Fake News Detection

Shaik MD.Ibnay Momen
dept. of CSE
East West University
email: 2016-2-60-100@std.ewubd.edu

Md.Mehedi Hasan
dept. of CSE
East West University
email: 2017-1-60-119@std.ewubd.edu

Md.Mukit Hasan
dept. of CSE
East West University
email:2017-1-60-113@std.ewubd.edu

Md.Jadid Mostafiz
dept. of CSE
East West University
email:2014-3-60-029@std.ewubd.edu

***Abstract*—Consuming news from social media is becoming increasingly popular nowadays. Social media brings benefits to users due to the inherent nature of fast dissemination, cheap cost, and easy access. However, the quality of news is considered lower than traditional news outlets, resulting in large amounts of fake news. Detecting fake news becomes very important and is attracting increasing attention due to the detrimental effects on individuals and the society. In this paper, we propose machine learning techniques, in particular supervised learning, for fake news detection. More precisely, we used a dataset of fake and real news to train a machine learning model using Scikit-learn library in Python. The outcome of our experiments was that the linear classification works the best with the TF-IDF model in the process of content classification. The Bi-gram frequency model gave the lowest accuracy for title classification in comparison with Bagof-Words and TF-IDF.**

I. Introduction

Due to the increasing amount of our time spent online, people tend to seek out and receive news from social media. The reasons for the fast increase of users' engagement in news online are mainly because of the nature of social media such as the easy access, less expensive, and fast dissemination. Despite these advantages, the quality of news on social media is considered lower than that of traditional news outlets. Large amounts of fake news, i.e., those low quality news with intentionally false information, are widely spread online. Fake news has significant detrimental effects on individuals and the society. fake news intentionally misleads people to believe false information. Fake news change the way people respond to real news. For example, people are confused about the news they read, which impedes their abilities to differentiating the truth from falsehood. The trustworthiness of entire news ecosystem is broken due to fake news. Besides detecting fake news articles, identifying the fake news creators and subjects will actually be more important, which will help completely eradicate a large number of fake news from the origins in online social networks. Therefore, it's generally not satisfactory to detect fake news only from news content, and auxiliary information is needed, such as user engagements on social media. In this paper, we present our preliminary experiments on applying machine learning techniques for fake news detection. In particular, we studied and developed methods and tools for detecting fake news, also, proposing a methodology for that purpose and implementing an algorithm which allows reporting, respectively detecting fake news articles. We used the machine learning library Scikit-learn in Python since it has built-in methods that implement different classification approaches. We recommend the users of our tool not to take the results of a fake news detection as a ground truth but to use also the filter of their critical thinking in order to decide the nature of the article. The objectives of this paper are: 1) to overview fake news corpora requirements 2) suggest pros and cons of varieties of fake news, as counterparts to serious genuine reporting. This research ultimately supports the development of an automated fake news detection system as part of a broader news verification suite. The main objective is to detect the fake news, which is a classic text classification problem with a straight forward proposition. It is needed to build a model that can differentiate between "Real" news and "Fake" news

II. Related Work

Most of works which discuss the detection of fake news and biased information are relevantly modern. Some of them are based on studying the credibility of a news source regardless of the news content. This process is not a good way because a news source could be classified as untrusted and at the same time it could publish a true fact. Another project that discusses fake news problem based on news sources and not on the article content is http://bsdetector.tech. This website offers an extension to be installed on Internet browsers and which verifies any website the user will access. Then it gives a notification if the website is classified as unreliable. In contrast, in our tool, we used a static dataset where no feedback will be used for training the model. We performed studies only on the text features without taking the visual features into consideration. We studied articles from news websites, and not only social media websites. Paper discusses the detection of fake news based on clickbait titles. The system studies the relation between the article title and its body. A similar approach to detect the similarity between

headline and article content using the same dataset offered by the fake news challenge. Other strategy to detect fake news and alternative facts like is crowdsourcing2. Based on social argumentations online, the authors present a prototype system to verify the credibility of a news article. They developed a graph-theoretic framework by using substantial discussion basics. This argumentation graph is filled with information from different Internet users (especially social networks users) through a web-based application. Another strategy for fake news detection is presented. The authors show that satirical news could be used as a guide in distinguishing between fake and real news. They present an automated tool which can indicate deceptive information fast and efficiently by analyzing the satire news text characteristics using Support Vector Machines models and 10-fold cross validation to train and evaluate a machine learning model. The main difference to our approach is that they focus on detecting satire articles.

III. Methodology

Predicting anything is always very interesting, Machine learning helps us very much in this sector, it can easily generate values of future with the help of many approaches. For our project, we did find stocks data from all available online and offline sources. The dataset we'll use for this python project- we'll call it news.csv. This dataset has a shape of 6335×4. The first column identifies the news, the second and third are the title and text, and the fourth column has labels denoting whether the news is REAL or FAKE. Further we will divide the data into two parts, training data and testing data, where 67 percent of the data will be used for training and 33 percent of the data will be used for testing.

We are thinking to solve the problem using below supervised learning techniques to build our model: Using sklearn, we build a TfidfVectorizer on our dataset. Then, we initialize a Passive Aggressive Classifier and fit the model. In the end, the accuracy score and the confusion matrix tell us how well our model fares. To solve the problem, we will follow below steps –

1) Using dataset or We can import data from web of any duration time.
2) Provide the data to the system.
3) Train the system.
4) System will predict the output.

.

The original dataset will be converted to a normal form in the data preprocessing phase in order to get better decision making and mitigating the computational overhead. Here, tittle column dropped from dataset. As it has been said in the contributions section, this has model provided "lower computational complexity on account of fake news detection". With respect to this reduction, in the first step passive Aggressive classifier has been applied and its computational complexity reflects as O(k*d), where k is number of test train split in the respective dataset and d indicates dimensionality of data. in this section given test size is 33% and random state is 8. At this step, Fake and Real find out how many data is given here. According to the find out vocabulary test case. Accuracy defines the working ability. In this study, it can be seen that the accuracy in the second layer for fake news is 93.5 percent % which is a decent percentage. Flow Chart is given below:

Fig. 1: Flow Chart.

IV. Implementation

A. Data Collection

There are several datasets of Fake News Detection available. We collected our data from Kaggle. This dataset has a shape of 6335×4. The first column identifies the news, the second and third are the title and text, and the fourth column has labels denoting whether the news is REAL or FAKE.

B. Data processing:

1) In our dataset there were a column name "title" which is not required for our project that's why we drop that column from dataset.
2) We check all the cell in our dataset whether there is any null cell.
3) 3) After that we split our data into train and test set, the percentage of 67 train data and 33 test data.

.

C. Model Development:

In our model, we used one machine learning algorithms which is Passive-Aggressive-Classifier and for the implementation work, we used Python 3.6.5 as our programmable language in anaconda jupyter notebook environment. And also, we used some built in function like as numpy, pandas, sklearn, matplotlib, stopwords for our project. We import our data and process it and after that we implement our model. Passive-Aggressive-Classifier: Passive Aggressive algorithms are online learning algorithms. Such an algorithm remains passive for a correct classification outcome, and turns aggressive in the event of a miscalculation, updating and adjusting. Unlike most other algorithms, it does not converge. Its purpose is to make updates that correct the loss, causing very little change in the norm of the weight vector.

V. Experimental result

We got an accuracy of 93.11 % with this model. Finally, let's print out a confusion matrix to gain insight into the number of false and true negatives and positives. So with this model, we have 973 true positives, 974 true negatives, 84 false positives, and 60 false negatives.

VI. Conclusion

The problems of fake news and disinformation play an important role on nowadays life. This is because the advanced level of technology and communication methods we have enabled information spreading among people without any verification. This is a reason why researchers started searching for solutions to stop fake news and disinformation from spreading easily. However, it is well known that controlling the flow of information online is impossible. In this paper, we performed an attempt to verify the news articles credibility depending on their characteristics. At this aim, we implemented an algorithm combining several classification methods with text models. It performed well, and the accuracy results were relatively satisfying. As future work, we plan to better study the combination between the feature extraction methods and the classifiers as we will be able to choose the text representation model that performs best with the classifier. Moreover, to achieve a higher accuracy, we will have to implement a more sophisticated algorithm which may use data mining technologies with big data, because creating a big dataset including more types of news articles with more class variables (labels) will help raising the accuracy score.