CS689A PROJECT PRESENTATION 2022-2023

SENTIMENT ANALYSIS ON HINDI TEXT

Presented By:
Allan Robey – 22111007
Avnish Tripathi – 22111014
Divyesh Tripathi - 22111020

CONTENTS

1. Introduction
2. Objectives
3. Problem Statement
4. Methodology
5. Experimental Set Up and Results Analysis
6. Conclusion
7. Future Scope

INTRODUCTION

There are enormous works in the field of sentiment analysis that are available in English. There hasn't been much effort done on Indian languages, though.

However, today there is an explosive growth of information on the web in native Indian languages such as Hindi, Marathi, Gujarati, Malayalam, Tamil, etc. This data will be of no use if they are not classified on the basis of their sentiments.

Sentiment analysis in Hindi is necessary since opinion-rich sites have exploded in Hindi during the past few years.

➢ **Challenges in performing sentiment analysis of Hindi text**

unavailability of well annotated standard corpus

scarcity of resources in Hindi language

this language lacks effective taggers and parsers

PRIMARY OBJECTIVES

To collect and create a well annotated Hindi dataset

To determine the most efficient methodology to annotate the Hindi Dataset

To determine the most efficient method of featurization among Bag of words, TF-IDF and BM25 for various classifiers

To determine the best classifier to perform sentiment analysis of Hindi text for different methods of featurization

PROBLEM STATEMENT

- There is an urgent need to analyse the data available on the web so as to determine the opinion of the people based on the sentiment of the text.
- It can then help the business people by tackling the customer's opinion about the particular product so that they can modify changes if it is required.
- In order to perform sentiment analysis, there is a need of a well annotated data. However, there is a scarcity of resources in Hindi language, which makes it difficult to collect the data and create the necessary annotated datasets.
- The project aims to create a well annotated corpora for Hindi language as well as determine the most efficient method for feature matrix generation.
- The project also aims at determining the most efficient method for classifying the Hindi text into different classes such as positive, negative and neutral based on the sentiment.

METHODOLOGY

1. DATASETS

2. DATA LABELLING PROCESS

3. DATA PRE-PROCESSING

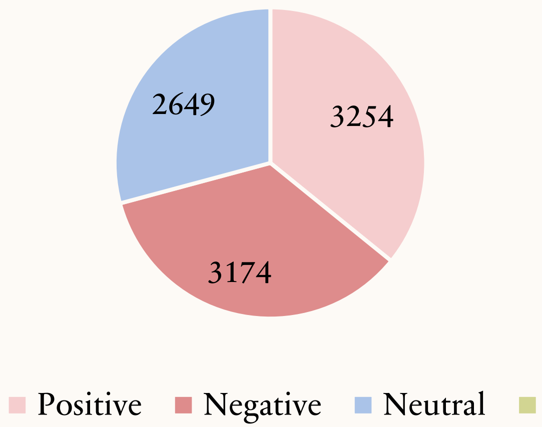
4. FEATURE MATRIX GENERATION

5. CLASSIFICATI ON

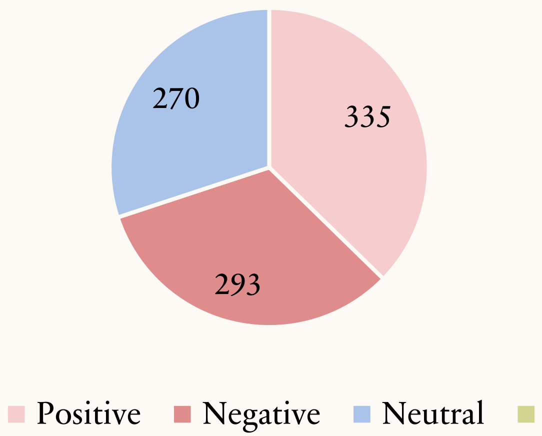
DATASETS

Dataset	No of Sentences	Purpose
News Dataset in Hindi	9077	To determine the most efficient model for data annotating task
Movie Reviews in Hindi **https://www.kaggle.com/datasets/disisbig/hindi-movie-reviews-dataset**	898	To perform classification task
Self – Annotated Hindi Dataset	1,00,000	Annotated using the determined most efficient deep learning model to perform the classification task

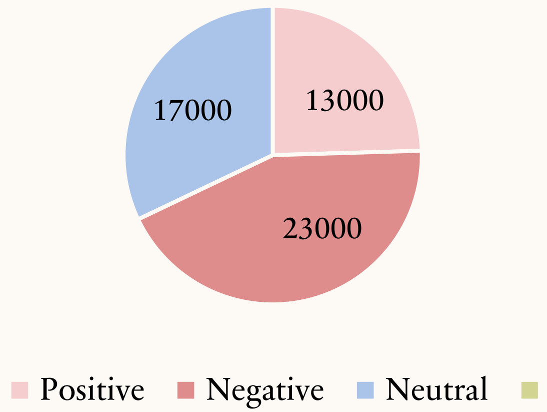
Hindi News Dataset

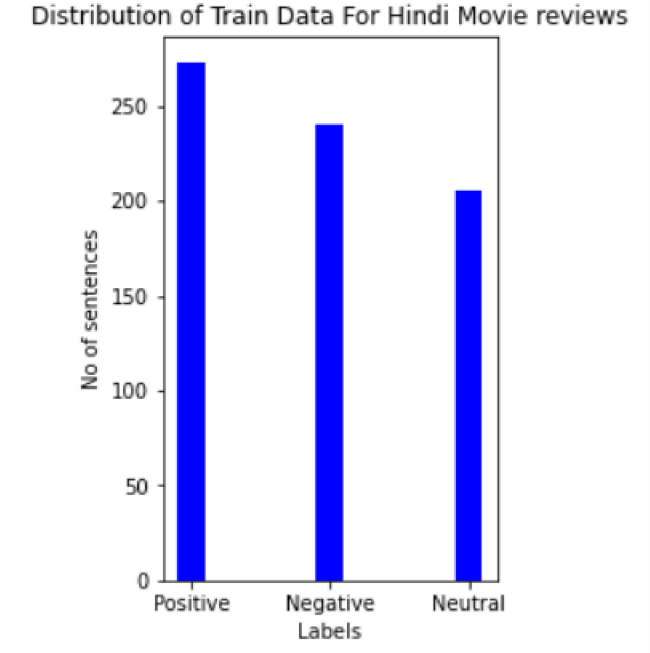


Movie Reviews Dataset

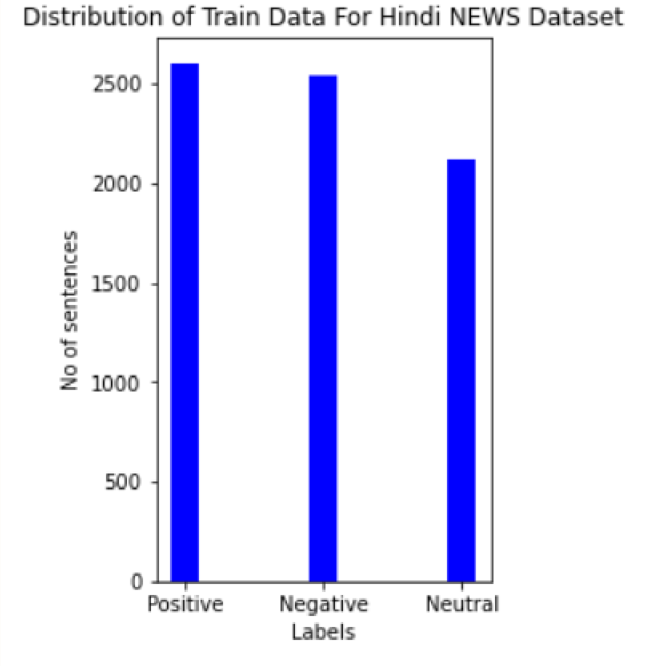


Self-Annotated Hindi Dataset

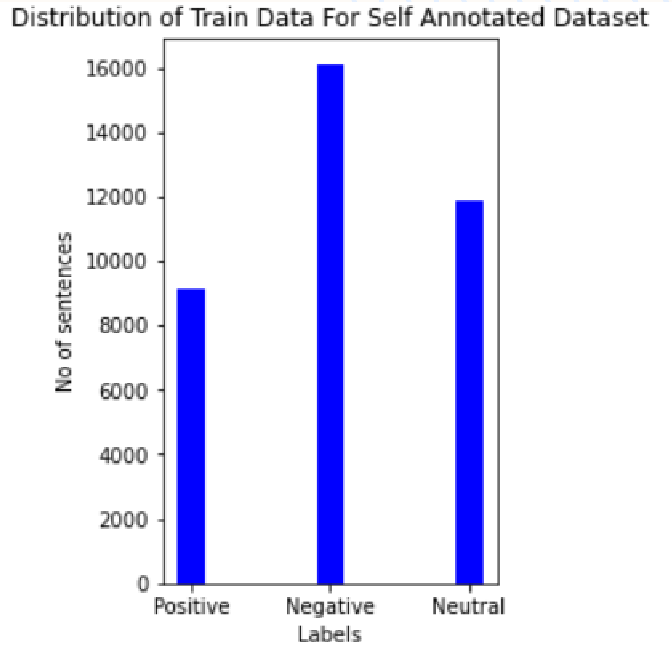




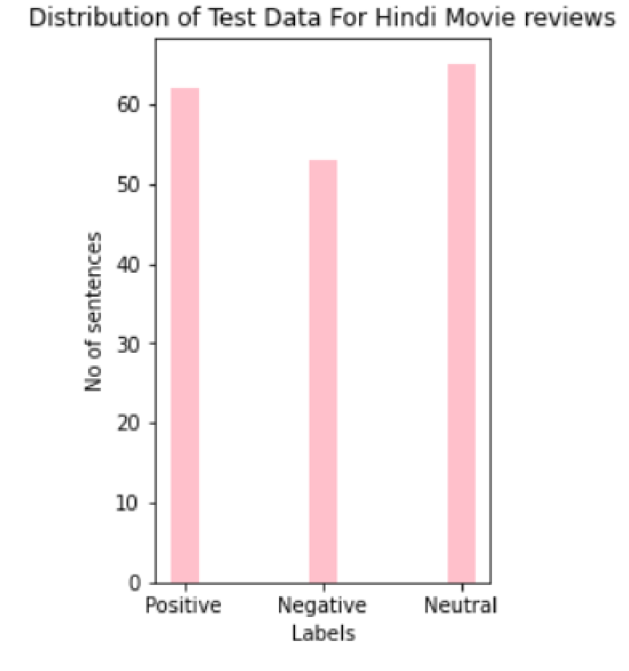
Distribution of Train Data For Hindi Movie reviews



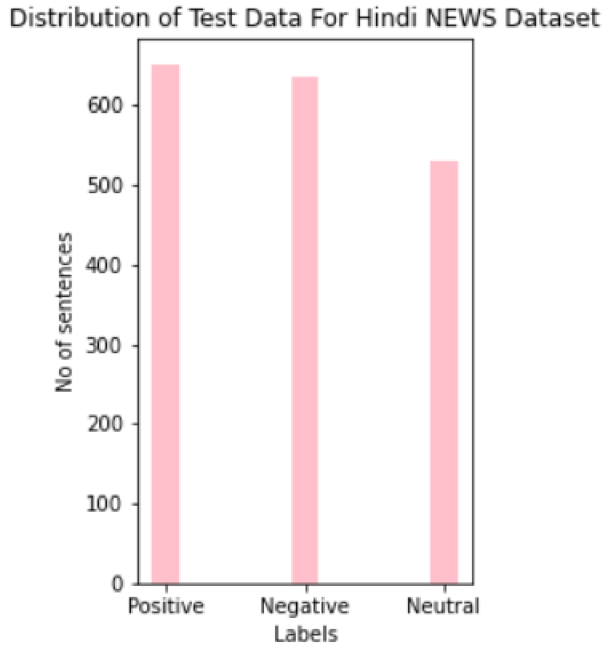
Distribution of Train Data For Hindi NEWS Dataset



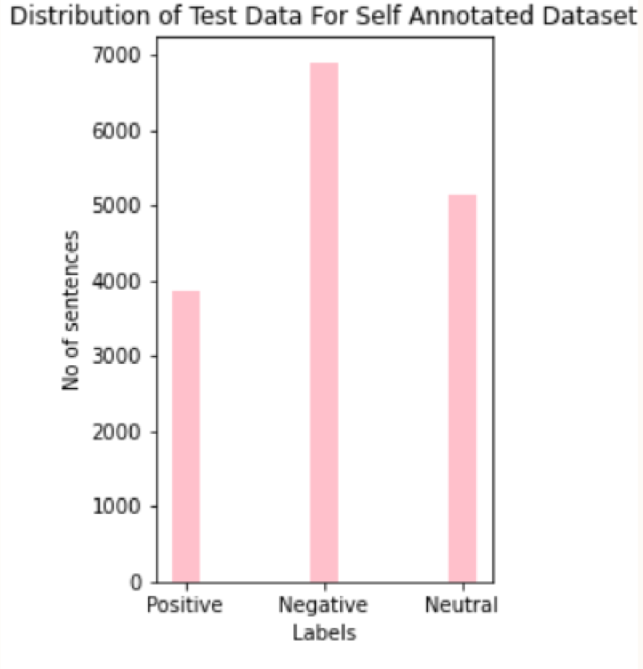
Distribution of Train Data For Self Annotated Dataset



Distribution of Test Data For Hindi Movie reviews



Distribution of Test Data For Hindi NEWS Dataset



Distribution of Test Data For Self Annotated Dataset

DATA LABELLING PROCESS

In order to annotate to Hindi text data different pre-trained deep learning models such as Vader, Roberta, Bert and TextBlob were used.

PRE-TRAINED MODELS	ACCURACY
1. Vader Sentiment Scoring	62.89%
2. Roberta Pre-trained Model	**73.56%**
3. bert-base-multilingual-uncased-sentiment	65.28%
4. Text Blob	48.53%

In order to annotate the data, the Hindi text were initially translated to English using GoogleTranslator from deep_translator.

Then the Roberta pre-trained deep learning model was applied on the annotated data. Out of 1,00,000 sentences 53,000 sentences were considered of which comprises of 13,0000 sentences belonging to positive class, 23,000 sentences belonging to negative class and

DATA PRE-PRCOSSEING - eliminated any punctuation, digits, one-length words, and stop words from each sentence that do not improve the classification's accuracy.

FEATURE MATRIX GENERATION

1. TF-IDF

2. BAG OF WORDS

3. BM25

CLASSIFICATION

1. GAUSSIAN NAÏVE BAYES

Naive Bayes is a classification method that relies on the independence of predictors assumption and is based on Bayes' Theorem.

2. LOGISTIC REGRESSION

The cumulative logistic distribution is the logistic function used in logistic regression to estimate probabilities

3. SVM

It looks for a hyperplane that can distinguish two classes of data with the greatest degree of precision

4. KERNELIZED SVM

It uses a set of mathematical operations that provide the kernel the ability to manipulate the data

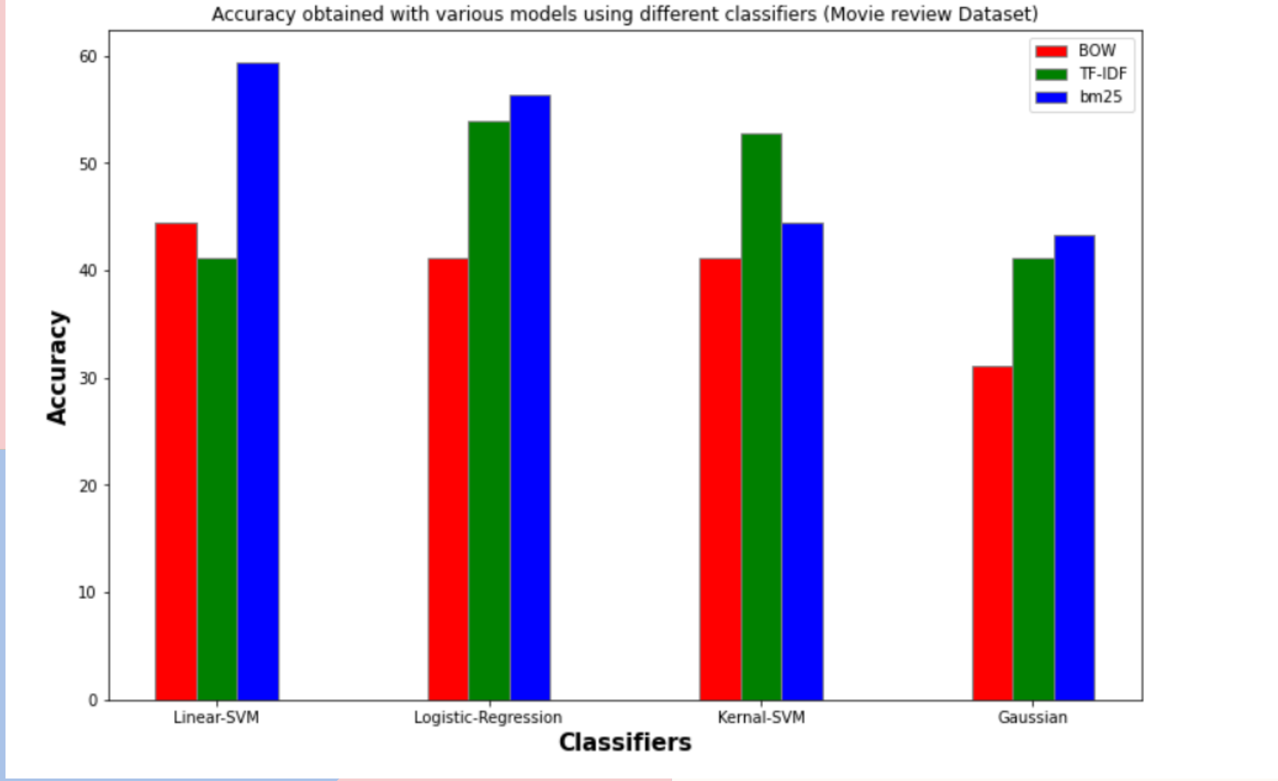
EXPERIMENTAL SET UP AND RESULT ANALYSIS

The pre-trained deep learning models such as Vader, Roberta, TextBlob and bert-base-multilingual-uncased-sentiment model were selected to annotate the Hindi unlabelled dataset. The performance of these deep learning models was determined on the Hindi News dataset. The pre-trained Roberta deep learning model has been determined as the most efficient model for annotating the Hindi Dataset comprising of 1,00,000 sentences as it gave the highest accuracy of %. In order to validate the accuracy of the model we have also manually checked the predicted label of 1,000 sentences of which 869 were found to be correct.

In order to annotate the data, the sentences had to be initially translated to English language from Hindi language using GoogleTranslator from deep_translator. However, translating 1,00,000 sentences using GoogleTranslator is a tedious and cumbersome task as it is extremely time consuming. Hence, the data has been split sequentially into different parts and has been translated on multiple systems in order to speed-up the process of translating 1,00,000 Hindi sentences. Once the data has been translated to English, the translated data has been fed into the Roberta Deep learning model in order to annotate the data. As it an intensive task, it has carried out using GPU in order to enhance the speed-up.

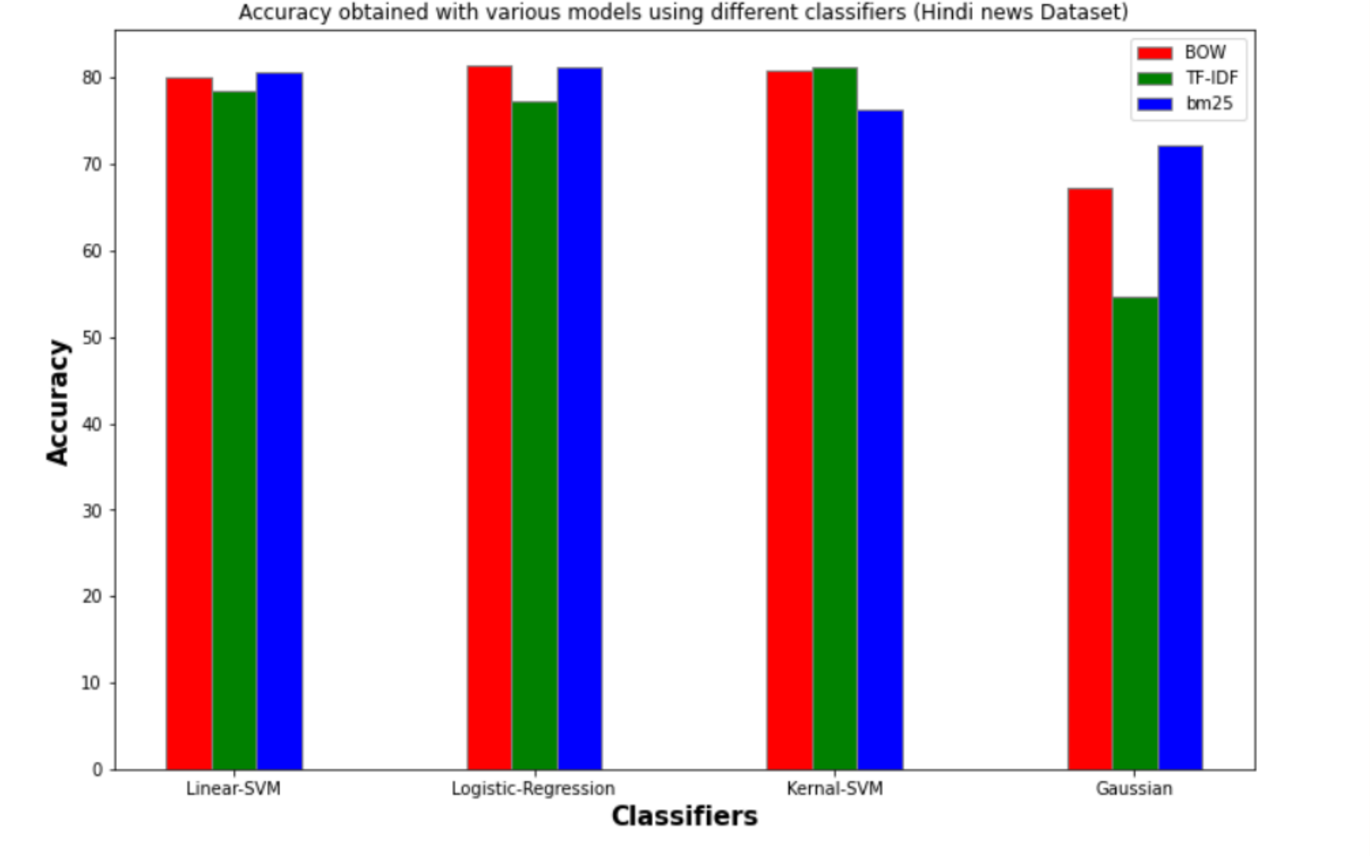
Once the Hindi data has been annotated as positive, negative and neutral using the deep learning model then the accuracy obtained by various models such as TF-IDF, Bag of Words and BM25 using different classifiers such as Gaussian Naïve Bayes, Logistic Regression, SVM and Kernalized SVM has been analysed.

ACCURACY OBTAINED WITH VARIOUS MODELS USING DIFFERENT CLASSIFIERS ON MOVIE REVIEWS DATASET



Models used	Gaussian Naïve Bayes	Logistic Regression	Support Vector Machine	Kernalized SVM
Bag of Words	31.11%	41.11 %	44.44 %	41.11 %
TF-IDF	41.11 %	53.33 %	53.88 %	52.77 %
BM25	43.33 %	58.33 %	59.44 %	44.44 %

ACCURACY OBTAINED WITH VARIOUS MODELS USING DIFFERENT CLASSIFIERS ON HINDI NEWS DATSET



Models used	Gaussian Naïve Bayes	Logistic Regression	Support Vector Machine	Kernalized SVM
Bag of Words	67.29 %	81.49 %	80.12 %	80.78 %
TF-IDF	54.625 %	77.2 %	78.48 %	81.27 %
BM25	72.68 %	81.27 %	80.61 %	76.21 %

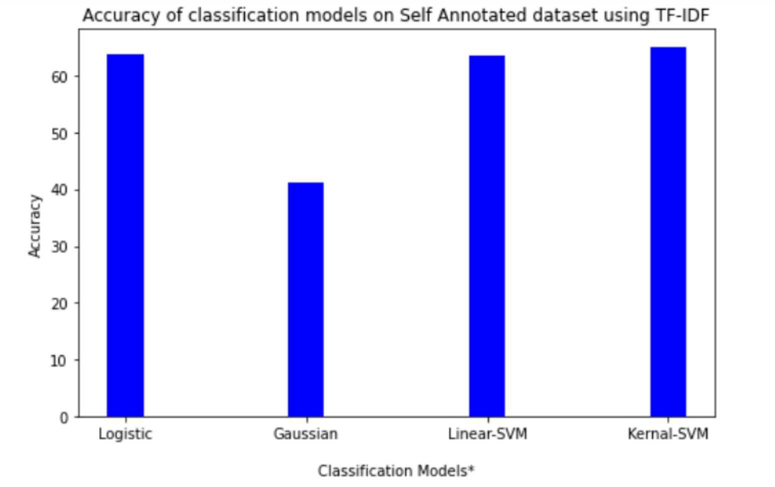
ACCURACY OBTAINED WITH VARIOUS MODELS USING DIFFERENT CLASSIFIERS ON SELF-ANNOTATED HINDI DATASET

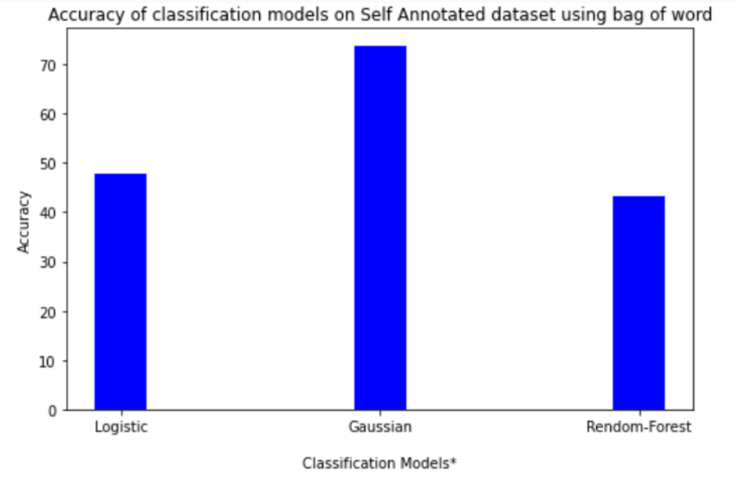
1. TF - IDF

Classifier	Accuracy
Gaussian Naïve Bayes	41.11 %
Logistic Regression	63.79 %
Linear SVM	63.52 %
Kernelized SVM	64.88 %

2. Bag of Words

Classifier	Accuracy
Gaussian Naïve Bayes	73.56 %
Logistic Regression	47.68 %
Random Forest	43.39 %





CONCLUSION

- ❖ In this project we have focused majorly on analysing three aspects of sentiment analysis. Firstly, collecting the Hindi dataset from various sources and creating a well annotated Hindi dataset.
- ❖ Secondly, determining the most efficient deep learning model to annotate the data among pre-trained deep learning models such as Vader, Roberta, TextBlob and bert-base-multilingual-uncased-sentiment model.
- ❖ The results have proven that Roberta is the most efficient deep-learning model to annotate the Hindi data among the mentioned models. Thus, in this project we have successfully annotated 1,00,000 Hindi sentences using Roberta pre-trained model.
- ❖ Finally, most efficient method of featurization among Bag of words, TF-IDF, unigram, bigram and BM25 for various classifiers such as Gaussian Naïve Bayes, Logistic Regression, SVM and Kernelized SVM has been determined successfully.
- ❖ The best classifier to perform sentiment analysis of Hindi text for different methods of featurization has also been determined successfully.

FUTURE SCOPE

- The project has been limited to perform sentiment analysis on Hindi language only.
- The project can be extended to perform sentiment analysis on many unrecognized and low-resourced Indian languages in the future.
- The size of the Hindi corpora can be extended to incorporate more data.
- The project can be extended to incorporate and determine other efficient methods in order to annotate the data.

SUMMARY

“ Thus in this project we have performed analysis on Sentiment analysis of Hindi text rather than sentiment analysis on Hindi Text by performing different approaches of Sentiment Analysis. “

MEET OUR TEAM

ALLAN ROBEY

22111007

AVNISH TRIPATHI

22111014

DIVYESH TRIPATHI

22111020

THANK YOU

- Allan Robey

allanrobey22@iitk.ac.in

- Avnish Tripathi

avnisht22@iitk.ac.in

- Divyesh Tripathi

divyeshdt22@iitk.ac.in