Exploring Challenges In Sentiment Analysis

By

Saidur Rahman
ID: CSE1602008001

Md. Sweet Reza
ID: CSE1603009043

Md. Abdur Razzaq
ID: CSE1701010004

Ferdousai Akter
ID: CSE1701010031

Submitted in partial fulfillment of the requirements for the degree of
Bachelor of Science in Computer Science and Engineering

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SONARGAON UNIVERSITY (SU)

October 2020

APPROVAL

The thesis titled "**Exploring Challenges in Sentiment Analysis**" submitted by SAIDUR RAHMAN (CSE1602008001), MD. SWEET REZA (CSE1603009043), MD ABDUR RAZZAQ (CSE1701010004) and FERDOUSAI AKTER (CSE1701010031) to the Department of Computer Science and Engineering, Sonargaon University (SU), has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering and approved as to its style and contents.

Board of Examiners

Supervisor

Arifur Rahaman
Lecturer & Co-ordinator,
Department of Computer Science and Engineering
Sonargaon University (SU)

Examiner 1

(Examiner Name & Signature)
Department of Computer Science and Engineering
Sonargaon University (SU)

Examiner 2

(Examiner Name & Signature)
Department of Computer Science and Engineering
Sonargaon University (SU)

Examiner 3

(Examiner Name & Signature)
Department of Computer Science and Engineering
Sonargaon University (SU)

DECLARATION

We, hereby, declare that the work presented in this thesis paper is the outcome of the investigation performed by us under the supervision of **Arifur Rahaman, Lecturer & Co-ordinator,** Department of Computer Science and Engineering, Sonargaon University, Dhaka, Bangladesh. We reaffirm that no part of this Thesis and thereof has been or is being submitted elsewhere for the award of any degree or diploma.

Countersigned	Signature
------------------------------	------------------------
(Arifur Rahaman)	**Saidur Rahman**
Supervisor	ID: CSE1602008001

	Md. Sweet Reza
	ID: CSE1603009043

	Md. Abdur Razzaq
	ID: CSE1701010004

	Ferdousai Akter
	ID: CSE1701010031

ABSTRACT

In this moment in the people of whole world are always facing different situation and explaining different ways for different moment. It explains with emotions. Maximum organization play their business with customer's emotions. They need to analysis the customer's mind. It is call Sentiment Analysis. SA is the process of determining whether a piece of writing is positive, negative or neutral. The benefits of SA is that it helps data analysts within large enterprises gauge public opinion, conduct nuanced market research, monitor brand and product reputation, and understand customer experiences. They generated sentiment content can be about books, people, hotels, products, research, events, company, rooms, foods, business etc. These sentiments become very beneficial for business, governments and individuals. While this content meant to be useful, a bulk of this written content require using the text mining techniques and sentiments analysis. Sentiment analysis is the practice of applying Natural Language Processing and Text Analysis techniques to identify and extract subjective information from text. For this reason every organization need to SA. It has many challenges. We have explored important challenges in this paper to analysis. This paper represents the Sentiments Analysis challenges relevant to their approaches and techniques with exploring challenges.

ACKNOWLEDGMENT

At the very beginning, we would like to express my deepest gratitude to the Almighty Allah for giving us the ability and the strength to finish the task successfully within the schedule time.

We are auspicious that we had the kind association as well as supervision of **Arifur Rahaman**, Lecturer & Co-ordinator, Department of Computer Science and Engineering, Sonargaon University whose hearted and valuable support with best concern and direction acted as necessary recourse to carry out our project.

We would like to convey our special gratitude to **Bulbul Ahmad,** Associate Professor & Head of Department, Faculty of Science and Engineering for his kind concern and precious suggestions.

We are also thankful to all our teachers during our whole education, for exposing us to the beauty of learning.

Finally, our deepest gratitude and love to my parents for their support, encouragement, and endless love.

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
LSTM	Long Short Term Memory
ML	Machine Learning
NLP	Natural Language Processing
OP	Opinion Mining
RNN	Recurrent Neural Networks
SA	Sentiment Analysis
SVM	Support Vector Machine

TABLE OF CONTENTS

Title				Page No.
DECLARATION …………………………………………………………..……..				iii
ABSTRACT …………………...……………………………………………				iv
ACKNOWLEDGEMENT …………………………………….…………………				v
LIST OF ABBREVIATION ……….…………………………………………				vi
CHAPTER 1				1-2
INTRODUCTION TO SENTIMEN ANALYSIS				
	1.1	Introduction ……………………………..……………….….....		1
	1.2	Objectives …………………………...………………………...		1
	1.3	Why need to Sentiment Analysis………………………. ……..………		2
CHAPTER 2				3 – 5
TYPES OF SENTIMENT ANALYSIS				
	2.1	Classification..…………………………………………….…......		3
		2.1.1	Fine grained Sentiment Analysis	3
		2.1.2	Emotions Detection	3
		2.1.3	Aspect based Sentiment Analysis	3
		2.1.4	Multilingual Sentiment Analysis	3
	2.2	Depth of Analysis ……………………………….……………….		4-5
		2.2.1	Document Level	4
		2.2.2	Sentence Level	4
		2.2.3	Phrase and Aspect Level	4
		2.2.4	User Level	5
CHAPTER 3				6 –10
TECHNIQUES OF SENTIMENT ANALYSIS				
	3.1	Methodology………………..………………………………......		6
		3.1.1	Data gathering	6
		3.1.2	Preprocessing	6
		3.1.3	Features	6
		3.1.4	Choosing and applying sentiment analysis methods	6

		3.1.5	Evaluating the Results	6
	3.2	Machine Learning……………………….……………….………………		7-8
		3.2.1	Supervised	7
		3.2.2	Un-supervised	8
		3.2.3	Semi-supervised	8
	3.3	Laxicon-based………………………………………………………		8
	3.4	Hybrid techniques…………………………………………………		9
	3.5	Classification algorithms for sentiment anlaysis………………………		9-10
CHAPTER 4				11 – 16
CHALLENGES IN SENTIMENT ANALYSIS				
	4.1	What is Challenges in Sentiment Analysis?…..………………………		11
	4.2	Challenges in Sentiment Analysis…………………………………..		11-16
CHAPTER 5				17-18
PROBABLE SOLUTIONS IN CHALLENGES				
	5.1	Probable Solutions Techniques………………………..………………		17
	5.2	Problems Definition………………………………………………...		18
CHAPTER 6				19
CONCLUSION AND FUTURE WORKS				
	6.1	Conclusion …………………………………………………..…..		19
	6.2	Future Works …………………………………………………..……		19
REFERENCES ……………………………………………………….………..				20

CHAPTER 1

INTRODUCTION TO SENTIMENT ANALYSIS

1.1 Introduction

Sentiment analysis is the process that which is analyzing people's opinions and emotions, generally using language clues. At first glance, classifications problems are been by it, but if we analysis deeply, we will find out that there are a lot of challenging problems which seriously affect sentiment analysis for business related companies. SA is a machine learning technique that detects polarity (e.g. a *yes* or *no* opinion) within text, whether a whole documents, paragraphs, sentences, or speech. Understanding people's emotions and feelings is essential for businesses since customers are able to express their thoughts and feelings more openly than ever before. By automatically analyzing customer feedback, from survey responses to social media conversations, brands are able to listen attentively to their customers, and tailor products and services to meet their needs. For example, using sentiment analysis to automatically analyze 50000+ reviews about their product could help them discover if customers are happy about their pricing plans and customer service. It will benefited for the organization or company. Otherwise they cannot successes in the business. When any product come to market, the production of this product depends on the customer's feedback. If customers give feedback in positive, production will be high. If customers give feedback in negative, production will be low or stop. It depend on the customer's like. For example, a new product from apple. The company give a post in their website, likely "This mobile phone is long lasting with strong battery backup". In this post, if the maximum customers are commented with positive thinking, which means the product is marketable. If the maximum customers are commented in negative thinking, that means is not marketable. So, any organization need to sentiment analysis for their productivity. They have their sentiment scores, what do they do with them? The simplest implementation is to measure the sentiment across each of their responses and take the average as a gauge for their overall sentiment. Track their average sentiment over time to get a feel for how their customers feel towards their business. From here, they can look at segmenting the data and comparing different segments. For example, if their business operates in different locations, or whether they have some demographic information they can use for segmenting customers. Further, they can use a text analytics solution, such as Thematic, to further split up the sentiment across different themes found in their data.

1.2 Objectives

Sentiment analysis is an analytical study which is help to us choose a better decision for goods production. First objective to find easily define challenges for analysis. Anyone can collecting information to analysis from this paper. Second objective to know the techniques and solution. Here we explained the challenges and solution. Anyone can study about these challenges after reading this paper with clear content. We can provide their organization with readmit analytical challenges and solution's techniques. For their organization. Which is effective for any business related organization, who want to success in marketing. There are many fields to SA. Examples, Artificial Intelligence, Data analytics, Data Science, Information Systems, Machine learning, Predictive modelling etc. Most important thing in the whole world is AI. Nowadays we think or pass at moment without AI. Every place and every sectors, government or non-government organizations are used AI directly or indirectly. For any organization need to data analysis which is related with SA. So, SA is the main and important rule in this ultra-modern age. Third objective to find out the problems without solution for analysis. We propose to employ and extract an objective text of images rather than the classic subjective Text provided by users which is extensively exploited in the state of the art to infer the SA associated to social images. SA has a huge of applications and benefits to any business and organization. It can be used to give them business valuable insights into how people feel about their product brand or service. In media channels, it can be used to identify spikes in sentiment, thereby allowing them to identify potential product advocates or social media influences. It can be used to identify when potential negative threads are emerging online regarding their business, thereby allowing them to proactive in dealing with it more quickly.

1.3 Why need to Sentiment Analysis?

SA is extremely useful in social media monitoring as it allows us to gain an overview of the wider public opinion behind certain topics. It will be able to quickly see the sentiment behind everything from forum posts to news articles means being better able to strategies and plan for the future. SA helps any organization or company manage customer's complains and avoid leaving them feeling ignored. When company has a good customer service and improved product quality which can be achieved through SA the sales revenue will automatically scale up. SA also helps organizations measure the ROI of their marketing campaigns and improve their customer service. Since sentiment analysis gives the organizations a sneak peek into their customer's emotions, they can be aware of any crisis that's to come well in time and manage it accordingly. Social media sentiment analysis can be an excellent source of information and can provide insights that can: Determine marketing strategy, Improve campaign success, improve product messaging, Improve customer service, Test business KPIs, and Generate leads etc. Nowadays according to some researchers, Sentiment Analysis of social communication website data can help in the prediction of stock market movements. Researches show that news articles and social media can hugely influence the stock market. Comments in posts positive sentiment has been observed to related to a large increase in price albeit for a short period of time. On the other hand, negative comments in the post is seen to be linked to a decrease in price but with more prolonged effects. Ideally, sentiment analysis can be put to use by any brand looking to: Target specific individuals to improve their services, Track customer sentiment and emotions over time, Determine which customer segment feels more strongly about your brand, Track the changes in user behavior corresponding to the changes in your product, Find out your key promoters and detractors etc. Clearly, SA gives an organization the much-needed insights on their customers. Organizations can now adjust their marketing strategies depending on how the customers are responding to it. SA also helps organizations measure the ROI of their marketing campaigns and improve their customer service. Since sentiment analysis gives the organizations a sneak peek into their customer's emotions, they can be aware of any crisis that's to come well in time and manage it accordingly.

CHAPTER 2

TYPES OF SENTIMENT ANALYSIS

2.1 Classification

SA models focus on polarity (positive, negative, neutral) but also on emotions (love, hate, angry) and feelings (happy, sad, unhappy, painful, etc.), and even on intentions (e.g. *interested* v. *not interested*). Here are some of the most popular types of sentiment analysis:

2.1.1 Fine grained Sentiment Analysis

If polarity precision is important to any companies' business, they might consider expanding their polarity categories to include: 1) Positive 2) Very Positive 3) Negative 4) Very Negative 5) Neutral. The company want get feedback with very positive. But it depends on customers.

2.1.2 Emotion detection

It aims at detecting emotions, like happiness, frustration, anger, sadness, and so on. Many emotion detection systems use lexicons (i.e. the emotions they convey and lists of words) or deeper machine learning algorithms. One of the downsides of using lexicons is that people express emotions in different ways. Some words that typically express anger, like *bad* or *kill* (e.g. *their product is so bad* or *their customer support is killing them*) might also express happiness (e.g. *this is bad ass* or *they are killing it*). Example, the company post in their website to know their product's feedback. The customers commented in any anger word. That means its feedback is not well. The customer's emotion detecting is most important for any organization.

2.1.3 Aspect based Sentiment Analysis

Usually, when analyzing sentiments of texts, let's say product reviews, they'll want to know which particular aspects or features people are mentioning in a positive, neutral, or negative way. That's where aspect-based sentiment analysis can help, for example in this text: *"The camera resolution of this mobile is too low"*, an aspect-based classifier would be able to determine that the sentence expresses a negative opinion about the feature mobile camera.

2.1.4 Multilingual sentiment analysis

It can be difficult. It involves a lot of preprocessing and resources. Most of these resources are available online (e.g. sentiment lexicons), while others need to be created (e.g. translated corpora or noise detection algorithms), but you'll need to know how to code to use them. Alternatively, you could detect language in texts automatically with language classifier, then train a custom sentiment analysis model to classify texts in the language of their choice.

2.2 Depth of Analysis

Sentiment analysis can be conducted on different levels of granularity: document, sentence, word or phrase, aspect or user levels. The aspects of these levels are explained in the following paragraphs.

2.1.1 Document Level

Document level sentiment analysis targets the whole document and assigns an over- all sentiment to it, assuming that the document expresses a single sentiment. This assumption is criticized as unrealistic because a text could hold more than one opinion and hence, the analysis should target finer levels. However, this assumption holds for some branches like reviews, where a final statement about the product is required which is a weighted conclusion arising from different aspects even if the review carries different opinions. Another case where this assumption is valid is in financial news where the news that carries a positive or negative sentiment reflects in a buy or sell signal. In this light, two types of texts, movie reviews and financial news are considered in this thesis.

2.1.2 Sentence Level

Sentence level is a finer level of analysis. It inspects sentences which express a single opinion and try to define its orientation. The assumption of each sentence carries a single sentiment. This assumption does not hold for all sentences in a text. In fact, many sentences do not carry a specific sentiment. It is important to distinguish subjective from non-subjective sentences, because non-subjective sentences add no information to the classifier. Subjective statements are those that express an opinion. Subjectivity detection is the separation of sentences that contain opinions from those that contain facts. Even in opinionated texts, tagging subjectivity prevents misleading and irrelevant sentences from affecting the sentiment classifier. Hence, subjectivity classification is used to improve the performance of sentence level sentiment classifiers. Sentence level is preferred over document level when different opinions in one document are needed to be captured. Texts consist of various types of sentences, each of those has several particular characteristics that make it possible to treat it differently and allow different types of special classification for them, for example, conditional or comparative sentences. It is claimed that there is not one classification strategy that fits all the types of sentences or even a whole text. Therefore, combining different strategies based on different types of sentences improves the classification accuracy. Generally, in an opinionated text or a review the opinion holder is likely to express an overall positive or negative opinion. However, the review author might mention good and bad features about the object.

2.1.3 Phrase and Aspect Level

Phrase or word level analysis investigates the polarity of texts on a finer level: the phrase level. This involves distinguishing polar phrases and then defining their sentiment. In finance, word level sentiment is used to detect polar words and measure their relationships to other variables like firm earnings or stock prices. Recently, many statistical models have been built for a deeper analysis of product reviews, that is, mining the customer's opinions about certain product features. This is commonly referred to as aspect level sentiment analysis. It is the process of extracting relevant aspects of the reviewed product and determining the sentiment of the corresponding opinion about them. The assumption here is that all opinions are generally directed at a specific topic/object which none of the above frames targets accurately and consistently. For example, in movie reviews, extracted aspects could be: music, actors, or lights. When the customers are writing about a movie, they express their opinions about these aspects like what they think of the actors or the music choice.

2.1.4 User Level

In addition to the previous levels of analysis, some studies carry out the analysis on a user level that looks into users' networks and predicts users sentiment based on the sentiment of neighbouring users. In addition, a number of studies use joint models that combine two or three different levels, where the knowledge about one level, say documents, helps predict the sentiment of another level, say sentences.

CHAPTER 3

TECHNIQUES OF SENTIMENT ANALYSIS

3.1 Methodology

SA has five main steps. They are the: (1) data gathering (2) preprocessing the data, (3) extracting the features, (4) applying the analysis (lexicon based or machine learning based), and (5) evaluating the results. A summarization of the challenges in sentiment analysis research is presented. We highlight the methodology steps, pros, and cons for each of the studies. The following subsections explain the sentiment analysis models steps and literature related to it in more detail.

3.1.1 Data gathering

Opinions can be collected from different sources. Nowadays, the most common source for collecting data is social media websites such as Facebook and Twitter. These are considered multi-domain and contain reviews about different topics. The data could also be collected from domain-specified websites such as Trip advisor for tourism related reviews, IMDB for movie reviews, and Amazon for product reviews.

3.1.2 Preprocessing

Preprocessing is the process of cleaning the data from unwanted elements. It increases the accuracy of the results by reducing errors in the data. Not using preprocessing, such as spelling corrections, may lead the system to ignore important words. On the other hand, over using preprocessing techniques may sometimes cause loss of important data. There are general preprocessing techniques and Twitter-related preprocessing techniques.

3.1.3 Features

Features give a more accurate analysis of the sentiments and detailed summarization of the results. The most common features in sentiment analysis are n and POS tagging. Due to the complexity of the challenges, some research explored other linguistic features, which are described in detail below. Additionally, there are other less common features related to lexicons, Twitter, and punctuation.

3.1.4 Choosing and applying sentiment analysis method

There are three main methods to sentiment analysis. They are the (1) machine learning, (2) lexicon-based, and (3) hybrid or combined methods. These methods are described below.

3.1.5 Evaluating the Results

As a classification problem, SA uses the evaluation metrics of Precision, Recall, F-score, and Accuracy. Also, average measures like macro, micro, and weighted F1-scores are useful for multi-class problems. Depending on the balance of classes of the dataset the most appropriate metric should be used.

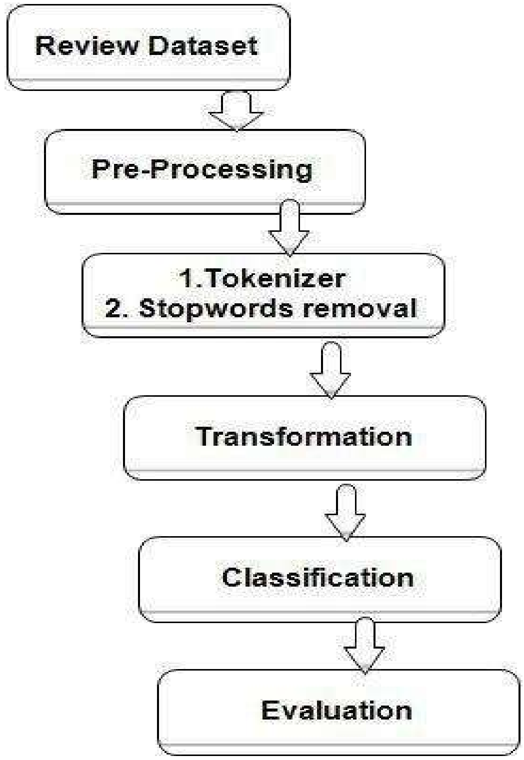

Figure 1. Steps-to-Evaluate-Sentiment-Analysis

3.2 Machine learning

Machine learning method is the most common method used in SA. This method uses classifiers to automatically detect the labels of the new data. It can be used only when the dataset is annotated. Many researchers annotate the data manually. Some annotate it automatically using. Due to the rapid growth of social media, bulk of user generated data is available now. Analyzing the sentiments and accurate classification of this gigantic amount of data is a very challenging task. Most of the data available on the internet is in the textual form as it is the most natural and readable form for presenting the thoughts and opinions to the users [1]. In this research Machine Learning algorithms and techniques for sentiment analysis are deeply analyzed. These algorithms are more adaptive to changing inputs. Unigrams (single word), Bigrams (dual word) and N-grams (multi Words) are used by different algorithms for data labelling and data processing. Machine learning techniques are generally used for binary classification and predictions of sentiments as either positive or negative. Machine learning algorithms are further classified in the following categories [2].

3.2.1 Supervised: In these algorithms training dataset with the pre-labeled classes are given and on the basis of this trained dataset the inputs are labeled with the output class/result [3]. These algorithms classify the input data set with the help of trained classifier. Training data is composed of a set of training examples, each of them comprise of input object and desired output results. An inferred function is created by analyzing the training data by supervised learning methods that can be later used for mapping new incoming data which is also called the test data. Mostly machine learning techniques use the supervised approach. It can be further categorized in two methodologies. I.e. Classification and Regression. Most common examples of supervised machine learning algorithms are Linear Regression, Random Forest and Support Vector Machines (SVM). Supervised learning includes two categories of algorithms: *regression* and classification algorithms. There's a significant difference between the two: Classification: Classification is a problem that is used to predict which class a data point is part of which is usually a discrete value. From the example I will buy this product, predicting whether a person is likely to default on a buy or not is an example of a classification problem since the classes we want to predict are discrete: "likely to pay a bill" and "not likely to pay a bill". Regression: Regression is a problem that is used to predict continuous quantity output. A continuous output variable is a real-value, such as an integer or floating point value. For example, where classification has been used to determine whether or not it will flood in this year or not, a regression algorithm will be used to predict the amount of flood's water.

3.2.2 Un-supervised: These type of machine learning algorithms takes the unlabeled input data and then with the help of different algorithms hidden structure/pattern is discovered unlike the supervised learning this technique does not use the pre-label data to train the classifier. Unsupervised machine learning can be further divided into clustering and association, the most common example of unsupervised machine learning algorithms are K-means and Apriority algorithm. Some use cases for unsupervised learning more specifically, clustering include: Customer segmentation, or understanding different customer groups around which to build marketing or other business strategies. Genetics, for example clustering DNA patterns to analyze evolutionary biology

3.2.3 Semi-Supervised: This is the middle class machine learning algorithms which mainly define solution by using comparison techniques. These type of algorithms deal with the both labeled and unlabeled data sets reviewed different lexicon based tools and techniques and mentioned the comparison between the features and accuracy results of different lexicon techniques. Taking it a step ahead, different Machine Learning techniques/algorithms are studied and analyzed in this research. A comprehensive analysis is also formulated between different techniques and accuracies.

3.3 Lexicon-based

The lexicon-based method is usually implemented when the data are unlabeled. Lexicons are sentiment dictionaries with the word and its occurring sentiment or sentiment score. Lexicons are used to label the data and to predict the polarity. Some researchers have created lexicons. However, most of these lexicons are not publicly available. There exist lexicons for MSA text, some of which is obtained using the translation of English lexicons. However, MSA lexicons are limiting as most social media users write informally. Therefore, it is important to research different languages dialects which have been very limited except for the Egyptian dialect, El-Makky et al, Al-Sabbagh and Girju etc. The authors created a lexicon consisting of 600 positive words/phrases and more than 900 negative words/phrases and 100 neutral words/phrases. The words and phrases are the most frequently Arabic words/phrases used over the web. In the next subsection, we present more examples of lexicons. The lexicon-based method is usually known as a weak method in comparison with machine learning method. To this end, only a few research articles used the lexicon-based method alone to analyze sentiment in the different languages. Indeed, most have adopted the lexicon-based approach along with machine learning—that is, the hybrid approach. This approach will be discussed in the next subsection. Pertaining to studies that solely used the lexicon-based approach, formal and informal different language's texts that were taken from online reviews and news articles have been examined. Al-Subaihin et al proposed the design of a lexicon based approach to conduct sentiment analysis on informal Arabic text. They used human-based computing to help build a lexicon. Their system consists of two different parts: the first part is a free online computer game that aimed to collect annotations of reviews from online players. The game has two players who are asked to highlight all the words or phrases that have positive and negative meaning. The aim of the game was to build a lexicon with positive and negative words. The second part was the sentiment analyzer which classified reviews according to their sentiments using sequence patterns and lexicons created from previous games. They tagged words to POS, NEG, ENT, or NO if it is positive, negative, entity, or a negation, respectively.

3.4 Hybrid techniques

The hybrid or combination approach uses both lexicon- and machine learning-based methods. This approach is more dominant in the relevant literature and is usually known to have a higher performance that lexicon-based method and machine learning method alone. The lexicon scores are typically used as features to input in the classifier. Pertaining to lexicons, the research explored word-level and sentence-level modern standard languages (MSL), dialectical or informal languages (DL), and MSL and DL combined. As for the machine learning classifiers, support vector machines and naive Bayes were the dominating methods. Yet, other approaches like K-nearest neighbour and entropy were also used. Specifically, El- Halees contrasted the accuracy rates of three methods: The lexicon-based only, combined lexicon-based and machine learning-based maximum entropy, and combined lexicon-based and machine learning-based K-nearest neighbour. Firstly, lexicon-based method was used to classify the documents. For the lexicon, they used SentiStrength software and translated words from English to different languages and an online dictionary. Then, the classified documents were used as a training set for maximum entropy method which classified other documents. Finally, the k-nearest method used classified documents from both the lexicon-based method and maximum entropy as a training set to classify the rest of the documents. The hybrid method in combination with the maximum and the k-nearest method led to the highest accuracy at 80%.

3.5 Classification Algorithms for Sentiment Analysis

There are various classification algorithms that are used to make predictions such as:

Neural Networks: These technique is used to read any text with AI. This is the latest algorithm in medical science for detecting skin cancer. Has various use cases. An example is in Computer Vision which is done through convolutional neural networks (CNN). We can read more on how Google classifies people and places using Computer Vision together with other use cases on a post on Introduction to Computer Vision that my student wrote.

K-NN: K-Nearest Neighbors is often used in search applications where you are looking for "similar" items. One of the biggest use cases of K-NN search is in the development of Recommender Systems.

Decision Trees: Decision trees are used in both regression and classification problems. A decision tree can be used to visually and explicitly represent decisions and decision making. They can be used to assess the characteristics of a client that leads to the purchase of a new product in a direct marketing campaign.

Random Forest: Random Forest algorithms can also be used in both regression and classification problems. It builds multiple decision trees and merges them together to get a more accurate and stable prediction. It can be used in a number of circumstances including image classification, recommendation engines, feature selection, etc.

Support Vector Machines (SVM): This is a fundamental data science algorithm which can be used for both regression and classification problems. However, it is mostly used in classification problems. It has a plethora of use cases such as face detection, handwriting recognition and classification of images just to mention a few.

Naive Bayes: This is a simple and easy to implement algorithm. A classical use case for Naive Bayes is document classification where it determines whether a given text document corresponds to one or more categories. It can be used in classifying whether an email is Spam or not Spam or to classify a news article about technology, politics or sports. I've also previously done sentiment analysis using Naive Bayes. You can find the notes and code here.

CHAPTER 4

CHALLENGES IN SENTIMENT ANALYSIS

4.1 What is challenges in Sentiment Analysis?

Challenges explain that a call to someone to participate in a competitive situation or fight to decide who is superior in terms of ability. But here challenges means it is possible to challenges the report's assumptions. Anyone can solve any problem with appropriate technique or algorithms. Only they accepted this challenges. In this paper we will represent some of problems for analysis with appropriate technique or algorithms. The main problems that exist in the current techniques are inability to perform well in different domains, inadequate accuracy and performance in sentiment analysis based on insufficient labeled data, incapability to deal with complex sentences that require more than sentiment words and simple analyzing. SA is the process of determining whether a piece of writing is positive, negative or neutral. SA helps data analysts within large enterprises gauge public opinion, conduct nuanced market research, monitor brand and product reputation and understand customer experiences. For these reasons they need to some of analysis with the customers emotions. That is called challenges for the sentiment analysis. Some of these has solution techniques, some of these are difficult to solve. For example, "This refrigerator will lasting for long life". Here need to customer's feedback. If customers give feedback with positive. It means the product is marketable. If customers give feedback with negative. It means the product is not perfect for marketing with maximum profit. It depends on SA. Percent of feedback is negative, positive or neutral.

4.2 Challenges in Sentiment Analysis

There are many challenges in the world to analysis and finding solution for every financial organization. Below some of challenges which is analytical and creative challenges. We face everyday life with these problems and it will be increase day by day. We also face new types of problems for globalization. Because every language has different kinds of sentiment words to explain their sentiments. So it will be difficult day by day to analysis new sentiments. We hope and think that it will be helped for the global commercial companies which companies marketing with globally. Some of important challenges below.

Sentiment Analysis Challenge No. 1: Language Problem

In Opinion Mining, English language is very well used because of its resource availability means lexicons, dictionaries and corpora but analyzed get attracted by using OP with language other than English (Arabic, Bengali, Chinese and German etc.). Therefore, analysis face a challenge for building resources i.e. lexicons, dictionaries and corpora for these languages [4]. There are more than twenty four thousands languages in the whole world and every language has different word to explain or react emotions. It is difficult to define every language's sentiment. So, it is the big and complex problem for the modern technology. Because if we solve this problem properly we get an ultra-modern age within a short time for development in AI sectors.

Challenge No. 2: Natural Language Processing (NLP)

Using NLP in the OP process needs more enhancements because it attracts the researches. And NLP provide better OP results and provides good language understanding. There is a need to pay more attention in the research of domain- dependent opinion mining or context- based opinion mining because domain specific OP gives good result than domain independent corpus. And domain specific OP is difficult or more complicated to build [5].

Challenge No. 3: Fake Opinion

It is also called Fake review and refers to bogus or fake reviews which misguide the readers or customers by providing them untruthful negative or positive opinions related to lower the reputation of any object. These spams make sentiment opinion useless in various application areas. This is a social challenge faced by the opinion mining and in spite of this challenge, OP made progress.

Challenge No. 4: Sarcasm Detection

To date, many sophisticated tools and approaches have been proposed to deal with sarcasm. More recently, [6] employ deep neural network (pre-trained convolutional neural network) for identifying sentiment, emotion, and personality features for sarcasm detection. Looking closer at these works, they mostly focus only on detecting the sarcasm in the text and not on how to cope with it in the sentiment analysis task. This raises the interesting question about how sarcasm may or may not affect the sentiment of the tweets and how to deal with sarcastic tweets in both the training and prediction phases. Rill of et al. [7] have proposed an algorithm to recognize the common form of sarcasm which flips the polarity in the sentence. These kinds of polarity- reverser sarcastic tweets often express the positive (negative) sentiment in the context of a negative (positive) activity or situation. However, Maynard et al. [8] show that determining the scope of sarcasm in tweets is still challenging. In fact, the polarity of sarcasm may apply to part of a tweet or its hashtags but not necessarily the whole. As a result, dealing with sarcasm in the task of sentiment analysis is an open research issue worth more work. Based on our observation, 7% of Trump's tweets and 6% of Clinton's tweets are sarcastic. Among these sarcastic tweets, 39% and 32% of them were classified incorrectly by our system. In terms of the training set, our hypothesis is that excluding the sarcastic instances from the training set will remove the noise and improve the quality of our training set.

Challenge No. 5: Emotion Analysis

Study of sentiment has evolved to the study of emotions, which has finer granularity. Positive, negative, and neutral sentiments can be expressed with different emotions such as joy and love for positive polarity; anxiety and sadness for negative; and apathy for neutral sentiment. Our emotion analysis on who tweeted #IVOTED in the 2016 US presidential election showed that Trump had many more tweets and individuals expressing joyful emotion compared to Clinton. Though the sentiment analysis favored Hillary in the early hours, emotion analysis was showing better support for Trump. We considered emotion as a better criterion for predicting people's action like voting and usually there are significant emotional differences in the tweets which belong to the same polarity. This was key to our successful prediction of the 2016 election.

Challenge No. 6: Hashtags

Recently, there has been a surge of interest in distant supervision, which is training a classifier on a weekly labeled training set [9]. In this method, the training data gets automatically labeled based on a set of heuristics. In the context of sentiment analysis, using the emoticons :) and :(and other similar emoticons as a positive and a negative label respectively is one way of using distant supervision. Hashtags are also widely used for different machine learning tasks such as emotion identification [20]. People use a plethora of hashtags in their tweets about the election. Due to the dynamic nature of the election domain, the quality, quantity, and freshness of labeled data plays a vital role in creating a robust classifier. It is therefore desirable to use popular hashtags that each candidate's supporters use as a weak label in our dataset. However, our analysis for the 2016 election showed that hashtags were widely used for sarcasm, so using popular hashtags for automatic labeling leads to incorrectly labeling instances. For example, through the election only 43% of tweets containing #Imwithher were positive for Clinton, while it was used sarcastically in 27% of tweets. Consequently, our experiments show that using those hashtags as a feature for our classifier will decrease accuracy rather than increase it [14].

Challenge No. 7: Links

All existing techniques for tweet classifiers rely merely on tweet contents and ignore the content of the documents they point to through a URL. However, around 36% of the 2016 election tweets contain a URL to an external link. In the 2012 election [11], we noticed d that 60% of tweets from very highly engaged users contain URLs. Those links are crucial as without them often the tweet is incomplete and inferring the sentiment is impossible or difficult even for a human annotator. Therefore, our hypothesis is that incorporating the content, keywords or title of the documents that a URL points to as a feature will cause a gain in our performance. To the best of our knowledge, there is no work on tweet classification that expands tweets based on their URLs. However, link expansion has successfully been applied to other problems such as topical anomaly detection [12] and distant supervision [13].

Challenge No. 8: Sentiment Analysis versus Emotion Analysis

Study of sentiment has evolved to the study of emotions, which has finer granularity. Positive, negative, and neutral sentiments can be expressed with different emotions such as joy and love for positive polarity; anxiety and sadness for negative; and apathy for neutral sentiment. Our emotion analysis on who tweeted #IVOTED in the 2016 US presidential election showed that Trump had many more tweets and individuals expressing joyful emotion compared to Clinton. Though the sentiment analysis favored Hillary in the early hours, emotion analysis was showing better support for Trump. We considered emotion as a better criterion for predicting people's action like voting and usually there are significant emotional differences in the tweets which belong to the same polarity. This was key to our successful prediction of the 2016 election.

Challenge No. 9: Vote counting versus engagement counting

Most/all of the aforementioned challenges affect the quality of our sentiment analysis approach. It is also very important to correlate a user's online behavior and opinion with their actual vote. Chen et al. [4] show the more important role of highly engaged users in result prediction of the 2012 election. There are two plausible explanations for this. First, the more a user tweets, the more reliably we can predict his/her opinion. Second, highly active people are usually more influential and more likely to actually vote in the real world. That is why an election monitoring system should report both user-level normalized sentiment in addition to a tweet-level one. It is the end user analyzer's task to consider both of these factors in prediction.

Challenge No. 10: Multilingual sentiment analysis

Work on multilingual sentiment analysis has mainly addressed mapping sentiment resources from English into morphologically complex languages. Mihalcea Banea, and Wilber (2007) use English resources to automatically generate a Romanian subjectivity lexicon using an English–Romanian dictionary. The generated lexicon is then used to classify Romanian text. Wan (2008) translated Chinese customer reviews to English using a machine translation system. The translated reviews are then annotated using rule-based system that uses English lexicons. A higher accuracy is achieved when using ensemble methods and combining knowledge from Chinese and English resources. Balahur and Turchi (2014) conducted a study to assess the performance of statistical sentiment analysis techniques on machine- translated texts. Opinion-bearing phrases from the New York Times Text (2002–2005) corpus were automatically translated using publicly available machine-translation engines (Google, Bing, and Moses). Then, the accuracy of a sentiment analysis system trained on original English texts was compared to the accuracy of the system trained on automatic translations to German, Spanish, and French. The authors conclude that the quality of machine translation is acceptable for sentiment analysis to be performed on automatically translated texts. Salameh,Mohammad, and Kiritchenko (2015) conducted experiments to determine loss in sentiment predictability when they translate Arabic social media posts into English, manually and automatically. As bench marks, they use manually and automat- ically determined sentiment labels of theArabic texts. They show that sentiment analysis of English translations of Arabic texts produces competitive results, w.r.t. Arabic sentiment analysis. They also claim that even though translation significantly reduces human ability to recover sentiment, automatic sentiment systems are affected relatively less by this. Some of the areas less explored in the realm of multilingual sentiment analysis include: how to translate text so as to preserve the degree of sentiment in the source text; how sentiment modifiers such as negators and modals differ in function across languages; under- standing how automatic translations differ from manual translations in terms of sentiment; and how to translate figurative language without losing its affectual gist.

Challenge No. 11: Negation Detection

In linguistics, negation is a way of reversing the polarity of words, phrases, and even sentences. Researchers use different linguistic rules to identify whether negation is occurring, but it's also important to determine the range of the words that are affected by negation words. There is no fixed size for the scope of affected words. For example, in the sentence "The show was not interesting," the scope is only the next word after the negation word. But for sentences like "I do not call this film a comedy movie," the effect of the negation word "not" is until the end of the sentence. The original meaning of the words changes if a positive or negative word falls inside the scope of negation—in that case, opposite polarity will be returned. The simplest approach for dealing with negation in a sentence, which is used in most state-of-the-art sentiment analysis techniques, is marking as negated all the words from a negation cue to the next punctuation token. The effectiveness of the negation model can be changed because of the specific construction of language in different contexts. There are several forms to express a negative opinion in sentences: Negation can be morphological where it is either denoted by a prefix ("dis-", "non-") or a suffix ("-less"). Negation can be implicit, as in "with this act, it will be his first and last movie"—it carries a negative sentiment, but no negative words are used. Negation can be explicit, as in "this is not good." Having samples with different types of described negations will increase the quality of a dataset for training and testing sentiment classification models within negation. According to the latest research on recurrent neural networks (RNNs), various architectures of LSTM models outperform all other approaches in detecting types of negations in sentences. In the paper Effect of Negation in Sentiment Analysis, a sentiment analysis model evaluated 500 reviews that were collected from Amazon and Trustedreviews.com. The authors show a comparison of the models with and without negation detection. Their evaluation demonstrates how considering negation can significantly increase the accuracy of a model.

Challenge No. 12: Word Ambiguity

Word ambiguity is another pitfall you'll face working on a sentiment analysis problem. The problem of word ambiguity is the impossibility to define polarity in advance because the polarity for some words is strongly dependent on the sentence context. Lexicon-based sentiment analysis approaches are popular among existing methods. An opinion lexicon contains opinion words with their polarity value. There are some public opinion lexicons available on the internet: SentiWordNet, General Inquirer, and SenticNet, among others. Because word polarity varies in different domains, it is impossible to develop a universal opinion lexicon that has a polarity for every word. For example: 1. "The story is unpredictable." 2. "The steering wheel is unpredictable." These two examples show how context affects opinion word sentiment. In the first example, the word polarity of "unpredictable" is predicted as positive. In the second, the same word's polarity is negative.

Challenge No. 13: Multipolarity

Sometimes, a given sentence or document—or whatever unit of text we would like to analyze will exhibit Multipolarity. In these cases, having only the total result of the analysis can be misleading, very much like how an average can sometimes hide valuable information about all the numbers that went into it. Picture when authors talk about different people, products, or companies (or aspects of them) in an article or review. It's common that within a piece of text, some subjects will be criticized and some praised. Here, the total sentiment polarity will be missing key information. This is why it's necessary to extract all the entities or aspects in the sentence with assigned sentiment labels and only calculate the total polarity if needed. Let's consider an example which consists of multiple polarities: “The audio quality of my new laptop is so cool but the display colors are not too good.” Some sentiment analysis models will assign a negative or a neutral polarity to this sentence. To deal with such situations, a sentiment analysis model must assign a polarity to each aspect in the sentence; here, “audio” is an aspect assigned a positive polarity and “display” is a separate aspect with a negative polarity. For a more in-depth description of this approach, I recommend the interesting and useful paper Deep Learning for Aspect-based Sentiment Analysis by Bo Wanf and Min Liu from Stanford University.

Challenge No. 14: Sentiment analysis of Arabic text

Arabic Language is a complex and heavy grammarian’s language. So, it is difficult to SA in Arabic Language. We identified the challenges and limitations that researchers highlight. While there are many challenges incurred when conducting sentiment analysis research, we limit this section to those challenges that are specific to the Arabic language. Because maximum or every languages in the whole are left to right. But Arabic language is right to left. So, it is difficult to detection and analysis with any methods.

CHAPTER 5

PROBABLE SOLUTIONS IN CHALLENGES

5.1 Probable Solutions Techniques

There are many different languages in the world for different nation or colony. For this reason a same method or algorithm is not work for the languages of whole world. Here an example, Informal language could be described as language that ignores the standard rules of grammar and spelling. In general, the Arabic language is written from right to left, while English is written from left to right. There is no capitalization in Arabic, unlike in English. It is difficult to analysis with Arabic language with SA. So, we used some step with via English language. To solve this problem here used Naïve Bayes Algorithm.SA uses various Natural Language Processing (NLP) methods and algorithms, which we'll go over in more detail in this section. The main types of algorithms used include: 1) Rule-based systems that perform sentiment analysis based on a set of manually crafted rules. 2) Automatic systems that rely on machine learning techniques to learn from data. 3) Hybrid systems that combine both rule-based and automatic approaches. Finally, SA is done using different Machine Learning approaches including maximum entropy, SVM, convolutional neural network and long short term memory to solve these problems. The classification step usually involves a statistical model like Naïve Bayes, Logistic Regression, Support Vector Machines, or Neural Networks: 1) Bayes: This is a probabilistic model based on the Bag of words module to store only the frequencies of each word and ignore their positioning with respect to each other. So it is a family of probabilistic algorithms that uses Bayes's Theorem to predict the category of a text. 2) Support Vector Machines: SVM works for multiple machine learning problems such as regression and classification. The main principle that works behind SVM is finding a particular linear classifier that separates all the classes in the search space in the best possible manner. It improve feature sets were used for the reviews after pre-processing. So it is a non-probabilistic model which uses a representation of text examples as points in a multidimensional space. Examples of different categories (sentiments) are mapped to distinct regions within that space. Then, new texts are assigned a category based on similarities with existing texts and the regions they're mapped to. 3) Deep Learning: This is a diverse set of algorithms that attempt to mimic the human brain, by employing artificial neural networks to process data. 4) Long Short Term Memory (LSTM): LSTM is a type of Recurrent Neural Networks that is capable of handling long term dependencies as otherwise it was difficult for RNN to connect multiple long term dependencies. RNN focus on the issue of considering the past information so as to understand the meaning of current and next words. Deep Learning techniques are also known as Artificial Neural Networks. These methods or techniques have given great advances in NLP in the last few years. One particular model known as the LSTM (Long Short-Term Memory) has been dominating most NLP tasks in the last few years achieving state of the art results. An LSTM approach reads text sequentially and stores relevant information to the task at hand. Within the LSTM there are cells which control what information is remembered and what is forgotten. In the case of sentiment analysis negation is very important. For example, the difference between “good” and “not good”. An LSTM trained to predict sentiment will learn that this is important and get good at understanding which words should be negated. By reading large amounts of text an LSTM can be thought of as ‘learning’ grammar rules. Deep learning architectures continue to advance with innovations such as the Sentiment Neuron which is an unsupervised system (a system that does not need labelled training data) coming from Open.ai. Google has developed the Transformer and recently added retraining (pre-training is where you train a model on a different task before fine tuning with your specialized dataset) to the transformer with a technique known as BERT , achieving state of the art results across many NLP tasks. Imagine that scenario: they’re the owner of a small delivery business and they receive about 20 responses to their email surveys every month. They could (and should), read these themself and perform their own analysis by hand. Now, imagine that they are receiving 30,000 responses per month. That’s more than a thousand responses each day! Needless to say this is *impossible* as a part of a business owner’s day job. Then, there’s the question of bias. Everyone knows “those days” where things go wrong and they’re in a foul mood even before reaching the office. The risk of you interpreting messages and any form of communication more negatively, is rife. They must also have their own, preconceived opinions about the topic at hand. All of this can influence how you interpret the text they need to analyze. They’ll also need to *summarize* the feedback into a few actionable insights, so that it is meaningful for their company to make use of. Also the insights need to be translated into presentable form so that it is easy to grasp. Sentiment analysis is important because companies want their brand being perceived positively, or at least more positively than the brands of competitors.

5.2 Problems Definition

This paper mainly focuses on the challenge of Sentiment Analysis, takes into techniques for solutions, and makes a comparative analysis of different challenges with their solution techniques. There are many difficulties in these. For Sarcasm Detection , combination of features like n-grams, capitalization, sentiment score, part of speech are used together to generate feature vectors. Emoji’s Handling and Slang Word Standardization are handled during the preprocessing stage. But computer programs have problems recognizing things like sarcasm and irony, negations, jokes, and exaggerations the sorts of things a person would have little trouble identifying. And failing to recognize these can skew the results

CHAPTER 6

CONCLUSION AND FUTURE WORKS

6.1 Conclusion

In this thesis paper, we have explored multiple challenges in sentiments analysis and techniques, which is useful for them who want to analysis with Sentiments. There is some case which has no solution or better and perfect techniques to get result in early and easily. So, this paper is useful for any financial companies and is effective for any companies. Text pre- processing techniques have been addressed widely in the information retrieval field. In particular the effect of various text pre-processing techniques is examined in the sentiment analysis field. The sentiment of online movie reviews is investigated. Here we try to explaining the techniques for solve the challenges. Which has solution or has no solution. SA is a rapidly growing field of analysis due to the explosive growth in digital information. In this age modern world of artificial intelligence, SA is one of the necessary tools to extract emotion information from massive data. It is applied to a variety of user data from customer's reviews from social networks posts. To the best of our knowledge, SA based on the categorization of users by demographics to less work.

6.2 Future Works

In the future work, we can also include exploration of new methods to solve new problems for analyzing of challenges in SA. The field of SA is an exciting new analytical direction due to large number of real world applications where discovering people's opinion is important in better decision making. SA is one of the significant components of the area of the development of techniques for the document level. Almost, some of people have started expressing their opinions on the Web that increased the need of analyzing the opinionated online content for various real world applications. There are many researches are present in literature and online for detecting sentiment form the text. Still, here are many scope of improvement of these existing SA models. SA models can be improved further with more semantic and commonsense knowledge. For this reason in future need to improve the methods or techniques to solve challenges. So, that result is being gotten early and within a short time.

REFERENCES

[1] I. Smeureanu and M. Zurini, “Spam Filtering For Optimization in Internet Promotions Using Bayesian Analysis,” J. Appl. Quant. Methods, vol. 5, no. 2, 2010.

[2] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, “A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle,” Proc. 50th Annu. Meet. Assoc. Comput. Linguist., no. July, pp. 115–120, 2012.

[3] P. Goncalves, B. Fabrício, A. Matheus, and C. Meeyoung, “Comparing and Combining Sentiment Analysis Methods (SVM) Categories and Subject Descriptors,” Proc. first ACM Conf. Online Soc. networks, pp. 27–38, 2013. 7

[4] Peng, Haiyun, Erik Cambria, and Amir Hussain. "A review of sentiment analysis research in Chinese language."

[5] Rajput, Adil. "Natural Language Processing, Sentiment Analysis, and Clinical Analytics." Academic Press, 2020.

[6] S. Poria, et al., "A deeper look into sarcastic Tweets using deep convolutional neural networks," arXiv preprint arXiv: 1610.08815, 2016.

[7] E. Riloff, et al., "Sarcasm as Contrast between a Positive Sentiment and Negative Situation," EMNLP, vol. 13, 2013.

[8] D. Maynard and M. A. Greenwood, "Who cares about Sarcastic Tweets? Investigating the Impact of Sarcasm on Sentiment Analysis," LREC, 2014.

[9] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," CS224N Project Report, Stanford, vol. 1, no. 12, 2009.

[10] W. Wang, et al., "Harnessing twitter ‘big data’ for automatic emotion identification," Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom), IEEE, 2012.

[11] L. Chen, et al., "Extracting Diverse Sentiment Expressions with Target-Dependent Polarity from Twitter," Sixth International AAAI Conference on Weblogs and Social Media, 2012.

[12] P. Anantharam, K. Thirunarayan, and A. Sheth, "Topical anomaly detection from twitter stream," Proceedings of the 4th Annual ACM Web Science Conference, ACM, 2012.

[13] W. Magdy, et al., "Distant Supervision for Tweet Classification Using YouTube Labels," ICWSM, 2015.

[14] M. Ebrahimi, A.H. Yazdavar and A. Sheth,"Challenges of Sentiment Analysis for Dynamic events," in *IEEE Intelligent Systems*, vol. 32, no. 5.September/October 2017.