FAKE NEWS DETECTION USING NLP

phase 2

Introduction

I. INTRODUCTION

Fake news is false or misleading information presented as news. The proposed study uses machine learning and natural language processing approaches to identify false news—specifically, false news items that come from unreliable sources.

Fake news and disinformation are ongoing problems that may be found all around us in biased software that amplifies just our viewpoints for a "better" and smoother user experience. Fake news and misinformation are becoming more of a problem as the internet and social media platforms become more main stream. A common goal of fake news is to harm someone or something's reputation or to profit through advertising. The propagation of these ideas may have been influenced by a variety of factors, but they all present humanity with the same underlying problem: a misunderstanding of what is real and what is false. This confusion could result in additional problems, such a medical emergency.

Satire websites or sensationalistic "click-bait" is where news site proprietors most frequently steal from fake. Satire websites frequently publish outlandish news parodies, and visitors to these websites are aware that they should not be taken seriously. Tabloids are what click-bait news articles resemble. Even though the actual tales are relatively mundane, their sensationalist and dramatic headlines entice people to click to learn more. These kinds of headlines draw readers to buzz feed and up worthy.

In this paper, we are focusing on the fake news detection in text media. Machine learning and deep learning techniques for fraud detection has been the subject of extensive study, most of which has concentrated on categorising online reviews and publicly accessible social media posts. Some of the drawbacks of the fake news are shift in public opinion, defamation, false perception and many more.

To overcome the drawbacks of the fake news, a model is created to distinguish the real news from the fake news. The proposed method uses the ISOT dataset. Web scraping is also done on 4 websites and the scraped data is further added into the data.

Edit with WPS Office

The data undergoes data pre-processing, feature extraction, dimensionality reductio n and finally the data is sent to the classification models i.e. Rocchio classification, B agging classifier, Gradient boosting classifier and Passive Aggressive Classifier to tr ain the model which is further used to detect the fake news.

The paper is organised as follows. The section covers the previous research on meth ods for identifying fake news. The proposed approach is described in the third sectio n. Section four describes the evaluation and analysis of the proposed methodology, and the final section provides the paper's conclusion.

II. LITERATURE SURVEY

There have been several previous works done to detect the fake news. Some of the p revious works are given below. A fair amount of research has been done different as pects of this project domain.

1. According to "Fake News Detection Using Machine Learning Ensemble Methods" paper [1] the goal is to develop a system or model that can use historical data to forecast if a news report is fake or not. The dataset used here is ISOT dataset. The model used in this method is Random Forest Classifier. A large number of decision trees are built during the training phase of the random forests or random decision forests ensemble learning approach, which is used for classification, regression, and other tasks. The class that the majority of the trees chose is the output of the random forest for classification problems. Accuracy is one factor to consider when evaluating categorization models. The accuracy of the proposed solution is 90.64.
2. The author T. Dinesh. et. al. [2] has introduced a training model that can be used to predict the class or value of the target variable by learning simple decision rules developed from training data. The approach of classifying the fake political news manually requires more knowledge of the domain. In this research, the problem of classifying fake political news articles using machine learning models is discussed. The mean accuracy and standard deviation for

Edit with WPS Office

the Decision Tree algorithm is 99.6990 and 0.10577. For Naive Bayes algorithm is 95.3870 and 0.00061. The accuracy of innovative fake news detection for political news detection using Decision Tree algorithms has better accuracy in comparison with Naive Bayes algorithms.

3. According to the author Kasra Majbouri. et. al. [3], the purpose is running K-Means clustering to see if the algorithm can successfully cluster the news into Real and Fake using just the words in the articles. The proposed method of choosing features and detecting fake news has four main steps. The first step is computing similarity between primary features in the fake news dataset. Then, features are clustered based on their similarities. Next, the final attributes of all clusters are selected to reduce the dataset dimensions. Finally, fake news is detected using the k-means approach. The accuracy of the K-means clustering algorithm in the detection of fake news is approximately 87%.
4. The goal of fake news identification, according to the author Monther Aldwahedi et al. [4], is to develop a method that users may use to identify and filter out websites that contain inaccurate and misleading information. A simple and carefully selected feature of the title and post to accurately identify fake posts. The experimental results showed a 99.4% accuracy using the logistic classifier.

III. FAKE NEWS DETECTION USING NATURAL LANGUAGE PROCESSING

Most text and documents contain many terms that are redundant for text classificati on, such as stop words, misspellings, slangs, and so on. Hence, data pre-processing has to be done before the data is sent to the classification models. After that, the dat aset's dimensionality is decreased in order to save time and storage space. When th e dimensions are reduced, it becomes easier to visualise. The data is then used to tr ain classification models, which can be used to predict whether or not the presented data is fraudulent.

Edit with WPS Office

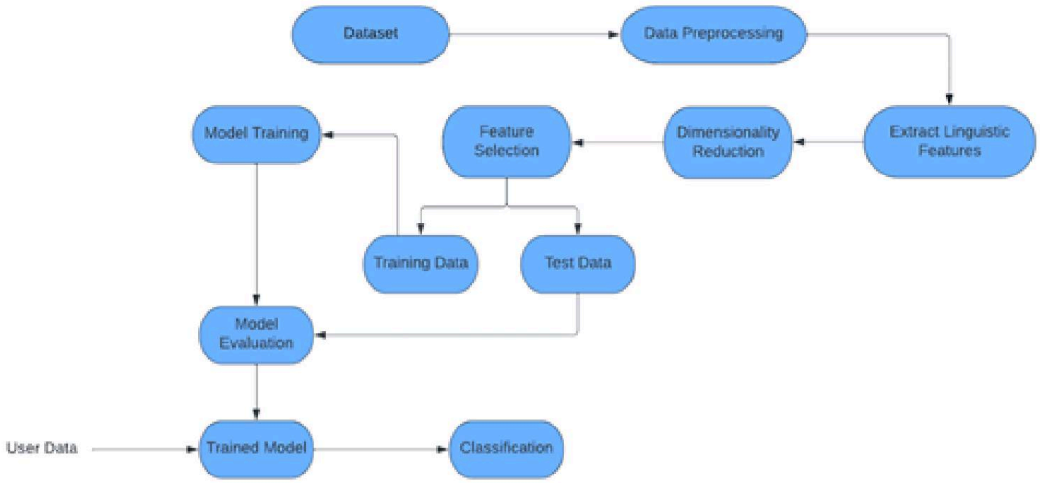

Fig 1. Flow Chart

A. *Description of Dataset*

The dataset used in this paper is ISOT dataset. In this dataset, there are two types of articles: fake news and real news. The dataset was gathered from real-world sources, and true articles were retrieved via crawling articles from Reuters.com. The fake news articles came from a variety of sources. Politifact and Wikipedia were used to gather the fake news items. Although the majority of the articles in the collection are about politics and foreign events, they cover a wide range of topics. The dataset consists of two CSV files. True.csv is the first file, and it contains almost 12,600 reuter.com stories. Fake.csv, the second file, comprises about 12,600 items obtained from various fake news sites.

???????B. *Web Scraping*

Large volumes of data can be automatically gathered from websites via web scraping. The majority of this data is unstructured in HTML format and is transformed into structured data in a database or spreadsheet so that it can be used in multiple applications.

Here, web scraping is done on 4 websites to get the present news. It is further added into the dataset to detect the present news as fake or not and also to increase the efficiency of detecting the fake news.

Edit with WPS Office

???????C. *Text Cleaning and Pre-processing*

To prepare the text data for the model building we perform text pre-processing. It is t he first step in the Natural Language Processing. Some of the pre-processing steps are:

1. *Tokenization:* Tokenization is the process of breaking down a stream of text into tokens, which can be words, phrases, symbols, or any other significant items. This step's major purpose is to extract individual words in a sentence. The tokenization is done on each text in the dataset.
2. *Stop Words:* Stop words are the commonly used words and are removed from the text as they do not add any value to the analysis. These phrases have little or no meaning. A list of terms that are regarded as stop words in the English language is included in the NLTK library. All the stop words from the texts are removed.
3. *Capitalization:* Sentences can have a combination of capital and lowercase letters. A written document is made up of multiple sentences. One of the method for reducing the issue space is to convert everything to lower case. This aligns all of the words in a document in the same location. Using the python function, all the words are converted to lower case.
4. *Stemming:* Stemming is the process of reducing the words to its root form by eliminating extraneous characters. PorterStemmer is one of the stemming model which is used here to convert the words into its root form.
5. *Lemmatization:* Text lemmatization is the process of removing a word's superfluous prefix or suffix and extracting the basic word. All the suffixes and prefixes from the words are removed to reduce space.

???????D. *Feature Extraction*

TF-IDF stands for Term Frequency-Inverse Document Frequency and it is a measure, used in the fields of information retrieval and machine learning that can quantify the i mportance or relevance of string representations in a document amongst a collectio n of documents. The Bag of Words technique, which is useful for text classification o r for assisting a machine read words in numbers, is outperformed by the TF-IDF tech

Edit with WPS Office

ni when it comes to understanding the meaning of sentences made out of words. Ea ch feature's TF-IDF weights are computed and recorded in a matrix with columns den oting features and rows denoting sentences.

???????E. *Dimensionality Reduction*

Dimensionality refers to how many input features, variables, or columns are present i n a given dataset, while dimensionality reduction refers to the process of reducing th ese features. In many circumstances, a dataset has a significant number of input fea tures, which complicates the process of predictive modelling.

In these situations, dimensionality reduction techniques must be used because it is h ighly challenging to visualise or forecast for the training dataset with a huge number of features. The curse of dimensionality, which is more commonly known, describes how adding more input features frequently makes a predictive modelling task more d ifficult to model. Since the TF-IDF matrix is a sparse matrix, Singular Value Decompo sition is used for dimensionality reduction.

1. *Singular Value Decomposition:* Singular Value Decomposition is one of several techniques that can be used to reduce the dimensionality, i.e., the number of columns, of a dataset. A matrix's Singular Value Decomposition is a factorization of that matrix into three other matrices. Finding the ideal set of variables that can most accurately predict the result is the aim of SVD. During data pre-processing prior to text mining operations, SVD is used to find the underlying meaning of terms in various documents.

Mathematics behind SVD,

The SVD of mxn matrix is given by the formula,

A = UWVT

where

U: mxn matrix of the orthonormal eigenvectors of AAT

V:transpose of mxn matrix containing the orthonormal eigenvectors of ATA

Edit with WPS Office

W: a nxn diagonal matrix of the singular values which are the square roots of the eig en values of ATA

The matrix from TF-IDF is given as input to the TruncatedSVD. The columns i.e. featu res denotes the dimensions whereas the rows in the matrix denotes the points in the space. The dimensions of the matrix are reduced using TruncatedSVD.

???????F. *Classification Techniques*

On the basis of training data, the Classification algorithm is a Supervised Learning te chnique that is used to categorise new observations. The classification algorithms u sed in this paper is,

1. *Rocchio Classification:* A type of Rocchio relevant feedback is Rocchio classification. The centroid of the class of relevant documents is the average of the relevant documents, which corresponds to the most important component of the Rocchio vector in relevance feedback. Rocchio classification, which uses centroids to define the boundaries, is used to compute good class boundaries. Rocchio classification calculates the centroid for each class. When a new text data is given, it calculates the distance from each of the centroid and assigns the data point to the nearest centroid.
2. *Bagging:* When the goal is to reduce the variance of a decision tree classifier, bagging is utilised. The goal is to construct different subsets of data from a training sample that was picked at random and replaced. Their decision trees are trained with each group of data. As a result, we have a collection of various models. The average of all the forecasts from various trees is used which is more robust than a single decision tree classifier.
3. *Gradient Boosting:* A method for creating a collection of forecasts is called boosting. In order to reduce training errors, boosting is an ensemble learning technique that combines a number of weak learners into a strong learner. A random sample of data is chosen, fitted with a model, and then trained successively in boosting; each model attempts to make up for the

Edit with WPS Office

shortcomings of the one before it. The weak rules from each classifier are joined during each iteration to create a single, powerful prediction rule. Gradient boosting is a type of machine learning boosting. It relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error. The key idea is to set the target outcomes for this next model in order to minimize the error. The target outcome for each case in the data depends on how much changing that case's prediction impacts the overall prediction error.

4. *Passive Aggressive Classifier*: For large-scale learning, passive-aggressive algorithms are commonly used. It is one of the few 'online-learning algorithms'. In contrast to batch learning, where the full training dataset is used at once, online machine learning algorithms take the input data in a sequential order and update the machine learning model step by step. This is quite helpful when there is a lot of data and training the entire dataset is computationally difficult because of the size of the data. Since, the web scraping is used in this method, it adds the data to the dataset, and the size of the dataset becomes large which makes the Passive Aggressive Classifier model to work efficiently.

IV. RESULT

To assess the effectiveness of the suggested technique on diverse datasets, we ran a number of simulations and experiments using different classifiers. The dataset was divided into training and test set. 80 percent of the dataset is regarded as the training data, and the remaining 20 percent is taken as the test data. The performance of several approaches was compared using the classification's accuracy as the criterion.

Models	Accuracy	Precision	Recall	F1-score
Rocchio Classification	88.0	0.93	0.84	0.89

Edit with WPS Office

	9%			
Gradient Boosting	86.75%	0.92	0.83	0.87
Bagging Classifier	94.67%	0.95	0.94	0.95
Passive Aggressive Classifier	93.34%	0.95	0.93	0.94

TABLE I. ANALYSIS OF CLASSIFICATION MODELS

Table I and Fig 2 depicts the comparison of the classification report of Rocchio classification, Gradient Boosting Classification, Bagging Classification and Passive Aggressive classification model. The values in the tables specify the accuracy, precision, recall and f1-score of the classification models in the proposed method.

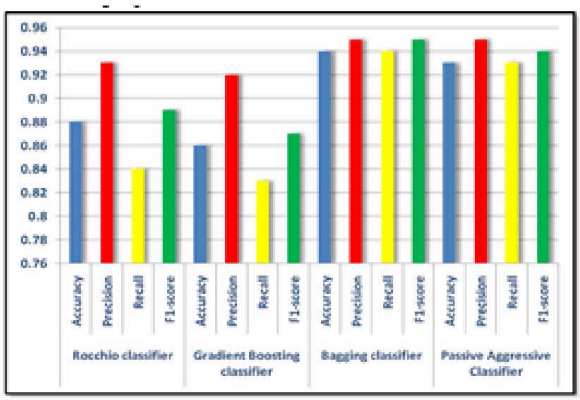

Fig 2. Bar graph for comparision of classification models

As a result of comparison, the BaggingClassifier model has higher accuracy of 94.67% than other classification models like Rocchio Classification model, gradient boosting model and passive aggressive classifier model.

V. ACKNOWLEDGMENT

We, the authors would like to express our profound appreciation to our guide Prof. Ashritha R Murthy for her invaluable mentoring, as well as for her helpful suggestions and encouraging words, which motivated us to work even harder. Due to her foretho

Edit with WPS Office

ught, appreciation of work involved and continuous imparting of useful tips, this research has been successfully completed.

Conclusion

The manual classification of false political news requires for a deeper understanding of the field. The problem of predicting and categorizing data in the fake news detection issue needs to be confirmed using training data. Reducing the amount of these features could increase the accuracy of the fake news detection algorithm because the majority of fake news datasets have many attributes, many of which are redundant and useless. As a result, this research suggests a technique for dimensionality reduction-based fake news detection. The dimension-reduced dataset is constructed using the final set of features. After specifying the final set of features, the next step involves utilizing classification models like Rocchio Classification, Bagging, Gradient Boosting Classifier, and Passive Aggressive Classifier to forecast the fake data. We assessed the performance of the suggested method on the dataset after it had been implemented. With the classification methods, we achieved the highest accuracy with the 94.67 percent accuracy of the TF-IDF feature extraction and the bagging classifier technique

Edit with WPS Office