Computational Linguistics (DSE 308)
Final Report

Members **-** 1) Sanket S Houde (18387)
2) Saumya Jain (18234)

Introduction

In the era of the world wide web, like most things, journalism has also seen a revolutionary change. Online media has altered the nature, distribution and representation of journalism. While it has helped boost accessibility, transparency and involvement of local communities, it has also posed problems about authenticity and manipulation of facts. Therefore, it is important not only to segregate fake news, but also opinionated news from fact-based news. This is not just a question of improving the quality of news and narrowing down our search for manipulative, possibly harmful news articles, but also to categorize different types of journalism into objective and subjective.

A fact is a statement that can be proven either right or wrong. On the other hand, an opinion is an expression of a group or individual's feelings. Opinionated news may be based on facts and factual news may contain opinions too. So, it is not always easy to have a clear distinction between the two. Our humanly biases may also come into play.

In this project, we have taken datasets of fact-based and opinion-based news articles, tried to analyse the similarities and differences between them, and built various models, training and testing them, and analysing their accuracies. This is a first step towards creating a fact vs opinion based news-classifier.

Linguistic Theories and Machine Learning models used

Assumptions/Theories:

1. Any article falls to either the opinion-based or the fact-based category.
2. Articles from certain categories were all assumed to be fact-based, as described in the Datasets section below.
3. It is assumed that the PoS tagger used tagged the training set with 100% accuracy.

Models:

We tried out quite a few conventional Machine learning models to classify the articles -

- *Multinomial Naive Bayes Classifier*
  Naive Bayes is based on Bayes' theorem, where the adjective Naïve says that features in the dataset are mutually independent. Occurrence of one feature does not affect the probability of occurrence of the other feature.
- *Random Forest Classifier*
  Random Forest is an ensemble method, meaning that a random forest model is made up of a large number of small decision trees, called estimators, which each produce their own predictions. The random forest model combines the predictions of the estimators to produce a more accurate prediction.
- *K Nearest Neighbors Classifier*
  KNN algorithm is used to classify by finding the K nearest matches in training data and then using the label of closest matches to predict. Traditionally, distance such as euclidean is used to find the closest match.
- *Support Vector Classifier*
  Support Vector Classifier is a supervised machine learning algorithm where each data item is plotted as a point in n-dimensional space (where n is number of features) with the value of each feature being the value of a particular coordinate. Then, classification is performed by finding the hyper-plane that differentiates the two classes very well.

Datasets

The datasets used were taken from two UK based news publishing houses - 'BBC' and 'The Guardian'. The *fact based articles* were 'scraped' from different sections like politics, technology, sports, entertainment and business on the BBC News website. The *opinionated articles* were extracted from the 'Opinion' section on The Guardian website. This consisted of editorials and opinion pieces from columnists.

For the final litmus test, we compiled a rather small dataset of 10 articles (due to time constraints) from 'The Hindustan Times' and the articles were specific to events revolving around India. *Opinionated articles* were compiled from the "Opinion" section and the *fact based articles* from sections under 'cities', 'tech', 'education' and 'cricket'.

A Comparative Study:

A. On Words

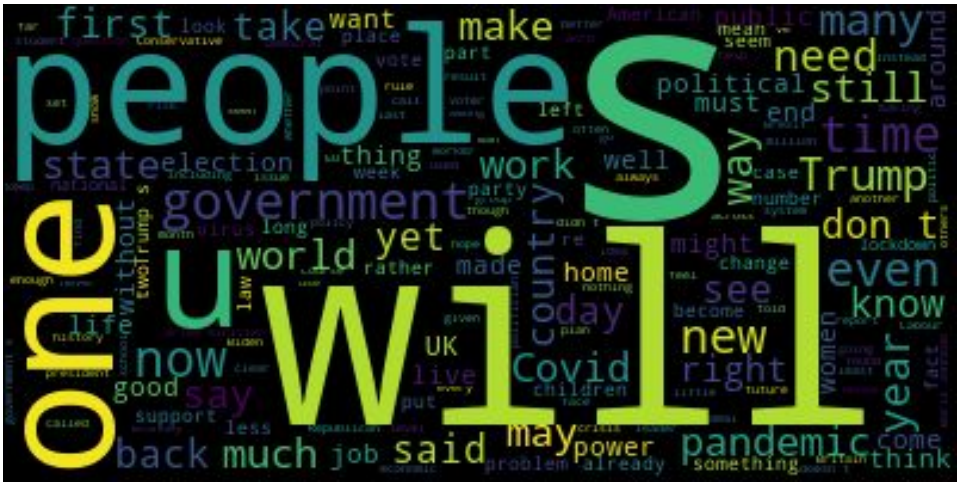
Preprocessing done here:
1. Punctuation removed. 2. Everything converted to lowercase.

An analysis on the most common words appearing in

1) a) Fact-based Articles:

Top words: ('said', 'would', 'year', ‘also', 'people')

Vs b) Opinion-based Articles:



Top Words: ('The', 'people', 'would', 'Trump', 'government')

Then the same analysis is done after removing stopwords.

2) a) Fact-based Articles:

Top words: (‘said’, ‘also’, ‘would’, ‘people’, ‘one’)

Note that this list is the same as the one when stop words were not removed.

b) Opinion-based Articles:

Top words: (‘people’, ‘one’, ‘would’, ‘government’, ‘Trump’)

*Note that there's a slight change in the list after removing stopwords.***A few remarks:**

- The average word-count of each article in fact-based dataset is 394 words, while it is 896 words for an opinion-based dataset.
- It is seen that time-denoting words like days, years and words like ‘since’, ‘currently’ etc are very common in fact-based articles. Some other top words are : ‘research’, ‘data’, ‘according’, ‘announced’.
- Words like ‘often’, ‘means’, ‘important’, ‘great’, ‘possible’, ‘show’, ‘believe’, etc were quite common in opinion-based articles.

B. On PoS Tags

An analysis on the most common words appearing in

1) **a) Fact-based Articles:**

Note that the most common tags are NN, JJ, DT which broadly represents Nouns, Adjectives, Determiners.

b) Opinion-based Articles

Most common tags are NN, NNP, DT.

Note that in opinion based, NNP tag is much more than in fact-based.

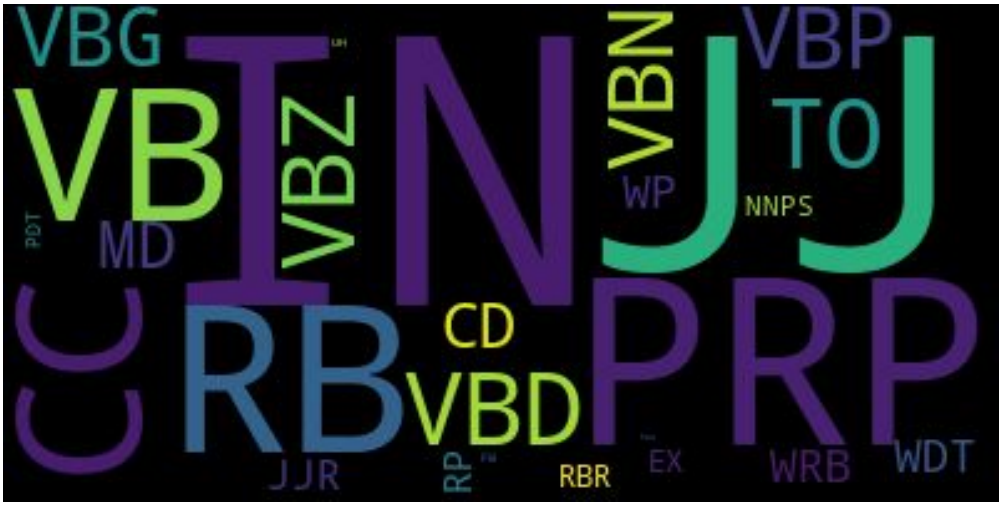
The high frequency of nouns and determinants are common and expected in any article. So we next remove these tags to analyze things more clearly.

2. Then the same analysis is done after **removing the tags = 'NN', 'NNS', 'NNP' and 'DT'.**

a) **Fact-based articles**

Note that the most common tags are IN, JJ, PRP, which represents prepositions, adjectives and personal pronouns.

b) Opinion-based Articles:



Most common tags are IN, JJ, PRP and RB, representing prepositions, adjectives, personal pronouns and adverbs respectively.

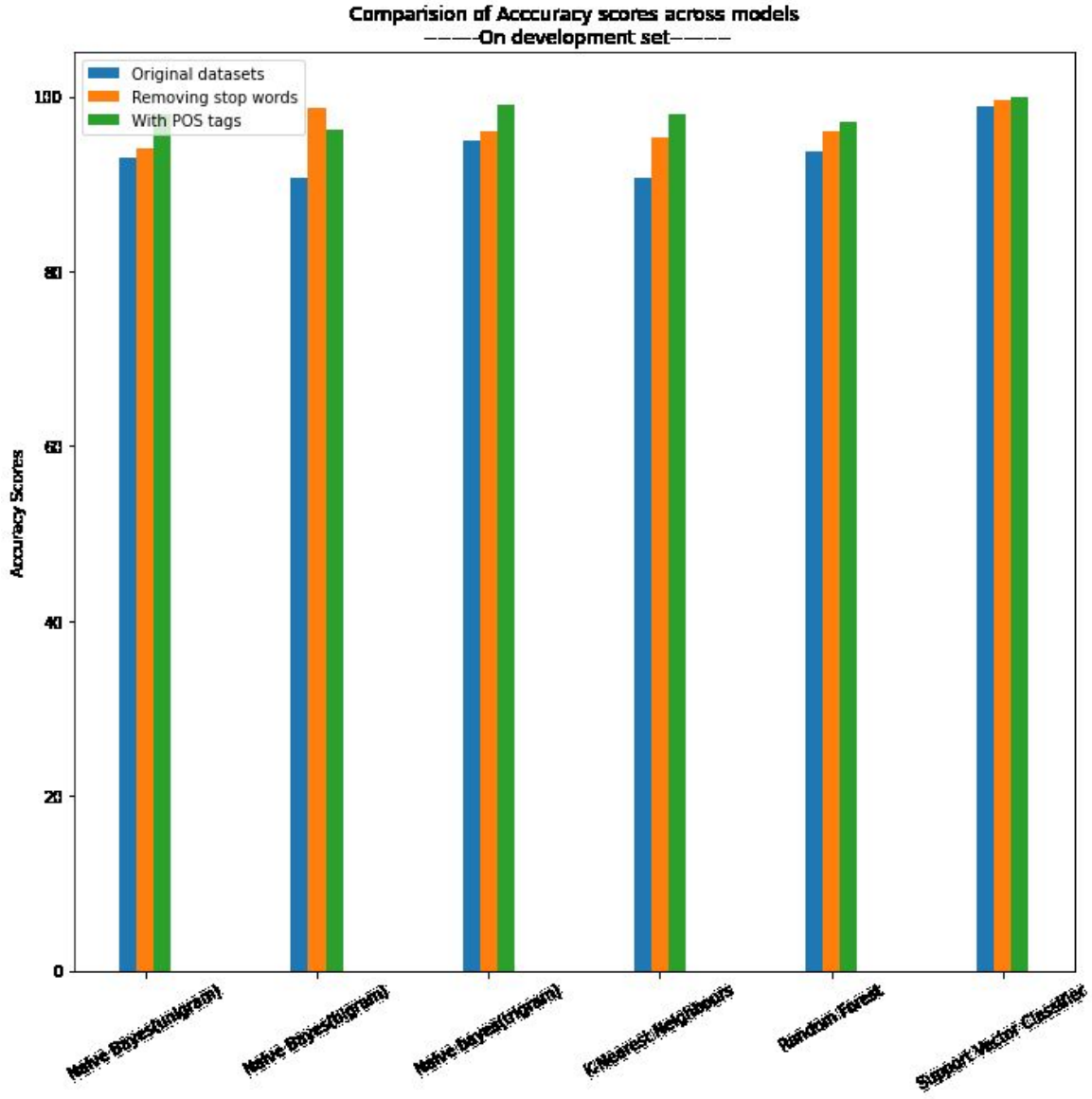
#Note that in opinion based, RB tag is much more than in fact-based. While fact-based, CD is more common than opinion-based. This makes sense, since use of adverbs (especially those describing other adjectives or adverbs) may be attributed to presence of opinions rather than plain facts, which tends to make more use of precise data (numbers, CD) than words like ‘very’, ‘little’, etc.

The above analysis tells us that there might be a set of words as well as PoS tags, whose frequency can be analyzed to categorize articles into different categories.

Model-Based Study:

A. On Development set

The development set was the subset of articles compiled from ‘BBC News’ and ‘The Guardian’ websites. The accuracy scores of all classification models are summarized in the snapshot below :



<u>Observations</u> :

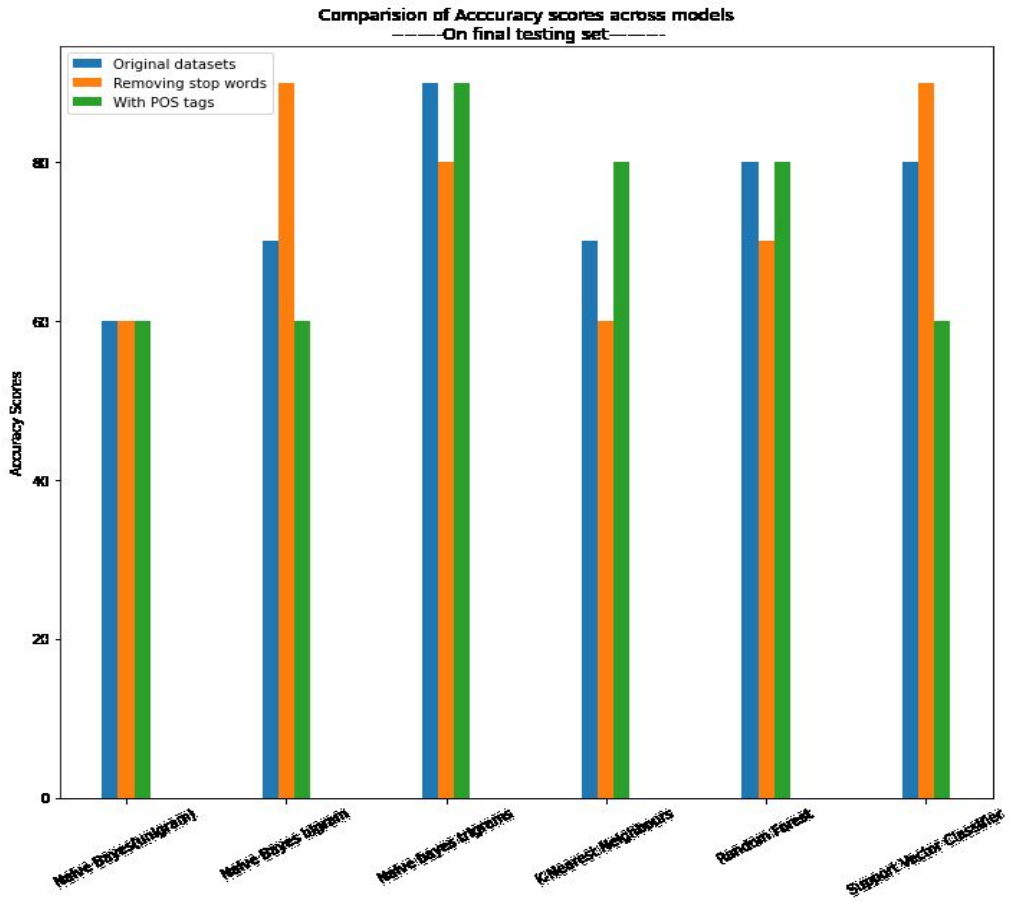
- The classification accuracies of the models seems to have *increased* after removing the stop words from the corpus. The stop words’ list has been obtained from the NLTK library.
- Further improvements in accuracy scores are obtained from attaching POS tags to the words and training the models. The only exception to this

trend is the Naive Bayes Classifier which took bigrams as the input features.

- Among the classifiers, Support Vector Machine classifier seems to have the highest accuracy scores with POS tagged words as input features having the highest accuracy score of **100%.**
- As for the most computationally efficient model with great accuracy scores, Naive Bayes classifier with trigrams as input features seems to take the lead.

B. On Final testing set

The final testing set consisted of articles compiled from the 'Hindustan Times'. The snapshot below summarizes all the accuracy scores :



Observations :

- The general trend of increasing accuracy scores after removing stop words and incorporating POS tags seen in development set seemed to have completely derailed in case of the final testing dataset.
- Naive Bayes classifier with unigrams as input features seems to be the worst model in terms of accuracy scores. There's no improvement at all even after all the text processing.
- The most promising classifier model seems to be the Naive Bayes classifier with trigrams as the input features 'cause they have high accuracy scores.
- More robust conclusions could have been made if the testing dataset had a fairly large number of constituting articles.

Limitations and Scope:

1. The analysis may be imprecise owing to the source of the articles, the size and nature of the dataset. This can be clearly seen in the list of top words of opinion-based articles.
2. The training set consisted of news articles pertaining to the UK and the US. This could have affected the machine learning models' accuracies while classifying Indian centric news articles. Expanding the final testing dataset could have enabled us to find robust conclusions regarding the generalizability of the models.
3. The possible solutions could be incorporating all kinds of regional news articles in the training set or finding out further linguistic differences between factual sentences and sentences expressing opinion viz. *sentence level classification of the articles.*
4. Models' accuracies and generalization capabilities could have been further increased if we could have removed stop words from the POS tagged news articles.