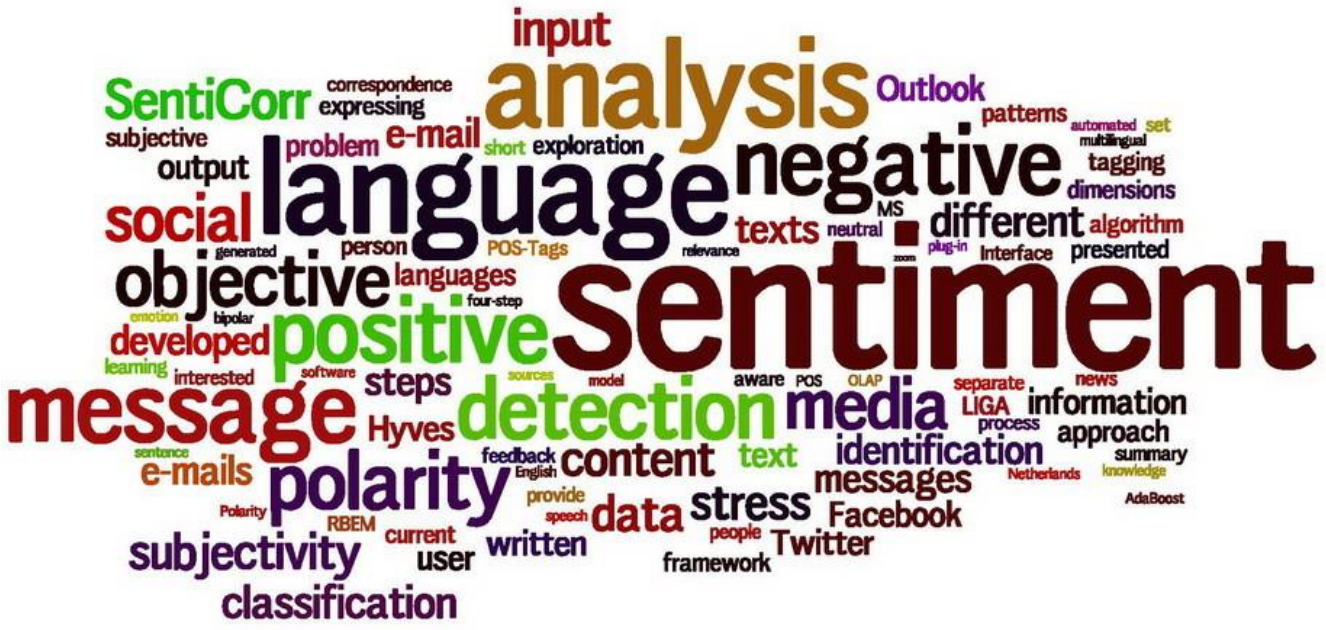
Symbiosis Skills and Professional University

Kiwale, Pune

PROJECT REPORT

On

"Twitter Sentiment Analysis"



Submitted by

1) **Pratik Dange**
2) **Aakanksha Bhavsar**
3) **Prarthana Walwadkar**

DA- Batch FST02:DA03

Under The Guidance of

Trainer's Name: 1) Ritviz Singh
2) Shiv Patel

STUDENT DECLARATION AND ATTESTATION BY TRAINER

I'm Pratik Dange, Aakanksha Bhavsar, Prarthana Walwadkar A student at **Symbiosis Skills and Professional University,** hereby declare and attest that we understand and agree to comply with the rules and regulation pertaining to the Twitter Sentiment Analysis implemented by the intuition.

Name of Students: - 1) Pratik Dange
2) Aakanksha Bhavsar
3) Prarthana Walwadkar

Course: Data Associate

CERTIFICATE

This is to certify that the report entitled, **"Twitter Sentiment Analysis**" submitted by "**1)Pratik Dange, 2)Aakanksha Bhavsar, 3)Prarthana Walwadkar**" to Symbiosis Skills and Professional University, Pune, Maharashtra, India, is a record of bonafide Project work carried out by him under my supervision and guidance and is worthy of consideration for the completion of certificate course in 'Data Associate".

Signature of Trainer
Name of Trainer
Date: / / 2024

Supervisor
Date:

Supervisor

ACKNOWLEDGEMENTS

We extend our sincere appreciation to all those who contributed significantly to the successful development of the "Twitter Sentiment Analysis" by TeamPratik Dange, Aakanksha Bhavsar, Prarthana Walwadkar, under the guidanceof **Ritviz Singh** and **Shiv Patel** at **Symbiosis Centre of Distance Learning.**

Our team members have exhibited unwavering dedication, bringing a wealth of diverse skills to the project, resulting in the creation of an innovative system that seamlessly integrates technology with practical application.

Heartfelt thanks are extended to our esteemed project guides, **Ritviz Singh** and **Shiv Patel**, for their invaluable mentorship and technical expertise. Their guidance has not only shaped the project's technical aspects but has also ensured its alignment with the high academic standards of **Symbiosis Centre of Distance Learning.**

Team -

1) Pratik Dange
2) Aakanksha Bhavsar
3) Prarthana Walwadkar

INDEX

ABSTRACT

A Sentiment is a generic term that describes an attitude or opinion that is often caused or influenced by emotion. Feelings such as pity, romantic love, sadness, etc. that influence somebody's action or behavior are also sentiments but majorly there are two types of sentiments — **"positive" and "negative"**. Currently work will be done on positive and negative comments.

People came across various social media platforms where the sentiments are emotions of the user are hurt by the comments of the other users so to avoid that a model which predict the sentiment of the statement should be applied on social media platforms show that if hate speech is recognized or something which may hurt the other user should be avoided so it will be beneficial for the users and their mental peace and social media will be used only for the purpose of entertainment and not for getting hurt or to hurt someone.

Existing work in this domain includes using ensemble machine learning methods and natural language processing to detect the sentiment of the text. Using this method was not completely useful as it was not able to understand the shortform or the patterns as well as the chat words and the accuracy of the model was not good. And bilingual data is tough to work on as it was difficult to machine to understand the text language and predict the sentiment. This project works on both that is it understands the short forms, chat words, emoticons as well as the code mixed comments and predict the sentiment of the user.

LIST OF ABBREVIATIONS

NLP	Natural Language Processing
BNB	Bernoulli Naïve Bayes’
SVM	Support Vector Machine
SVC	Support Vector Classifier
MNB	Multinomial Naïve Bayes

CHAPTER 1

INTRODUCTION

This project discusses the different classifiers that can be used for sentiment analysis of twitter data, to classify the tweets as positive or negative. The challenge of Hindi-English Code-mixed Social Media Text is focused on here. An existing labelled Kaggle dataset is used for this study. Forty thousand rows of this dataset are randomly selected and then cleaned. Adjectives, adverbs and abstract nouns are selected as features and extracted for each cleaned tweet. Then seven different classifiers, namely, Naïve Bayes', Multinomial Naïve Bayes', Bernoulli's Naïve Bayes', Logistic Regression, Stochastic Gradient Descent, Support Vector Machines and Maximum Entropy classifiers are trained on 85% of the dataset. A hybrid model is created using these seven classifiers by implementing the voting based ensemble model. Then a function is created which uses TextBlob to identify the language of the word and in case, the language is 'hi', i.e. Hindi, then the Google Machine Translator is used to convert that word to its English form. The function also tries to handle the challenge of language ambiguity which occurs when a word exists in both the English and Hindi dictionaries. The translated final string is passed to another function that uses the string as a test case for the seven base classifiers and the hybrid model and the sentiment predicted by each of these classifiers is printed along with the extracted features and the translated final tweet/string given by the user.

PROBLEM STATEMENT

A CSV (comma-separated values) file is having comments from the social media platform Twitter, and the problem is that classification has to be done on the text on the basis of sentiments so that the model can predict the comment nature.

Classification is one of the central problems of machine learning. Much work has been done in the field of classification. The sentiment analysis also essentially boils down to a classification problem.

MOTIVATION

Humans ourselves are not able to understand how exactly language is processed by brains. So, it is possible for us to teach a machine to learn human language through extensive research, a lot of methods have been developed that could help machines understand human languages. NLP or Natural Language Processing is the field of study that focuses on the interactions between human language and computers. One subproblem of NLP is sentiment analysis, i.e. classifying a statement as positive or negative. Let's take an example of Amazon website. On Amazon, it's users can leave a comment about a product stating whether it was good, bad or it could even be neutral. Now, using a human to read all the comments and obtaining the overall customer feedback on the product would be expensive and time-consuming. Enter machine learning model. The machine learning model can churn through a vast amount of data, making inferences and classifying the comment. Using this ML model, Amazon can better its products through the customer reviews which would bring in more revenue for the company.

Sentiment analysis isn't as straightforward as it may seem. If one thinks that the comments which contain the words "good", "awesome", etc can be classified as a positive comment and the comments which the words "bad", "miserable" etc can be classified as a negative comment, he should think again. E.x: "Completely lacking in good taste" and "Good for a quick meal but nothing special" represent a negative and neutral feedback respectively even though they have the word "good" in them. Therefore, the task may not be as easy as it may seem.

CHAPTER 2

LITERATURE REVIEW

CSV FILE

A CSV is a comma-separated values file, which allows data to be saved in a tabular format. CSVs look like a garden-variety spreadsheet but with a .csv extension.
CSV files can be used with most any spreadsheet program, such as Microsoft Excel or Google Spreadsheets. They differ from other spreadsheet file types because user can only have a single sheet in a file, they cannot save cell, column, or row. Also, user cannot not save formulas in this format.

These files serve a number of different business purposes. They help companies export a high volume of data to a more concentrated database, for instance. They also serve two other primary business functions:
CSV files are plain-text files, making them easier for the website developer to create
Since they're plain text, they're easier to import into a spreadsheet or another storage database, regardless of the specific software you're using
To better organize large amounts of data

BILINGUAL SENTIMENT ANALYSIS

Bilingual sentiment analysis is an approach to conduct sentiment analysis on the 2 mixed languages that is Hindi and English. In India, majorly Hindi language is spoken and used on social media platforms as well as in chatting or messaging but those Hindi words are type in English language. As well as reviews on various platforms such as Amazon, Flipkart and other shopping apps are in hinglish language. The computer cannot directly understand the hinglish language therefore sentiment analysis is necessary on this hinglish data.

CODE MIXING

Code in sociolinguistics refers to a language or a language variety. Code-mixing can be defined as simply mixing of two or more varieties of the same language or of different languages altogether. Example of Code-Mixed Hindi-English text is – “tu apne saath college bag leja raha hai?” or “arey waah! I am very proud of you”.

OVERVIEW

MATERIAL AND METHODS

Most of the research work is targeted in the direction of event detection and performing the sentiment analysis of events. The researchers have considered the text written in one language as English, Hindi, Chinese, Romanian, French and German etc. and bilingual like English & Chinese, English & Romanian etc along with different multilingual communities. However Indian researchers have focused on unilingual and bilingual communities. People have worked on code mixing and linguistic switching in Hindi-English bilingual combination. The presence of ambiguous words and inconsistent spellings has rendered the analysis rather incomplete and less effective. Words written in roman can have many different appearances. For example words such as mei, mein, main, etc. refer to one word in devnagri script. There are many words which are present in English and in Hindi with same phonetics but different meaning like bus, suffer, Holi etc. In the proposed strategy, the aim to provide more accurate results for social feeds in Hindi-English (Hinglish) and bridge the mentioned research gaps. To cover gaps, a large amount of data set is required. For English language, python is well-equipped with rich library of sentiment words and natural language processing functions. On the other hand, dealing with Hindi and Hinglish, utf encode and decode functions are not very helpful in translating each and every word.

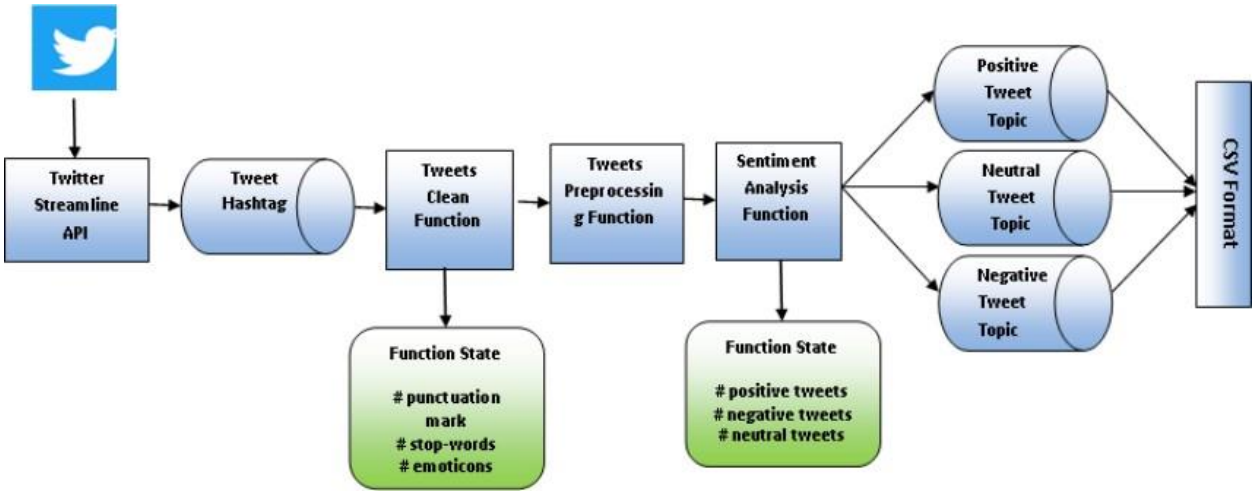


Figure 2.1: Functional Model Of Twitter Sentiment Analysis

CHAPTER 3

METHODOLOGIES

MODEL WORKFLOW

The model takes the data set and reads it. All the instances and attributes are equally considered. Encoding is applied on the data to convert it. Preprocessing such as dropping unwanted columns, removal of HTML tags, converting alphabetical values, replacing concatenated words, removal of punctuation, spell checker, Lemmatization and doing feature extraction. Training and testing is done on the dataset and the classifiers are trained. Then accuracy is checked and then the text is classified.

Naïve Bayes' Classifier

Naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features (see Bayes classifier). They are among the simplest Bayesian network models, but coupled with kernel density estimation, they can achieve high accuracy levels.

- Multinomial naive Bayes

With a multinomial event model, samples (feature vectors) represent the frequencies with which certain events have been generated by a multinomial(p1,….pn) where pi is the probability that event i occurs (or K such multinomials in the multiclass case). A feature vector **x** = (x1,…,xn) is then a histogram with xi counting the number of times event i was observed in a particular instance. This is the event model typically used for document classification, with events representing the occurrence of a word in a single document (see bag of words assumption). The likelihood of observing a histogram x is given by

$$p(\mathbf{x} \mid C_k) = \frac{(\sum_{i=1}^{n} x_i)!}{\prod_{i=1}^{n} x_i!} \prod_{i=1}^{n} {p_{ki}}^{x_i}$$

The multinomial naive Bayes classifier becomes a linear classifier when expressed in log-space:

$$\log p(C_k \mid \mathbf{x}) \propto \log\left(p(C_k)\prod_{i=1}^{n} {p_{ki}}^{x_i}\right)$$
$$= \log p(C_k) + \sum_{i=1}^{n} x_i \cdot \log p_{ki}$$
$$= b + \mathbf{w}_k^\top \mathbf{x}$$

where $b = \log p(C_k)$ and $w_{ki} = \log p_{ki}$.

If a given class and feature value never occur together in the training data, then the frequency-based probability estimate will be zero, because the probability estimate is directly proportional to the number of occurrences of a feature's value. This is problematic because it will wipe out all information in the other probabilities when they are multiplied. Therefore, it is often desirable to incorporate a small-sample correction, called pseudocount, in all probability estimates such that no probability is ever set to be exactly zero. This way of regularizing naive Bayes is called Laplace smoothing when the pseudocount is one, and Lidstone smoothing in the general case.

Rennie et al. discuss problems with the multinomial assumption in the context of document classification and possible ways to alleviate those problems, including the use of tf–idf weights instead of raw term frequencies and document length normalization, to produce a naive Bayes classifier that is competitive with support vector machines.

- Bernoulli naive Bayes

In the multivariate Bernoulli event model, features are independent Booleans (binary variables) describing inputs. Like the multinomial model, this model is popular for document classification tasks, where binary term occurrence features are used rather than term frequencies. If xi is a boolean expressing the occurrence or absence of the i'th term from the vocabulary, then the likelihood of a document given a class Ck is given by

$$p(\mathbf{x} \mid C_k) = \prod_{i=1}^{n} p_{ki}^{x_i}(1 - p_{ki})^{(1-x_i)}$$

where pki is the probability of class Ck generating the term xi. This event model is especially popular for classifying short texts. It has the benefit of explicitly modelling the absence of terms. Note that a naive Bayes classifier with a Bernoulli event model is not the same as a multinomial NB classifier with frequency counts truncated to one.

STOCHASTIC GRADIENT DESCENT

Stochastic gradient descent (often abbreviated SGD) is an iterative method for optimizing an objective function with suitable smoothness properties (e.g. differentiable or subdifferentiable). It can be regarded as a stochastic approximation of gradient descent optimization, since it replaces the actual gradient (calculated from the entire data set) by an estimate thereof (calculated from a randomly selected subset of the data). Especially in high-dimensional optimization problems this reduces the very high computational burden, achieving faster iterations in trade for a lower convergence rate.

CHAPTER 4

EXPERIMENTS

DATASET

The dataset that was used was obtained from “Kaggle” called the Sentiment140 dataset.

It contains 1,600,000 tweets extracted using the twitter API. The tweets have been annotated (0 = negative, 2 = neutral, 4 = positive) and they can be used to detect sentiment. But only 40000 rows where randomly selected from this dataset, with equal distribution of positive and negative tweets, i.e. neutral tweets were ignored as this study focuses on binary classification.

PREPROCESSING

In order to make sure that the work was carried out in the most efficient manner, the dataset used needed to be preprocessed before actual usage.

Steps :

- Drop unwanted columns
- Removal of HTML tags and question marks
- Convert all the text to lowercase
- Replace concatenated words with full forms
- Removal of punctuation marks, stopwords
- Spellchecker for spelling correction
- Stemming process
- Lemmatization of words and substitution of negated words by antonym
- Feature Extraction

DATA CLEANING

"In our Twitter feelings project, we made the words people wrote simpler so we could understand them better. We used a special tool called PorterStemmer to do this. First, we cleaned up the words by getting rid of strange symbols and making everything lowercase. Then, we separated the words to look at them one by one. The important part was using PorterStemmer to change big words into smaller ones, like turning 'running' into 'run.' We also took out common words that don't give us much information, like 'the' or 'and.' After doing this process for each

tweet in our Twitter data, we made a new section called 'steemed_content.' This section has the words in a way that's easier for a computer to understand, and it helps us figure out how people feel in a smarter way."The final datasetlooks like the following:

```
1 twitter_data.head()
```

	lable	id	date	flag	user	tweet	steemed_content
0	ï»¿Label	number	date	no_query	name	Tweet	tweet
1	0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, t...	switchfoot http twitpic com zl awww bummer sho...
2	0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by ...	upset updat facebook text might cri result sch...
3	0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Man...	kenichan dive mani time ball manag save rest g...
4	0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire	whole bodi feel itchi like fire

Figure 4.1: Dataset Overview after Preprocessing

FEATURE EXTRACTION

focused on and the rest of the words were removed as they did not add any value to the sentiment. This was done as part for the feature identification and extraction. This was implemented by checking each word in the cleaned tweet with the words in a file, which was prefilled with most common adjectives, adverbs and abstract nouns. If the word was not present in this file, it was not chosen as a feature and if it matched a word in this file, it was selected as one of the features. All these features of each text was stored in another column. This method was used because the existing method is not completely accurate and is outdated. For example, the word ‘clever’ was being identified as a ‘noun’ by the pos_tag function provided by the nltk library.

FINAL DATASET

This dataset has three columns, namely Tweet, Label and steemed_tweets. The first two are the ones taken from the original dataset. The last column is a derived column from all the preprocessing and contains just the features extracted from each of the tweet texts, as string.

PROPOSED SYSTEM

After getting the pre-processed dataset, this dataset is split into training and testing data with a ratio of 85%. This approximates to 34000 tweets in training data and 6000 tweets in testing data. Then the above mentioned seven classifiers are trained on this data and then tested against the testing data and the performance of these classifiers is evaluated. Next a hybrid model as a voting based ensemble model of these seven classifiers is constructed. Following suit, this model is trained and then tested to evaluate its performance. After this, a translation mechanism is used to handle the challenge of the Hinglish text using the Google Translator Machine and a function is created that takes in text as input and translates it if required and then uses the seven trained base classifiers and the hybrid model to predict the sentiment of the input.

CHAPTER 5

RESULTS

The dataset is used to train the model. Many classifying models are applied on the data set. To evaluate the performance of all classifier models, the precision, recall and f1 score.

Bernoulli Naïve Bayes'

Classification Report:

	precision	recall	f1-score	support
0	0.74	0.82	0.78	4002
4	0.80	0.71	0.75	3999
accuracy			0.76	8001
macro avg	0.77	0.76	0.76	8001
weighted avg	0.77	0.76	0.76	8001

Multinomial Naïve Bayes'

Classification Report:

	precision	recall	f1-score	support
0	0.74	0.82	0.78	4002
4	0.80	0.71	0.75	3999
accuracy			0.76	8001
macro avg	0.77	0.76	0.76	8001
weighted avg	0.77	0.76	0.76	8001

Logistic Regression

```
Classification Report (Training Data):
              precision    recall  f1-score   support

           0       0.86      0.84      0.85     15998
           4       0.84      0.86      0.85     16002

    accuracy                           0.85     32000
   macro avg       0.85      0.85      0.85     32000
weighted avg       0.85      0.85      0.85     32000
```

Stochastic Gradient Descent

```
Classification Report:
              precision    recall  f1-score   support

           0       0.77      0.74      0.75      4002
           4       0.75      0.77      0.76      3999

    accuracy                           0.76      8001
   macro avg       0.76      0.76      0.76      8001
weighted avg       0.76      0.76      0.76      8001
```

Support vector machine

```
Classification Report:
              precision    recall  f1-score   support

           0       0.77      0.76      0.76      4002
           4       0.76      0.77      0.77      3999

    accuracy                           0.76      8001
   macro avg       0.76      0.76      0.76      8001
weighted avg       0.76      0.76      0.76      8001
```

Hybrid Model

```
Accuracy: 0.7515310586176728
              precision    recall  f1-score   support

           0       0.74      0.77      0.76      4002
           4       0.76      0.74      0.75      3999

    accuracy                           0.75      8001
   macro avg       0.75      0.75      0.75      8001
weighted avg       0.75      0.75      0.75      8001
```

ACCURACY

The graph showing comparison of the accuracies:

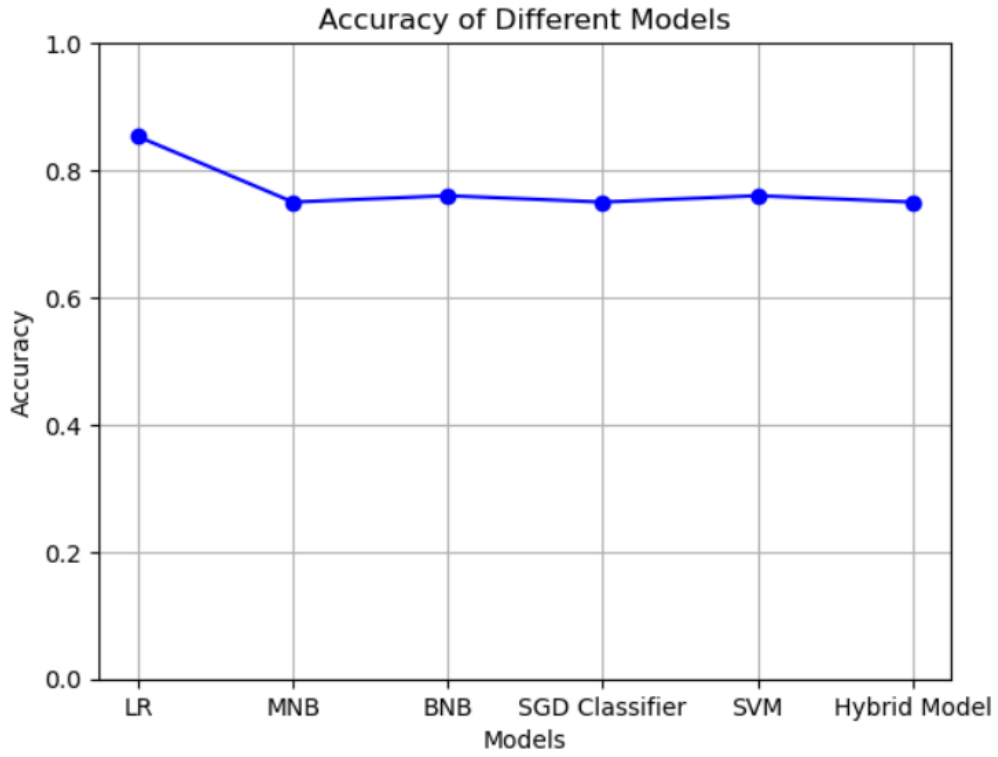
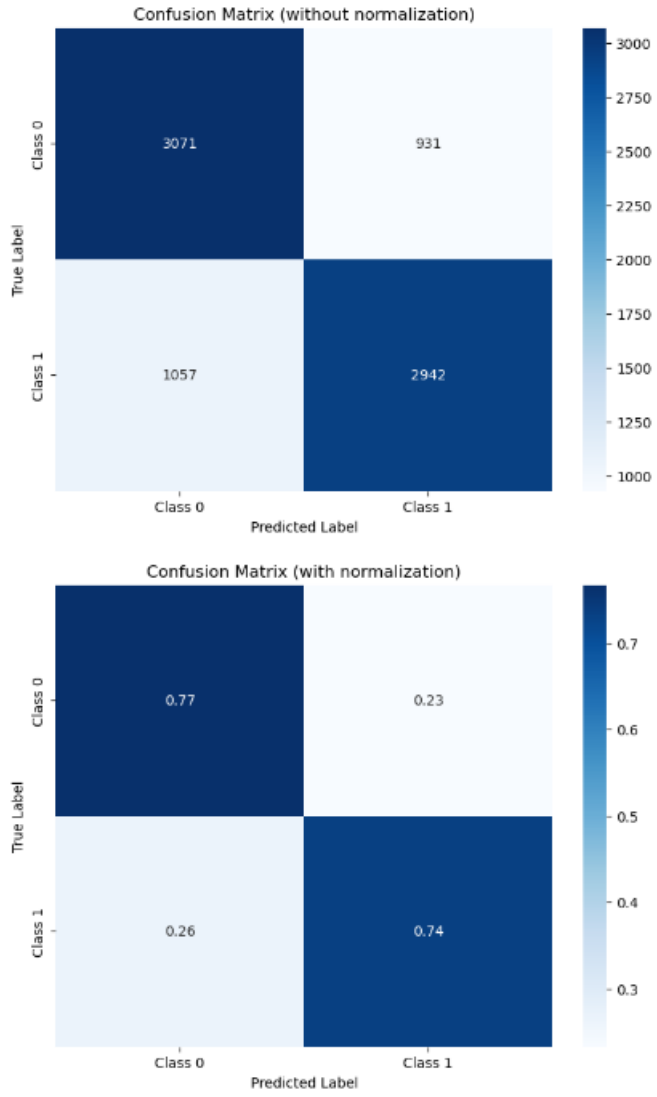


Figure 5.1: Graphical Comparison of Accuracies

```
Confusion Matrix (without normalization):
[[3071  931]
 [1057 2942]]

Confusion Matrix (with normalization):
[[0.76736632 0.23263368]
 [0.26431608 0.73568392]]
```



The confusion matrix for the hybrid model before and after normalization have beenshown below

Figure 5.2: Confusion Matrix without and with Normalization

CHAPTER 6
CONCLUSION

In this work, various machine learning techniques are used. Models such as Naïve Bayes', Bernoulli Naïve Bayes', Multinomial Naïve Bayes', Logistic Regression, Stochastic Gradient Descent and Support Vector Machine are implemented. All models are showing good accuracy but to get the accuracy comparison, Hybrid model is being applied. Hybrid Model basically compares accuracies of all models and then user can classify which model is better.

The tweet text can now be classified according to it's sentiment and accuracy is observed. Hybrid model and Bernoulli Naïve Bayes' are two best models with comparatively highest accuracy and this can be clearly observed from the graphical representation. Model successfully classifies the sentiment of hinglish comments.

REFERENCES

[1] Aditya Joshi, Ameya Prabhu Pandurang, Manish Shrivatsava and Vasudeva Varma, "Towards Sub-Word Level Compositons for Sentiment Analysis of Hindi-English Code Mixed Text," 26th International Conference on Computational Linguistics, December 2017.

[2] R. Mahesh, K. Sinha and Anil Thakur, "Machine Translation of Bilingual Hindi-English (Hinglish) Text," January 2005.

[3] Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed S. Akhtar and Manish Shrivatsava, "A Dataset of Hindi-English Code-Mixed Social Media Text for Hate Speech Detection," 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics, June 2018.

[4] Sakshi Gupta, Piyush Bansal and Radhika Mamidi, "Resource Creation for Hindi-English Code Mixed Social Media Text," July 2016.

[5] Kumar Ravi and Vadlamani Ravi, "Sentiment classification of Hinglish text," March 2016.

[6] [Sentiment Analysis] (https://monkeylearn.com/sentiment-analysis/)

[7] [Code Mixing] (http://languagelinguistics.com/2017/06/27/code-mixingsociolinguistics/)

[8] [Machine Learning Concepts] (https://deepai.org/machine-learning-glossary-and-terms/classifier)

[9] [Machine Learning Classifiers] (https://towardsdatascience.com/machine-learning-classifiersa5cc4e1b0623)

[10] [Classification Algorithms] (https://analyticsindiamag.com/7-types-classification-algorithms/)

[11] [Logistic Regression] (https://dataaspirant.com/2017/03/02/how-logistic-regression-modelworks/)

[12] [Naive Bayes'] (https://en.wikipedia.org/wiki/Naive_Bayes_classifier#Multinomial_naive_Bayes)

[13] [Ensemble Learning] (https://www.datacamp.com/community/tutorials/ensemble-learningpython)