



Cyberspace News Prediction of Text and Image with Report Generation

Mrs. Bihade mam, MR. Ali Sir, Ankita vibhute, priti chandane.

Dept Of Computer Engineering

Abstract:

The cyberspace news consumption is increasing day by day all over the world. The main reason for cyber space news consumption is due to its low cost, easy access and easy sharing facility which lead people to consume news rapidly without the knowing whether the news is fake or true. Thus, the wide spread of fake news which have the serious negative impacts on society. Therefore, false news prediction on cyberspace is attracting a tremendous attention.

This model works on processing the text and images together by providing an interactive Application Interface (API), i.e., text by applying the model Logistic regression classifier and image by applying self-consistency algorithm. The natural language tool kit (NLTK) model is used for these implementations through python. Once the news is predicted fake, a report is redirected to the authorized website (cybercrime department) to take the immediate necessary actions required to stop this news from spreading.

Index Terms — Cyberspace, fake-news, text and image, Logistic regression classifier, self-consistency algorithm, report, redirect.

I. INTRODUCTION

Now a days the cyberspace news consumption is increasing day by day all over the world. The main reason for cyber space news consumption is due to its rapid spread of information and its easy access which lead people to consume news rapidly without the knowledge of whether the news is fake or real.

It leads to the wide spread of fake news which leads to the negative impacts on society. Therefore, false news prediction on cyberspace is attracting a tremendous attention.

The main purpose of this project is to classify the news as truthful or fake using various data mining techniques.

This model is a solution to all these problems of fake news in cyberspaces that is fast growing. In particular the datasets which are trained by various machine learning techniques like data pre-processing, feature selection, self-consistency etc. and all these are implemented by natural language processing in python.

Here we detect both forms of fake news, i.e., both text and image streams. Once the news is Predicted as fake then the report is generated and it is immediately redirected to the authorized page (cybercrime department) insisting the seriousness of the news for which the actions will be taken accordingly.

Through this we try to bring a safe and trustable cyberspace experience to people who rely on this. They can now verify news

before they are believing or forwarding them to others.

This model works on processing the text and images together by providing an interactive Application Interface (API), i.e., text by applying the model Logistic regression classifier and image by applying self-consistency algorithm.

A. Logistic Regression Classifier

logistic regression is an example of supervised learning. It is used to calculate or predict the probability of a binary (yes/no) event occurring. An example of logistic regression could be applying machine learning to determine if a person is likely to be infected with COVID-19 or not.

B. Self Consistency Algorithm

The k-means algorithm and the principal curve algorithm are special cases of a self-consistency algorithm. A general self-consistency algorithm is described and results are provided describing the behavior of the algorithm for theoretical distributions, in particular elliptical distributions. The results are used to contrast the behavior of the algorithms when applied to a theoretical model and when applied to finite datasets from the model.

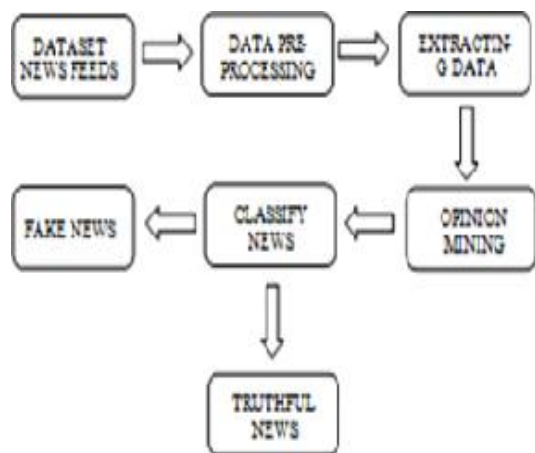
II. DATASET AND METHODS

A. Description of Dataset

The Cyberspace news prediction of multimedia content is fairly a new problem. Therefore, there are only a smaller number of data sets that are public. Usually, the news prediction can be classified based on the accuracy of news as completely-true, completely-false, obscure-true, obscure-false and soon. To make it simple and clear we are using only two variables i.e., true and false along with probability of accuracy of news. Hence, the labels are clearly differentiated into only two forms i.e., True and False.

B. Methods

There are various models involved in implementation of each module starting from data pre-processing to classification of news.



III. EXPERIMENTAL MODEL

A. Data Pre-processing

The dataset we are using for the model needs subjective refinement of noisy data which involves punctuation, stop-word, tokenization, casing. These are ultimately done to reduce the original size of data and remove the noisy (irrelevant) information from the dataset.

An accurate processing function is created in order to refine the dataset by removing the punctuation and non-alphabet characters from each sentence in the dataset. N-gram model is a natural processing technique which involves text-based and word-based features. We use word-based features to classify sentence in a dataset. The tokenizer in N-gram feature is used to slice the sentence to the length of n.

Stop words are the unnecessary information in the sentence which contributes to a greater number of noises. Some of the stop words are conjunctions, preposition, articles etc. Therefore, these are words are removed from each sentence and trained accordingly.

The trained document after the removal of noisy words is used further. The final output of data pre-processing classifier is shown in Fig. 2. The data pre-processing model is implemented using Natural Language Toolkit (NLTK), in which the output of the dataset trained after pre-processing is shown in the form of graph. The most specific part of pre-processing model used here is stemming and tokenization.

Stemming is the natural language processing technique that is used to reduce words to their base form, also known as root form. It is used to normalize text and make it easier to process.

Tokenization is the process of dividing the text into a set of meaningful pieces. These pieces are called tokens.



Fig. 2. Data Pre-Processing

B. Feature Extraction

feature extraction is needed to reduce the size of text features and volume of dataset.

Term Frequency (TF), takes the input of number of words from the sentence to find out the similarity of words among

the sentences. Each sentence in the dataset is represented based on count of words. These word counts are then converted into probability which exists in the dataset

The Term Frequency-Inverted Document Frequency (TF-IDF) is basically used for the retrieval of information based on the statistical feature from the dataset. Therefore, this increases the frequency of occurrence of words in the dataset. Thus, one of the main features of TF-IDF is that it classifies the frequent occurrence of words than the occurrence of rare words in a dataset.

The most important characteristic of feature extraction is that it is able to build a vocabulary of words, which is used further for the classification. Therefore, the final output consists of the entire vocabulary of words in a dataset and the count for the occurrence of words which is shown in Fig. 3.

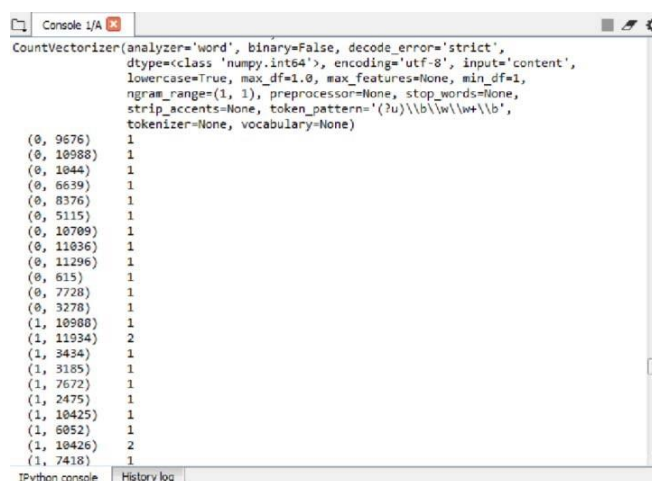


Fig. 3. Feature Extraction.

C. Classification Process

The final step of the model is classification process. In order to find the accurate model, we carry out all the five-classification process namely Naïve-Bayes classifier,

Logistic regression classifier, Linear Support Vector Machines (LSVM), Stochastic Gradient Descent (SGD), Random Forest classifier. The implementation of these classifiers is done using python Natural Language Toolkit (NLTK). These five classifiers are validated using K-fold cross validation and the accuracy of the output are found for all these models. Out of all these the best two performing models is taken which is known as candidate models. The candidate models are

1. Logistic regression classifier
2. random forest classifier

Logistic Regression Classifier predicts the output of categorical dependent variable therefore the outcome must be a categorical or discrete value. It can be either yes or no, zero or one, true or false etc.

Random forest classifier contains a number of decision trees on various subsets of the given datasets and take the average to improve the predictive accuracy of that dataset.

Further, logistic regression and random forest classifier are performed to find out the best parameter with Grid Search method. Grid Search method is essentially an optimization algorithm which lets you to select the best parameters at which the model gives the best accuracy.

By knowing the accuracy of these models, the best model is saved and it is used further for our classification.

D. Report Generation

The most important part of this model is generation of report, when the news is predicted as fake then it is redirected to the authorized website to stop this news from spreading.

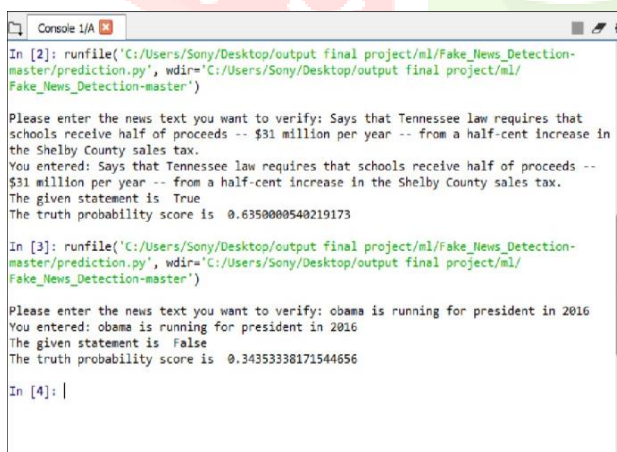
This sample of the prediction of news for both true and false news is shown in Fig 6.

This final output is achieved by the processing of above mentioned model in an efficient way and the maximum accuracy is met with.

Therefore, in this model we have found a way to detect such fake news in both the forms of text and image. By redirecting the fake news to the authorized website (cyber-crime department), thus it reduces the spreading of fake news distinctly. This model can be further discussed for the future improvement in fake news detection which can be in audio, video streams and commercialize the field to other applications.

V. REFERENCES

- [1] Faiza Masood, Ghana Ammad, Ahmad Almogren, Assad Abbas, Hasan Ali Khattak, Ikram Ud Din, Mohsen Guizani and Mansour Zuair, "Spammer Detection and Fake User Identification on Social Networks," IEEE Trans. Inf. Translations and content mining, vol. 7, pp. 2169- 3536, 2019.
- [2] Himank Gupta, Mohd. Saalim Jamal, Sreekanth Madisetty and Maunendra Sankar Desarkar, "A framework for realtime spam detection in Twitter," IEEE Int. Conf. Communication Systems and networks, pp. 2155-2509, 2018.
- [3] K.Sakthidasan, G.Srinithya, V.Nagarajan (FEB 2014), "Enhanced Edge Preserving Restoration for 3D Images Using Histogram Equalization Technique", International Journal of Electronic Communications Engineering Advanced Research, Vol.2, SP-1, Feb.2014, pp. 40-44
- [4] S. Kwon, M. Cha, K. Jung, W. Chen and Y. Wang, "Prominent features of rumor propagation in online social media," IEEE Int. Conf. Data Mining, pp. 1103-1108, 2013.



```

In [2]: runfile('C:/Users/Sony/Desktop/output final project/ml/Fake_News_Detection-master/prediction.py', wdir='C:/Users/Sony/Desktop/output final project/ml/Fake_News_Detection-master')

Please enter the news text you want to verify: Says that Tennessee law requires that schools receive half of proceeds -- $31 million per year -- from a half-cent increase in the Shelby County sales tax.
You entered: Says that Tennessee law requires that schools receive half of proceeds -- $31 million per year -- from a half-cent increase in the Shelby County sales tax.
The given statement is True
The truth probability score is 0.6350000540219173

In [3]: runfile('C:/Users/Sony/Desktop/output final project/ml/Fake_News_Detection-master/prediction.py', wdir='C:/Users/Sony/Desktop/output final project/ml/Fake_News_Detection-master')

Please enter the news text you want to verify: obama is running for president in 2016
You entered: obama is running for president in 2016
The given statement is False
The truth probability score is 0.34353338171544656

In [4]: |
  
```

Fig 2. Output Of Prediction

IV. CONCLUSION

The consumption of news is increasing day by day in cyberspace than the traditional media. Due to its increasing popularity and user-friendly access, it leaves a huge impact on individuals and society.