Machine Learning End Sem Project Report

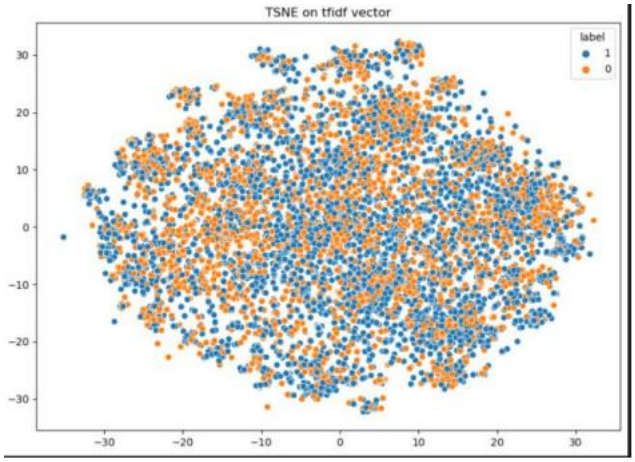
FAKE AND REAL NEWS DETECTION

Mayank Bajpai

In the recent years of information transfer we have seen how major shortcomings in the field of technology have affected the lives of the people. The times of social media has catalysed the process of propagating a lot of fake news from anti-social elements all across the world.

1. INTRODUCTION TO THE PROBLEM STATEMENT
   Though, technology has been the reason for the recent positive developments in the human history it also has had its fair share of disadvantages too. One can see that there was a time when we had to search books for gathering information or maybe read newspapers for reading news but now people have both information and news in their pockets in the form of mobile phones. With regards to news which comes from various sectors in the form of social media, digital news etc. people tend to rely on certain things which are not true. This results in the propagation of information which is wrong. This is happening extensively nowadays due to bias with which journalists are reporting incidents due to their involvement of a particular political organization. Just recently we saw how there were riots in India due to circulation of news where a person belonging from a certain community was accused killing someone from the other communities.
2. DATA PREPROCESSING [USE OF NLP] As our data is in the form of text, we need to convert it in the form of numerical data and vectorization. For doing so we will take the help of natural language processing. So first we need to refine the data for actually converting the text to numbers. We first remove all the punctuation marks, links and extra white spaces except the commas by normal methods. Next, we do our first NLP where we tokenize the data. For tokenization we use the library as it is where it's work is to split paragraphs and sentences into smaller units giving it an actual meaning.
3. DATA DESCRIPTION The dataset had 16 columns namely index, the ID of the statement([ID].json), label, statement, subject, speaker, speaker's job title,, the state info, party affiliation, the total credit history account, including current statement(comprises of 5 columns together which are barely true counts, false counts, half true counts, mostly true counts, pants on fire counts), context(venue /location of speech statement) and extracted justification. There are a total of 12788 rows. There are 10239 rows for testing,1266 rows for testing and 1283 rows for validation take the help of label encoder which converts label into numeric form thereby converting it into machine readable form. We concatenate the data available into four parts where no speaker and no party is there and we are using either tfidf and bow separately. In the next we take both the speaker and the part he belongs and apply both the NLP algorithms. Next Grid

search would take care of the hyper parameters being optimized.



4. METHODOLOGY The methodology used in determining whether the data is a real or not is a combination of data cleaning, data processing. NLP and the different algorithms that I am applying to get the best accuracy. In data cleaning as written above we saw that we dropped certain columns and rows, tinkered with the classification as well.