International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 10 Issue 6 ǁ 2022 ǁ PP. 1349-1352

Fraud News Identification Using NLP

Dr. Ramya p[1], Yeswanth Sai Bezawada[1†‡], Ashraf AkramN[1§], BalajiD[1¶], * Bharath Doosan J.R[1‖]

[1] *Computer Science, Mahendra Engineering College, Salem-Tiruchengode Highway,,,Mahendhirapuri, Mallasamudram West, Namakkal, 637503, Tamil Nadu, India.*
Tel.: 04288288522

Abstract
One of the most serious issue in the presenttime is rapid spread of fake news all around world which has the power to shape perceptions and affect decisions. This happens due tohuge volume of datawhichis nearly 26,00,000 terabytes has been generated each and every day because of the usage of World Wide Web and the massive adoption of social media plat-forms (such as Instagram and Twitter) have tunneled the way for information circulation that has never been witnessed in the human history before and spread of bogus news on social media and the Internet is alarming.
With the current rate of usage in social media, the users are generating more information than ever beforewhich has made difficultto classify the accuracy of the news . People are being deceived to such an extent that it needsto be stopped.
Even, Classification of a text article as misinformation or disinformation is a challengingtask. Even thoughusing numerous keywords from previously detected fake contents, the authenticity of a news piece must be assessed from multiple perspectives. So we propose to classify news items using a machine learning ensemble approach combined with NLP. We'll use a variety of textual features and techniques to help identify between fake and authentic information.
By using those properties, we train our model with various machine learning algorithms and evaluates their performance on realworld data set. The Machine Learning algorithms we used in our research are Logistic Regression, Decision tree, Random Forest,XGBoost which are good for predicting the categorical dependent variable using a given set of independent variables.

Date of Submission: 06-06-2022 Date of acceptance: 21-06-2022

I. INTRODUCTION

Over the past decade, there has been a fast surge in the propagation of fake news, most notably during election seasonsandmorerecentlyduringthe pandemic. The widespreaddistribution of inaccurate information on the internet hasresulted in a slew of issues ranging from sports to health to science.

The financial market is one of those areas which has been affected by fake news, where a rumor can have severe implications and evenbring the market to a halt. Our ability to makedecisions is largelydetermined by the informationwe absorb; our worldview is influenced bythe information we consume.

False information can reach a much largeraudience than it appears. There is mounting evidence that people have reacted irrationallyto news that afterwards turned out to be false.Even in 2016, When India's Prime Minister, Mr. Narendra Modi, said that the majority of people's cash had become useless, and that all old currency had to be deposited in bankswithin a month. As a result, a succession of fake news stories were published, primarilyforclick bait and political advantage. The news of new paper bills with a GPS tracker or anincrease in the daily limit of the amount thatmay be deposited in banks spread like wildfire.Now, this may not seem like a big deal, but the impact of such publications was so great that the Ministry of Finance hadto issue official statements assuring thepublic citizens that the information they were reading was untrue. This is only one example of how incor-rect information may propagate influence a farlarger audience than you might think.

Recent example is the transmission of the corona virus, which saw bogus stories regard-ing the virus's origin, nature, and behavior spread over the Internet. As more individuals learned about the bogus content online, the situation became worse.

With this, Social media has plagued by the development of fraudulent accounts and fake news. Furthermore, fraudulent users use theiridentities for a variety of purposes, including spreading rumors that harm a specific eco-nomic sector or perhaps the entire society. In order to classify a text article we gather the data into a dataset , analyze the data in it and use classification algorithm to identify thenews is fake or legit.

1.1 LITERATURE SURVEY

While some existing tools, such as BS Detec-tor and Politifact, can help users identify falsenews to some extent,they require human par-ticipation and in the case of BS Detector, the domain is limited because it does not tell theuser how false the story is.

gives a data mining viewpoint on detect-ing false news on social media, which incorp rates fake news categorization based on psy- chology and social theories. Naive Realism andConfirmation Bias are two significant elements that contribute towidespread user acceptanceof bogus news in this article .Further, it pro-poses a two-phase general data mining frame- work which includes 1) Feature Extraction and 2) Model Construction and discusses the datasets and evaluation metricsfor the fake news detection research.

They propose in [2] a method to detect online deceptive testsbyusing a logisticregression classifier that is based on POS tags extracted from a corpus of deceptive and truthful texts and achieves an accuracy of 72 per-cent, which could be improved further by per-forming cross-corpus analysis of classificationmodels and reducing the size of the input feature vector.

In [3] they present an SVM-based algorithmthat uses satirical cues todetect deceptive newsandincludes five predictive features: absurdity,humour, grammar, negative affect, and punctuation. The paper uses 87 percent accuracy totranslatetheories of comedy, irony, and satireinto apredictive model for satireidentification.

They used linguistic cues and network anal-ysis methodologies in [4] to create arudimen-tary fake news detector thathas high accuracyin terms of classification tests. They propose ahybrid system that includes elements such as multi- layer linguistic processing and network behavior.

The purpose of this paper is to propose a newmodel for fake news detection which is using the dataset gathered fromKaggle , along with multiple classification algorithms like Decisiontree, Logistic Regression , XGBoost and Ran- domForest . By using Random Forest ,there isan advantage of handling binary features and moreover, they do not expect linear features.

1.2 PROPOSED APPROACH

The goal of this project is to accurately iden-tify the legitimacy of a news article's contents.We created a process to retrieve the results for this purpose. First, we choose a dataset which has already collected data from World wide web which includes details about the arti-cles unique id, title, author of the article, text and the reliable label. We used this datasetto train the model and make predictions usingthe Machine Learning Algorithm. After getting the dataset ,we will be performing the data visualization process in-order to look into the dataset . After visualization, the data prepro- cessing process Imputation will be performedto fill the missing values in the dataset. As,a ML model cannot handle null values while the time of training. Following it, stemming process is performed . The process of stem- ming is the reduction of a word to its stem or root form. For example. "actor", "actress" and "acting". They can all be summed up withthe word "act". After all, they all convey the same meaning.This reduces

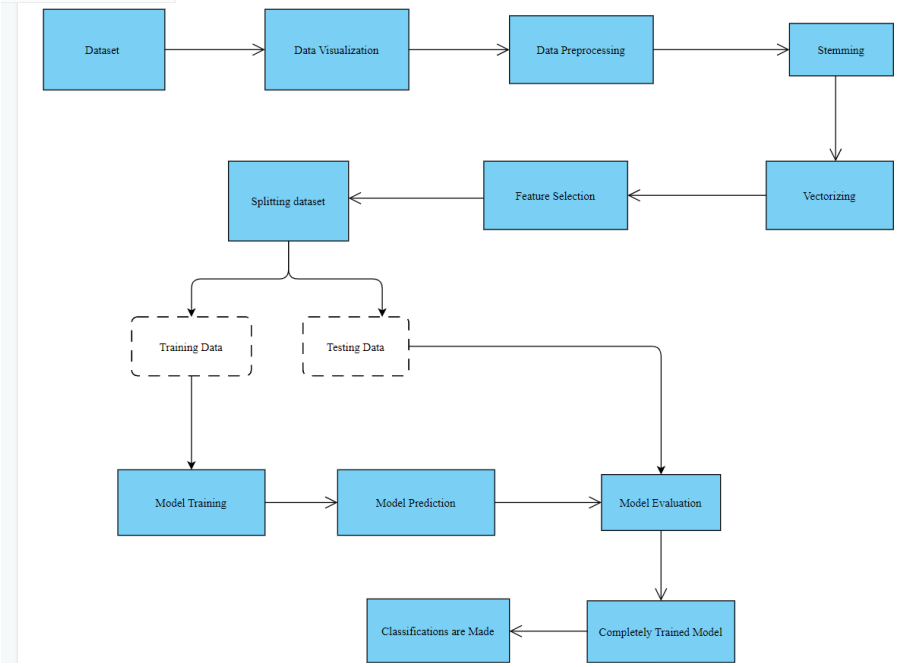


FIGURE 1
Architecture of Proposed System

complexity while preserving the essence of meaning conveyed by these three words. By using this we could avoid over training the model with large number of same meaning words and fast and crude operation carried out by applying very simple style rules of search and replace. Followed by Word Embeddings, also known as Word Vectorization, is an NLP technique used as a traditional method of turning raw data (text) into vectors of real numbers, which is the format that ML models allow, which can then be used to derive word predictions and semantics. As,using Encoding technique instead of vectorization creates more problem as encoding the entire text in the dataset , name of the author will creates confusion. So, we are preferring vectorization to make the text data be easily understood by the model. We use train test split method in order to divide the entire data into test and train data. As we use training dataset to feed the model and make it learn about the dataset with the help of a ML Algorithm. Then we test the accuracy of the model using the test dataset, where predictions are made. Based on the accuracy score the model is evaluated and the classifications are made about the given data

II. EXPERIMENTAL RESULTS & DISCUSSION

The Project results are evaluated based on two factors 1) Accuracy score 2) F1 score, the two evaluation techniques are used to deter- mine the efficiency of the model and Machine Learning Algorithm that we are using.

2.1 Dataset:

The dataset which has been used int his project was collected from one of the 'Kaggle Competition. It holds the details about the Article Id, Article Title, Article Content and Author name which tells that the model has only 4 input features. These dataset has contained information which has been collected from World wide web which are already classified(Determined whether realistic or Unrealistic) .

2.2 Result:

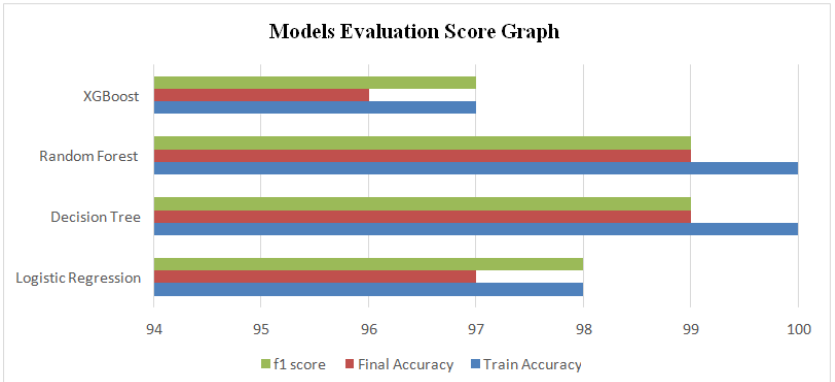
We will be splitting the dataset into train and test . After training the model using train data along with the ML Algorithm, we will check the score of the model by predicting the out- come of the test dataset. Here in this project, we have used 4 different classification algo- rithms for our model. At first we have trained our model with logistic regression, which pro-vides the result based on probability of a dat- apoint belonging to the class so, using that a datapoint is classified. After training the model with Logistic Regression, the training accuracy score is 98%, test accuracy is 97 % and the F1 score is 98%. While with Decision Tree the training, test accuracy and F1 scores are 99% .For Random Forest similarly like Decision Tree all the three evaluation scores are 99% . When the model is trained with XGBoost the training accuracy is 97%, test accuracy is 96% and the f1 score is 97%.

The proposed approach solves fake news problems by using the author's name and the title of the article. It detects fake users by ignoring the news that they provide (detect fake content of news), but if the user is not fake, it uses the title to classify the news and uses similarity measures and machine learn- ing algorithms to increase the credibility of the news.

Table -1: Algorithms Accuracy

Algorithm	Accuracy
Logistic Regression	97%
Decision Tree	99%
Random Forest	99%
XGBoost	96%

Figure 2



2.3 Definitions:

2.3.1 Logistic Regression:

A categorical dependent variable's output is predicted using logistic regression. As a result, the final outcome canbe a discrete or categorical value. It can be Yes or No, 0 or 1, true or false, and so on, but instead of giving exact values like 0 and 1, it delivers probabilistic values that are somewhere between 0 and 1.

Unlike the Linear Regression, instated of fitting to a linear line , we will be having 'S' Shaped graph with two values (0 &1)

$$\log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

2.3.2 Random Forest:

Random Forest is a collection of multiple number of Decision Trees which takes the average for ach subset of given dataset and makes predictions based on it.

If the Forest have More Number of Tress, then it will have better the accuracy.

2.3.3 Decision Tree:

Decision Tree is a flowchart like Tree structure which is used to estimate the target value, decision trees learn how to break the information into smaller subsets. In Decision tree the results are represented as a "leaf" (node), while the alternative results are represented as "branches" (edges) and the decisions are made using decision node which has branches.

2.3.4 XGBoost:

XGBoost is also known as Extreme Gradient Boosting which is Used to implement Gradient Boosted Decision Tree. Here, Decision Tress are created in sequential order and weights are assigned to input variable which helps thedecision tree to make predictions. Weights plays a major role in xbgoost.

III. Conclusion:

In this project , the author's identity and thetitle of the article are used to tackle fake news issues. It detects false users by disregarding the news they offer (identify fake news con- tent), but if the user is genuine, it utilizes the title to categories the news and employs simi-laritymetricsandmachinelearningtechniquesto boost the news' trustworthiness. The best accuracy score of test data is recorded by both Decision Tree and Random Forest which is of 99%. We developed this project to avoid the growing problem of fake news which onlycomplicates matters by attempting to sway people's opinions and attitudes against the usage of digital technologies.

When a person is duped by fake news, one of two things can happen. People begin to believe that their assumptionsabout a given topic are correct. Another issue is that, even if there is a news story accessible that contradicts an allegedly false one, people trust the words that simply reinforce their beliefs with- out considering the facts.

References

[1]. T CLu, T Yu, and S HChen. (eds) DecisionEconomics: In the Tradition of Herbert A.Simon's Heritage. DCAI 2017. Advances in Intelligent Systems and Computing, 618, 2018.

[2]. L Ball and J Elworthy, 2014. https://doi. org/10.1057/jma.2014.15

[3]. Victoria & Rubin, Niall & Con- roy, Chen, and Sarah Cornwell, FakeNewsor Truth? Using Satirical Cues to Detect Potentially Misleading News, 2016.

[4]. Niall, Conroy, Victoria, Rubin, and YiminChen. Automatic Deception: methods for finding Fake News.USA,2015