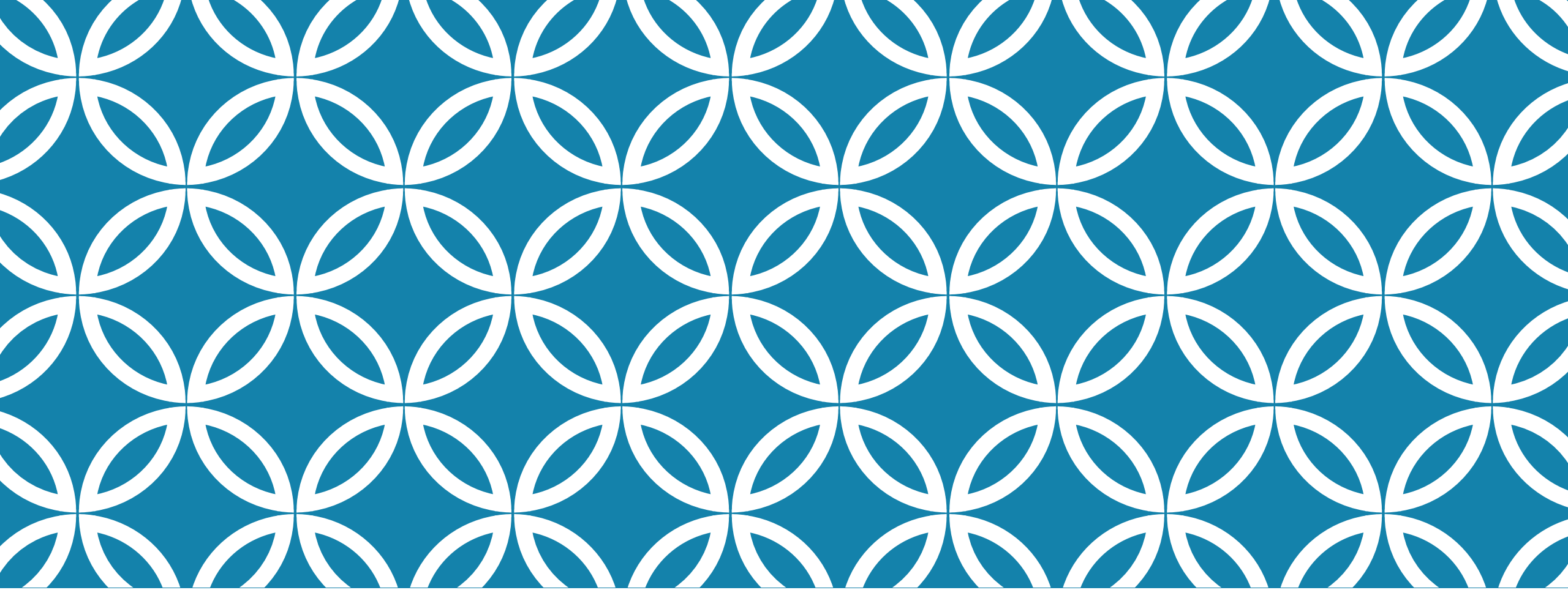


# Bilingual Sentiment Analysis

Sakshi Goel PES1201700148
Suhail Rahman PES1201701420
UE17CS333 Project Submission

# ABOUT THE PROJECT

- The main aim of the project is to develop a sentiment analyzer that can be used on twitter data to classify it as positive or negative.

- Our project takes care of the challenge of bilingual comments, where people tweet in two languages, in this case Hindi and  English, in the English Alphabet.

# UNIQUENESS AND ANALYSIS

- We created an aggregated model consisting of all the classifiers used during the process. The ensemble model created worked to our advantage as we saw in the previous slides that it provided one of the highest accuracy compared to other classifiers.

- When a sentence is in Hindi, we use Google Translate to directly convert it to English. If the sentence consists of a combination of Hindi and English, we make use of TextBlob to identify that.

- We can observe that using this approach of both the platforms, increased our accuracy significantly when compared to using them individually.

# DATASET SOURCE

- 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, 4 = positive) and they can be used to detect sentiment.

- The two columns that we mainly need are as follows:
  - The Label
  - The Tweet

# DATASET SOURCE

- The format of the Tweet column was not useful and had to be cleaned and tokenized. We also limited the number of tweets to 40 thousand.

| Label | number | date | no_query | name | Tweet |
|---|---|---|---|---|---|
| 0 | 1467810369 | Mon Apr 06 22:19:45 PDT 2009 | NO_QUERY | _TheSpecialOne_ | @switchfoot http://twitpic.com/2y1zl - Awww, t... |
| 0 | 1467810672 | Mon Apr 06 22:19:49 PDT 2009 | NO_QUERY | scotthamilton | is upset that he can't update his Facebook by ... |
| 0 | 1467810917 | Mon Apr 06 22:19:53 PDT 2009 | NO_QUERY | mattycus | @Kenichan I dived many times for the ball. Man... |
| 0 | 1467811184 | Mon Apr 06 22:19:57 PDT 2009 | NO_QUERY | ElleCTF | my whole body feels itchy and like its on fire |
| 0 | 1467811193 | Mon Apr 06 22:19:57 PDT 2009 | NO_QUERY | Karoli | @nationwideclass no, it's not behaving at all.... |
| ... | ... | ... | ... | ... | ... |
| 4 | 1960186342 | Fri May 29 07:33:44 PDT 2009 | NO_QUERY | Madelinedugganx | My GrandMa is making Dinenr with my Mum |
| 4 | 1960186409 | Fri May 29 07:33:43 PDT 2009 | NO_QUERY | OffRoad_Dude | Mid-morning snack time... A bowl of cheese noo... |
| 4 | 1960186429 | Fri May 29 07:33:44 PDT 2009 | NO_QUERY | Falchion | @ShaDeLa same here say it like from the Termi... |
| 4 | 1960186445 | Fri May 29 07:33:44 PDT 2009 | NO_QUERY | jonasobsessedx | @DestinyHope92 im great thaanks wbuu? |
| 4 | 1960186607 | Fri May 29 07:33:45 PDT 2009 | NO_QUERY | sugababez | cant wait til her date this weekend |

# DATASET PREPROCESSING

- Chose the relevant columns that were required for our study, which were the tweet and the sentiment associated.

- If there were any emoticons used, we converted them into their equivalent emotion that they are trying to signify, while emojis were removed.

- We also expanded some words which were joined together such as “Can’t” was changed to “Can not”.

# DATASET PREPROCESSING

- Removal of numbers, URLs, html tags and symbols, the "@" symbol followed by the account handle.

- These were all some data cleaning steps that were important to the study to function effectively. Finally, the dataset contained the cleaned tweets which we converted to lowercase for simplicity.

- Certain features, like adjectives, abstract nouns and adverbs were focused on and the rest of the words were removed as they did not add any value to the sentiment.

# LITERATURE REVIEW - TABLE 1

| Papers | Title | Authors | Methodology Used |
|---|---|---|---|
| Paper 1 | Machine translation of bi-lingual Hindi-English (Hinglish) text | R. Mahesh, K.Sinha, Anil Thakur | Makes use a system designed specifically to separate out the Hindi and English parts of a word that has a combination of the two. |
| Paper 2 | Towards Sub-Word Level Compositions for Sentiment Analysis of Hindi-English Code Mixed Text | Aditya Joshi,Ameya Prabhu Pandurang, Manish Shrivatsava and Vasudeva Varma | Introduces a constantly learning sub-word level representation in LSTM (Subword-LSTM) architecture instead of character-level or word-level representations. |

# LITERATURE REVIEW - TABLE 1

| | | | |
|---|---|---|---|
| Paper 3 | A Dataset of Hindi-English Code-Mixed Social Media Text for Hate Speech Detection | Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed S. Akhtar and Manish Shrivatsava | Makes use of a system created that classifies a tweet having a combination of Hindi and English to negative or not. |
| Paper 4 | Resource Creation for Hindi-English Code Mixed Social Media Text | Sakshi Gupta, Piyush Bansal and Radhika Mamidi | Proposes a method to successfully aggregate data to form a dataset of words that have a multilingual characteristic. |
| Paper 5 | Sentiment classification of Hinglish text | Kumar Ravi and Vadlamani Ravi | Made use of different combinations of feature selection methods and a host of classifiers using term frequency-inverse document frequency feature representation. |

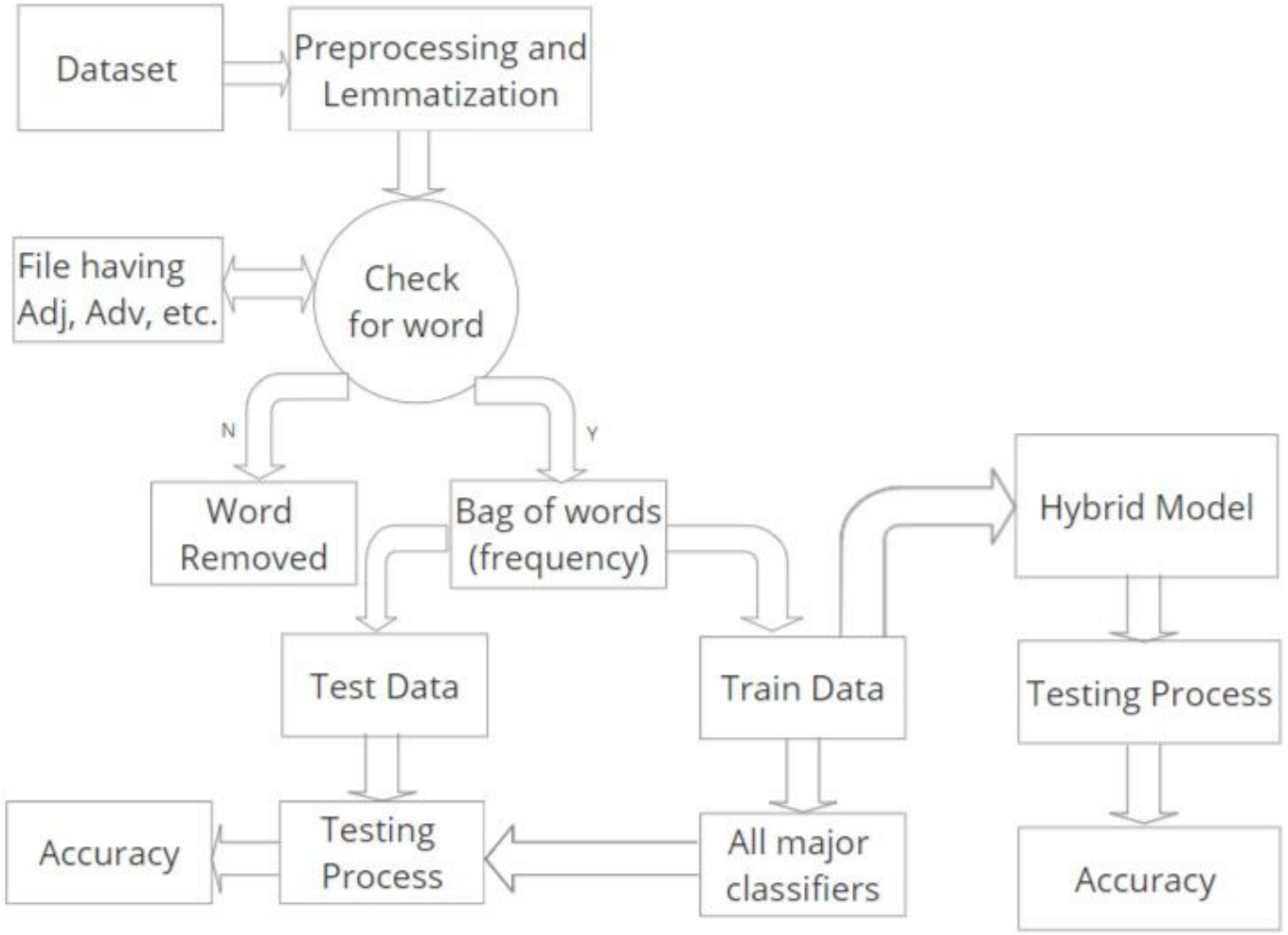
# LITERATURE REVIEW - TABLE 2

| Papers | Accuracy | Benefits | Drawbacks |
|---|---|---|---|
| Paper 1 | 90% | The strategy described here is equally applicable to all Indian languages as these are verb ending languages and have similar mixture of lexicons as in case of Hindi. | Elaborate testing is not possible as these languages are used in verbal communication. |
| Paper 2 | 69.7% | Sub-Word LSTM interprets sentiment based on morpheme-like structures and the results thus produced are significantly better than baselines. | The lexicon lookup approach didn't perform well owing to the heavily misspelt words in the text, which led to incorrect transliterations. |

# LITERATURE REVIEW - TABLE 2

| | | | |
|---|---|---|---|
| Paper 3 | 71.7% | The features used in the classification system are character n-grams, word n-grams, punctuations, negation words and hate lexicon which are integrated in the SVM as the classification system. | The corpus was not annotated with part-of-speech tags at word level which would have yield better results. |
| Paper 4 | 89.94% | They have used an existing language identification system, and improved a normalisation system, achieving a higher accuracy than the base system. | Have not taken into consideration the sentence-level context for word disambiguation. |
| Paper 5 | AUC = 0.8601 | Proposed a triumvirate of TF-IDF, GR, and RBFNN, which is found as the best combination for classifying sentiment expressed in the Hinglish text. | Did not employ sentence parser for considering relation between different parts-of-speech of a sentence. |

# BLOCK DIAGRAM FOR IMPLEMENTATION



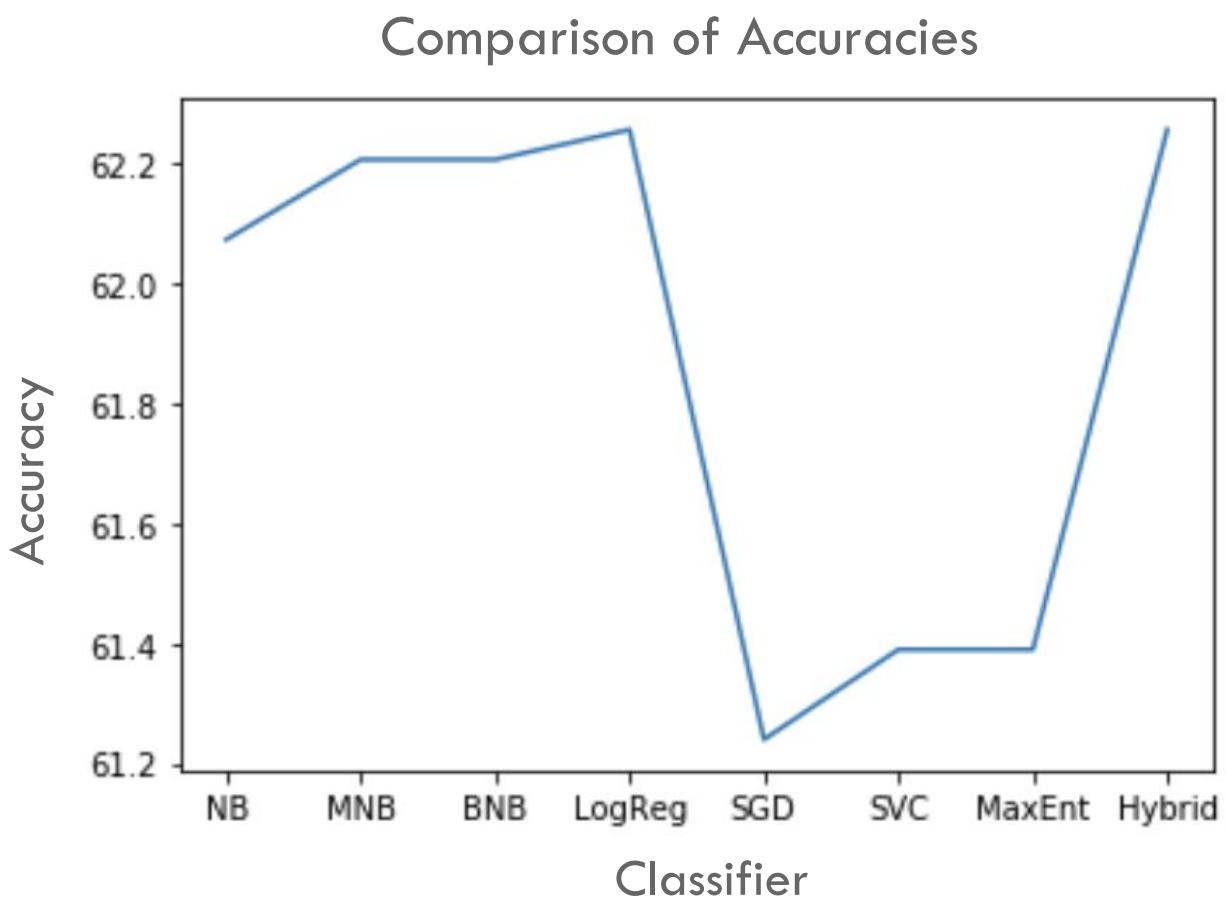
# QUANTITY OF WORK – THE MAIN CODE MODULES

| Sl. No. | Code Module Description | Status (% completed) | Comments |
|---|---|---|---|
| 1. | func(test_text) | 100% | The master module |
| 2. | hinglish(test_text) | 100% | Takes care of text translation |
| 3. | text_classify(text) | 100% | Classifies text using all 8 models |
| 4. | hybrid(test_set_formatted) | 100% | Builds the hybrid model classifier |
| 5. | features(test_text) | 100% | Filters features from the text |
| 6. | start(text) | 100% | Preprocessing module |

# QUALITY OF WORK – MILESTONES THAT ARE DONE AND WORKING

| Serial no | Milestone description | Status (% complete) | Comments |
|---|---|---|---|
| 1. | Dataset Selection | 100% | A better dataset can be used. |
| 2. | Preprocessing | 100% | Cleaning done efficiently. |
| 3. | Feature Selection | 100% | Adjectives, Abstract Nouns, Adverbs |
| 4. | Choice of Classifiers | 100% | 7 Classifiers chosen. |
| 5. | Building Classifiers | 100% | Successfully built |
| 6. | Training Classifiers | 100% | Trained on 85% data. |
| 7. | Creation of Hybrid Model | 100% | Voting Based Ensemble Model. |
| 8. | Translation Challenge | 100% | Google Translate Machine, TextBlob |
| 9. | Creating a controller module | 100% | func module combines all functionality. |

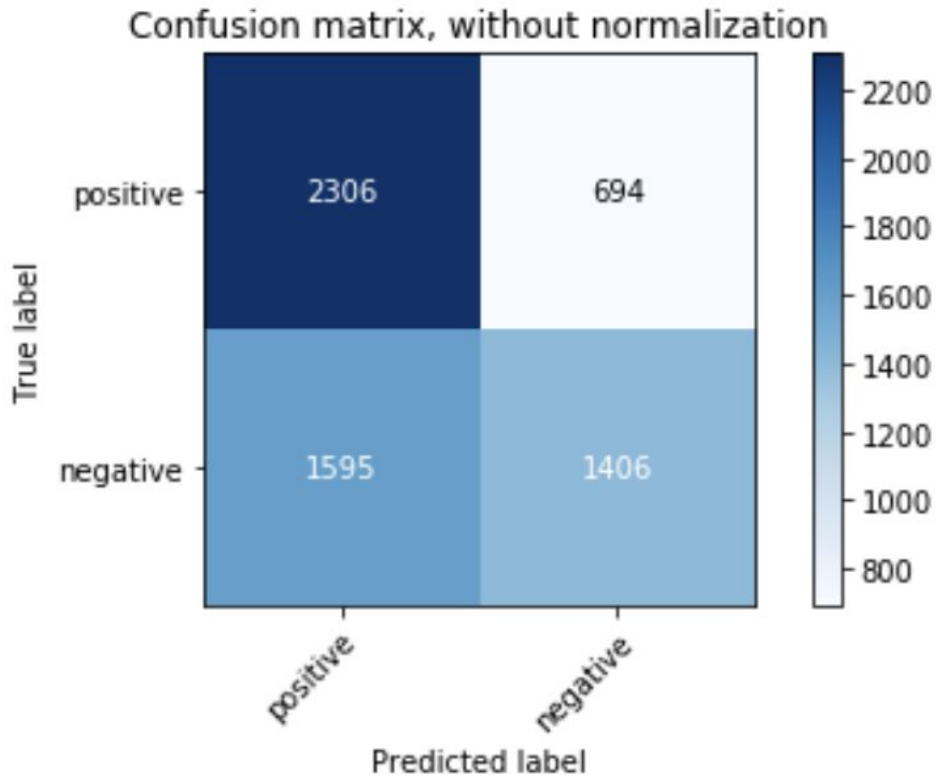
# RESULTS OBTAINED - Accuracy

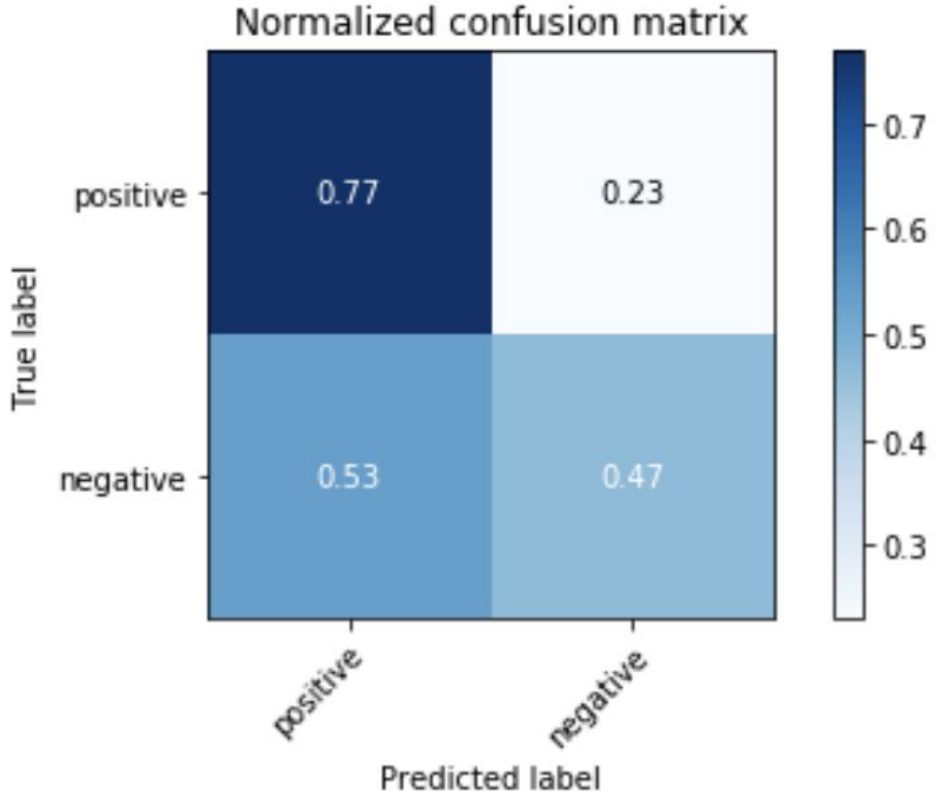



| Classifier Used | Accuracy |
|---|---|
| Naive Bayes | 62.0729 |
| Multinomial Naive Bayes | 62.2062 |
| Bernoulli Naive Bayes | 62.2062 |
| Logistic Regression | 62.2562 |
| SGD | 61.2397 |
| SVC Classifier | 61.3897 |
| Max Entropy | 613897 |
| Hybrid Model | 62.2563 |

# RESULTS OBTAINED - Confusion Matrix

For Hybrid Model:



# RESULTS OBTAINED - F1 Score

**Naive Bayes' Classifier:**

| | precision | recall | f1-score |
|---|---|---|---|
| positive | 0.59 | 0.78 | 0.67 |
| negative | 0.68 | 0.46 | 0.55 |

**Bernouille's Naive Bayes' Classifier:**

| | precision | recall | f1-score |
|---|---|---|---|
| positive | 0.59 | 0.79 | 0.67 |
| negative | 0.69 | 0.46 | 0.55 |

# RESULTS OBTAINED - F1 Score

**Multinomial Naive Bayes' Classifier:**

| | precision | recall | f1-score |
|---|---|---|---|
| positive | 0.59 | 0.78 | 0.67 |
| negative | 0.68 | 0.46 | 0.55 |

**Logistic Regression Classifier:**

| | precision | recall | f1-score |
|---|---|---|---|
| positive | 0.59 | 0.78 | 0.67 |
| negative | 0.68 | 0.47 | 0.55 |

# RESULTS OBTAINED - F1 Score

**Stochastic Gradient Descent Classifier:**

| | precision | recall | f1-score |
|---|---|---|---|
| positive | 0.69 | 0.40 | 0.51 |
| negative | 0.58 | 0.82 | 0.68 |

**Support Vector Machines Classifier:**

| | precision | recall | f1-score |
|---|---|---|---|
| positive | 0.58 | 0.79 | 0.67 |
| negative | 0.68 | 0.44 | 0.53 |

# RESULTS OBTAINED - F1 Score

**Maximum Entropy Classifer:**

| | precision | recall | f1-score |
|---|---|---|---|
| positive | 0.68 | 0.40 | 0.51 |
| negative | 0.58 | 0.82 | 0.68 |

**Hybrid Model:**

| | precision | recall | f1-score |
|---|---|---|---|
| positive | 0.59 | 0.79 | 0.67 |
| negative | 0.69 | 0.46 | 0.55 |

# OUR TOP THREE LEARNING IN THIS PROJECT

1. We were able to get familiar with the usage and implementation of different classifiers.

2. Understanding which classifiers work when used on a certain type of data. Learning the advantages and drawbacks of the used classification models.

3. Getting the opportunity to create an ensemble model to give us optimal results.

# TOP CHALLENGES UNRESOLVED SO FAR

1. Accuracy for the testing of the models was around 60%, even after several efforts to increase it.

2. Two separate modules, instead of one, used for translation.

3. Dataset used for training could be a better one.

# OUR GOING FORWARD PLAN (IF ANY)

1. Find a better dataset to work with.

2. Try more complex machine learning models for the classification of text.

3. Use better translation techniques.