National University of Computer and Emerging Sciences

“Fake News Detection”

STUDENTS NAME and REGISTRATION NUMBER:

Maaz Asad	i18-0474
DEGREE PROGRAM:	BS(CS)
Section:	A
SUBJECT NAME:	NLP
DATE OF SUBMISSION:	May 30, 2021
SUBMITTED TO:	Doctor Omer Beg

Overview

In this Assignment I have implemented Fake New Detection using Naive Bayes to train the model and Laplace smoothing to test the model on the given data.

File Reading

We provided separate testing and training data. Both having real and fake news data. We were also provided with stop words. Data was loaded into the program into lists and cleaned with the help of the URDUHACK library to remove unnecessary white spaces and characters.

Removing Stop Words and Duplicates

It was evident that by removing stop words and duplicates from the training and test data, it increased the accuracy of the result. With the raw unfiltered data (containing duplicates and stop words) the program achieved an accuracy of 0.7366 and precision of 0.768. Whereas after removing duplicate words and stopwords the accuracy increased to 0.77, however the precision remained unchanged. We can get better results if we have a larger training data that will ensure that the model has calculated the probabilities more accurately which would enable the model to produce more accurate results

Naive Bayes Model

This function in my implementation takes Classes, Real News Data, and Fake News Data as parameters. Classes have 2 elements (Real, Fake) which is used to distinguish between the real news and the fake news data. Then a vocabulary is created using the data sets provided regardless if it is fake or real and count of words are stored. For each word in Vocabulary its conditional probability is calculated and returned. In my implementation conditional probability of real and fake are return separately

Laplace Smoothing

This Function accepts Classes, Conditional probability of real and fake news and a test string. It first extracts the words from the vocabulary that are present in the test string, and using the conditional probabilities it scores the test string with logarithmic values of the probabilities. In the end it returns whether the test string is considered to be a fake or as real news

Evaluation of the Model

The training and test data has been processed into 4 different forms

1. Raw. only removed unnecessary white spaces
2. Stop words removed
3. Duplicate words removed
4. Both Duplicate and Stop words Removed

Also we have taken into consideration the False positives and the False Negatives to evaluate the model. The 4th data set(with duplicates and stop words removed) had the highest accuracy