# Fake News Detection

**Soumam Banerjee** MT20043
**Shivam Sharma** MT20121
**Jaswanth Naidu** MT20061

## Abstract

This report shows fake news detections using various ML approaches like Naive Bayes classifier, Logistic Regression, Decision Tree Classifier and SVM. The report also contains a roadmap to the planned baseline implementation based on few research papers and how we plan to contribute and improve it further. All the above mentioned models are trained and tested on 2 datasets which we have acquired from Kaggle. Received results that are discussed in a subsequent section suggests that fake news detection problem can be addressed with artificial intelligence methods. This is in coherence with what we have gathered from various research papers as well.

## 1 Project Introduction

Internet and social media made the access to the news information much easier and comfortable. Often Internet users can follow the events of their interest in online mode, and spread of the mobile devices makes this process even easier.
Mass media have a huge influence on the society, and as it often happens, there is someone who wants to take advantage of this fact. Sometimes to achieve some goals mass-media may manipulate the information in different ways. This leads to producing of the news articles that are not completely true or even completely false. There even exist lots of websites that produce fake news almost exclusively. They deliberately publish hoaxes, propaganda and disinformation purporting to be real news – often using social media to drive web traffic and amplify their effect. The main goal of fake news websites is to affect the public opinion on certain matters (mostly political). Fake news is now a global issue as well as a global challenge. (Mykhailo Granik, 2017)

## 2 Importance Of The Project

With the increasing rate in people using social media, they are exposed to new information everyday but using social media as a medium for news updates is a double-edged sword. The main advantage is social media provides free access and spread of information at an impressive rate. The main disadvantage is social media provides the ideal place for creation and spread of fake news which can become highly influential and has ability to spread at exponential rate. The problem is this misinformation is hard to correct and may cause problems. For example, on October 03, 2008 Apple's stock took a 10-point hit because of a false report surfaced on CNN's iReport that Steve Jobs had a heart attack. The fake news will leads to a lot of confusion and uncertainty because in a short time it will spread a lot. The fake news will cause a imbalance in news ecosystem, for instance *"The 2016 US Presidential Election"*, the most popular fake news was even more widespread on Facebook. This describes how people will pay more attention to misinformation rather than authenticated facts. The fake news can change the the way the people interpret and respond to real news. To mitigate the negative effects caused by fake news both to benefit the society and news ecosystem it's important to develop methods that will automatically detects fake news on social media.(T S, 2020)

## 3 Definitions

### 3.1 Count Vectors

Count Vectors: Count Vector represents a notation in the form of a matrix data set matrix notation in which corpus document is represented by each row, each column represents a corpus term, and each cell represents the frequency count of a particular term in a particular document.

### 3.2 Bag Of Words

The Bag of Words model is simplifying the representation of text(such as sentence or text dataset or document), which is represented in bag(multiset) of its words, disregarding the grammar and even word order but keeping multiplicity. In simple words Bag of words is a simplest form of representing text interms of numbers. This is often used in Text Classification because of its simplicity and less expensive to compute.

### 3.3 TF-IDF

TF-IDF is an acronym of Term Frequency-Inverse Document Frequency. TF-IDF is a numerical statistic that is intended to reflect how important a word/term is to a document in a corpus. TF-IDF is the product of the TF and IDF scores of the term. TF means the number of times the word appears in the document to the total number of words in the document. IDF means the the number of times a word occurs in a corpus of documents. In simple words TF-IDF is a score to highlight each word's relevance in the entire document. So using TF model makes sense of important words in a document and using IDF model makes sense of important words throughout all documents. (George, 2020)

### 3.4 Stemming

Stemming is NLP technique used to normalize the word by truncating the word to its stem word. Stemming may not give us a dictionary, grammatical word for a particular set of words it only gives the stem word.

### 3.5 Lemmatization

Lemmatization technique is also like stemming. However, what makes it different from stemming is that lemmatization finds the dictionary word or meaningful word instead of truncating to the original word. Because of this Lemmatization is slower than stemming.

### 3.6 Doc2vec

Doc2vec is a unsupervised algorithm which computes a feature vector for every document in the corpus. These vectors can be used to find similarity between sentence or paragraph or documents.

### 3.7 Word2vec

Word2vec is a combination of models used to represent distributed representations of words in a corpus. Word2vec works on the idea of Distributional semantics which means that we can understand the meaning of a word by understanding the context that a word keeps. In simple words, Word vector algorithms use the context of the words to learn numerical representations for words, so that words used in the same context have similar looking word vectors.

### 3.8 GloVe

Global Vectors for Word Representation or GloVe is an unsupervised learning algorithm for obtaining vector representations for words. GloVe is a count based model, which takes the advantage of global count statistics instead of only local information. GloVe uses co-occurrence matrix to derive semantic relationships between words. (Jeffrey Pennington, 2014)

## 4 Literature Survey

There has been many work done till date on fake news as this is an important issue of the decade and with each passing year the type and definition of Fake news are also changing and thus it is becoming more difficult day by day. There are various aspects of fake news detection ranging from using chat-bots for spread of misinformation to use of click baits for the rumor spreading. Nowadays internet is is full of click-baits and spams in each and every other websites and social media.

Some researchers have described Linguistic Cue Approaches with Machine Learning , Bag of words approach ,Rhetorical Structure and discourse analysis ,Network analysis approaches and SVM classifiers. These are models are text based only and have very little or negligible improvement on existing methods.

Some have classified every tweet/post as binary classification Problem. The Classification is purely on the basis of source of the post/tweet. The Authors used manually collected data sets using twitter API , DMOZ . The following algorithms where used on data sets

1. Naïve Bayes
2. Descision trees
3. SVM
4. Nueral Networks
5. Random Forest
6. XG Boost.
7. LSTM.
8. Bidirectional LSTM.

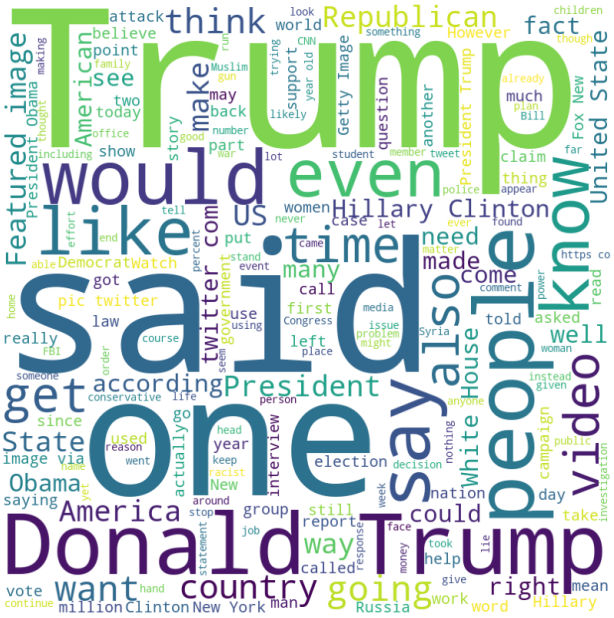
Now a days lots of deep learning based techniques are also used like RNN , lstm and CNN which have inturn provided a better accuracy than our traditional ML based techniques.

**Out-of-Vocabulary Words**

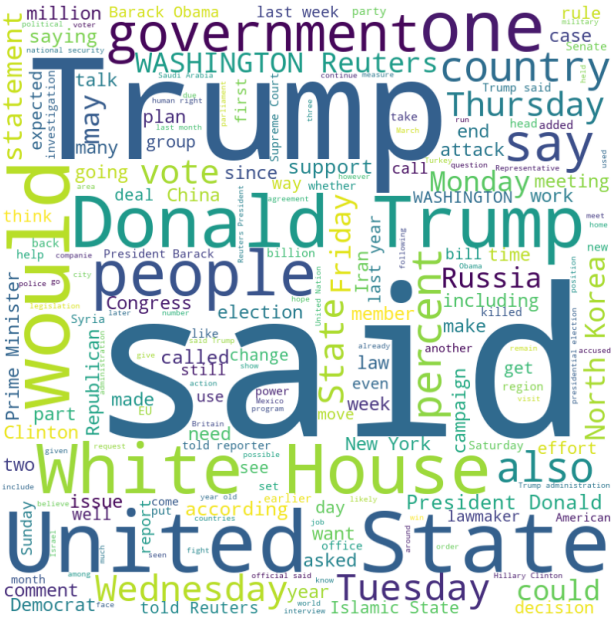
Word Embeddings encode the relationships between words through vector representations of the words. These word vectors are analogous to the meaning of the word. A limitation of word embeddings are that, they are learned by the Natural Language Model (word2vec, GloVe and the like) and therefore words must have been seen in the training data before, in order to have an embedding. (Kandi, 2018) (Manzoor and Nikita, 2019)

## 5 Data Analysis

### 5.1 Word Count - Fake




### 5.2 Word Count - Real




## 6 Dataset

We have worked on dataset taken from kaggle competition Kaggle dataset

## 7 Model Analysis - Midsem

| Model | Accuracy |
|---|---|
| Naive Bayes | 54.00 |
| Logistic Regression | 52.80 |
| Decision Tree | 51.0 |
| Linear SVM | 53.60 |
| RBF SVM | 55.20 |

Table 1: **Bag Of Words & Stemming**

| Model | Accuracy |
|---|---|
| Naive Bayes | 56.60 |
| Logistic Regression | 55.20 |
| Decision Tree | 50.80 |
| Linear SVM | 54.50 |
| RBF SVM | 55.70 |

Table 2: **TF-IDF & Lemmatization**

### 7.1 Naive Bayes

Naive Bayes are mostly used in natural language processing. Naive Bayes classifier algorithm is a family of algorithms which use Bayes Theorem. It uses the naive assumption that all the features are independent of each other. Bayes theorem calculates the probability P(c—x) where c is the class

of possible outcomes and x is the given instance which has to be classified.

### 7.2 Logistic Regression

Logistic Regression is one of the classification algorithm that is used to predict the probability of a categorical Target variable. In logistic regression, the Target variable is a binary variable that contains data coded as 1(yes, True) or 0 (no, fake). Logistic Regression is quite good in solving binary classifications due to its predictive power in probability values. Logistic Regression can also be used for solving multiclass classification.

### 7.3 Decision Tree

Decision Tree is one of the most widely used classifiers. It is a supervised learning algorithm. Decision tree can perform both classification and regression. Decision Tree classifiers are more popular because tree analysis is easy to understand. It divides the given data set into small parts and a decision tree is incrementally constructed. The leaf nodes of a decision tree represent the classification. Decision trees are comfortable with numeric and categorical data.

### 7.4 Support Vector Machine

Support Vector Machine or SVM is a linear model for classification and regression problems. SVM model takes the data in the training set, and maps it to data points in space so that there is a clear gap between points belonging to different categories. This gap is made as wide as possible to improve the performance of the model. Whenever a new data point is given to the model, it maps the point to the same space, and predict the category based on the side of the gap on which they fall.

## 8 Model Analysis - Endsem

| Model | Accuracy |
|---|---|
| LSTM | 98.59 |
| LSTM+Attention | 89.40 |
| BiLSTM | 98.21 |

Table 3: **Word2Vec**

| Model | Accuracy |
|---|---|
| LSTM | 95.98 |
| LSTM+Attention | 88.65 |
| BiLSTM | 96.31 |

Table 4: **Glove**

### 8.1 Long Short-Term Memory

Long short-term memory is an artificial recurrent neural network architecture used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. It can not only process single data points, but also entire sequences of data.(Wikipedia)

### 8.2 LSTM + Attention

Attention Mechanism is an attempt to implement the same action of selectively concentrating on a few relevant things, while ignoring others in deep neural networks.

### 8.3 Bidirectional RNN with LSTMs

As a final RNN model, a bidirectional recurrent network with LSTM cells was also implemented. The final states of the forward and backward layers are concatenated and passed through an affine transform before being input to the sigmoid to generate the prediction.

## 9 Conclusion

Even though there is a bit of success in detection of fake news using some of the Machine learning techniques yet since with every passing day characteristics and definition and types of fake news in social media networks is changing drastically causing a challenge in classification of fake news . But with the advent of deep learning techniques and applications in recent past, most of research works is implementing deep learning methods, like CNN,, Deep neural network and Deep auto encoder model,LSTM, RNN , in various applications, like audio and speech processing, Natural language processing and modelling, information retrieval, objective recognition and computer vision, hence post midsem we worked particularly on modern DL techniques like LSTM ,attention models, Bi-Lstm and we saw a huge improvement in accuracy than our old traditional ML techniques as we can see from the tables itself. (Henil Chopada, 2020) (Bajaj, 2017)

## 10 Appendix

Data and project code available here

## References


Samir Bajaj. 2017. "the pope has a new baby!" fake news detection using deep learning.

Joyce Annie George. 2020. Fake news detection using nlp techniques.

Rushikesh Desai Divya Ebenezer Nathaniel Henil Chopada, Maitri Patel. 2020. Fake information detection techniques.

Christopher D. Manning Jeffrey Pennington. 2014. Glove: Global vectors for word representation.

Shabeel M. Kandi. 2018. Language modelling for handling out-of-vocabulary words in natural language processing.

J. Singla Manzoor and Nikita. 2019. Fake news detection using machine learning approaches: A systematic review. *3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India*.

Volodymyr Mesyura Mykhailo Granik. 2017. Fake news detection using naive bayes classifier. *IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*.

SREEJA T S, Steni P S. 2020. Fake news detection on social media-a review. *Test Engineering and Management*.

Wikipedia. Long short-term memory.