Fake News Detection

Ms. Juhi Jain
Department of Computer Engineering
Delhi Technological University

Delhi, India - 110042

Sourabh (2K18CO355)
Vishruth Khare (2K18CO393)
Department of Computer Engineering

Delhi Technological University
vishruthkhare_2k18co393@dtu.ac.in
sourabh_2k18co355@dtu.ac.in

Abstract— **With the inception of big data, there is a lot of information that goes unfiltered contributing to Fake news. This false information can lead to catastrophic results in the longer run. Hence to curb this spread, we aim to provide constructive techniques that can help detect such kind of misleading news. With the help of this paper, we aim at exploring and implementing machine learning algorithms like Naive Bayes, Linear Support Vector Classifier, and Passive Aggressive Classifiers. The analysis is based on the dataset extracted from diverse sources and contains over 6,000 news samples. Further, methods like Count Vectorizer and Tf-Idf Transformer were incorporated for the pre-processing. The proposed work results in an accuracy ranging from 85.7% to 94.35% which is good enough to classify most of the information but still there is a scope of improvement. With advancements in models and other self-learning models can yield more precise results.**

Index Terms— **Naïve Bayes, Linear SVM, Passive Aggressive Classifier, Count Vectorizer, TF-IDF Vectorizer.**

I. INTRODUCTION

We all know that in the nearby world, everyone has been affected by the ubiquitous fake news which is spread by all and sundry, deliberately or unwittingly. Frequently it is originate that bogus information shared has malafide intentions of political gain or for spreading propaganda against particular casts, creeds, race, or religion.
The increased use of social media has led upto, in sequence being generously reachable to all and sundry. With the evolution of internet and spread of social media, it provides a massive quantity of in order however the trustworthiness of information depends ahead several factors. Mammoth sum of in rank is available daily via online along with media, excluding it cannot trouble-free to tell whether the information is a accurate or not. “Fake News” is the expression labeled for kind of untrue information. The fictitious information is a premeditated try with the goal in arrange to spoil or errand an association, entity or individual‟s status or it can be simply with the motive to gain fiscally or politically.

“The aim of project is to develop a model that is able to predict whether the users claim is fake or real.” This project “Fake News Detection System” uses machine learning algorithms and natural language processing techniques. Natural –language processing is an area of 2 computer science and artificial intelligence concerned with connections flanked by computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data.
It is a well known fact that social media plays momentous responsibility in fake news dissemination and so to maintain the trustworthiness and reliability of social media, timely detection, and extermination of fake news is necessary [1].
The topic of fake news and its detection came into the limelight during the US presidential elections of 2016 when the entire social media was flooded with fake news and further studies even suggested that the results might have been different if not for the inception of fake news [2].

II. LITERATURE REVIEW

In recent time fake news has become a hot topic of discussion. So I have looked upon few works on the same topic and concluded some of the major topics below.

2.1 Impact of Fake News

Fake news influences people’s decisions regarding whom to make your choice for during

elections. According to the researchers at the Oxford Internet Institute, in the run up to 2016 US Presidential election,
Fake newswas prevalent and spread rapidly with the help of social media bots" [1]. "A communal bot refers to an account on social media that is programmed to manufacture content and interact with humans or other wicked bots" [6]. "Studies reveal that these bots influenced the election online discussionslargely" [1]. "Fake news hinders serious media coverage and makes it more difficult for journalists to cover important news stories" [2]. "An analysis done by Buzzfeed revealed thatthe top 20 Fake news stories about the 2016 US Presidential election established more concentration on Facebook than the top 20 election stories from 19 major media outlets" [3].

2.2 Definitions and Types

Fake news is news that is intentionally and verifiably false and has the potential to mislead viewers/readers. There are two important dimensions of this definition: "intention" and "authenticity". First, fake news propagates misinformation that can be verified. Second, most of the fake news is formed with fraudulent intent to deceive community. This definition is widely adopted in recent research analysis [11; 12; 13; 14]. In general, Fake news can be categorized into three groups. In first group - "Actual Fake News", we can put those types of news, which are false and made up by the author of the article. The second group – "Fake news that is actually satire" is created purely to amuse rather than mislead its audience. Therefore, intentionally misleading and deceptive fake news is different from obvious satire or parody. The third group is "Poorly reported news that fits an agenda". This type of news has some real content but is not entirely correct and is designed especially for some political propaganda. Many researchers have streamlined the types of Fake news to simplify their research. For instance, According to definitions given by [1], there are a few types of news that cannot be called as "Fake"- (1) "Satire news having proper context." (2) "Misinformation that is created unintentionally." (3) Conspiracy theories those are difficult to put in true/false dichotomies. This paper [1] has presented two main aspects of fake news detection problem: "characterization" and "detection".

2.3 Different types of Detection Methods

1. **Knowledge Based Detection**: It aims to use external sources to fact-check the claims made in the news content. Two typical external sources are open web and knowledge graph. Open web sources are compared to the claims in terms of consistency and frequency [18, 19], whereas Knowledge graph is used to check whether the claims can be inferred from existing facts in graph or not [20, 21, 22].Many fact-checking websites (For eg.AltNews, Snopes, Smhoaxslayer, Boomlive) are using domain experts to determine manually the news veracity. Facebook has recently partnered with Indian fact-checking agency Boomlive to spot false news circulation on its website [23].
2. **Stance based detection**: This technique compares how a series of posts on social media or a group of reputable sources feels about the claim -Agree, Disagree, Neutral or is a unconnected. In [10], the writers old lexical as well as similarity features fed through a multi-layer preceptor (MLP) with one hidden layer to detect the stance of the articles. They hard-coded reputation score feature (Table 2.1) of various sources based on nationwide research studies. Their model achieved 82% accuracy for pure stance detection on their dataset. Another paper [13] used "wisdom of crowd" feature to improve news verification by discovering conflicting viewpoints in micro blogs with the help of topic model method - Latent Dirichlet Allocation (LDA). Their overall news veracity accuracy reached up to 84%.

3. **Other related works**: In implement Document comparison scrutiny, that calculate the Jaccard match, a far and wide used correspondence evaluate, between a hearsay „n‟ in test set with each news in Fake news instruction set „F‟ and real news instruction set „R‟. The outcome obtain were very shows potential. In [16] the authors have exploited the distribution patterns of in sequence to detect the hoax. Many delve into papers have used different linguistic and word embed features. The most common ones are tf-idf, word2vec, punctuations, ngrams, PCFG.

III. Methodology

This project was developed in Python using hard codedt libraries. Python has a colossal set of libraries in adding together to extensions, which can be easily used in Machine Learning. Sci-Kit Learn library is the optimum foundation for machine learning algorithms where practically all types of machine learning algorithms are with pleasure offered for Python, thus straightforward and rapid educated guess of ML algorithms is probable.

Dataset

The dataset used for the project was downloaded from https://www.kaggle.com/c/fake-news/data train.csv: A full training dataset with the following attributes:

- id: unique id for a news article
- title: the title of a news article
- text: the text of the article; could be incomplete
- label: a label that marks the article as potentially unreliable

Text Pre Processing

The data on social media is extremely amorphous – bulk of it is casual announcement with typing errors, slangs and poor-grammar. For better impending, it is essential to clean the data before it can be used for extrapolative modeling. So, fundamental pre-processing was completed on the training data. This step was comprised of

1. Conversion to Lower case: Initial step was to change the text into lower case, this is done to avoid multiple copies of the same words. For e.g. while ruling the word count, "Basic" and "basic" is full as unlike words.

2. Removal of Punctuations: Punctuations do not have a great deal consequence while treating the text data. So, removing them helps to trim down the size by and large text.

3. Stop-words removal: Stop-words are the largest part regularly taking place old words in a quantity. These are for e.g. a, the, of, on, at etc. They routinely classify the constitution of a text and not the context. When treated as feature, it would result in deprived presentation. So, Stop-words are uninvolved from the training data as the part of text onslaught process.

4. Tokenization: It refers to dividing the text into a sequence of words or group of words like bigram, trigram etc. Tokenization was done so that frequency-based vectors values could be obtained for these tokens.

5. Lemmatization: It converts the words into its word root. With the help of a vocabulary, it does morphological analysis to pick up the root word. In this work, Lemmatization was performed to improve the values of frequency-based vectors.

Feature generation

Count Vector

Count Vector can be understood as matrix memo of the dataset, in which rows correspond to the credentials in the quantity, columns symbolize a term from the amount, and cells stand for the count of that scrupulous term in a exacting file. The glossary is fashioned using the inventoryof only one of its kind tokens or words in the body.

TF-IDF vectors

TF-IDF weight depicts the comparative significance of an expression in the article and entire quantity.

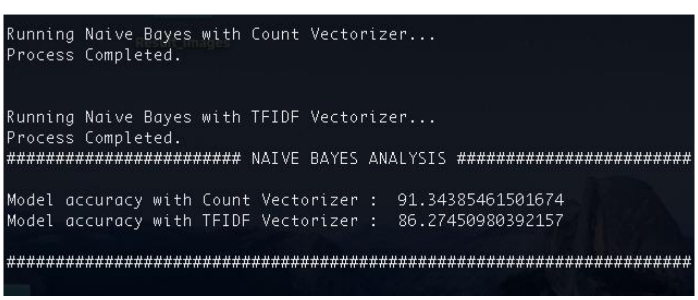
TF stands for Term Frequency: It studies how recurrently an expression appears in a file. As mostly all the file size varies, an expression may become visible more in an extensive sized file, than that in a short one. So the measurement lengthwise of the text often divides Term frequency.

$$TF(t,d) = \frac{Number\ of\ times\ t\ occurs\ in\ a\ document\ 'd'}{Total\ word\ count\ of\ document\ 'd'}$$

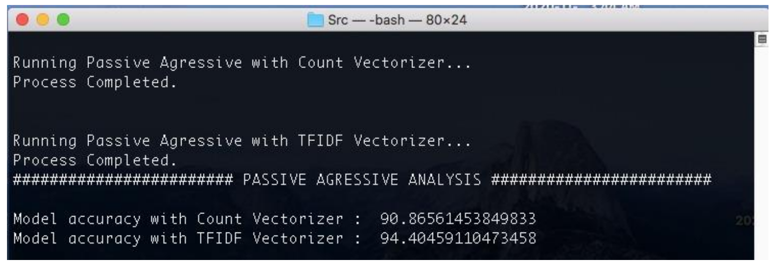
Algorithms used for classification

Now we proceed further by training the chosen classifiers. Training of classifiers is done using three specific algorithms Multinomial Naïve Bayes Passive Aggressive Classifier and Linear Support Vector machines.

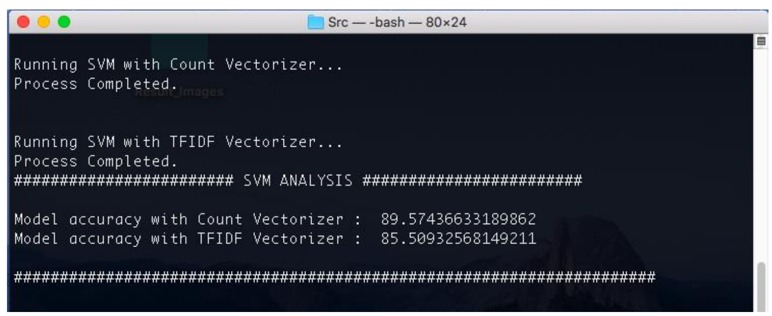
Naïve Bayes: This organization system is based on Bayes theorem, which presumes that the being there of a scrupulous attribute in a group of pupils is sovereign of the attendance of any additional trait. We can use it for calculating posterior probability.



Passive Aggressive Classifier: The Passive Aggressive Algorithm is an online algorithm; perfect to differentiate enormous streams of statistics. It is easier to implement and very fast. It works by taking an example, educating from it and then putting it away.



Support Vector Machine: In this algorithm, each fact item is plotted as a point in n-dimensional space. Values of each attribute are the value of all coordinate. It exclusively extracts an unsurpassed promising hyper-plane or a set of hyper-planes in a high dimensional space that segregate two class. Linear kernel is also sometimes used for SVM in this craft.



	Count Vectorizer	TF-IDF Vectorizer
Naïve Bayes	0.0727	0.0527
Linear SVC	0.5155	0.1264
Passive Aggressive	0.2913	0.1028

Table 1 : Classifier Time

Metrics used to access the Performance of Model

This section deals with the tools used to measure the performance of the model. Some of the metrics used in the project are described below.

Classification Accuracy: It is one of the mainstream familiar evaluations metric for organization problems. It can be defined as the number of accurate predication as against the number of whole predictions. However, this metric alone cannot give sufficient in sequence to come to a decision whether the model is a good quality one or not. It is apposite when in attendance are equal numbers of inspection in each class.

Confusion Matrix: It is also known as Error matrix, which are table illustrations that show the recital of the representation. It is special kind of Contingency table having two dimensions- “actual”, labeled on x-axis and “predicted” on the y-axis. The cells which are present on table are the number of predictions done with the help of the algorithm.

Predicted/Actual	1	0
1	True Positive	False Positive
0	False Negative	False Negative

Table 2 : Confussion Matrix format

Classification Report: Some platform provides an expediency description when operational on organization troubles which outputs precision, recall, F1 score and support for each class.

Precision: Precision is the ratio of appropriately predicted constructive instances to the entirety predicted constructive instances. High precision means low False Positive rate.

$$Precision = \frac{TP}{TP + FP}$$

IV. Result and Analysis

To analyze the outcome a few experiments were conducted Count Vectors and using Tf-Idf vectors at Word level and Ngram-level. Correctness was observed and is mentioned ahead for all models.

Count Vector and Tf-Idf vector were used to classify the responses as “Word level”i.e Only one word was preferred as indication for this experiment.

	Count – Vectorizer	TF-IDF Vectorizer	Remarks
Naïve Bayes	91.34%	86.27%	Simple and Efficient algorithm
Linear SVC	89.57%	85.51%	Worked well for Binary classification
Passive Aggressive	90.86%	94.40%	Highly accurate and but computationally extensive

Table 3 : Result Analysis and discussion (I)

Parameters comparable

- Precision – percent of positive predictions that were correct.
  - (TP)/(TP+FP)*100
- Sensitivity – percent of positive outcomes as observed by the model.
  - (TP)/(TP+FN)*100
- Specificity – Percent of negative outcomes as observed by the model.
  - (TN)/(TN+FP)*100
- Accuracy – Percent of the correct outcomes as observed by the model
  - (TP+FN)/(TP+TN+FP+FN)*100

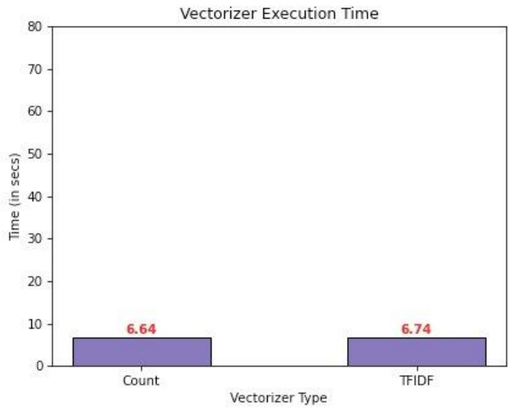
	Precision	**Sensitivity**	**Specificity**
Naïve Bayes	0.84	0.90	0.83
LinearSVM	0.89	0.88	0.90
Passive Agressive	0.86	0.92	0.86

Table 3 : Result Analysis (II)

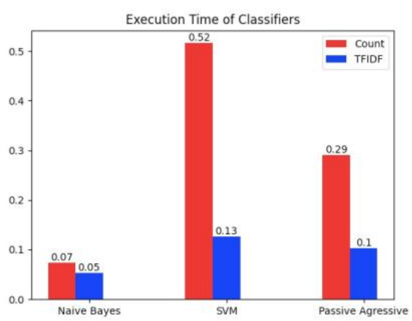
Comparison

This section deals with comparison of count vector and TF-idf vectors on the three algorithm naïve bayes, passive aggressive and linear SVM.

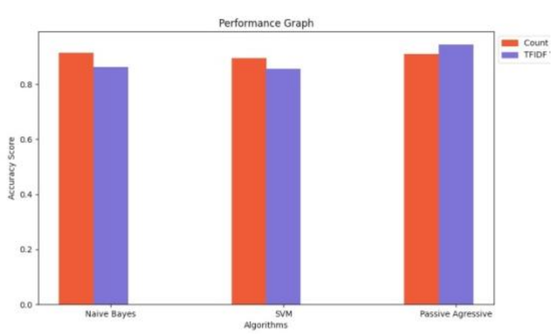
1. Vectorizer Execution Time



2. Execution Time of Classifiers



3. Performance Graph



V. VALID THREATS

Since there's a surplus of data out there, there's still a possibility of not being able to gather enough information on the concerned topic and hence the model might predict unacceptable results. Moreover since the accuracy isn't a 100%, there is still a need for manual intervention to actually check the credibility of the data. Going of high order models will cost extra time and space and may require high time of execution. Hence the time/value trade-off will always be there. Despite gathering a lot of information on a certain topic to actually predict the credibility of the news, there are chances that the news might be a Hoax at the very inception at the grass-root level and hence Manual Error is not accounted for in the model. The aforementioned points are the flaws that can be encountered in the process of detecting Fake News and more accountable Sources and Supplies will definitely yield more precise results in the longer run.

VI. CONCLUSION

Some hypotheses can be made on why same models works very well on one dataset and does not work well on the other one. The first thing we can think of is that the original hypothesis on different styles of writing between fake and reliable news is only verified in one dataset, the Fake News Corpus, and it is the most logical one, as these texts are coming from online newspapers (or pretending to be), and thus capitalize on advertisements for making money.

The second dataset, Liar-Liar Corpus is described by its authors as a collection a short sentence coming from various contexts such as political debate, interviews, TV ads and so on, thus it induces a lot of variety in writing style. For instance, it contains a transcription of vocal messages, which have in essence a different style from written one.

The data exploration chapter had already given an insight about this fact, as 2D data projection of the Liar-Liar Corpus shows no clear sign of separation, when Fake News Corpus shows one at the first look.

Future works

Studying fake news recognition only on supervised models on the data has revealed to be insufficient in most of the cases. To solve this setback, the majority of the research emphasizes on supplementary information such as about the author.

Most of the flourishing loom would be regular fact checking model, which is much more convincing of a model with some kind of data base, the purpose of the model would then be to haul out information from the data and confirm the information from the database. The only issue with this method would be that the awareness base would need to be persistently and physically renew to keep on up to date.

References

[1]. Nunes, M., &Correia, J. (2013). Improving trust using online credibility sources and social network quality in P2P marketplaces. 2013 8th Iberian Conference on Information Systems and Technologies (CISTI) , 1-4.

[2]. Ahmad, F. and Lokeshkumar, R. (2019). A Comparison of Machine Learning Algorithms in Fake News Detection. International Journal on Emerging Technologies, 10(4): 01–07.

[3].https://gadgets.ndtv.com/social-networking/news face book-twitter-step-up-fight-against-misinformation-on-u-selections-2278979

[4]https://en.wikipedia.org/wikiNatural_language_processing

[5]FAKE NEWS IDENTIFICATION CS 229: MACHINE LEARNING : GROUP 621 SohanMone, Devyani Choudhary, Ayush Singhania

[6]Emilio Ferrara, OnurVarol, Clayton Davis, FilippoMenczer, and Alessandro Flammini. The rise of social bots. Communications of the ACM, 59(7):96{104, 2016.

[7] Carlos Merlo (2017), "Millonarionegocio FAKE NEWS", Univision Noticias

[8] Chang, Juju; Lefferman, Jake; Pedersen, Claire; Martz, Geoff (November 29, 2016). "When Fake News Stories Make Real News Headlines". Nightline. ABC News.

[9] Read & Win: Machine Learning - 2019 - Wiley Analytical Science. https://analyticalscience.wiley.com/do/10.1002/gitlab.19010

[10]/kkkkkkkkkkkkkkkkkkkkkkkk.txt - Natural language processing https://www.coursehero.com/file/70457569/kkkkkkkkkkkkkkkkkkkkkkkktxt/