International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 09 | Sep 2020 www.irjet.net p-ISSN: 2395-0072

FAKE NEWS DETECTOR

Akshata Gayakwad[1], Ruchi Rautela[2]

[1]PG Student, Vivekanand Education Society's Institute of Technology, Dept. of MCA, Mumbai, India
[2]Assistant Professor, Vivekanand Education Society's Institute of Technology, Dept. of MCA, Mumbai, India

---

Abstract - *Fake news on online social media is effectively spreading misinformation which breaks up the trust of people on traditional presses and journalisms. Detecting fake news is a difficult challenge to differentiate between real and fake news. Fake news is written to mislead readers to believe false information, which makes it difficult and significant to detect based on news content. All information should be included, such as using social engagements on social media, to help decide. Fake news produces data that is big, incomplete, unstructured, and noisy. Technical Skills of Fake news are Jupyter Notebook, Google Collab, Python, NLP, Nltk, TF-IDF, Cosine Similarity, Data-Cleaning, MachineLearning Classification Models, ROC Curves. For Fake News Detection we have used four algorithms like Logistic Regression, Random Forest, Naive Bayes, XGBoost, and found the best algorithm with the highest accuracy. We have created two datasets for comparing id's of two datasets whether the data is right, wrong, related, unrelated.*

Key Words*: *News Analysis, Information, detection, Logistic Regression, Random Forest, XG Boost, Datasets, social media, presses, Journalism, Algorithm, NLP, Machine Learning, Data Mining, stance Detection.

1. INTRODUCTION

Towards effectively detecting, characterizing, and modeling Internet fake news on online social media, this paper proposes a new framework which systematically characterizes the Web sites and reputations of the publishers of the fake and real news articles, analyzes the similarity and dissimilarity of the fake and real news on the most important terms of the news articles via term frequency-inverse document frequency (tf-IDF) & Latent Dirichlet Allocation(LDA) topic modeling. Fake news and distrust in the media are growing problems with huge complications in our society. The term 'fake news' became common for the issue, particularly to describe factually incorrect and misleading articles published mostly to make money through page views. In this paper, it is sought to produce a model that can accurately predict the likelihood that a given article is fake news. A given algorithm must be unbiased or impartial. Since fake news exists on both ends of the spectrum and also gives equal balance to the right news sources on either end of the spectrum. Also, the question of authenticity is a difficult one. To solve this problem, it is necessary to have an understanding of what Fake News is. Later, it is needed to look into how the techniques in the fields of machine learning, natural language processing(NLP) help us to detect fake news.

Stance identification is one aspect that affects the identification of false news[3]. To fact check any piece of knowledge generated by the media is not realistic for humans. The purpose of this project is therefore to use natural language processing (NLP) techniques to automate the detection of a position to determine the quality of the news source[2].

1.1 Objectives

The goal of this project is to use natural language processing techniques to automate stance detection since it is not practical for humans to fact check every piece of information produced by the media. Stance detection is a method used to determine the quality of a news article by taking into consideration what other organizations write about the same headline. A body of the text is claimed to agree, disagree, discuss, or be unrelated to a headline. Stance detection is the method that will be used to determine the quality of a news source[3]. The main objective is to detect the fake news with a straightforward proposition, which is a classic text classification problem. A model that can distinguish between "Real" news and "False" news is required.

2. FAKE NEWS DETECTION TYPES

- Knowledge-Based
- Style Based
- Stance Based
- Propagation Based

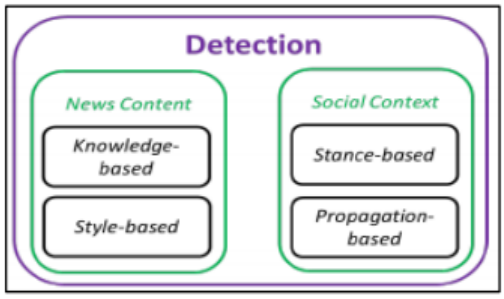


Fig-1: Types of Detection

3. NATURAL LANGUAGE PROCESSING (NLP)

NLP is a branch of artificial intelligence that helps computers understand, interpret, and manipulate human language. The NLP draws from many backgrounds to fill the void between human communication and computer comprehension, including computer science and computational language. Although natural language processing is not a new science, technology is evolving rapidly due to increasing interest in human-to-machine communications, as well as the availability of big data, powerful computing; and enhanced algorithms[2].As a human, you may speak and write in English, Spanish, or Chinese. But a computer's native language known as machine code or machine language is largely understandable to most people. Communication occurs not with words but through millions of zeros and ones that produce logical actions.

3.1 Importance of NLP

Large volumes of textual data Natural language processing helps computers communicate with humans in their language and scales other language-related tasks. For instance, NLP allows computers to read a text, hear speech, interpret it, calculate feelings, and decide which sections are significant. Today's machines can analyze more language-based data than humans, without tiredness and in a consistent, unbiased way.

Natural language processing includes many different techniques for interpreting human language, ranging from statistical and machine learning methods to rules-based and algorithmic approaches. Basic NLP tasks include tokenization and analyzing, part-of-speech tagging, language detection, and identification of logic.

3.2 NLP Techniques

- **Content Categorization :**

A description of a linguistic text, including search and indexing, content warnings, and identification of duplicates.

- **Topic Discovery and Modeling :**

Accurately capture the meaning and themes in text collections, and apply advanced analytics to text, like optimization and predict.

- **Contextual Extraction :**

Automatically convert structured information from text-based sources.

- **Sentiment Analysis :**

Identifying the mood or subjective opinions within large amounts of text, including average sentiment and opinion mining.

- **Speech-to-text and text-to-speech Conversion :**

It Transforms voice commands into written text and vice versa.

- **Document Summarization :**

It Automatically generates a summary of large bodies of text.

4. CHARACTERISTICS

The contributions of this paper can be summarized as follows:

We systematically characterize the Web sites and reputations of the publishers of the fake and real news articles on their registration patterns, Web site ages, domain rankings, domain popularity, and the probabilities of news disappearance from the Internet. We analyze the similarity and dissimilarity of the fake and real news on the most important terms of the news articles via the-IDF and LDA topic modeling[1].

We explore document similarity between fake, real, or hybrid news articles via the Jaccard similarity coefficient to distinguish, classify, and predict fake and real news.

The output of this paper is organized as follows.

Section 1: It describes the background of the fake news problem over online social media and also describes datasets.
Section 2: It characterizes the Websites and reputations of the publishers of the fake and real news articles,
Section 3: It focuses on analyzing the similarity and dissimilarity of the fake and real news on the most important terms of the news articles.
Section 4: We show the promising direction of leveraging document similarity to distinguish fake and real news by measuring their document similarity.
Section 5: Summarizes related work in detecting and analyzing fake news and highlights the difference between this effort with existing studies.
Section 6: It concludes this paper and outlines our future work.

5. ANALYTICS MODEL

This section describes the conceptual and mathematical ideas of the proposed analytical model developed for establishing the reliability of news articles. We start this section by describing the composition of complete and incomplete sentences. Next, we define events and topics extracted from complete sentences. Boolean-value

functions that distinguish fake events and topics. Finally, we describe the mathematical formulation used for quantifying the reliability of news articles[5].

6. DATASETS

A dataset that consists of a headline and a body of the text. Maybe this body of the text is from another post. Allowing the body of text from different articles allows this system to take into account what the other organizations are saying about the same headline. The performance of the method will be the title-related body of text location.

There are following stance types which system supports:

- Agrees
- Disagrees
- Discuss
- Unrelated

With this system, for a set of news headlines, statistics can be gathered concerning the stances. With these statistics, a user can come to their conclusion of whether a new organization has reputable news sources. This data will provide the location along with the headline and body to allow the system to learn which word combinations lead to which location. This framework will train on the data generated by the fake news challenge to achieve these stances. For testing, data will be provided without the stances. The body of text may be from a different article. The device output will be the location of the text body relating to the title.

Two CSVs that we have used to implement the system are:

Train_bodies.cs: ContainsBodyIDandBodyText-noofrecords:25
Train_stances.csv: ContainsHeadline,BodyIDandStance-noofrecords:49972

7. DATA COLLECTION

There must be two parts to the data-acquisition process: “fake news” and “real news”.

It was easy to compile the fake news as Kaggle released a fake news dataset containing 13,000 articles written during the 2016 election cycle. Now the later part has difficulty in it. That is to get the real news for the fake news dataset. It requires a huge workaround for many Sites because it was the only way to do web scraping thousands of articles from many websites. A real news dataset was created with the aid of web scraping a total of 5279 posts, mostly from media organizations (New York Times, WSJ, Bloomberg, NPR, and The Guardian) which were published around 2015–2016.

REQUIREMENTS:

- Python
- NumPy
- pandas
- Matplotlib
- Sklearn

8. CLASSIFICATION MODELS

Most classification models have been simple Naive Bayes approaches, which have worked great well. .We implement the following models:

- LogisticsRegression(LF)
- RandomForests(RF)
- Multinomial NaiveBayes
- XGBoost

The code relies upon the implementation of these various methods in SciKit-Learn. After generating the news feature vector, now we classify the vector to whether it is fake or real.

8.1 Logistic Regression :

In order to analyse results, logistic regression is used to explain the relationship between a binary dependent variable and one or more nominal , ordinal, interval, or ratio-level variables.

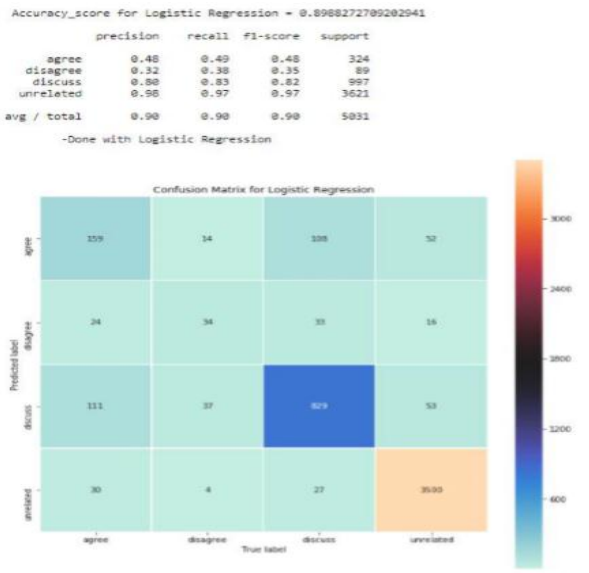


Chart1: Logistic Regression-Accuracy-90%

8.2 Random Forests :

Random Forest is a bagging type of ensemble model in which the base model for bagging is a decision tree. In addition to bagging (i.e taking random samples of total data and train those samples independently and then take the majority voting of numerous samples of the total dataset to find the output of the total dataset), there is feature bagging (i.e column sampling in which not all the columns/features are taken into consideration while training but rather random samples of features are considered while training different samples.

Algorithm :

1. import csv,random, numpy, os, nltk
2. from sklearn.ensemble import Random Forest Classifier
3. print("\t-RandomForestClassifier")
4. rf=Random Forest Classifier(n_estimators=num1,random_state=num2)
5.y_pred_rf=rf.fit(training_features,targets_tr).predict(test_features) #Evaluation
6. print("Evaluatingmodel..")
7.report_score(targets_test,y_pred_rf,'RandomForestClassifier')
8. print("\t-Done with Random Forest Classifier")

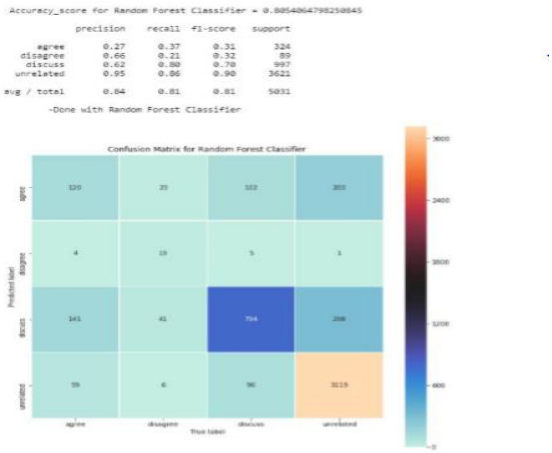


Chart 2: Random Forest got 80% accuracy.

8.3 Multinomial Naïve Bayes :

The multinomial Naive Bayes classifier is suitable for classification with discrete features e.g., word counts for text classification. The multinomial distribution normally requires integer feature counts. Fractional counts such as tf-idf may also work.

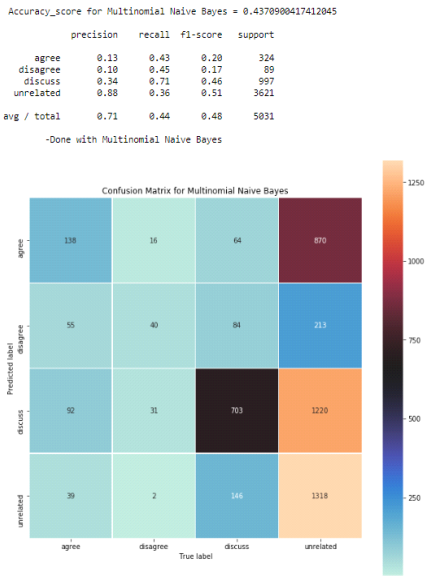


Chart 3: MultinomailNaïve Bayes-Accuracy-43%

8.4 XGBoost :

XGBoost is a Machine Learning algorithm based on a decision-tree ensemble that uses a gradient boosting system. It uses unstructured data for prediction problems (images, text, etc.).

Various applications can be used to solve regression, classification, ranking, and user-defined prediction problems.

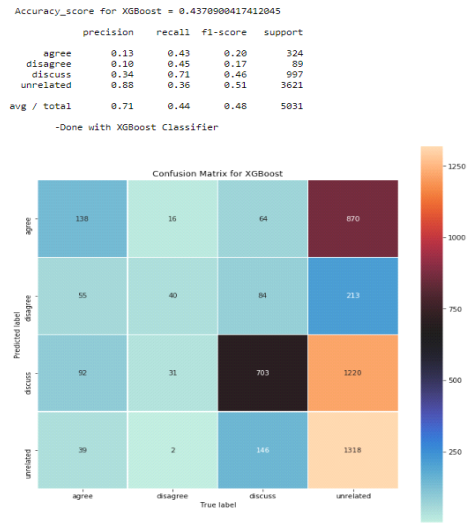


Chart 4: XG Boost-Accuracy-43%

In this way, we have done four algorithms and got the output with various accuracy

9. CHALLENGES TO DETECT WEBSITES

There are different ways to tell if a website is good or reliable. It is challenging to determine whether a website we are using is good, but here are a few things to look for[3]:

Author – Information on the internet with a listed author is one indication of a believable site. The fact that the author is willing to stand behind the information presented (and in some cases, include his or her contact information) is a good indication that the information is believable.

Sources –There are acceptable websites like books and scholarly articles, which should state the source of the information presented.

Date – The website enables readers to determine, by adding a date, if the information is sufficiently recent for their intent. It is important to know the date of any research information including the information found on the Internet.

Domain – Some domains such as .com, .org, and .net can be purchased and used by any individual. However,
a) The domain .edu is reserved for colleges and universities,
b) .gov denotes a government website.

These two are typically reliable sources of knowledge (although sometimes a university will allocate a.edu address for personal use to one of its students, in which case use caution when citing). Be careful with the domain .org, because .org is usually used by non-profit organizations which may have an agenda of believing something rather than education.

Writing Style – Poor spelling and grammar are an indication that the site may not be trustworthy. To make the information presented easy to understand, trustworthy sites watch writing style closely.

Site Design – A well-designed site can be an indication of more reliable information. Good design helps make information more easily to give access.

10. LITERATURE SURVEY

In this literature survey, we present a study/review of detecting fake news on social media. It includes fake news characterizations on psychology and social theory, current data mining tools, measurement measures, and representative datasets.We also discuss related research areas, open problems, and future research directions for fake news detectors. This paper comes up with the applications of NLP (Natural Language Processing) techniques for detecting the 'fake news' or 'false news',i.e. misleading news stories that come from the non-reputable sources. Only by building a model based on a count vectorizer using word tallies) or a (Term Frequency Inverse Document Frequency) tf-idf matrix, (word tallies relative to how often they are used in other articles in your dataset)[4].But these models don't consider important qualities like word ordering and context. It is very possible that two articles that are similar in their word count will be completely different in their meaning. To fight against the fake news is a classic text classification project with a straightforward proposition. Is it possible for you to build a model that can differentiate between “Real “news and “Fake” news? So a proposed work on assembling a dataset of both fake and real news and a Naive Bayes classifier in order to create a model to classify an article into fake or real based on its words and phrases.

11. DETECTION OF FAKE NEWS

Fake news is created by fabricating non-existent news or modifying legitimate news. The credibility of fake news is boosted by

(1) Imitating well-known authors’ writing styles or
(2) Expressing opinions with a tone frequently used in real news.

Very recently, an increasing number of fake news detection methods have been developed. All existing detection schemes may be grouped into two distinct classes, namely,

a) Linguistic-based methods
b) Network-based methods

The existing sentiment and syntax analysis schemes are customized for special data types, thereby being inadequate for fake news detection systems.

features such as:

a) Content-based features (e.g., words and segments appearance, part of speech),
b) Network-based features (i.e., re-tweets or tweets propagation)
c) Twitter-specific Memes (i.e., Hashtag or shared URLs).

Traditional fake news detection approaches pay more attention to reducing content leakage, which may provide misleading information when original articles are imitated or modified. Open information extraction (OIE) is a task of extracting factual information from textual data such as Twitter posts, spams, and articles in social media. More recently, OIE tools have been used for producing grammatical clauses, which can be used for topic extraction purposes.

For example, Ghosh and Ghosh (2016) demonstrated their approach by extracting topics from a collection of 50,000 microblogs during a disaster event. Similar to the above semantic knowledge-based approach, our proposed approach aims to grasp an understanding of news through complete comparisons of news content. This approach no longer relies solely on statistical, sentiment, or syntax analysis to detect fake news but uses topic and event-level analysis to understand patterns that are deeply embedded within the news for improved detection accuracy.

12. APPLICATIONS

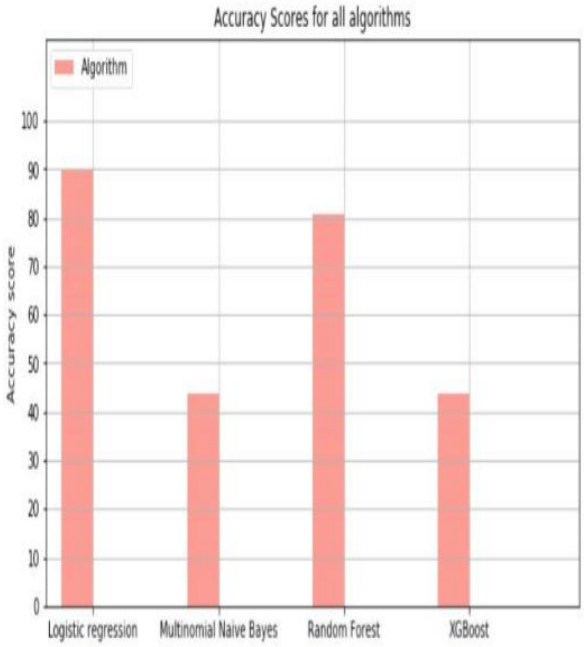
Recent efforts on fake news detectors have focused on varied approaches.

B.S. Detector – Alerts users of unreliable news sources by searching all links of a given webpage for sources that have been collected in an unreliable-news database, which includes samples of fake news, extreme bias, state news, junk science.

Fake News Detector AI – Identifies fake-news websites by measuring similarity to existing fake-news websites using artificial intelligence techniques as a BlackBox. This system uses a neural network-based feature analysis (e.g., headline, code structures, site popularity) approach on known websites, thereby yielding the credibility of the tested websites.

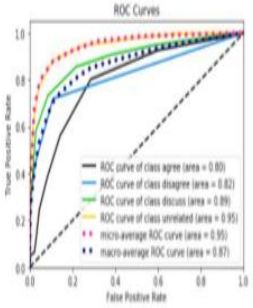
13. ACCURACY OF ALL ALGORITHMS

We can conclude that Logistics Regression is the best algorithm with the highest accuracy around 90 percent.




```
In [46]: # ! pip install scikit-plot
         import scikitplot as skplt
         import matplotlib.pyplot as plt
         import warnings
         warnings.filterwarnings("ignore")

         # y_true = classes# ground truth labels
         y_probas = rf.predict_proba(test_features)# predicted probabilities generated by sklearn classifier
         skplt.metrics.plot_roc_curve(targets_test, y_probas)
         plt.show()
```



14. CONCLUSION AND FUTURE WORK

This paper characterizes the applications of NLP (Natural Language Processing) techniques for detecting the 'fake news', that is, misleading news stories that come from non-reputable sources. as well as the important terms of each news and their word embeddings. As fake news and disinformation continue to grow in online social media, it becomes imperative to gain an in-depth understanding of the characteristics of fake and real news articles for better detecting and filtering fake news. Our analysis shows that a model that we have built can differentiate between “Real “news and “Fake” news. Naive Bayes classifier to create a model to classify an article into fake or real based on its words and phrases. As we have used four algorithms and analyzed the best algorithms which have the highest accuracy. On the other hand, the difference in the topics and word embeddings shows little difference between fake and real news.

Future Work:

Our future work is centered on exploring the word2vec algorithm, a computationally-efficient predictive model based on neural networks for learning the representations of words in the high-dimensional vector space, to learn word embedding of the important words or terms discovered via the aforementioned tf-idf analysis. Rather than comparing the few important words of each new article, word2vec will allow us to compare the entire vector and embeddings of each word for broadly capturing the similarity and dissimilarity of the content in the fake or real news.

REFERENCES

[1] Articles are ubiquitous. "How Fake News Goes Viral: A Case Study," The New York Times,November20,2016,accessed1/22/2017at https://www.nytimes.com/2016/11/20/business/media/how-fake-news-spreads.html?_r=0.2NewsfeedFYIBlog,December12,2016: http://newsroom.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/.

[2] "What Is Natural Language Processing?" SAS, www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html.

[3] "Fake News Challenge Stage 1 (FNC-I): Stance Detection." Fake News Challenge, fakenewschallenge.org/.

[4] "CountVectorizer, TfidfVectorizer, Predict Comments." Kaggle, www.kaggle.com/adamschroeder/countvectorizer-tfidfvectorizer-predict-comments.

[5]"KDnuggets." KDnuggets Analytics Big Data Data Mining and Data Science, www.kdnuggets.com/2017/04/machine-learning-fake-news-accuracy.html.

[6] "UKnowIT (Self Service)." How Can I Tell If a Website Is Credible?, uknowit.uwgb.edu/page.php?id=30276.