Tribhuvan University

Institute of Science and Technology

Seminar Report

On

"Comparative Analysis between Logistic Regression, SVM and Naive Bayes for Twitter Sentiment Analysis"

Submitted to

Central Department of Computer Science and Information Technology
Tribhuvan University, Kirtipur
Kathmandu, Nepal

In the partial fulfilment of the requirement for Master's Degree in Masters in Information Technology (MIT) Second Semester

Submitted by

Pralhad Khanal
Roll No. 7925018

June, 2024

Tribhuvan University

Institute of Science and Technology

SUPERVISOR'S RECOMMENDATION

I hereby recommend that this seminar report is prepared under my supervision, by Pralhad Khanal entitled as "**Comparative Analysis between Logistic Regression, SVM and Naive Bayes for Twitter Sentiment Analysis**" in partial fulfillment of the requirement for Master's Degree in Masters in Information Technology (MIT) be prepared for evaluation.

Supervisor

Assoc. Prof. Mr. Nawaraj Paudel

Central Department of Computer Science and Information Technology

LETTER OF APPROVAL

The undersigned certify that they have read, recommended and accepted to the seminar topic entitled as "**Comparative Analysis between Logistic Regression, SVM and Naive Bayes for Twitter Sentiment Analysis**" in partial fulfillment of the requirement for Master's Degree in Masters in Information Technology (MIT).

Evaluation Committee

……………..……………………………..

Asst. Prof. Sarbin Sayami

(H.O.D)

Central Department of Computer Science

And Information Technology

………..……………………………………

Assoc. Prof. Nawaraj Paudel

(Supervisor)

Central Department of Computer Science

and Information Technology

……………………………

(Internal)

Acknowledgement

It is with great pride and satisfaction after a prolonged tactics and assistance of different people I present a seminar report under the "**Comparative Analysis between Logistic Regression, SVM and Naive Bayes for Twitter Sentiment Analysis**", during second semester, in partial fulfillment of the requirement for Master's Degree in Masters in Information Technology (MIT).

I am thankful to Assoc. Prof. Nawaraj Paudel for being supervisor during this endeavor. He has been the cornerstone of our project and has guided me during periods of doubt and uncertainties. His idea and inspirations have helped me making this nascent idea of topic into a full- fledge seminar.

I am also thankful to **Asst. Prof. Sarbin Sayami**, HOD of Central Department of Computer Science and Information Technology for his constant support throughout the period.

Pralhad Khanal (7925018)

ABSTRACT

The paper is basically prepared on twitter sentiment analysis using different classification algorithms based on the twitter user’s views. The study is based on the comparative analysis of different algorithms namely logistic Regression, Naive Bayes and Support Vector Machine using bag of words and tf-idf features which is verified with obtaining accuracy, precision, recall and f1 score. In order to obtain the desired outcome this project includes data preprocessing steps followed by feature extraction and model building, which classifies the tweets into positive and negative sentiment. On initiating the project various different other classifiers are also analyzed and ongoing with the factors like: accuracy, precision, recall and f1 score, logistic regression is considered as the best and this project is device with the same. The main aim of the project is to assist consumer on classifying the products based on positive and negative sentiments and make a purchase decision.

Keywords: Sentiment, logistic regression, Naive Bayes, Support Vector Machine, bag of words, features, accuracy, precision, recall, f1 score, classifier.

Contents

LIST OF FIGURES

LISTS OF TABLES

ABBREVIATIONS

AI: Artificial Intelligence

BoW: Bag of Words

ML: Machine Learning

NLP: Natural Language Processing

NLTK: Natural Language Toolkit

SVM: Support Vector Machine

TF-IDF : Term Frequency-Inverse Document Frequency

URL: Uniform Resource Locator

CHAPTER ONE

INTRODUCTION

1.1 Introduction

Having been born in an era of technology most people try to express their views on social media sites in order to make other people decide based on their view or some people express their views in order to make the organization address what they like. Similarly there are various different social media sites in these days where each of them have their own features and users segment. Unlike all other twitter tends to be different and possess unique user's count, where each of them expresses their views based on scenario and their perspective for betterment of self, any other organizations, political views on current scenario, etc. Twitter is a micro blogging platform incorporated with different views and each of the views can have different sentiments like: positive, negative and neutral.

Sentiment analysis dives into the realm that explores emotions reactions and the choices shaped by them. In the vast realms of social media analytics along with web and data mining analytics messages play a crucial role. It is widely believed that feelings are the key elements in assessing human actions. Sentiment analysis can be described as a tool that helps in finding out the "positive" or "negative" from each sets of datasets. Sentiment analysis can ascertain the goodness of the information in definite manner. The sentiment analysis is modeled for two classifications: positive and negative which is based on three: data preprocessing, feature extraction and modelling of classifiers.

Sentiment analysis on twitter data set plays a crucial role in decision making of the customers for their product purchasing ability along with the organization wanting to take an insight in order to explore more of the customers need and their customization need so that the organization explores more to cope in the surrounding. This seminar project is based on the dataset where twitter dataset are downloaded and analyzed based on the classifiers with eliminating necessary words, stop word, urls, etc.

1.1.1 Defining Sentiment

Sentiment is the expression of the feelings on the product, topic or the objects. Sentiments can be positive, negative or neutral based on each piece of topic, object. Various political parties are using these sentiments as a tool to enhance their election campaign. Alongside, not only political parties but a majority of decision in a large scale industries are made based on the sentiments of the viewers, their posts and engagement.

Table 1.1: Example Tweets

Sentiment	**Tweet**
Positive	A great place to get news.
Negative	A worst place to get news.

1.1.2. Characteristic of Tweets

Those characteristics of tweets are:

Length: The maximum number of characters a twitter tweet consists of is 4000 characters.

Data availability: With various forms and contents of data, the available format of data is easy to obtained but with certain validations and necessary parameter feedings.

Language model: Large language are deviced and stored with necessary parameters.

Domain: Data are grouped and characterized based on the user domain.

Username: Twitter handle consists of username starting with "@".

1.2 Problem Statement

Various different users provides their views on products, objects or circumstances like: political, religious, etc. With majority of data are unstructured and requires some sorts of structuring for proper decision making. When the unstructured data are pre- processed using different tools and mechanisms so that these structured resultant data will overcome the problem of classifying the proper set of data, i.e twitter data into positive or negative which will help in product development to analyzing insight about products or services. The major problem statements are:

This seminar project is based on the dataset where twitter dataset are trained, tested and analyzed based on the classifiers like: logistic Regression, Naive Bayes and Support Vector Machine with eliminating necessary words, stop word, urls, etc.

1.3 Objective

The major objectives of this seminar report is to:

- To compare the performance of logistic Regression, Naive Bayes and Support Vector Machine.

CHAPTER TWO

BACKGROUND STUDY and LITERATURE REVIEW

2.1 Background Study

With the advancement of technology, various different tasks like speech to text conversion, text annotations, image to text conversion, one language to next language conversion, sentiment analysis, etc. are performed due to use of Natural Language Processing (NLP) which is a subset of Artificial Intelligence (AI). With the AI and NLP gaps are bridged with tremendous use of technology to communicate between human and computers. Today's innovation and progress in the field of healthcare, business, etc. came from the advancement of Machine Learning (ML) and AI. Analyzing the sentiments is the major subject of concern in today's age to overcome the users views required to mitigate the negative comments and increment the positive sentiments towards product to get business reach the destined. Not only the business but to all those product business whose product is based on individual requires sentiments of the users to be analyzed for future progress and overcoming the shortcomings.

Among the various different methods, tools and device of analyzing the twitter data this project tries to deviced a ML method of classifying the sentiments of twitter data from various different users and compare which of the deviced classifier ML algorithms provides better result.

2.1.1 Sentiment Analysis

The availability of sentiment analysis has made it possible for businesses with huge amounts of unstructured data to analyze and take out important meanings from them in a fast and efficient way. With the large volume of text created by customers on digital channels, human teams can be easily flooded with information. Strong AI-enhanced cloud-based customer sentiment analysis tools help organizations scale their business intelligence delivery from customer data without using unnecessary resources. Sentiment analysis is a natural language processing (NLP) technique which involves the automated analysis of textual data to determine the sentiment expressed within it.

Sentiment analysis is used in a variety of applications, including social media monitoring, customer feedback analysis, and market research. By using ML models, sentiment analysis categorizes a text into positive or negative sentiments based on the underlying sentiments expressed by words, phrases, or the entire text. This is highly beneficial for businesses and organizations as it enables them to make informed decisions based on public sentiment, enhance customer experiences, and adapt to emerging trends in the digital landscape. Sentiment analysis is widely used in areas like market research, customer feedback analysis, social media monitoring, and brand reputation management. It's an important and foremost tool which provides valuable insights from the vast amount of textual data and application of it can make a positive change in an organization, its products and overall 360 degree of any organizations.

2.2 Literature Review

Sentiment analysis is a trending field of NLP where various different classifications, regressions algorithms are deviced to analyze their polarity. Among various different classification algorithms each of the classifications are analyzed individually but not all of them are analyzed in any of the previous report. There are various different numbers of research carried out for sentiment analysis and various different authors have different opinions regarding each of the classification algorithms and ML projects on sentiment analysis.

Various technique for sentiment analysis for the twitter message. They compare the different classification algorithm like Naive Bayes classification, Baseline and Support Vector Machine. They also consider the different factor affecting the sentiment like unigrams, bigrams, Part of speech (POS) etc. They achieved accuracy of above 83% for all three algorithm using unigrams + bigrams [1].

Technique for sentiment analysis for the movies review. They compare the different classification algorithm like Naive Bayes classification, Maximum Entropy classification and Support Vector Machine. They also consider the different factor affecting the sentiment like unigrams, bigrams, Part of speech (POS) etc. They achieved accuracy of above 80% for all three algorithm using unigrams + bigrams [2].

Mainly the paper is based on the Naive Bayes classifier where they take the baseline for their research. They display the result on pie chart for positive, negative and neutral for the specific keyword [3].

Topic based classification based on the Logistic Regression. They also have used the confusion matrix as a classifier model. They achieved the accuracy of 92% for the tweets classification into selected topics [4].

People are used to check something before they do it- like checking it films, pubs, shopping online and more. The present work is followed by the concept of the Twitter Sentiment Analysis, whereby the person who thinks of every particular tweet they provided is identified.

They used natural language processing (NLP) has been used in current work. They found that further research work and decision making may be used. They also used Random forest and Decision tree, but they provide more precision than the SVM techniques [5].

Machine learning classifier required the trained data set to work. For this project 10, 00,000 data sets were used.

CHAPTER THREE

METHODOLOGY

3.1 Introduction

Methodology refers to the way how the process related to the projects are serialized and sequenced and what sorts of algorithms and tools are used, are properly described. Sentiment Analysis of twitter data uses different classifiers i.e. logistic Regression, Naive Bayes and Support Vector Machine using bag of words and tf-idf features. The detail architecture of the system is presented below.

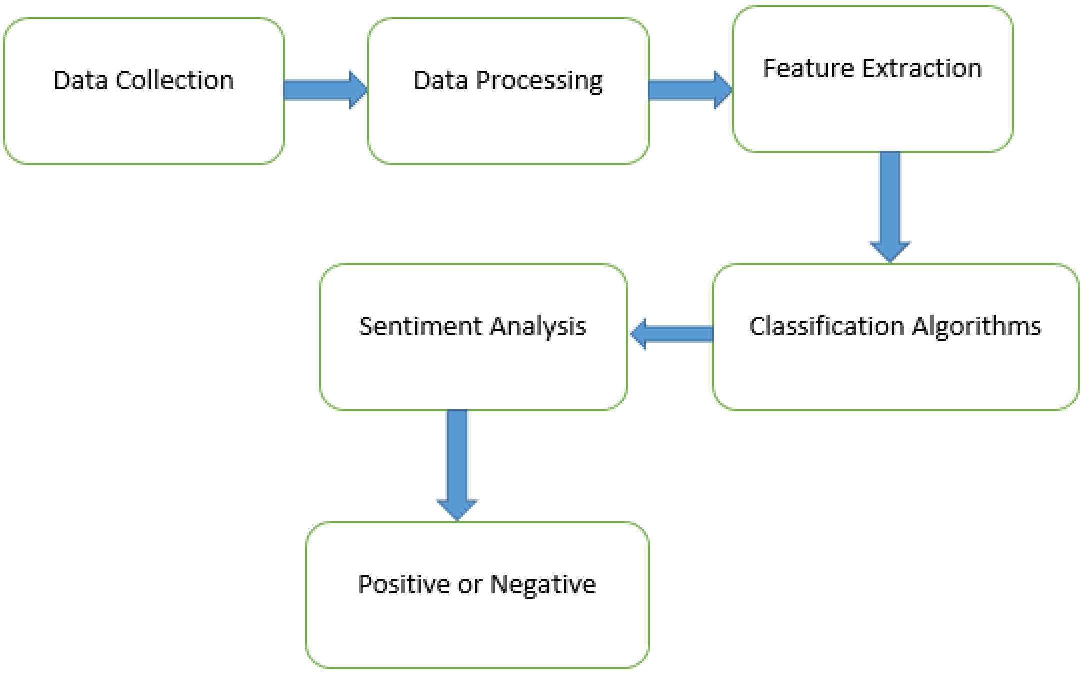
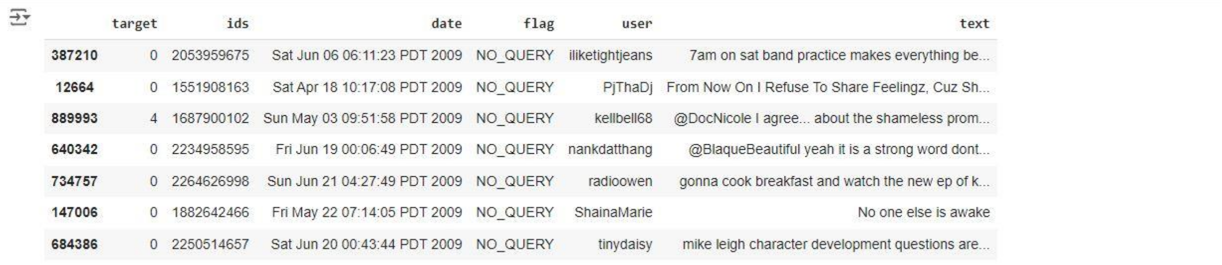


Figure 3.1: System Architecture

3.2 Data Set Description

The dataset used for Sentiment Analysis of Twitter Data using logistic Regression, Naive Bayes and Support Vector Machine using bag of words and tf-idf features is dataset which is collected from Kaggle, a repository known for its diverse collection of data. Kaggle datasets are online sourced freely obtained dataset, with not much elimination of words, usernames, urls, etc. The dataset consists of 10, 00,000 training data in csv format and is labeled as 0 for negative and 1 for positive, previously it was labeled as 0 and 4 accordingly. So, this project uses only the positive and negative datasets. The data sample dataset used in this project is shown below:



	target	ids	date	flag	user	text
387210	0	2053959675	Sat Jun 06 06:11:23 PDT 2009	NO_QUERY	iliketightjeans	7am on sat band practice makes everything be...
12664	0	1551908163	Sat Apr 18 10:17:08 PDT 2009	NO_QUERY	PjThaDj	From Now On I Refuse To Share Feelingz, Cuz Sh...
889993	4	1687900102	Sun May 03 09:51:58 PDT 2009	NO_QUERY	kellbell68	@DocNicole I agree... about the shameless prom...
640342	0	2234958595	Fri Jun 19 00:06:49 PDT 2009	NO_QUERY	nankdatthang	@BlaqueBeautiful yeah it is a strong word dont...
734757	0	2264626998	Sun Jun 21 04:27:49 PDT 2009	NO_QUERY	radioowen	gonna cook breakfast and watch the new ep of k...
147006	0	1882642466	Fri May 22 07:14:05 PDT 2009	NO_QUERY	ShainaMarie	No one else is awake
684386	0	2250514657	Sat Jun 20 00:43:44 PDT 2009	NO_QUERY	tinydaisy	mike leigh character development questions are...

Figure 3.2: Twitter Sentiment Analysis Dataset

3.3 Data Pre-processing

Pre-processing is the processing of the obtained data in order to get insight and eliminate redundant features and empty spaces along with different symbols, numbers, etc. as per requirement. Various different pre-processing stages includes various different types of techniques to eliminate the non-required properties. Thus removing of stop words, urls, usernames, etc. are placed under preprocessing stages. Data pre-processing includes various step.

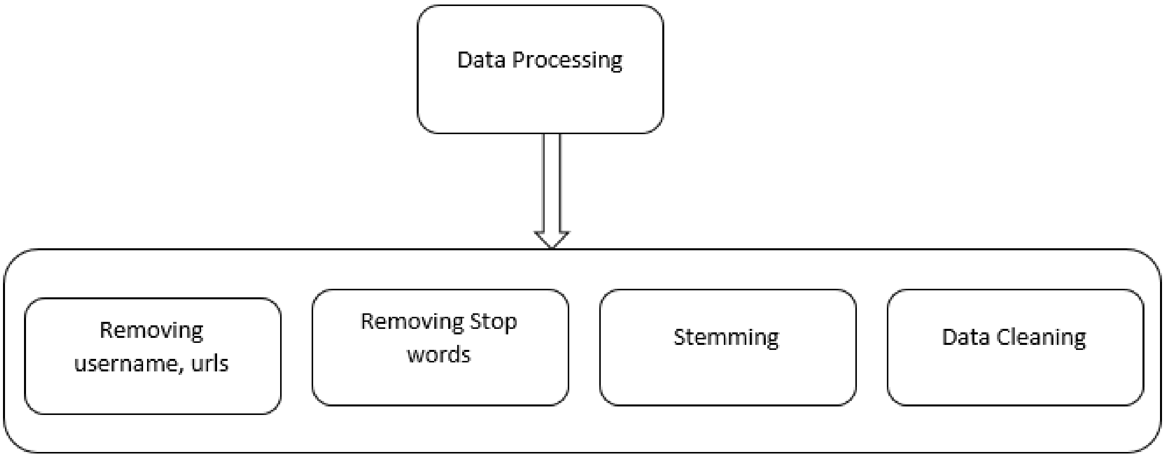


Fig: 3.3 Steps in Pre-Processing

1. Usernames: Username refers to the name of the twitter user. Twitter username starts with @ symbol and during the pre-processing stage it is replaced by removing @ symbol in data sets as it has no role during sentiment analysis rather it can create obstacle.

2. URL: URL, termed as Unified Resource Locator is the address of a specific webpage or file that locates to the particular set of data. Within the twitter dataset url has no such meaning of itself thus they are removed as a noise.

3. Stop Words: Stop word are those filler words which are not useful in for sentiment as they have no such meaning of themselves. These words includes most repeated word like a, an, the, for, etc. These words does not give any sentiment hence they are filtered out form the datasets.

4. Removing Hash-tags: Hash-tag in a tweet is used to emphasize a particular word or sentence, for example #thisisrain. Removal of these hash-tags is important because these hash-tags do not define any sentiment. Thus pre-processing is done and hash-tags before any word are removed.

5. Repeated Letters: Tweets contain the very causal language so the word such as happyyyy is replaced with actual word happy. The letter repeated more time reduced to the one.

6. Stemming: Change a word in the text into its base term or root term. Example, happiness to happy. This helps in reduction of dimensionality for feature set. This seminar used Porter. Porter Stemmer eliminates prefixes and suffixes from words, transforming them into their fundamental or root form, however, its applications are only limited to English words.

7. Data Cleaning Digits and special characters present in tweets doesn’t show any sentiment. In some cases they were mixed with words. Removal of those words helps in association of two words. Otherwise that words are considered as different.

3.4 Feature Extraction

Feature extraction refers to the conversion of texts into the feature vectors. With the reduction of dimension of data the accuracy of the prediction comes to an definite result, thus the feature extraction helps in mapping of the feature vectors for their accuracy. After pre-processing the tweets, we get features which have equal weights. Various different feature extraction methods are used:

3.4.1 Bag of Words

Bag of words represents the frequency of a word occurrence in the text as a feature. It is used to extract features from text. BoW ignores word order, its semantic meaning, and its grammatical structure. Instead of counting the frequency of the individual words, BoW can apply a slightly more sophisticated approach with N-grams, adding a meaningful meaning to a feature. It may give a high score to a word that appeared frequently in a document, while it may not contain relevant information to the document.

3.4.2 TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF), is one of the generally used techniques for feature extraction in text processing. It consists of two approaches: Term Frequency (TF), the frequency of occurrence of a feature term in the text set and Inverse Document Frequency (IDF) measures a term's importance within a document. It solves the problem of BoW.

3.5 Machine Learning Classifiers

3.5.1 Logistic Regression

The Logistic Regression algorithm is used as a predictive analysis model based on binary classification. It classifies the given tweets based on the probability to that particular class. In order to model the tweets into positive and negative from a large set of data. This seminar have used label dataset with probability value as 1 for the positive tweets and 0 for negative one with removing necessary outliers.

Logistic Regression is a discriminative model which means computing probability P(y|x) by discriminating among the different possible values of the class y based the given input x. The equation for this is as shown below:

$$P(C|x) = \sum_{i=1}^{N} w_i . f_i$$

In order to obtain a value of probability, P(y|x) which lies is in between 0 and 1, exponent function are described as:

$$P(C|x) = \frac{1}{z} exp \sum_{i} w_i . f_i$$

In order to normalization and specify the number of features logistic regression uses the mentioned expression:

$$P(C|x) = \frac{\exp(\sum_{i=1}^{N} w_i \cdot f_i)}{\sum_c \exp(\sum_{i=1}^{N} w_i \cdot f_i)}$$

Probability of x being y of class c are modelled in logistic regression as:

$$P(C|x) = \frac{\exp(\sum_{i=1}^{N} w_i \cdot f_i\,(c,x))}{\sum_{c' \in C} \exp(\sum_{i=1}^{N} w_i \cdot f_{i(c',x)})}$$

3.5.2 SVM classification algorithm

SVM is a supervised ML classification algorithm that draws a decision boundary line called a hyperplane to separate and classify data. In sentiment analysis, SVM is used to classify text, for example, into positive, negative, or neutral sentiment. It excels at both predicting continuous values (regression) and distinguishing between discrete categories (classification). The hyperplane is chosen to maximize the margin between the two classes, the distance between the hyperplane, and the closest data points from each class. SVM is a popular choice for sentiment analysis because they are relatively simple to implement and can achieve high accuracy on a variety of datasets. Support vector machines (SVMs) have been shown to be highly effective at traditional text categorization, generally outperforming Naive Bayes.

3.5.3 Naïve Bayes classifier

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles. The classifier works similar to the Multinomial classifier, but the predictor variables are the independent Booleans variables. Such as if a particular word is present or not in a document. This model is also famous for document classification tasks.

CHAPTER FOUR

IMPLEMENTATION AND TESTING

The implementation section explains the twitter sentiment analysis using classification of different algorithms namely logistic Regression, Naive Bayes and Support Vector Machine using bag of words and tf-idf features, is a data based analysis where each of the algorithm calculates the necessary precision, accuracy and f1 score and all those factors are analyzed and among them the best one will be chosen.

All the algorithms for the application are written in Python. Algorithms used in Sentiment Analysis of Twitter Data Using Logistic Regression, Naive Bayes and Support Vector Machine is Predictive analysis model. Logistic regression, Naive Bayes and Support Vector Machine is used to describe data and to explain the relationship between one dependent binary variable and one or more independent variables. The algorithm is implemented using python programming language.

4.1 Description of Major Function

The major function in the application are:

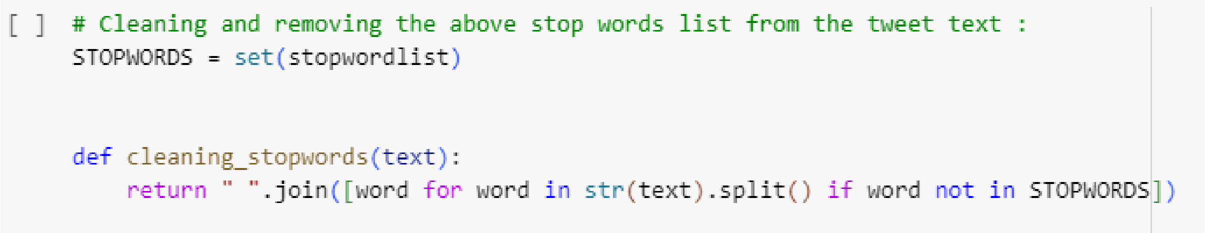
4.1.1 Preprocessing

This is the function which is run for processing the tweets.

Input: It takes the inputs as tweet data.

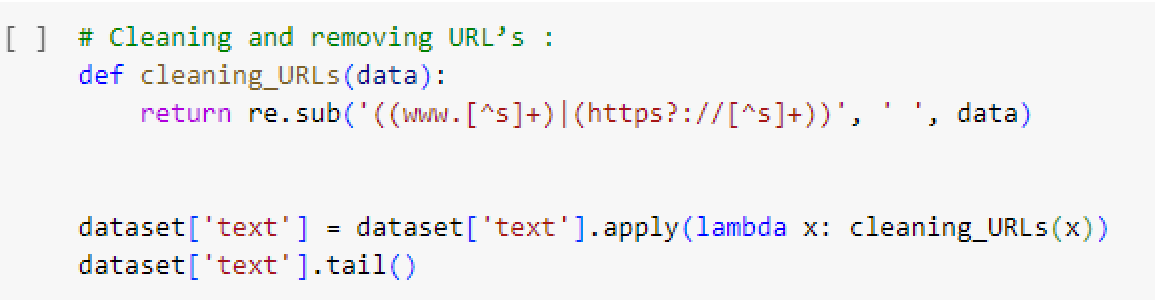
Pre-Processing: It call other function like stop words, urls, usernames, etc. are placed under preprocessing stages, etc.

Output: It gives either positive or negative sentiments based on twitter data with elimination of pre-processing.



```
# Cleaning and removing the above stop words list from the tweet text :
STOPWORDS = set(stopwordlist)


def cleaning_stopwords(text):
    return " ".join([word for word in str(text).split() if word not in STOPWORDS])
```



```
[ ]  # Cleaning and removing URL’s :
     def cleaning_URLs(data):
         return re.sub('((www.[^s]+)|(https?://[^s]+))', ' ', data)


     dataset['text'] = dataset['text'].apply(lambda x: cleaning_URLs(x))
     dataset['text'].tail()
```

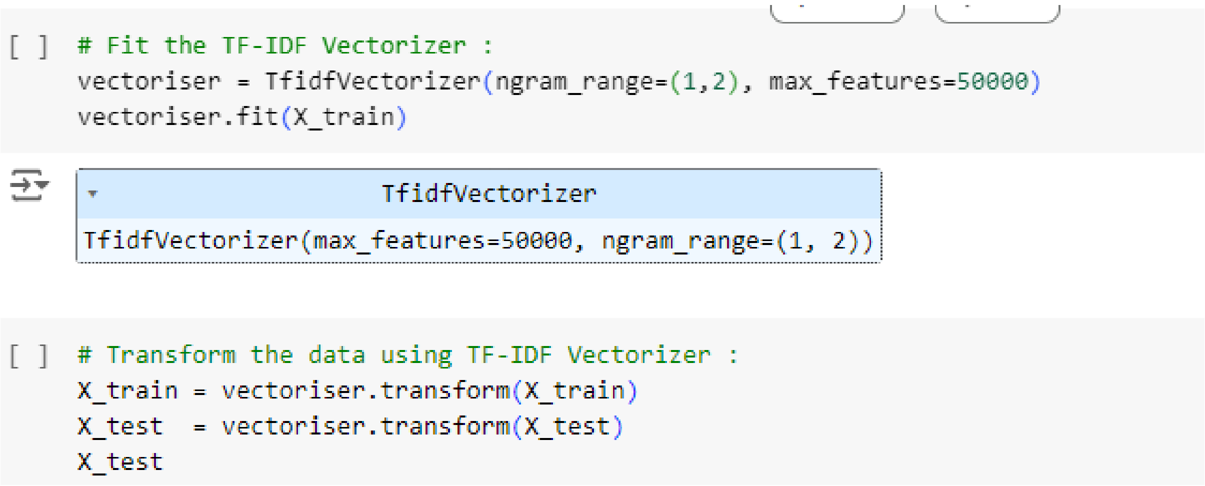
Figure 4.1: Data Preprocessing

4.1.2 Feature Extraction

This is implemented after the preprocessing of data. Once the data are cleaned using pre-processing the features of the data are extracted using the model.

Input: This takes pre-processed data as in input.

Process: It then uses the method, to process the taken input, process and extract the feature.



```
[ ]  # Fit the TF-IDF Vectorizer :
     vectoriser = TfidfVectorizer(ngram_range=(1,2), max_features=50000)
     vectoriser.fit(X_train)
```

```
▾                 TfidfVectorizer
TfidfVectorizer(max_features=50000, ngram_range=(1, 2))
```

```
[ ]  # Transform the data using TF-IDF Vectorizer :
     X_train = vectoriser.transform(X_train)
     X_test  = vectoriser.transform(X_test)
     X_test
```

Figure 4.2: Feature Extraction

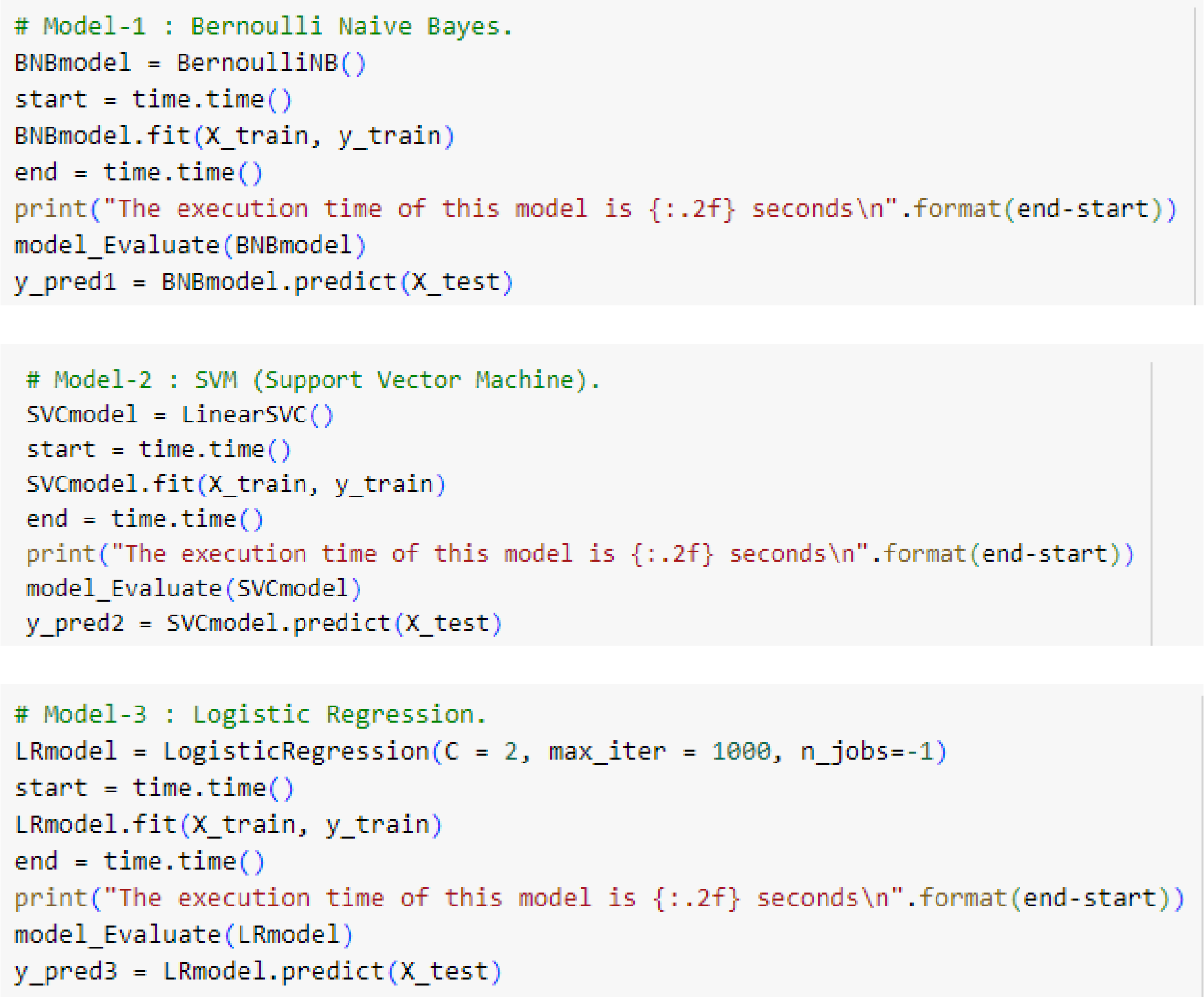
4.1.3 Model Training

With the elimination of stop words, urls and feature extraction the model is trained using logistic regression, Naive Bayes and Support Vector Machine, after that model is used for prediction.

Input: It takes input from the feature extractor.

Process: It classify the tweets as positive or negative and return the list of tweets with its sentiment value.

Output: It gives classified tweets with positive and negative tweets value.



```
# Model-1 : Bernoulli Naive Bayes.
BNBmodel = BernoulliNB()
start = time.time()
BNBmodel.fit(X_train, y_train)
end = time.time()
print("The execution time of this model is {:.2f} seconds\n".format(end-start))
model_Evaluate(BNBmodel)
y_pred1 = BNBmodel.predict(X_test)
```

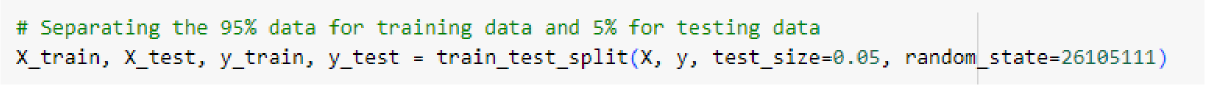
```
# Model-2 : SVM (Support Vector Machine).
SVCmodel = LinearSVC()
start = time.time()
SVCmodel.fit(X_train, y_train)
end = time.time()
print("The execution time of this model is {:.2f} seconds\n".format(end-start))
model_Evaluate(SVCmodel)
y_pred2 = SVCmodel.predict(X_test)
```

```
# Model-3 : Logistic Regression.
LRmodel = LogisticRegression(C = 2, max_iter = 1000, n_jobs=-1)
start = time.time()
LRmodel.fit(X_train, y_train)
end = time.time()
print("The execution time of this model is {:.2f} seconds\n".format(end-start))
model_Evaluate(LRmodel)
y_pred3 = LRmodel.predict(X_test)
```

Figure 4.3: Model Training

4.2 Testing

Among the total data 95% of the data is used for training and 5% is used for testing. For testing hit and trial method is followed and the testing module from Scikit-learn is used for testing.

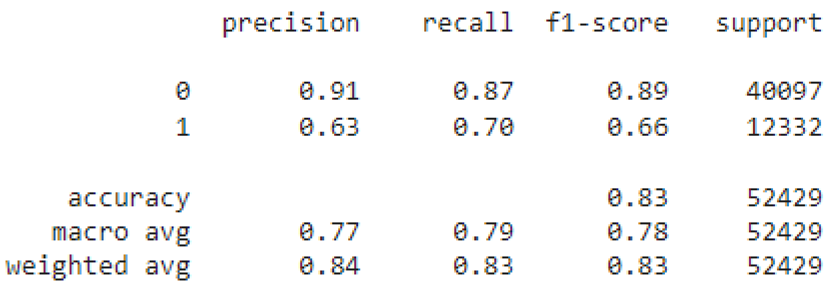


```
# Separating the 95% data for training data and 5% for testing data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.05, random_state=26105111)
```

Figure 4.4: Train-Test Split

4.3 Model Evaluation

After training, the model's performance was evaluated on the test dataset. Metrics such as accuracy, precision, recall and f1-score were calculated from the confusion matrix to assess its effectiveness in classifying positive and negative sentiments.



```
              precision    recall  f1-score   support

           0       0.91      0.87      0.89     40097
           1       0.63      0.70      0.66     12332

    accuracy                           0.83     52429
   macro avg       0.77      0.79      0.78     52429
weighted avg       0.84      0.83      0.83     52429
```

Figure 4.5: Model Evaluation

CHAPTER FIVE

RESULT and FINDINGS

5.1 Evaluation Metrics

Confusion Metrics: The confusion metrics provides a detailed breakdown of the model's predictions, showing the number of true positive, true negative, false positive, and false negatives. The confusion matrix results provide a detailed view of the classification performance of the model.

Confusion matrix of Naive Bayes

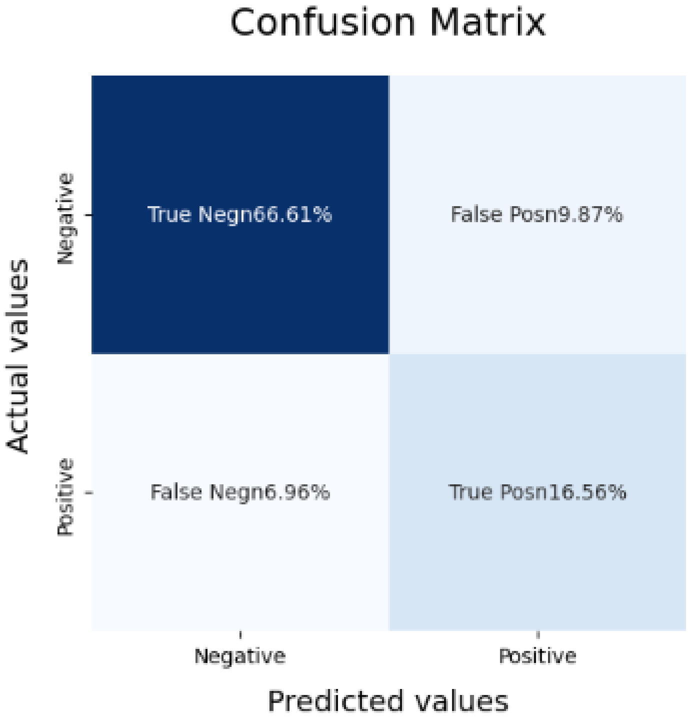


Figure 5.1: Confusion matrix of Naive Bayes

Confusion matrix of Support Vector Machine

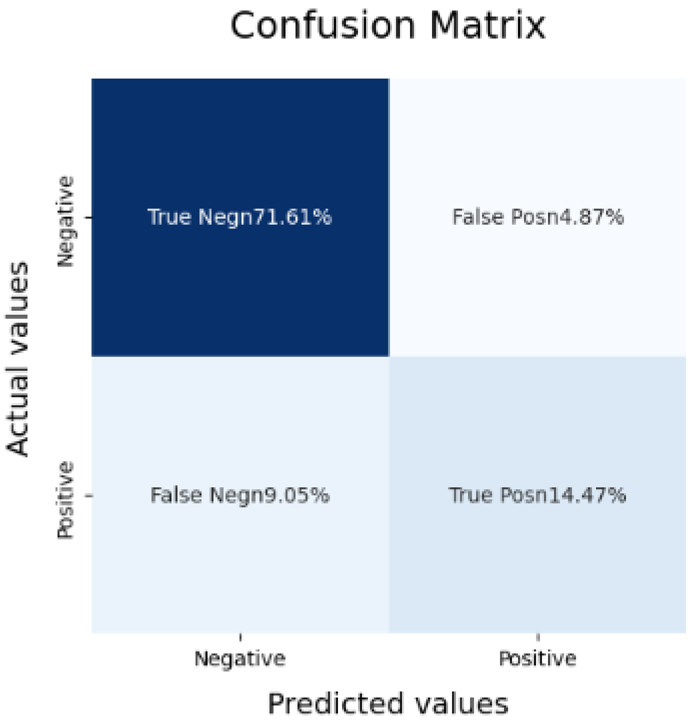


Figure 5.2: Confusion matrix of Support Vector Machine

Confusion matrix of Logistic Regression

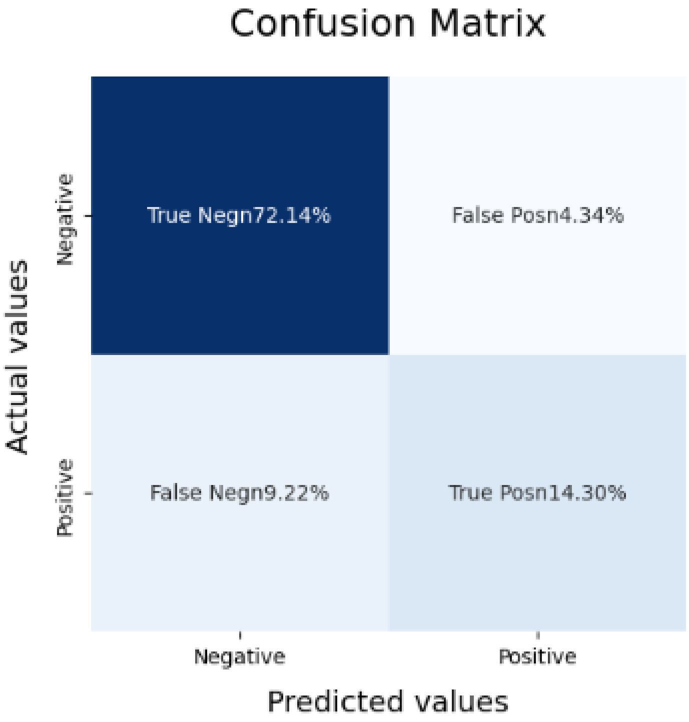


Figure 5.3: Confusion matrix of Logistic Regression

Accuracy: Accuracy will be a good evaluator for the models. This is most commonly use metric to evaluate how well the model predicts. Accuracy is the ratio of the number of correct predictions made against the total number of prediction made.

$$\text{Accuracy} = \frac{(\text{True Positive} + \text{True Negative})}{\text{Total Number of Samples}}$$

Precision: Precision calculates the ratio of correctly predicted positive instances to the total number of instances predicted as positive by the model. It assesses how well the model performs when it predicts a positive sentiment.

$$\text{Precision} = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Positive})}$$

Recall: Recall measures the model’s ability to correctly identify positive instances (true positives) out of all the actual positive instances. It is also known as Sensitivity or True Positive Rate.

$$\text{Recall} = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Negative})}$$

F1-Score: The F1-score is the harmonic mean of precision and recall. It provides a balance between precision and recall, especially when you want to find a single metric to evaluate your model’s performance.

$$\text{F1} - \text{Score} = \frac{2 * \text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})}$$

5.2 Model Performance Evaluation

Performance Model of Each output are evaluated and are deviced to compare each set of classifiers output and among all the most accurate data will be considered as a useful one. The logistic regression works the best as their accuracy is the most among all.

Table 5.2 Model Performance Evaluation

	Positive		
Classifier	Naive Bayes	SVM	Logistic Regression
Accuracy	83%	86%	87%
Precision	91%	89%	89%
Recall	87%	94%	95%
F1 Score	89%	91%	92%
Average F1 score	78%	79%	80%

Through this seminar report a comparison of different ML classifiers is performed on twitter Sentiment Analysis. The Logistic Regression shows the highest accuracy of 87% with an average F1 score of 80% for all the two types of tweets. The Naive Bayes and SVM classifiers display the near similar pattern, as both the classifiers analyze both positive and negative tweets with much accuracy in the positive sentiments tweets than the negative tweets. The Naive Bayes classifier also performed fairly well with the accuracy rate of 83%. The least performer among all the base classifiers is Naive Bayes with the accuracy rate of just 83% and an average F1 score rate further slips to 78%. The results thus obtained can be helpful for the companies to analyze their product related customer opinions and also to customers to choose the best product based on public reviews.

CHAPTER SIX

CONCLUSION AND FUTURE RECOMMENDATIONS

6.1 Conclusions

The study compared the performance of logistic Regression, Naive Bayes and Support Vector Machine for Tweets Sentiment Analysis classification. Evaluation metrics such as accuracy, precision, recall, and F1 score were used.

The logistic Regression emerged as the top performer, exhibiting the highest accuracy, precision and recall, it demonstrated superior overall performance in accurately classifying positive sentiments tweets.

In conclusion, the logistic Regression stands as the optimal choice for classifying and analyzing twitter sentiment analysis tasks, offering superior accuracy, precision and recall.

6.2 Future Recommendation

There are many potential directions for future research and development. Investigating different classification algorithms, such as random forests, or neural networks, to see how well they perform in sentiment analysis tasks and determine which approach is best for this specific issue is one possible direction.

In order to record the semantic relationships between words and possibly increase classification accuracy, more advanced feature extraction techniques like word embedding's could be investigated.

Another area of future research is the incorporation of advanced natural languages processing techniques, such as sentiment lexicons or deep learning-based models like transformers, to better understand and capture the nuances within textual data.

Finally, it would be advantageous to assess the model's performance on bigger and more varied datasets, covering a wider variety of sentiment expressions and domains, to increase the generalizability of the model.

References

[1] R. B. L. H. Alec Go, "Twitter Sentiment Classification using Distant Supervision," *Twitter Sentiment Classification using Distant Supervision,* p. 6, Jan 1, 2009.

[2] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," *arXiv preprint cs/0205070,* 2002 May 28.

[3] P. Waykar, K. Wadhwani, P. More and A. Kollu, "Sentiment Analysis of Twitter tweets using supervised classification technique," *International Journal of Engineering Research and Applications,* pp. 32-34, 2016.

[4] I. S.T, L. Wikarsa, B. MComp, R. Turang and S. MKom, "Using logistic regression method to classify tweets into the selected topics," *international conference on advanced computer science and information systems (icacsis),* pp. 385-390, 2016 Oct.

[5] A. P. D. B. Shikha Tiwari, "Social Media Sentiment Analysis," *International Conference on Advanced Computing & Communication,* 2020.