Fake News Detection Using Machine Learning

Savita[1], Vaishnavi N[2], Shruti Suman[3], Vaishnavi S[4]
[1,2,3,4]*Guru Nanak Dev Engineering Collage Bidar-585128*

Abstract **- In our modern era where the internet is ubiquitous, everyone relies on various online resources for news. Along with the increase in the use of social media platforms like Facebook, Twitter, etc. news spread rapidly among millions of users within a very short span of time. The spread of fake news has far-reaching consequences like the creation of biased opinions to swaying election outcomes for the benefit of certain candidates. Moreover, spammers use appealing news headlines to generate revenue using advertisements via click baits. This project aim to perform binary classification of various news articles available online with the help of concepts pertaining to Artificial Intelligence, Natural Language Processing and Machine Learning. We aim to provide the user with the ability to classify the news as fake or real and also check the authenticity of the website publishing the news.**

Due to easy access, rapid growth, and proliferation of the information available through regular news mediums or social media, it is becoming easy for people to look for news and consume it. These days a lot of information is being shared over social media and we are not able to differentiate between which information is Fake and which is legitimacy. For publishing a news in social media the cost is low, easy access. The extension spread of fake news has the potential for extremely negative impact on individuals and society. The goal of this project is to create an efficient machine learning algorithm for identifying the fake news.

INTRODUCTION

As an increasing amount of our lives is spent interacting online through social media platforms, more and more people tend to hunt out and consume news from social media instead of traditional news organizations. The explanations for this alteration in consumption behaviours are inherent within the nature of those social media platforms:

(i) it's often more timely and fewer expensive to consume news on social media compared with traditional journalism, like newspapers or television; and

(ii) it's easier to further share, discuss, and discuss the news with friends or other readers on social media. For instance, 62 percent of U.S. adults get news on social media in 2016, while in 2012; only 49 percent reported seeing news on social media.

It had been also found that social media now outperforms television because the major news source. Despite the benefits provided by social media, the standard of stories on social media is less than traditional news organizations. However, because it's inexpensive to supply news online and far faster and easier to propagate through social media, large volumes of faux news, i.e., those news articles with intentionally false information, are produced online for a spread of purposes, like financial and political gain. it had been estimated that over 1 million tweets are associated with fake news

Pizza gate" by the top of the presidential election. Given the prevalence of this new phenomenon, ―Fake news" was even named the word of the year by the Macquarie dictionary in 2016. The extensive spread of faux news can have a significant negative impact on individuals and society. First, fake news can shatter the authenticity equilibrium of the news ecosystem for instance; it's evident that the most popular fake news was even more outspread on Facebook than the most accepted genuine mainstream news during the U.S. 2016 presidential election. Second, fake news intentionally persuades consumers to simply accept biased or false beliefs.

LITERATURE SURVEY

Due to easy access, rapid growth, and proliferation of the information available through regular news mediums or social media, it is becoming easy for people to look for news and consume it. These days a lot of information is being shared over social media and we are not able to differentiate between which information is Fake and which is legitimacy.

For publishing a news in social media the cost is low, easy access. The extension spread of fake news has the potential for extremely negative impact on individuals and society. The goal of this project is to create an efficient machine learning algorithm for identifying the fake news.

Types of Algorithms:
There are several algorithms for detecting the fake news:

- Random Forest
- CNN
- KNN
- Logistic Regression
- Naive Bayes
- Hybrid models

Random Forest:
It is a combination of decision trees. Here each tree will build a random subset of a training dataset. In each decision tree model, a random subset of variables is used to partition the data set at each node. Bhavika, Bhutani, Neha, Rastogi, Priyanshu, Sehgal, Archana and Pulwar implemented a Sentiment Analysis technique for Fake news detection. They used LIAR, George McIntire, Merged Datasets and they classified those datasets by using Random Forest, Naive Bayes classifiers. For detecting the fake news they proposed a new solution by taking Sentiment as an important feature to improve the accuracy.

Convolutional neural networks:
In this by increasing the depth of the network the accuracy is increased when using the CNN method. In this by using k-nearest neighbour algorithm the accuracy is decreased and also precision, recall, f1-score values are reduced. They collected data from Kaggle: https://www.kaggle.com/c/fake-news for detecting the fake news by CNN with text only, CNN with text + title, CNN with text + author. They used layers present in convolution network for data pre-processing and feature extraction. The merged CNN reached accuracy of
96%.

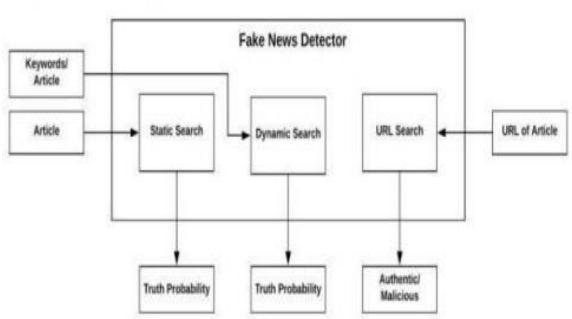
K-Nearest Neighbour:
Ankit Kesar Wani, Sudakar Singh Chauhan and Anil Ramachandran Nair developed a K-Nearest Neighbour Classifier technique for Fake News Detection on social media. In this they use Buzz Feed news. It contains the information about the Facebook news. In this the model has achieved maximum accuracy when the value of K taken between15 to 20. In this they gain the maximum accuracy of 79% tested against Facebook news dataset.

Logistic Regression:
It is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes and success) or 0 (no and failure). The logistic regression algorithm when implemented after extracting feature with term frequency and inverse document frequency gave the highest accuracy of 71% while testing the model

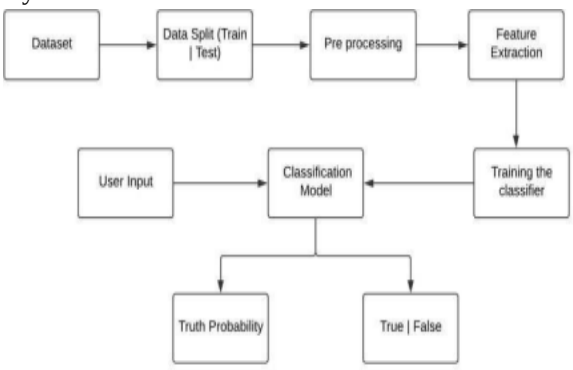
Naive Bayes:
It uses probabilistic approaches and based on Bayes theorem. They deal with probability distribution of variables in the dataset and predicting the response variable of value. An advantage of naïve Bayes classifier is that only requires less bulk of training data to access the parameters necessary for classification. Mykhailo Granick and Volodymyr Meshuga developed a Naive Bayes classifier technique for fake news detection. In this they use Buzz feed news which contains the information of Facebook content. In this the classification accuracy for true is 75.59% and for false is 71.73% and accuracy for total is 75.40%.Rahul M, Monica R, Mamatha N, Krishana R developed a machine learning model for fake news detection by using FND-Ju, Pontes Rout, News Files datasets.

Hybrid models:
Ms. Smita Vinit implemented a hybrid model for fake news detection. It is a combination of SVM and Naïve Bayes techniques [19]. They used manual dataset for training the model and Word Count, Gloat techniques are used for feature extraction. The hybrid model gave a good accuracy of 78%. Shlok Gilda implemented a Evaluating Machine Learning Algorithms for Fake News Detection [20].

System design:



Dynamic Search:



The second search field of the site asks for specific keywords to be searched on the net upon which it provides a suitable output for the percentage probability of that term actually being present in an article or a similar article with those keyword references in it.

SYSTEM ANALYSIS AND IMPLEMENTATION

Data collection and analysis:

We can get online news from different sources like social media websites, search engine, homepage of news agency websites or the fact-checking websites. On the Internet, there are a few publicly available datasets for Fake news classification like Buzzfeed News, LIAR, BS Detector etc. These datasets have been widely used in different research papers for determining the veracity of news. In the following sections, discussed about the sources of the dataset used in this work.

Online news can be collected from different sources, such as news agency homepages, search engines, and social media websites. However, manually determining the veracity of news is a challenging task, usually requiring annotators with domain expertise who performs careful analysis of claims and additional evidence, context, and reports from authoritative sources.

LIAR: A Benchmark Dataset for Fake News Detection William Yang Wang, ―Liar, Liar Pants on Firel: A New Benchmark Dataset for Fake News Detection, to appear in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017), short paper, Vancouver, BC, Canada, July 30-August 4, ACL.

Below are the columns used to create 3 datasets that have been in used in this project-

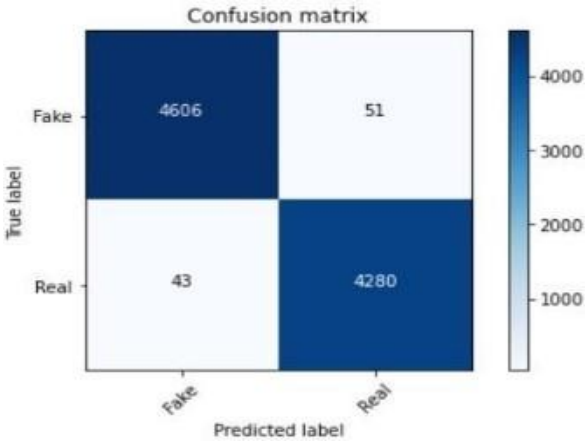
- Column1: Statement (News headline or text).
- Column2: Label (Label class contains: True, False)

The dataset used for this- project were in csv format named train.csv, test.csv and valid.csv.

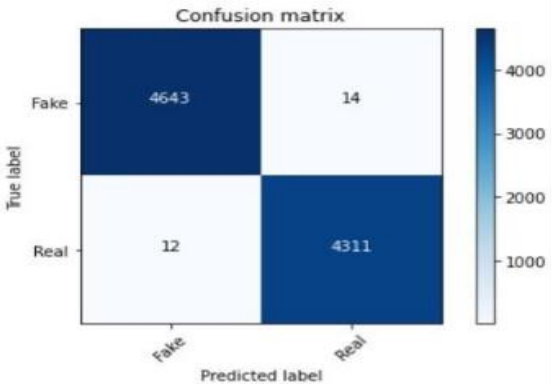
REAL_OR_FAKE.CSV: we used this dataset for passive aggressive classifier. It contains 3 columns viz 1- Text/keyword, 2-Statement, 3-Label (Fake/True).

OUTPUT

Logistic Regression The nature of target or dependent variable is dichotomous, which means there would be only two possible classes. Mathematically, a logistic regression model predicts P(Y=1) as a function of X as we said above we used counter vectorizer, tfidf transformer by creating confusion matrix we found an accuracy of or score=98.95%.



That is you explain what the input is and what the corresponding output is in the training data. where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. As we said above we used counter vectorizer, tfidf transformer by creating confusion matrix we found an accuracy of or score =99.71%.



CONCLUSION AND FUTURE SCOPE

In these days more people are continuously consuming news from social media rather than the traditional media. This fake news develops a strong negative impact on individual users and the society. Therefore, for detecting the fake news we analyse different research papers and identifies Word Embedding, Tokenization and Parts of speech tagging are best for Pre-Processing of data and also identifies TF-IDF and Count Vectorizer are best for feature extraction. So, for better approach Further we want to use those methods for Pre-processing, feature extraction and also, we want to implement the Random Forest classifier, Convolutional Neural Networks, Long Short-Term Memory for high accuracy and an Ensemble Learning Approach for high accuracy.

REFERANCES

[1] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu, ―Fake News Detection on social media: A Data Mining Perspective‖ arXiv:1708.01967v3 [cs.SI], 3 Sep 2017

[2] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu, ―Fake News Detection on social media: A Data Mining Perspective‖ arXiv:1708.01967v3 [cs.SI], 3 Sep 2017

[3] M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kiev, 2017, pp. 900-903.

[4] Fake news websites. (n.d.) Wikipedia. [Online]. Available:https://en.wikipedia.org/wiki/Fake_news_website. Accessed Feb. 6, 2017

[5] Cade Metz. (2016, Dec. 16). The bittersweet sweepstakes to build an AI that destroys fake news.

[6] Conroy, N., Rubin, V. and Chen, Y. (2015). ―Automatic deception detection: Methods for finding fake news‖ at Proceedings of the Association for Information Science and Technology, 52(1), pp.1-4.

[7] Markines, B., Cattuto, C., & Menczer, F. (2009, April). ―Social spam detection‖. In Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web (pp. 41-48)