IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Fake News Detection using Machine Learning

[1]Shruti Sunil Ajankar, [2]Sejal Rajesh Raka, [3]Asra Fatema Zubair Quazi, [4]Ketaki Ashok Thakare,

[5]Shweta Sanjay Jadhao, [6]Dr. D. N. Chaudhari

Students, Department of Computer Science and Engineering [1,2,3,4,5]

HOD, Department of Computer Science and Engineering [6]

Jawaharlal Darda Institute of Engineering and Technology, Yavatmal, Maharashtra, India

Abstract: Due to the immense use of social media and online news media there has been a large surge of fake news in recent times. Spreading fake news has become much easier. Such spreading of fake news may have a severe effect. Hence it is extremely essential that certain measures should be taken in order to reduce or distinguish between real and fake news. Information overload and a general absence of comprehension about how the web functions by individuals have additionally added to an expansion in fake news or hoax stories. Social media sites can have a major influence in expanding the span of this kind of story. Fake news is a news created to intentionally misguide or mislead readers. Fake news is spread mainly for gaining political or financial incentives. They submit this well-crafted news stories and also recruit social bots or paid scammers to spread the news more rapidly and different approaches using text-based analysis for detecting fake news. Ever since the birth of social media and online news media, the spread of fake news has increased drastically. Social media sites such as Facebook, Google Plus etc are one of the biggest sources of spreading fake news. As the spread of such fake news can be intentional or unintentional but this affects society. Thus, an increasing number of fake news has to be controlled by using the computational tool which predicts such misleading information as if it is fake or real. Here we have focused on developing such computational tool to help classify news using machine learning algorithm. Which helps the model to be more trustworthy. Which describes the pre-processing, feature extraction, classification and prediction process in detail. We've used Logistic Regression to classify fake news. The pre-processing functions perform some operations like tokenizing, lemmatization and exploratory data analysis like response variable distribution and data quality check (i.e., null or missing values). Simple Count Vectorization, TF-IDF is used as feature extraction techniques. The logistic regression is used as a classifier for fake news detection with a probability of truth.

***Index Terms* - Fake news detection, Logistic regression, TF-IDF, count vectorization, NLP, feature selection**

I. INTRODUCTION

Data or information is the most valuable asset. The most important problem to be solved is to evaluate whether the data is relevant or irrelevant. Fake data has a huge impact on lot of people and organizations as well. Fake news is untrue information presented as news. It often has the aim of damaging the reputation of a person or entity or making money through advertising the revenue. It is extremely difficult to decide whether the news we come across is real or not. There are very few options to check the authenticity and all of them are sophisticated and not accessible to the average person. There is an acute need for a web-based fact-checking platform that harnesses the power of Machine Learning to provide us with that opportunity. These days fake news is creating different issues from sarcastic articles to a fabricated news and plan government propaganda in some outlets. Fake news and lack of trust in the media are growing problems with huge ramifications in our society. Obviously, a purposely misleading story is "fake news" but lately blathering social media's discourse is changing its definition. Some of them now use the term to dismiss the facts counter to their preferred viewpoints. As human beings, when we read a sentence or a paragraph, we can interpret the words with the whole document and understand the context. In this project, we teach a system how to read and understand the differences between real news and the fake news using concepts like natural language processing, NLP and machine learning and prediction classifier like the Logistic Regression which will predict the truthfulness or fake-news of an article.

II. AIM AND OBJECTIVE

Our main aim of the project is to make a machine learning model, with the help of which news can be classified as fake or real with help of machine learning classification algorithms and text feature extraction methods for classifying news.

To create a system that can use the data of past news reports and predict the chances of a news report being fake or not. It provides the user the ability to classify the news as fake or real. It provides the user the ability to classify the news as fake or real.

To achieve the goal of developing machine learning model to classify news as fake or real, we need perform following tasks in the same order as stated. Data Collection and Analysis, Preprocessing the data, Text feature extraction, Using different classification algorithms, Taking the best classification algorithm and feature extraction method, Classifying the news as fake or real, Deploying the model.

The objectives are, to develop a system that is capable of reading datasets, to implement an algorithm for automatic classification of text into positive and negative, to design the system in such a way that it can easily predict the false news as soon as the user enters the data, to process the system to obtain the better accuracy results.

III. METHODOLOGY

The methodology of fake news detection system includes, collecting the data from various news articles, followed by pre-processing that data and after that feature extraction is done, later the classifiers are fed with the training dataset to train them and lastly the content is classified i.e testing is done.

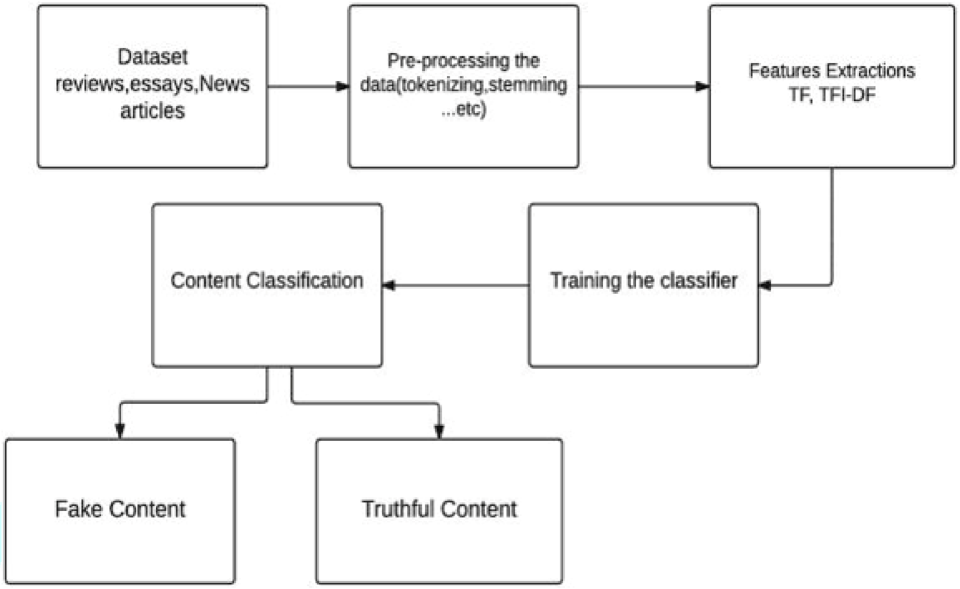

Dataset reviews,essays,News articles
Pre-processing the data(tokenizing,stemming ...etc)
Features Extractions TF, TFI-DF
Training the classifier
Content Classification
Fake Content
Truthful Content

Fig.1. Typical framework for fake news detection using machine learning techniques

A] Data set

The dataset is simple. It contains the titles of the news, the body text and a label field, which, if the news is authentic, shows REAL and if inauthentic, shows FAKE. There are 3 main segments of the methodology: The core Machine Learning model, the web interface, the common platform that brings the model and the interface together. Train and test datasets are the two key concepts of machine learning, where the training dataset is used to fit the model, and the test dataset is used to evaluate the model. The training data is the biggest (in -size) subset of the original dataset, which is used to train or fit the machine learning model. Firstly, the training data is fed to the ML algorithms, which lets them learn how to make predictions for the given task. Once we train the model with the training dataset, it's time to test the model with the test dataset. This dataset evaluates the performance of the model and ensures that the model can generalize well with the new or unseen dataset. The test dataset is another subset of original data, which is independent of the training dataset. In the project the dataset used is from Kaggle website where real news and fake news are in two separate datasets we combined both the datasets into one and trained with the machine learning classification algorithm to classify the news as fake or not. Here we will be ignoring attributes like the source of the news, whether it was reported online or in print, etc. and instead focus only the content matter being reported. We aim to use different machine learning algorithms and determine the best way to classify news[2].

B] Pre-processing the data

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first step while creating a machine learning model. Data preprocessing is required tasks for cleaning the data for a machine learning model which also increases the accuracy and efficiency of a machine learning model.
The various preprocessing steps that are involved are:

- Removing Stop words - Stopwords are the commonly used words and are removed from the text as they do not add any value to the analysis.

- Tokenization - Tokenization is essentially splitting a phrase, sentence, paragraph, or an entire text document into smaller units, such as individual words or terms. In Tokenization the text is differentiate into smaller units.

- Lemmatization - The Lemmatization is pre-defined dictionary that stores the context of words and checks the word in the dictionary.

C] Feature Extraction

After the pre-processing of data, feature extraction is done. In this module we have performed feature selection or the extraction methods from sci-kit learn python libraries. For feature selection, we have used methods like count Vectorization term frequency like tf-idf weighting. Text needs to be converted into numbers before it is used with a machine learning algorithm. For classification of documents, documents are taken as input and a class label is generated as output by the predictive algorithm. The documents need to be converted into fixed-length vectors of numbers for the algorithm to take them as input. The input for the machine learning algorithm are the words encoded as integers or floating point values.

- Count Vectorization - Vectorization is a methodology in NLP to map the words or phrases from vocabulary to a corresponding vector of real numbers which is used to find word predictions or the word similarities. The documents' corpora must first be converted into some numerical structure to make them more relatable for computers. Vectorization is a process of converting the text data into a machine-readable form. The words are represented as vectors. we cannot pass text directly to train our models in Natural Language Processing, thus we need to convert it into numbers, which machine can understand and can perform the required modelling on it. Count Vectorizer tokenizes (tokenization means breaking down a sentence or paragraph or any text into words) the text along with performing very basic pre-processing like removing the punctuation marks, converting all the words to lowercase, etc. The vocabulary of known words is formed which is also used for encoding unseen text later. An encoded vector is returned with a length of the entire vocabulary and an integer count for the number of times each word appeared in the document[2]. It counts the number of times a token shows up in the document and uses this value as its weight.

- TF-IDF - TF-IDF stands for "term frequency-inverse document frequency", meaning the weight assigned to each token not only depends on its frequency in a document but also how recurrent that term is in the entire corpora. This is a technique to quantify a word in documents, we generally compute a weight to each word which signifies the importance of the word in the document and corpus. This method is a widely used technique in Information Retrieval and Text Mining. TF is individual to each document and word, hence we can formulate TF as follows.

Term Frequency (TF) = (Number of times term t appears in a document) / (Number of terms in the document)

This measures the importance of document in whole set of corpus, this is very similar to TF. The only difference is that TF is frequency counter for a term t in document d, whereas DF is the count of occurrences of term t in the document set N.

df(t) = occurrence of t in documents

IDF is the inverse of the document frequency which measures the informativeness of term t. When we calculate IDF, it will be very low for the most occurring words such as stop words (because stop words such as "is" is present in almost all of the documents, and N/df will give a very low value to that word). This finally gives what we want, a relative weightage[3].

Tf-idf(t) = N / df

D] Training the Classifiers

The extracted features are then fed into different classifiers. Here we have used Logistic Regression from sci-kit learn. This computational tool uses the logistic regression classifiers so that user could get more accurate results.

- Logistic Regression : Logistic regression is a machine learning algorithm used for binary classification[2]. Logistic regression is an example of supervised learning. It is used to calculate or predict the probability of a binary event occurring. In a classification problem, the target variable (or output), y, can take only discrete values for given set of features (or inputs), X.

  $$LR(z) = \frac{1}{1+e^{z}}$$

  Logistic regression becomes a classification technique only when a decision threshold is brought into the picture. The setting of the threshold value is a very important aspect of Logistic regression and is dependent on the classification problem itself. As per feature selection used for the data set here the best threshold value for logistic Regression is 0.6. Here the 80% data was used for training and 20% data was used for testing on the logistic regression classifier it gives us mean score of 0.93 and best score of 0.94[2].

  Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression[1] (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 and 1, hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a logit, from logistic unit, hence the alternative names. Analogous models with a different sigmoid function instead of the logistic function can also be used, such as the probit model; the defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively scales the odds of the given outcome at a constant rate, with each independent variable having its own parameter; for a binary dependent variable this generalizes the odds ratio. The logistic regression model itself simply models probability of output in terms of input and does not perform statistical classification (it is not a classifier), though it can be used to make a classifier, for instance by choosing a cutoff value and classifying inputs with probability greater than the cutoff as one class, below the cutoff as the other; this is a common way to make a binary classifier[3].

E] Content Classification

Once the model has been trained the testing needs to be done. Testing simply implies the content classification i.e classifying the news into true or fake.

IV. Proposed Work

1] Front-End

User Module:

- User can use the system functionalities.
- User can enter a news and get predicted whether the entered news is true or fake by the system.

Prediction Module:

- Here, the system reads the input entered by the user.
- Performs comparison of the entered news with the news in the stored database.
- Predicts the entered input news is true or fake.
- Displays the predicted result as output

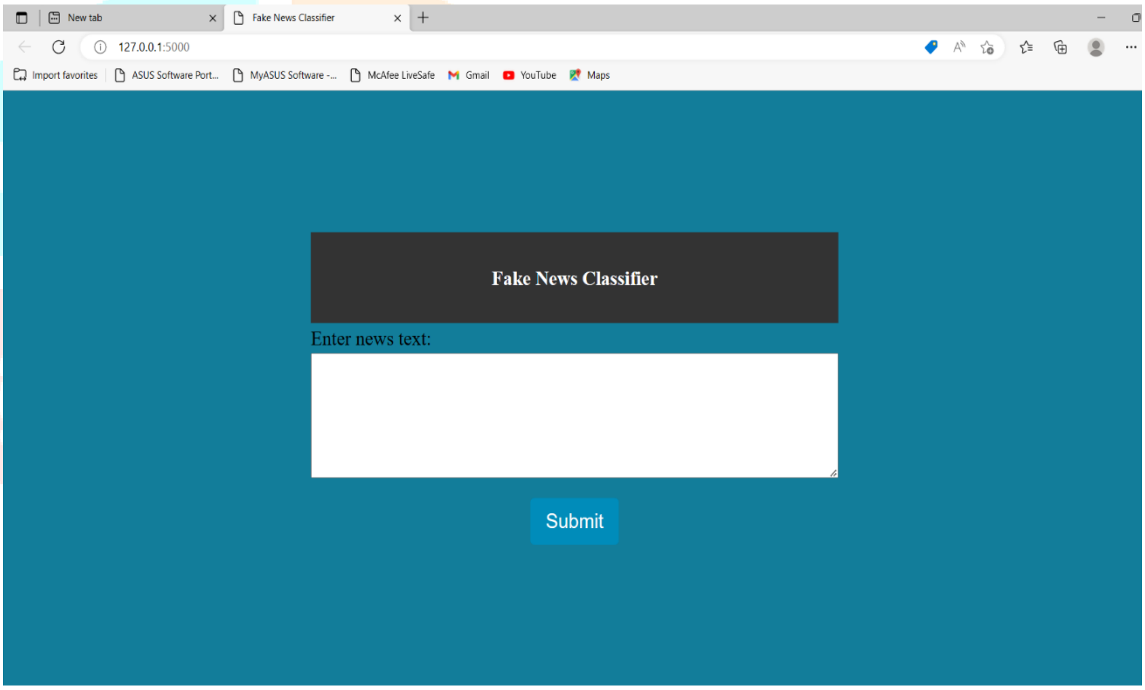


Fig.2. Design of front-end module

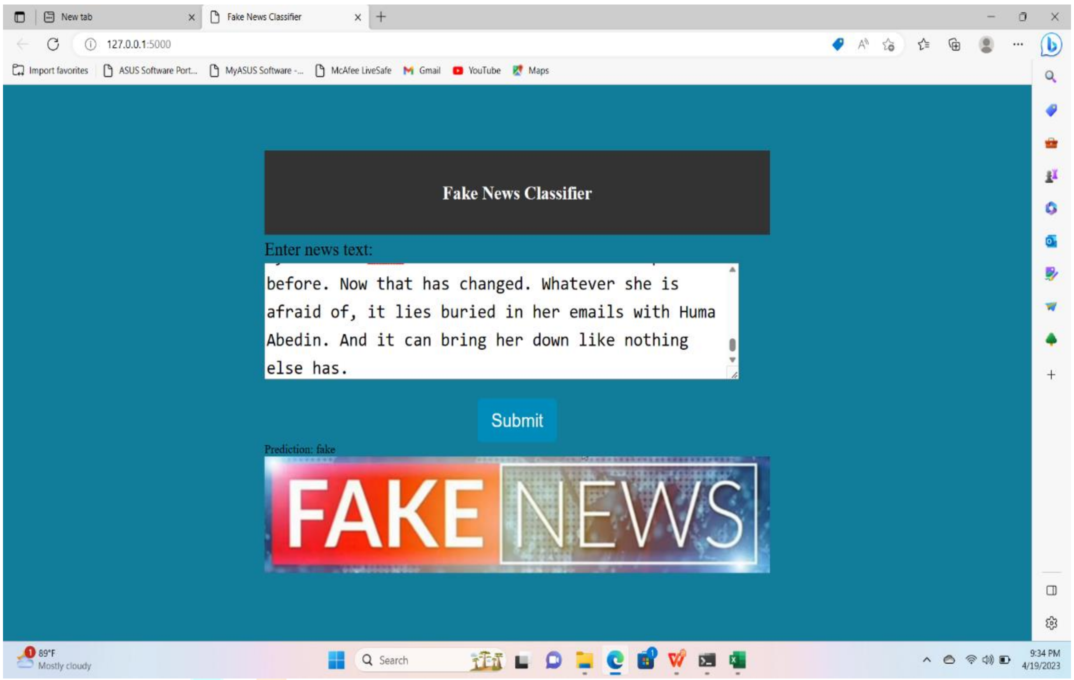
2] Back end

In our work, We have implemented some machine learning methods for detecting the news whether it is fake or real. For that, first we collect the data and after that raw texts of news requires some pre-processing. So the data needs to be pre-processed first. The performance of machine learning models depends on a great deal on feature design. So we have to extract a wide range of features and then by using methods we have to train the data for classifying the data. After that the data is classified into fake or real.

In our work, we used Logistic Regression. Logistic Regression 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).

Result :

To check whether the news is fake or real, we need to select the text of any one of the news from any dataset and paste it into the space provided in the design page. And then click on the submit button. After this, at the back-end the classifiers will be testing the news on the basis of the trained dataset and display the result whether its real or fake.



Fake News Classifier

Enter news text:

before. Now that has changed. Whatever she is afraid of, it lies buried in her emails with Huma Abedin. And it can bring her down like nothing else has.

Prediction: fake

Fig.3. Result

V. Conclusion

• Machine Learning uses a statistical technique to give the computer the ability to learn with data hence it is widely used in the detection of fake news. Methods used for taking parameters and for categorizing the type of news are also discussed. It was observed that the dataset is first preprocessed using preprocessing techniques such as stop word removal, tokenization and stemming. The techniques TF-IDF and probabilistic context-free grammar that are used to extract features are also identified.

• In this paper, we have used Logistic Regression classifier which will predict the truthfulness of user input news, here we have presented a prediction model with feature selection used as Count Vectorization, TF-IDF which helps the model to be more accurate. We have investigated the classifier model with feature extraction as Count Vectorization, TF-IDF. The proposed Logistic Regression model achieves the mean accuracy of 0.91.

• Firstly the data set is selected from the news articles or reports. But these data can't be directly processed further as it is not readable by the machine. So the data undergoes some NLP pre-processing techniques which makes the data relatable to the machine. After this feature extraction occurs. It's the process of transforming raw data into numerical features that can be processed while preserving the information in the original data set. Feature extraction increases the accuracy of the learned models by extracting features from the input data. After this, the classifiers are fed with the training data set. Once the classifiers are trained, they are tested by giving the testing data set.

References

[1] S. M. Metev and V. P. Veiko, Laser Assisted Microtechnology, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.

[2] Akshay Jain, Amey Kasbe, "Fake News Detection", The Institute of Electrical and Electronics Engineers, Published 2018

[3] Abhishek Singh, Aditya Ugale, Niraj Shah and Prof. Amruta Sankhe, "Fake News Detection using Logistic Regression and Multinomial Naïve Bayes", International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 08 Issue: 04 | Apr 2021

[4] Kelly Stahl, "Fake news detection in social media", California State University Stanislaus, 2018.

[5] Subhadra Gurav, Swati Sase, Supriya Shinde, Prachi Wabale, Sumit Hirve[3] , Survey on Automated System for Fake News Detection using NLP & Machine Learning Approach, International Research Journal of Engineering and Technology (IRJET), 2019

[6] Fake News Detection Using Naive Bayes Classifier by Mykhailo Granik, Volodymyr Mesyura. http://ieeexplore.ieee.org/document/8100379/

[7] Automatically Identifying Fake News in Popular Twitter Threads by Cody Buntain Available: http://ieeexplore.ieee.org/abstract/document/7100738/

[8] Aphiwongsophon, S., & Chongstitvatana, P. (2018, July),"Detecting Fake News with Machine Learning Method",In 2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTICON) (pp. 528-531). IEEE.

[9] Conroy, N. J., Rubin, V. L., & Chen, Y. (2015, November),"Automatic deception detection: Methods for finding fake news", In Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community (p. 82). American Society for Information Science.

[10] Gilda, S. (2017, December),"Evaluating machine learning algorithms for fake news detection",In 2017 IEEE 15th Student Conference on Research and Development (SCOReD) (pp. 110-115). IEEE.

[11] Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S., & de Alfaro, L. (2017). Some like it hoax: Automated fake news detection in social networks. arXiv preprint arXiv:1704.07506.

[12] Bourgonje, P., Schneider, J. M., & Rehm, G. (2017). From clickbait to fake news detection: an approach based on detecting the stance of headlines to articles. In Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism (pp. 84-89).

[13] Ahmed, H., Traore, I., & Saad, S. (2017, October).,"Detection of online fake news using n-gram analysis and machine learning techniques" In International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments (pp. 127-138). Springer, Cham.