A Unique Approach for Fake News Detection Using Natural Language Processing

Yash Shukla, Nalini Yadav, Akshaya Hari
[1]*Department Of Computer Engineering*
Dr.D.Y.Patil Institute Of Engineering,Management & Research, Pune

***Abstract*—**Fake news, alternative facts are associated to each other since the time news was transmitted using news papers or radio. The fake news most recently came to light when the US Presidential election was underway. There have been several hoax stories where citizens, governments as well all other social elements are all affected by these stories. Facebook has been the amidst the controversy by the media houses for targeting the audiences and showing them posts to their support. This paper focus on detecting fake news with the help of various python libraries in association with Natural Language Processing. The system will be taking input from the user and then compare them with an existing data-set. We could spend multiple data-sets for increasing the accuracy of the algorithm. We will then show how the links and various sources of the news affects on it being true or false.

Keywords*—Natural Language Processing, Fake News Detection, Machine Learning, News sources.*

I. INTRODUCTION

In today's times there are various social media messaging and share applications that give users the power to share a piece of information with millions of people at the click of the button. The real problem is when people start to accept that rather than any of the news being "fake" theirs might have a new perspective on this. The problem begins where the masses begin to believe the fake news without checking its authenticity. There are very few tools or websites that tell the public about the news and its authenticity. Social media platforms like Facebook and Whats app are widely used in India for conversations and obtaining news. Many of the people believe that the news on these platforms have been published by a proper authenticated source whereas most of the news are hoax stories. The application which is proposed should be neutral in all ways that is there should be no political favoritism as the news is published by all of the parties for examples. We also have to identify the source of the article and then give a reason for not trusting the source. Some of the portals where fake articles are published are as follows:

Unreliable News Source	**Note**
70 News	A WordPress-hosted site that published a false news story, stating that Donald Trump had won the popular vote in the 2016 United States presidential election; the fake story rose to the top in searches for "final election results" on Google News.
Before It's News	Cited by US President Donald Trump at his 2016 campaign rallies. Before It's News and InfoWars were described as "unabashedly unhinged 'news' sites" in 2014 by *The Washington Post* following its promotion of conspiracy theories relating to Malaysia Airlines Flight 17.
bizstandardnews.com	Its stories have been mistaken as real-news then shared and cited as real-news. Its disclaimer says the stories "could be true" because "reality is so strange nowadays". But the disclaimer also says it is "a satirical site designed to parody the 24-hour news cycle."
Breaking-CNN.com	Responsible for publishing numerous death hoaxes, including one for former First Lady Barbara Bush one day after her announcement that she would halt all further medical treatment in 2018. Designed to emulate CNN.
CountyNewsroom.info	The fake news website, registered to Tbilisi, Georgia, makes "a minimal attempt to look official" and is used to spread malware on readers' computers.

Table 1: Sample Sources of fake news.[1]

There are no laws of sharing online content and there is no agency at present to which we can report fake news or flag the news, by doing an online web portal system will try to address this issue. In developed countries like the us the election are affected by the fake news but in developing countries where there are no means to cross reference and check the news that has been spread, due to this there have been killings in the year 2018 [2]. The paper has compared various models of machine learning and shown that how these models perform using random data sets.

II. LITERATURE SURVEY

Many of existing fake news detection methods massively relies on feature extraction. In [3], [4], [5], [6], [7], [8], authors have proposed approaches that are based on feature extraction. Here we will discuss the some of the algorithms that will be used in this project:

1. Bounded Decision Trees:

In computational complexity and communication complexity theories the decision tree model is the model of computation or communication in which an algorithm or communication process is considered to be basically a decision tree, i.e., a sequence of branching operations based on comparisons of some quantities, the comparisons being assigned the unit computational cost. The branching operations are called "tests" or "queries". In this setting the algorithm in question may be

viewed as a computation of a Boolean function where the input is a series of queries and the output is the final decision. Every query is dependent on previous queries. Several variants of decision tree models have been introduced, depending on the complexity of the operations allowed in the computation of a single comparison and the way of branching.[8]

2. Gradient Boosting:

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.[9]

3. Support Vector Machine

In machine learning, support vector machines (SVMs, also support vector networks[1]) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. [10]

III. PROPOSED MODELS

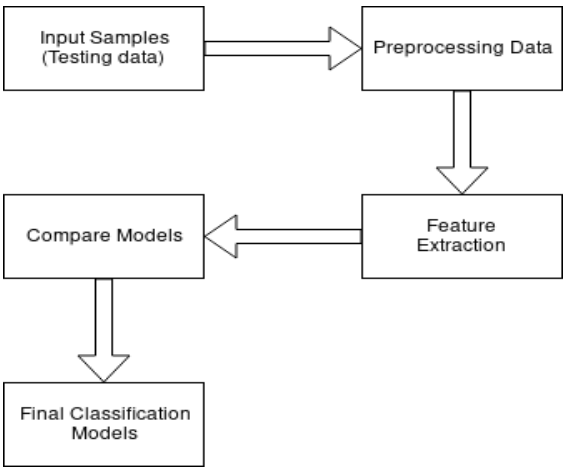

fig 1- Basic flow of model for fake news detection.

There is a large number of research done on machine learning algorithms. Here there is a proposition to compare various algorithms on the basis of the characteristics and calculate their precession recall and accuracy. By comparing these models with respect to the Probabilistic Context Free Grammar (PCFG) and Term frequency-inverse document frequency (TF-IDF) vs the various characteristics of machine learning algorithms. The above is the basic structure. Here we have taken the data-set and split it into training and testing data. Then the training samples are sent to the pre-processing unit. Then features are extracted. Henceforth the algorithms are compared on the basis of their statistical values. A final classification is chosen and then Probabilistic value or true and false values are calculated. This is done for every algorithm and then a final method is selected.
Here is an abstract of the methods:

Probabilistic Context Free Grammar(PCFG):
The process of recursive generation of strings from a grammar. The formal definition is as follows

Similar to a CFG(Context Free Grammar), a probabilistic context-free grammar G can be defined by a quintuple: $G = (M, T, R, S, P)$

where
M is the set of non-terminal symbols
T is the set of terminal symbols
R is the set of production rules
S is the start symbol
P is the set of probabilities on production rules[13]

TF-IDF (Term frequency-inverse document frequency)

Tf-idf stands for term frequency-inverse document frequency, and the tf-idf weight is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Variations of the tf-idf weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query. One of the simplest ranking functions is computed by summing the tf-idf for each query term; many more sophisticated ranking functions are variants of this simple model. Tf-idf can be successfully used for stop-words filtering in various subject fields including text summarization and classification.[14]

IV. COMPARISON OF MODELS

Model	Area Under Curve	Precision	Recall	Accuracy
Bounded Decision Trees	65.9%	66.9%	37.9%	67.6%
Gradient Boosting	75.6%	40.2%	16.1%	65.7%
Random Forests	80.0%	84.2%	18.4%	64.8%
Stochastic Gradient Descent	87.5%	74.1%	71.1%	65.7%
Support Vector Machine	84.3%	80.9%	44.5%	73.6%
Baseline	-	32.18%	32.18 %	67.89%

Table-2: Average model performance with both PCFG and TF-IDF bi-gram features at 0.7 score threshold for categorization. [5].

Model	Area Under Curve	Precision	Recall	Accuracy
Bounded Decision trees	60.7%	58.5%	23.3%	66.1%
Gradient Boosting	79.4%	41.0%	22.3%	68.7%
Random Forests	78.8%	82.9%	25.3%	67.6%
Stochastic Gradient Descent	88.3%	88.8%	45.3%	77.2%
Support Vector Machine	85.6%	81.3%	48.1%	76.2%
Baseline	-	32.18%	32.18%	67.89%

Table-3: Average model performance with only TF-IDF bi-gram features at 0.7 score threshold for categorization.

Model	Area Under Curve	Precision	Recall	Accuracy
Bounded Decision trees	50.0%	40.9%	10.8%	60.1%
Gradient Boosting	50.0%	40.9%	10.8%	60.1%
Random Forests	50.0%	40.9%	10.8%	60.1%
Stochastic Gradient Descent	50.0%	40.9%	10.8%	60.1%
Support Vector Machine	50.0%	40.9%	10.8%	60.1%
Baseline	-	32.18%	32.18%	67.89

Table-4: Average model performance with only PCFG features, classifying the top 5% of scores as positive (k = 0.05).

Here k cross fold validation is calculated to check whether the model is over fitting or not. The k is selected on the following terms:
For a small k, we have a higher selection bias but low variance in the performances.
For a large k, we have a small selection bias but high variance in the performances.

V. CONCLUSION

The paper has illustrated various rapprochements as to how the system will be implemented. But while implementing the system consideration of other algorithms is necessary. By combining different methods and comparing them with their base qualities better accuracy would be achieved. Though this approach might not give us the precise method , but the results of the prediction are promising. Implementation of this system will be done and then future algorithms to be compared will be added to the list. Also larger data sets might perform superior due to the combination of word feature extraction algorithms.

VI. REFERENCES

[1] https://en.wikipedia.org/wiki List_of_fake_news_websites#List
[2] https://tribune.com.pk/story/1757899/3-death-fake-news-social-media-fuelled-lynchings-shock-india/
[3] V. L. Rubin, N. J. Conroy, and Y. Chen, "Towards news verification: Deception detection methods for news discourse," in Hawaii International Conference on System Sciences, 2015.
[4] Y. Chen, N. J. Conroy, and V. L. Rubin, "Misleading online content: Recognizing clickbait as false news," in Proceedingsof the 2015 ACM on Workshop on Multimodal Deception Detection, pp. 15–19
[5] B. Markines, C. Cattuto, and F. Menczer, "Social spam detection," in Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web. ACM, 2009, pp. 41–48.
[6] N. J. Conroy, V. L. Rubin, and Y. Chen, "Automatic deception detection: Methods for finding fake news," Proceedings of the Association for Information Science and Technology, vol. 52, no. 1, pp. 1–4, 2015.[7] D. S. K. R. Vivek Singh, Rupanjal Dasgupta and I. Ghosh, "Automated fake news detection using linguistic analysis and machine learning," in International Conference on SocialComputing, Behavioral-Cultural Modeling, & Prediction and Behavior Representation in Modeling and Simulation (SBP-BRiMS), 2017, pp. 1–3.
[7] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," ACM SIGKDD Explorations Newsletter, vol. 19, no. 1, pp. 22–36,2017.
[8] https://en.wikipedia.org/wiki/Decision_tree_model
[9] https://en.wikipedia.org/wiki/Gradient_boosting
[10] https://en.wikipedia.org/wiki/Support_vector_machine
[11] Evaluating Machine Learning Algorithms for Fake News DetectionShlok GildaDepartment of Computer EngineeringPune Institute of Computer Technology, Pune, India.2017 IEEE 15th Student Conference on Research and Development (SCOReD)
[12] http://www.cs.columbia.edu/~mcollins/courses/nlp2011/notes/pcfgs.pdf

.

[13] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[14] H. Wu and R. Luk and K. Wong and K. Kwok. "Interpreting TF-IDF term weights as making relevance decisions". ACM Transactions on Information Systems, 26 (3). 2008