Real and Fake News Classification Using Natural Language Processing

Shivam Kumar[1], C. Santhana Krishnan[2], M. Ramya[3]

[1]Department of CSE, SRMIST, Kattankulathur, India. E-mail: sk2941@srmist.edu.in

[2]Department of CSE, SRMIST, Kattankulathur, India. E-mail: santhanc@srmist.edu.in

[3]Department of CSE, SRMIST, Kattankulathur, India. E-mail: rm0394@srmist.edu.in

Abstract

The concept of Real and Fake news Classification and Detection is a domain which is still in the initial-development stage as compared to other projects of similar kind in this domain.ML or Machine Learning is a useful part of this project. The purpose of using these algorithms is to help the users to understand the various difficult and unyielding problems and to build Smart Artificial Intelligence and Machine Learning Systems to tackle problems for this concept. For the purpose of this research, we have used the concept of NLP along with two popular Machine Learning Algorithms for the purpose of the classification of real and fake news. They are Logistic Regression and Decision Tree Classifier. Other Algorithms like Random Forest, Support Vector Machine can also be used for this. The purpose of the project that has been built here is not to simply perform classification of the news articles as one cannot simply implement the ML algorithms and then predict whether the news is real or not. No, what has been done here is a clear-cut implementation and a mix of Data Science Tools as well as ML concepts for the classification as well as prediction of fake news. Various ML models will be implemented here for the prediction of news. The process of the classification will focus on using data science tools for pre-processing of the text and then using the results of the pre-processed dataset to build a improved model for the project. The major obstacle which was tackled whilst project was the lack of a properly processed dataset as well as a pre-defined model to differentiate between the two categories of news as mentioned in the title of the paper. For simplicity's sake, some of the more commonly known ML algorithms and classifiers have been implemented on some datasets that are available on the internet. The results, when the ML models were implemented on the dataset, have been very encouraging and can prove to be very useful if any future work is done on this project or in this particular domain.

Keywords: Natural Language Processing, Classification Purpose, Proposed System.

DOI: 10.47750/pnr.2022.13.S03.236

INTRODUCTION

Let us take a look at what Fake news might be. FAKE news can be the spread of any sort of misinformation, whether it be through online sources, newspapers, magazines, websites etc. The most prevalent form of spreading fake news is through news channels and through news reporters as it is very easy to twist facts in such a way as to spread an entirely different point as compared to what we want to say. It is not only a problem of recent times, as Fake news have long been in existence.

In this modern era, it is possible for anyone to spread any sort of news through electronic means which might not be credible as in today's era, it is highly possible that fake news will spread much faster than real news. Especially on the internet, the amount of fake news is very high which is easily accessed and is deemed believable by most of the people today. This problem of detecting the fake news is resolved through implementating the mentioned ML models by using the concept of NLP [7]. The major purpose of the projects that were built previously was to classify and detect the online news and social media posts in this era where there has been a massive increase in the amount of fake news prevailing on the Web. To provide an overview, what has been done in this project is that a dataset is taken, and then after pre-processing, NLP and ML algorithms have been implemented to develop a model which is then used on several publicly available datasets to show the accuracy of the models which have been built in this project.

Another important aspect that has to be taken into account is the number of times that a given word occurs in the given dataset which is being for the real and fake news classification. The instance of Cloud Visualization is used in the implementation of this project. This representation of words in the form of a cloud represent various words. Let us consider some words such as Russia, Ukraine, War, Political etc. These words will occur a maximum number of times in the dataset as the probability of news generation is very high, whether it be real or fake. As already mentioned, various detests have been used which can contain several types of stories, whether they be real or fake. The Classification Model is generated using the ML models and the word cloud is used for the classification and prediction of the category of news articles i.e., to classify whether they are real or fake. All of this will be explained properly int the implementation aspect of the project.

Literature Survey

The purpose of this project is to create a real or fake news detection Model. This is quite an important and innovative idea, especially in these times, when the problem of fake news is at its peak. Various Studies, Projects and Experiments have been conducted in this field. There are various methods which all have the same outcome, the classification of various types of news. Let's look into some of the works which have already been developed and implemented int this field.

The first method is the oldest, the classification which is performed by people themselves who are knowledgeable in this field. Natural Language Processing has been used for the implantation of concepts such as stop-words etc. The developers of the prototype project proposed a model in which, the categorization of news was performed live, i.e., at that time only [4]. The total process took about 40 seconds in which the news can be categorized, whether is based on a particular propaganda, or a particular political debate, or for a specific person. Naïve Bayes Classifier was used for this. In this ML algorithm, a small dataset is used for the processing and only a limited amount of space is required for the storage of the models. Facebook post prediction using genuine or fraudulent labeling is used through Bayes [2]. The results have shown that this approach is quite successful. The method which was proposed here is capable of segregating the news which is not real into different components: fabrication, hoaxes and hilarious fake. It also provides an means to separate the various categories of fake news. There was a research which was sponsored by EU Commission. This was a three year long project. The purpose of this project was to deal with rumors, fake news on social media, rumors and the analysis of their frequency. It was sponsored from 2014 to 2017. It is a well known fact that fake news generation rate is very high and is often quite believable, due to which it is a very tedious task to detect fake news. Different Social, Behavioral and social judgmental cues have been used for the prediction of fake news by Facebook. The model implemented by Facebook uses the given parameters to detect fake news. SVM is another popular method which has been used for real and fake news classification on the organizational requirements [1]. Stance Detection is another method which can be used for the classification process. It uses the concept of n-gram matching for detecting whether the given news is fake or not. It uses the news articles present and checks whether the given articles are connected to each other or not. This method has also been very popular for identification of fake news, especially clickbait detection. The dataset which has been worked on in the given project is available on public sources [8]. Another concept that can be used for detection of Fake News is Deep learning. This is also a very innovative model, which has been used extensively in the past [5]. When this model was implemented with concepts of Natural Language Processing, it produced excellent results.

Existing System

There are various models which exist for Real and Fake news Detection. The most prevalent system consists of a model that detects fake news based on keywords as well as the headlines, simultaneously. The FDML model was such news with certain topics have high probabilities to be classified as fake news and a few authors have high probability to publish fake news. FDML introduce a task gate to selective integrate representations supported different tasks and style a dynamic weight strategy to balance the importance between two tasks. The imbalance learning problem to propose an easy dynamic weighting strategy where the load of every task is dynamically adjusted in each iteration. This model trains the data which is present in the dataset and splits into training and test set. Then, the classification and prediction of news takes place. As mentioned, this classification takes place on the basis of segregation of words as per their presence in the word cloud and as per the headlines of the news article.

Proposed System

The proposed model is to create a machine learning model that's capable of classifying the news into a category – Real or Fake. The fake news are considered to be widespread and controlling them is extremely difficult because the world is developing toward digital everyone now has access to internet and that they can post whatever they need. So there's a greater chance for the people to urge misguided. The machine learning is usually build to tackle these sort of complicated task love it takes more amount of your time to analyse these sort of data manually. The machine learning are often wont to classify the news is fake or not by using the previous data and make them to know the pattern and improve the accuracy of the model by adjusting parameters and use that model because the classification model. Different algorithms are often compared and therefore the best model are often used for classification purpose.

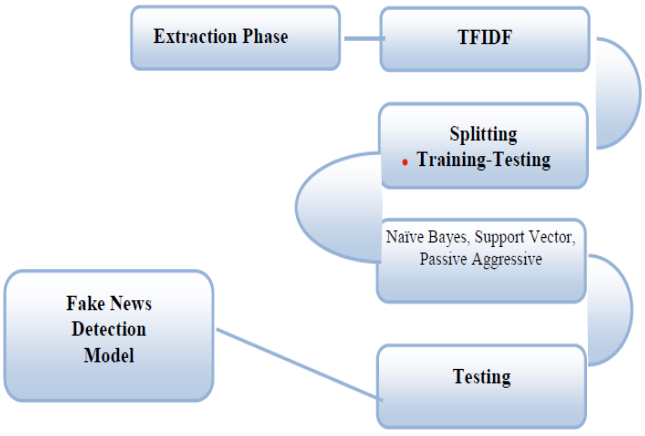

Fig. 1 Proposed System

The proposed system has several advantages over the prevailing one. The classification model build had the power to classify the new whether it's real or not. By classifying the new by using this process the prospect of individuals getting misguided are often reduced.

METHODOLOGY

- **Detailed Description of given Dataset**

The dataset which has been used for this project has been taken from various sources, all of which are available on the internet, free of cost. The sources of the dataset for the news articles can be Kaggle, GitHub or some other portal [3]. These articles were collected from various news sources and that they were labelled as Real and faux. The news articles which have collected into the dataset are categorized into two categories – Real and Fake. The preprocessing of the dataset has to take place for the process of classification to take place. This division are often seen within the project where the news articles have been sorted into the a separate category and others in a separate category. Some of the articles have not been classified as Real or Fake in the pre-processing module as some of their information such as ID, Label etc. is missing [15]. The pre-processing of this dataset itself is quite a tedious task and if proper steps are not taken, can result in quite an imbalanced dataset. This problem is not only in this project but in the others as well. For the rectification of this error, we skip the records with missing data.

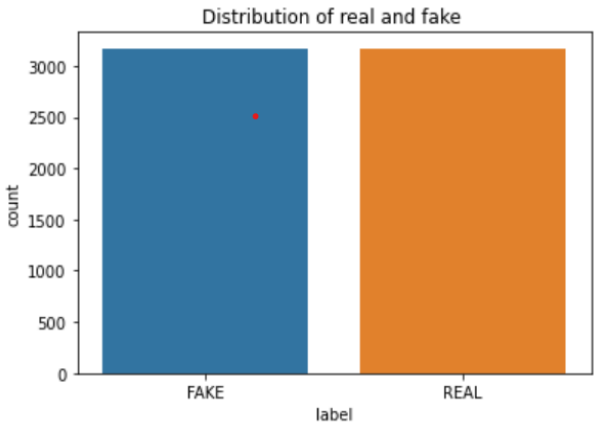

Fig. 2 Description of Dataset

- **Exploration of different ML Models**

There are several types of Machine Learning Algorithms which can be used for the purpose of this project. After rigorous experimentation and testing of different algorithms, different ML models were deployed for the classification process. Some models which had already proven to be effective were taken into consideration whereas other models which did not give correct, or desired results, were removed from the proposed work. The models which have been retained for the project work are Logistic Regression and Decision Tree Classifier as they have given a higher accuracy as compared to other models like SVM, Naïve Bayes etc. After the identification of the models to be deployed, the NLP concepts were implemented on them for the prediction of Fake news. For safety, nltk library of NLP module was also implemented on some of the higher performing rejected models. But still, the accuracy of the two mentioned models has proven to be greater as compared to other Machine Learning Models.

- **Naïve Bayes:** The Naïve Bayes Algorithm was implemented for the classification of Real and Fake News because, as per the previous works, its performance was quite good. But while implementing this model along with NLP concepts, its performance was found to be lacking as compared to others [9]. This can be explained by the algorithmic formula for the Naïve Bayes as well as the classification report that was generated.
- **Support Vector Machine:** Support Vector Machine is another model that was implemented for the purpose of Real and Fake news classification. This model is very useful and has several merits. The rate of training of this model is relatively higher as compared to others. Also, this model was used by previous works for the Real and Fake news classification [12]. However, this model also did not perform accurately. Simply, the accuracy of this model was lower as compared to others. This model has a number of advantages, such as the ability of tolerance to unimportant material in the dataset [13]. But still, for the purpose of this project work, this model has been discarded.
- **Passive Aggressive:** The Passive Aggressive model is not as much used for the Real and Fake News Classification, but still it is an important and upcoming model in various other domains, due to which this was also taken into consideration. The implementation of this model is easier as compared to other complex models and has been demonstrated by several works of different authors [6]. But still, seeing the low accuracy score of this model, it was also discarded from the given project work.
- **Logistic Regression:** The Logistic Regression is quite a popular model and has been used in various domains of project work. The accuracy score of this model was quite high as compared to other algorithms and the prediction was also accurate. The Logistic Regression Model is capable of handling large amounts of data which made it an essential part of the prediction and classification model. This model has been used in this project for the classification purpose.

RE - INITIALIZATION

This process is known by various names such as re-coding, re-initialization or pre-processing. This is the initial module of the project. The purpose of this module is to process or re-initialize the given data into a suitable format which can be used by the Machine Learning Model. There are various methods of proper re-initialization such as assigning a proper ID, cleaning of repeated data, filtering unwanted data etc. All of this is managed in this module of the project [14]. The sklearn and NumPy libraries are used for this. Along with this, various NLP concepts are also used in the pre-processing module. The major purpose of this module is to ensure proper categorization and representation of data i.e., the labels of real and fake are accurately assigned to each row, the rows are unique etc. Then the concept of NLP is used to ensure that words are present int their base format for easier processing. There are various other concepts of Data Science as well as Machine Learning that have been taken into account for this project. Tokenization is an important aspect which is used to split the text in any row in the dataset into singular words. The concept of stop words is also used to remove the words which do not carry any meaning in identifying whether the given news is real or not.

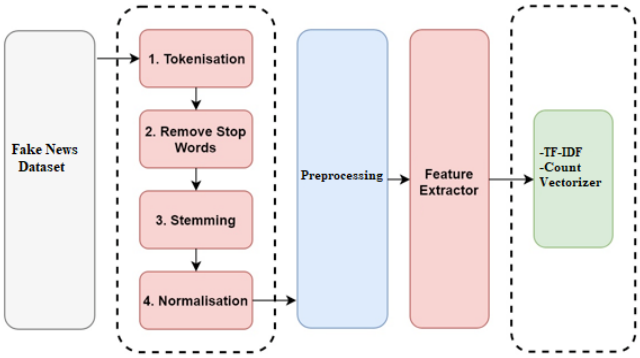

Fig. 3 Processing of Dataset for ML

MACHINE LEARNING

In the 21st century, in the modern era, there is a rapid boom in the generation of fake news, rumors etc. whether it be online or offline. Due to this rapid increase, the need of a device or a software to detect fake news has also increased. Machine Learning Models like Decision Tree Classifier and Logistic Regression have been developed and implemented along with Natural Language Processing for the classification and prediction of the fake news which has been generated. The goal of this project is to merge the data science as well the concepts of Machine Learning along with NLP to sort the category of news based on the headlines as well as the news article. This means that the classification can take place either by the title of the news article or the content. This module has been successfully implemented in the project.

- **Natural Language Processing**

The vast amount of repetitive and irrelevant features which are present in the given dataset often result or cause a significant negative affect on the accuracy of the result as well as the performance of the classifier. This causes a huge problem in the prediction and the classification module. In order to tackle these issues, feature extraction is used to reduce the length of the text as well as to implement stop-words and other modules. The process of NLP consists of various steps and many functions are implemented in this. Some of them are conversion of the casing of the alphabets, sum of the number of words in a particular part of the dataset, removal of punctuations etc. The Natural Language Toolkit or nltk of the NLP library is used in this project. The concept of stop-words, stemming, tokenization etc. has been implemented in this module and has been successfully integrated into the project

- **Count Vectorizer**

This is an important part of NLP and ML algorithms. The purpose of CountVectorizer is to take a particular aspect of the dataset and then removing punctuation marks, conversion of the casing of the words, pre-processing of the text in the dataset etc. The first step is to gather the data into a word cloud using the implemented modules [10]. Then, the generated vocabulary is used for the classification and prediction of fake news. By using the mentioned module i.e., vectorizer, a table is generated which stores the number of times a word occurs in the dataset and where it occurs.

- **Term Frequency- Inverse Doc. Frequency**

TF-IDF is another important aspect of NLP which has been used in this project. It stands for Term Frequency. The IDF stands for Inverse document Frequency. The number of times a word occurs in the dataset is stored after the re-initialization of the dataset. The above module is used to calculate the importance of a given string as per the classification point of view [11]. Its purpose is to convert sentences into an array of integers. For their usage in further modules. The formula which can be used for the measurement of this and for the measurement of corpus is :

$$TF\text{-IDF} = T(\text{w})\text{d} \times IDF(\text{w})\text{D}.$$

Let's try to understand this concept through a example. Consider a situation in which we have a news report which consists of over 500 words. To determine the TF and IDF for the given word "president". Term "president" is present in the document for a total number of 15 times. By this logic, the Term Frequency will come out to be 15/500 i.e., 0.03. Now we calculate the IDF for 300 reports for the given word. By using the formula, the IDF and TF-IDF is calculated.

IMPLEMENTATION

The various modules which have been described in this research paper are implemented in the project of Real and Fake News Classification. Different tools, functionalities have been used for the implementation of this project. First off, the Data Science tools such as sklearn, NumPy have been used for the re-initialization. But first, the pandas library of python is used. It is a freely available, open-source library that has significant uses in the field of Data Science and Machine Learning, one of them being the classification as done in thus project. The accuracy of different models is also represented through graphs by using matplotlib to plot graphs. Various graphs can be used such as bar graph, scatter plot etc. Here, a simple Bar Graph has been used for the visualization.

Prior to the implementation of the various ML algorithms, the concepts of NLP such as Count Vectorizer and TF are added in the project. This is done using the nltk library. NLTK stands for Natural Language Toolkit. It is used in the re-initialization phase and also in the subsequent steps. The project was created using the software known as Anaconda Jupyter. So, all the importing of libraries as well as implementation of models takes place in Jupyter notebook.

While ensuring that no alterations were made in the processing of the training as well as the testing data, we have further attached the data that has been tested with certain algorithms of Machine Learning and Natural Language Processing.

One of the pre-requisites of this project was to deploy a model that is able to calculate the Term Frequency as well as Inverse Doc. Frequency, which has been successfully implemented. The purpose of this project was to classify the given news articles into two categories- Real and Fake and also to train the model on the given dataset to predict whether the news which is prevailing at that moment is authentic or not. Sometimes, the classification models also give errors in predicting the results of real or fake news. The primary goal is to create a model that supports count vectorization and TF. -IDF. To overcome the problem of errors, we can segregate the text present in the documents or in the dataset into separate entities for the classification module. Also, before the actual classification or prediction actually takes place, we have to perform feature extraction as well. After these steps are complete, we can deploy the NLP and Machine Learning models to detect their accuracy as well as to predict whether the given news article is real or not. This phase's goal is to reduce the size of the information by eliminating unnecessary data that isn't necessary for categorization. The information was then altered so that it wouldn't produce impartial findings when applying ML techniques if the first half of the information had a false label set and the second half had a true label. Another useful function of NLP is tokenization which after taking a string of text, splits it up into individual words. For clarity, the words have been returned to their original form. Then stemming was used, which reduces the number of words based on word type and sophistication. Let's say we have three terms in the dataset that are comparable to each other, such as "running," "ran," and "runner." These words will be combined to form the word "run." Different stemming methods exist, but Porter's excellent accuracy rate led us to use it. Because stop word removal eliminates frequent terms found in articles, prepositions, and conjunctions, we used it.

RESULTS

The project of Real and Fake news Classification has been successfully deployed. Various types of Machine Learning Algorithms were used to classify the news into the category of Real or Fake- whether they be supervised or unsupervised. As seen from the previous works, the most popular topic for the generation of fake news is politics. So, it might be possible that the models which have been deployed give accurate results for the news in politics but there might be a slight margin of error when comparing to news of other topics. From what has been observed in this project and the previously developed models, it has been induced that it is a very tedious task to generate a ML model that is capable of working on all categories of news. Here, the best work has been done to deploy the models accurately for the classification to take place. We observed that the Random Forests algorithm with an easy term frequency-inverse document frequency vector gives the simplest output compares to others. Our study examines various text properties which will be wont to distinguish fake and real content, and we trained a mixture of various machine learning algorithms using these properties.

Table 1: Observed Results

Model Deployed	Accuracy
Logistic Regression	0.81%
Random Forest	0.79%
Passive Aggressive	0.92%

CONCLUSION

The models which were used for the classification process have been deployed accurately as per the accuracy score. The results show that that the project has been implemented successfully and the prediction module also works correctly 97% of the time. There is a slight margin of error which an be rectified in the future works. If this project is taken up in the future, then those errors can be taken into account and can be rectified by connecting the developed models to the cloud for the classification and prediction of real and fake news. Other factors like the writers of the particular news articles, the score of how much fake news or rumors are given by a particular writer, popular topics etc. can be taken into account as well.

REFERENCES

Deshpande, G.C. Fake News Detection Using Deep Learning Techniques. 2019 (ICAIT) (pp. 411-415). IEEE.

Kasbe, A. Fake news detection. In 2018 IEEE International Students' Conference on (SCEECS) (pp. 1-5). IEEE.

Automated Fact Checking News Corpus. 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop (pp. 357-362) by Pathak A.

Lorent, S. (2019). Master thesis: Fake news detection using machine learning.

Schneider, J. M. From clickbait to fake news detection: an approach based on detecting the stance of headlines to articles.2017 EMNLP Workshop: Natural Language Processing meets Journalism (pp. 84-89).

Nykanen, M., Oussalah, M. Meta-terrorism: identifying linguistic patterns in public discourse after an attack. 2018 IEEE International Conference (ASONAM) (pp. 1079-1083). IEEE.

Sajid Ahmed, Development of Fake News Model, International Journal of Computer and Information Engineering, 2020.

Defining fake news by Tandoc, E.C. in Digital Journalism.

Study of hoax news detection in ICTS (page 73-78) by Pratiwi, IVR.

Fake news Detection with models by Vijayaraghavan, arXiv preprint arXiv: 2003.

Gilda, Evaluating ML Algorithms for Fake News Detection in 2017 SCOReD (page 110-115) IEEE.

Davuth N, Classification of malicious domain names using SVM, Journal of Security and Its Applications 7(51-58).

Banerjee S, Using supervises learning to classify on authentic online reviews in 9th International Conference on Ubiquitous Information Management and Communication (pp. 1-7).

Ganiz, M.C., Analysis of preprocessing methods on classification of Turkish texts in International Symposium on Innovations in Intelligent Systems and Applications (page 112-117).

Patwari A, TATHYA: A Multi- Classifier System for Detecting Check-Worthy Statements in Political Debates' in Conference on Information and Knowledge Management. doi:10.1145/3132847.3133150.