**Fake News Detection**

Advanced Database Topic (COMP8157)

Group 3: Technocrats

Avi Mistry1, Dhara Kishorbhai Hirpara2, Urvil Patel3, Vraj Soni4

Master of Applied Computing, University of Windsor

UID: 1110070746, 2110071938, 3110030912, 4110060926

Email: 1[mistry43@uwindsor.ca](mailto:mistry43@uwindsor.ca), 2[hirpara3@uwindsor.ca](mailto:hirpara3@uwindsor.ca), 3[patel8g1@uwindsor.ca](mailto:patel8g1@uwindsor.ca), 4[soni84@uwindsor.ca](mailto:soni84@uwindsor.ca)

1. **ABSTRACT**

Because of the massive distribution of information via social media platforms, individuals and society can distinguish the authenticity of any news. Fake news is not a new concept, but it has become a pervasive problem in recent times. Fake news can have an influence, from just being irritating to manipulating and misleading entire communities or even nations. There are several methods for detecting fake news. Through a comprehensive literature study, we identify many techniques currently available to identify false news and how these approaches are helpful in different circumstances. This proposed solution seeks to give multiple classifier models and strategies that may be utilised to find a better way to detect fake news.

**KEYWORDS**

Fake news, Ensemble model Natural language processing, Machine learning, Classification Algorithm

1. **INTRODUCTION**

The main aim is to identify whether a given text-based news is fake or not. The proposed approach is discussed in detail in the next section, followed by the formal problem definition. News articles can be represented as N = {n1, n2,n3,...,n𝑚}. The credibility label is denoted for each news article by n𝑖 ∈ N. The credibility label is from set the set Y where Y = {True, Mostly True, Half True, Mostly False, False}.

* 1. **MOTIVATION**

Although fake news is created and spread intentionally, it is difficult to understand the intention of the news creator due to the lack of information available through the public dataset. In this proposal, we propose to detect the veracity of news using the Natural Language Processing approach by considering it efficient for all the domains. Moreover, providing category-wise detection by labeling the class as partly true or false. Thus, to make it more efficient by providing statements along with labels in multi-domain.

* 1. **MOTIVATING EXAMPLE**

Opinion Spam Detection, Sentiment Analysis, and Fact-checking are related domains that facilitate the process of false news detection.

**Opinion Spam Detection**: Opinion spam means influencing the mind of users by altering the commercial, political, or social reviews. The aim is to mislead the buyers to purchase or avoid certain items or manipulate the citizens about a particular political party [1][.](#_bookmark7)

**Sentiment Analysis**: It is the computational study of people’s opinions, sentiments, emotions, and attitudes [2]. The writer’s intentions are mostly legitimate, but they are misinterpreted, and misleading information is conveyed. Sometimes, a writer might spread fake information in an attempt to impress the audience. As a result, emotion-based news verification is possible.

1. **RELATED WORK**

A lot of work has been done to detect the truthfulness of the news in the past. Most of the approaches used towards fake news detection are utilizing natural language processing and machine learning. Majorly, fake news detection has resulted as a binary classification but that would not suffice the truthfulness of the news. Multi-classification of news becomes a necessity to understand the correctness of news, for instance true, majorly true, or false, false, partially true or false, etc. Majorly, four approaches are proposed in the literature: (1) knowledge-based approach, (2) language-based approach, (3) machine learning approach, and (4) hybrid approach. The knowledge-based approach is seen more as compared to the other two. Language-based approaches make use of linguistics to determine between fake and real news. The propagation-based approach makes use of the network to determine the veracity of the news sources. However, language-based is recommended more because of the real-time feedback i.e they are not restricted to being applied only to posterior and the other thing is that it was scalable.

Chart

Description automatically generated

Figure 1: Current accuracy using different approach

1. **PROPOSED MODEL**

In the proposed model, we are expanding the current work by introducing an ensemble technique to classify if a given news article is fake or not from multiple domains. Using a combination of various ensemble methods that are not thoroughly explored in the existing literature, we train a variety of machine learning algorithms. These techniques facilitate the training of different machine learning algorithms in an effective and efficient manner. We conducted extensive experiments on the LIAR dataset. Currently, we have included Logistic Regression, Naive Bayes, Support Vector Machine (SVM), Random Forest, and Stochastic Gradient Descent as base classifiers and finally build the ensemble model using the best candidate classifiers. The proposed model architecture is shown in the figure 2.

**4.1 Feature Extraction**

Our proposed method of analysis analyses the features extracted from the news and assigns them a credibility score. In our model, we approach this problem in two ways: first, by extracting lexicographic features from the news, and second, by calculating a credibility score for the news article.

Graphical user interface

Description automatically generated

Figure 2: Proposed Model Architecture

As a result, a vector matrix is formed by linking the highly deceptive word list to each source. For feature selection, we have used methods like bag-of words and n-grams and then term frequency like TF-IDF (Term Frequency-Inverse Document Frequency) weighting. We have also used word2vec and POS tagging to extract the features. Below are some details on each of the features extraction models.

**Bag-of-words**: It is a fundamental way to represent text data and extract features from the text. In this, we tokenize words for each observation and find their frequency [9].

**N-grams**: An n-gram is any contiguous sequence of n tokens or words. A bag of n-grams can be more informative than a bag of words because they capture more context around each word [10].

**TF-IDF**: It is a technique for information retrieval in which its value increases if that token occurs frequently in the document and decreases if it occurs frequently in the corpus, thus giving an accurate metric value for that token. It gives a word frequency score that tries to highlight more interesting words. In the j document, for term i:

*W i,j = tf i,j \* log ( N / df i )* where,

*tf i,j* = number of occurrence of i in j

*df i* = number of documents containing i

*N* = total number of documents [2]

**Wordvec**: In this model, it transforms words into numeric vectors called embedding, where semantically similar words are close distance-wise in the embedded space [3].

**POS-Tagging**: Part-of-Speech tagging is defined as the process of assigning particular parts of speech corresponding to that word based on its context and its meaning. It can be used for removing syntactic ambiguity or for word sense disambiguation to get the sense of the authenticity of the news [1].

There are linguistic features that can be used to identify misinformation as how they are written is to obscure the real information by highlighting or using certain non-descriptive and inaccurate words to attract readers to click baits or fake news stories [5]. Finding a credibility score would be aided by analysing the bag of words in articles. This analysis is done on textual information from the news in our experiment. The linguistic features give you a list of words that are commonly used in spam communications. The majority of spam communications are fake news. We create a typical vector matrix of these phrases after analysing and extracting them. Statements with a high frequency of ambiguous and deceptive statements are given a poor credibility rating.

**4.2 CLASSIFICATION MODEL**

We have used Naive Bayes, Random Forest, Logistic Regression, Stochastic Gradient, SVM classifiers to analyze news segments. Accuracy, precision, recall, and F-score were the main metrics on which the performance of different algorithms was evaluated.

***4.2.1 Random Forest Classifier***

*Random Forest is built on the concept of building many decision tree algorithms, after which the decision trees get a separate result. The results, which are predicted by a large number of decision trees, are taken up by the random forest. To ensure a variation of the decision trees, the random forest randomly selects a subcategory of properties from each group [7]. The applicability of random forest is best when used on uncorrelated decision trees. If applied to similar trees, the overall result will be more or less similar to a single decision tree. Uncorrelated decision trees can be obtained by bootstrapping and featuring randomness. For the classification problem, we have used the Gini index as a cost function to estimate a split in the dataset. The Gini index is calculated by subtracting the sum of the squared probabilities of each class from one. The mathematical formula to calculate the Gini index G ind:*

*G ind =* 1- 𝜎(𝑝i)2

***2.2.2 Logistic Regression***

*It 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, success, etc) or 0 (no, failure, etc.). The hypothesis function can be defined mathematically as follows:*

h𝜃 . = 1/1+e−(𝛽0 + 𝛽1𝑥)

*A sigmoid function transforms the output into a probability. Logistic regression produces a logistic curve that is restricted to values that are between 0 and 1 by utilizing a sigmoid function [4].*

***2.2.3 Support Vector Machine***

*SVM Support Vector Machine is a machine learning-based supervised algorithm that is primarily used for classification problems. The goal of the SVM algorithm is to find the best line in 2D or the best hyper-plane in an N dimensional space that classifies the points. To distinguish between the two classes of points, such a hyper-plane is chosen to maximize the margin between the classes. Data points can be classified into different classes based on their position on either side of the plane. The hyper-plane is defined by:*

*w x+ b = 0*

*where b is the bias and w is weight vector.*

L(w)=𝜎𝑚𝑎𝑥 (0, 1 − *𝑦𝑖 [w𝑡 x𝑖 + |b|*]) + 𝜆||𝑤||2

*The first term is the Loss function that is used to calculate all the errors due to data points being in the vicinity of the classification boundary than margin. The second term is known as the regularization function used to avoid over-fitting.*

***2.2.4 Stochastic Gradient Descent***

*Stochastic gradient descent uses an iterative method to optimize an objective function through appropriate smoothness properties such as differentiable or subdifferentiable. The method consumes randomly shuffled or selected samples to gauge the gradients. Therefore, stochastic gradient descent “can be regarded as a stochastic approximation of gradient descent optimization” [8]. Gradient descent is applied iteratively to find the parameter values of a function that will minimize the function value with the maximum quantity. Therefore, the objective is to determine optimal parameter values in order to obtain the minimum value of the cost function.*

***2.2.5 Naive Bayes***

***------------------------------------------------------------------------------***

**Algorithm 1** pseudo code for fake news detection using ensemble approach

***------------------------------------------------------------------------------***

**procedure**

Split the dataset;

Construct the base models using Naive Bayes, Logistic Regression, Random Forest, SVM, Stochastic Gradient;

Train the blending ensemble;

Repeat

Fit on the training set.

Make prediction on holdout set.

Store the predictions as input for blending until the end of

base models.

Build a 2D array using the stored predictions;

Create the ensemble model;

Fit ensemble model on the predictions from base models;

Make predictions with the ensemble model;

***------------------------------------------------------------------------------***

*It is one of the most popular probabilistic models used for classification. It is based on the Bayes algorithm for calculating unconditional probability. This algorithm works on Bayes theory under the assumption that its free from predictors and is used in multiple machine learning problems [6]. Mathematically, it is represented as:*

*P(c|x) = P(x|c) P(c)/ P(x)*

*We are finding the probability of event c if event x has occurred. We make 2 assumptions while using the Naive Bayes Algorithm that all the features extracted are independent and they have an equal effect. We can extend the result for n events.*

*P(c|x) = P(x1|c)\* P(x2|c)\*....\*P(x2|c)\*P(c)*

*where, P(c|x) is posterior Probability*

*P(x|c) is the Likelihood*

*P(x) is the Class Prior Probability.*

*P(c) is the Predictor Prior Probability.*

1. **RESULTS**

Firstly, we are dividing the whole dataset into training and testing dataset. Below are the top five data entries from the training dataset and they are labelled based on their truthiness. After data preprocessing, we will be applying different classifiers.

Table

Description automatically generated

Figure 3: Top 5 data entries

Figure 5 displays the amount of data entries which are true and false.

Chart, bar chart

Description automatically generated

Figure 4: Data classification

Below mentioned classification reports of different techniques displayed their precision and accuracy value.

Table

Description automatically generated

Figure 5: Naïve-Bayes classification report

Table

Description automatically generated

Figure 6: Logistic Regression classification report

Table

Description automatically generated

Figure 7: SVM classification report

Table

Description automatically generated

Figure 8: SGD classifier classification report

Table

Description automatically generated

Figure 9: Random Forest classification report

From the above generated reports, we will be generating confusion matrix of each output and after comparing their values we are selecting Random Forest and Logistic Regression classification techniques.

Chart

Description automatically generated

Figure 10: Random Forest Confusion Matrix

Chart

Description automatically generated

Figure 11: Logistic Regression Confusion Matrix

Below mentioned are their cross-validation score graphs.

Chart, line chart

Description automatically generated

Figure 12: Logistic Regression Training data accuracy graph

Chart, line chart

Description automatically generated

Figure 13: Random Forest Training data accuracy graph

1. **LIMITATIONS**

Dataset – Before they can provide usable results, many machine learning algorithms require a vast amount of data. A neural network is a good illustration of this. Neural networks are data-hungry machines that require a lot of data to train. To get viable findings, the greater the design, the more data is required. Reusing data is a bad idea, and data augmentation might be helpful in some cases, but having more data is always the best option. Also, the lack of some good feature could be the reason behind the poor performance of the algorithm. Not have truth dataset can also limit the performance.

Hardware – Hardware issue can also be one of the major drawbacks for poor performance due as ML algorithms are hardware hungry and consume most of the resources for decent performance. And for the proper performance as mentioned above larger dataset are required which in turn requires higher computer resources.

1. **CONCLUSION AND FUTURE WORK**

During the process we discovered to develop an NLP based application using to classify the truthiness of the new based on dataset provided. We used five different algorithms and compared their accuracy and selected Logistic Regression and Random Forest as they provide comparatively higher level of accuracy. Specifically, we learned to develop a testing model, performing data analysis and generating a classification model. For future work, this could be a great initiative to also fight the deepfakes. As it can help one create the evidence of scenes that never took place in the first place. Additionally, it can help detecting some interesting patterns to support the growth in future technologies that will help consumers receiving filtered information.

**ACKNOWLEDGEMENT**

We’re privileged that Dr. Shafaq Khan gave us a chance to study this topic and taught us various related topics to this. Without her guidance and proper knowledge about research, it would be more difficult to complete the study. The project code and final results are accessible on [GitH ub](https://github.com/dharace/Fake-News-Detection).

**REFERENCES**

|  |  |
| --- | --- |
| [1] | M. Wendling, "bbcnews," [Online]. Available: https://www.bbc.com/news/blogs-trending-42724320. |
| [2] | A. S. S. W. J. T. a. H. L. K. Shu. [Online]. Available: https://dl.acm.org/doi/10.1145/3137597.3137600. |
| [3] | S. P. a. P. Atrey, "Media-Rich Fake News Detection: A Survey," [Online]. Available: https://ieeexplore.ieee.org/document/8397049. |
| [4] | Q. Su, "Information quality: Linguistic cues and automatic judgements," [Online]. Available: https://www.taylorfrancis.com/chapters/edit/10.4324/9781315625157-33/information-quality-qi-su. |
| [5] | R. Oshikawa, J. Qian and W. Y. Wang, "A Survey on Natural Language Processing for Fake News Detection". |
| [6] | Y. Lin, "Opinion Spam," [Online]. Available: https://www.worldscientific.com/doi/10.1142/9789813100459\_0007?utm\_source=TrendMD&utm\_medium=cpc&utm\_campaign=World\_Scientific\_Book\_TrendMD\_0. |
| [7] | "Wikipedia," [Online]. Available: https://en.wikipedia.org/wiki/Fact-checking. |
| [8] | J. Soll, "The Long and Brutal History of Fake News," *The Long and Brutal History of Fake News,* 2016. |
| [9] | M. Wendling, "The (Almost) Complete History of ‘Fake News," *The (Almost) Complete History of ‘Fake News.* |
| [10] | Merriam-Webster, "The Real Story of ‘Fake News". |
|  |  |