Real or Fake News

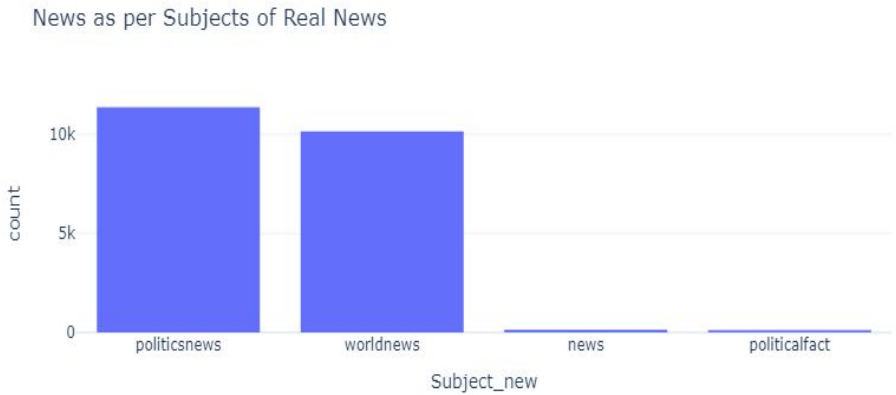
Problem Statement

To build a Predictive model to classify the news whether the news are fake or real and also produce some insights from the various news which words/phrases are important to classify the real and fake news.

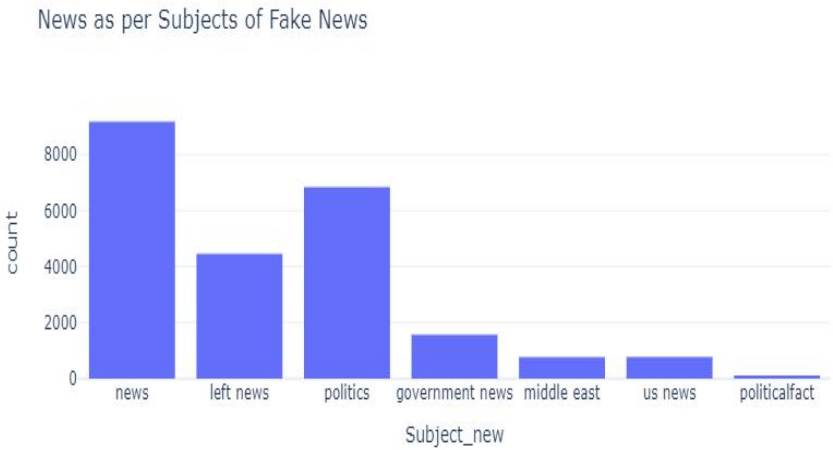
Description of the Dataset

Dataset consist of real and fake news both the dataset consist of three columns Title,News and Subject.

Distribution of Real News with respect to Subject

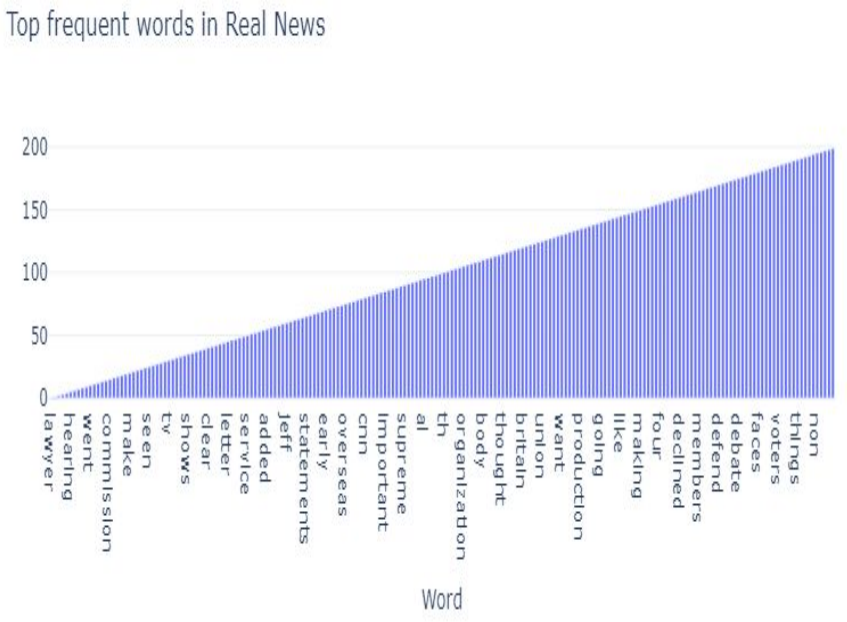


Distribution of Fake News with respect to Subject

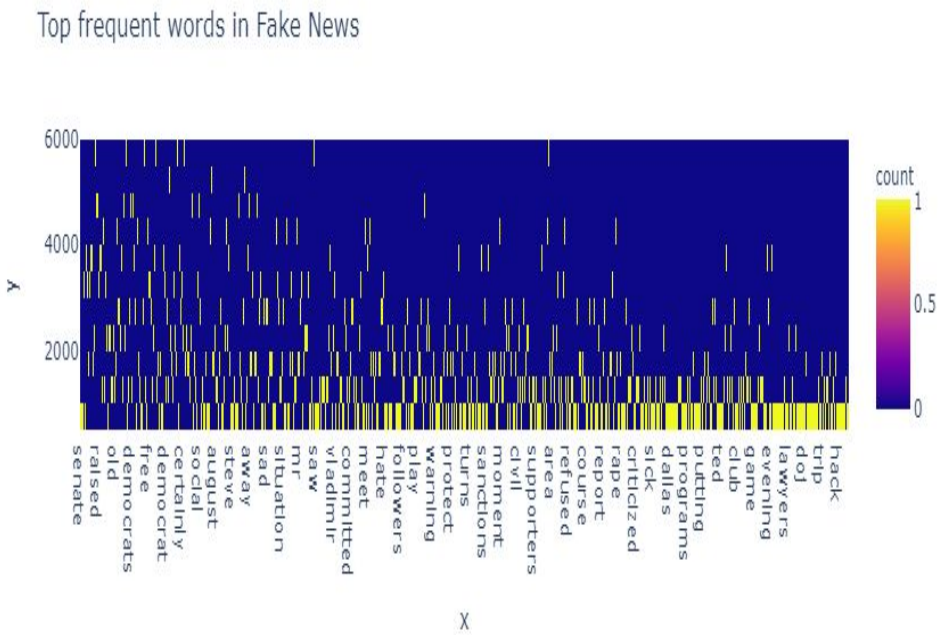


Top Frequent words in Real & Fake News Article

Top frequent words in Real News

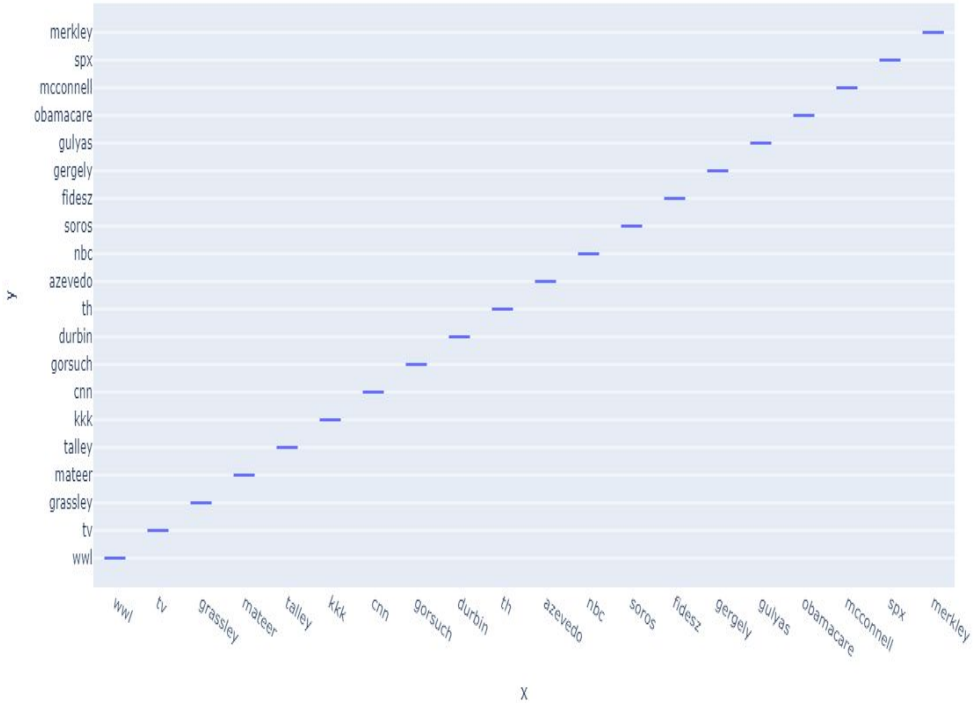


Top frequent words in Fake News

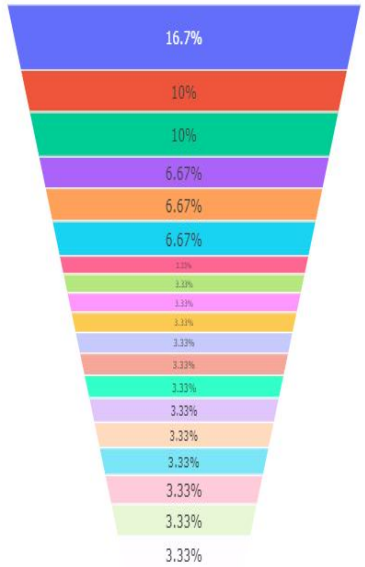


Graphs of words Before & After Spell Check

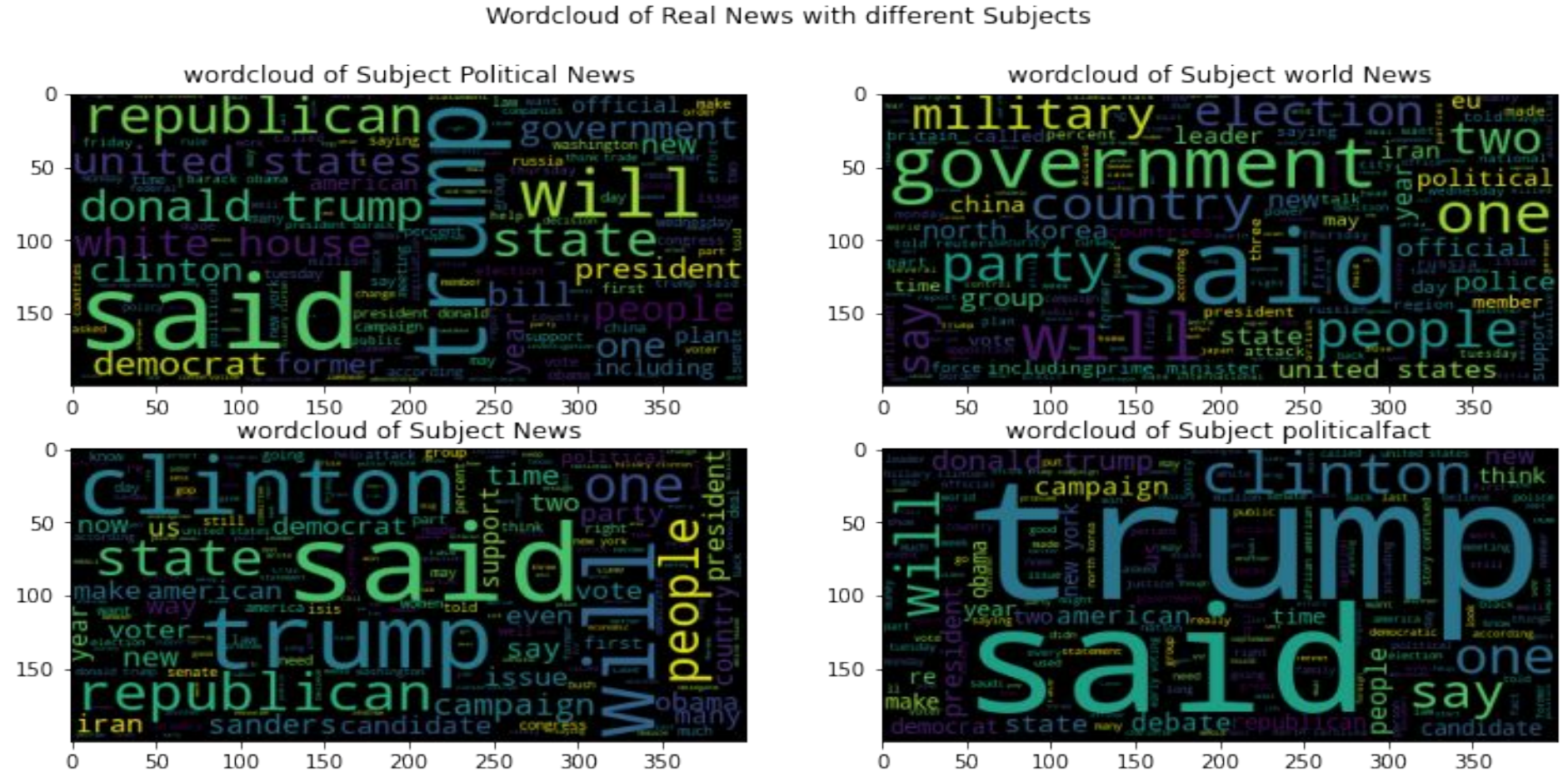
Before SpellCheck



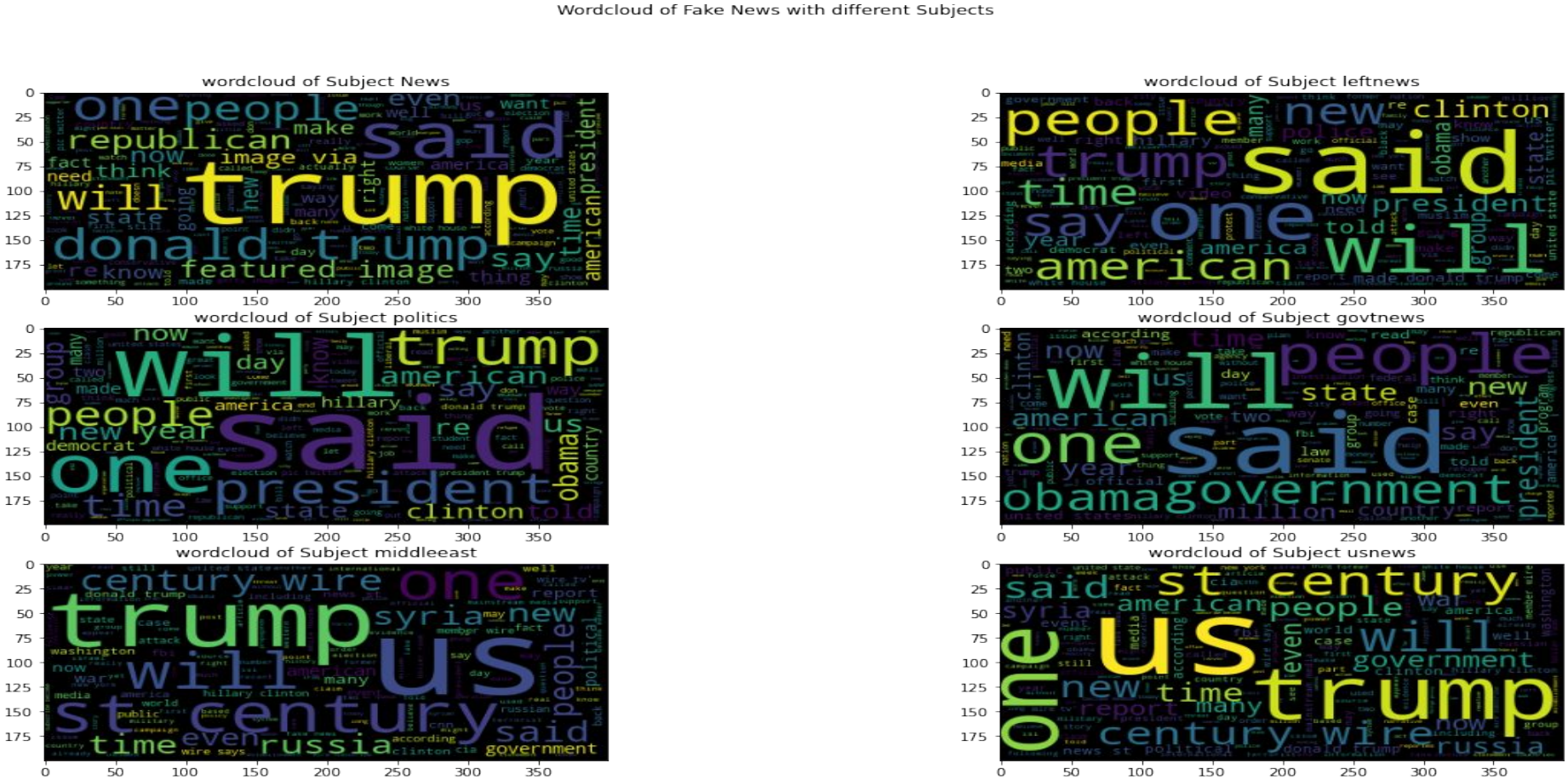
After SpellCheck



Wordcloud of Real News with Different Subjects



Wordcloud of Fake News with different Subjects



Preprocessing of Real & Fake News Text

Real & Fake news contains many non alphanumeric characters and also consists of wildcard characters, we have used rule based patterns to clean the text with Regex function and also removed stopwords as because these stopwords are meaningless and can outperform the model .Instead of Porterstemmer ,i have used Wordnet Lemmatizer Now the basic difference Lemmatizier is because Stemming just removes or stems the last few characters of a word, often leading to incorrect meanings and spelling. Lemmatization considers the context and converts the word to its meaningful base form, which is called Lemma.

Some Pharses of Real News

'foreign policy adviser',
'fight iran aligned',
'republican leader mitch',
'last month risks',
'behalf three people',
'former secretary state',
'will make federal',
'director general roberto',
'decades public service',
'summit next week',
'stand nuclear crisis',
'delegates take nomination',
'left democrats fuming',
'considering whether increase',
'human rights lawyers'

Some Pharses of Fake News

'evidence support argument',
'though still appears',
'2016 presidential race',
'spector vote us',
'spill administration becomes',
'matter fbi working',
'raw story facebook',
'department office indian',
'national security reid',
'jewish didn stop',
'fact ve got',
'news washington post',
'isn left trump',
'energy department office',
'boehner good company'

Top Bigram,Trigram & Qudragram words

	Bigram words
0	(united, states)
1	(donald, trump)
2	(white, house)
3	(president, donald)
4	(north, korea)
5	(prime, minister)
6	(new, york)
7	(said, statement)
8	(trump, said)
9	(islamic, state)

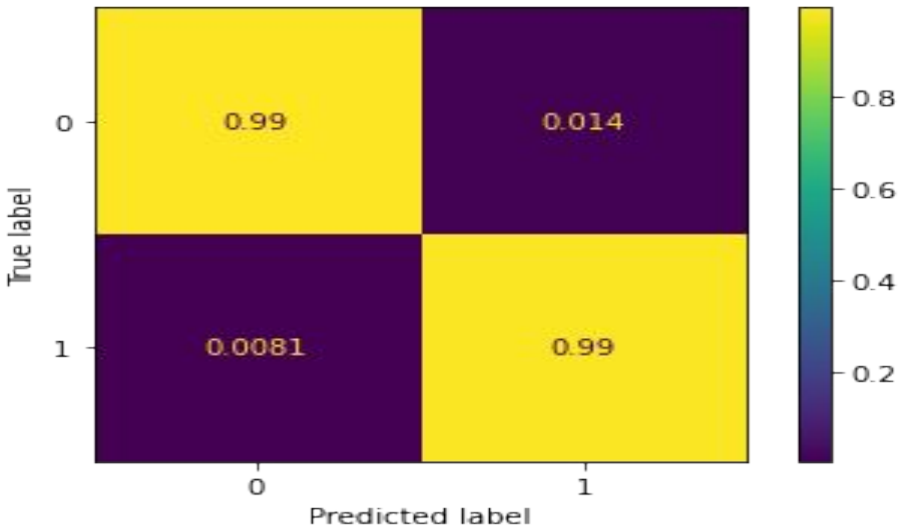
	Trigram words
0	(president, donald, trump)
1	(president, barack,obama)
2	(white, house, said)
3	(elect, donald, trump)
4	(president, elect, donald)
5	(president, vladimir, putin)
6	(state, rex, tillerson)
7	(secretary, state, rex)
8	(former, president, barack)
9	(speaker, paul, ryan)

	Quadrogram words
0	(president, elect, donald, trump)
1	(secretary, state, rex, tillerson)
2	(former, president, barack, obama)
3	(russian, president, vladimir, putin)
4	(prime, minister, theresa, may)
5	(democratic, president, barack, obama)
6	(majority, leader, mitch,mcconnell)
7	(senate, majority, leader,mitch)
8	(vice, president, mike, pence)
9	(president, bashar, al, assad)

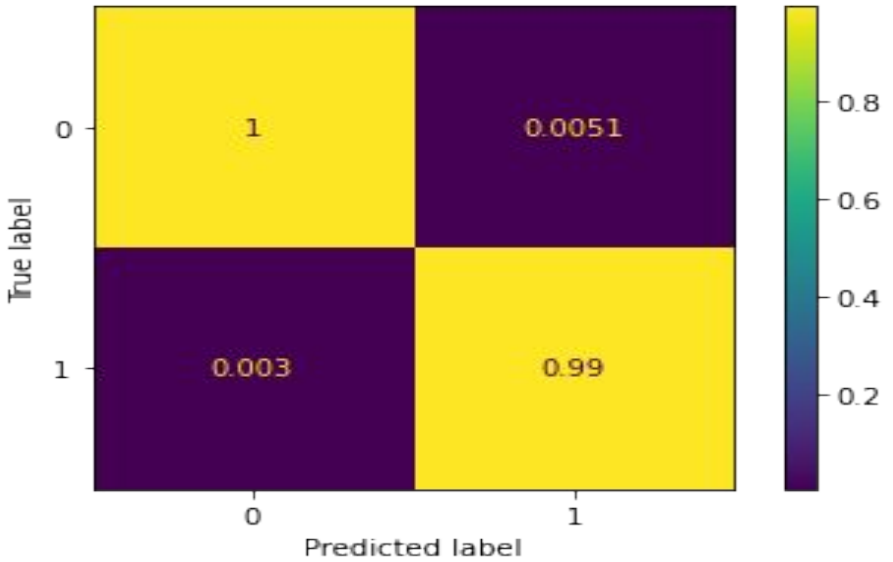
Model Buiding & Validation

We have made procession with Data Preprocessing and Data Cleaning, Remove all punctuations & stopwords ,replaced misspelt words with corrected words and used TFIDFVectorizer to vectorize the text and convert it for Machine Learning Algorithms. Algorithms used are XGBClassifier & LGBMClassifier.Best Model came up is LGBClassifier.

Confusion Matrix of XGBClassifier



Confusion Matrix of LGBClassifier



Conclusion

Approching these dataset, challenges i faced was cleaning the text adequately As also there was punctuations i have done count of punctuations of each article and feature engineered a column and replacing the mistaken words with correctly spelt words and also one column created on length of each article.while using vectorization of the article lots of hit & trial done to minimize the loss of the models with hyperparameter tuning. So, i concluded that for any nlp classification extracting new features based on No of Nouns ,No of verbs,No of each Punctuation,Length of the text, Spell check , and trying with different types of Vectorization is very important to achieve a good accuracy model with loss minimum.