Detection of Fake News Based on Typical Machine Learning Models

Ruining Yang *

Department of Data Science, Mount Holyoke College, MA, USA

*Corresponding author: yang26r@mtholyoke.edu

Abstract. With the rapid expansion of the network, the glut of news spread everywhere. Because of the obscurity of news sources and the unrestricted types of viewers, the harmful impact of false news is more pervasive than ever before. The goal of this study is to evaluate the efficacy of five machine learning models, namely Decision Tree, Logistic Regression, Random Forest, Multilayer Perceptron (MLP) and Naive Bayes to detect false news using a dataset obtained from Kaggle. Following the application of five models for predicting false news based on the news' title and comparison of the training and testing accuracies of each model, the results indicate that Random Forest is the best model, with Decision Tree and MLP models also having very high testing accuracies. Surprisingly, the Naive Bayes model, widely recognized as the optimal classifier for text data, had the lowest testing accuracy in this study, implying that more research is required to explain this outcome. Finally, the limits of current machine learning algorithms, as well as the possibility of bias in datasets, provide a good direction for future studies.

Keywords: Fake News, Machine Learning Models, Random Forest.

1. Introduction

Social media has proliferated with the widespread adoption of digital technology and mobile Internet. Online news removes the restriction on the identity of information publishers that existed in traditional news media, allowing people to publish information while receiving it. The previous ethical framework of information dissemination was impacted when many low-threshold self-media intervened in news production and dissemination. Some fake news will blind people's minds and even cause social panic due to the difficulty of validation and the diversity of motivation. For example, the emergence of the pandemic and measures of social confinement caused social media and search queries to obtain information about the progression of the disease to be constantly expanding, including on Twitter, Facebook, Instagram, and other popular sources [1]. It allows fake news to multiply rapidly and act as narratives that omit or add information to facts. Moreover, some fake news stems from conspiracy theories, such as a biological weapon produced in China will cause social panic or accelerate the spread of the virus [2]. The development of artificially intelligent algorithms enhances the human ability to detect fake news and provides a feasible way to deal with the chaos of fake news on the Internet since there have been lot of studies regarding the excellent application of machine learning [3-5]. Many researchers distinguish fake news from real news by news source and audience type, but the source of online news is not always clearly indicated, nor is the reach of the audience. Moreover, many people pay less attention to the source of information and only focus on the news itself.

Hakak et al. proposed a method based on the ensemble classification model for detecting fake news [6]. This study was conducted on two different datasets called ISOT and Liar. First, they identified 26 linguistic-based textual features e.g. the number of word and characters. Then, an ensemble classification model consisting of three machine learning supervised algorithms, namely Random Forest, Decision Tree, and Extra Tree Classifier, was built. Finally, on the Liar dataset, they got training and testing accuracy of 99.8% and 44.15% training and testing accuracy. For the ISOT dataset, they achieved the training and testing accuracy of 100%. However, it is rare in reality to get 100% accuracy in test data. Moreover, the accuracy of the training dataset is super high but for the testing, the dataset is low in the Liar dataset. Thus, this study may suffer an overfitting problem.

Another related work experimented with and used the Liar dataset to examine four typical machine learning algorithms (i.e. random forest, the neural network, the Naïve Bayes, and the decision trees), to train them in the fake news dataset and validate the effectiveness of these machine learning algorithms [7]. The result showed that the Naïve Bayes classifier defeats the other algorithms remarkably in this dataset. However, the Liar dataset is relatively well-established, so it may be easier to get high accuracy of the model by using this dataset, and the potential biases or unseen errors may appear in this dataset. On the other hand, the accuracy for Naïve Bayes in five K-fold cross-validations is unchanged, but the accuracy for another method is slightly changed. So, there may need to be more K-fold cross-validations to justify the results.

Therefore, this study aims to find optimal machine learning models to detect fake news by comparing the accuracy of several classifiers.

2. Methods

2.1 Dataset description and preprocessing.

The public dataset used for this study is on Kaggle [8], which includes 20, 387 items of data with five categories: id, title, author, text, and label (1: true news, 0: fake news). Specifically, this paper combined news titles and authors for 20, 387 news in the training file for analysis. The sample data is shown in Table 1.

Table 1. Dataset Sample

id	title	author	label
I	House Dem Aide: We Didn't Even See Comey's Let...	Darrell Lucus	1
II	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	0
III	Why the Truth Might Get You Fired	Consortiumnews.com	1
IV	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	1

Because there are relatively minimal missing values with 558 missing titles, this study decided to remove them from the dataset. From the rest of the data, the number of news was divided by the ratio of 2:8 and set them as train and test sets, respectively. In the preprocessing step, this study would like to amplify signals from literal strings to convert words to some common form. So, a stemmer was deployed, which uses some decision rules to chop off endings to get to some common base representation of a word. Following that, tokens were lemmatized, a more complex way of identifying common roots. Finally, title and author were combined as our x variable and label as our y variable.

2.2 Machine learning techniques

The analysis of this study applied five machine learning models, including Logistic Regression, Decision Tree, Random Forest, MLP, and Naive Bayes, and compared their accuracies.

The logistic regression model is a statistical model that predicts the likelihood of an event occurring by making the recorded probability of an event a linear combination of one or more independent variables [9]. The sigmoid function serves as the foundation for this approach. The

logistic regression model, as opposed to the linear regression model, provides a binary output and is superior for categorization.

The Decision Tree model is a computing paradigm in which an algorithm is essentially a decision tree, i.e., a succession of inquiries or tests that are done adaptively, such that the results of earlier tests might impact the test that is conducted next [10]. As a result, regular people will find it easier to perceive and comprehend this model.

Random forest that can complete regression and classification tasks, is an ensemble learning model containing multiple decision trees, each reflecting a unique instance of the random forest's classification of data input. The random forest approach examines each case independently, selecting the one with the most votes as the chosen forecast.

A multilayer perceptron is a simple type of the artificial neural network (ANN). In general, it consists of three parts, namely an input layer, a hidden layer, and an output layer. With the exception of the input nodes, each node will be combined with the activation function for output [12].

Naive Bayes is a conditional probability model based on applying Bayes' theorem to characteristics with strong independence requirements. As a result, naive Bayes classifiers are extremely scalable, needing a small number of parameters that are proportional to the number of features in the dataset [13].

3. Results

The training and testing accuracy for each classifier will be taken into account when evaluating the performance of machine learning models. The accuracy of each classifier is compared in Table 2.

Table.2. Training and testing accuracy for machine learning models

Models	Training Accuracy	Testing Accuracy
Logistic Regression	98.69%	97.76%
Decision Tree	100%	99.28%
Random Forest	100%	99.30%
Multi-layer Perceptron Neural Network	100%	99.18%
Naive Bayes	94.72%	80.41%

In general, all models have high training and testing accuracy. Except for the Naive Bayes model, each model's training accuracies are greater than 95%, and some even reach 100%. The Random Forest classifier, which has roughly 99.30% testing accuracy, is the most accurate of these algorithms. The testing accuracies of the Decision Tree and Multi-layer Perceptron models are also similar, at 99.28% and 99.18%, respectively. Naive Bayes has the lowest testing accuracy of 80.41% and the lowest training accuracy of 94.72%.

4. Discussion

According to Table 2, the Random Forest model is the best model for this dataset. The Decision Tree and Multi-layer Perceptron models are also ideal models because the differences in training accuracy among these three models are less than 0.01. It is reasonable because, due to its parallel architecture, the Random Forest achieve more satisfactory result than other machine learning models [14]. Furthermore, the Decision Tree and Random Forest models share many similarities. Both the forest and the decision tree have nearly identical hyperparameters and generate randomly partitioned data. Compared to previous models, logistic regression has a lower training accuracy, but it is still greater than 95%. Surprisingly, the Naive Bayes model has the lowest training and testing accuracy; additionally, the difference in testing accuracy between Naive Bayes and the other models exceeds 10%. Because hidden messages or text correlations can be detected in the Naive Bayes model, it is commonly employed for text analysis. However, this study demonstrates the Naive Bayes model's text categorization limitations. The appearance of zero frequency could be the cause. Suppose a test data

set contains a categorical variable from a category that was not present in the training data set. In that case, the Naive Bayes model will give it zero probability and will be unable to make any predictions. The low Nave Bayes prediction accuracy may support this theory. Another possibility is that the textual data used in this study is the title of the news rather than the content of the news. Because Nave Bayes used conditional probability to detect the hidden message from word and sentence correlations, and news titles have fewer correlations to detect, Nave Bayes' accuracy may not be expected. Furthermore, because both prediction and training accuracy are very high, and the difference between the two types of accuracy is small, overfitting issues can be neglected in this study.

5. Conclusions

In sum, using computational tools for text analysis has been a relatively new and popular subject in recent years. This study uses a dataset from Kaggle to perform false and genuine new classifications and to evaluate five machine learning classifiers, including logistic regression, decision trees, random forest, multi-layer perceptron neural networks and Naïve Bayes. The findings demonstrated that the random forest model has the highest training accuracy, whereas the Naïve Bayes model has the lowest. It is a new attempt to detect false news by using the title of the news. However, due to the title's short duration, additional data is needed in machine learning models for in-depth analysis. In addition, as previously noted, the poor accuracy of the Naïve Bayes model in this work may imply the necessity for an extensive dataset for future relevant research.

Furthermore, the classification of false and factual news, which is the 'label' column, is done by hand so some unconscious bias may be present in this dataset. The definition of false news varies depending on the circumstances, and individuals must discern between fake and true news based on their own judgment, resulting in prejudice. Indeed, ethics is a common worry when discussing machine learning, especially when using machine learning to evaluate text data. Regardless of their background, every choice or thought produced by a human being is considered subjective. This issue is more of a concern for text data because human involvement is higher when analyzing text data than when analyzing numeric data. As a result, researchers must reduce bias when detecting false news using machine learning approaches.

References

[1] Depoux A, Martin S, Karafillakis E, Preet R, Wilder-Smith A, Larson H, The pandemic of social media panic travels faster than the COVID-19 outbreak [J]. J Travel Med 27(3), 2020.

[2] Hua J, Shaw R. Corona virus (Covid-19) “infodemic” and emerging issues through a data lens: the case of China [J]. Int J Env Res Public Health. 2020.

[3] Patel, Jigar, et al. Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques [J]. Expert systems with applications 42.1 (2015): 259-268.

[4] Y. Qiu, et al. Clustering Analysis for Silent Telecom Customers Based on K-means++ [C]. 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). Vol. 1. IEEE, 2020.

[5] Kononenko, Igor. Machine learning for medical diagnosis: history, state of the art and perspective [J]. Artificial Intelligence in medicine 23.1 (2001): 89-109.

[6] Hakak, Saqib, et al. An ensemble machine learning approach through effective feature extraction to classify fake news [J]. Future Generation Computer Systems 117 (2021): 47-58.

[7] Albahr, Abdulaziz, and Marwan Albahar. "An empirical comparison of fake news detection using different machine learning algorithms. [J]" Int. J. Adv. Comput. Sci. Appl 11.9 (2020): 146-152.

[8] Kaggle, Fake News, https://www.kaggle.com/c/fake-news/data?select=train.csv, 2018.

[9] Wikimedia Foundation, Logistic regression [R]. Retrieved April 24, 2022, from https://en.wikipedia.org/wiki/Logistic_regression, 2022.

[10] Wikimedia Foundation. Decision tree [R]. Retrieved April 24, 2022, from https://en.wikipedia.org/wiki/Decision_tree, 2022.

[11] Wikimedia Foundation. Random Forest [R]. Wikipedia. Retrieved April 24, 2022, from https://en.wikipedia.org/wiki/Random_forest, 2022.

[12] Wikimedia Foundation. Multilayer Perceptron [R]. Wikipedia. Retrieved April 24, 2022, from https://en.wikipedia.org/wiki/Multilayer_perceptron, 2022.

[13] Wikimedia Foundation. Naive Bayes classifier [R]. Wikipedia. Retrieved April 24, 2022, from https://en.wikipedia.org/wiki/Naive_Bayes_classifier, 2022.

[14] Jalal, N., Mehmood, A., Choi, G. S., & Ashraf, I. A novel improved random forest for text classification using feature ranking and optimal number of trees [J]. Journal of King Saud University - Computer and Information Sciences. Retrieved April 24, 2022.