

**Comparison of Naive Bayes and SVM Algorithm Based on Sentiment
Analysis using Review Dataset**

BY

Abdul Mohaimin Rahat

ID: 161-15-7174

&

Abdul Kahir

ID: 161-15-7173

This Report Presented in Partial Fulfillment of the Requirements for the Degree of
Bachelor of Science in Computer Science and Engineering

Supervised By
Mr. Saiful Islam
Senior Lecturer
Department of CSE
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
DECEMBER 2019

APPROVAL

This Project titled “**Comparison of Naive Bayes and SVM Algorithm Based on Sentiment Analysis using Review Dataset**”, submitted by Abdul Mohaimin Rahat, ID No: 161-15-7174, Abdul Kahir, ID No: 161-15-7173, to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 05 December 2019.

BOARD OF EXAMINERS



Dr. Syed Akhter Hossain
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Chairman



Nazmun Nessa Moon
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Gazi Zahirul Islam
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Dr. Mohammad Shorif Uddin
Professor
Department of Computer Science and Engineering
Jahangirnagar University

External Examiner

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **MR. Saiful Islam** Senior Lecturer, Department of CSE Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:



MR. Saiful Islam

Senior Lecturer

Department of CSE

Daffodil International University

Submitted by:



Abdul Mohaimin Rahat

ID: 161-15-7174

Department of CSE

Daffodil International University



Abdul Kahir

ID: 161-15-7173

Department of CSE

Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **MR. Saiful Islam, Senior Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “*Data mining*” to carry out this project. His endless patience ,scholarly guidance ,continual encouragement , constant and energetic supervision, constructive criticism , valuable advice ,reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to Mr. Saiful Isalm and Head, Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Now a day's sentiment analysis is the most used research topic. The sentiment analysis result is based on different investigation for example politics, terrorism, economy, international affairs, movies, fashion, justice, humanity. Social media are the main resource for collecting people's opinion and their sentiment about a different trending topic. People use many abusing words in social media to express their emotion. Using sentiment analysis, we will build a platform where one can easily identify the opinions are either positive or negative or neutral. This research paper will contain supervised learning which is under the machine learning approach. We run an experiment on different queries from humanity to terrorism and find out an interesting result. First of all, we have preprocessed the dataset to convert unstructured airline review into structured review form. After that, we convert structured review into numerical value. We have to preprocess the data before using it. Stop word removal, @ removal, Hashtag removal, POS tagging, calculating sentiment score have done in preprocessing part. Then an algorithm has been applied to classify the opinion as either it is positive or negative. In this research paper we will briefly discuss supervised machine learning. Support vector machine as well as Naïve Bayes algorithm and compares their overall accuracy, precession, recall value. The result shows that in case of airline reviews Support vector machine gave way better result than Naïve Bayes algorithm.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiner	ii
Declaration	iii
Acknowledgement	iv
Abstract	v
Table of content	vi
List of figures	viii
List of tables	ix
 CHAPTER	
 CHAPTER 1: INTRODUCTION	 01-03
1.1 Introduction	01
1.2 Motivation	02
1.3 Research Questions	02
1.4 Expected Outcome	03
 CHAPTER 2: BACKGROUND STUDY	 04-07
2.1 Introduction	04
2.2 Related Works	04
2.3 Research Summary	06
2.4 Scope of the problem	06
2.5Challenge	06

CHAPTER 3: Research Methodology	08-13
3.1 Introduction	08
3.2 Data Collection Procedure	08
3.3 Methodology	08
3.4 Implementation Details	10
CHAPTER 4: Experimental Results and Discussion	14-18
4.1 Introduction	14
4.2 Experimental Results	14
4.3 Summary	18
CHAPTER 5: CONCLUSION & FUTURE SCOPE	19-19
5.1 Conclusion	19
5.2 Future Scope	19
REFERENCE	20
APENDIX	

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.2.1: Data distribution of the review dataset	10
Figure 3.2.2: Workflow of Sentiment Analysis	11
Figure 4.2.1: Accuracy Comparison of Both Algorithm	16
Figure 4.2.2: Accuracy Comparison of Both Algorithm	18

LIST OF TABLES

TABLES	PAGE NO
Table 1: Accuracy of SVM and Naïve Bays Algorithm	15
Table2: Result prediction of both Algorithms	17

CHAPTER 1

Introduction

1.1 Introduction

Social Networking site is the best source to express the innovative and creative thought as well as the review. Different activities occur around us. People use social media to give their opinion on that. Social media is the largest platform and it is open and free for the users. This huge number of networking site contains different thoughts. People express their opinion, their sentiment, their platform, their idea, their feeling through this platform. The events that occur around us are such as politics, injustice, inhumanity, international affairs, economy, natural disaster, terrorism, and various upcoming topics. Social networking site is a place where people can give their personal opinion and think about those issues. Nowadays people can also protest about injustice using this networking platform. Suppose a soccer team can know how much support it has got on social media. But it is harder to do analysis such a big amount of data using simple techniques and methods. So a new thing should be introduced to the analysis of the data. The process should be much more reliable and efficient way than the previous. Again we can imagine a political leader want to know how much popular he or she among the people while his campaigning but it would be quite hard to know if we use general methods or techniques. A machine learning methods proposal could be applied. So that the political leader or the candidate will be able to know his popularity analyzing the sentiment of people.

People use social networking site as a platform to express their emotions. Twitter is very popular among the social networking site. One can face many problems while using the sentiment analysis it could be like data scarcity and so on. This is occurring because of the slang language. There are various number of challenges in Sentiment Analysis. An opinion word expressed by people might be positive in one situation. Again the same opinion word might turn into negative in another situation. Another challenge is that people use different expressing ways to give their opinions on various issues. Those expressing ways are not always the same. Sometimes it is challenging to understand because of its lack of context. For example, "The Game was as good as its previous Game". This entire sentence depends upon the thoughts of the person about the previous version game. People use twitter for multiple purposes. Among the people around the world, twitter reported that it has a number of active users nearest 321 million monthly. People of all age

use twitter. Twitter contains a huge number of information obtained from users. Twitter contains opinion from different part of view. The opinion is an expression of people is a particular subject. Different people have different mind and opinion. People share their opinion on twitter using the tweet and retweet. So we will consider Twitters tweet as our data set and production system that will able to analyze and gives us an accuracy as our output. In this research paper, we are using two types of machine learning algorithms and those are Support Vector Machine algorithm and Naïve Bayes algorithm. This will calculate accuracies, precisions and recall values. There will be three types of output in our result. Those are negative, neutral, positive using supervised learning and unsupervised learning.

1.2 Motivation

Social media is one of the most important communication medium today. Social media technology enables the message to be sent quickly, become widespread and even viral if the topic attracts public attention. Unfortunately, this also means that hate speech can also spread easily and quickly that it can lead to conflicts between groups in society. Hate speech especially concerning religion, nationality that we have seen in the recent past. There are several industries, companies they change or upgrade their marketing strategy on the basis of customer review of their service. Customer express their review on social media. Those companies run a query on how much they satisfy a customer and what are the things they need to apply to fulfil customer satisfaction.

1.3 Research Questions

Our research paper is based on review dataset of customer in a particular topic. Where we analysis customer review on airline service. Customer express their emotion through social networking site. Analyzing the customer review airline companies upgrade their services to provide customer satisfaction. Airline Company's analysis the negative review and according to that they change their service. While the positive review encourage the airline Company's to perform better.

1.4 Expected Outcome

Machine learning approach with prediction algorithm to solve the problem. Here we make an attempt to find out the positive, negative and neutral sentiment on Twitter data. Those positive, negative and neutral divided comment is used to analyze the quality of the airline. This project is divided into two main segments. Firstly the main categories of the reviews were identified. Secondly, sentiment analysis is performed for the categories/aspect detected from the reviews. The result of the reviewed work will provide a robust decision support for the customers which will not only help them in making an effective decision for choosing their airlines but also help the airline companies to look after the areas of improvement. This will also help them to get a competitive edge over their other rivals/competitors in the airline industry.

CHAPTER 2

Background

2.1 Introduction

In this new era, Judgment is based on public reviews. People set up their mind on something by just seeing some review of the product. So product review plays an important part for people judgment or decision making. And Internet is the source where people can get the expected reviews. A user can get knowledge about anything that is available on internet such as any place, country, food, restaurant, hotels and tourist place and so on. The best and for most popular way to get reviews is social networking site. Where other user give their opinion about the things.

2.2 Related Works

This part of our paper will review different research paper works that are similar to ours. Sentiment analyze is the process that can classify human expression in different matters. There could be many sources for sentiment analysis. Text, image, video, audio, are the possible source for sentiment analysis. Using any of those sources analyzing sentiment is possible. But each of those sentiment analyzing source has different possible ways to approach. There are many authors who describe sentiment analyze. Some authors describe that people sentiment will detect automatically using the tweets from twitter. Firstly, get the data set from twitter. Then those data have to be extracted. Then those data have to be extracted. The next step in preprocessing the data. Some authors categorized tweets both as an objective and subjective category. These authors compose trained data instead of manually annotated data. Again they use a word tweet into its part of speech using POS tagging. A significant and must effective sentiment detection is approached by the author.

Bhumika M.Jadav and Vimalkumar B.Vaghela [2] worked on a research paper on Sentiment Analysis. In this research paper, they researched on movie review dataset [2]. They converted unstructured movie review into structured form. Then, they calculated the score of structured word [2]. They gave the score as input to Support Vector Machine and produce output as either positive or negative.

S.M. Mazharul Haque also published a research paper on sentiment analysis. They used big data and R platform to analyze the advanced level of data [3]. Again, the algorithm they used was the Naïve Bayes algorithm along with the Lexicon Based analyzer [3]. Thus, they find out an effective and fruitfully way to analyze the sentiment of people.

A research paper from Sharvil Saha K Kumar was published on sentiment analysis. They first retrial the data before they began preprocessing the data. They used hashtag classification to find out the trending topic and the most used tweet and re-tweet [10]. Finally, the heart of the paper polarity classifier is used. They use Naïve Bayes classified Bigram models for their classification of polar data [10].

Sanjay Chakraborty worked on a research paper on Sentiment Analysis. In their research paper, they researched on movie and hotel review [4]. Naïve Bayes and K-NN classifier are used on this research paper [4]. Finally, they arrive at an accuracy of 80% on movie review, while the accuracy rate for hotel review was lower.

Raheesa Safrin published a research paper on sentiment analysis. They researched on product review based on the customer comments on their own website [1]. In the website, customers give their opinion in the comment section. The comments are of two types, one being positive and the other negative. The positive tagging inspired the company and extracted the most effective characteristic to get the best result. They used K-NN algorithm. K-NN is a technique of classification and K-means is a clustering technique [1].

Zohreh Madhoushi et al [12] worked on a research paper on Sentiment Analysis Techniques in recent words. They used three types of techniques to classify the sentence. Firstly they used machine learning approaches which contain both Support vector machine and Naïve Bayes classification. SVM analyze and recognize data patterns while Naïve Bayes calculates maximum likelihood. Secondly they used lexicon-based approach which is used for parts of speech tagging and WordNet. Finally they used hybrid approaches which contains both machine learning and lexical resource. That's how they find an effective way to analysis sentiment of people.

Divya Sehagal et al [13] worked on real-time Sentiment Analysis of big data application using the help of twitter data. In this paper, they analyze and process big data. They used Hadoop technology for analyzing and preprocessing. They also stream the data API from twitter. They split the root

word and remove the unwanted and unnecessary extra storage of derived words. All those steps are necessary to analysis Big Data Application using twitter data.

2.3 Research Summary

In this paper we are analyzing the review dataset of airline. Those dataset of airline. Those dataset contains both positive, negative, neutral reviews about the airline companies. There are both supervised and unsupervised approaches applied in this paper. The research paper will contain supervised learning which is under machine learning algorithm. Here the unstructured airline reviews are converted into a structured form. We also preprocess the data before using it. Stop word removal, @ removal, Hashtag removal, removal URL, POS tagging have done in preprocessing part. Finally the algorithm is applied to produce an output. Both Support Vector Machine and Naïve Bayes Algorithm is applied for a better outcome.

2.4 Scope of the Problem

There will be a lot of problems in our research paper. There are many types of sentiment problems analyzing text review, exact meaning of the review, expressions analyzing and in the evaluation of sentiment analysis. In this research paper, we will briefly discuss different machine learning prediction algorithms. Many people are same review to find out the sentence is in which class level, we have faced the problem the similarity and dissimilarity between the customer review. The sentiment problems evaluation and detection polarities for customer reviews and find the most effected algorithms solutions for the highest accuracy for text.

2.5 Challenges

There are several challenges of our research paper, first of all fake detection and spam of the user's emotions. Example Review1: "The airline is good" and Review2: "The airline is not good". In this example, the first review sentiment score [good] with positive expressions of customers and the second review sentiment score [not good] with negative expressions of customers. When we evaluate the sentiment score for determine airline the first expression will be useful and the second expression will be spam and fake detection. There are many similar review in our dataset. There

are most difficult to show the sentiment scores: positivity, negativity, objectivity. Another problem of temporal review when the airplane services is good then the customers review is positive, unfortunately some of the reason this airline service services is not good then the customers review is negative. This types of sentiments that are changed with time may enhance the sentiment analysis performance.

CHAPTER 3

Research Methodology

3.1 Introduction

There are already many works that were done according to this topic. This research field is the mostly used and popular field to work on. Natural Language Processing is the area where all those research were done. Sentiment analysis is the mostly used topic under Natural Language Processing area. Sentiment analysis is basically opinion mining of people. In this research we will apply necessary algorithm to calculate sentiment and produce a good accuracy.

3.2 Data Collection procedure

Executing the research paper, the Data Set that is considered here was Airline Reviews, those tweets were collected from the Twitter platform. We collected around 14000 Tweets from the twitter. Where positive and negative reviews of Airline were given. We analysis those reviews to get our result.

3.3 METHODOLOGY

I. DATA PREPROCESSING

Twitter is the data source for our research. There are several reasons that are why we choose twitter data. According to Twitter, it has almost around 200 million active users. Those users' tweets almost 500 million per day report says. There is also an important reason to choose twitter. We could collect data from various social media like Facebook, Instar gram etc. Those social media might contain videos, pictures, many web sharing link and so on. But twitter is all about text. Twitter allows the user to publish text with a limitation of 140 characters. Besides the positive side, it also has a negative site. Because twitter contains many hash tagging, slangs etc. Again twitter contains URL links in the texts.

A. '@' Remove The first thing we should do is to remove '@' as well as the URLs. Again '@' mention a username. It is used to tag something. So the total word that starts with '@' could be removed.

B. Hashtag Remove data Twitter text contains hashtag ‘#’. The hashtag is responsible to trend a topic. Hashtag might contain some slugger. Example: #we_want_justice. So we decide to remove the ‘#’ but keep the entire word after the hashtag.

C. Stop words Remove The major part of the data preprocessing is stopping words removal means to filter out useless data from the text. Stop words is known as the removal of unnecessary data. Natural language toolkit in python has a list of stop words in 16 different languages. Example “I love eating, So I eat”. Here after stopping words this sentence keyword will look like, “love”, “eating”, “eat”.

D. Stemming Stemming is a basic part of text preprocessing module. Stemming is basically the operation of transforming the targeted words to its source form. It reduces the word into their base form. Stemming is a technique that is used to preprocess the data. It converts word like “plays”, “playing”, “played”, “player” to its root or base word form which is “play”. By using a stemmer, the word that is converted into the root form will be a unique word before further processing.

Sentiment analysis is a famous research area in NLP. So, previously many research works have done based on different kinds of methodology. But the study of the previous result said that machine learning algorithm such as SVM and Naive Bayes provides a better result. In our working process firstly we collect the dataset using twitter API. Then preprocess the review text for the trained model. Finally, check the result of the working model of this research. This section will be discussed by applying the methodology for this research work.

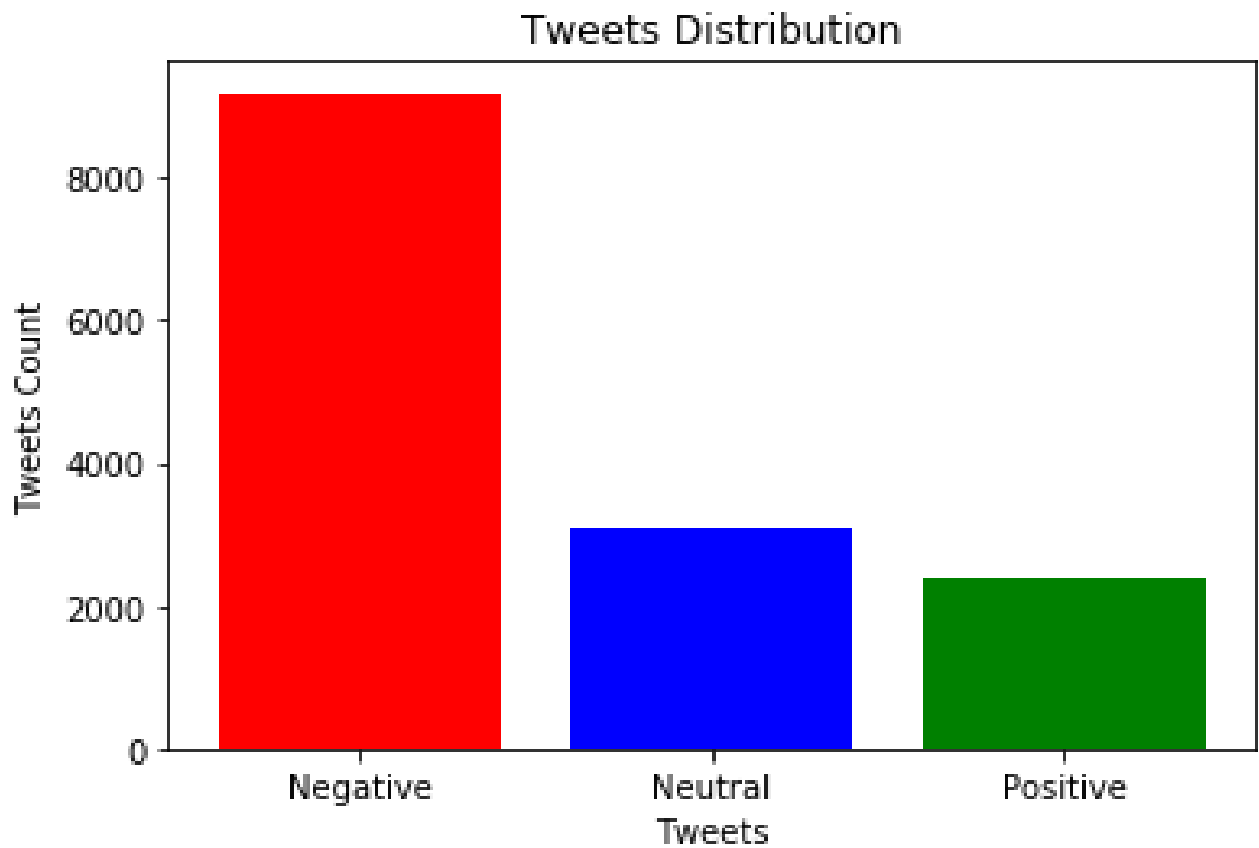


Figure 3.2.1: Data distribution of the review dataset

3.4 Implementation Details

The motive of the research paper is to analyze data with both supervised learnings which is under the machine learning approach. A graphical representation of this sentiment analysis. In this flow helps to understand the whole working process of this research work. The short view of the whole working flow is given below.

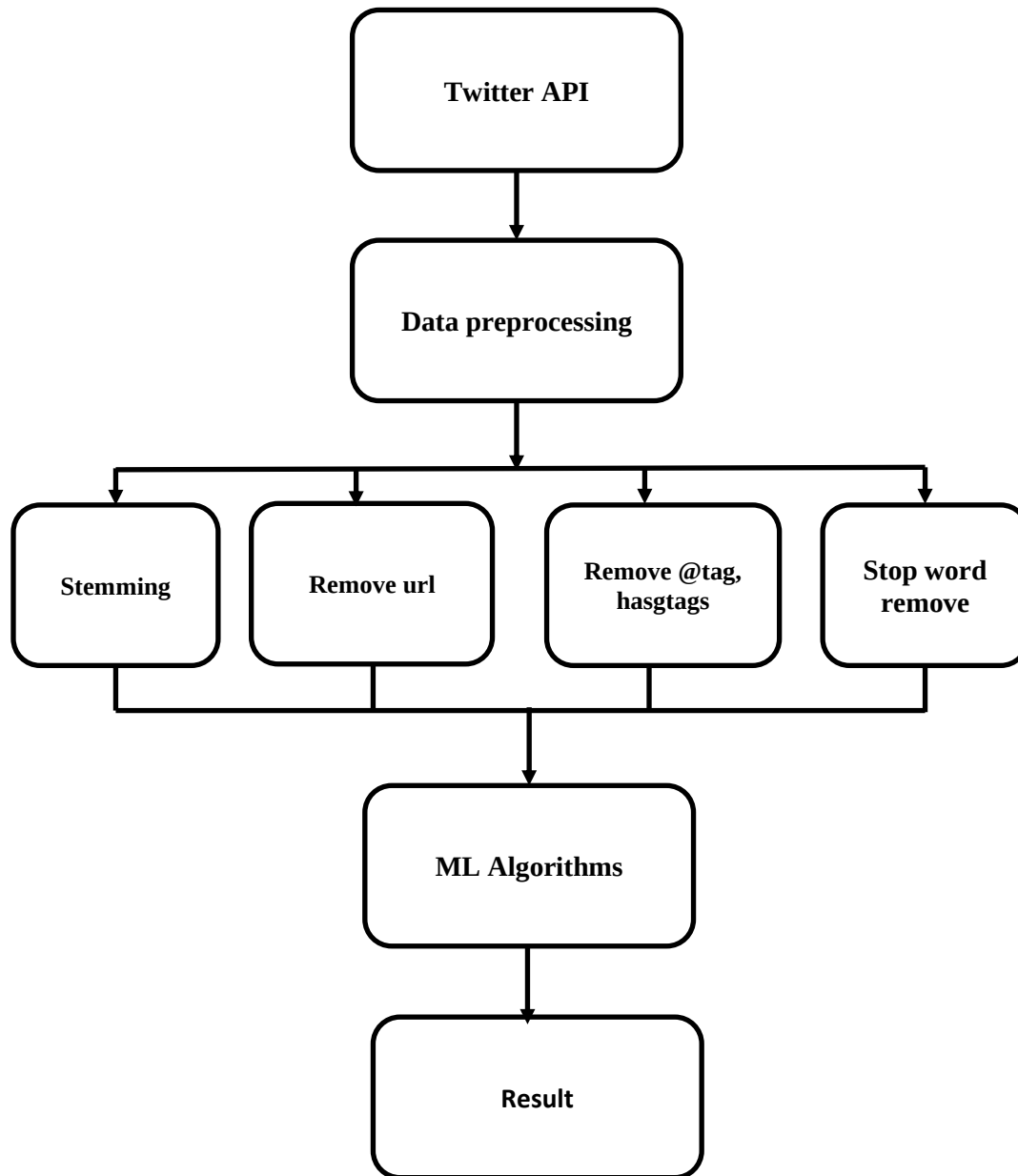


Figure 3.2.2: Workflow of Sentiment Analysis

The step we followed to execute this research paper firstly we collect our datasets from twitter. It's an airline review of the passengers. Passengers give their opinion on a different perspective. We use natural language toolkit (NLTK) for this research paper. We also use Naïve Bayes and SVM for these methods.

A. Naïve Bayes Approach

Naive Bayes is a collection of classification algorithms which are based on Bayes Theorem. Naive Bayes classifier gives us an excellent result when one uses it for text data analysis. Such as Natural Language Processing. Naive Bayes algorithm gives us a probability analyzing the data set we have given. Naïve Bayes classifier is used as a probabilistic classifier. To perform the classifier, it uses the concepts of mixture models. A mixture model is capable of establishing the probability of the component that it consists of Bayes theorem to perform as a probabilistic classifier. Another name that a naïve Bayes is known as simple Bayes or independence Bayes. The probability P is defined as follows:

$$P(m|n) = \frac{P(n|m) P(m)}{P(n)} \dots\dots\dots(i)$$

Above,

$P(m | n)$ is the probability of class x. Where x is the target and predictory is the attribute.

$P(m)$ is the prior probability of class.

$P(n | m)$ is the probability of predictor of the given class.

$P(n)$ is the prior probability of predictor.

B. Support Vector Machine

Support Vector Machine is a universal learner. Support Vector Machine has defined both input and output format. The output is either positive or negative and input is vector space. The text document is not suitable for learning. Those texts are transformed into a structured format. The text is transformed into a format which matches into the input of machine learning algorithm. The score of the texts are calculated and then the score is given as input to Support Vector Machine. Support Vector Machine has been proved one of the most powerful learning algorithms for text Categorization. But text categorization sometime may produce occurs. To decide which one is better between texts a comparison of text classifier is required. The performance measure is used in this case.

CHAPTER 4

Experimental Result and Discussion

4.1 Introduction

In this section we have discussed about our experimental result about sentiment analysis on airline review. In this paper we are working on airline sentiment review, our main features was to develop a system that provide higher accuracy and easy to implement in machine learning algorithm. The result of the airline sentiment reviewed work will provide a best decision support for the customers choosing their airlines and it also helps the airline companies to improve their airline services.

4.2 Experimental Result

We have collected data to twitter and preprocessed the data according to our interest. We have used 14000 text airline review where 9000 is negative, 3000 is positive and 2000 positive review. We have used 10000 review from training the model and 4000 for testing

Predicted Data \ Actual Data	(Positive)	(Negative)
(Positive)	True Positive	False Positive
(Negative)	False Negative	True Negative

Accuracy: Accuracy is a performance measure again it is a ratio of accurately predicted

Observation from the overall observation.

F1 Score: F1 Score is an overall measure of accuracy that includes both precision and recall. It's the weighted average of both Precision & Recall.

Precision: Precision is the measure of accurately determined positive events from the overall predicted positive events.

Recall: Recall is the measure of accurately determined positive events from the actual positive events.

- Accuracy (A) = $\frac{(TP+TN)}{(TP+TN+FP+FN)}$
- AUC (Area under the Curve) = $\frac{1}{2} \cdot \frac{T-P}{TP+FN} + \frac{TN}{TN+FP}$
- Precision = $\frac{TP}{TP+FP}$
- Recall = $\frac{TP}{TP+FN}$
- F1 Measure = $\frac{Precision+Recall}{2}$

Before the experiment result set the train sample size 67% of the whole dataset and set the random state is 40. Then used the SVM and naive bays algorithm for result prediction. For SVM we used SVC linear kernel and for Naïve Bayes used multinomial Naive Bayes. The classification report of both algorithms is given in table1.

Support Vector Machine give accurate result than Naïve Bayes algorithm. After the train, we test both algorithm prediction by the review twitter and define the best performer. The classification report of both algorithms is given in table1.

Table 1: Accuracy of SVM and Naïve Bays Algorithm.

	SVM	Naïve Bayes
Accuracy	82.48	76.56
Precision	90.33	89.00
Recall	81.79	83.75
F1 Measure	85.85	86.37

In table 1 displays the result comparison of both applying algorithms. The accuracy of SVM is 82.48 where Naïve Bayes accuracy is 76.56. Precision and Recall value for SVM is 90.33 and 81.79 also 89.00 and 83.75 for Naïve Bayes.

Now in figure3 shows the comparison of both algorithms. Where SVM provides almost 83% accuracy and Naïve Bayes provide 77% accuracy in this experiment.

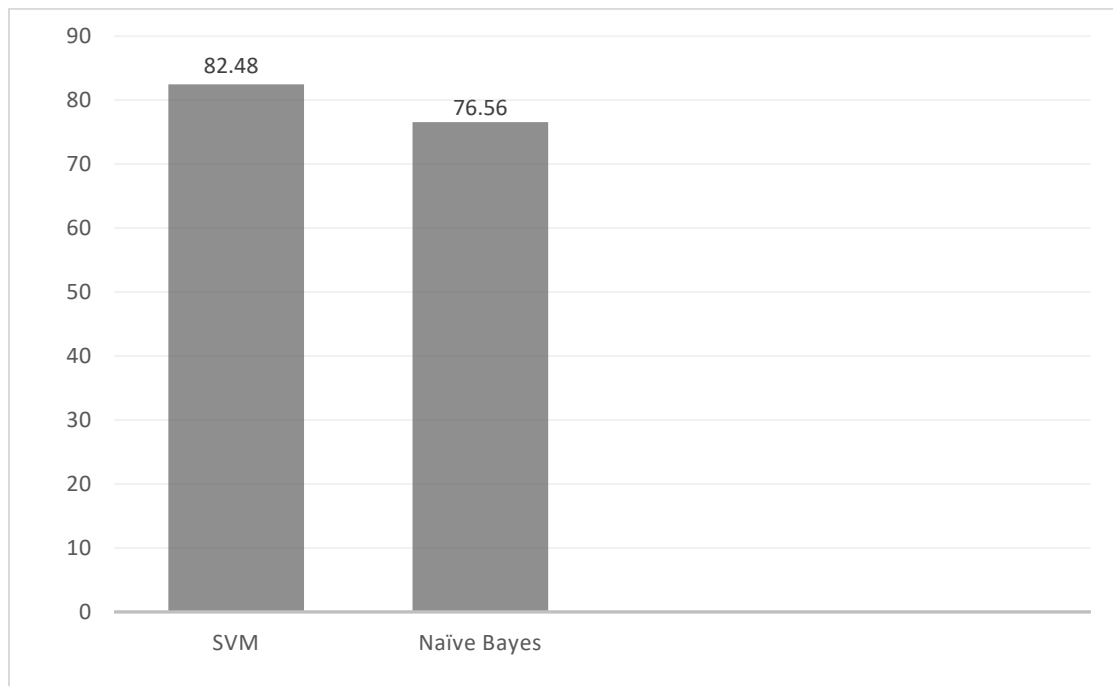


Figure 4.2.1: Accuracy Comparison of Both Algorithm

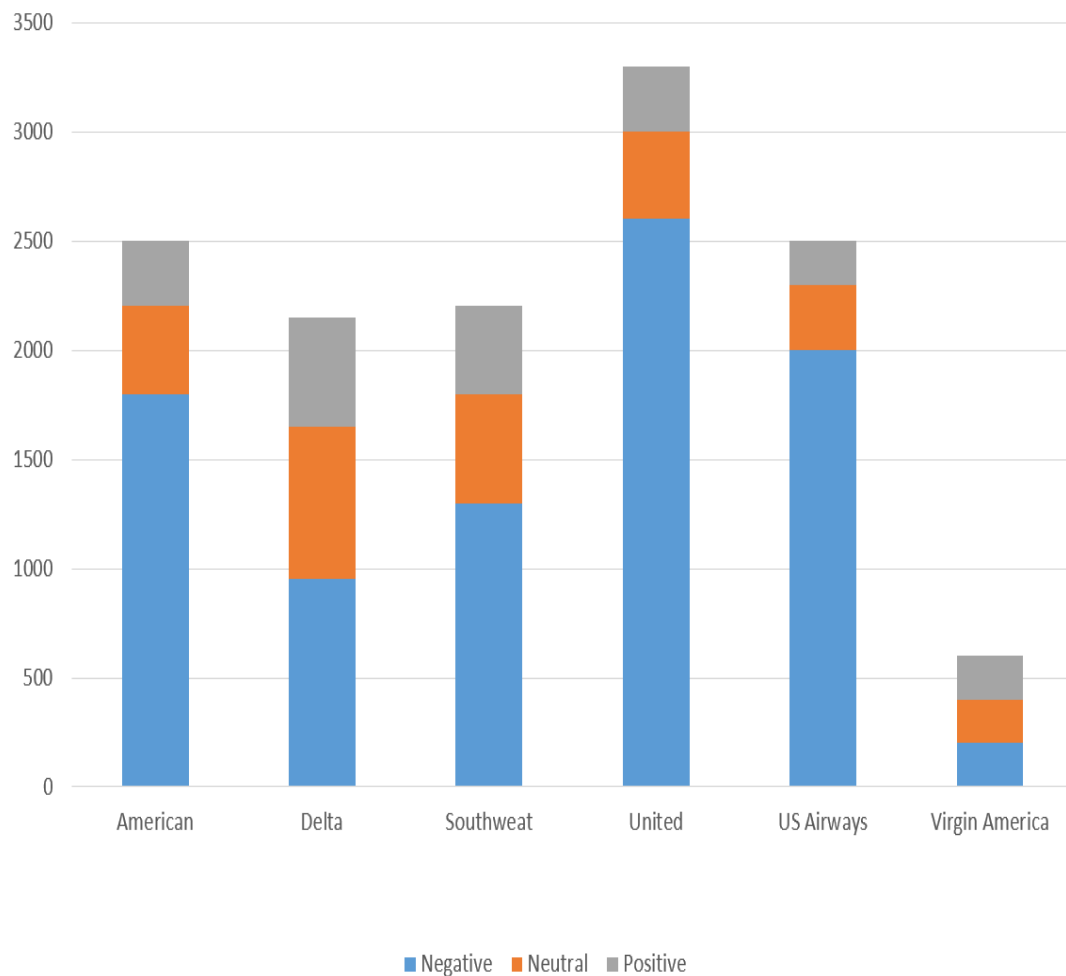
After that, check the model output using the review from the dataset and predict the output using both algorithms. Check the result of model performance we test the prediction of the model by using the review. Take the review as input from the user and then classify review using both algorithms. Maximum time the SVM classifier gives the accurate result where Naïve Bayes is not. In table2 shows the predicted output result for providing a review.

Table2: Result prediction of both Algorithms.

	SVM Prediction	Naive Bayes Prediction
Prediction	0	1
Actual Sentiment	Positive Review	Positive Review
Response Sentiment	Positive Review	Negative Review

Review Text: “flying with @united is always a great experience”. Here, 1 contain negative review and 0 contain positive review.

Figure 4.2.2: Accuracy Comparison of Both Algorithm



Summary:

We have tried all the possible ways to get the best possible result. We have applied two different algorithm and compared between them to get the highest possible accuracy. We have divided the sentence into positive, negative and neutral. We also calculate the precision, recall, F1 Measure and accuracy and applied it. Comparison to other related model about sentiment analysis, we were possible to achieve good rate of accuracy.

Chapter 5

Conclusion & Future Scope

5.1 Future Scope

We research on a specific platform for sentiment analysis system. Through our research results a good output but it has some limitations too. This research could be done better if some more algorithm applied. We compare two algorithm and results the maximum output. But if some more algorithm could applied we would be able to compare all the results and use the best accuracy for final output. KNN is an algorithm that can be introduced. Applying all those three algorithm will take this research to another level. We are confident that our research is benefits to all mankind. It will make a revolution change to many industries and companies as they will get to know what people think about their companies and industries.

5.3 Conclusion

In this paper, we use machine learning algorithms of Naïve Bayes and Support Vector Machine for sentiment classification of airplane reviews. In this paper we used the perform measurement of people's sentiment analysis. This paper is used to perform evaluation measure on comments obtained from the passengers. As an example, a positive sentiment is "Good flight" and a negative sentiment is "Late flight!" In this paper, we take an attempt to find out the positive and negative sentiments on Twitter data. The separation of positive and negative comments is used to analyze the quality of the airplane. The result of the reviewed work will provide a robust decision support for the customers which will not only help them in making an effective decision for choosing their airlines but also help the airline companies to look after the areas of improvement.

REFERENCES

- [1] Raheesa Safrin, K.R.Sharmila, T.S.Shri Subangi, and E.A.Vimal “Sentiment Analysis on online PRODUCT review” International Research Journal of Engineering and Technology (IRJET) Volume: 04 Issue: 04 Apr -2017.
- [2] Vimalkumar B. Vaghela and Bhumika M. Jadav “Sentiment Analysis using Support Vector Machine based on Feature Selection and Semantic Analysis” International Journal of Computer Applications (0975 – 8887) Volume 146 – No.13, July 2016.
- [3] S. M. Mazharul Hoque Chowdhury, Priyanka Ghosh, Sheikh Abujar, Most. Arina Afrin, and Syed Akhter Hossain “Sentiment Prediction Based on Lexical Analysis Using Deep Learning” Springer Nature Singapore Pte Ltd. 2019 A. Abraham et al. (eds.), Emerging Technologies in Data Mining and Information Security, Advances in Intelligent Systems and Computing 814
- [4] Lopamudra Dey, Sanjay Chakraborty, Lopamudra Dey, Beepa Bose and Sweta Tiwari ”Sentiment Analysis of Review Datasets Using Naïve Bayes‘ and K-NN Classifier” Information Engineering and Electronic Business, 2016, 4, 54-62, July 2016.
- [5] Ali Selamat Faculty and Nurulhuda Zainuddin “Sentiment Analysis Using Support Vector Machine” 2014 IEEE 2014 International Conference on Computer, Communication, and Control Technology (I4CT 2014), September 2 - 4, 2014 - Langkawi, Kedah, Malaysia978.
- [6] Ms. Gaurangi Patil, Ms. Varsha Galande, Mr. Vedant Kekan, and Ms. Kalpana Dange “Sentiment Analysis Using Support Vector Machine” International Journal of Innovative Research in Computer and Communication Engineering, (An ISO 3297: 2007 Certified Organization), Vol. 2, Issue 1, January 2014.
- [7] Geetika Gautam, Divakar Yadav “Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis”
- [8] Yan Luo ,Wei Huang “Product Review Information Extraction Based on Adjective Opinion Words” 2011 Fourth International Joint Conference on Computational Sciences and Optimization IEEE computer society.
- [9] Bin Ren, Lianglun Cheng “Research of Classification System based on Naive Bayes and MetaClass” 2009 Second International Conference on Information and Computing Science IEEE computer society.

[10] Shah, Sharvil, Kumar, K.; Saravanaguru, Ra. K “Sentiment Analysis of Twitter data using classifier algorithm”

International Journal of Electrical & Computer Engineering (2088-8708). Feb2016, Vol. 6 Issue 1, p357-366. 10p.

[11]. M.A. Hearst,” Support vector machines,”IEEE Intelligent Systems, pp. 18-28, 1998.

[12] Zohreh Madhoushi, Abdul Razak Hamdan, Suhaila Zainudin, “Sentiment Analysis Techniques in Recent Works “, Science and Information Conference 2015 July 28-30, 2015 London, UK.

[13] Divya Sehgal , Ambuj Kumar Agarwal, “Real-time Sentiment Analysis of Big Data Applications Using Twitter Data with Hadoop Framework”, Springer Nature Singapore Pte Ltd. 2018, Soft Computing: Theories and Applications, Advances in Intelligent Systems and Computing 584

Turnitin Originality Report

Processed on: 31-Oct-2019 14:09 +06

ID: 1204151410

Word Count: 5051

Submitted: 1

Similarity Index

27%

Similarity by Source

Internet Sources: N/A
Publications: N/A
Student Papers: 27%

Sentiment Analysis By Abdul
Mohaimin Rahat

4% match (student papers from 01-Sep-2019)

[Submitted to Daffodil International University on 2019-09-01](#)

3% match (student papers from 02-Apr-2019)

[Submitted to Daffodil International University on 2019-04-02](#)

3% match (student papers from 11-Dec-2017)

[Submitted to National College of Ireland on 2017-12-11](#)

3% match (student papers from 02-Apr-2019)

[Submitted to Daffodil International University on 2019-04-02](#)

2% match (student papers from 14-Dec-2017)

[Submitted to Indian Institute of Technology Roorkee on 2017-12-14](#)

1% match (student papers from 25-Jun-2018)

[Submitted to Chandigarh University on 2018-06-25](#)

1% match (student papers from 02-Apr-2019)

[Submitted to Daffodil International University on 2019-04-02](#)

1% match (student papers from 07-Apr-2018)

[Submitted to Daffodil International University on 2018-04-07](#)

1% match (student papers from 02-Sep-2019)

[Submitted to Griffith College Dublin on 2019-09-02](#)

1% match (student papers from 13-May-2019)

[Submitted to Universiti Teknologi Malaysia on 2019-05-13](#)

1% match (student papers from 05-Oct-2018)

[Submitted to Monash University on 2018-10-05](#)

< 1% match (student papers from 02-Sep-2010)

[Submitted to University of Oxford on 2010-09-02](#)

< 1% match (student papers from 03-Apr-2019)

[Submitted to Daffodil International University on 2019-04-03](#)

< 1% match (student papers from 02-Apr-2019)

[Submitted to Daffodil International University on 2019-04-02](#)

< 1% match (student papers from 02-Apr-2019) Submitted to Daffodil International University on 2019-04-02
< 1% match (student papers from 30-Jul-2018) Submitted to Universiti Teknologi Petronas on 2018-07-30
< 1% match (student papers from 27-May-2018) Submitted to University of Melbourne on 2018-05-27
< 1% match (student papers from 09-Oct-2018) Submitted to University of Hertfordshire on 2018-10-09
< 1% match (student papers from 16-Oct-2018) Submitted to University of Melbourne on 2018-10-16
< 1% match (student papers from 27-Apr-2019) Submitted to Indian Institute of Technology, Madras on 2019-04-27
< 1% match (student papers from 03-Sep-2019) Submitted to Higher Education Commission Pakistan on 2019-09-03
< 1% match (student papers from 04-Dec-2016) Submitted to University of Ghana on 2016-12-04
< 1% match (student papers from 23-May-2018) Submitted to Sharda University on 2018-05-23
< 1% match (student papers from 14-Sep-2017) Submitted to University of Warwick on 2017-09-14
< 1% match (student papers from 20-May-2019) Submitted to South Bank University on 2019-05-20
< 1% match (student papers from 30-May-2019) Submitted to National Institute of Technology, Hamirpur on 2019-05-30
< 1% match (student papers from 23-Mar-2019) Submitted to Aston University on 2019-03-23
< 1% match (student papers from 14-Mar-2018) Submitted to Queen Mary and Westfield College on 2018-03-14
< 1% match (student papers from 09-Apr-2018) Submitted to Banaras Hindu University on 2018-04-09
< 1% match (student papers from 12-Nov-2015) Submitted to University of Sunderland on 2015-11-12
< 1% match (student papers from 23-Feb-2019) Submitted to Assumption University on 2019-02-23
< 1% match (student papers from 12-Sep-2017) Submitted to Haberdashers' Aske's Boys' School on 2017-09-12
< 1% match (student papers from 19-Apr-2016)

[Submitted to University of Hertfordshire on 2016-04-19](#)

< 1% match (student papers from 02-Apr-2019)

[Submitted to Daffodil International University on 2019-04-02](#)

< 1% match (student papers from 05-Apr-2018)

[Submitted to Daffodil International University on 2018-04-05](#)

< 1% match (student papers from 04-Sep-2017)

[Submitted to University College London on 2017-09-04](#)

< 1% match (student papers from 05-Jun-2018)

[Submitted to University of Southern California on 2018-06-05](#)

< 1% match (student papers from 07-Sep-2018)

[Submitted to University of Stirling on 2018-09-07](#)

< 1% match (student papers from 22-Aug-2018)

[Submitted to University of Kent at Canterbury on 2018-08-22](#)

[Sentiment Analysis of Review Dataset Using Naive Bayes and Support Vector Machine](#) BY Abdul Mohaimin Rahat ID: 161-15-7174 & Abdul Kahir ID: 161-15-7173 [This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering Supervised By Mr. Saiful Islam Senior Lecturer Department of CSE Daffodil International University DAFFODIL INTERNATIONAL UNIVERSITY DHAKA, BANGLADESH DECEMBER 2019 APPROVAL This Project titled "Sentiment Analysis of Review Dataset Using Naive Bayes and Support Vector Machine", submitted by Abdul Mohaimin Rahat ID: 161-15-7174 and Abdul Kahir ID: 161-15-7173 to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on. BOARD OF EXAMINERS \(Name\). \[Font-12, Bold\] Designation Department of CSE \[Font-12\] Faculty of Science & Information Technology DaffodilInternationalUniversity Chairman \(Name\). Designation Department of CSE Faculty of Science & Information Technology DaffodilInternationalUniversity Internal Examiner \(Name\) Designation Department of ----- Jahangirnagar University External Examiner](#)

DECLARATION We hereby declare that, this project has been done by us under the supervision of MR. Saiful Islam Senior Lecturer, Department of CSE Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma. Supervised by: MR. Saiful Islam Senior Lecturer Department of CSE Daffodil International University Submitted by: Abdul Mohaimin Rahat ID: 161 -15- 7174 Department of CSE Daffodil International University Abdul Kahir ID: 161 -15- 7173 Department of CSE Daffodil International University

ACKNOWLEDGEMENT First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully. We really grateful and wish our profound our indebtedness to MR. Saiful Islam, Senior Lecturer, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "Data mining" to carry out this project. His endless patience ,scholarly guidance ,continual encouragement , constant and

energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this project. We would like to express our heartiest gratitude to Mr. Saiful Isalm and Head, Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University. We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work. Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT Abstract— Now a day's sentiment analysis is the most used research topic. The sentiment analysis result is based on different investigation for example politics, terrorism, economy, international affairs, movies, fashion, justice, humanity. Social media are the main resource for collecting people's opinion and their sentiment about a different trending topic. People use many abusing words in social media to express their emotion. Using sentiment analysis, we will build a platform where one can easily identify the opinions are either positive or negative or neutral. This research paper will contain supervised learning which is under the machine learning approach. We run an experiment on different queries from humanity to terrorism and find out an interesting result. First of all, we have preprocessed the dataset to convert unstructured airline review into structured review form. After that, we convert structured review into numerical value. We have to preprocess the data before using it. Stop word removal, @ removal, Hashtag removal, POS tagging, calculating sentiment score have done in preprocessing part. Then an algorithm has been applied to classify the opinion as either it is positive or negative. In this research paper we will briefly discuss supervised machine learning. Support vector machine as well as Naïve Bayes algorithm and compares their overall accuracy, precession, recall value. The result shows that in case of airline reviews Support vector machine gave way better result than Naïve Bayes algorithm.

TABLE OF CONTENTS CONTENTS Board of examiners Declaration Acknowledgements Abstract CHAPTER 1: Introduction 1.1 Introduction 1.2 Motivation 1.3 Rationale of the Study 1.4 Research Questions 1.5 Expected Output 1.6 Report Layout CHAPTER 2: Background 2.1 Introduction 2.2 Related Works 2.3 Research Summary 2.4 Scope of the Problem 2.5 Challenges PAGE i ii iii iv 1- 5 1 2 CHAPTER 3: Research Methodology 3.1 Introduction 3.2 Research Subject and Instrumentation 3.3 Data Collection Procedure 3.4 Statistical Analysis 3.5 Implementation Requirements CHAPTER 4: Experimental Result and Discussion 4.1 Introduction 4.2 Experimental Result 4.3 Descriptive Analysis 4.4 Summary CHAPTER 5: Summary, Conclusion, Recommendation and Implication for Future Research 5.1 Summary of the Study 5.2 Conclusions 5.3 Recommendations 5.4 Implication for Further Study APPENDIX REFERENCES CHAPTER 1 Introduction 1.1 Introduction Social Networking site is the best source to express the innovative and creative thought as well as the review. Different activities occur around us. People use social media to give their opinion on that. Social media is the largest platform and it is open and free for the users. This huge number of networking site contains different thoughts. People express their opinion, Their sentiment, their platform, their idea, their felling through this platform. The events that occur around us are such as politics, injustice, inhumanity, international affairs, economy, natural disaster, terrorism, and various upcoming topics. Social networking site in a place where people can give their personal opinion and think about those issues. Nowadays people can also protest about injustice using this networking platform. Suppose a soccer team can know how much support it has got on social media. But it is harder to ta analysis such a big amount of data using simple techniques and methods. So a new thing should be introduced to the analysis of the data. The process should be much more

reliable and efficient way than the previous. Again we can imagine a political leader want to know how much popular he or she among the people while his campaigning but it would be quite hard to know if we use general methods or techniques. A machine learns methods proposal could be applied. So that the political leader or the candidate will be able to know his popularity analyzing the sentiment of people. People use social networking site as a platform to express their emotions. Twitter is very popular among the social networking site. One can face many problems while using the sentiment analysis it could be like data scarcity and so on. This is occurring because of the slang language. There is the various number of challenges in Sentiment Analysis. An opinion word expressed by people might be positive in one situation. Again the same opinion word might turn into negative in another situation. Another challenge is that people use different expressing ways to give their opinions on various issues. Those expressing ways are not always the same. Sometimes it is challenging to understand because of its lack of context. For example, "The Game was as good as its previous Game". This entire sentence dependence upon the thoughts of the person about the previous version game. People use twitter for multiple purposes. Among the people around the world, twitter reported that it has a number of active users nearest 321 million monthly. People of all age use twitter. Twitter contains a huge number of information obtained from users. Twitter contains opinion from different part of view. The opinion is an expression of people is a particular subject. Different people have different mind and opinion. People share their opinion on twitter using the tweet and retweet. So we will consider Twitters tweet as our data set and production system that will able to analyze and gives us an accuracy as our output. In this research paper, we are using two types of machine learning algorithms and those are Support Vector Machine algorithm and Naïve Bayes algorithm. This will calculate accuracies, precisions and recall values. There will be three types of output in our result. Those are negative, neutral, positive using supervised learning and unsupervised learning.

1.2 Motivation: Social media is one of the most important communication medium today. Social media technology enables the message to be sent quickly, become widespread and even viral if the topic attracts public attention. Unfortunately, this also means that hate speech can also spread easily and quickly that it can lead to conflicts between groups in society. Hate speech especially concerning religion, nationality that we have seen in the recent past. There are several industries, companies they change or upgrade their marketing strategy on the basis of customer review of their service. Customer express their review on social media. Those companies run a query on how much they satisfy a customer and what are the things they need to apply to fulfil customer satisfaction.

1.4 Research Questions: Our research paper is based on review dataset of customer in a particular topic. Where we analysis customer review on airline service. Customer express their emotion through social networking site. Analyzing the customer review airline companies upgrade their services to provide customer satisfaction. Airline Company's analysis the negative review and according to that they change their service. While the positive review encourage the airline Company's to perform better.

1.5 Expected Outcome: Machine learning approach with prediction algorithm to solve the problem. Here we make an attempt to find out the positive, negative and neutral sentiment on Twitter data. Those positive, negative and neutral divided comment is used to analyze the quality of the airline. This project is divided into two main segments. Firstly, the main categories of the reviews were identified. Secondly, sentiment analysis is performed for the categories/aspect detected from the reviews. The result of the reviewed work will provide a robust decision support for the customers which will not only help them in making an effective decision for choosing their airlines but also help the airline companies to look after the

[areas of improvement. This will also help them to get a competitive edge over their other rivals/competitors in the airline industry.](#) CHAPTER 2

Background 2.1 Introduction In this new era, Judgment is based on public reviews. People set up their mind on something by just seeing some review of the product. So product review plays an important part for people judgment or decision making. And Internet is the source where people can get the expected reviews. A user can get knowledge about anything that is available on internet such as any place, country, food, restaurant, hotels and tourist place and so on. The best and for most popular way to get reviews is social networking site. Where other user give their opinion about the things. 2.2 Related Works This part of our paper will review different research paper works that are similar to ours. Sentiment analyze is the process that can classify human expression in different matters. There could be many sources for sentiment analysis. Text, image, video, audio, are the possible source for sentiment analysis. Using any of those sources analyzing sentiment is possible. But each of those sentiment analyzing source has different possible ways to approach. There are many authors who describe sentiment analyze. Some authors describe that people sentiment will detect automatically using the tweets from twitter. Firstly, get the data set from twitter. Then those data have to be extracted. Then those data have to be extracted. The next step in preprocessing the data. Some authors categorized tweets both as an objective and subjective category. These authors compose trained data instead of manually annotated data. Again they use a word tweet into its part of speech using POS tagging. A significant and must effective sentiment detection is approached by the author. Bhumika M.Jadav and Vimalkumar B.Vaghela [2] worked on a research paper on Sentiment Analysis. In this research paper, they researched on movie review dataset [2]. They converted unstructured movie review into structured form. Then, they calculated the score of structured word [2]. They gave the score as input to Support Vector Machine and produce output as either positive or negative. S.M. Mazharul Haque also published a research paper on sentiment analysis. They used big data and R platform to analyze the advanced level of data [3]. Again, the algorithm they used was the Naïve Bayes algorithm along with the Lexicon Based analyzer [3]. Thus, they find out an effective and fruitfully way to analyze the sentiment of people. A research paper from Sharvil Saha K Kumar was published on sentiment analysis. They first retrial the data before they began preprocessing the data. They used hashtag classification to find out the trending topic and the most used tweet and re-tweet [10]. Finally, the heart of the paper polarity classifier is used. They use Naïve Bayes classified Bigram models for their classification of polar data [10]. Sanjay Chakraborty worked on a research paper on Sentiment Analysis. In their research paper, they researched on movie and hotel review [4]. [Naïve Bayes and K-NN classifier](#) are used [on](#) this research paper [4]. Finally, they arrive at an accuracy of 80% on movie review, while the accuracy rate for hotel review was lower. Raheesa Safrin published a research paper on sentiment analysis. They researched on product review based on the customer comments on their own website [1]. In the website, customers give their opinion in the comment section. The comments are of two types, one being positive and the other negative. The positive tagging inspired the company and extracted the most effective characteristic to get the best result. They used K-NN algorithm. K-NN is a technique of classification and K-means is a clustering technique [1]. Zohreh Madhoushi et al [12] worked on a research paper on Sentiment Analysis Techniques in recent words. They used three types of techniques to classify the sentence. Firstly they used machine learning approaches which contain both [Support vector machine and Naïve Bayes classification](#). SVM analyze [and](#) recognize data patterns while Naïve Bayes calculates maximum likelihood. Secondly they used lexicon-based approach which is used for parts

of speech tagging and WordNet. Finally they used hybrid approaches which contains both machine learning and lexical resource. That's how they find an effective way to analysis sentiment of people. Divya Sehagal et al [13] worked on [real-time Sentiment Analysis of big data](#) application [using](#) the help of [twitter data](#). In this paper, they analyze and process big data. They used Hadoop technology for analyzing and preprocessing. They also stream the data API from twitter. They split the root word and remove the unwanted and unnecessary extra storage of derived words. All those steps are necessary to analysis Big Data Application using twitter data.

2.3 Research Summary

In this paper we are analyzing the review dataset of airline. Those dataset of airline. Those dataset contains both positive, negative, neutral reviews about the airline companies. [There are both supervised and unsupervised approaches](#) applied [in](#) this paper. [The](#) research paper will contain supervised learning which is under machine learning algorithm. Here the unstructured airline reviews are converted into a structured form. We also preprocess the data before using it. Stop word removal, @ removal, Hashtag removal, removal URL, POS tagging have done in preprocessing part. Finally the algorithm is applied to produce an output. Both Support Vector Machine and Naïve Bayes Algorithm is applied for a better outcome.

2.4 Scope of the Problem

There will be a lot of problems in our research paper. [There are many types of sentiment problems](#) analyzing [text](#) review, exact [meaning of the](#) review, expressions analyzing [and in the evaluation of sentiment analysis](#). In this research paper, [we](#) will briefly [discuss](#) different machine learning prediction algorithms. Many people are same review to find out the sentence is in which class level, we have faced the problem the similarity and dissimilarity between the customer review. The sentiment problems evaluation and detection [polarities for customer reviews and find the most effected algorithms solutions for the highest accuracy for text](#).

2.5 Challenges

There are several challenges of our research paper, first of all fake detection and spam of the user's emotions. Example Review1: "The airline is good" and Review2: "The airline is not good". In this example, the first review [sentiment score \[good\] with positive](#) expressions of customers [and the second review](#) sentiment score [not good] with negative expressions of customers. When we evaluate the [sentiment score for determine](#) airline [the first](#) expression [will be useful](#) and [the second](#) expression [will be spam](#) and fake detection. There are many similar review in our dataset. There are most difficult to show the sentiment scores: positivity, negativity, objectivity. Another problem of temporal review when the airplane services is good then the customers review is positive, unfortunately some of the reason this airline service services is not good then the customers review is negative. This types of sentiments that [are changed with time may](#) enhance [the](#) sentiment [analysis performance](#).

CHAPTER 3 Research Methodology Introduction: There are already many works that were done according to this topic.

This research field is the mostly used and popular field to work on. Natural Language Processing is the area where all those research were done. Sentiment analysis is the mostly used topic under [Natural Language Processing](#) area. [Sentiment analysis](#) is basically [opinion mining of](#) people. In this research we will apply necessary algorithm to calculate sentiment and produce a good accuracy.

A. Data Collection procedure

Executing the research paper, the Data Set that is considered here was Airline Reviews, those tweets were collected from the Twitter platform. We collected around 14000 Tweets from the twitter. Where positive and negative reviews of Airline were given. We analysis those reviews to get our result.

METHODOLOGY: A. DATA PREPROCESSING

Twitter is the data source for our research. There are several reasons that are why we choose twitter data. According to Twitter, it has almost around 200 million active users. Those users' tweets almost 500 million per day report says. There is also an important reason to choose

twitter. We could collect data from various social media like Facebook, Instagram etc. Those social media might contain videos, pictures, many web sharing link and so on. But twitter is all about text. Twitter allows the user to publish text with a limitation of 140 characters. Besides the positive side, it also has a negative side. Because twitter contains many hashtagging, slangs etc. Again twitter contains URL links in the texts.

A. '@' Remove: The first thing we should do is to remove '@' as well as the URLs. Again '@' mention a username. It is used to tag something. So the total word that starts with '@' could be removed.

B. Hashtag Remove data: Twitter text contains hashtag '#'. The hashtag is responsible to trend a topic. Hashtag might contain some slugger. Example: #we_want_justice. So we decide to remove the '#' but keep the entire word after the hashtag.

C. Stop words Remove: The major part of the data preprocessing is stopping words removal means to filter out useless data from the text. Stop words is known as the removal of unnecessary data. [Natural language toolkit in python has a list of stop words](#) in 16 different languages. Example "I love eating, So I eat". Here after stopping words this sentence keyword will look like, "love", "eating", "eat".

D. Stemming: Stemming is a basic part of text preprocessing module. Stemming is basically the operation of transforming the targeted words to its source form. It reduces the word into their base form. Stemming is a technique that is used to preprocess the data. It converts word like "plays", "playing", "played", "player" to its root or base word form which is "play". By using a stemmer, the word that is converted into the root form will be a unique word before further processing.

Sentiment analysis is a famous research area in NLP. So, previously many research works have done based on different kinds of methodology. But the study of the previous result said that [machine learning algorithm such as SVM and Naive Bayes](#) provides a [better result](#). In our working process firstly we collect the dataset using twitter API. Then preprocess the review text for the trained model. Finally, check the result of the working model of this research. This section will be discussed by applying the methodology for this research work.

Figure1: Data distribution of the review dataset B. Proposed Work The motive of the research paper is to analyze data with both supervised learnings which is under the machine learning approach. A graphical representation of this sentiment analysis. In this flow helps to understand the whole working process of this research work. The short view of the whole working flow is given below.

Twitter API
Data preprocessing
Stemming
Remove url
Remove @tag
Stop word
hashtags
remove
ML Algorithms
Result

Figure2: Workflow of Sentiment Analysis

The step we followed to execute this research paper firstly we collect our datasets from twitter. It's an airline review of the passengers. Passengers give their opinion on a different perspective. We use natural language toolkit (NLTK) for this research paper. We also use Naïve Bayes and SVM for these methods.

i. [Naïve Bayes](#) Approach: [Naive Bayes is a collection of classification algorithms](#) which are [based on Bayes Theorem](#). Naive Bayes classifier gives us an excellent result when one uses it for text data analysis. Such as Natural Language Processing. Naive Bayes algorithm gives us a probability analyzing the data set we have given. [Naive Bayes classifier is used as a probabilistic classifier](#). To perform [the](#) classifier, it uses the concepts of mixture models. A mixture model is capable of establishing the probability of the component that it consists of Bayes theorem to perform as a probabilistic classifier. Another name that a naïve Bayes is known as simple Bayes or independence Bayes. The probability P is defined as follows: $P(m|m) = \frac{P(m)P(m|m)}{P(m)}$

(i) Above, $P(m|m)$ is the probability of class x. Where x is the target and predictor is the attribute. $P(m)$ is the prior probability of class. $P(n|m)$ is the probability of predictor of the given class. $P(n)$ is the prior probability of predictor.

ii. [Support Vector Machine](#): [Support Vector Machine is a](#) universal learner. Support Vector Machine has defined both input and output format.

The output is either positive or negative and input is vector space. The text document is not suitable for learning. Those texts are transformed into a structured format. The text is transformed into a format which matches into the input of machine learning algorithm. The score of the texts are calculated and then the score is given as input to [Support Vector Machine](#). [Support Vector Machine](#) has been proved [one of the most powerful learning algorithms](#) for text Categorization. But text categorization sometime may produce occurs. To decide which one is better between texts a comparison of text classifier is required. The performance measure is used in this case.

[CHAPTER 4 Experimental Result and Discussion 4.1 Introduction](#) In this section [we have discussed about our experimental result](#) about sentiment analysis on airline review. In this paper we are working on airline sentiment [review, our main features was to develop a system that provide higher accuracy and easy to implement](#) in machine learning algorithm. The [result of the airline sentiment reviewed work will provide a best decision support for the customers choosing their airlines](#) and it [also helps the airline companies to improve their airline services](#). [4.2 Experimental Result](#) [We have collected data](#) to twitter [and preprocessed the data according to our interest](#). [We have used](#) 14000 text airline review where 9000 is negative, 3000 is positive and 2000 positive review. We have used 10000 review from training the model and 4000 for testing. Predicted Data \ Actual Data [\(Positive\) \(Negative\)](#). [\(Positive\) True Positive False Positive \(Negative\) False Negative True Negative](#)

? Accuracy $(A) = \frac{TP + (TN + TF + FN)}{TP + TN + FP + FN}$? AUC (Area under the Curve) = 12 . $TPT + -FPN + TNT + NFP$? Precision = $\frac{TPT + PFP}{TPT + PFP}$? Recall = $\frac{TPT + PFN}{TPT + PFN}$? F1 Measure = $\frac{Precision \times Recall \times 2}{Precision + Recall}$

Before the experiment result set the train sample size 67% of the whole dataset and set the random state is 40. Then used the SVM and naive bays algorithm for result prediction. For SVM we used SVC linear kernel and [for Naïve Bayes](#) used [multinomial Naive Bayes](#). The classification report of both algorithms is given in table1. Support Vector Machine give accurate result than Naïve Bayes algorithm. After the train, we test both algorithm prediction by the review twitter and define the best performer. The classification report of both algorithms is given in table1. SVM Accuracy 82.48 Precision 90.33 Recall 81.79 F1 Measure 85.85 Table 1: Accuracy of SVM and Naïve Bays Algorithm. Naïve Bayes 76.56 89.00 83.75 86.37 In table 1 displays the result comparison of both applying algorithms. The accuracy of SVM is 82.48 where Naïve Bayes accuracy is 76.56. Precision and Recall value for SVM is 90.33 and 81.79 also 89.00 and 83.75 for Naïve Bayes. Now in figure3 shows the comparison of both algorithms. Where SVM provides almost 83% accuracy and Naïve Bayes provide 77% accuracy in this experiment. 90 82.48 80 76.56 70 60 50 40 30 20 10 0 SVM Naïve Bayes

Figure 3: Accuracy Comparison of Both Algorithm After that, check the model output using the review from the dataset and predict the output using both algorithms. Check the result of model performance we test the prediction of the model by using the review. Take the review as input from the user and then classify review using both algorithms. Maximum time the SVM classifier gives the accurate result where Naïve Bayes is not. In table2 shows the predicted output result for providing a review. Review Text: "flying with @united is always a great experience". Here, 1 contain negative review and 0 contain positive review. SVM Prediction Naïve Bayes Prediction Prediction 0 1 Actual Sentiment Positive Review Positive Review Response Sentiment Positive Review Negative Review Table2: Result prediction of both Algorithms. Summary: We have tried all the possible ways to get the best possible result. We have applied two different algorithm and compared between them to get the highest possible accuracy. We have divided the sentence into positive, negative and neutral. We also calculate the precision, recall, F1 Measure and accuracy and applied it. Comparison to other related model about sentiment analysis, we were possible to achieve good rate of accuracy. [Chapter 5](#)

Summary, Conclusion, Recommendation and Implication for Future Research

Future Scope: We research on a specific platform for sentiment analysis system. Through our research results a good output but it has some limitations too. This research could be done better if some more algorithm applied. We compare two algorithm and results the maximum output. But if some more algorithm could applied we would be able to compare all the results and use the best accuracy for final output. KNN is an algorithm that can be introduced. Applying all those three algorithm will take this research to another level. We are confident that our research is benefits to all mankind. It will make a revolution change to many industries and companies as they will get to know what people think about their companies and industries.

Conclusion: In this paper, we use machine learning algorithms of Naïve Bayes and Support Vector Machine for sentiment classification of airplane reviews.

In this paper we used the perform measurement of people's sentiment analysis. This paper is used to perform evaluation measure on comments obtained from the passengers. As an example, a positive sentiment is "Good flight" and a negative sentiment is "Late flight!" In this paper, we take an attempt to find out the positive and negative sentiments on Twitter data. The separation of positive and negative comments is used to analyze the quality of the airplane. The result of the reviewed work will provide a robust decision support for the customers which will not only help them in making an effective decision for choosing their airlines but also help the airline companies to look after the areas of improvement.

REFERENCES [1] Raheesa Safrin, K. R. Sharmila, T.S. Shri Subangi, and E.A.Vimal "Sentiment Analysis on online PRODUCT review" International Research Journal of Engineering and Technology (IRJET) Volume: 04 Issue: 04 Apr -2017. [2] Vimalkumar B. Vaghela and Bhumika M. Jadav "Sentiment Analysis using Support Vector Machine based on Feature Selection and Semantic Analysis" International Journal of Computer Applications (0975 – 8887) Volume 146 – No.13, July 2016. [3] S. M. Mazharul Hoque Chowdhury, Priyanka Ghosh, Sheikh Abujar, Most. Arina Afrin, and Syed Akhter Hossain "Sentiment Prediction Based on Lexical Analysis Using Deep Learning" Springer Nature Singapore Pte Ltd. 2019 A. Abraham et al. (eds.), Emerging Technologies in Data Mining and Information Security, Advances in Intelligent Systems and Computing 814 [4] Lopamudra Dey, Sanjay Chakraborty, Lopamudra Dey, Beepa Bose and Sweta Tiwari "Sentiment Analysis of Review Datasets Using Naïve Bayes' and K-NN Classifier" Information Engineering and Electronic Business, 2016, 4, 54-62, July 2016. [5] Ali Selamat Faculty and Nurulhuda Zainuddin "Sentiment Analysis Using Support Vector Machine" 2014 IEEE 2014 International Conference on Computer, Communication, and Control Technology (I4CT 2014), September 2 - 4, 2014 - Langkawi, Kedah, Malaysia 978. [6] Ms. Gaurangi Patil, Ms. Varsha Galande, Mr. Vedant Kekan, and Ms. Kalpana Dange "Sentiment Analysis Using Support Vector Machine" International Journal of Innovative Research in Computer and Communication Engineering, (An ISO 3297: 2007 Certified Organization), Vol. 2, Issue 1, January 2014. [7] Geetika Gautam, Divakar Yadav "Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis" [8] Yan Luo ,Wei Huang "Product Review Information Extraction Based on Adjective Opinion Words" 2011 Fourth International Joint Conference on Computational Sciences and Optimization IEEE computer society. [9] Bin Ren, Lianglun Cheng "Research of Classification System based on Naive Bayes and MetaClass" 2009 Second International Conference on Information and Computing Science IEEE computer society. [10] Shah, Sharvil, Kumar, K.; Saravanaguru, Ra. K "Sentiment Analysis of Twitter data using classifier algorithm" International Journal of Electrical & Computer Engineering (2088-8708). Feb2016, Vol. 6 Issue 1, p357-366. 10p. [11]. M.A. Hearst, "Support vector machines," IEEE Intelligent Systems, pp. 18-28, 1998. [12] Zohreh

Madhoushi, Abdul Razak Hamdan, Suhaila Zainudin, "Sentiment Analysis Techniques in Recent Works ", Science and Information Conference 2015 July 28-30, 2015 London, UK. [13] Divya Sehgal , Ambuj Kumar Agarwal, "Real-time Sentiment Analysis of Big Data Applications Using Twitter Data with Hadoop Framework", Springer Nature Singapore Pte Ltd. 2018, Soft Computing: Theories and Applications, Advances in Intelligent Systems and Computing 584