Fake News Detection Using Different Machine Learning Models

Saman Sarker Joy
Department of CSE
BRAC University
Dhaka, Bangladesh
20101114

Sabrina Tabassum
Department of CSE
BRAC University
Dhaka, Bangladesh
18301135

Nizbath Ahsan
Department of CSE
BRAC University
Dhaka, Bangladesh
20101227

Khaled Mushahed Hossain
Department of CSE
BRAC University
Dhaka, Bangladesh
20101297

Abstract

Fake news is the propaganda of inflammation of spreading misinformation among the people through broadcasting of news or through social media. This project will assist in detection of accuracy of false and real news. We have two different datasets. For pre-processing our dataset, we went through five crucial stages. After implementing the algorithm we get results of each of the 3 classifiers. From the results of our algorithm we can see that Support Vector Classification has the most Accuracy, which is 100%.

1 Introduction

Fake news is the propaganda of inflammation of spreading misinformation among the people through broadcasting of news or through social media. It is to be noted that the fake news rushes faster than real news. Fake news – defined by the researchers as stories debunked by six major fact-checking services – can spread 10 times faster than legitimate news stories, according to the study by researchers at the Massachusetts Institute of Technology. While some U.S. lawmakers and other critics have blamed automated bots for the spread of fake news before the 2016 election, the MIT researchers filtered out tweets spread by bots for their study.

However, establishment of artificial intelligence actually paved the way to stop spreading fake news. AI trains machines according to the ability of completing desired purposes. It introduced with different models to detect the alarming news with fake and real state. These models actually help to decrease the longevity of false news. This project will assist in detection of accuracy of false and real news with collection of data of different types by implementing some machine learning models. This will take the input of the users about any news and analysis it and upgrade the output with informing about real or fake news.

2 Methodology

2.1 Dataset Description

Datasets Link: https://cutt.ly/zXEMtvV
This datasets we used were previously used for two journals authored by H Ahmed, I Traore and S Saad (2) (3).

Two types of separate dataset basically used to detect about fake and real. One dataset is used for fake and the other one for real. 17903 unique values are presented in dataset in combination of fake or real news. More about 2000 data are contained in dataset to visualize the real or fake on with the help of four type's model.

2.2 Dataset Restructure

We have two different datasets and before merging them we make a new column for each of the dataset named "Label". Now, we label all of Fake News = 0 and Real News = 1. Then we merge both of the datasets into one dataset, shuffle them and do the rest of our work with that dataset.

2.3 Pre-Processing Techniques Applied

For pre-processing our dataset, we went through five crucial stages.

Null Checking: A null indicates that a variable doesn't point to any object and holds no value. It is commonly used to denote or verify the

non-existence of something. Here, we checked if any data is missing or not by using a basic 'if' statement. As we did not find any null values, dropping of any rows were not needed.

Making "Content" Column: It is basically merging all information of a news article into a single text. It contains the values of Title, Subject and Text.

Dropping unnecessary columns: It means reducing the number of columns without affecting the main information. Such as: title, subject, text are merged into the content column. So, they no longer need to occupy separate columns in the dataset. Also, the date columns values do not have much significance to our project. So, we dropped these columns. And kept only the necessary columns, Content and Label.

Stemming: Stemming is the process of reducing inflection in words to their root forms such as mapping a group of words to the same stem even if the stem itself is not a valid word in the Language. It is a kind of normalization for words.(4) It is a technique where a set of words in a sentence are converted into a sequence to shorten its lookup. The words which have the same meaning but have some variation according to the context or sentence are normalized. By stemming a user-entered term, more documents are matched, as the alternate word forms for a user-entered term are matched as well, increasing the total recall. This comes at the expense of reducing the precision. However, before stemming, we removed all non-alphabet characters, converted all letters into lowercase then stemmed all Words. And after all that, we removed stop words.

Vectorization: Vectorization is a technique of implementing array operations without using for loops. (5) Instead, we use functions defined by various modules which are highly optimized that reduce the running and execution time of code. Vectorization is used to get some distinct features out of the text for the model to train on by converting text to numerical vectors.

2.4 Models applied

Logistic Regression Score: The logistic regression function $p(x)$ is the sigmoid function of $f(x)$: $p(x) = 1/(1 + exp(f(x))$. (6) As such, it's often close to either 0 or 1. A logistic regression (LR) model is used because it offers the intuitive equation to categorize issues into binary or multiple classes. This is because we are classifying text on the basis of a large feature set, with a binary output (true/false or true article/fake article). While several parameters were tested before obtaining the maximum accuracies from the LR model, we did hyperparameter tuning to acquire the best outcome for each particular dataset. The logistic regression hypothesis function can be defined mathematically as follows:

$$h_\theta(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}} \tag{1}$$

The output of logistic regression is converted to a probability value using a sigmoid function; the goal is to minimize the cost function to arrive at the best probability. As demonstrated in the cost function calculation:

$$Cost\left(h_\theta(x), y\right) = \begin{cases} \log\left(h_\theta(x)\right), & y = 1 \\ -\log\left(1 - h_\theta(x)\right), & y = 0 \end{cases} \tag{2}$$

Support Vector Classification: Another model for the binary classification problem is the support vector machine (SVM), which comes in different kernel functions. An SVM model's goal is to calculate a hyperplane (or decision boundary) based on a feature set in order to categorize data points. (7) The goal is to locate the hyperplane that separates the data points of two classes with the greatest margin, which may take many different forms in an N-dimensional space. The cost function for the SVM model is illustrated in a mathematical form:

$$J(\theta) = \frac{1}{2}\sum_{j=1}^{n} \theta_j^2 \tag{3}$$

$$\theta^T x^{(i)} \geq 1, \quad y^{(i)} = 1 \tag{4}$$

$$\theta^T x^{(i)} \leq -1, \quad y^{(i)} = 0 \tag{5}$$

A linear kernel is used in the code above. Kernels are typically used to fit multidimensional or difficult to easily separate data points. We have applied sigmoid SVM, kernel SVM (polynomial SVM), Gaussian SVM, and fundamental linear SVM models in our situation.

Naive Bayes Classification: It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. (8) In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

It is a classification algorithm that is suitable for binary and multiclass classification. It is also called diot Bayes because the calculations of the probabilities for each class are simplified to make their calculations tractable. Naive Bayes performs well in cases of categorical input variables compared to numerical variables. It is useful for making predictions and forecasting data based on historical results.

3 Results

This is Logistic Regression's Result:

Logistic Regression				
	precesion	recall	f1-score	support
0	0.99	0.99	0.99	4682
1	0.99	0.99	0.99	4298
avg	**0.99**	**0.99**	**0.99**	**8980**

This is SVM's Result:

SVM				
	precesion	recall	f1-score	support
0	1.00	1.00	1.00	4682
1	1.00	1.00	1.00	4298
avg	**1.00**	**1.00**	**1.00**	**8980**

This is Naive Bayes's Result:

Naive Bayes				
	precesion	recall	f1-score	support
0	0.94	0.95	0.95	4682
1	0.95	0.93	0.94	4298
avg	**0.94**	**0.94**	**0.94**	**8980**

After the deployment of one of the test data through our Machine Learning algorithm. The algorithm presents us with the result that it detected by learning from the training data that we have stored earlier. We learnt from data exploration that articles with "No author, No title, No Image and bs" types are more likely to be fake news. In all the classifier cases we split the dataset into 20% test data and 80% training data.

The model's positive predictive values(precision) represent the appropriate text among the repossessed text documents, whereas sensitivity(recall) represents the fraction of the total number of related text documents that were actually retrieved. F1-score represents the harmonic mean of the combination of precision and recall.

Classifiers	**Precision**	**Recall**	**f1-Score**
LR	0.99	0.99	0.99
SVM	**1.00**	**1.00**	**1.00**
NB	0.94	0.94	0.94

After implementing the algorithm we get the Precision, Recall, f1-score of each of the 3 classifiers. From the results of our algorithm we can see that Support Vector Classification has the most Accuracy, which is 100%.

References

[1] "Fake News Spreads Fast, but Don't Blame the Bots." *Internet Society*, 21 Mar. 2018, https://www.internetsociety.org/blog/2018/03/fake-news-spread-fast-dont-blame-bots/.

[2] Ahmed H, Traore I, Saad S. *"Detecting opinion spams and fake news using text classification"*, Journal of Security and Privacy, Volume 1, Issue 1, Wiley, January/February 2018.

[3] Ahmed H, Traore I, Saad S. (2017) *"Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In: Traore I., Woungang I., Awad A. (eds) Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments.* ISDDC 2017. Lecture Notes in Computer Science, vol 10618. Springer, Cham (pp. 127-138).

[4] "Stemming and Lemmatization in Python." *DataCamp, DataCamp*, 23 Oct. 2018, https://www.datacamp.com/tutorial/stemming-lemmatization-python.

[5] "Vectorization in Python - A Complete Guide." *AskPython*, 5 Sept. 2021, https://www.askpython.com/python-modules/numpy/vectorization-numpy.

[6] Real Python. "Logistic Regression in Python." *Real Python, Real Python*, 18 Aug. 2022, https://realpython.com/logistic-regression-python/.

[7] Contributor, TechTarget. *"What Is Support Vector Machine (SVM)? - Definition from Whatis.com."* WhatIs.com, TechTarget, 29 Nov. 2017, https://www.techtarget.com/whatis/definition/support-vector-machine-SVM.

[8] Brownlee, Jason. "Naive Bayes Classifier from Scratch in Python." *Machine Learning Mastery*, 24 Oct. 2019, https://machinelearningmastery.com/naive-bayes-classifier-scratch-python/.