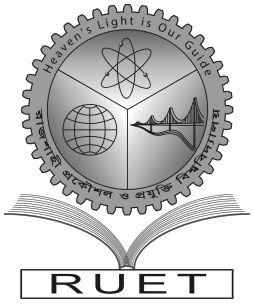
Heaven's Light is Our Guide



DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Rajshahi University of Engineering & Technology, Bangladesh

Sentiment Analysis: Comparative Analysis On Different Machine Learning and Deep Learning Approaches.

Author

Nibaron Kumar Das

Roll No. 1603104

Department of Computer Science & Engineering

Rajshahi University of Engineering & Technology

Supervised by

Dr. Mir Md.Jahangir Kabir

Professor

Department of Computer Science & Engineering

Rajshahi University of Engineering & Technology

ACKNOWLEDGEMENT

At first, we would like to thank the Almighty for giving us the opportunity and enthusiasm along the way for the completion of our thesis work.

Foremost, I would like to express our sincere appreciation, gratitude, and respect to our supervisor Dr. Mir Md. Jahangir Kabir, Professor of Department of Computer Science and Engineering, Rajshahi University of Engineering and Technology, Rajshahi. Throughout the year he has not only given us technical guidelines, advice and necessary documents to complete the work he has also given us continuous encouragement, advice, helps and sympathetic co-operation whenever he deemed necessary. His continuous support was the most successful tool that helped us to achieve our result. Whenever we were stuck in any complex problems or situation he was there for us at any time of the day. Without his sincere care, this work not has been materialized in the final form that it is now at the present.

I am also grateful to all the respective teachers of Computer Science and Engineering, Rajshahi University of Engineering and Technology, Rajshahi for good valuable suggestions and inspirations from time to time.

Finally, I convey my thanks to my parents, friends, and well-wishers for their constant inspirations and many helpful aids throughout this work.

November 21, 2022 Nibaron Kumar Das

RUET, Rajshahi

Heaven's Light is Our Guide

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Rajshahi University of Engineering & Technology, Bangladesh

CERTIFICATE

This is to certify that this thesis report entitled ***"Sentiment Analysis: Comparative Analysis On Different Machine Learning and Deep Learning Approaches."*** *submitted by* ***Nibaron Kumar Das, Roll:1603104*** *in partial fulfillment of the requirement for the award of the degree of Bachelor of Science in Department of Computer Science & Engineering of Rajshahi University of Engineering & Technology, Bangladesh is a record of the candidate own work carried out by him under my supervision. This thesis has not been submitted for the award of any other degree.*

Supervisor

Dr. Mir Md.Jahangir Kabir
Professor
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi-6204

External Examiner

Bayezid Islam
Assistant Professor
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi-6204

ABSTRACT

Sentiment analysis has drawn a lot of interest with the growing popularity of the internet.People share their ideas on social media platforms, personal blogs, and online review sites. The sentiment analysis approach captures positive and negative opinions toward individuals, organizations, locations, events, and ideas. The main concentrated research terminology used for sentiment analysis are Natural Language Processing (NLP) and Data Mining approaches to harvest this special insights from the linguistic data.Different traditional techniques like machine learning, lexicon based technique and hybrid techniques have been applied in sentiment analysis or opinion mining. In the developed strategy, Twitter data analysis is used to identify global sentiment using machine learning techniques.In this work, we have considered different approaches such as Logistics Regression, Random forest, MultinomialNB, Support Vector Machine(SVM) and Bi-LSTM to extract sentiment from 3 level of sentiment output.The objective is to compare both implementation outcomes and demonstrate the most effective method for sentiment analysis on social media for semi-structured data.

Key-Words: Natural Language Processing (NLP),Logistics Regression, Random forest,Support Vector Machine(SVM), MultinomialNB, Bi-LSTM.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Introduction

The purpose of this chapter is to provide a introduction of sentiment analysis and the overall finding of this research. In section 1.2 and 1.3, the problem statement and motivation for this paper thesis are described briefly which will solve the problem that is being dealt with.The difficulties and sentiment challenges are highlighted in section 1.4. Then, in section 1.5, contributions of the thesis are outlined.The structure of the thesis book is provided in section 1.6. Finally, a conclusion will end the chapter.

1.2 Problem Statement

In modern time, Technology has revolutionized our daily life style and data has grown at a breakneck speed to such a big value. People express their opinions, ideas, thoughts, feed-backs and communicate with each other through different social media platforms, blogging, e-commerce site and so on. But, the more we express, the more the date increases to great extent. This great extent of data is unstructured, noisy and complex. To get knowledge from these data, data mining techniques can be used for mining interesting, valid and significant patterns for the users[1].

Sentiment analysis, another term, opinion mining used interchangeably, is the field of study that examine human's opinions, sentiments, appraisals, evaluations, emotions and attitudes towards entities for example services, products, individuals, organizations, issues, topics, events, and the attributes [2].

Sentiment analysis has become one of the most important sources in decision making.You can't just know what matters most to the people you want to like your brand if you don't have the resources to study all of the common feeling.Sentiment is the easy idea to comprehend. It's just a thought,an attitude, or a point of view. The tone or emotions expressed during a brand mention on social media may also expose the sentiment of a message. Sentiment analysis is an automated method for deducing a person's feeling about a topic from written or spoken words.

A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level—whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. Advanced, "beyond polarity" sentiment classification looks, for instance, at emotional states such as enjoyment, anger, disgust, sadness, fear, and surprise [3].

Figure 1.1: Sentiment analysis with polarities

We can classify our sentiment analysis problem into two major categories. first one refers as problem identifications that emphasize on the fact, the root of our problem, so that it can be solved. In sentiment analysis, our sentiment falls into two major buckets, binary polarities like the positive sentiment or the negative one.

The second major problem is the sentiment problem evaluation, an approach to analyze the reviewed data to classify them into multi-level. Appropriate sentiment evaluation to understand computer and human linguistic so that the research gap between sentiment challenges and sentiment evaluation can be minimized.

1.3 Motivation

In recent time, the sentiment analysis of user generated text is interesting for different practical reasons. It can be exploited in marketing, business, politics and social analysis

and other commercial aspect. Unstructured and complex data can be organized into structured manner so that it will become easy to understand and analysis, will make sense.

The most considered factor in sentiment analysis or opinion mining is the polarity. Polarity score determines if the opinion is positive, negative or neutral. Subjectivity is also an important factor in sentiment analysis. Subjectivity refers the weight of any opinions, views, issues, products or services. It specifies how much emphasis should be given on a specific reviewed data. For most explicit review, if any sentence is opinionated or not can be determined using subjectivity [4]. So, we have to combined the polarity and subjectivity for the better and more accurate result classifying the reviews data into different levels.

1.4 Research Challenges

In order to provide a more complete image, natural language processing uses machine learning and data mining, but the inherent complexity of language makes it hard to ensure that algorithms analyze tone and meaning correctly. Grammatical nuances, inferred meaning from facial expressions and body language, misspellings, uncertainty, and geographical or cultural differences in language are factors that restrict these algorithms.

Following are the common sentiment issues related to the opinionated text and should be addressed efficiently:

- Lack of comparative opinion analysis in detection techniques from various data types.
- Ensuring faster but accurate detection.
- Subjective words not expressed any opinion.
- Lack of proper tuning of hyper parameters.
- Maintaining high quality multi-class classification results.
- Multilingual opinion handling.
- Sarcasm and ironic detection.
- Lack of implementation on large datasets.
- Finding a feature extraction method that is both effective and practical.

1.5 Research Contributions

The main contributions of this are:

i.Development of Recurrent Neural Network (RNN) based model to analysis the social media sentiment.

ii.Analysis the sentiment into 3 major polarities: positive, negative and neutral.

iii.Consideration of the three types of data- text, summary and appended text with summary as input for sentiment classification.

iv.Comparative analysis between machine learning approach and Neural network approach.

1.6 Structure of the Thesis Book

Chapter 1: Introduction

Introductory discussion about the thesis topics.

Chapter 2: Literature review

This chapter gives an overview of the sentiment analysis and the research concluded in this field.

Chapter 3: Background on Sentiment Analysis

Overview for sentiment analysis and the important definitions in this domain. It examines the differentiate sentiment analysis techniques architecture, the importance of sentiment analysis and the sentiment analysis challenges.

Chapter 4: Recurrent Neural Network and LSTM

This chapter gives an overview of the Recurrent Neural Network and LSTM concluded in this field.

Chapter 5: Applied Method

This chapter presents the proposed technique and describe a hybrid model and how to evaluate sentiment and system score.

Chapter 6: Implementation and Result

This chapter presents their experimental results we achieved through comparing our proposed approach to existing relevant techniques.

Chapter 7: Conclusion and Future Work

It concludes the thesis and mentions the possible direction for future work.

1.7 Conclusion

Sentiment analysis is very promising research topic in demand under Natural Language processing (NLP), machine learning and Neural Network approach. The fundamental task here is to extract the polarity of the text where we express our opinions and emotions.

Chapter 2

Literature Review

2.1 Introduction

This chapter has described about previous work on this field. Improvement on this field was also mentioned and the way different machine learning approaches and deep neural network algorithms has been implemented on different datasets to gauge the performance of various proposed models. In section 2.2, The basic work in sentiment analysis and a review on the literature are presented to support the motivation of the proposed work. In the end, a conclusion in section 2.3 about the whole chapter will end the chapter.

2.2 Related Works and Contributions

'The pen is mightier than the sword' indicates that free speech (written language in particular) is a more powerful weapon than direct violence [5]. The study of sentiments is a collection of approaches, techniques and tools for detecting and extracting subjective knowledge from language,such as opinions and attitudes [6]. Sentiment analysis has historically been about opinion polarity,i.e. whether someone has a positive, neutral or negative view of something. A good or a service whose evaluation has been made public on the Internet has generally been the focus of sentiment analysis.

In recent years, we have seen a huge rise in the number of papers based on sentiment analysis and opinion mining. According to our data, nearly 7,000 papers were published on this subject and, more interestingly, 99 percent of the papers appeared after 2004, making sentiment analysis one of the research areas with the fastest growth. While the

present paper focuses on the research articles of sentiment analysis, we can see that in the general public, the subject is also gaining interest.

A large number of papers focused on addressing the problem of sentiment classification. They used different techniques for classifying the sentiments of individuals at various levels. Different machine learning [7] and lexicon-based approaches have been applied for sentiment analysis. Not only supervised techniques but also unsupervised techniques can achieve a good accuracy which was presented by Turney [8]. Dave, Lawrence, and Pennock [9] designed a model of semantic orientation for positive and negative words with scoring which is used to classify the review data.

Different machine Learning algorithms such as Naive Bayes Classifier,Max Entropy Classifier,Boosted Trees Classifier, support Vector Machine (SVM) classifier, Linear Regression(LR) classifier and Random Forest Classifier were implemented on text review data and after classification, an aggregate score was presented [10]. Rule-based and lexicon-based approaches were compared with different machine learning approaches and a summary was represented in [11].

Tripathy et al, proposed a hybrid approach[12] combining the SVM and ANN where SVM is used for feature selection and ANN to perform accuracy measurement. This hybrid approach provided better accuracy than single machine learning approach.A hybrid approach was proposed by Zhang et al [13] for sentiment analysis by combining both lexicon and learning based approaches and implemented on Twitter data.

In past years, the fuzzy-based approach has become a matter of attention for sentiment analysis. As opinions are fuzzy in nature, it allows refined classification based on strength of sentiment. Though fuzzy logic is a powerful tool, much work has not been done for multi-label sentiment analysis using this approach. Jefferson, Liu, and Cocea,proposed a fuzzy-based sentiment analysis technique that helps to define different degrees of sentiment with a limited number of classes[14].

In recent times, the use of Neural Network such as CNN, ANN and RNN playing a significant rule in sentiment analysis and opinion mining, the accuracy and result have improved with the help of Neural Network Model. X.Ouyang, P. Zhou, C. H. Li and L. Liu suggested a system called Word2vec + CNN (CNN). Firstly, Google's suggested word2vec was used to measure vector word representations which were a CNN input to use word2vec, to perform the word vector representation and to reflect the space of

terms. In this case, the parameters of CNN were initialized,which could increase the network output efficiently during this crisis. Secondly, they have crafted a fitting CNN architecture to evaluate your sentiment. During this architecture they used three pairs of convolutional layers and reservoirs. You assess your algorithm and create your neural network on your general public dataset.Then the results of the tests were related and compared to previous computer algorithms[15].

D.Goularas and S. Kamis used configurations of deep learning methods based on networks from CNN and LSTM are tested for Sentiment analysis of data from Twiter.Slightly, this assessment gave slightly with the state-of-the-art processes, inferior but comparable outcomes,it thus enables credible conclusions to be made about the Various configurations. The fairly poor efficiency of these the shortcomings of CNN and LSTM networks demonstrated the structures and on the ground.Regarding their configuration, it was noted when the CNN and LSTM networks are merged,they did better than they do alone when used. This is attributed to the Efficient method of minimizing dimensionality for CNN's and the When using LSTM, protection of word dependencies with networks. In addition, using multiple CNNs and LSTM Networks improve the system's efficiency. It demonstrates that having an adequate dataset, as expected,the main component for improving the performance of such products is Uh, programs. As a consequence, it seems like investing more time and time efforts to build good training sets are more challenging. Benefits rather than playing with various CNN and LSTM network combinations or configurations [16].

Kai Zhang,Wei Song,Lizhen Liu,Xinlei Zhao and Chao Du have identified a Bidirectional Long Short Term Memory model,which is used to extract the features of comments and measure the parameters optimized by the SoftMax classifier. Their model performs better on three distinct domain product statements, as demonstrated by experimental findings.There are still some barriers to the ideal solution [17].

S. Naz, A. Sharan and N. Malik proposed an SVM current Based classifier that hybridizes internal n-gram based Features with an external vector of emotions to maximize the A regular classifier based on n-grams. They performed experiments with four different sets of n-gram characteristics and three different techniques for weighting. Compared to other feature sets, the unigram characteristics show the best results in terms of precision outcome. Next, they have an external function with n-grams hybridized and they observed,

in terms of precision, the SVM classifier performed better.With regard to future work, they intended to discuss some of them. More specific external information that can function as a characteristic to provide more enhanced efficiency,set [18].
R.Monika,S.Deivalakshmi and B.Janet proposed an emotion analysis using deep feeling analysis Techniques for learning through the model of the LSTM network.They analyze data on tweets in social media for the airline industry Which in the training set shows better results. In addition,precision can be improved using the Network bidirectional LSTM(Bi-LSTM)[19].

2.3 Conclusion

Statistical methods with different machine learning algorithms are very good at retrieving text,splitting them into parts, counting their words and finally giving a positive or negative result.Recently, machine learning based sentiment analysis models are gaining prominence in the field.But high dimensionality is a curse for the performance of the algorithms. They reduce the highdimensional feature space with the help of feature selection and feature extraction techniques. Researchers have discovered that text generated by long and short users should be handled differently. A relaxing finding shows that reviewing short forms often makes it better than creating a long form,since filtering noise through the text is easier, the increasing length not always resulted in an improvement in the quantity of the text of the long form.

Chapter 3

Background on Sentiment Analysis

3.1 Introduction

This chapter provides information and discussion on sentiment analysis and different sentiment analysis extraction methods. In section 3.3, different Sentiment analysis approaches were discussed. Different levels on sentiment were described in section 3.6. Finally, in section 3.7, the overview of the whole chapter is given.

3.2 Sentiment Analysis Classification

Sentiment analysis, also known as opinion mining, is a computer research that examines how ratings, ideas, perceptions, assessments, behaviors, subjective opinions, emotions, etc. are mirrored in the text. We'll talk about sentiment analysis and the factors that affect it in the following sections.

It is crucial for the company or the event organizers to understand the opinions and sentiments of the targeted consumers or individuals who have reacted to it via social media in order to have a successful and well-established business or event. Social networking sites like Facebook, twitter etc make it simpler to convey thoughts, feelings, and opinions about any scenario in today's technologically advanced society. Social media responses from attendees and consumers are free-form and may include written text comments from them. There is therefore no better approach than through social media platforms to track the success of a product's advertising, a famous person, an event, or an organization's accomplishment.

Sentiment analysis has become one of the most important sources in decision making.You can't just know what matters most to the people you want to like your brand if you don't have the resources to study all of the common feeling.Social media blogs provide information on the general consensus regarding how well-liked the company is doing as a result. Through social media platforms, analysts can extract useful information from these blogs to aid in decision-making.

3.2.1 Sentiment and Emotion

A mental attitude is what is meant by a sentiment. A sentiment enables the speaker to express his emotion through gesture.On the other hand, Emotions are complex psychological states that can be very raw and natural.Sentiments are the way that emotions are expressed when they are connected to a social object.Sometimes, sentiment called emotions with actions.Sentiment analysis is one of the more creative yet useful applications of binary classification.

It evaluates text such as a product review, a tweet, or a comment left on a website and rates it on a scale of 0 to 1, with 0 denoting very negative sentiment and 1 denoting very positive sentiment.Sentiment can also be classified into three classes, such as positive, negative and neutral, we can use 0 as negative, 1 for neutral and 2 for positive sentiment.

(a) Sentiment

(b) Emotions

Figure 3.1: Sentiment & Emotion

3.3 Sentiment Analysis Approaches

There have many approaches for sentiment analysis on the linguistic data.Different algorithms can be incorporated into sentiment analysis models depending on the volume of data you need to examine and the level of precision you require for your model.Most research carried out in the field of sentiment analysis employs lexicon-based analysis or machine learning techniques. Machine learning techniques control the data processing by the use of machine learning algorithm and by classifying the linguistic data by representing them into vector form.

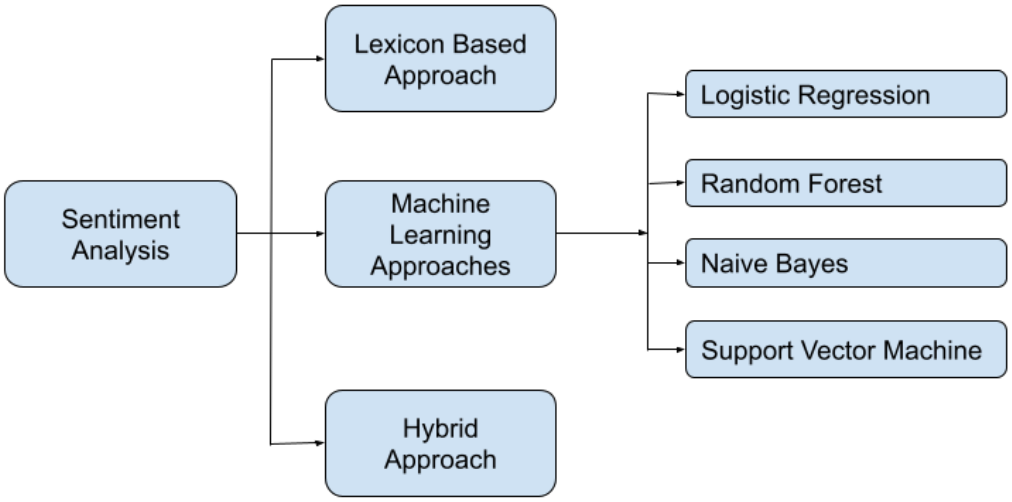


Figure 3.2: Sentiment Analysis Approaches

In this section we will discuss, major sentiment algorithm approaches.Sentiment analysis algorithms are basically classified into one of three types.

i. **Lexicon based Approach**:Using a set of manually created criteria, these systems carry out sentiment analysis automatically.

ii. **Machine Learning Algorithms**:Here system depends on different machine learning techniques like Logistic Regression, Random forest, Naive bayes, Support Vector Machine to learn from data.

iii. **Hybrid approach**:Here, we use both Ruled based Algorithm and Machine Learning Algorithms.

3.3.1 Lexicon Based Approach

This approach works with a simple rule engine, sometimes it is called rule Based approach.System uses set of rules crafted by human to identify polarity and subjectivity from the object.To support new phrases and terminology, it is possible to apply more sophisticated processing techniques and add additional rules. However, introducing new rules can affect prior results, and the system as a whole could grow exceedingly complex.

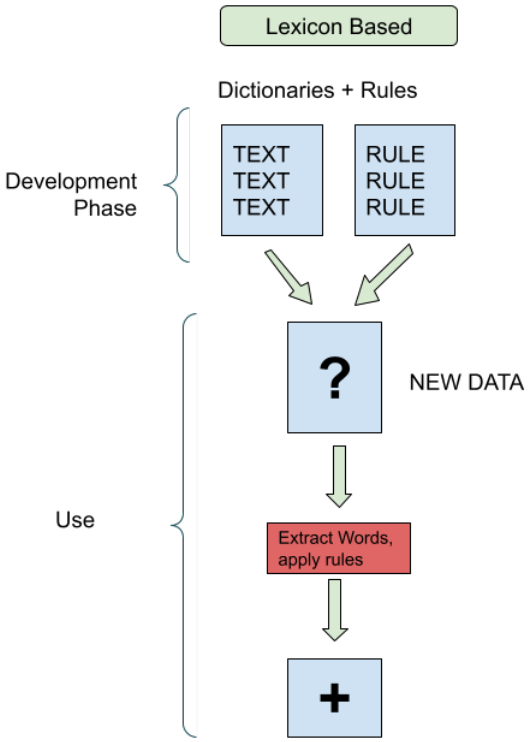


Figure 3.3: Lexicon Based Approach

Here is a simple illustration of how a rule-based system functions:

1. Establishes two lists of polarized words, one with negative words like awful, worst, and ugly and the other with positive phrases like good, best, and beautiful.

2. Determines the proportion of positive and negative terms in a given text.

3. The system returns a good attitude if there are more positive word appearances than negative word appearances, and vice versa. In the event if the numbers are equal, the algorithm will produce a neutral result.

3.3.2 Machine Learning Algorithms

Unlike rule-based systems, machine learning approaches rely more on machine learning techniques than on rules that have been manually created. In order to perform sentiment analysis, a text is often supplied to a classifier, which then outputs a category, such as positive, negative, or neutral.

Here's how a machine learning classifier can be implemented:

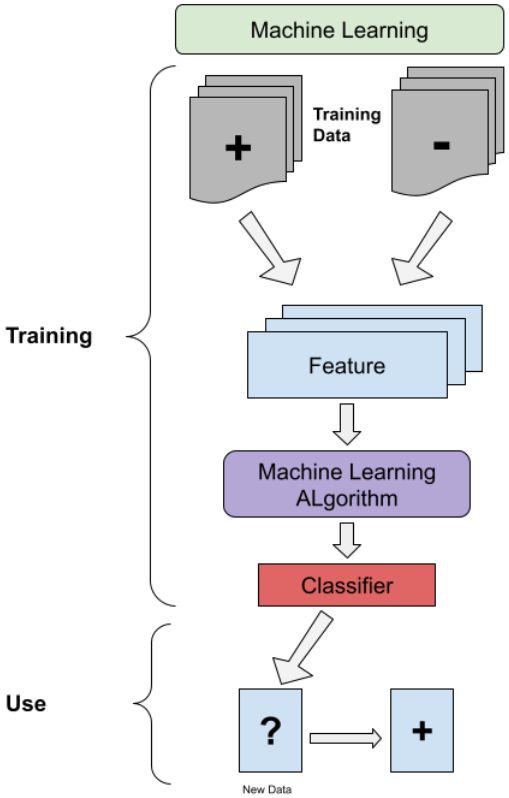


Figure 3.4: Machine Learning Algorithms

Feature Extraction From Text

The bag-of-words, bag-of-n-grams, or Word2Vec methods were traditionally used to vectorize or convert text as the initial step of a text classifier for machine learning.

Recently, novel features extraction techniques based on word embeddings have been employed (also known as word vectors). This kind of representation enables concepts with comparable meanings to have equivalent representations, which can increase the effectiveness of the classifier.

Classification Algorithms

A statistical model like Naive Bayes, Logistic Regression, Support Vector Machines, or Neural Networks are typically used in the classification step:

• **Naive Bayes**: a group of probabilistic algorithms that employs Bayes' Theorem to determine the genre of a document.

• **Linear Regression**: it is a fairly popular statistical approach for predicting a value (Y) given a set of features (X).

• **Random Forest**: A supervised learning algorithm that is used for both classification and regression is Random Forest. But it is, however, primarily used for problems with classification. As we know, a forest is made up of trees, and more trees make the forest stronger. It is an ensemble technique that is better than a single decision tree because by combining the result, it decreases the over-fitting.

• **Support vector Machine**: A non-probabilistic model called Support Vector Machines represents text instances as points in a multidimensional space. Regions in that space are mapped to examples of several types (sentiments). Then, a category is given to a new text depending on how closely it resembles an existing text and the regions it is mapped to.

• **Deep Learning**: a variety of methods that use synthetic neural networks to try to replicate the human brain.

3.3.3 Hybrid Approach

In this approach, we use both the Ruled Based Lexicon approaches and machine Learning Approaches at the same time. The tremendous advantage of these methods is that the effects are always more precise and more accurate.

Some research methods advise mixing lexicon-based and machine learning methods to increase the effectiveness of sentiment classification. We can take use of the best of both worlds with this hybrid approach, which is its main advantage. It is known that combining lexicons with machine learning improves accuracy. Therefore, it will improve the algorithm's recall and accuracy when combined with a classifier for machine learning. Some academic publications concentrate on sentiment analysis utilizing NLP and machine learning.

3.4 Importance of Sentiment Analysis

It is estimated that 90% of the world's data is unstructured, which means that it is unorganized. Every day, huge amounts of business data are generated: emails, help tickets, talks, interactions on social media, surveys, posts, documents, etc[20]. But it is difficult to analyze sentiment in a timely and efficient way.

By automating business processes, gaining actionable insights, and saving hours of manual data processing, sentiment analysis solutions enable businesses to make sense of this sea of unstructured text by increasing team productivity.Brands may discover what makes consumers happy or irritated by automatically evaluating customer feedback, such as comments in survey replies and social media conversations, in order to modify products and services to suit their requirements.

Some of the advantages of sentiment analysis include the following:

• **Scalability**:Imagine going through thousands of tweets, customer service discussions, or customer evaluations by hand. To manually process the data would just be impossible.Sentiment analysis makes it possible to analyse data at scale in a productive and economical manner.

• **Real Time Analysis**:In order to get situational awareness in real time during particular events, sentiment analysis can be utilized to find crucial information. Is there an impending PR catastrophe in social media? A client who is going to erupt in rage? These types of issues may be quickly recognized and addressed with the use of a sentiment analysis system.

• **Opinion Aggregation**:Another novel aspect to our work concerns the type of aggregation that can be applied to opinions to be extracted from various sources and co-referred. This can be applied to the derived information in classical information extraction in a straightforward manner: data can be combined if there are no contradictions, e.g. on the properties of an object.

• **Consistence criteria**:Humans don't have a set standard by which to judge a text's emotion. When assessing the sentiment for a given piece of writing, it is estimated that various people only agree 60–65% of the time. Businesses can use a centralized sentiment analysis system to analyze all of their data using the same standards. This enhances data consistency and reduces mistakes.

3.5 Sentiment Analysis Challenges

3.5.1 Relevance

Even if a crawler is limited to specific topics and correctly defines relevant pages, this does not mean that it would not be relevant to any comment on those pages.It's a particular problem with social media since discussions and comment threads may easily get off topic, unlike product evaluations, which don't often do that. For instance, we have seen remarks about a TV show that was seen just after the Rock am Ring event on the Rock am Ring forum. Similar to this, a user's interests on Twitter may span a wide range of topics, therefore it makes little sense to define "interesting" tweets for all users using a single lexical model. We may approach the relevance problem in different ways.

3.5.2 Negation

Identify the negation is called the hardest challenge in sentiment analysis.The disadvantage of the bag-of-words' simpler emotion classifiers is that they do not handle negation effectively; in a unigram model, the distinction between "good" and "not good" is largely missed despite the fact that they have wholly different meanings. The phrase "Surprisingly, the build quality is well above par, taking into account the rest of the features" is one example where longer range features, such as higher order n-grams or dependency structures, could help catch more comprehensive, subtle trends. However, the word "surprisingly" could partially negate the overall positive feeling.

3.5.3 Irony and Sarcasm

People use positive phrases to convey their bad feelings when utilizing irony and sarcasm, which makes it challenging for machines to grasp without having a complete understanding of the scenario in which a feeling was conveyed.

3.5.4 Multi-Language Issues

Sentiment analysis that is multilingual involves sentiment grading across multiple languages.Given that just 25.9% of internet users are native English speakers, sentiment analysis in non-English-speaking languages is becoming more and more crucial. Social

analytics and social listening are significantly impacted by the variety of languages and cultures on the internet. Because of this, sentiment analysis in English alone is insufficient. Multilingual sentiment analysis, which acknowledges that sentiment is innately influenced by language and culture, enables social analytics and social listening in several languages, guaranteeing that businesses are able to overcome language barriers and get insightful data in real-time.

3.5.5 Volatility Over Time

Twitter in particular, has a very high temporal dynamic. More specifically, attitudes can abruptly shift from positive to negative or the other way around.By using data extraction techniques similar to those mentioned in this paper and semantic annotation techniques similar to those in (Maynard and Greenwood, 2012) aimed at managing the evolution of entities over time, it is possible to address this problem by associating the various types of possible opinions with the classes representing entities, facts, and events as ontological properties.Instead than adding fresh information to an old perspective, the challenge for the agency in question is how to recognize emerging new opinions. In order to track trends over time, contradictions and changes must also be recorded and utilised, particularly through opinion merging, which is the topic of our next section.

3.6 Different levels on Sentiment Analysis

Many applications need to identify the tone of a text, which could be a sentence, paragraph, clause, or even just a single word in a document. In sentiment analysis, four main text granularity levels have been examined:

• **Document Level**:This level of analysis seeks to ascertain whether the overall tone of the document is good or negative. This level makes the assumption that each document only discusses one thing, like a movie review. Therefore, if the document discusses several entities, it does not apply.

• **Sentence Level**:This degree of analysis assesses if a sentence is positive, negative, or neutral.

• **Word Level**:This degree of analysis examines the word. It establishes if a word conveys a neutral, positive, or either. The term "polarities" is often reserved for sentiment tasks at higher levels. This level's task might be viewed as the creation of a sentiment lexicon.

3.7 Conclusion

Sentiment analysis' goal is to assess a writer's opinions and sentiments on a single issue or a range of related topics. This feeling may be expressed by his or her conclusions, assessments, or other emotional reviews. The difficulty of finding other people's opinions has grown along with the availability of opinion resources like social networks and review websites.It may be beneficial to combine information retrieval and natural language processing techniques in sentiment analysis.

Chapter 4

Recurrent Neural Network and LSTM

4.1 Introduction

This chapter presents the terms we have used during this research. We have discussed Recurrent Neural Network(known as RNN) and Long Short Term Memory(known as LSTM).

4.2 Neural Network

A neural network is a collection of algorithms that aims to identify underlying links in a set of data using a method that imitates how the human brain functions. Neural networks are able to adjust to changing input, they can produce the optimal results without having to change the output criterion. As trading systems are developed, the artificial intelligence-based notion of neural networks is quickly gaining favor.

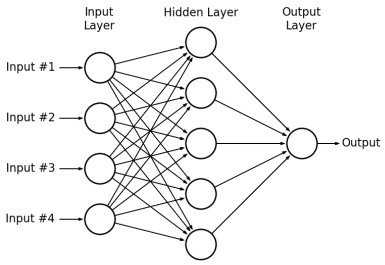


Figure 4.1: Neural Network

4.3 Recurrent Neural Network

Traditional neural networks are feed-forward, which go from the input layer through the hidden layer to the output layer.All the inputs and outputs are independent of each other, but in cases like when it is required to predict the next word of a sentence, the previous words are required and hence there is a need to remember the previous words. RNN can handle sequential data, it can memorize the previous inputs, The result at any given time is retrieved back to the network to be improved. The main and most important feature of RNN is Hidden state, which remembers some information about a sequence.

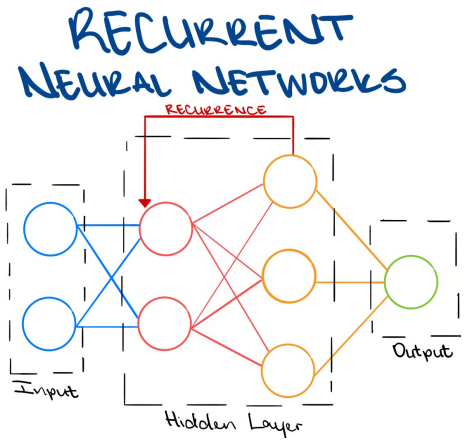


Figure 4.2: Recurrent Neural Network

4.4 LSTM

LSTM can be viewed as a modification of Recurrent Neural Networks (RNNs).When the impacts of earlier states are still being felt in the current state, long short-term memory, or LSTM, is frequently used. A cell state mechanism controls the flow of information in the LSTM. Cell state can also be referred to as memory. This system allowed the LSTM to recall or forget a piece of data.

Down the entire sequence chain, the cell state transmits relative information. A device known as a gate can add or subtract information. The input gate unit guards against erroneous information affecting the cell state or memory contents. An output gate unit shields other units from disruption by currently irrelevant memory contents. Figure 3.3 displays the physical makeup of an LSTM cell.[21]

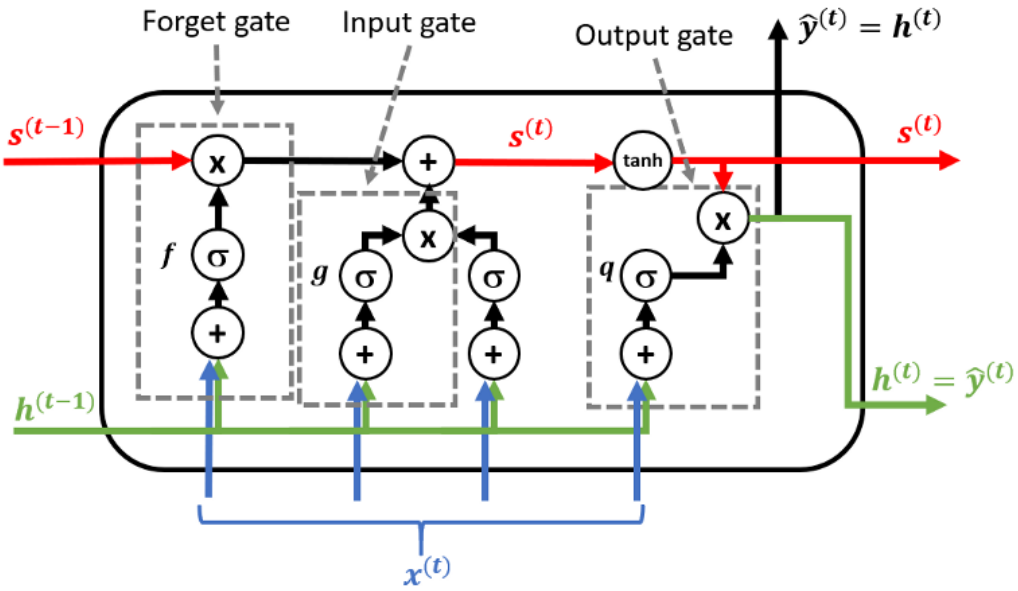


Figure 4.3: Structure of a LSTM cell

Forge gate and input gate create a input gate unit where forge gate is responsible for removing memory from the LSTM.Forge gate have the authority to through away or store the data.It makes use of the input for the current cell as well as the hidden state from the previous cell as inputs.

Forge Gate

The input x and prior hidden state h are used to build the forget gate, which determines what data from the previous cell state is forgotten (t-1).

$$f_i^{(t)} = \sigma\left(b_i^f + \sum_j U_{i,j}^f x_j^{(t)} + \sum_j W_{i,j}^f h_j^{(t-1)}\right) \tag{4.1}$$

Input Gate

The external input gate g(t) and the activated combination of the input vector and previous concealed state, each having their own parameters, are added to the product of the forget gate and previous cell state. The input of the cell and the concealed state of the previous cell again determine how the external input gate operates.[22]

$$s_i^{(t)} = f_i^{(t)} s_i^{(t-1)} + g_i^{(t)} \sigma\left(b_i + \sum_j U_{i,j} x_j^{(t)} + \sum_j W_{i,j} h_j^{(t-1)}\right) \tag{4.2}$$

$$g_i^{(t)} = \sigma\left(b_i^g + \sum_j U_{i,j}^g x_j^{(t)} + \sum_j W_{i,j}^g h_j^{(t-1)}\right) \tag{4.3}$$

Output Gate

The output gate q(t) and the cell state together make up the cell output h(t). In some notations, the terms "output vector y(t)" and "output hidden state h(t)" will both refer to the same thing in an LSTM.[23]

$$q_i^{(t)} = \sigma\left(b_i^o + \sum_j U_{i,j}^o x_j^{(t)} + \sum_j W_{i,j}^o h_j^{(t-1)}\right) \tag{4.4}$$

$$h_i^{(t)} = \tanh\left(s_i^{(t)}\right) q_i^{(t)} \tag{4.5}$$

4.5 Conclusion

Neural Network models are becoming more famous in recent years. researchers exploring Artificial Neural network model(ANN), Convolution Neural network model(CNN) and recurrent Neural Network model(RNN) with Natural Language Processing(NLP) in various research field to make their model tore efficient and precise one.

Chapter 5

Methodology

5.1 Introduction

This chapter is all about work flow and implementation of the applied methodology.The basic work flow is explained in section 5.2. Section 5.3 and section 5.4 describes the data cleaning process and scaling the data into the desired range. Feature selection approaches were briefly discussed in section 5.5. Different classification models were introduced in section 5.6. Then, section 5.7 covers various evaluation matrix to measure the accuracy and error of the models.Finally, in section 5.8, the overview of the whole chapter is given.

5.2 Workflow of the Research Work

To extract polarity from social media tweet, blog, comment etc. we have applied a framework step by step.Firstly, we collect raw data from our desired data sets. these data are noisy and unsuitable to process. so we have to apply data preprocessing methods to clean the data.After data cleaning, we convert it into vector which is numeric manner, because we can't process textual data for sentiment analysis.Then, scaling should be done between +1.0 to -1.0.

Finally, The classification model is applied to the processed vector data.Machine Learning classifier such as Support Vector Classifier, Ada boast , Random forest,K-Nearest Neighbour(KNN) or, Neural Network Classifier such as Bi-LSTM-1, CNN etc. applied to the trained data to classify and gives the desired outcome.

The systematic flow of our applied approach is depicted in the following figure 5.1.

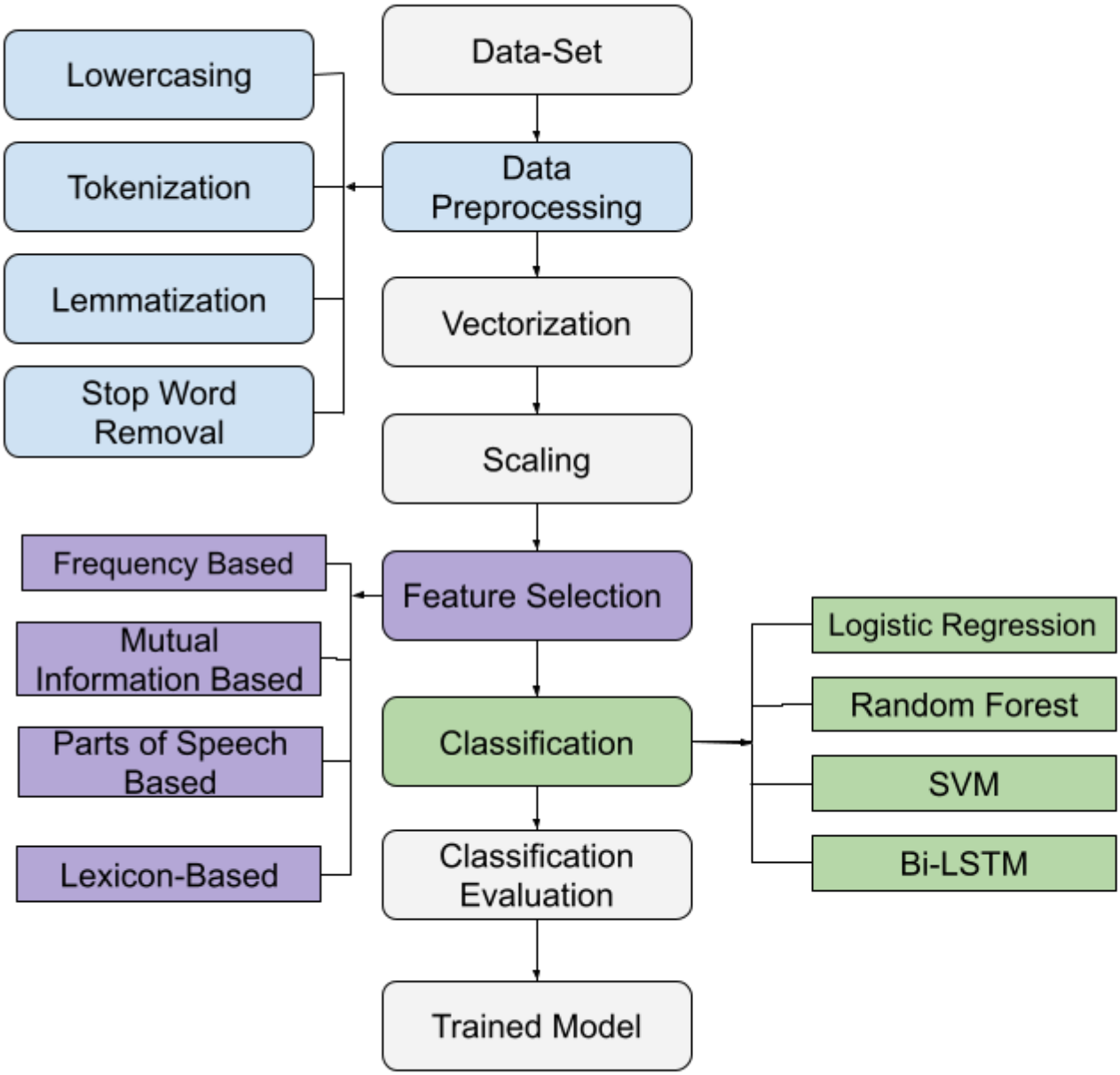


Figure 5.1: Applied Methodology

5.3 Data Cleaning

Real world data is usually noisy, erroneous , inconsistent and incomplete.To extract useful knowledge and information, elimination of noise and insignificant is very important.We have applied different cleaning technique on the datasets. This steps reduce the size of the data set and help to handle the easily.

5.3.1 Lowercasing

One of the most straightforward and efficient methods of text preprocessing is lowercasing. For instance, even though "Book" and "book" have the same meaning, the vector space model counts them as two different words. As a result, the dimension is decreased because these two words are now deemed to be one after lowercasing. For certain natural language processing (NLP) procedures like frequency, term frequency-inverse document frequency (tf-idf), etc., lowercasing is more efficient. For subsequent processing, this phase minimizes duplications and ensures accurate counting.

5.3.2 Tokenization

For text analytics, tokenization is a crucial preprocessing step. It is the process of dividing a text into tokens, which may be phrases, paragraphs, or single words. First, sentences from the review text data are separated using sent tokenize() in nltk (Natural Language Toolkit). Using the word tokenize() function from the nlik package, each tokenized sentence was then once more broken down into its component words. The resulting word list is utilized for sophisticated processing.

5.3.3 Lemmatization

The process of lemmatization involves morphologically analyzing words. The idea behind this method is to identify a word's fundamental root from a variety of word forms. For instance, the verbs "travel," "travels," and "traveling" are essentially different verb tenses. These word forms are condensed during the lemmatization process, which identifies the shared root as "journey". In this case, lemmatizing is done using WordVetLemmatizer() from the nltk package. This method aids in noise reduction and gets the text ready for more examination.

5.3.4 Stop Word Removal

Stopwords are frequently used words that offer no helpful information in any language. Stopwords include words like ”the,” ”a,” ”is,” ”are,” etc. in English. It is feasible to lower the bulk of the data and increase the attention on vital terms by deleting these unnecessary words. The nltk program can be used to remove stopwords in a variety of languages.

Sample Text with stop Words	**Without stop Words**
I like reading, So i read.	Like, reading, read
You look okay to me.	look, okay
The job should be done.	job, done
So what are you waiting for?	waiting

Table 5.1: Stop Word Removal

5.4 Vectorization and Scaling

Inputs for machine learning algorithms must be numerical feature vectors in order for the program to comprehend.In machine learning, the aim of scaling and standardization is to transfer data range to [0, 1].

```
w2v.wv.most_similar('thank')
```

```
[('much', 0.9658550024032593),
 ('thanks', 0.9616288542747498),
 ('appreciate', 0.9582455158233643),
 ('awesome', 0.9505729675292969),
 ('amazing', 0.9487380981445312),
 ('response', 0.9363716840744019),
 ('quick', 0.9362509250640869),
 ('love', 0.9337433576583862),
 ('twitter', 0.9307826161384583),
 ('care', 0.9266414642333984)]
```

```
w2v.wv.most_similar('bad')
```

```
[('lack', 0.9135125875473022),
 ('disappointed', 0.9110564589500427),
 ('worse', 0.9047411680221558),
 ('poor', 0.9026240110397339),
 ('awful', 0.896759033203125),
 ('staff', 0.8953446745872498),
 ('excellent', 0.8864026665687561),
 ('communication', 0.8859975934028625),
 ('horrible', 0.8827396631240845),
 ('seen', 0.8804244995117188)]
```

Figure 5.2: Vectorization and Scaling

5.4.1 Vectorization

Machine learning methods require inputs to be numerical feature vectors in order for the program to function. Thus, while working with text documents, it must be converted into a numeric vector. This procedure is known as text vectorization. In this thesis, the Word2vec model was chosen and employed for vectorization. A collection of related models called the Word2vec model are used to create embedded terms. There are two patterns in Word2vec, CBOW, and skip gram algorithms. Generally speaking, the models are shallow, two-layer neural networks that can reconstruct words in linguistic situations.

5.4.2 Scaling and Standardization

Scaling and standardization in machine learning are used to reduce the data range to [0, 1, 2]

The purpose of this method and its benefits are as follows:

(1) it expedites the convergence of gradient descent;

(2) it avoids the problem of different feature dimensions and makes the features similar.

5.5 Feature Selection

5.5.1 Frequency-Based Selection

It is common practice in text modeling to exclude words that are infrequently used in the corpus. These are likely misspellings, which make it difficult to classify them in a broad way.However, it has been discovered that terms that appear just once within a particular corpus are highly accurate predictors of subjectivity [24]. Due to the potential significance of rare terms in classification, we explore several cutoffs using frequency counts.

5.5.2 Mutual Information Based Selection

Eliminating some of the less valuable attributes may help boost the classifier's performance. Expected Mutual Information is one of the often used feature selection metrics [46].The anticipated MI often scores the traits, with the top few being considered the most helpful for categorization. We likewise follow a similar strategy.

5.5.3 Part of Speech(POS) Based Selection

Certain POS have been found to be more beneficial in categorization tasks, especially for SA. Show, for instance, that utilizing adjectives and adverbs is more effective than using only adjectives. Use verbs to classify emotions as well. Limiting the feature space to to these adjectives, provided adjectives are actually crucial variables in predicting sentiment polarity [25], may enhance classifier performance by omitting less helpful words. By keeping just words that are adjectives, verbs, and nouns separately and together, we put this idea to the test.

To achieve annotated part-of-speech in the approach used, POS tagger class available in the NLTK package has been used to develop algorithm to obtain word sense for only English language tags from the sentence. It analyzes the lowest level of syntactic structure of the sentence and tags them with their related part-of-speech, which categorizes the word lexically with its POS label and gloss together for further classification.

5.5.4 Lexicon-Based Selection

For feature selection, sentiment-annotated lexicons may be utilized similarly. Less useful characteristics may be left out of the feature list by choosing phrases that express strong opinion.

Popular lexicons are the English language's equivalent of WordNet, a sizable lexical database (http://wordnet.princeton.edu/). SentiWordNet, for example, contains polarity and objectivity labels for the terms [21]. Utilize WordNet synsets, which are collections of words, in WordNet-Affect [80] to label31 each synset with emotional labels. Both lexicons are used in our analysis, and both have been widely used in the community. Additionally, we make use of Affect Control Theory vocabulary.

5.6 Classification Model

Classification in machine learning algorithms involves determining which class a new observation falls into using training data. supervised classification and unsupervised classification are the two basic types of classification methods.

Clustering is the more common name for unsupervised classification. Below are some examples of well-known classification methods that were effective in this investigation.

5.6.1 Logistic Regression

When a dependent variable is dichotomous, the proper regression analysis to use is logistic regression (binary). The logistic regression is a predictive analysis, just like all regression analyses. To describe data and explain the relationship between one dependent binary variable and one or more independent nominal, ordinal, interval, or ratio-level variables, we employ logistic regression. [27]

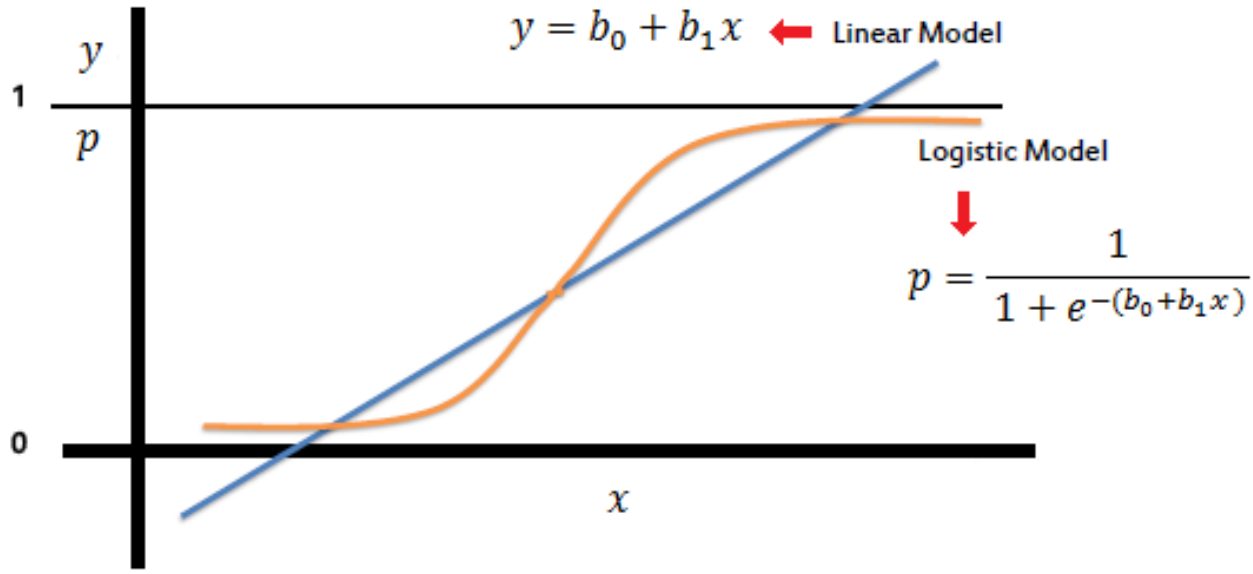


Figure 5.3: Step function curve obtained from Logistic Regression[26]

Types of Logistic Regression:

1. Binary Logistic Regression

The categorical response has only two 2 possible outcomes. Example: Spam or Not.

2. Multi-nomial Logistic Regression

Three or more categories without ordering. Example: Predicting which food is preferred more (Veg, Non-Veg, Vegan).

3. Ordinal Logistic regression

Three or more categories with ordering. Example: Movie rating from 1 to 5.

5.6.2 Random Forest

Random Forest is a supervised learning technique that is applied to both classification and regression. However, categorization issues are its main application. A forest is made up of trees, as we all know, and a forest gets stronger as more trees are added. Similar to this, the random forest method builds decision trees from data samples, obtains predictions from each one, and then votes on the best answer. It is an ensemble strategy that is superior to a single decision tree since it reduces over-fitting by mixing the results.

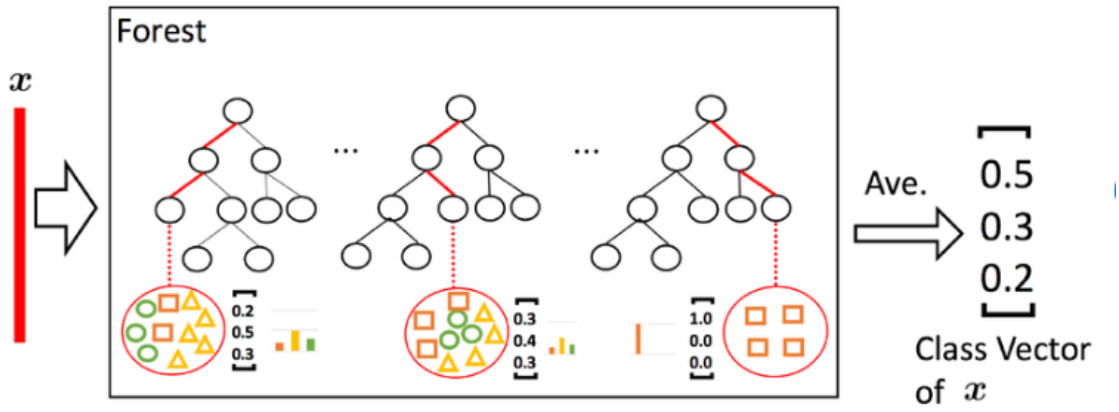


Figure 5.4: Random forest Classifier[27]

Random Forest Algorithm in Action

The steps below can help us understand how the Random Forest algorithm functions:

Step 1:

Choose random samples at random from the provided dataset.

Step 2:

The second step of this technique is to create a decision tree for each sample. The forecast outcome from each decision tree will then be obtained.

Step 3:

Voting will be conducted for each predicated outcome.

Step 4:

Finally, choose the prediction result that received the most votes as the final prediction result.

5.6.3 Support Vector Machine (SVM)

Although it can be used for regression, the Support Vector Machine is a supervised learning technique that is primarily used for classification. The key concept is that the algorithm searches for the best hyperplane that may be used to categorize new data points based on the labeled data (training data). The hyperplane is a straightforward line in two dimensions.

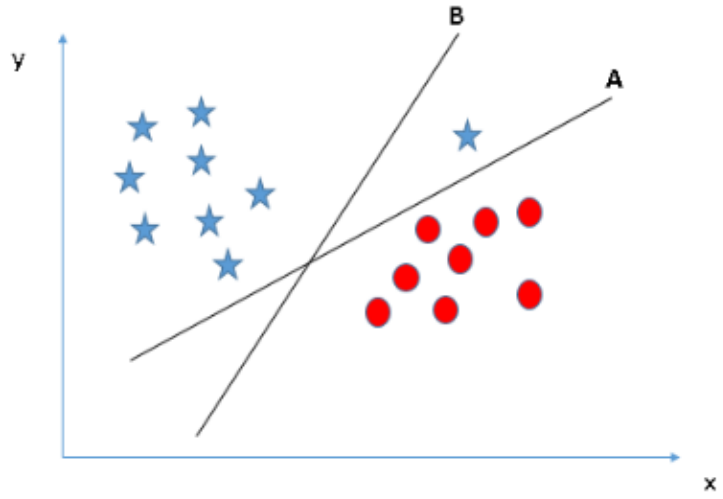


Figure 5.5: Support Vector Machine (SVM)[28]

Given that hyper-plane B has a bigger margin than hyper-plane A, some of you may have chosen it. The catch is that SVM chooses the hyper-plane that accurately classifies the classes before maximizing margin. In this case, hyper-plane B made a classification mistake, but A classified everything correctly. A is, thus, the proper hyper-plane.[28]
Pros and Cons associated with SVM:-

PROS:

1.With a distinct margin of separation, it functions incredibly effectively.

2.In situations where there are more dimensions than samples, it is effective.

CONS:

1.With larger data-sets the performance drops, as it takes time to train the data.

2.With noisy data, overlapping occurs.

5.6.4 Naive Bayes

The Bayes theorem forms the basis of it. It is most frequently used in text classification since it has a large training dataset. The naive bayes classifier is a straightforward and efficient classification technique that helps develop quick machine learning models that can make precise predictions. It is a probabilistic classifier, which means that it bases its predictions on the likelihood that an object will occur. Spam filtration, sentiment analysis, and article classification are examples of naive bayes algorithm implementations. The equation is:

$$P(\mathbf{A}|\mathbf{B}) = P(\mathbf{A})\frac{P(\mathbf{B}|\mathbf{A})}{P(\mathbf{B})} \tag{5.1}$$

Where,

P(**A** I **B**) is posterior probability:Probability of hypothesis A on the observed event B.

P(**B** I **A**) is likelihood probability:Probability of the evidence given that the probability of a hypothesis is true.

P(A) is prior probability:Probability of hypothesis before observing the evidence.

P(B) is marginal probability:Probability of evidence.

Types of Naive Bayes:

1.GAUSSIAN:

The Gaussian model presumes a regularly distributed distribution of features. The model presumptively assumes that the values come from a Gaussian distribution if the predictors are continuous rather than discrete.

2. MULTINOMIAL:

In cases where the data is multinomially distributed, the multinomial Nave Bayes classifier is employed. It is frequently used to handle classification issues with documents, such as determining whether a document belongs in the sports, politics, or education categories. The classifier's predictions are reliant on the frequency of terms.

3. BERNOULLI:

The Bernoulli classifier functions similarly to the multinomial classifier; the only distinction is that the predictor variables are all independent boolean variables instead of the multinomial classifier's boolean variables.

5.7 Evaluation Matrix

Various measures have been applied to gauge how well emotion classification performed. This covers metrics like Confusion matrix, precision-recall curves, F1 scores, accuracy-error curves, and recall metrics. We will briefly cover various metrics in this section.

5.7.1 Confusion Matrix

A performance indicator for classification issues involving machine learning is the confusion matrix or error matrix. The matrix's rows and columns each stand for a predicted class and an actual class, respectively. The confusion matrix can be shown using the representation in table 3.1.

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive	False Negative
Actual Negative	False Positive	True Negative

Table 5.2: Confusion Matrix

5.7.2 Accuracy and Error Rate

One parameter for assessing classification models is accuracy.Accuracy can be defined as the number of correctly predicted data points out of all predicted data points[29].So,

$$\textbf{Accuracy} = \frac{\textbf{Numbers of Correct Predictions}}{\textbf{Total Numbers of Predictions}} \tag{5.2}$$

Accuracy can also be determined in terms of positives and negatives for binary classification, as seen below:

$$\textbf{Accuracy} = \frac{\textbf{TP + TN}}{\textbf{TP + TN + FP + FN}} \tag{5.3}$$

Where,

TP = True Positives,

TN = True Negatives,

FP = False Positives, and

FN = False Negatives.

As we have defined accuracy, error rate can be defined as:

$$\textbf{Error Rate} = \textbf{1 - Accuracy} \tag{5.4}$$

In terms of confusion matrix,

$$\textbf{Error Rate} = \frac{\textbf{FP + FN}}{\textbf{TP + TN + FP + FN}} \tag{5.5}$$

Therefore, the percentage of accurately anticipated data points out of all expected data points is known as the error rate.

Accuracy or error rate is the most straightforward and understandable indicator for assessing classifiers. Accuracy, however, is unable to differentiate between the kinds of errors (False Positives or False Negatives) that a classifier commits. This is allowed, though, if the dataset has an equal number of cases for each class. Even though the accuracy is very high, it is difficult to conceive a situation when an application wouldn't need to differentiate between the different types of errors. Even disregarding a fact occasionally can have disastrous effects, as in the case of a medical classification issue. Precision and Recall can help solve this issue.

5.7.3 Recall and Precision

Recall, also referred to as Sensitivity, measures the amount to which all of the data points that required to be identified as positive really were. This refers to the proportion of positive samples that have accurate labels.Recall is the ratio of True Positives and total True Positives (TP) False Negatives (FN). [30].

$$\textbf{Recall} = \frac{\textbf{TP}}{\textbf{TP + FN}} \tag{5.6}$$

Recall seeks to catalog every instance of a True Positive. Recall score can be very important in situations like the medical industry, where discovering the real positive is of utmost importance.

Precision, also know as Confidence assesses to what extent the algorithm or classifier is correct.[31] This refers to how many samples that are labeled as positive are in fact positive and how many cases that are labeled as negative are in fact negative. The ratio of True Positives to all True Positives (TP) and False Positives is known as precision (FP).

$$\textbf{Precision} = \frac{\textbf{TP}}{\textbf{TP + FP}} \tag{5.7}$$

Precision also can be called True Positive Accuracy.[32]

5.7.4 F1 measure

The harmonic mean of Precision and Recall is computed using the F1 score. The metrics' weighted average is what it is.Thus,

$$\textbf{F1} = \textbf{2 x } \frac{\textbf{Precision x Recall}}{\textbf{Precision + Recall}} \tag{5.8}$$

False positives and false negatives are also considered in the F1 score. A large number of True Negatives or an uneven class distribution can both significantly contribute to accuracy. If we want to strike a compromise between Precision and Recall as well as an uneven class distribution, F1 is typically more helpful than accuracy.

5.8 Conclusion

Machine learning-based tools and approaches for sentiment analysis and categorization are utilized in a number of studies and research projects. We describe well-known algorithms and strategies that are frequently employed in sentiment analysis. A comparison of accuracy across several datasets is provided that can be utilized as a quick reference in further research projects.

Chapter 6

Implementation and Result

6.1 Introduction

This chapter presents the experimental results of sentiment classification for various levels.The dataset and sample input output have been explained in section 6.2 and section 6.3. Exploratory Data Analysis,section 6.4, has assessed graphical evidence which demonstrate the effectiveness of our strategy .In section 6.5 and section 6.6, the reviews, text, summaries, and other factors were taken to calculate accuracy, recall, precision for sentiment categorization.Section 6.7 illustrates different evaluation matrix for each classification, and the outcomes were compared using a machine learning and neural network approach.Section 6.8 shows the comparison chart between the existing model and our applied model, while section 6.9 provides an overview of the entire chapter.

6.2 The Dataset

DATASET: Twitter-airline-sentiment.csv

The dataset was gathered from an open source. It has 14640 text documents in it. The two columns in the dataset are "text" and "Sentiment." The content is in the "text" column, while the "sentiment" column designates whether the document is positive, neutral or negative. Each document has a sentiment value on it. Each document has a label1 (negative) , label2(neutral) or label3(positive) designation. Two data sets, Training and Test, were created from the dataset. 70% of the dataset consists of training data. The remaining 30% are from test data.

airline_sentiment		negativereason		airline		text
negative neutral Other (2363)	63% 21% 16%	[null] Customer Service ... Other (6268)	37% 20% 43%	United US Airways Other (7905)	26% 20% 54%	**14427** unique values
neutral				Virgin America		@VirginAmerica What @dhepburn said.
positive				Virgin America		@VirginAmerica plus you've added commercials to the experience... tacky.
neutral				Virgin America		@VirginAmerica I didn't today... Must mean I need to take another trip!
negative		Bad Flight		Virgin America		@VirginAmerica it's really aggressive to blast obnoxious "entertainment" in your guests' faces &...
negative		Can't Tell		Virgin America		@VirginAmerica and it's a really big bad thing about it
negative		Can't Tell		Virgin America		@VirginAmerica seriously would pay

Figure 6.1: twitter-airline-sentiment-data-set

6.3 Input and Output

6.3.1 Review Input and Output types

The output can have different level, 2-level, 3-level or 5-levels. In this experiment, the 3-level sentiment analyses have been used to evaluate the applicable approach. Three different forms of review data are regarded as input for the data-set. the primary analysis of the text data, the text data summary, and the attached text and summary result.

Input	**Text**	*when can I book my flight to Hawaii??, everything was fine until you lost my bag, you rock, very satisfying!!*
	Summary	*Wonderful, satisfy.*
	Text + summary	*Wonderful, when can I book my flight to Hawaii??, everything was fine until you lost my bag, you rock, very satisfying!!*
Output	**3-level**	*Positive, negative, neutral*
	2-level	*Positive, negative.*

Figure 6.2: Review Input and Output types

6.3.2 Sample Input and output

As described above, based on polarity, the review sentences are classified into 3 different categories. Sample inputs and output of our approach are given in table 4.3.

3-level sentiment analysis:

Review Sentences	**Polarity**	**Sentiment**
everything was fine until you lost my bag	0	Negative
when can I book my flight to Hawaii??	1	Neutral
You Guys are awesome.	2	Positive

Table 6.1: Confusion Matrix

6.4 Exploratory Data Analysis(EDA)

Exploratory Data Analysis (EDA) is a method for data analysis that uses a range of (mainly graphical) tools to extract key variables, identify outliers, build parsimonious models, and discover the best factor settings while maximizing insight into a data set[33].

RATIO OF NEGATIVE,POSITIVE AND NEUTRAL:

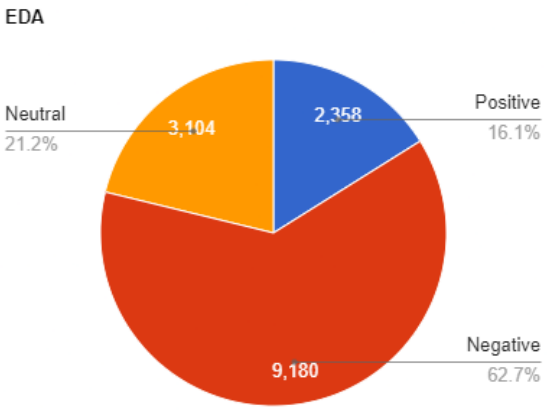


Figure 6.3: Portion of Sentiment

KERNEL DISTRIBUTION OF NUMBER OF WORDS:

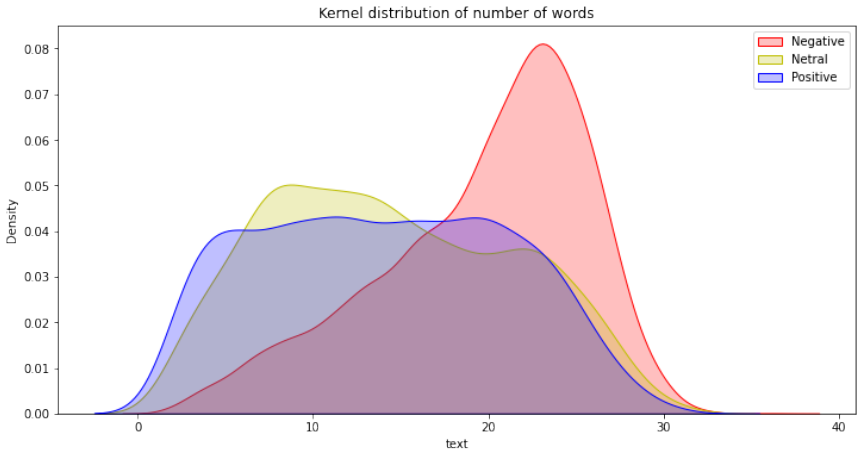


Figure 6.4: Kernel distribution of number of words

The result shows that the majority of tweets (63%) are negative, followed by 21% of neutral tweets and lastly (16%) of positive tweets.

DISTRIBUTION OF TWEETS:

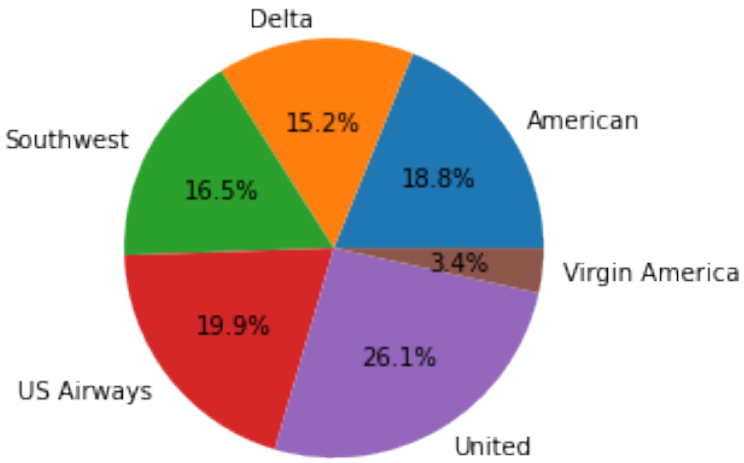


Figure 6.5: distribution of tweets

The percentage of tweets that were made public for each airline is displayed in the output. The most tweets, or 26%, belong to United Airlines, followed by US Airways (20%).

PROPORTION OF SENTIMENT:

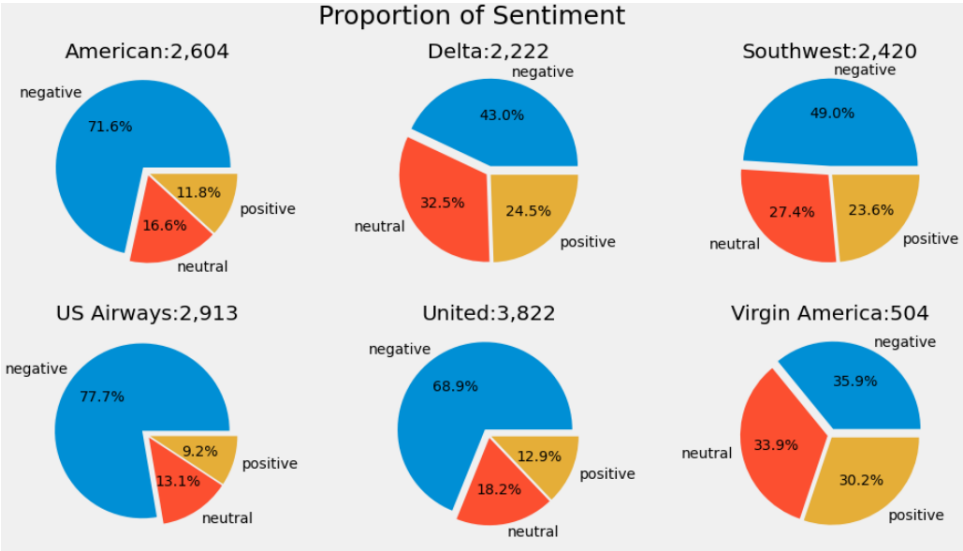


Figure 6.6: Proportion of Sentiment

American,US Airways: The sentiment for these two airlines shows negative in general.

United: 68.9% of people felt it negative.

Delta, Southwest: Better than other as negative sentiment is less than 50%.

Virgin America: The proportion of sentiments is very well-balanced.

WORD CLOUD - Keyword Analysis:

Word Cloud is one of the easiest way to show which word mainly(frequently) appears in the set of sentences.

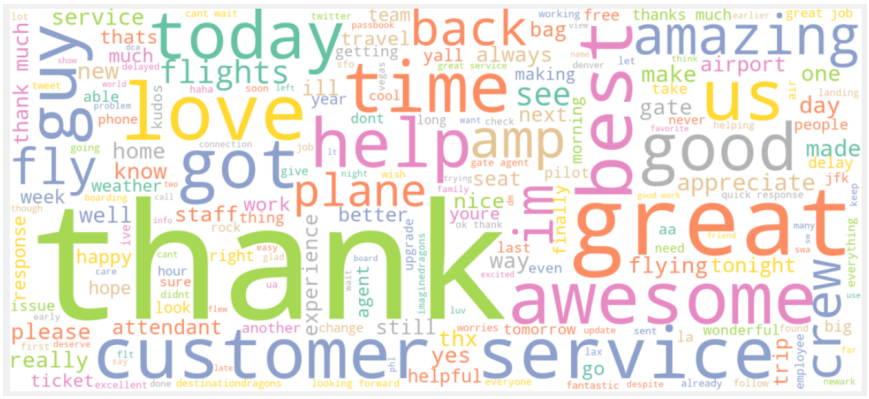


Figure 6.7: Positive Word Cloud

The main words we can see in the Word Cloud of positive sentiment is 'thank', 'great', 'awesome', 'great', 'love' etc.

Figure 6.8: Negative Word Cloud

As, we have seen before, most of our data has negative sentiment. The main words we can see in the Word Cloud of negative sentiment is 'plane', 'bag', 'time', 'customer service', etc. Plane refers the delay of planes, bag refers the lost of bags during travelling, time for late flight, customer service refers poor customer service mainly.

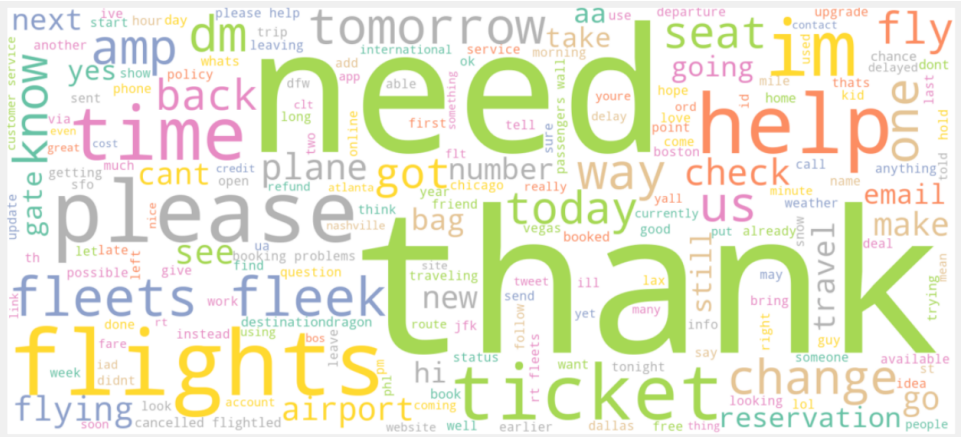


Figure 6.9: Neutral Word Cloud

Neutral is just neutral. The only positive word we can see at a glance is 'thank'. Almost no negative and positive words.

SENTIMENTS OF EACH AIRLINE:

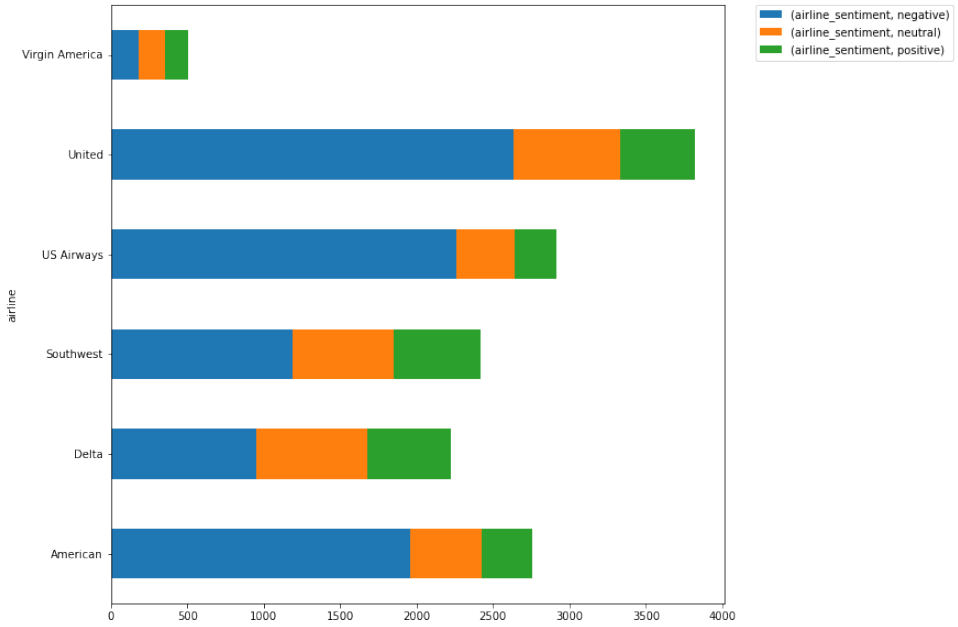


Figure 6.10: sentiments of each airline

This diagram demonstrates that positive, negative and neutral sentiment of each airline. Here,The most negative sentiment belongs to United Airlines, followed by US Airways.

PLOTTING ALL THE NEGATIVE REASONS:

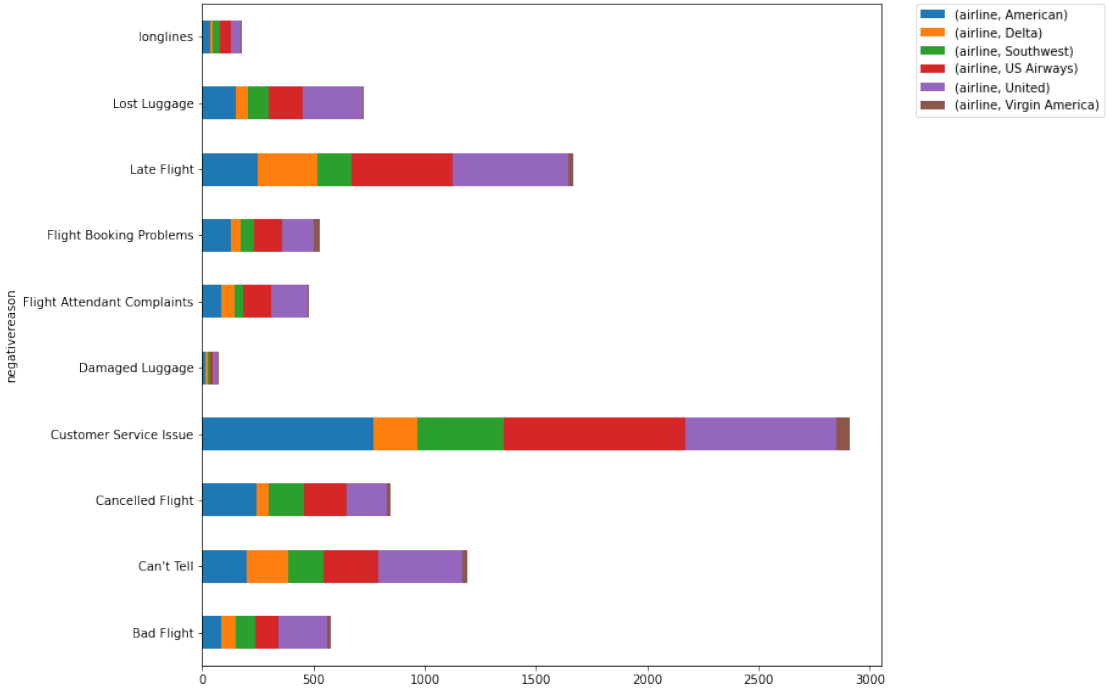


Figure 6.11: distribution of sentiments

This diagram demonstrates that poor customer service and delayed flights are the leading causes of problems. And occasionally, individuals give something a bad rating without explaining why.

6.5 Data Preprocessing

A supervised ML algorithm's performance is frequently significantly impacted by the preprocessing of the data[34]. Following are the actions that are taken during data preprocessing:

a) Case Conversion

b) Stop-words Removal

c) Stemming & Lemmatization

d) Spelling Correction

PREPARE TRAIN AND TEST DATA

Two separate data sets—Training data set and Test data set—will be created. The model will be fitted using the training data set, and the test data set will be utilized to test the predictions.

Test data	Training Data	Total data
10,248	4,392	14,640
70%	30%	100%

Table 6.2: Amount of Train and Test data

REPRESENTING TEXT IN NUMERIC FORM

Our Machine Learning and deep Learning Model work with numeric data, so first we have to convert the textual data in numeric manner.To do so, we have used-

- **TF-IDF Transformer**
- **CountVectorizer**
- **Word2Vec**.

6.6 Classification Model for LSTM

After training the models with training data, prediction has been made evaluating the classification model. The structure of the Applied model is given below:

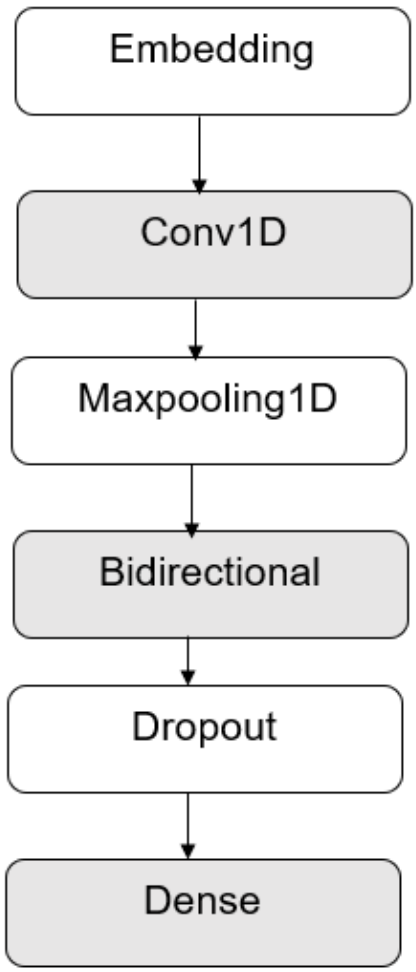


Figure 6.12: Applied Model Layer for LSTM

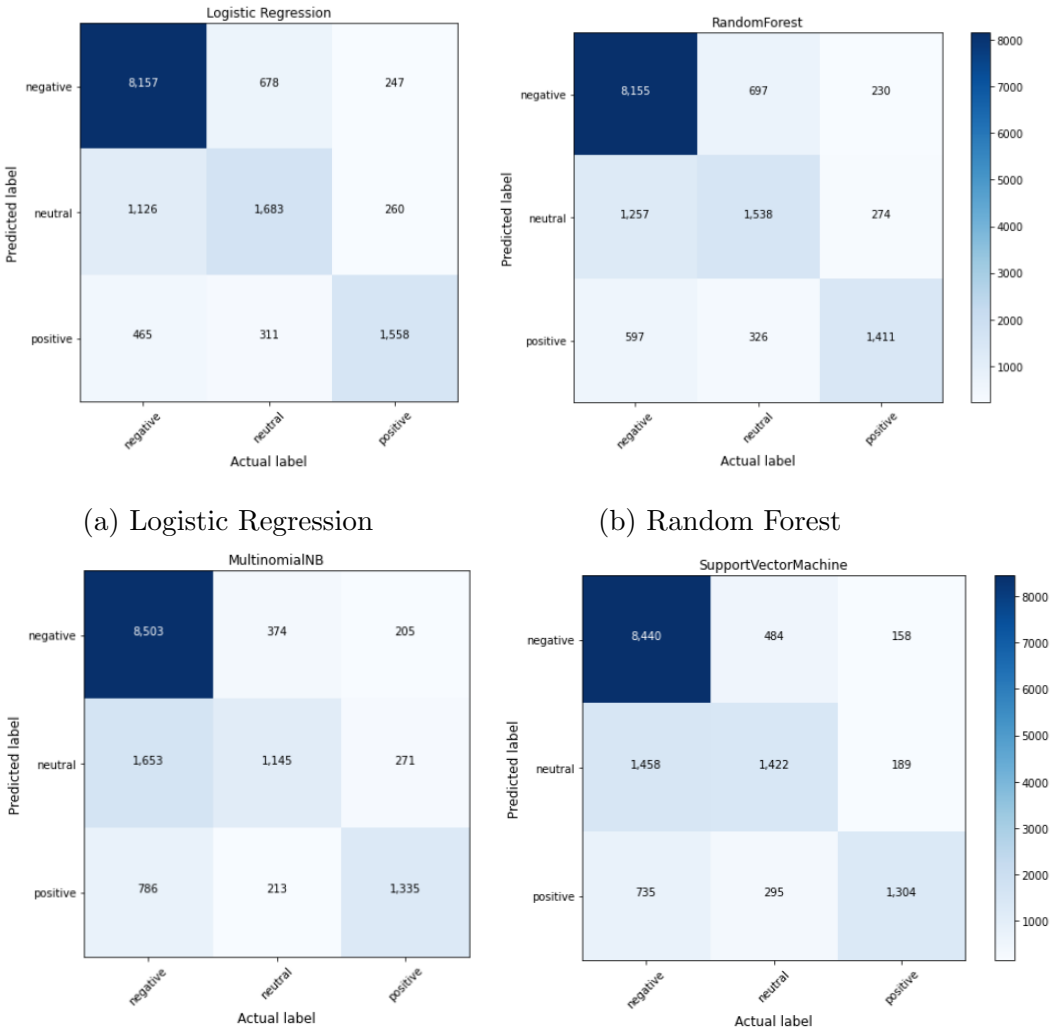
6.7 Results

Our data sets are now prepared for feeding into several classification algorithms. The test dataset will be used to apply the MutinomialNB, Support Vector Machine, Bi-LSTM, and Logistic Regression classifier models, and results will be shown for each classifier in terms of accuracy, precision, recall.

6.7.1 Using CountVectorizer:

Model	Accuracy	Precision	Recall
Logistic Regression	**0.786**	0.740	**0.704**
Random Forest	0.766	0.717	0.667
MultinomialNB	0.758	0.725	0.627
Support Vector Machine	0.770	**0.743**	0.650

Figure 6.13: Evaluation Measurement(CountVectorizer)



(a) Logistic Regression

(b) Random Forest

(c) MultinomialNB

(d) Support Vector Machine

Figure 6.14: Confusion Matrix(CountVectorizer)

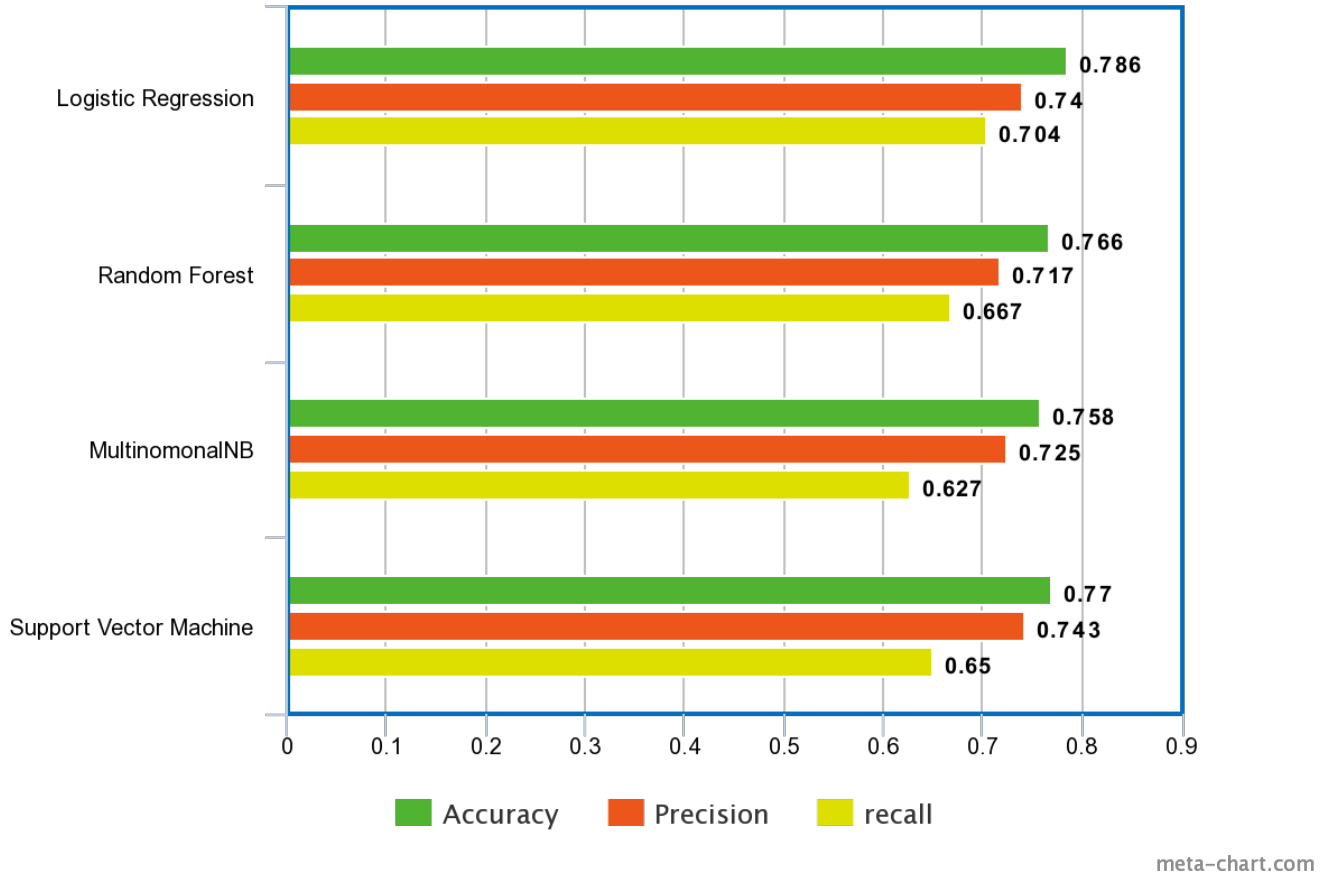


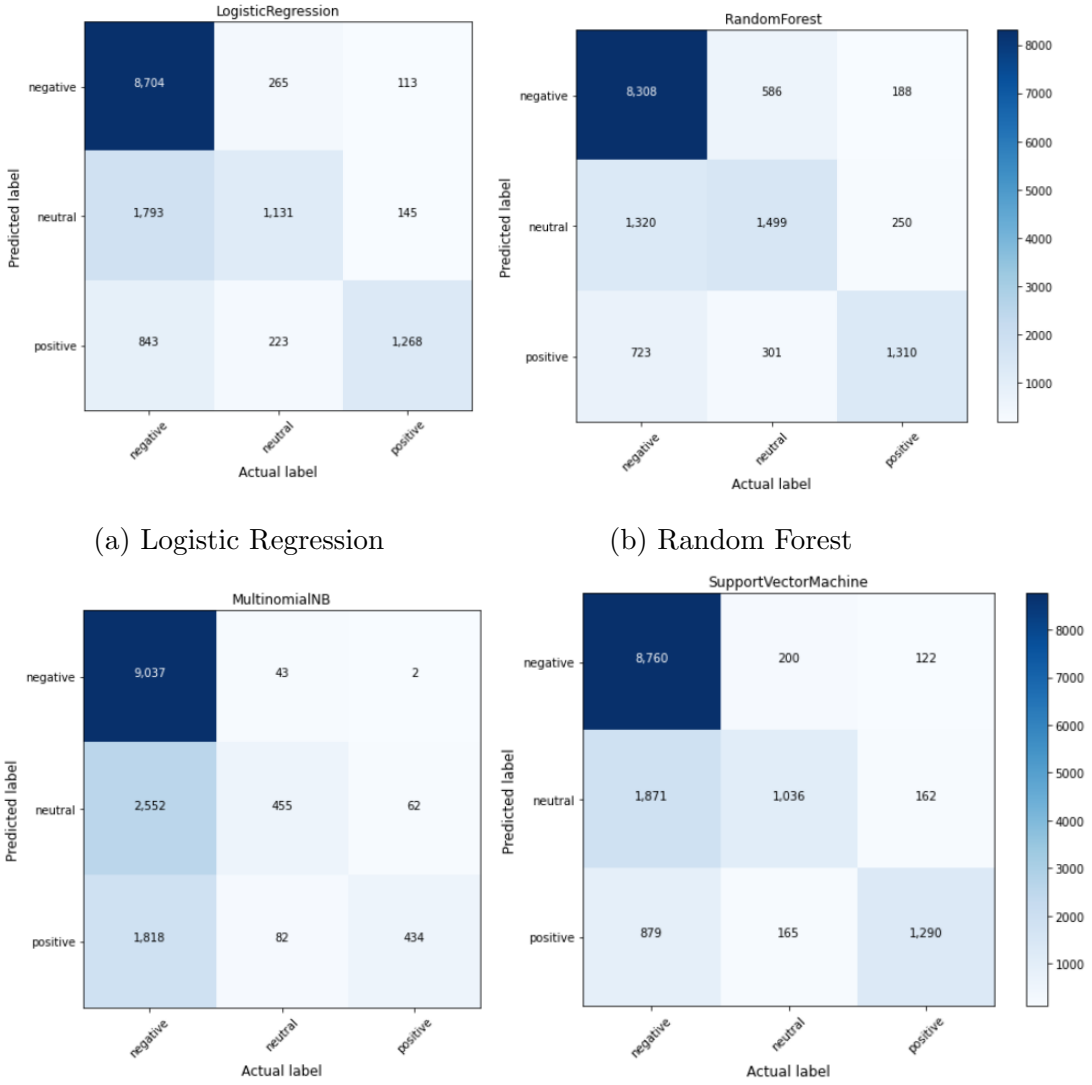
Figure 6.15: Evaluation Measurement chart

As you can see All of the models produce findings of a similar nature. The effectiveness of support vector machines (SVM) and logistic regression is slightly higher.

6.7.2 Using TF-IDF Vectorizer:

Model	Accuracy	Precision	Recall
Logistic Regression	0.766	0.765	0.623
Random Forest	**0.767**	0.726	**0.654**
MultinomialNB	0.685	**0.776**	0.443
Support Vector Machine	0.765	0.773	0.618

Figure 6.16: Evaluation Measurement(TF-IDF)



(a) Logistic Regression

(b) Random Forest

(c) MultinomialNB

(d) Support Vector Machine

Figure 6.17: Confusion Matrix(TF-IDF)

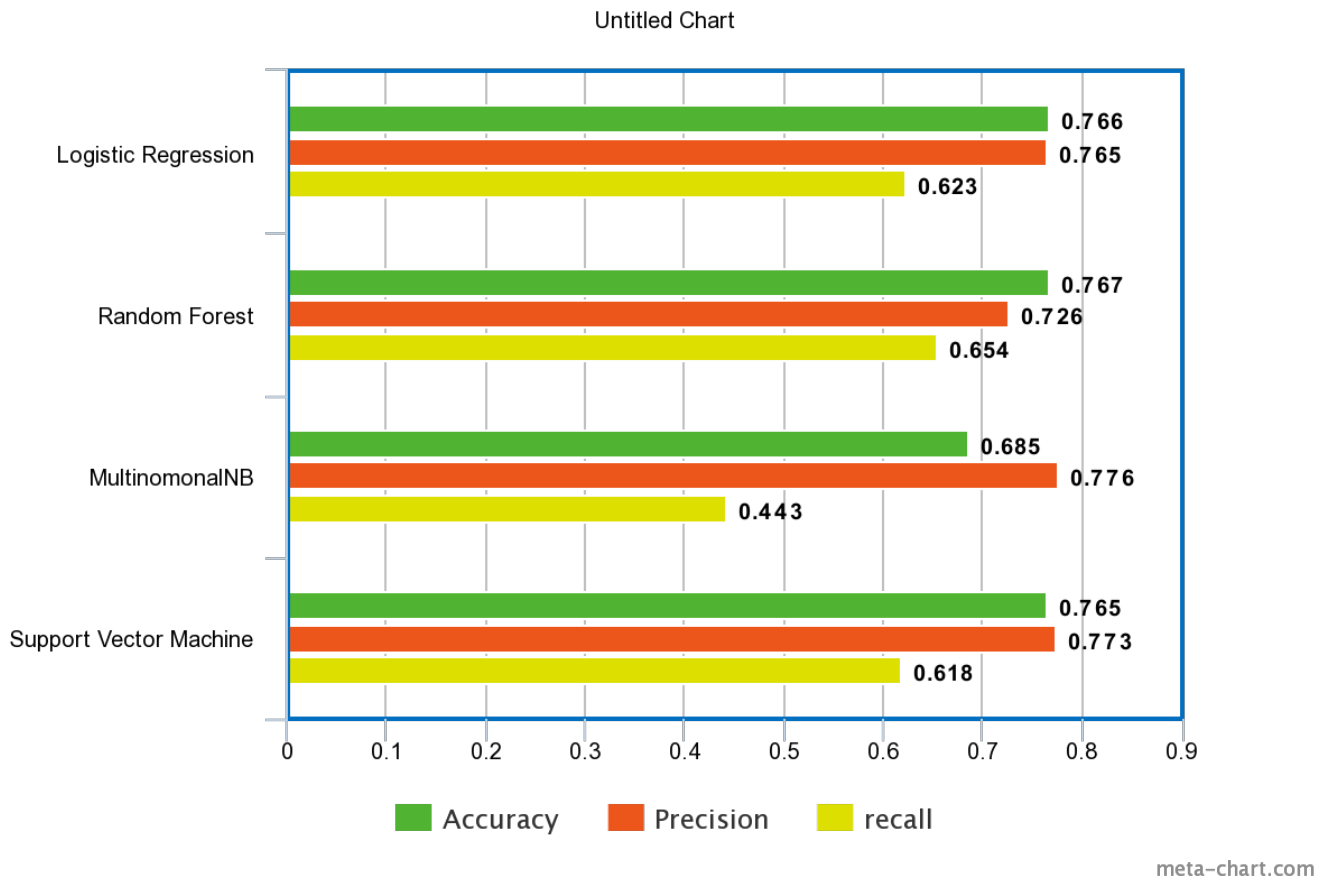


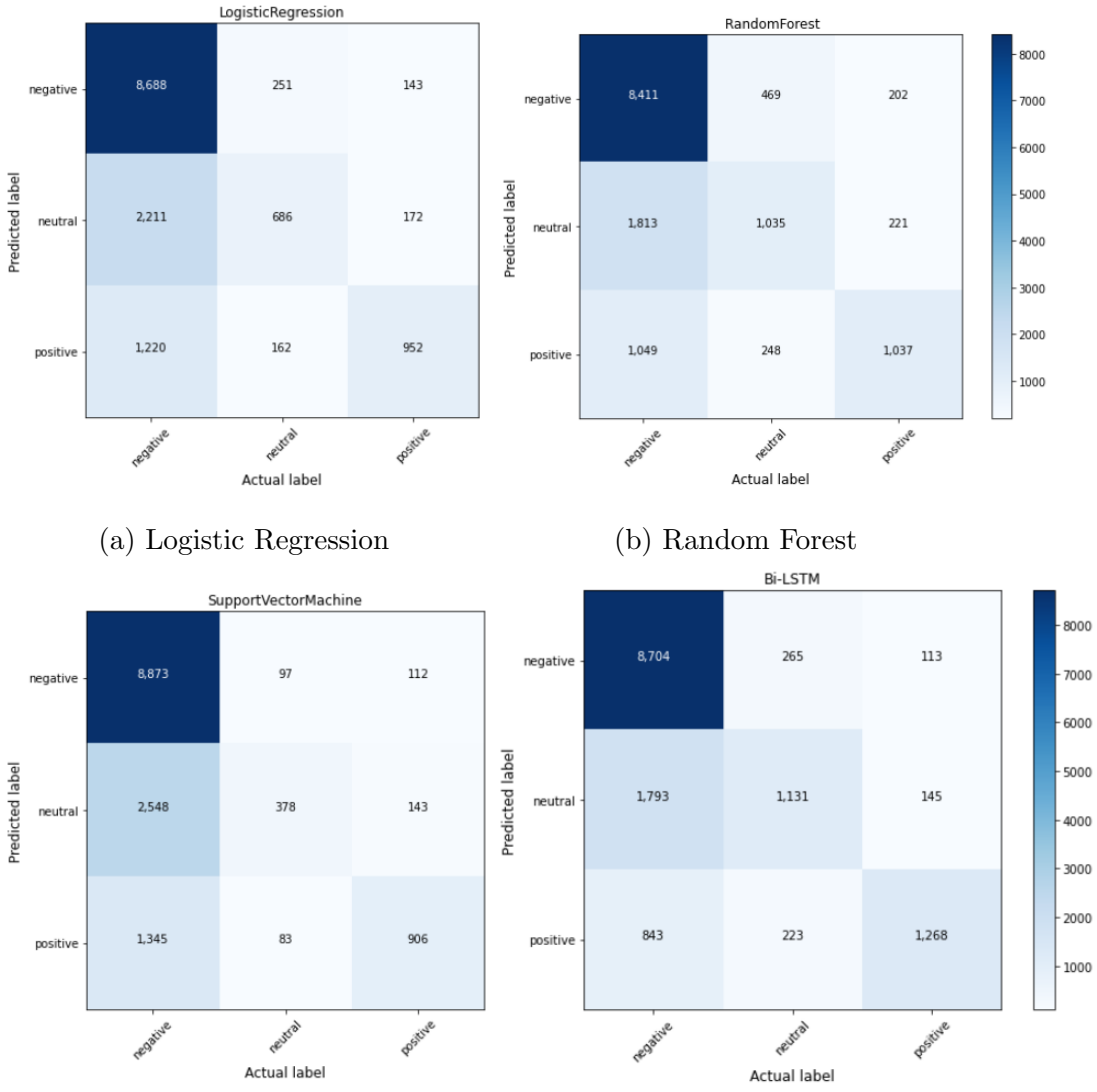
Figure 6.18: Evaluation Measurement chart

The effectiveness of the Nominal Naive Bayes technique has been significantly reduced by the usage of TF-IDF vectorization.

6.7.3 Using Word2Vec:

Model	Accuracy	Precision	Recall
Logistic Regression	0.712	0.694	0.530
Random Forest	0.726	0.685	0.570
Support Vector Machine	0.701	**0.716**	0.497
Bi-LSTM	**0.760**	0.772	0.654

Figure 6.19: Evaluation Measurement(Word2Vec)



(a) Logistic Regression (b) Random Forest

(c) Support Vector Machine (d) Bi-LSTM

Figure 6.20: Confusion Matrix(Word2Vec)

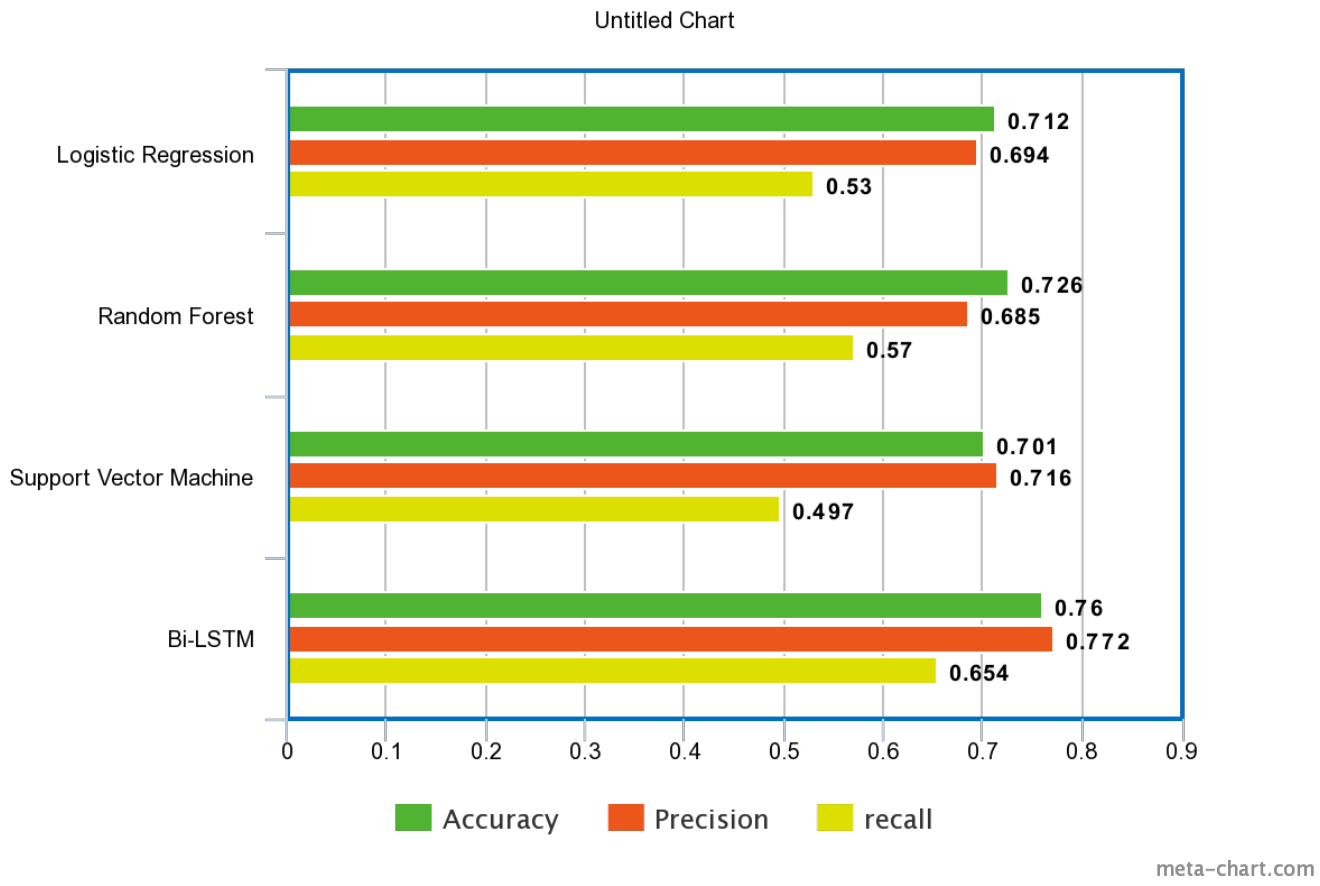


Figure 6.21: Evaluation Measurement chart

With deep learning models like Bi-LSTM, Word2Vec often performs better, but all machine learning techniques have significantly suffered here.

6.7.4 Model Accuracy Summary

Finally, all the accuracy, precision and recall have been shown in one large table to get the complete image of our applied model performance.

	Model	Accuracy	Precision	Recall
CountVectorizer	Logistic Regression	0.786	0.74	0.704
	Random Forest	0.766	0.717	0.667
	MultinomialNB	0.758	0.725	0.627
	Support Vector Machine	0.77	0.743	0.65
TF-IDF	Logistic Regression	0.766	0.765	0.623
	Random Forest	0.767	0.726	0.654
	MultinomialNB	0.685	0.776	0.443
	Support Vector Machine	0.765	0.773	0.618
Word2Vec	Logistic Regression	0.712	0.694	0.53
	Random Forest	0.726	0.685	0.57
	Support Vector Machine	0.701	0.716	0.497
	Bi-LSTM	0.76	0.772	0.654

Figure 6.22: Model accuracy Table

Here, the best result for each model has been measured using the above table.We can determine that the Logistic Regression classifier provided the maximum accuracy after measuring all the parameters using the machine learning technique discussed above.

Model	Accuracy
Logistic Regression	78.6%
Random Forest	76.7%
MultinomialNB	75.8%
Support Vector Machine	77.0%
Bi-LSTM	76.0%

Figure 6.23: Model Accuracy

6.8 Comparison Chart

The base paper have used Logistic Regression, Random Forest, Decision Tree, Adaboost, Support Vector Machine algorithm as Machine Learning approaches and CNN model for Deep Learning Approach. In this paper, Logistic Regression, Random Forest, MultinomialNB, Support Vector Machine as Machine Learning approaches and LSTM for Deep Learning approach have beed used, comparison chart shows our applied models have beeter accuracy in every aspects.

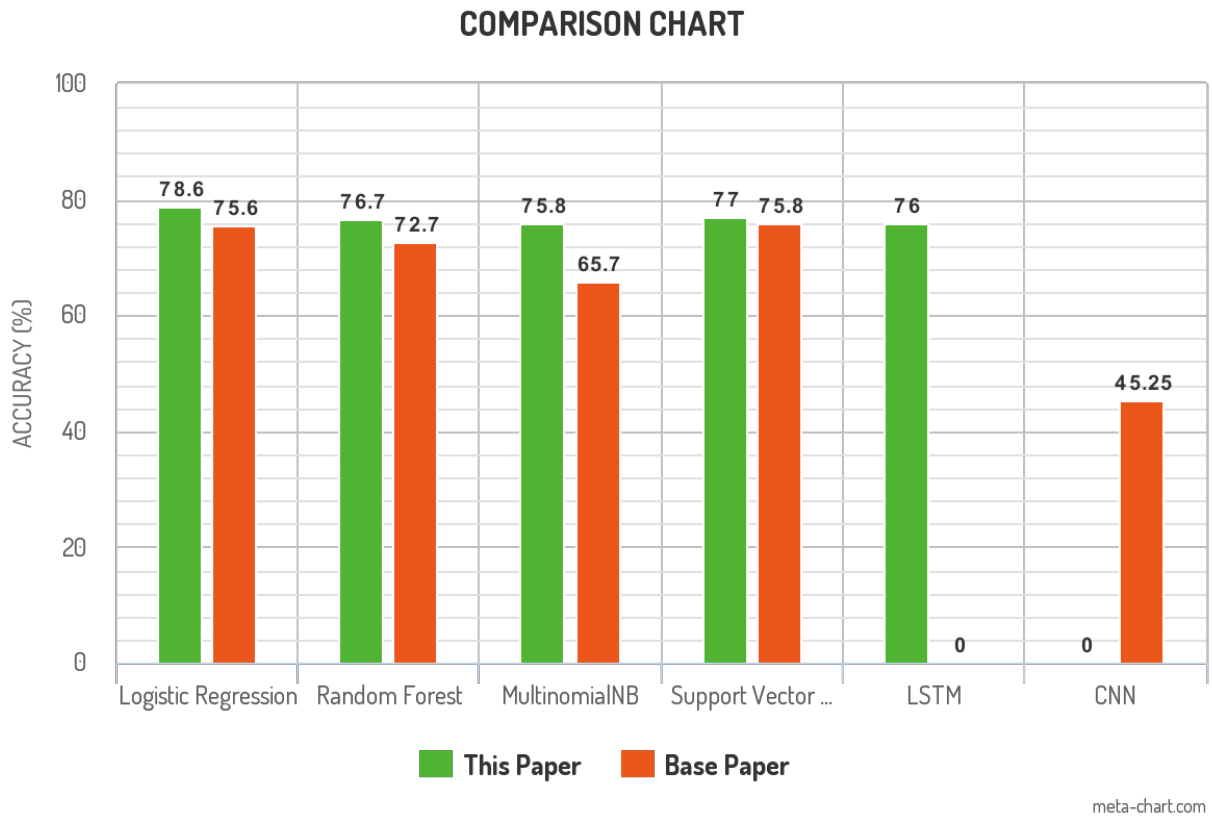


Figure 6.24: Model Accuracy Comparison

6.9 Conclusion

A lexicon-based technique, a machine learning-based approach, or a hybrid approach can all be used for sentiment analysis. The weakness of the sentiment categorization depends on the quantity of the vocabulary, which is a drawback of the lexicon-based technique . This method gets more time-consuming and incorrect as the lexicon's size grows. The various processes for performing sentiment analysis on Twitter data using

machine learning algorithms are explained in detail in this study.

A labeled dataset that is split into a train set and a test set is necessary for a machine learning classifier. Once the right dataset has been gathered, the next step is to preprocess the data (tweets) using NLP-based techniques, then use a feature extraction method to extract features that are important to the sentiment of the dataset.

A model is then developed using machine learning classifiers that have been put to the test on test data. Accuracy, precision, recall, and F-score are four metrics that can be used to assess the model's performance. The framework uses Multinomial NB, Support Vector Machine, Logistic Regression, Random Forest, and LSTM to perform sentiment analysis. Due to the utilization of the Apache Spark framework, the suggested text analytics framework is also real-time, quick, scalable, and reliable.

Chapter 7

Conclusion And Future Work

7.1 Conclusion

In this thesis,machine learning and deep learning methods have been applied for sentiment analysis on linguistic data sets[35]. Using Natural Language Processing (NLP) approaches, the created algorithms for removing noise or data filtering and pre-processing linguistic data are exhibited. Additionally, a series of pre-processing operations must be carried out in order to remove the noise from the textual Twitter data. The input tweets are screened and processed during this procedure to produce more accurate data and shrink the dataset.
The applied method use combined polarity and subjectivity which can be used to classify sentiment more accurately and precisely. The idea behind the work are as follows:

- polarity to evaluate sentiment score.
- Subjectivity to measure the weight of each word or sentence.
- Applied Model to extract sentiment based on both polarity and subjectivity.

It could effectively profile the products, assess trends, and forecast utilizing the sentiment scores for sentiments surrounding a certain product or service and the user's information. As a result, the system is able to predict how a group of people in a specific age range, geographic location, and profession will feel about a specific good or service in the future, which is the most relevant knowledge in the commercial sector.

7.2 Future Works

Issues during this project lead us to take a break in order to mining opinion of the suggested model. Sometimes, despite the fact that it was impeding the model and reducing its accuracy and efficacy, we chose to ignore the issue rather than overcome it.
For future, the system limitation will be eradicated by completing the following tasks:

- We have faced some problems in handling phrases and we will try to improve it in future work.
- Advance Neural Network and Deep learning technique will be applied in our model. So, the design will be easy and more precise.
- Multilingualism is becoming more and more prevalent on social media every day. The future sentiment analysis model can easily handle it.
- Emoji and GIF will be considered and converted to text for better sentiment analysis.
- The business user will be able to access real-time sentiment analysis for the linguistic data with the help of future model.
- The present era is evolving into an age of trolling and sarcasm. Sarcasm will be simple to recognize by future models, and sentiment level misinterpretation can be reduced.

REFERENCES

[1] M. M. J. Kabir, S. Xu, B. H. Kang, and Z. Zhao, "A new multiple seeds based genetic algorithm for discovering a set of interesting boolean association rules," *Expert Systems with Applications*, vol. 74, pp. 55–69, 2017.

[2] B. Liu, "Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1–167, 2012.

[3] V. A. Ho, D. H.-C. Nguyen, D. H. Nguyen, L. T.-V. Pham, D.-V. Nguyen, K. V. Nguyen, and N. L.-T. Nguyen, "Emotion recognition for vietnamese social media text," in *International Conference of the Pacific Association for Computational Linguistics*, pp. 319–333, Springer, 2019.

[4] B. Liu, "Sentiment analysis and subjectivity," 01 2010.

[5] Wikipedia, "The pen is mighter than the sword.." `en.wikipedia.org/wiki/The_pen_is_mightier_than_the_sword/`.

[6] B. Liu, "Handbook chapter: sentiment analysis and subjectivity. handbook of natural language processing," *Handbook of Natural Language Processing. Marcel Dekker, Inc. New York, NY, USA*, 2009.

[7] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," *arXiv preprint cs/0205070*, 2002.

[8] P. D. Turney, "Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews," *arXiv preprint cs/0212032*, 2002.

[9] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," in *Proceedings of the 12th international conference on World Wide Web*, pp. 519–528, 2003.

[10] A. Gupte, S. Joshi, P. Gadgul, A. Kadam, and A. Gupte, "Comparative study of classification algorithms used in sentiment analysis," *International Journal of Computer Science and Information Technologies*, vol. 5, no. 5, pp. 6261–6264, 2014.

[11] M. Devika, C. Sunitha, and A. Ganesh, "Sentiment analysis: a comparative study on different approaches," *Procedia Computer Science*, vol. 87, pp. 44–49, 2016.

[12] A. Tripathy, A. Anand, and S. K. Rath, "Document-level sentiment classification using hybrid machine learning approach," *Knowledge and Information Systems*, vol. 53, no. 3, pp. 805–831, 2017.

[13] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu, "Combining lexicon-based and learning-based methods for twitter sentiment analysis," *HP Laboratories, Technical Report HPL-2011*, vol. 89, pp. 1–8, 2011.

[14] C. Jefferson, H. Liu, and M. Cocea, "Fuzzy approach for sentiment analysis," in *2017 IEEE international conference on fuzzy systems (FUZZ-IEEE)*, pp. 1–6, IEEE, 2017.

[15] X. Ouyang, P. Zhou, C. H. Li, and L. Liu, "Sentiment analysis using convolutional neural network," in *2015 IEEE international conference on computer and information technology; ubiquitous computing and communications; dependable, autonomic and secure computing; pervasive intelligence and computing*, pp. 2359–2364, IEEE, 2015.

[16] D. Goularas and S. Kamis, "Evaluation of deep learning techniques in sentiment analysis from twitter data," in *2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML)*, pp. 12–17, IEEE, 2019.

[17] K. Zhang, W. Song, L. Liu, X. Zhao, and C. Du, "Bidirectional long short-term memory for sentiment analysis of chinese product reviews," in *2019 IEEE 9th international conference on electronics information and emergency communication (ICEIEC)*, pp. 1–4, IEEE, 2019.

[18] S. Naz, A. Sharan, and N. Malik, "Sentiment classification on twitter data using support vector machine," in *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pp. 676–679, IEEE, 2018.

[19] R. Monika, S. Deivalakshmi, and B. Janet, "Sentiment analysis of us airlines tweets using lstm/rnn," in *2019 IEEE 9th International Conference on Advanced Computing (IACC)*, pp. 92–95, IEEE, 2019.

[20] E. Cambria, P. Gastaldo, F. Bisio, and R. Zunino, "An elm-based model for affective analogical reasoning," *Neurocomputing*, vol. 149, pp. 443–455, 2015.

[21] E. Kang, "Long short-term memory (lstm): Concept." `"https://www.medium.com/@kangeugine/long-short-term-memory-lstm-concept-cb3283934359/`.

[22] T. Sullivan.

[23] S. Hochreiter and J. Schmidhuber, "Lstm can solve hard long time lag problems," *Advances in neural information processing systems*, vol. 9, 1996.

[24] B. Liu, "Sentiment analysis and subjectivity," 01 2010.

[25] V. Hatzivassiloglou and J. Wiebe, "Effects of adjective orientation and gradability on sentence subjectivity," 01 2003.

[26] D. S. Sayad, "An introduction to data science." `"https://www.saedsayad.com/logistic_regression.htm/`.

[27] Z. Lateef, "A comprehensive guide to random forest." `"https://www.edureka.co/blog/random-forest-classifier/`.

[28] S. Ray, "Support vector machine algorithm in machine learning." `https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code//`.

[29] T. O. Nelson, "A comparison of current measures of the accuracy of feeling-of-knowing predictions.," *Psychological bulletin*, vol. 95, no. 1, p. 109, 1984.

[30] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," *arXiv preprint arXiv:2010.16061*, 2020.

[31] N. Japkowicz, "Why question machine learning evaluation methods," in *AAAI workshop on evaluation methods for machine learning*, pp. 6–11, 2006.

[32] F. J. Provost, T. Fawcett, R. Kohavi, *et al.*, "The case against accuracy estimation for comparing induction algorithms.," in *ICML*, vol. 98, pp. 445–453, 1998.

[33] J. W. Tukey *et al.*, *Exploratory data analysis*, vol. 2. Reading, MA, 1977.

[34] J. Huang, Y.-F. Li, and M. Xie, "An empirical analysis of data preprocessing for machine learning-based software cost estimation," *Information and software Technology*, vol. 67, pp. 108–127, 2015.

[35] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1253, 2018.