SentiMix Hindi-English

I.SEMEVAL-2020 TASK 9: OVERVIEW OF SENTIMENT ANALYSIS OF CODE-MIXED TWEETS

Ayesha Ateeq Burney
CS Department
FAST NUCES University
Islamabad, Pakistan
i170010@nu.edu.pk

***Abstract*—This paper discusses the different classifiers that can be used for sentiment analysis of code-mixed twitter data, to classify the tweets as positive, negative, or neutral. The challenge of bilingual comments (English and Hindi mixed) is focused on here. An existing labelled dataset is used for this study. Forty thousand rows of this dataset are randomly selected and then cleaned. Most frequent words are selected as features and extracted for each cleaned tweet. Then six different classifiers, namely, Multinomial Naïve Bayes', Bernouille's Naïve Bayes', Logistic Regression, Stochastic Gradient Descent, Linear Support Vector Classifier and Random Forest classifiers are trained and tested on the dataset. The cleaned test tweets are passed to the six base classifiers and the sentiment predicted by each of these classifiers is printed along the accuracy of the classifier**

Keywords— twitter, dataset, Naïve Bayes', Logistic Regression, Stochastic Gradient Descent, Support Vector Classifier, Random forest, accuracy

II. INTRODUCTION

Sentiment Analysis is the interpretation and classification of emotions (positive, negative and neutral) within text data using text analysis techniques. It combines natural language processing (NLP) and machine learning techniques to assign weighted sentiment scores.

It allows us to gain an overview of the wider public opinion behind certain topics. Some of the most popular types of sentiment analysis are Fine-grained Sentiment Analysis, Emotion detection, Aspect-based Sentiment Analysis and Multilingual Sentiment Analysis.

This paper focuses on bilingual sentiment analysis, which is a subset of multilingual sentiment analysis. This involves a lot of pre-processing and resources. On social media websites like Twitter, users comment in various languages and in different alphabets. This makes it a challenge to analyse the sentiment of the text and is referred to as code-mixing.

Code-mixing can be defined as simply mixing of two or more varieties of the same language or of different languages altogether. Out of the codes that are mixed, the one whose structure or syntax is followed is generally called the matric language. Language contact and bilingualism are the prime causes of code mixing

With these thoughts in mind, this paper provides a method in which sentiment analysis can be done for multiple languages. In particular, the paper focuses on two languages – English and Hindi, written in the Latin alphabet.

Data is taken from an existing dataset for this study. Above fifteen thousand rows are randomly selected from the dataset rows with label as positive, negative, or neutral respectively. This dataset is then pre-processed, i.e. cleaned and features are extracted.

Six different base models – Multinomial Naïve Bayes', Bernouille's Naïve Bayes', Logistic Regression, Stochastic Gradient Descent, Support Vector Classifier, Random Forest - are trained on train dataset, i.e. 15,130 rows and tested on the test data 3,000 rows to evaluate each of their performances The performance of these classifiers is measured by calculating their accuracies, confusion matrix and classification report.

The next part discusses the 'Background and Related Work,' followed by 'Data' that describes the 'Dataset' and 'Data Pre-processing'. These are followed by the 'Proposed System', and the 'Implementation'. The 'Classifier Evaluation' and 'Results and Discussion' are then presented. Finally, the 'Conclusion' ends the paper with 'Acknowledgment' and 'References'.

III. BACKGROUND AND RELATED WORK

Advances in this area are impeded by the lack of a suitable annotated dataset. The authors of introduce a Hindi-English (Hi-En) code-mixed dataset for sentiment analysis and perform empirical analysis comparing the suitability and performance of various state-of-the-art SA methods in social media. They also introduce learning sub-word level representations in LSTM (Subword-LSTM) architecture instead of character-level or word-level representations. This linguistic prior in their architecture enables us to learn the information about sentiment value of important morphemes.

Aditya Joshi et al [1] mainly focuses on introducing a constantly learning sub-word level representation in LSTM (Subword-LSTM) architecture instead of character-level or word-level representations. This linguistic prior in their architecture enables us to learn the information about sentiment value of important morphemes. Some points to stress upon are that it works well in noisy text containing misspelling of words as well.

In order to determine the sentiment polarity of Hinglish text written in Roman script, the authors of [2] experimented with different combinations of feature selection methods and a host of classifiers using term frequency-inverse document frequency feature representation. They carried out in total 840 experiments in order to determine the best classifiers for sentiment expressed in the news and Facebook comments written in Hinglish. Some conclusions are a triumvirate of term frequency-inverse document frequency-based feature representation, gain ratio-based feature selection, and Radial

Basis Function Neural Network as the best combination to classify sentiment expressed in the Hinglish text.

All these papers discussed here do not handle the specific problem of sentiment analysis of Hindi-English Code-Mixed Social Media Text. The few papers that handle this take a different and more complex approach that is not completely necessary. For example, [2] converts the Hindi text written in the Latin Alphabet in the test case to Hindi written in the Devanagari Alphabet and then it performs sentiment analysis for Hindi Language separately and English language separately. This just adds an extra step and makes it rather complicated.

IV. DATA

A. DATASET

The dataset that was used was obtained from "CodaLab" called the Hindi-English CodeMix Data. It contains 17,000 tweets the train data has 15,131 tweets and the validation data has 3000 tweets extracted from the twitter API. The tweets have been annotated (negative, neutral, positive) and they can be used to detect sentiment

Given below is a snippet of the dataset that was used for this study:

SN	Tweet	label
3	@ AdilNisarButt pakistan ka ghra tauq he Pakistan Israel ko tasleem nahein kerta Isko Palestine kehta he - OCCUPIED PALESTINE	negative
41	Madarchod mulle ye mathura me Nahi dikha tha jab mullo ne du ko iss liye mara ki vo lasse ki paise mag liye the â€¦ https // t . co / oxf8tr3bly	negative
48	@ narendramodi Manya Pradhan Mantri mahoday Shriman Narendra Modi ji Pradhanmantri banne par Hardik Badhai tahe Dil â€¦ https // t . co / pRnoMsKkn1	positive
64	@ Atheist _ Krishna Jcb full trend me chal rahi aa	positive
66	@ AbhisharSharma _ @ RavishKumarBlog Loksabha me janta sirf modi ko vote de rahi thi na ki kisi MP or BJP ko without m â€¦ https // t . co / sHtBWCB7FM	positive
68	@ noirnaveed @ AngelAhana6 @ cricketworldcup Bhosdike tum pechvade ki tatti hi rahoge bc	negative
90	Love u Bhaijan ...â™¥â™¥ Father + son .. # Bharat # IAmBharat # BharatThisEid BEST Pic From Entire # Promotions .... Mashallah â€¦ https // t . co / S2xhwU6LUD	positive
99	@ manojgajjar111 Tumhara pass abh deemagh hai nahi islea Google ko apna deemagh banaya hua hai . Har koi tumhari tarh â€¦ https // t . co / BxuEug3xsN	negative
101	@ Mahlogo _ nolo Weni ankere o gae this weekend ðŸ˜‚ ðŸ˜‚ ðŸ˜‚ ðŸ˜‚ ðŸ˜‚	positive
104	@ Aurangzeb _ AIMIM @ Sachin540805591 Lage raho mullo tumhre issi Quran faad gyan ki kayal hain duniya Allah bhi khus â€¦ https // t co / GHa9dWNZ6u	negative
106	@ reema _ omer @ BushraGohar @ QasimKhanSuri @ Ali _ MuhammadPTI Pakhtoon k haqooq ki jang nahi punjab k haqooq ki jung pak â€¦ https // t . co / R7pjojNwre	negative
146	@ cooladitaya @ khanumarfa Ye sale is desh ke kabhi sage nahin ho sakte Ye gaddar hain Inke sath kadi se kadi kanun â€¦ https // t co / 0HuukV4zK2	negative
150	@ HYUN7IN saf baby ((i â€™ m sorry i wasn â€™ t online i hope you â€™ re sleeping well though i love you	positive
157	@ Nevveon Real life k log chor diye hain mene already jo snakes thy ab internet k chorne hain .... i'm sad for no reason solmon (((	negative
167	@ mjdawar Mohsin DawarTweets daikho aur kuch kha ker mar jao Bayghayrat insanTum aur Ali Wazir harami ho . Pakistan â€¦ https // t . co / i9v80emof3	negative
174	@ AnupamPKher Kuch media channel dikh rahe hai ki inke upar kai prakaar ke criminal cases hai shame on this type of media channel	negative
175	@ EkThaEngg @ AnupamPKher Abe mc .. nhi malum to ammi mat chuda bsdk .. grahm stains ko dara singh ne mara tha adivasiy â€¦ https // t . co / cL4elcDUe8	negative
182	@ tamashbeen _ Well chara Chor ke chele this news is a year old	negative
216	Twitter k baghair apna roza mumkin nahi hota ? Apna chutiyaap dusron per thopna band karo Bhai ! https // t . co / APKD4G8Ih0	positive
252	Oh haha the soldier is reading American Gods hardy har har thanks for reminding everyone of your OTHER clusterfuck	negative
266	@ asadowaisi @ PMOIndia Mp me govt kamalnath ki h .... Wese ek or video viral h jisme aapke shantidut bhai ek dukandar â€¦ https // t . co / doLm3ODDCs	negative
282	@ idevadhikari Dev da tmr new movie song Ami tmk valobasi I really loved â€ ðŸ˜˜ a lot ðŸ˜˜ðŸ˜˜	neutral

SN	Tweet	Label
20803	@O 454dkhan @O Heisunberg _O Agr kse ko itni importantce chaeay ni tou Ã°Å¸Ëœâ€¦EMT	neutral
20187	logon ko alloo pyaz tomator me toh allah pak ka naam nazar aa jata hai pr aankhon k samne allah pak ke bande nazar Ã¢â‚¬Â¦O https //O t	neutral
20953	@O LambaAlka Wafadaar bane rahane ka nayab tarika hai lamba ji aap us party se Judi hai jiska sardar @O ArvindKejriwal Ã¢â‚¬Â¦O https //C	neutral
13777	@O varnishant @O narendramodi Chup bhosdike .O He has been exonerated of all the charges .O Tum log ghanta prove nahi kar paaye .O	negative
20990	RT @O HardeepSPuri Ã Â¨Â¦Ã Â¨Â¾Ã Â¨Â¤Ã Â¨Â½O Ã Â¨Â¸Ã Â©â€˜O Ã Â¨Â¸Ã Â¨Â¾Ã Â¨Â²Ã Â¨Â¾Ã Â¨Â¹Ã Â©â‚¬Ã Â¨ÂO Ã Â¨ÂªÃ Â¨Â¿O Ã Â¨Â¸Ã Â¨	positive
15075	@O bomanirani #O HNY2 ki announcement hogyi to humka iss gola me nhi rehna hum nikal jaounga dusre gola me	neutral
32292	RT @O SharmakeF We're almost in the final three-days of Ramadan Ã¢â‚¬â€O very close to the finish line .O So let's make sure that we dout	neutral
4622	@O 6thrat This song makes me super happy .O I sing it for myself .O Take !O Khush raho abaad raho .O Hemant Da blessings !O https //O t	positive
22338	@O sadifabi @O Shahidmasooddr tm dikha rhy ho apni asli auqaat wese btao tou ye propaganda kyu krta hai shahid masood ?O	negative
23996	@O TheVampsCon you made Nadi so so so happy today .O Thank you so much for this .O I really love seeing one of my best friends this ha	positive
10426	RT @O DHEERAJ48968067 @O nirahua1 @O msunilbishnoi Wah bhai Shandaar jabab	positive
3038	RT @O deepakrajput97 @O sohailm08283995 @O SIR2712 @O krishan _O goa @O reema _O sahanii suar tere jaiso ko sikhaya ja hi nahi sakta	negative
37328	@O UdarOfficial Tere jesy chotiya sath hoge jo k conceler banny k laiq be nhi hain to koj nai badalna Pakistan da	negative
11313	Samarthan ke aabhav aur rajnitik swartho ke karan iska kabhi bhi sahi nyay nahi Kiya Gaya roshavash dharmik imarat Ã¢â‚¬Â¦O https //O t co	neutral
32158	RT @O AnujVer43198765 @O narendramodi Congratulations sir ji Dobara pm banee ki hardik subhkamnaye aapko	positive
2598	RT @O sahiljoshii Now kisan samman yojana for all no 5O Hector criteria anymore pension scheme for farmers 3000O rs per month and vacci	neutral
12837	RT @O mis _O bahu No Enlish word has double "O oo "O except Good .O Prove me wrong ?Ã°Å¸Â¤â€O https //O t .O co /O dNFMsFJNcF	neutral
26050	Modi ka lehar kisano pe baras raha kehar tut rahe haath pao road ki buri halato ka aasar @O aajtak @O BBCBreaking Ã¢â‚¬Â¦O https //O t .C	negative
30898	@O ravindraj12 @O TarekFatah Yarr India me aise log b hain ???O Life me 1st time India ka koi banda Positive dekha Appreciate u bro Ã°Å¸â€	positive
22656	@O BaijnathMourya @O mssirsa @O Aapkamanoj5 @O ArvindKejriwal Haan Mai free Wi-Fi dunga free water dunga Sasti Bijli doo Ã¢â‚¬Â¦O h	positive
2755	Log kitna bara chootiya hy jhoot ko such manker vote krdiya bcz of govt foult terror bring 1000 +O kg rdx and killed 5Ã¢â‚¬Â¦O https //O t c	negative
16061	@O asifnkhan08 @O SAUMEN _O BANERJ _O E @O JhaSanjay Time and circumstances ke sath jo change nhi hota vo sabse bada bewakuf ho	negative

Fig. 1. Snapshot of the Dataset

It contains three columns, which are described as follows:

- Label: The polarity of the tweet, that is the sentiment of the text. It has 3 major classifications that signify a different polarity. "negative" indicates a negative polarity, "neutral" indicates a neutral polarity and "positive" indicates a positive polarity.
- Uid: The SN of the tweet. An example is "4".
- Tweet: The text of the tweet. An example can be "@ noirnaveed @ AngelAhana6 @ cricketworldcup Bhosdike tum pechvade ki tatti hi rahoge bc". This is the most important field for this study along with the 'Label' field, as this is the text that will be passed into the classifiers.

B. DATA PREPROCESSING

- Drop unwanted text (i.e. meta)
- Removal of HTML tags, marks, twitter handles, and URLs
- Convert lower case to upper case
- Remove emojis and Unicode
- Removal of stop words
- Feature extraction

The snapshot of the dataset after removal of unwanted columns is as follows:

	SN	Tweet	label	cleaned_tweets
0	3	@ AdilNisarButt pakistan ka ghra tauq he Pakis...	negative	pakistan ka ghra tauq pakistan israel ko tasle...
1	41	Madarchod mulle ye mathura me Nahi dikha tha j...	negative	madarchod mulle ye mathura nahi dikha tha jab ...
2	48	@ narendramodi Manya Pradhan Mantri mahoday Sh...	positive	manya pradhan mantri mahoday shriman narendra ...
3	64	@ Atheist _ Krishna Jcb full trend me chal rah...	positive	krishna jcb full trend chal rahi a
4	66	@ AbhisharSharma _ @ RavishKumarBlog Loksabha ...	positive	loksabha janta sirf modi ko vote de rahi thi n...
...	...	...	...	...
15125	44345	Wow this is so sad ð□□ð□□ð□□ I don â□□ t AL...	negative	wow sad alika like da
15126	44673	@ rohitsharmawpg @ asadowaisi @ narendramodi W...	negative	shame shame na kro yeh narendra modi third chi...
15127	45085	@ Prof _ Hariom @ JKgrievance Who is BIJLI man...	negative	hariom bijli mantri people dont pay bills kash...
15128	45096	@ amjedmbt @ bandisanjay _ bjp @ cpkarimnagar ...	negative	bjp musalman ke naam pe kalan
15129	45164	@ Sunju _ Mishra To phir bjp ke leader vikas k...	negative	mishra phir bjp ke leader vikas ke bare kyon n...

15130 rows × 4 columns

	SN	Tweet	Label	cleaned_tweets
0	20803	@\tO 454dkhan\t @\tO Heisunberg\t _\tO Ag...	neutral	dkhan heisunberg agr kse ko itni importantce c...
1	20187	logon\t ko\t alloo\t pyaz\t tomator\t me\...	neutral	logon ko alloo pyaz tomator toh allah pak ka n...
2	20953	@\tO LambaAlka\t Wafadaar\t bane\t rahane\...	neutral	lambaalka wafadaar bane rahane ka nayab tarika...
3	13777	@\tO varnishant\t @\tO narendramodi\t Chup...	negative	varnishant narendramodi chup bhosdike exonera...
4	20990	RT\t @\tO HardeepSPuri\t Ã Â¨Â¦Ã Â¨Â¾Ã Â¨Â¤...	positive	hardeepspuri praise great giver gives sustenan...
...	...	...	...	...
2995	10425	@\tO i\t _\tO yogesh22\t @\tO atulreellif...	neutral	yogesh atulreellife haha atulreellife jai ho b...
2996	14162	Ã°Å¸â□¢Â□Ã°Å¸â□¢Â□Ã°Å¸â□¢Â□\tO you\t da\t r...	neutral	da realest thank putting track can't wait ne...
2997	15860	RT\t @\tO Nimra\t _\tO Khan241\t Fahad\t ...	neutral	nimra khan fahad bhaiiiii give us bilal abbas ...
2998	25435	RT\t @\tO fralaliciouxxe\t tbh\t i\t have...	neutral	fralaliciouxxe tbh bad sides say bad horribly...
2999	25183	@\tO BhadasManKi\t @\tO mithleshkumarmi\t ...	negative	bhadasmanki mithleshkumarmi kiranja bjpindia s...

3000 rows × 4 columns

Fig. 2. Dataset after dropping unwanted text

Given below is snippet of the dataset after feature extraction using TfidfVectorizer:

```
from nltk.corpus import stopwords
from sklearn.feature_extraction.text import TfidfVectorizer

vectorizer =TfidfVectorizer(sublinear_tf=True ,max_features = 2500, min_df=1, max_df=0.8)

processed_features_train = vectorizer.fit_transform(data['cleaned_tweets']).toarray()
processed_features_test = vectorizer.transform(data2['cleaned_tweets']).toarray()

print("processed_features_train\n",processed_features_train)
print("processed_features_test\n",processed_features_test)

```

```
processed_features_train
 [[0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 ...
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]]
processed_features_test
 [[0.         0.         0.         ... 0.         0.         0.        ]
 [0.23586814 0.         0.         ... 0.         0.         0.        ]
 [0.         0.         0.         ... 0.         0.         0.        ]
 ...
 [0.         0.         0.         ... 0.         0.         0.        ]
 [0.         0.         0.         ... 0.         0.         0.        ]
 [0.         0.         0.         ... 0.         0.         0.        ]]
```

```python
feature_names = np.array(vectorizer.get_feature_names())
sorted_by_idf = np.argsort(vectorizer.idf_)
_features = list(feature_names[sorted_by_idf[:100]])
counts = 0
for i in _features:
  print(i, sorted_by_idf[counts])
  counts += 1
```

```
hai 886
ki 1278
ko 1309
nahi 1627
ka 1170
ke 1236
se 2037
ye 2471
ho 963
bhi 335
hi 944
aur 171
ji 1139
ne 1650
modi 1571
jo 1159
sir 2106
aap 13
koi 1311
ha 872
nhi 1663
```

Fig. 3. Feature extraction of the cleaned tweets

V. Proposed System

To explain the basic system that is used during the course of this study, we can take assistance from the flowchart given below:

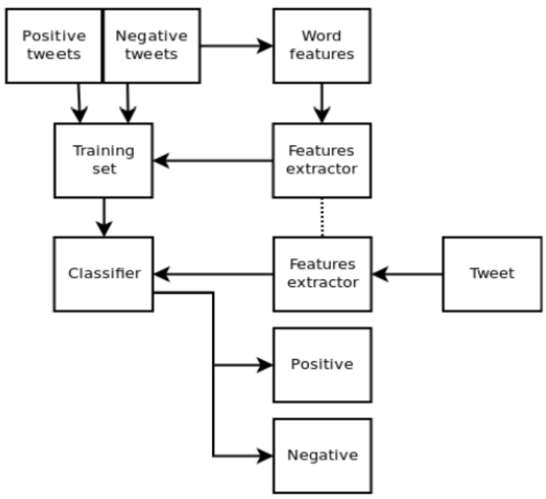

Fig. 4. Block Diagram depicting the flow

Initially, the dataset is cleaned as mentioned under 'Data Preprocessing' section under the heading 'Data'. Unwanted text, empty rows and stop words along with useless symbols are discarded. This approximates to 15,130 tweets of training data and 3,000 tweets of testing data. The training data is then passed into all the major classifiers that were used, namely Naive Bayes' Classifier, Multinomial Naive Bayes' Classifier, Bernoulli's Naive Bayes' Classifier, Logistic Regression, Stochastic Gradient Descent Classifier, Support Vector Classifier, and the Random Forest Classifier. Following this step, the testing data is passed into the respective trained models and determine the resulting accuracies of each classifier. From the above step, the classifier that has the highest accuracy will act as a benchmark accuracy for the study.

VI. Implementation

The following six classifiers were made to learn from the data given to it, also called the training data and make new classifications on the test data -

A. *Naïve Bayes' Classifier*

Naïve Bayes' classifiers are a collection of classification algorithms based on Bayes' Theorem. It is a family of algorithms where every pair of features being classified are independent of each other [4].

Assumptions of Naïve Bayes' classifiers are that each feature must be independent and equal. The formula used which is derived from Bayes' Theorem is as follows [4]:

$$y = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y) \qquad (1)$$

Some advantages of this classifier are [6]:

- Easy to implement.
- Fast.
- Does not require much training data.

Use Cases [6]:

- Document Classification
- Spam Filters
- Sentiment Analysis
- Disease Predictions

B. *Multinomial Naïve Bayes' Classifier*

Multinomial Naïve Bayes' Classifier estimates the conditional probability of a particular word, given a class as the relative frequency of term t in documents belonging to class (c) [7]. The variation takes into account the number of occurrences of term t in training documents from class (c), including multiple occurrences [7].

It is a specialised version of Naïve Bayes' that is designed more for text documents and explicitly models the word counts and adjusts the underlying calculations to deal with it [7].

The formula used is [7]:

$$p(x|C_k) = \frac{(\sum_i x_i)!}{\prod_i x_i!} \prod_i pki^{x_i} \qquad (2)$$

C. *Bernoulli's Naïve Bayes' Classifier*

The Bernoulli Naïve Bayes' classifier assumes that all our features are binary such that they take only two values (example – a nominal categorical feature that has been one-hot encoded) [7].

The formula derived is [7]:

$$p(x|C_k) = \prod_{i=1}^{n} p_{k_i}^{x_i}(1 - p_{k_i})^{(1-x_i)} \tag{3}$$

In the above formula, p_{k_i} is the probability of a class C_k generating term x_i .

This event model is especially popular for classifying short texts [7]. It has the benefit of explicitly modelling the absence of terms [4].

D. Logistic Regression

Logistic Regression is a statistical method for analysing a dataset in which there are one or more independent variables that determine an outcome [4]. The outcome is measured with a dichotomous variable (in which there are only two outcomes) [6].

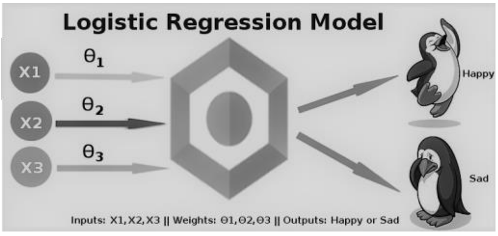

Fig. 5. Logistic Regression [5]

Some advantages are [6]:

- Performs well when data is linearly separable.
- Less prone to overfitting.
- Easy to implement and interpret.

Use Cases [6]:

- Identifying risk factors for diseases
- Word Classification
- Weather Prediction
- Voting Applications

E. Stochastic Gradient Descent Classifier

Stochastic Gradient Descent (SGD) is a simple yet very efficient approach to discriminative learning of linear classifiers under convex loss functions such as (linear) Support Vector Machines and Logistic Regression [4].

Stochastic gradient descent refers to calculating the derivative from each training data instance and calculating the update immediately [6]. It is particularly useful when the sample data is in a larger number. It supports different loss functions and penalties for classification [6].

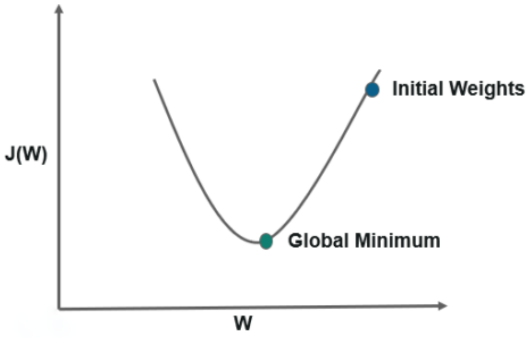

Fig. 6. Stochastic Gradient Descent [6]

Only recently received a lot of attention even though it has been around for a long time, some of its advantages are as follows [6]:

- Efficiency.
- Ease of implementation.

F. Linear Support Vector Machines

Generally, Support Vector Machines (SVM) is considered to be a classification approach, but it can be employed in both classification and regression problems [4]. It can easily handle multiple continuous and categorical variables [4].

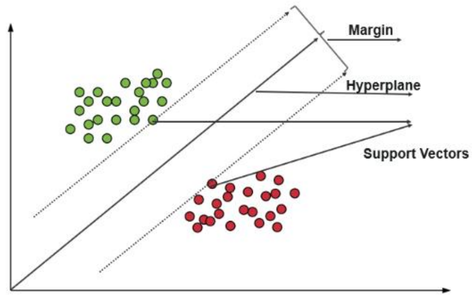

Fig. 7. Support Vectors [6]

Some advantages of SVMs are as follows [6]:

- Effective in high dimensional spaces.
- Effective where no. of dimensions > no. of samples

G. Random Forest

The Random Forest classifiers are suitable for dealing with the high dimensional noisy data in text classification. A Random Forest model comprises a set of decision trees each of which is trained using random subsets of features. Given an instance, the prediction by the Random Forest is obtained via majority voting of the predictions of all the trees in the forest.

Some advantages are [6]:

- They do not overfit.
- Provide a better accuracy than other classification algorithms.

VII. CLASSIFIER EVALUATION

The most important part after the completion of any classifier is the evaluation to check its accuracy and efficiency. There are a lot of ways in which a classifier can be evaluated. Here,

- Class 1: Positive
- Class 2: Negative

Definition of some terms [3]:

- Positive (P): Observation is positive.
- Negative (N): Observation is not positive (i.e. negative).
- True Positive (TP): Observation is positive and is predicted to be positive.
- False Positive (FP): Observation is negative but is predicted positive.
- True Negative (TN): Observation is negative and is predicted to be negative.
- False Negative (FN): Observation is positive but is predicted negative.

Now, the evaluation methods used here are listed below:

A. *Accuracy*

Accuracy is a ratio of correctly predicted observations to the total observations [6].

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

[3] (4)

However, there are problems with accuracy. It assumes equal costs for both kinds of errors. A 99% accuracy can be excellent, good, mediocre, poor or terrible depending upon the problem [3].

B. *Confusion Matrix [3]*

- A confusion matrix is a summary of prediction results on a classification problem.
- The number of correct and incorrect predictions are summarized with count values and broken down by each class.
- In other words, it shows the ways in which a classification model is confused when it makes predictions.
- It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made.

	Class 1 Predicted	Class 2 Predicted
Class 1 Actual	TP	FN
Class 2 Actual	FP	TN

Fig. 8. Confusion Matrix Depiction [3]

VIII. RESULTS AND DISCUSSIONS

The Confusion matrix, and Accuracy, of the six base classifier models are as follows:

A. *Multinomial Naïve Bayes' Classifier:*

```
[[642 223  35]
 [262 576 262]
 [ 42 245 713]]
              precision    recall  f1-score   support

    negative       0.68      0.71      0.70       900
     neutral       0.55      0.52      0.54      1100
    positive       0.71      0.71      0.71      1000

    accuracy                           0.64      3000
   macro avg       0.65      0.65      0.65      3000
weighted avg       0.64      0.64      0.64      3000

0.6436666666666667
```

Fig. 9. Performance of Multinomial Naïve Baiyes' Classifier

B. *Bernouille's Naïve Bayes' Classifier:*

```
[[667 189  44]
 [325 452 323]
 [ 60 155 785]]
              precision    recall  f1-score   support

    negative       0.63      0.74      0.68       900
     neutral       0.57      0.41      0.48      1100
    positive       0.68      0.79      0.73      1000

    accuracy                           0.63      3000
   macro avg       0.63      0.65      0.63      3000
weighted avg       0.63      0.63      0.62      3000

0.6346666666666667
```

Fig. 10. Performance of Bernouille's Naïve Baiyes' Classifier

C. Logistic Regression:

```
[[573 288  39]
 [197 709 194]
 [ 37 248 715]]
              precision    recall  f1-score   support

    negative       0.71      0.64      0.67       900
     neutral       0.57      0.64      0.60      1100
    positive       0.75      0.71      0.73      1000

    accuracy                           0.67      3000
   macro avg       0.68      0.67      0.67      3000
weighted avg       0.67      0.67      0.67      3000

0.6656666666666666
```

Fig. 11. Performance of Logistic Regression Model

D. Stochastic Gradient Descent Classifier:

```
[[574 294  32]
 [244 651 205]
 [ 56 258 686]]
              precision    recall  f1-score   support

    negative       0.66      0.64      0.65       900
     neutral       0.54      0.59      0.57      1100
    positive       0.74      0.69      0.71      1000

    accuracy                           0.64      3000
   macro avg       0.65      0.64      0.64      3000
weighted avg       0.64      0.64      0.64      3000

0.637
```

Fig. 12. Performance of SGD Classifier

E. Linear Support Vector Classifier:

```
[[517 284  99]
 [248 575 277]
 [ 88 294 618]]
              precision    recall  f1-score   support

    negative       0.61      0.57      0.59       900
     neutral       0.50      0.52      0.51      1100
    positive       0.62      0.62      0.62      1000

    accuracy                           0.57      3000
   macro avg       0.58      0.57      0.57      3000
weighted avg       0.57      0.57      0.57      3000

0.57
```

Fig. 13. Performance of SVM Classifier

F. Random Forest Classifier:

```
[[574 294  32]
 [244 651 205]
 [ 56 258 686]]
              precision    recall  f1-score   support

    negative       0.66      0.64      0.65       900
     neutral       0.54      0.59      0.57      1100
    positive       0.74      0.69      0.71      1000

    accuracy                           0.64      3000
   macro avg       0.65      0.64      0.64      3000
weighted avg       0.64      0.64      0.64      3000

0.637
```

Fig. 14. Performance of RF Classifier

The accuracies of all the seven base classifiers and the hybrid classifier together are as follows:

TABLE I. CLASSIFIER ACCURACIES

Classifier	***Accuracy***
Multinomial Naïve Bayes'	64.3667
Bernouille's Naïve Bayes'	63.4667
Logistic Regression	66.5667
Stochastic Gradient Descent	63.7000
Linear Support Vector Classifier	57.0000
Random Forest	63.7000

Comparing the above-mentioned accuracies of all these classifiers, the following line plot can be derived:

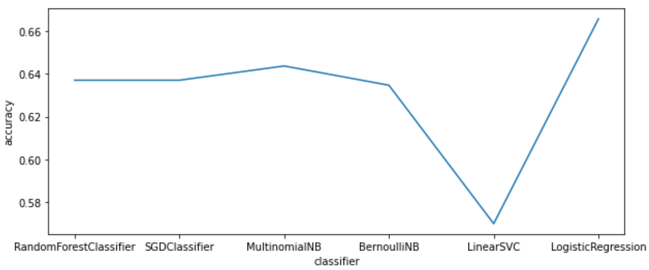

Fig. 15. Comparison of Classifier Accuracies Graph

It can be observed from the Accuracy graph that the Logistic Regression provides the best accuracy, while Linear Support Vector Classifier has the least accuracy.

IX. CONCLUSION AND FUTURE WORK

Some key features that were brought out during the course of this study helped us to get a broader understanding of the subject and gave us the necessary encouragement to dive deeper. The study conducted is unique in multiple ways which act as a differentiation factor when compared to similar work. The methodology followed, helped to create an aggregated model consisting of all the classifiers used during the process. On a personal note, I was able to get familiar with the usage and implementation of different classifiers. We also understood which classifiers work when used on a certain type of data. This also helped us to familiarize ourselves with the advantages and disadvantages of each particular classifier. If we were to look towards the future to enhance the study, we could find a larger and better dataset to work with. We can try to use better translation techniques and give a try at more complex machine learning models for the classification of text. The complexities of each of these models can also be found and compared to give the best efficiency. Furthermore, this study can be extended to multiple regional languages and can be translated into a multilingual sentiment analysis problem, making it more generic.

ACKNOWLEDGMENT

The satisfaction and euphoria that accompany the successful completion of any task would be but incomplete without mentioning the people who made it possible, whose constant guidance crowned our efforts with success. With a profound sense of gratitude, I acknowledge the guidance and support extended by Dr. Labiba Fahad CS Department, FAST NUCES University. I would also like to thank our instructor, Mr. Umair Arshad, FAST NUCES University for providing me with this opportunity to work on this particular project work. The guidance I received gave me the environment to enhance my knowledge, skills and to reach the pinnacle with sheer determination, dedication and hard work.

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

[3] Sahil Swami, Ankush Khandelwal, Vinay Singh, Syed S. Akhtar and Manish Shrivastava, "A Corpus of English-Hindi Code-Mixed Tweets for Sarcasm Detection," 19th International Conference on Computational Linguistics and Intelligent Text Processing, March 2018.

[4] https://towardsdatascience.com/advanced-ensemble-classifiers-8d7372e74e40

[5] https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623

[6] http://languagelinguistics.com/2017/06/27/code-mixing-sociolinguistics/

[7] https://www.edureka.co/blog/classification-in-machine-learning/

[8] https://en.wikipedia.org/wiki/Naive_Bayes_classifier#Multinomial_naive_Bayes