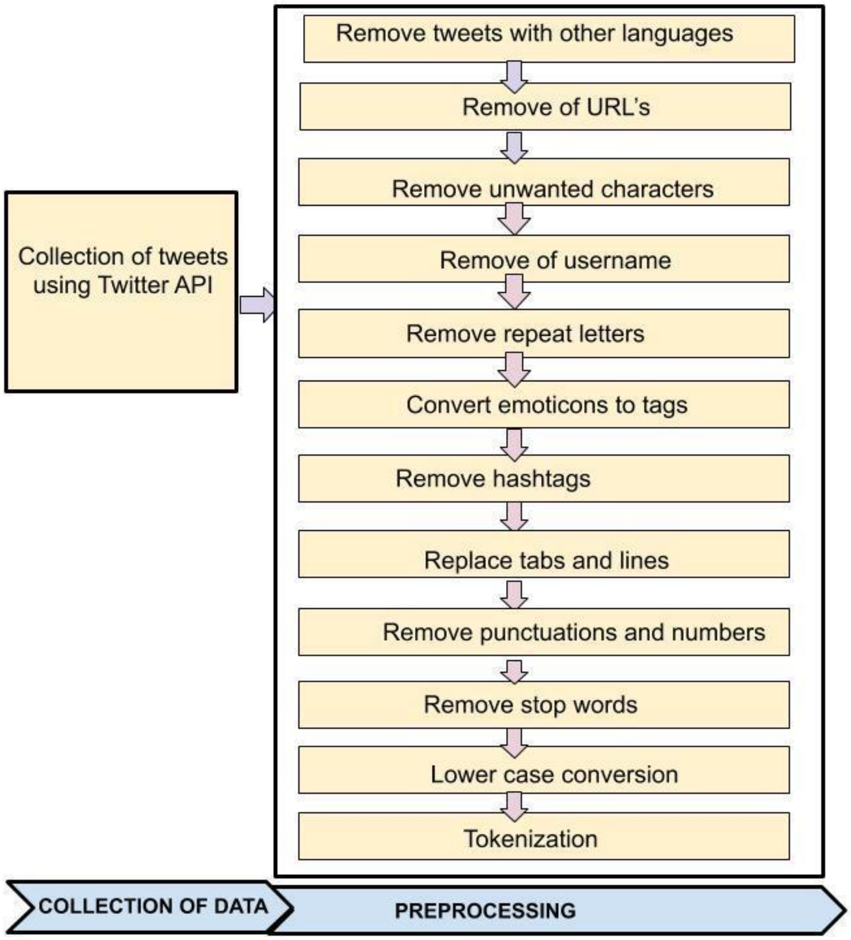


Figure 12 Data preprocessing steps

The above preprocessing steps explained in detail below.

Remove tweets with other languages

Here only consider the tweets in English language only.

Removing Unwanted Characters

As explained above all the tweets need to be cleaned to remove unwanted characters, punctuations, hyperlinks, tags, mentions that do not need to get the hidden sentiment in the tweet. The regular expressions used to remove the urls. Most of the users get the use of hash tags in their tweets and all hashtags are removed in order to remove the noise associated with tweets. Those hashtags are not important for the SA analysis of twitter data as it does not contribute to the opinion that carries with twitter.

For this we have used the re.py library package [58] to match particular strings where need to be replaced.

Mention replace with a whitespace:

```
tweets = re.sub('(@[A-Za-z0-9]+)', ' ', tweets)
```

Hashtags replace with a whitespace:

```
tweets = re.sub('(#[A-Za-z0-9]+)', ' ', tweets)
```

URLs replace with a whitespace:

```
tweets = re.sub('(\w+:\/\/\S+)', ' ', tweets)
```

Remove Repeated Letters

Some people use more than one character to emphasize on some words and those are needed to be removed or reduced into a single or double letters. Someone can argue that those words may emphasize and carry a strong meaning for the tweet. However, for the SA, it requires a homogeneous clean corpus. For example the word “good” is in English and doesn't have words like “gooooooooood”. These words come because users like a product/opinion, then he/she gives the review as “gooooooooooood”. In his point of view, that person likes the product so much. So in the data preparation process of SA, it is needed to remove the repeated letters and make it as good. The word “huuuuuuuungry” makes it as “hungry”. So it is really important to remove the letters from a word which are occurring more than two times.

Convert emoticons to tags

We cannot forget the contribution of emoticons towards SA and cannot just forget them. They express someone's emotions without using the language. Most of the people use emoticons since the provided text limit in twitter for a particular message is very less. Therefore it is really important to get the use of them. In Table 2 it shows some of the important emoticons used in twitter with the sentiment of whether they are as positive, negative or neutral [59].

Table 2 Emoticons’ with their sentiment annotations as positive, negative or neutral

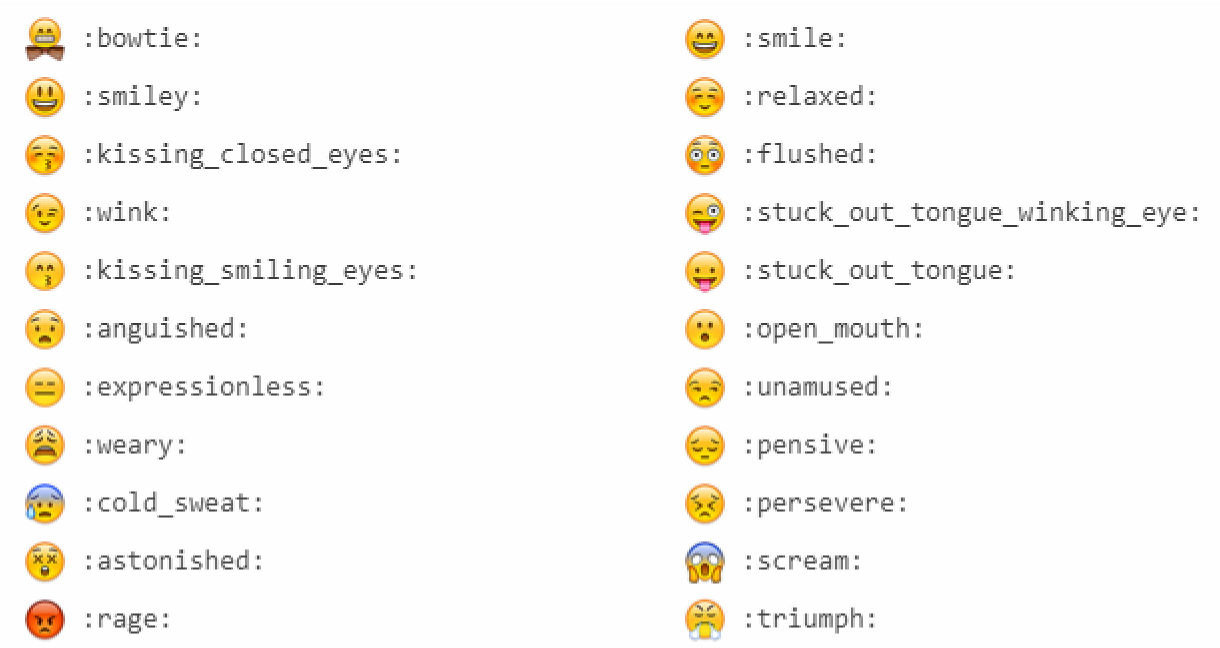
Emoticon	**Sentiment**
:) :-)	Positive
:(>:[:-(	Negative
>:\ >:/ :-/	Neutral

For the above task get the use of emoji package for python and install it by using “pip install emojis” [60]. All the emojis decoded as follows:

emojis.decode('This text contain a emojis 😄')

'This text contain a emojis :smile:

Emoji cheat sheet [61] contains all the character information about emojis as shown in figure 13.



```
🙇 :bowtie:
😃 :smiley:
😚 :kissing_closed_eyes:
😉 :wink:
😙 :kissing_smiling_eyes:
😧 :anguished:
😑 :expressionless:
😩 :weary:
😰 :cold_sweat:
😲 :astonished:
😡 :rage:
😄 :smile:
☺️ :relaxed:
😳 :flushed:
😜 :stuck_out_tongue_winking_eye:
😛 :stuck_out_tongue:
😮 :open_mouth:
😒 :unamused:
😔 :pensive:
😣 :persevere:
😱 :scream:
😤 :triumph:
```

Figure 13 Emoji cheat sheet retrieved from WebFX

(source: https://www.webfx.com/tools/emoji-cheat-sheet/)

The table 3 shows some of the interjections used in the tweets. Interjections can be found listed on http://www.vidarholen.net/contents/interjections/.

Table 3 Interjections' Lexicon

Interjection	**Sentiment**
Wow, waw	Positive
Haha, hihi, hehe	Positive
Oh dear	Negative
No way	Negative

Replace tabs and lines

There are more than one whitespaces and those are removed with the use of reg package discussed earlier.Tabs and newline characters also removed in the twitter corpus.

whitespace and new line characters replace with a whitespace:

tweets = re.sub('\s\s+', ' ', tweets)

Removing Punctuations and Numbers

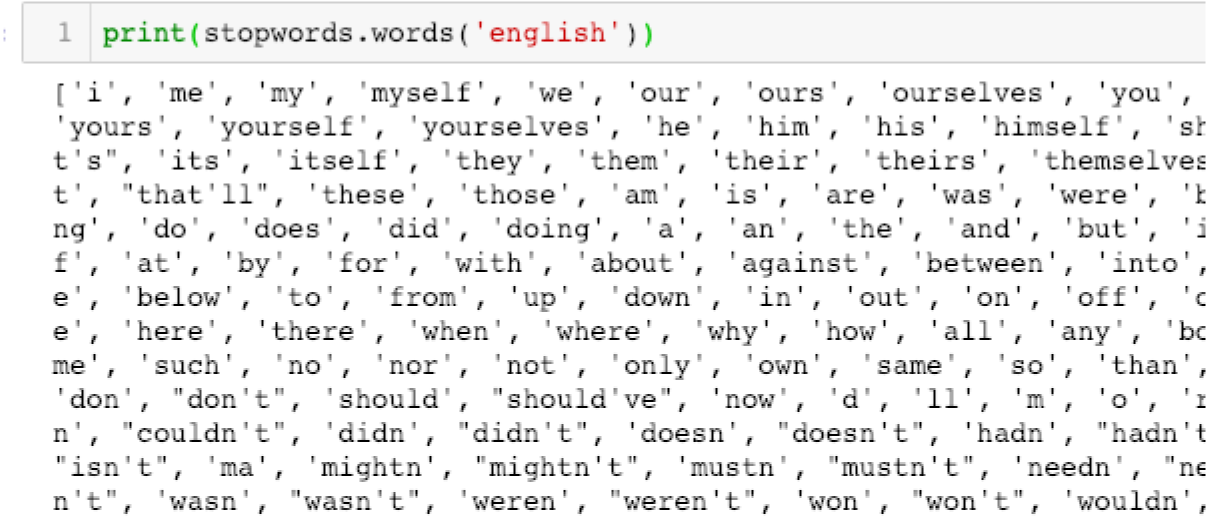
Punctuations such as question marks, quotation marks, dash, semicolon, brackets and numbers are removed again with the use of re python library. This makes more cleaner the data set, increasing the accuracy and the performance. All the numbers removed with the twitter corpus in order to get a clean data set.

Punctuations replace with a whitespace:

tweets = re.sub('[\.\,\!\?\:\;\-\=]', ' ', tweets)

Remove Stop words

This is a main step has followed with the data pre-processing steps. Those words are removed as they are very frequent in the corpus. There is no semantic meaning for them. Stop words consist of pronouns, articles, etc. with words such as “of, are, the, it, is”. Some of the stops words in English language are shown in the figure 14.



```
print(stopwords.words('english'))
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', 
'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'sh
t's", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves
t', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'b
ng', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'i
f', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into',
e', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'c
e', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'bo
me', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than',
'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 'r
n', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't
"isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "ne
n't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn',
```

Figure 14 Some default stop words used in English and removed in our twitter corpus

Lower case conversion

All the letters in the corpus need to normalize as in some cases they have mixed case letters. For instance: “FreeDom” is converted to “freedom”. String matching cannot proceed with mixed letters. Therefore, all the words which contain the same letters carry one meaning. By following the following step, the accuracy of the classifier can be improved with a letter casting approach.

```
tweet.lower();
```

Tokenization

Tokenization is used to get a token from tweets. At this stage a larger paragraph converts to tokens. Token can be defined as individual words separated by spaces in a text document. In our corpora we tokenize all the tweets into tokens which contains a list of words for each and every tweet. When input the sentence “the dog sat on the table” to the tokenization process it outputs “the”, “sat”, “on”, “the”, “table”. NLTK library has a TweetTokenizer module which can be utilized to get tokens from tweets where it breaks all words with punctuation.

```
from nltk.tokenize import word_tokenize

tokenized_tweet=word_tokenize(tweet)

print(tokenized_tweet)
```

Stemming and lemmatization

Stemming is the process of removing morphological affixes that can be associated in a given word in the corpus by leaving only the word stem. For instance, all the words “small”, “smaller”, “smallest” are converted into “small”. There are many stemmers developed in many programming languages and one of the stemmers in Weka is IteratedLovinsStemmer [62]. Stemming and lemmatization has the same capability of generating the root form of a word. The only difference with the stemming and lemmatization is that the stem may not be an actual word always whereas lemma is always an actual word in the language. The porter stemmer is very popular where it removes common morphological words from English language.

```
from nltk.stem import PorterStemmer

from nltk.stem import WordNetLemmatizer

ps = PorterStemmer()

sentence="They have intelligently working to pass the examinations"

ps.stem(sentence)

lemmatizer = WordNetLemmatizer()
```

3.3.3 Prepare train and test datasets

The Corpus divided into two sets as training and testing data sets. The training data set is to train the classifier and the testing data set is to evaluate the classifier in terms of accuracy and the performance. The split of the dataset is done with allocating 70% of data for training and the remaining 30% for testing. This has been done using the sklearn library.

```
import numpy as np

from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.3, random_state = 101)
```

Train_X (Training data predictors), Test_X (Testing predictors), Train_Y (Training data target), Test_Y (Testing data target) is clearly shown in the figure 15.

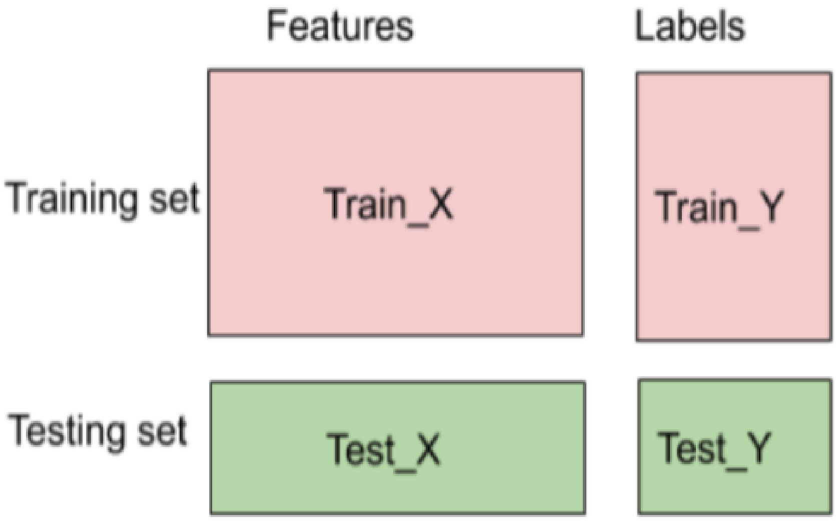


Figure 15 Training and testing data test

Encoding

This is used to convert the target variables which are in the form of string format into numerical values.

```
        Encoder = LabelEncoder()

        Train_Y = Encoder.fit_transform(Train_Y)

        Test_Y = Encoder.fit_transform(Test_Y)
```

3.3.4 Feature extraction process

In a classification problem, the training must be carried along with the selected features. Therefore, the feature extraction plays a vital role in order to train the classifier.

We have used Term Frequency and Inverse Document Frequency (TF-IDF) in this process with tweets. The classification algorithms applied with the output generated with TF-IDF which consists of weights for the word frequencies.

3.3.5 TF-IDF weighted Word Count: Feature Extraction

When generating the word count vector to represent the frequency of the words appearing for all the words used in all tweets, the zero appears in many places as all the words are not appearing in all the tweets.

After doing a vital literature review I was able to find many approaches to quantify a word in tweets. The frequency is one method which is the number of times a word appears divided by total number of words. POS is another method which has been already explained. We have used TF-IDF technique, which is based on a weighted scheme. This gives a high frequency for the words that appear more frequently.

Term Frequency(TF)

The number of occurrences of the term 't' in document 'd'is denoted as TF(t,d). This is increasing with the frequency of the word. This is using a logarithmic value to control the growth of exponential values.

$$TF(t,d) = \log(F(t,d))$$

Inverse Document Frequency(IDF)

$$IDF(t,D) = \log(N/Nt \in d)$$

where: D= Corpus, N = total of files in the corpus D, Nt ∈ d = number of files that appears the term 't'

$$TF(Term) = \frac{term\ frequency\ which\ appears\ in\ the\ document}{Total\ number\ of\ terms\ in\ a\ document}$$

$$ITF(Term) = \frac{total\ number\ of\ documents}{Total\ number\ of\ documents\ where\ that\ particular\ term\ appears}$$

Finally, the TF-IDF can be calculated as:

$$TF\text{-}IDF(t,d,D) = TF(t,d) \,.\, IDF(t,D)$$

Scikit-learn ML library provides this feature.

l**ibrary: sklearn.feature_extraction.text**

TF-IDF can be calculated with the help of sklearn.feature_extraction.text library class used in sklearn library for extraction of text features.

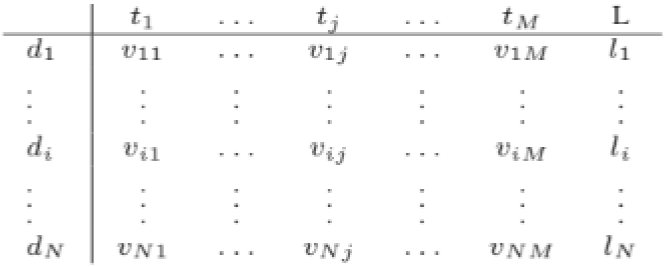
3.3.6 Word Vectorization with TF-IDF

We have considered a maximum of 10000 unique words/features. For each row all the words are appearing. The values will contain only the words that are in a particular tweet only. Below shows how Train_X and Test_X are transformed to TF-IDF vectorized values Train_X_Tfidf and Test_X_Tfidf.

```
Tfidf_vect = TfidfVectorizer(max_features=10000)

Tfidf_vect.fit(Corpus['text_final'])

Train_X_Tfidf = Tfidf_vect.transform(Train_X)

Test_X_Tfidf = Tfidf_vect.transform(Test_X)
```

Following table 4 depicted the dataset matrix where l_i = label/sentiment of a certain tweets d_i , v_{ij} = value of the term t_j , j ∈ {1,...,M} in d_i, i ∈ {1,...,N}. Now the v_{ij} can be calculated using TFIDF.

Table 4 Data set matrix with TF-IDF values for each term in a tweet



	t_1	...	t_j	...	t_M	L
d_1	v_{11}	...	v_{1j}	...	v_{1M}	l_1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
d_i	v_{i1}	...	v_{ij}	...	v_{iM}	l_i
⋮	⋮	⋮	⋮	⋮	⋮	⋮
d_N	v_{N1}	...	v_{Nj}	...	v_{NM}	l_N

3.4 Classification using ML methods

There are two approaches to build a classifier as: supervised and unsupervised approach. We have used a supervised learning algorithms based classifier for the SA. There are three classes as

positive, negative and neutral. Most of the literature shows only two classes as positive and negative. However, we have used the neutral as a class as in our political domain there are people with neutral sentiment as well. Here it uses a rough fuzzy classifier and some common classification algorithms such as SVM and NB which are used already by the previous scholars. The three sentiment classes described above are:

- Positive: Positive attitude towards the GR and SP. Politicians try to increase the tweets with positive sentiment.
- Negative Sentiments: Negative attitude towards the GR and SP. Politicians try to decrease the tweets with positive sentiment. If politicians have more negative tweets, it means that the people are not happy about his works.
- Neutral Sentiments: This does not reflect any sentiment towards the political party. Politicians try to convert those candidates positively towards them if they need to win from the election.

3.4.1 Fuzzy-Rough Classifier

The proposed system is primarily combined of data collection and cleaning, giving sentiment index value, generating the information system and predicting the sentiment with the classifier. Data cleaning will be done based on text SA. The lexicon-based approach, used to originate and analyze positivity, negativity and uncertainty text posted in twitter social media on GR and SP. Further, this extracts the feature from posting and computes the sentiment index value for each.

The fuzzy rough classifier predicts the sentiment of people towards the candidates before commencing the election by analyzing the sentiment of tweets of people towards the GR and SP. This classifier has two major branches as the learning path (the classifier is trained with training data) and the prediction path. At the same time the performance of the classifier is tested with testing data. After the twitter data preprocessing stage (to clean the twitter data which are noisy) the feature extraction will be done. After that, the Fuzzy-Rough Feature Selection (FRFS) will be applied by using Fuzzy-Rough Quickreduct inorder to get a minimum number of features. The output gained is fed to the ML classification algorithm. After finishing the training part, the preprocessed testing data feed to the classifier through the feature extraction module which is simpler than the training phase. The sentiment can be predicted from the testing data with the sentiment analyzer and the output generated is compared with the existing sentiment index value which has been manually annotated in order to predict the performance of the analyser. Now the classifier is ready for the prediction of the sentiment of the twitter data which has been extracted from twitter. The above predicted results are used for decision making. In this case study, the political sentiment towards the main political candidates can be analyzed before the election in order to give them an insight of their results at the election where they can take necessary precautions accordingly such as changing their campaign styles where necessary.

Attribute selection

The attribute selection task plays a vital role due to the vast amount of data collected in social media. Attribute selection is considered as a dimensionality reduction method which is defined as "the process of finding a best subset of features, from the original set of features in a given data set, optimal according to the defined goal and criterion of feature selection (a feature goodness criterion)" [9]. The selected subset of originally available attributes is much more relevant than taking the entire features. A smaller subset of attributes directly affecting the performance of the classifier. It improves the performance since the search space is less. Further, it reduces the risk of overfitting. Moreover, it creates less chances of creating wrong decisions. Attribute set with few is very easy to monitor, efficient, safe and understand. Rough set [3, 10, 11] based approach is considered as an efficient method where it gets the necessary information from the dataset itself and is capable of handling the vagueness in the information system. Fuzzy set theory has certain limitations with uncertainty. Dubois & Prade proposed an approach where it is capable of handling uncertainty and vagueness in the dataset [12]. Attribute reduction using fuzzy rough sets is often called FRFS.

FRFS reduces real valued features of the data set since it builds on the notion of fuzzy lower approximation. This is similar to the crisp approach. If we compared this with the traditional rough set theory, the crisp positive region is represented with the union of the lower approximation. The fuzzy positive region can be defined by:

$$\mu_{POS_{P(Q)}}(x) = \sup_{X \in \mathbb{U}/Q} \mu_{\underline{P}X}(x)$$

FRFS with Fuzzy-Rough Quickreduct

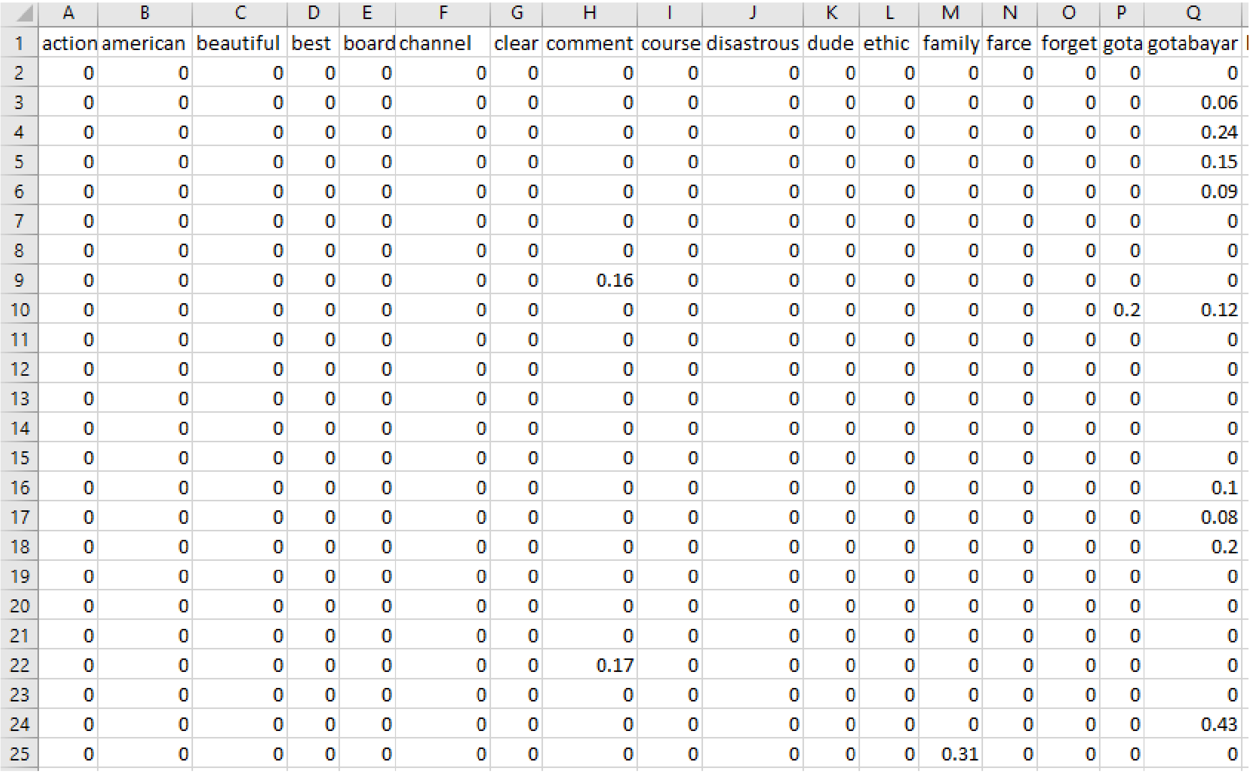
FRFS works fine with both continuous and nominal data which are considered as attributes. Therefore, we can use this for our classification data set. This is also possible with the regression data sets as well.

The accuracy of the classifier is very important as this is basically for the prediction of a future activity. Attribute selection plays a vital role to ensure the accuracy. The collected twitter corpus have higher dimensionality of microarray datasets. This diminishes the performance of the classifier. The quick reduct algorithm is the best solution for that. It begins with a set and fills the attributes by comparing the dependency with each other. It does one at a time. It selects the best attribute to produce the reduct set which compromises with the least dependency values [63].

Sentiment classification with ML algorithms

Now the feature extraction process is completed. Therefore, the dataset is now ready to feed for the classification algorithm.

According to the scenario of the presidential election 2019, the results after following the TFIDF vectorization is saved to a csv file as in figure 16.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	action	american	beautiful	best	board	channel	clear	comment	course	disastrous	dude	ethic	family	farce	forget	gota	gotabayar
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.06
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.24
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.15
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.09
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0.16	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.2	0.12
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.08
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.2
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0.17	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.43
25	0	0	0	0	0	0	0	0	0	0	0	0	0.31	0	0	0	0

Figure 16 The TFIDF values for 17 attributes for 25 tweets from the twitter data set for GR

The last column of the above data set represented in the csv file is the label which is manually labeled as either negative - 0 , positive - 1 or neutral - 2.

Based on the FRST, the Fuzzy-Rough Nearest Neighbours (FRNN) algorithm used as the ML algorithm in the classifier to classify the tweets [64, 66, 67, 68]. The use of fuzzy lower and upper approximations enhance the FNN. a target instance t was allotted to a class by the algorithm as follows:

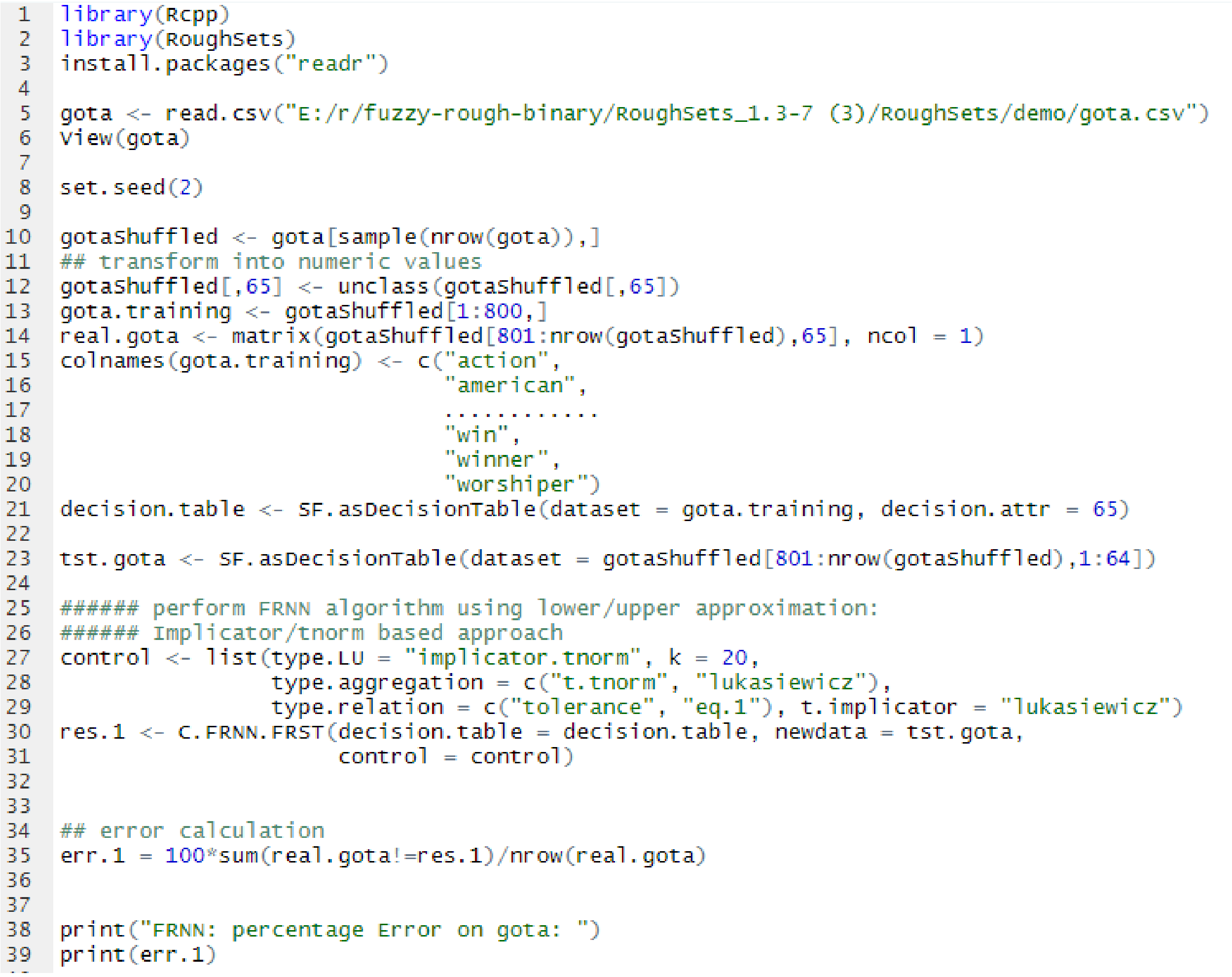
- K nearest neighbours - This is decided by the equivalence of new patterns.
- The maximal value is considered in the fuzzy lower and upper approximations in order to allot new patterns. However if the (R↓C)(y) is high means that the most of all neighbours of y is in C. If the (R↑C)(y) is high means at least one neighbour of y is in c.

There are two approaches considered here. It depends on the fuzzy lower and upper approximations. "implicator.tnorm" is based on the implicator/t-norm and "vqrs" is based on vaguely quantified rough sets. The FRNN algorithm is described as follows.

U, the training data; *C*, the set of decision classes;
y, the object to be classified.

```
(1) N ← getNearestNeighbors(y,K)
(2) μ1(y) ← 0, μ2(y) ← 0, Class ← Ø
(3) ∀C ∈ C
(4)     if ((R↓C)(y) ≥ μ1(y) && (R↑C)(y) ≥ μ2(y))
(5)             Class ← C
(6)             μ1(y) ← (R↓C)(y), μ2(y) ← (R↑C)(y)
(7) output Class
```

When the R(x, y) is growing less the impact of x is very less with (R↓C)(y) and (R↑C)(y). Therefore, with the FRNN-FRS it is not mandatory to use the K value. The RoughSets library of R Package used to develop the model of FRNN-FRS as shown in figure 17.



```
library(Rcpp)
library(RoughSets)
install.packages("readr")

gota <- read.csv("E:/r/fuzzy-rough-binary/RoughSets_1.3-7 (3)/RoughSets/demo/gota.csv")
View(gota)

set.seed(2)

gotaShuffled <- gota[sample(nrow(gota)),]
## transform into numeric values
gotaShuffled[,65] <- unclass(gotaShuffled[,65])
gota.training <- gotaShuffled[1:800,]
real.gota <- matrix(gotaShuffled[801:nrow(gotaShuffled),65], ncol = 1)
colnames(gota.training) <- c("action",
                             "american",
                             ............
                             "win",
                             "winner",
                             "worshiper")
decision.table <- SF.asDecisionTable(dataset = gota.training, decision.attr = 65)

tst.gota <- SF.asDecisionTable(dataset = gotaShuffled[801:nrow(gotaShuffled),1:64])

###### perform FRNN algorithm using lower/upper approximation:
###### Implicator/tnorm based approach
control <- list(type.LU = "implicator.tnorm", k = 20,
                type.aggregation = c("t.tnorm", "lukasiewicz"),
                type.relation = c("tolerance", "eq.1"), t.implicator = "lukasiewicz")
res.1 <- C.FRNN.FRST(decision.table = decision.table, newdata = tst.gota,
                     control = control)


## error calculation
err.1 = 100*sum(real.gota!=res.1)/nrow(real.gota)


print("FRNN: percentage Error on gota: ")
print(err.1)
```

Figure 17 the development of the FRNN-FRS algorithm using the R package

3.4.2 Naive Bayes

The NB classifier is based on the Bayes Theorem.

$P(\frac{A}{B}) = \frac{p(\frac{B}{A}).P(A)}{P(B)}$ where; A and B represents events while $P(B) \neq 0$

There are many types of NB models available such as Multi-variate Bernoulli NB , Multinomial NB and Binarized Multinomial NB models are popular among researchers [68].

X is a dataset

$$X = \{\overline{x}_1 , \overline{x}_2 , \ldots , \overline{x}_n \} \; where \; \overline{x}_i \in \mathbb{R}^m$$

Feature vectors as follows:

$$\overline{x}_i = \; [x_1 , x_2 , \ldots , x_m]$$

Target dataset is as follows:

$$Y = \{y_1 , y_2 , \ldots , y_n\} \; where \; y_n \in (0,1,2, \ldots , P)$$

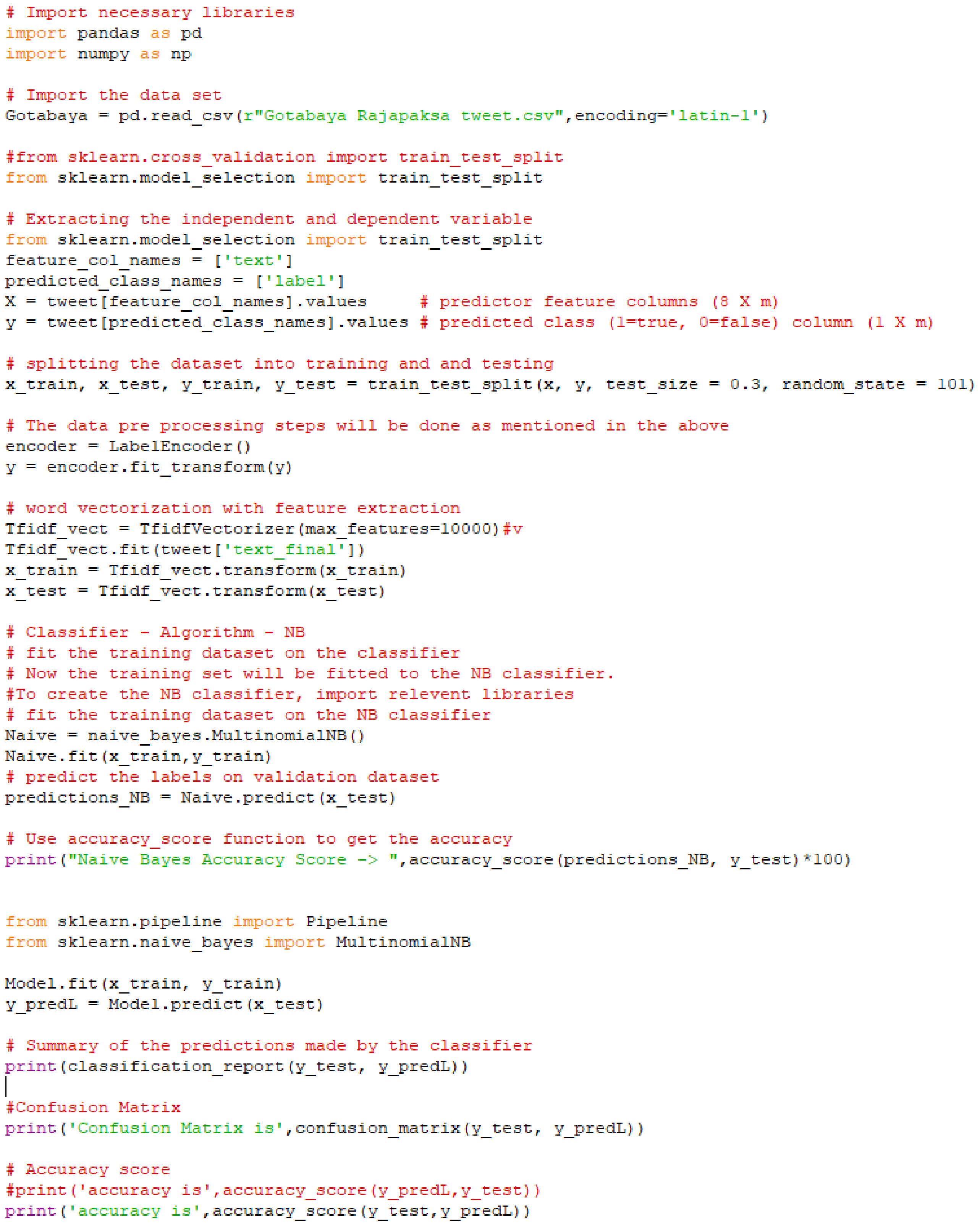
Therefore:

$$P(y|x_1 , x_2 , \ldots , x_m) = \; \alpha P(y) \prod_i P(x_i|y)$$

In case of a feature vectors consists of n with k different value:

$$P(X_1 = x_1 \; \cap X_2 = \; x_2 \cap \; \ldots \; \cap \; X_k) = \; \frac{n!}{\prod_i x_i!} \prod_i {P_i}^{x_i}$$

After doing the data preprocessing steps below is the implementation of Multinomial Naive Bayes classification in figure 18:



```
# Import necessary libraries
import pandas as pd
import numpy as np

# Import the data set
Gotabaya = pd.read_csv(r"Gotabaya Rajapaksa tweet.csv",encoding='latin-1')

#from sklearn.cross_validation import train_test_split
from sklearn.model_selection import train_test_split

# Extracting the independent and dependent variable
from sklearn.model_selection import train_test_split
feature_col_names = ['text']
predicted_class_names = ['label']
X = tweet[feature_col_names].values     # predictor feature columns (8 X m)
y = tweet[predicted_class_names].values # predicted class (1=true, 0=false) column (1 X m)

# splitting the dataset into training and and testing
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.3, random_state = 101)

# The data pre processing steps will be done as mentioned in the above
encoder = LabelEncoder()
y = encoder.fit_transform(y)

# word vectorization with feature extraction
Tfidf_vect = TfidfVectorizer(max_features=10000)#v
Tfidf_vect.fit(tweet['text_final'])
x_train = Tfidf_vect.transform(x_train)
x_test = Tfidf_vect.transform(x_test)

# Classifier - Algorithm - NB
# fit the training dataset on the classifier
# Now the training set will be fitted to the NB classifier.
#To create the NB classifier, import relevent libraries
# fit the training dataset on the NB classifier
Naive = naive_bayes.MultinomialNB()
Naive.fit(x_train,y_train)
# predict the labels on validation dataset
predictions_NB = Naive.predict(x_test)

# Use accuracy_score function to get the accuracy
print("Naive Bayes Accuracy Score -> ",accuracy_score(predictions_NB, y_test)*100)


from sklearn.pipeline import Pipeline
from sklearn.naive_bayes import MultinomialNB

Model.fit(x_train, y_train)
y_predL = Model.predict(x_test)

# Summary of the predictions made by the classifier
print(classification_report(y_test, y_predL))

#Confusion Matrix
print('Confusion Matrix is',confusion_matrix(y_test, y_predL))

# Accuracy score
#print('accuracy is',accuracy_score(y_predL,y_test))
print('accuracy is',accuracy_score(y_test,y_predL))
```

Figure 18 Multinomial Naive Bayes classification

3.4.3 SVM

The classification and regression challenges are common with SA and SVM provides a better solution for that. SVM is a supervised ML technique. This classifier is capable of handling a large set of features.

It is very important to get the best possible decision boundary which is known as a hyperplane. The below figure 198 shows a decision boundary which separates into two different categories.

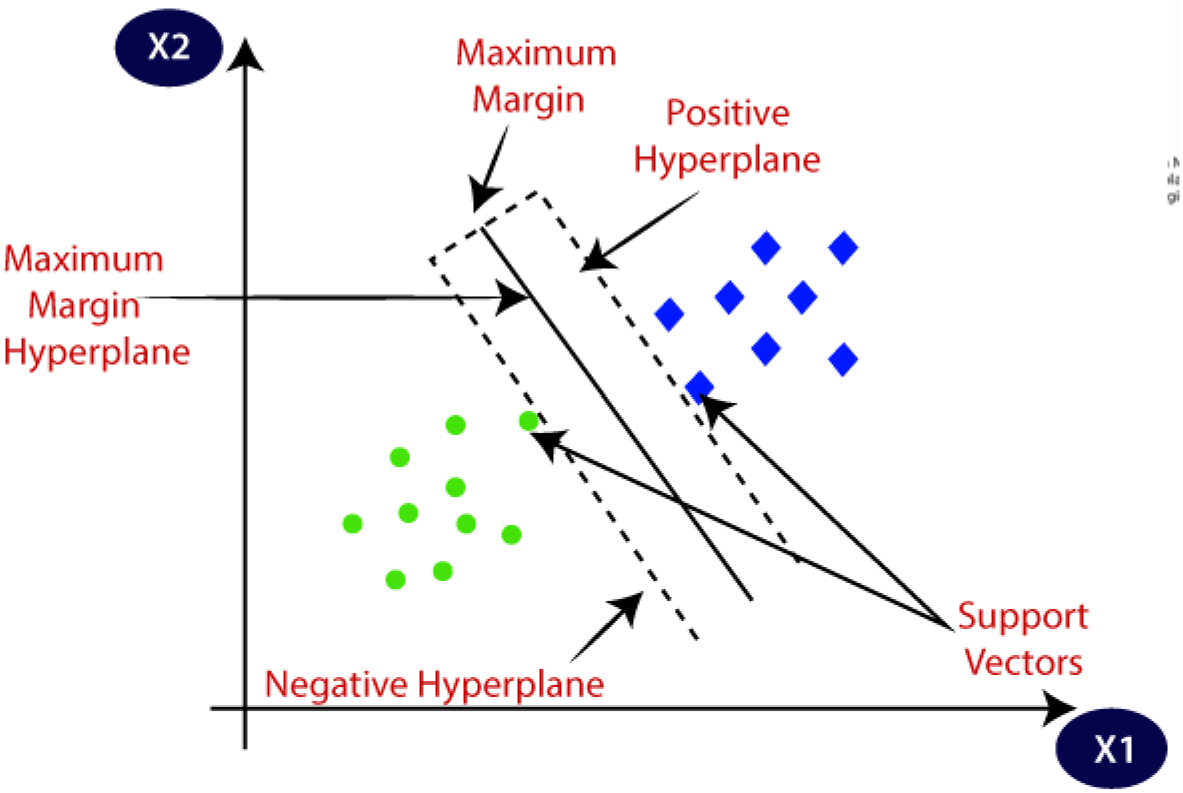
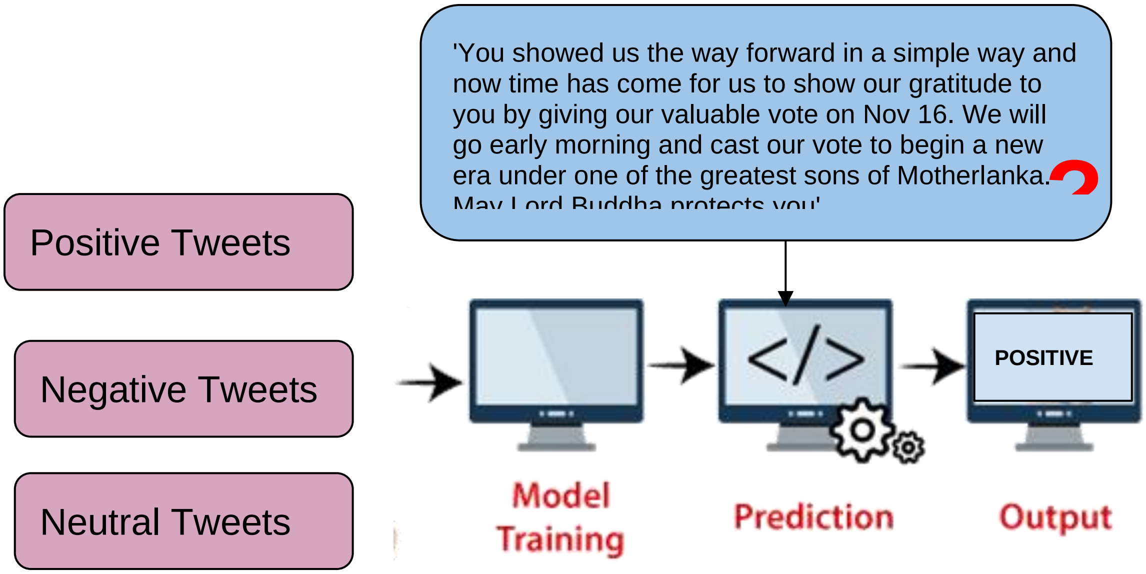


Figure 19 Two different categories that are classified using a decision boundary or hyperplane

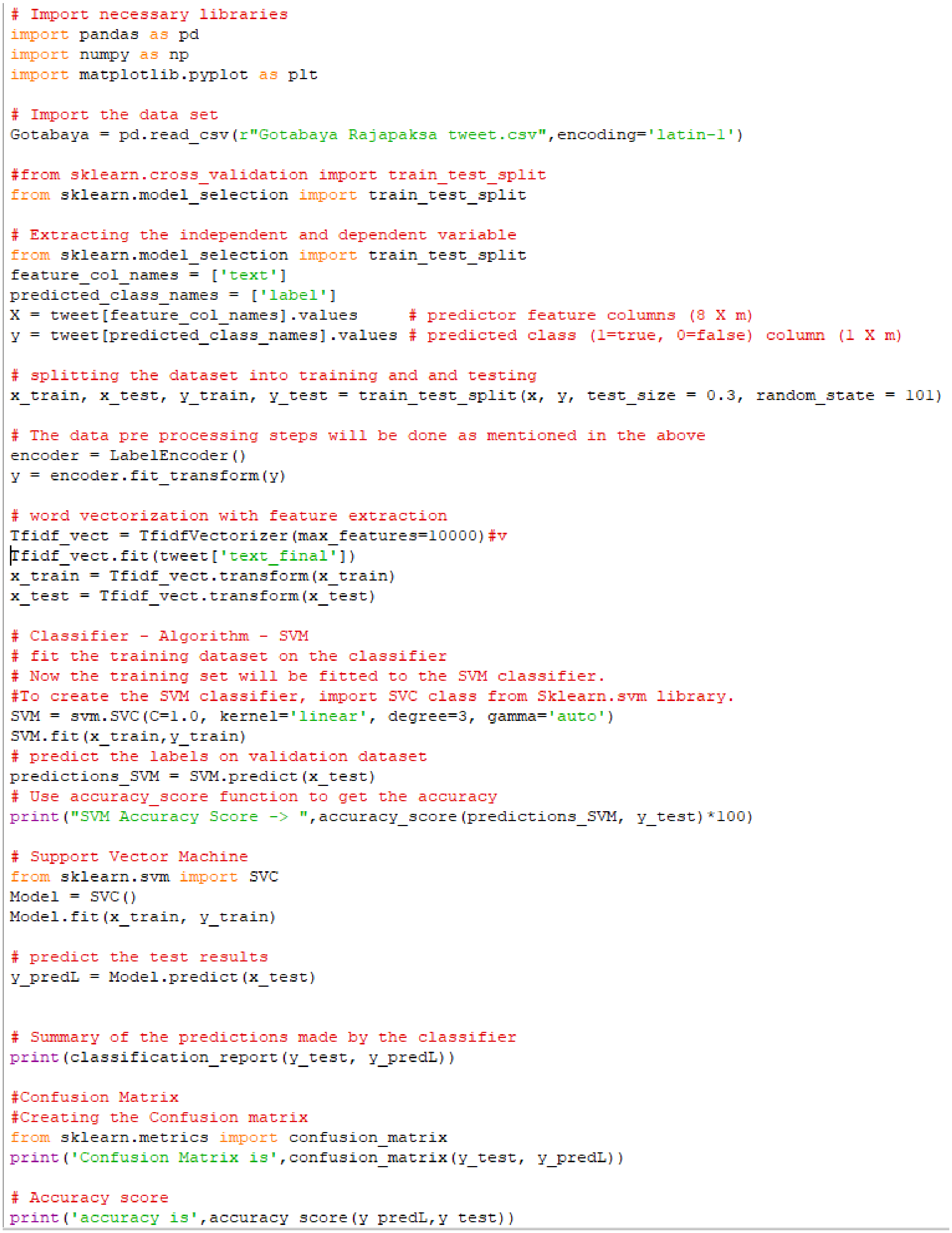
In the scenario addressed in this research, at the second phase twitters should be classified as either positive, negative or neutral. SVM is used for this with labeled tweets (0- Negative, 1 - positive and 2 - Neutral) to train the classifier and test with unlabeled tweets collected in the political domain. It classifies tweets as either positive, negative or neutral as shown in the below diagram depicted in the figure 20.



'You showed us the way forward in a simple way and now time has come for us to show our gratitude to you by giving our valuable vote on Nov 16. We will go early morning and cast our vote to begin a new era under one of the greatest sons of Motherlanka. May Lord Buddha protects you'

Figure 20 Model generation and prediction

The SVM is also available in the scikit-learn library. After doing the data preprocessing steps below is the implementation of SVM classification as in figure 21.



```
# Import necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# Import the data set
Gotabaya = pd.read_csv(r"Gotabaya Rajapaksa tweet.csv",encoding='latin-1')

#from sklearn.cross_validation import train_test_split
from sklearn.model_selection import train_test_split

# Extracting the independent and dependent variable
from sklearn.model_selection import train_test_split
feature_col_names = ['text']
predicted_class_names = ['label']
X = tweet[feature_col_names].values     # predictor feature columns (8 X m)
y = tweet[predicted_class_names].values # predicted class (1=true, 0=false) column (1 X m)

# splitting the dataset into training and and testing
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.3, random_state = 101)

# The data pre processing steps will be done as mentioned in the above
encoder = LabelEncoder()
y = encoder.fit_transform(y)

# word vectorization with feature extraction
Tfidf_vect = TfidfVectorizer(max_features=10000)#v
Tfidf_vect.fit(tweet['text_final'])
x_train = Tfidf_vect.transform(x_train)
x_test = Tfidf_vect.transform(x_test)

# Classifier - Algorithm - SVM
# fit the training dataset on the classifier
# Now the training set will be fitted to the SVM classifier.
#To create the SVM classifier, import SVC class from Sklearn.svm library.
SVM = svm.SVC(C=1.0, kernel='linear', degree=3, gamma='auto')
SVM.fit(x_train,y_train)
# predict the labels on validation dataset
predictions_SVM = SVM.predict(x_test)
# Use accuracy_score function to get the accuracy
print("SVM Accuracy Score -> ",accuracy_score(predictions_SVM, y_test)*100)

# Support Vector Machine
from sklearn.svm import SVC
Model = SVC()
Model.fit(x_train, y_train)

# predict the test results
y_predL = Model.predict(x_test)


# Summary of the predictions made by the classifier
print(classification_report(y_test, y_predL))

#Confusion Matrix
#Creating the Confusion matrix
from sklearn.metrics import confusion_matrix
print('Confusion Matrix is',confusion_matrix(y_test, y_predL))

# Accuracy score
print('accuracy is',accuracy score(y predL,y test))
```

Figure 21 SVM classification

Those algorithms separately train for GR and SP. The developed classifier model use to predict the sentiment of collect tweets and analyze the prediction of the presidential election 2019.

3.5 Visualization

Word Cloud is a data visualization technique. The size of the text is compatible with the frequency of its term. This can be used to identify the significant words among a series of words. In python, we used matplotlib, pandas and wordcloud with following commands to create the word cloud for each politician.

```
pip install matplotlib

pip install pandas

pip install wordcloud
```

The drawback for WordCloud is that the graphics only reflect the frequency of words, which can cause some uninformative words frequently appearing in the text can be highlighted on the cloud instead of informative words which are less frequently appeared in the text. These kinds of uninformative words could be stopwords or just some words frequently appeared in documents that are particularly longer than other documents. Although WordCloud is not the best visualization method to show all the aspects of the data, it is worth plotting them so that we can quickly and intuitively see what the text is about.

Chapter 4

Results and Evaluation

4.1 Introduction

This section discusses the experimental results and the evaluation. All the results obtained with the three different algorithms were evaluated to ensure the performance and the accuracy. This project is based on a case study and used ML based algorithms to develop a SA classifier. Further, it shows that the developed classifier provides accurate results and automatic sentiment classification of tweets posts collected before the election by predicting the election results of the presidential election 2019, Sri Lanka accurately. The project is to develop a model for pre-processing of harvested un-structured noisy data from social networks and analyze ML algorithms used in other classification models and evaluate their suitability for the problem of classifying data for SA. The performance and accuracy of the fuzzy-rough classification model will be evaluated with the existing models.

The performance of a new implementation needed to be evaluated. Most of the literature evident that many people use IRIS, MINST like a dataset to evaluate their developed classifiers. In this study the classifier accuracy is tested with the collected twitter data in order to measure the sentiment of a person towards GR and SP. Hence it requires an evaluation plan to measure the accuracy and the performance of the classifier.

This section focuses on comparing and evaluating the performance of the fuzzy-rough classifier with NB and SVM. For the above process we use the training and testing data set. According to the results, we select the best classifier for the task of twitter sentiment classification for the political data analysis in the presidential election 2019, Sri Lanka towards the GR and SP.

4.2 Results

For the implementation purpose, get the use of R language and python for implementation. Further, we use Jupiter Notebook, an open-source web application and RStudio which is an integrated development environment (IDE) for R. Those languages and tools offer maximum support when it comes to data mining with ML techniques. For the purposes of NLP, Python offers the NLTK and I used it as well.

During the election period, people used words like “premadasa, sajith, gamudawa, hambantota, election, village, people, likes, good, srilanka, citizenship” for SP as shown in the figure 22 wordcloud.

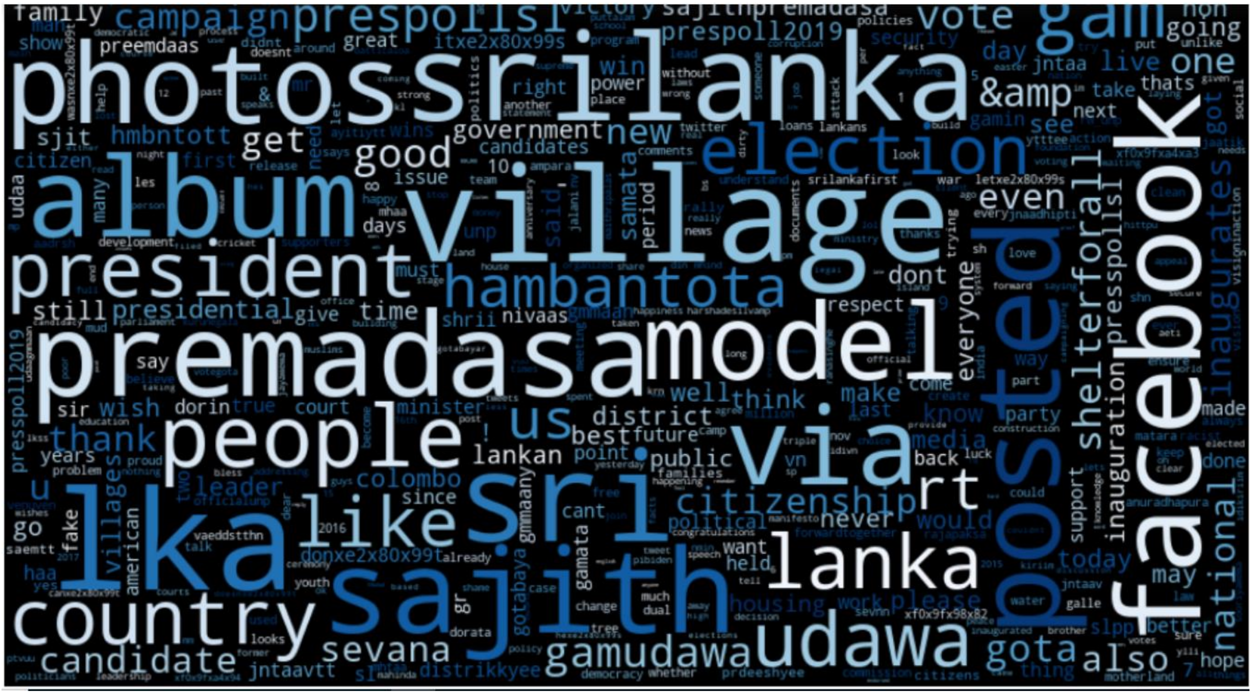


Figure 22 Wordcloud for SP

During the election period, people used words like “srilanka, president, like, election, country, vote, citizenship, gota” for GRas shown in the figure 23.

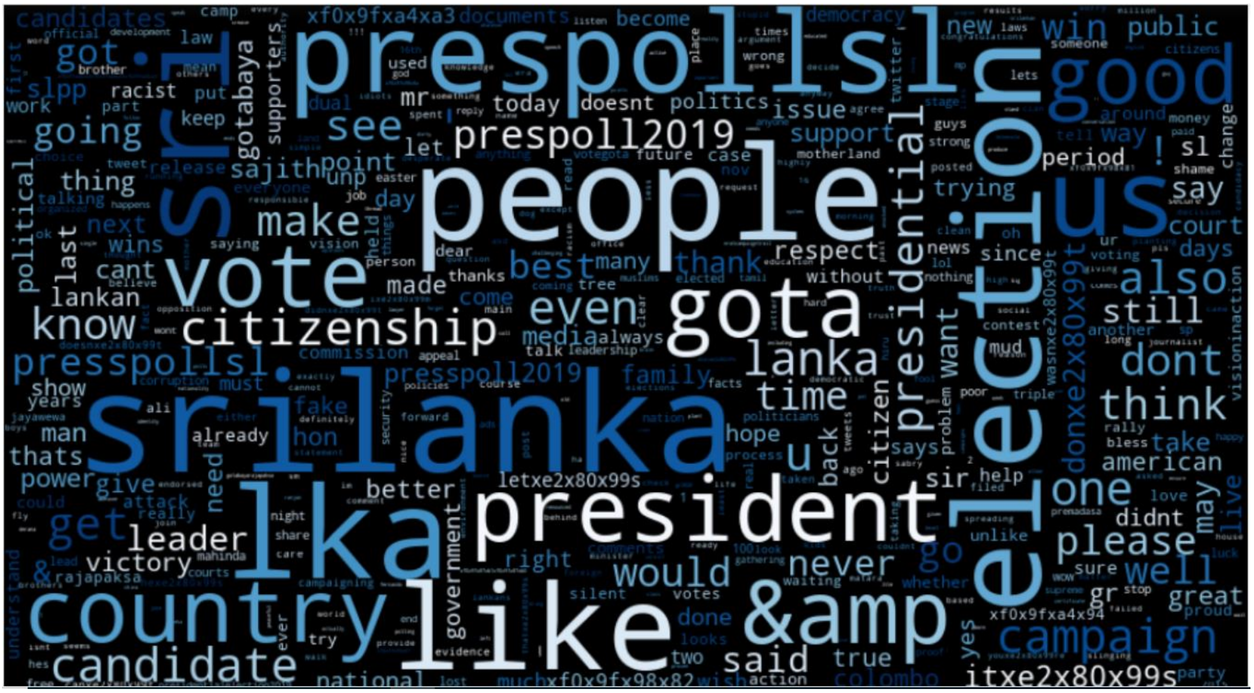


Figure 23 Wordcloud for GR

The training data set which was manually labeled as shown in the figure 24 and it consists of 1000 for GR and 1000 for SP.

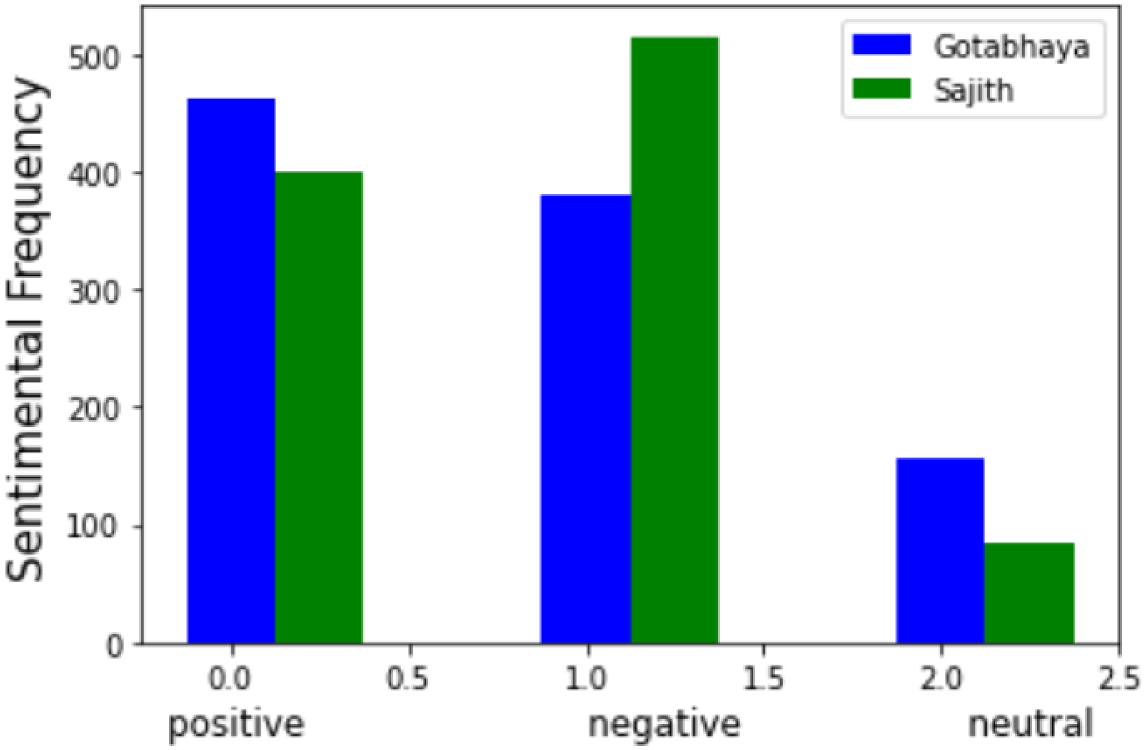


Figure 24 The sentiment of the training and testing data set

In the second phase the trained classifier classifies the tweeter data and final sentiment analysis for both GR and SP for 2000 tweets for each distributed among the three classes positive, negative and neutral as in the figure 25 and 26. The positive sentiment towards the GR is higher than the SP but it is not comparatively high. The difference between the negative comments towards the SP than GR is higher than the difference between the positive comments towards the GR than SP. However the negative sentiments towards the SP are higher than the GR. Many people have a negative sentiment for GR than the SP. According to the case study that shows the sentiment towards the two candidates GR and SP with tweeter data before the presidential election 2019 shows that people have positive sentiment for GR than SP and a negative sentiment for SP than GR. The actual results of the presidential election of 2019 Sri Lanka also tally with the predicted results of the classifier.

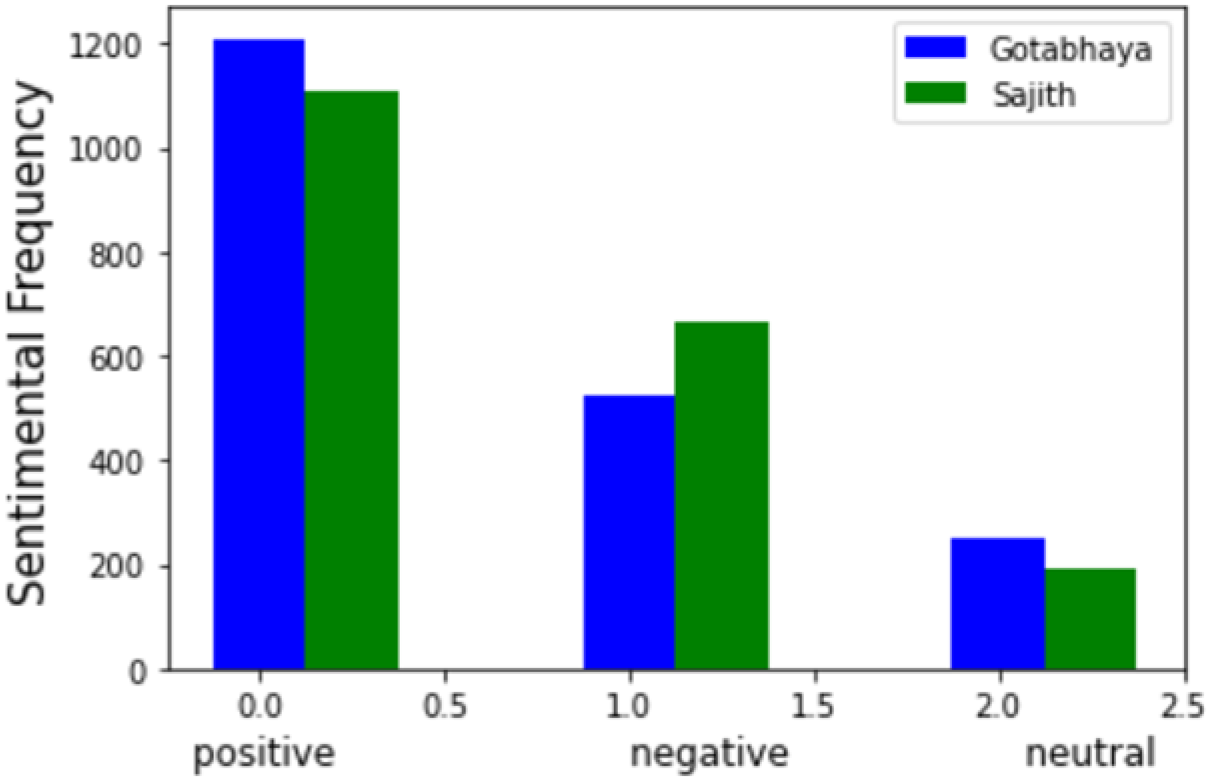


Figure 25 The overall sentiment for GR and SP

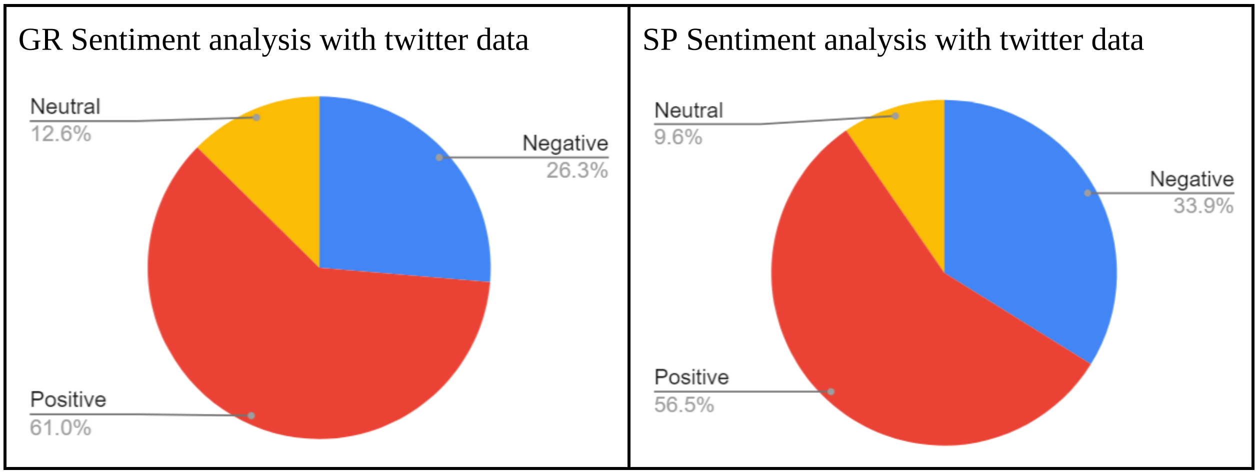


Figure 26 The overall sentiment for GR and SP as percentages

4.3 Experimental evaluation

All the classifiers were evaluated using cross-validation to omit the common mistake of overfitting. The flowchart of cross validation workflow in model training that has been adapted for this project has depicted in figure 27.

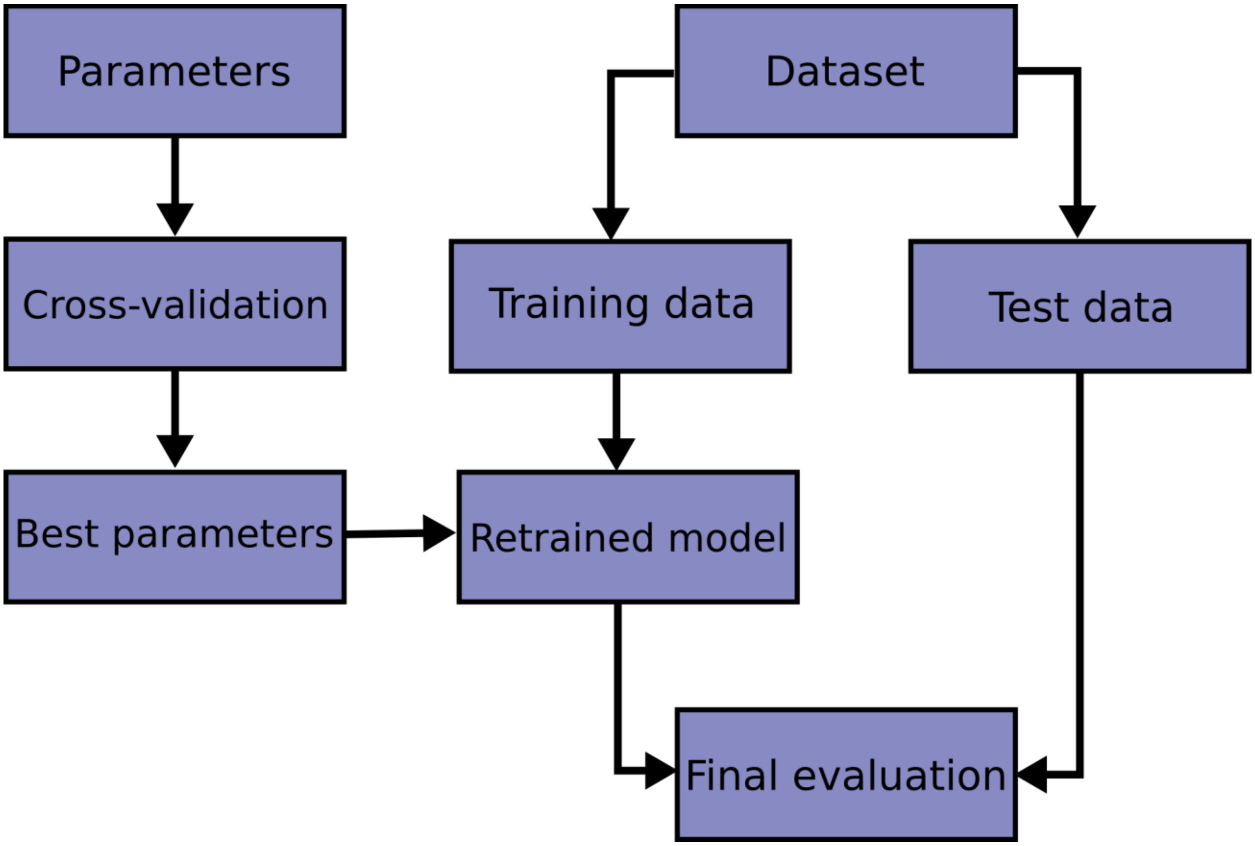


Figure 27 The flowchart of cross validation workflow for classifiers
(Source: https://scikit-learn.org/stable/modules/cross_validation.html?highlight=sklearn%20cross_validation).

Data Set

The collected twitter data set during 01st of September 2019 to 15th of November 2019 during the presidential election 2019 of Sri Lanka have been split into training and testing partitions. The splitting ratio is 70% to 30%. All the collected tweets manually annotated into sentiment labels as:

Negative - 0

Positive - 1

Neutral - 2

This uses an experimental base evaluation by aiming to estimate the contribution of the proposed rough-fuzzy classifier model. All the classifier were evaluated based on four criterias: accuracy, precision, recall, F-measure and Confusion Matrix.

Compute the accuracy

The accuracy is the percentage of tweets that are correctly classified under the each class.

Compute the precision

The precision is the,

$$Precision \; = \frac{Number\ of\ true\ positive}{(Number\ of\ true\ positive\ + Number\ of\ false\ positive)}$$

This measures the closeness where it is close to the true value.

Compute the recall

The recall is the,

$$Recall \; = \frac{Number\ of\ true\ positive}{(Number\ of\ true\ positive\ + Number\ of\ false\ negatives)}$$

This is a measurement where it indicates the relevance data in the dataset classified by the ML algorithm.

Compute the F1 score (F-measure)

F1 score reaches its best value at 1 and worst score at 0. The formula for the F1 score is:

F1 = 2 * (precision * recall) / (precision + recall)

The results for the above measures are listed in the table 3.

Table 5 Precision, Recall and F1 Measure for the classifiers to their three classes: Positive, Neutral and Negative.

	Fuzzy-Rough classifier					**NB**					**SVM**				
	Precision	**recall**	**F1-Measure**	**support**	**Accuracy**	**Precision**	**recall**	**F1-Measure**	**support**	**Accuracy**	**Precision**	**recall**	**F1-Measure**	**support**	**Accuracy**
Positive	*0.45*	*0.33*	*0.38*	*113*	*0.53*	*0.39*	*0.21*	*0.28*	*113*	*0.51*	*0.41*	*0.35*	*0.38*	*113*	*0.50*
Negative	*0.56*	*0.76*	*0.65*	*160*		*0.55*	*0.79*	*0.65*	*160*		*0.55*	*0.70*	*0.62*	*160*	
Neutral	*0.00*	*0*	*0*	*27*		*0.29*	*0.07*	*0.12*	*27*		*0.00*	*0.00*	*0.00*	*27*	

It shows that the accuracy is less for all the classifiers with the high variations with the opinion mining with twitter data. However the developed fuzzy rough classifier has a higher accuracy than the other two classifiers.

CHAPTER 5

Conclusion and Future work

5.1 Discussion

In conclusion, we have used a fuzzy-rough sets based SA classifier for analysing political Twitter data where all the classifiers based on ML technique. The twitter data during the Sri Lankan presidential election period in 2019 were collected as training and testing to perform the SA. Based on the classifiers on each candidate the results of their election can be predicted. Hereafter they are used to construct corpora for Positive, negative and neutral tweets. The bases for the comparison of the Fuzzy Rough Classifier against SVM and Multinomial NB. We used some of the evaluation measures such as: F1-measure, Recall, Precision and accuracy. TF-IDF is used as the weighting scheme. The classification is also done with Multinomial NB and SVM to evaluate the accuracy and the performance of the suggested classifier. The classifiers are able to group the entire corpus into three classifiers as negative, positive and neutral. The experiment results of accuracy, results show that the Fuzzy Rough Classifier yielded better results and is a popular technique for this application which outperforms others.

In the case study it shows that there are more positive opinions for GR than SP when analyzing the tweet data before the election. Those analyses can be used to improve their campaigns before the election if they can get an idea about the general public before the last moment with the use of social media data where people talk publicly and present their opinions. SA applications are broad and powerful where those help for marketers to promote their social media campaigns when there is an election. On the other hand they can evaluate their own campaign.

There are some weaknesses and limitations with the domain of the sentiment classification with political tweets associated with this case study as well. The detection of spam and fake reviews is always difficult and the duplicates of the same tweet presents. Since this is human centric and opinion based, a word can cater both positive and negative in two different places. People express their opinions in different ways. There might be a small difference between the text, however the meaning can be drastically different from the opinion. Identifying the entity is one of the challenges in opinion mining. A text may have multiple entities associated with it. For instance, “Coward is a more suitable word than foreign candidate", here there is a positive opinion for GR but negative for SP. In the same tweet it represents both negative and positive polarity. Sudden changes can be seen in the same tweet. Further, the domain is different and it has hidden meaning for some of the words where context matters. Demographics bias is associated with the tweets and however it was neglected with this analysis. Further, those tweets were produced by those politically active which represents only a part of the people who have the voting power in the election. Furthermore, the classifier should develop in terms of the accuracy of SA of political tweets must improve with more training data. Humor and sarcasm plays a major role and should take precautions for that.

The electoral predictions based on social media are analogous to traditional polls where people still use polls to predict the results. Therefore, the current state of art for the prediction of political sentiment with microblogging which is probable with the social media data as witnessed with this case study.

5.2 Future work

For further work, the accuracy of the algorithms can be increased by implementing a new algorithm utilizing the benefits of the two or three algorithms so that it can be used effectively in prediction and forecasting. This can be further developed with the use of new features according to the desirable level of people. The model also can be developed with the use of different stemmers. The data purity can be increased and guaranteed by taking only tweets by users eligible to vote. Bias in the data can be at least acknowledged and analyzed even attempting to remove demographic bias can be encouraged. The model tested with unigram and it can be expanded to compare the result between unigram, bigram and trigram.

References

[1] Pang, Bo and Lee, Lillian and Vaithyanathan, Shivakumar, Thumbs up? Sentiment Classification using Machine Learning Techniques, Proceedings of EMNLP 2002

[2] Peter Turney, “Thumbs up or thumbs down?,” in Semantic orientation applied to unsupervised classification of reviews, pages 417424, 2002

[3] Z. Pawlak, Rough sets, International Journal of Computer &amp; Information Sciences 11(5) (1982), 341–356.

[4] D. Dubois, H. Prade, “Rough Fuzzy Sets and Fuzzy Rough Set”, International Journal of General Systems, vol.17, 1990, pp. 191-209.

[5] D. Dubois and H. Prade. Putting rough sets and fuzzy sets together. In [171], pp. 203–232. 1992.

[6] Jensen, R., &amp; Shen, Q. (2007). Fuzzy-Rough Sets Assisted Attribute Selection. IEEE Transactions on Fuzzy Systems, 15(1), 73–89. doi:10.1109/tfuzz.2006.889761

[7] R.Jensen, Q.Shen, Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches, IEEE computational intelligence society

[8] W. Kasemsiri, and C. Kimpan, “Printed thai character recognition using fuzzy-rough sets”. *In TENCON2001. Proceedings of IEEE Region 10 International Conference on Electrical and Electronic Technology* (Vol. 1, pp.326-330). IEEE.

[9] Y. Liu and M. Schumann, &quot;Data mining feature selection for credit scoring models,&quot; Journal of the Operational Research Society, vol. 56, pp. 1099 - 1108, 2005.

[10] Pawlak, Z. (2012). Rough sets: theoretical aspects of reasoning about data: Vol. 9.

Dordrecht, Netherlands: Springer Science &amp; Business Media.

[11]Pawlak, Z., &amp; Skowron, A. (2007). Rough sets: Some extensions. Information Sciences,

177(1), 28–40.

[12] Dubois, D., &amp; Prade, H. (1990). Rough fuzzy sets and fuzzy rough sets. International

Journal of General System, 17(2-3), 191–209.

[13] Bollen, J., Mao, H., &amp; Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1--8. doi: 10.1016/j.jocs.2010.12.007

[14] Augustyniak, L., Kajdanowicz, T., Kazienko, P., Kulisiewicz, M., &amp; Tuliglowicz, W. (2014). An Approach to Sentiment Analysis of Movie Reviews: Lexicon Based vs. Classification. Hybrid Artificial Intelligence Systems, 168–178. doi:10.1007/978-3-319-07617-

1_15

[15]Shivaprasad, T. K., &amp; Shetty, J. (2017). Sentiment analysis of product reviews: A review. 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT). doi:10.1109/icicct.2017.7975207

[16]Choy, M., Cheong, L. F. M., Ma, N. L., &amp; Koo, P. S. (2011). A sentiment analysis of Singapore Presidential Election 2011 using Twitter data with census correction

[17] Payne, J. G. (2010). The Bradley effect: Mediated reality of race and politics in the 2008 US

presi-dentialelection.American Behavioral Scientist,54(4), 417–435.

[18] Ceron, A., Curini, L., Iacus, S. M., &amp;Porro, G. (2014). Every tweet counts? How sentiment analysisof social media can improve our knowledge of citizens'political preferences with an applicationto Italy and France.New Media &amp; Society,16(2), 340–358.

[19] K. Srinivas, G. R. Rao, A. Govardhan (2014), Rough-Fuzzy Classifier: A System to Predict the Heart Disease by Blending Two Different Set Theories. Arabian Journal for Science and Engineering, 39(4), 2857–2868.doi:10.1007/s13369-013-0934-1

[20] U.Keerthika,R.Sethukkarasi,A.Kannan. (2012), A ROUGH SET BASED FUZZY INFERENCE SYSTEM FOR MINING TEMPORAL MEDICAL DATABASES, International Journal on Soft Computing (IJSC) Vol.3, No.3, August 2012 DOI: 10.5121/ijsc.2012.3304

[21] K.Srividya, A.MarySowjanya,(2017), Sentiment analysis of facebook data using naïve bayesclassifierInternational Journal of Computer Science and Information Security (IJCSIS), Vol. 15, No. 1, January 2017

[22]P. Salunkhe, S. Deshmukh, (2008), Twitter Based Election Prediction and Analysis, International Research Journal of Engineering and Technology (IRJET) ,Volume: 04 Issue: 10 | Oct -2017

[23] Lewis Beck, M. S. (2005). Election forecasting: principles and practice. The British Journal of Politics &amp; International Relations, 7(2), 145-164.

[24] Fumagalli, L. &amp;. (2011). The total survey error paradigm and pre-election polls: The case of the 2006 Italian general elections. ISER Working Paper Series. 2011-29.

[25] F. M.F.Wong, C.Wei.Tan, S.Sen and M.Chiang, (2013) Quantifying Political Leaning from Tweets, Retweets, and Retweeters, Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media.

[26] M. Vadivukarassi, N. Puviarasan and P. Aruna, Sentimental Analysis of Tweets Using Naive Bayes Algorithm,World Applied Sciences Journal,India. 35 (1): 54-59, 2017, DOI:10.5829/idosi.wasj.2017.54.59

[27] A. Pak, P. Paroubek, Twitter as a Corpus for Sentiment Analysis and Opinion Mining, Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta, European Language Resources Association 2010, ISBN 2-9517408-6-7

[28] H.Schmid, Probabilistic part-of-speech Tagging Using Decision Tree, International Conference on New Methods in Language Processing,, 1994, Manchester, UK

[29] C.E.Shannon, W.Weaver, (1963). A Mathematical Theory of Communication. The

University of Illinois Press, Champaign, IL, USA.

[30] B. P. AniruddhaPrabhu, B. P. Ashwini, Tarique Anwar Khan and A. Das, Predicting Election Result with Sentimental Analysis Using Twitter Data for Candidate Selection, Springer Nature Singapore Pte Ltd. 2019 H. S. Saini et al. (eds.), Innovations in Computer Science and Engineering, Lecture Notes in Networks and Systems 74, https://doi.org/10.1007/978-981-13-7082-3_7

[31]C.S.Rao, G.S.Prasad, V.V.Rao, Prediction and Analysis of Sentiments on Twitter Data using Machine Learning Approach, International Journal of Computer Science and Information Security (IJCSIS), Vol. 16, No. 8, August 2018

[32] Wei, W. and Gulla, J. (2019). Sentiment Learning on Product Reviews via Sentiment Ontology Tree. In: 48th Annual Meeting of the Association for Computational Linguistics. [online] Sweden: Association for Computational Linguistics, pp.404–413. Available at: https://www.researchgate.net/publication/220874363 [Accessed 11 Oct. 2019].

[33] A. Neviarouskaya, H. Prendinger and M. Ishizuka, &quot;SentiFul: A Lexicon for Sentiment Analysis,&quot; in IEEE Transactions on Affective Computing, vol. 2, no. 1, pp. 22-36, Jan.-June 2011. doi: 10.1109/T-AFFC.2011.1

[34] Pang, Bo and Lee, Lillian, A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts, Proceedings of the ACL, 2004

[35] Wang, Hao et al., “A system for real time twitter sentiment analysis of 2012 US presidential election cycle” in Proceedings of the ACL 2012 System Demonstrations. Association for Computational Linguistics, 2012.

[36] Mostafa, Mohamed M. “More than words: Social networks text mining for consumer brand sentiments”, in Expert Systems with Applications, pages 4241-4251, 2013.

[37] J. Kamps, M. Marx, R. J. Mokken, and M. De Rijke, “Using wordnet to measure semantic orientations of adjectives,” 2004.

[38] R. Xia, C. Zong, and S. Li, “Ensemble of feature sets and classification algorithms for sentiment classification,” Information Sciences: an International Journal, vol. 181, no. 6, pp. 1138–1152, 2011.

[39] L. Barbosa and J. Feng. Robust Sentiment Detection on Twitter from Biased and Noisy Data. In Proceedings of the international conference on Computational Linguistics (COLING), 2010.

[40] N. Zainuddin and A. Selamat, “Sentiment Analysis Using Support Vector Machine” , International Conference on Computer, Communication, and Control Technology (I4CT 2014), September 2 - 4, 2014 - Langkawi, Kedah, Malaysia

[41] M.J. Bhumika and B.V. Vimalkumar, “Sentiment Analysis using Support Vector Machine based on Feature Selection and Semantic Analysis”, International Journal of Computer Applications (0975 – 8887), vol. 146, no.13, July 2016,

[42] M. Ahmad and S. Aftab, “Analyzing the Performance of SVM for Polarity Detection with Different Datasets”, Int. J. Mod. Educ. Comput. Sci., vol. 9, no. 10, pp. 29–36, 2017.

[43] A. Tarlekar and M.K. Kodmelwar, “Sentiment Analysis of Twitter Data from Political Domain Using Machine Learning Techniques”, International Journal of Innovative Research in Computer and Communication Engineering, vol. 3, no. 6, 2015, doi: 10.15680/ijircce.2015.0306084

[44] T. Elghazaly, A. Mahmoud and H.A. Hefny, “Political Sentiment Analysis Using Twitter Data”, ICC &#39;16: Proceedings of the International Conference on Internet of things and Cloud Computing, 2016, doi: 10.1145/2896387.2896396

[45] M. Ringsquandl and D. Petkovic, “Analyzing Political Sentiment on Twitter”, 2013 AAAI Spring Symposium, 2013

[46] P.Kassraie, A. Modirshanechi and H.K. Aghajan, “Election Vote Share Prediction using a Sentiment-based Fusion of Twitter Data with Google Trends and Online Polls”, Proceedings of the 6th International Conference on Data Science, Technology and Applications, pp: 363-370, 2017

[47] A.S. Raghuwanshi and S.K. Pawar, “Polarity Classification of Twitter Data using Sentiment Analysis”, International Journal on Recent and Innovation Trends in computing and Communication, vol. 5, no. 6, pp: 434 – 439, 2015

[48] L. Dey, S. Chakraborty, A. Biswas, B. Bose and S. Tiwari, “Sentiment Analysis of Review Datasets Using Naïve, Bayes‘ and K-NN Classifier”, International Journal of Information Engineering and Electronic Business, 2016, doi: 10.5815/ijieeb.2016.04.07

[49] T. Chen, P. Su, C. Shang, R. Hill, H. Zhang and Q. Shen, “Sentiment Classification of Drug Reviews Using Fuzzy-rough Feature Selection”, 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2019. Available: 10.1109/fuzz-ieee.2019.8858916 [Accessed 11 May 2020].

[50] H. Zhao, P. Wang, Q. Hu and P. Zhu, “Fuzzy Rough Set Based Feature Selection for Large-Scale Hierarchical Classification”, IEEE Transactions on Fuzzy Systems, vol. 27, no. 10, pp. 1891-1903, 2019. Available: 10.1109/tfuzz.2019.2892349 [Accessed 11 May 2020].

[51] L. Meenachi and S. Ramakrishnan, “Evolutionary sequential genetic search technique-based cancer classification using fuzzy rough nearest neighbour classifier”;, Healthcare Technology Letters, vol. 5, no. 4, pp. 130-135, 2018. Available: 10.1049/htl.2018.5041 [Accessed 11 May 2020]

[52] R. Jensen, C. Cornelis, (2008), “A New Approach to Fuzzy-Rough Nearest Neighbour Classification”, In: Chan CC., Grzymala-Busse J.W., Ziarko W.P. (eds) Rough Sets and Current Trends in Computing. RSCTC 2008. Lecture Notes in Computer Science, vol 5306. Springer, Berlin, Heidelberg

[53] S.J.Soman , P. Swaminathan, R. Anandan, and K. Kalaivani, “A comparative review of the challenges encountered in sentiment analysis of Indian regional language tweets vs English

language tweets”, *International Journal of Engineering & Technology*, vol. 7 (2.21) (2018), pp. 319-322

[54] V.A. Kharde, S.S. Sonawane, “Sentiment Analysis of Twitter Data: A Survey of Techniques”, *International Journal of Computer Applications (0975 – 8887),* Vol. 139 – no.11, April 2016

[55] K. Meghashree, S. Radhika, A. Shilpashree,S.Dinni, P.Monika, “Survey Paper on Algorithms used for Sentiment Analysis”, International Journal for Research in Applied Science &amp; Engineering Technology (IJRASET), vol. 8, no. V, May 2020

[56] M.D. Devika, C. Sunitha and A. Ganesha,”Sentiment Analysis:A Comparative Study On Different Approaches ” Fourth International Conference on Recent Trends in Computer Science &amp; Engineering, Procedia Computer Science 87 (2016), pp. 44 – 49, 2016, doi: 10.1016/j.procs.2016.05.124

[57] F.Zhang, Z.Zhang and M.Lan, “ECNU: A Combination Method and Multiple Features for Aspect Extraction and Sentiment Polarity Classification”, *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014),* Dublin, Ireland, pp. 252–258, 23–24 August (2014).

[58] Twitter Sentiment Analysis using fastText, https://towardsdatascience.com/twitter-sentiment-analysis-using-fasttext-9ccd04465597, retrieved on: 12/05/2020

[59] W. Wolny, “SENTIMENT ANALYSIS OF TWITTER DATA USING EMOTICONS AND EMOJI IDEOGRAMS”, University of Economics, Faculty of Informatics and Communication Department of Informatics, ISSN 2083-8611 Nr 296, 2016

[60] Emoji for Python, https://pypi.org/project/emoji/, retrieved on: 12/05/2020

[61] Emoji cheat sheet, https://www.webfx.com/tools/emoji-cheat-sheet, retrieved on: 12/05/2020

[62] Fornacciari, P., Mordonini, M., Tomaiuolo, M.: A case-study for sentiment analysis on twitter. In: Proceedings of the 16th Workshop ”From Objects to Agents”-WOA (2015)

[63] Hannah Inbarani H., Azar A.T., Jothi G.: ‘Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis’, Comput. Methods Programs Biomed., 2014, 113, (1), pp. 175–185 (doi: 10.1016/j.cmpb.2013.10.007) [PubMed] [Google Scholar]

[64] Jensen, R., & Cornelis, C. (2010). Fuzzy-rough instance selection. International Conference on Fuzzy Systems. doi:10.1109/fuzzy.2010.5584791

[65] A.M. Radzikowska and E.E. Kerre, “A comparative study of fuzzy rough sets,” Fuzzy Sets and Systems, vol. 126, no. 2, pp. 137–155, 2002.

[66] R. Jensen and C. Cornelis, "Fuzzy-rough Nearest Neighbour Classification and Prediction", Theoretical Computer Science, vol. 412, p. 5871 - 5884 (2011).

[67] N.Verbiest, C.Cornelis, and R.ensen, “Fuzzy Rough Positive Region based Nearest Neighbour Classification”, WCCI 2012 IEEE World Congress on Computational Intelligence June, 10-15, 2012, Brisbane, Australia

[68] C. Aggarwal and C. X. Zhai, “A survey of text classification algorithms,” pp. 163–222, 2012. [Online]. Available: http://www.time.mk/trajkovski/thesis/ text-class.pdf