Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/40/20230661

Identification and analysis of real and fake news by XGBoost algorithm of machine learning

JueYu Chen

School of Artificial Intelligence, Southeast University, Nanjing, 211189, China

jueyuc@seu.edu.cn

Abstract. As the internet develops rapidly, fake news has become increasingly easy to propagate. Numerous academics acknowledge the perilous nature of this phenomenon, particularly in the context of the contemporary "post-truth era," highlighting its substantial risk to the public. Hence, the detection and halting of fake news dissemination are absolutely vital. This study utilizes machine learning's eXtreme Gradient Boosting (XGBoost) algorithm to construct a model that can differentiate between genuine and fake news. The model is compared with others utilizing different algorithms and is ultimately selected. The study successfully constructs a model with an accuracy rate of approximately 95% in identifying real and fake news. This model provides the public with a convenient way to differentiate between real and fake news and gradually diminishes the threat of fake news. Additionally, this project's implications extend beyond merely identifying real and fake news. The model can be further developed to detect fake information, providing greater societal benefits.

Keywords: Fake News, eXtreme Gradient Boosting (XGBoost) Algorithm, Machine Learning, Identification.

1. Introduction

In 2004, Ralph Keyes proposed the concept of, the "post-truth era", and deemed that, ambiguities will become a new view of reality in the future [1]. Nowadays, people are more inclined to things that cater to their interests or appetites, and this phenomenon contributes to the neglect of the reality of news we are confronted with [2]. A study conducted by Craig Silverman and Jeremy Singer-Vine revealed that 75% of American adults confronted with misinformation believed the news they encountered to depict true and accurate information [3]. Moreover, contemporary individuals exhibit a heightened propensity to disseminate misinformation compared to past tendencies [4]. In recent days, fake news is easier to spread than before and more difficult to detect, which confuses the public's sights and brings harm to society. Hence, to curtail the propagation of fake news, the necessity for a model that discerns the veracity of news becomes unequivocally indispensable.

Indeed, paralleling the swift advancement of the Internet, the issue of fake news transcends local confines, emerging as a menace of global proportions. [5]. In 2013, An alarm was put on "digital wildfire"----which brings the "viral spread" of misleading information to the public intentionally or unintentionally----by the World Economic Forum [6]. More and more scholars have realized the threat that fake news may bring, therefore, there has been some related prior work and research on real and fake news identification. Some scholars discussed and used deep-learning models to perform

DOI: 10.54254/2755-2721/40/20230661

misinformation detection. For instance, Islam, Liu, Wang, and Xu have made an excellent conclusion about the performance of various deep-learning models in the misinformation detection field [7]. Thota, Tilak, Ahluwalia, and Lohia, carried out fake news detection by using neural network [8]. Girgis, Amer, and Gadallah had a try on increasing accuracy by applying a hybrid model of the Gate Recurrent Unit (GRU) and Convolutional neural network (CNN) techniques [9]. Besides, other scholars applied machine learning models to real and fake news identification. For example, Aphiwongsophon and Chongstitvatana detected fake news by Naive Bayes, Support Vector Machine, and Neural Network models and made a comparison between them [10]. Iftikhar Ahmad and his team ingeniously amalgamated an array of machine learning methodologies (including Logic Regression, Random Forests) to tackle the nuanced challenges of fake news across diverse contexts and disciplines. [11]. The prior work has provided some inspiration in the early stage of this research, and in the progress of trial, this research finds out the eXtreme Gradient Boosting (XGBoost) model illustrates a brilliant performance in coping with genuine and counterfeit news identification.

By leveraging machine learning techniques, this research aims to effectively recognize the authenticity of news. Specifically, the study employs the XGBoost algorithm to construct the model. XGBoost, a powerful algorithm based on gradient boosting, exhibits exceptional accuracy and efficiency when applied to large-scale datasets. This capability proves valuable in discerning between true and false news. Moreover, XGBoost facilitates automatic feature selection, enabling the model to automatically identify and rank crucial features for distinguishing between real and fake news. This enhances the training process efficiency and enables the model to better comprehend meaningful features. Additionally, XGBoost excels at capturing non-linear relationships and complex interactions, which is advantageous for detecting fake news that often possesses intricate textual patterns. The model achieves an impressive average accuracy of 94.9% in discerning real and fake news. The experimental results demonstrate the model's efficacy in detecting both types of news. Furthermore, this paper proposes potential avenues for the future development and enhancement of the model.

2. Methodology

2.1. Dataset description

In this study, the researcher leverages an openly accessible dataset from Kaggle comprising 72,134 news articles, inclusive of 35,028 genuine and 37,106 counterfeit news pieces from numerous sources (for instance, Kaggle, McIntire, Reuters, Buzzfeed Political) [12]. This inclusion provides the model with a substantial data foundation and mitigates potential over-fitting issues induced by classifiers. The dataset consists of a diverse collection of articles spanning diverse topics and domains, e.g.politics, economics, culture, etc. It encompasses reliable and unreliable information to create a balanced representation of real-world news scenarios. To classify the real and fake news clearly, this dataset labels real news by number 1 and fake news by number 0.

2.2. Proposed approach

The chief objective of this project is to construct a resilient model that discerns real from fake news, empowering the public to identify fake news effectively and thereby curtail its propagation. This project introduces the XGBoost algorithm, a powerful machine learning method, to cope with real and fake news identification. In the modelling process, the classifier is instantiated with specific evaluation metrics and label encoding settings. The classifier thrives on data meticulously earmarked for training, thereby making predictions on the test set. Its efficacy is evaluated by calculating the accuracy score. To gain valuable insights into the classifier's performance, a confusion matrix is generated which meticulously analyses the spread of predicted and actual labels. For better interpretability, this matrix is visually transformed into a heatmap. Overall, this approach allowed for the development of an effective predictive model using XGBoost, facilitating further analysis and decision-making.

DOI: 10.54254/2755-2721/40/20230661

2.2.1. Data pre-processing. Before feeding the data set into our models, several preprocessing steps are applied. Firstly, basic data exploration is performed. The author prints the first 10 rows of the dataset to illustrate a quick overview of the data's structure and format and carries out codes to generate a statistical summary of the dataset to get the distribution of the data, outliers, and missing values. The author also prints the basic information of the data, including the total number of rows, the number of non-null values for each feature, and the data types of the features. Next, the author removes unnecessary columns to avoid redundant information and noises influencing the model's performance. Besides, the author replaces the missing values with empty strings, which can prevent missing values from posing challenges in data analysis and modeling, which may lead to biased results or errors. Followingly, the author combines titles and texts into a single feature, calculates the body length of each news article, and performs data visualization to vividly illustrate the dataset. Subsequently, the author partitions the data set into training and testing subsets, proceeding to vectorize the text data. This optimization paves the way for a seamless integration of data into the forthcoming model. The whole data preprocessing process is illustrated in Figure 1.

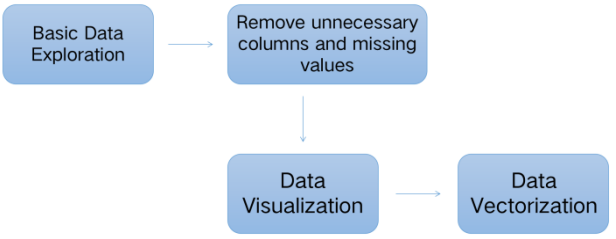

Figure 1. The process of data preprocessing.

2.2.2. XGBoost Algorithm. XGBoost, an abbreviation for eXtreme Gradient Boosting, represents a formidable machine learning technique that builds upon the principles of the Gradient Boosting Decision Tree, thereby exhibiting a comprehensive machine learning algorithm. It is applied extensively in the classifier and regression problems. In this project, the problem of identifying real and fake news can be understood as classifying the news set into two categories (real or fake), which in nature is a problem of classification. Thus, in essence, to build an efficient model to identify the real and fake news is to build an efficient classifier to divide the news set into real and fake news sets, which is actually what the XGBoost algorithm is skilled in.

XGBoost creates a strong predictive model by combining the efforts of several weaker models, often decision trees, in an iterative manner. Beginning with a foundational model, XGBoost progressively incorporates new models with each iteration to predict and rectify the residuals - that is, discrepancies between the foreseen and actual values stemming from the preceding model. It employs a gradient descent algorithm to minimize these residuals. The process is shown in Figure 2. Meanwhile, unlike other gradient-boosting algorithms, XGBoost contains a regularization term in its objective function, which helps control the model's complexity and prevents overfitting. It also supports parallel computing for enhanced performance and provides built-in handling for missing values, outlier detection, and tree pruning. It shows excellent performance in dealing with classifier problems, which is superbly fit for this project to cope with the identification of real and fake news.

DOI: 10.54254/2755-2721/40/20230661

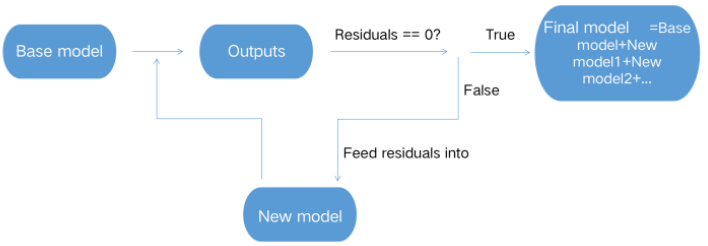

Figure 2. The basic process of the XGBoost.

In this project, XGBoost is applied to build the model of real and fake news identification. Firstly, the author instantiates the XGBoost classifier, using the root mean square error as the evaluation metric, and handles the label encoding appropriately. Subsequently, the author employs the training data set to train the model. Once the model undergoes comprehensive training, it is employed to make predictions on the test data set. The author subsequently gauges the model's accuracy by computing the accuracy score. The predicted labels are contrasted with the actual ones, leveraging the accuracy score for this purpose. The subsequent accuracy score furnished an insightful appraisal of the model's predictive prowess. Furthermore, to gain a more detailed understanding of the classifier's performance, the author generates a confusion matrix to analyse the distribution of predicted labels compared to the actual labels, using a heatmap to vividly visualize the confusion matrix.

2.3. Implemented details

This study uses Google Colab with Python 3.10.12 to carry out the code running on a macOS device with Apple M2 SoC. In the data preprocessing stage, the study applies matplotlib, seaborn library, and CountVectorizer model to realize the data visualization and data vectorization. The study uses the XGBoost Classifier to build the classifier for the model and calculate the accuracy of the predicted results by using the accuracy score function. Additionally, the study generates the confusion matrix that visualizes the comparison between the predicted and actual results by plotting a heatmap.

3. Result and discussion

In this project, the process of building the model of real and fake news identification can be divided into two main parts, the data pre-processing and the model constructing through the XGBoost algorithm. In the first main part----the data pre-processing part, the author carries out the basic data exploration to dig out the features of the data to help better understand the mathematical features of the real and fake news data. As shown below, the Figures 3-4 illustrate the result of the data exploration. In the Figure 3, it shows the statistical summary and basic information of the dataset. The Figure 4 shows the proportion of fake and real news. In the Figures 5 and 6, they demonstrate the result of the exploration of the text of the news dataset for fake and real news through the form of word cloud. After the data preprocessing stage, the author can get a general understanding of the features of the data set.

DOI: 10.54254/2755-2721/40/20230661

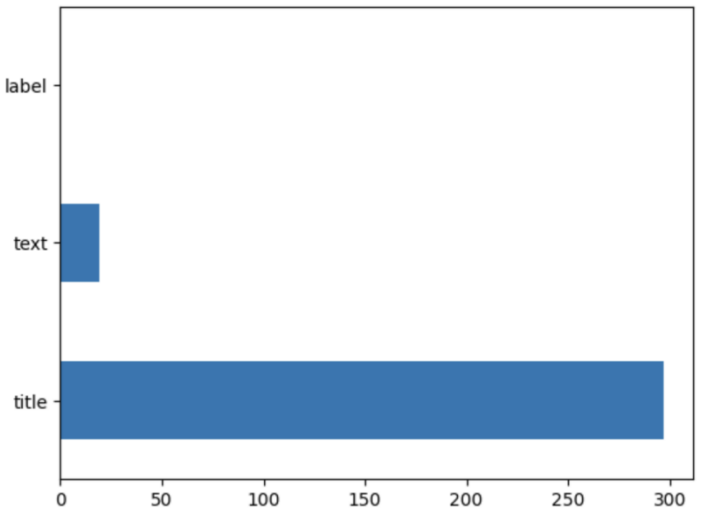

Figure 3. The result of basic data exploration.

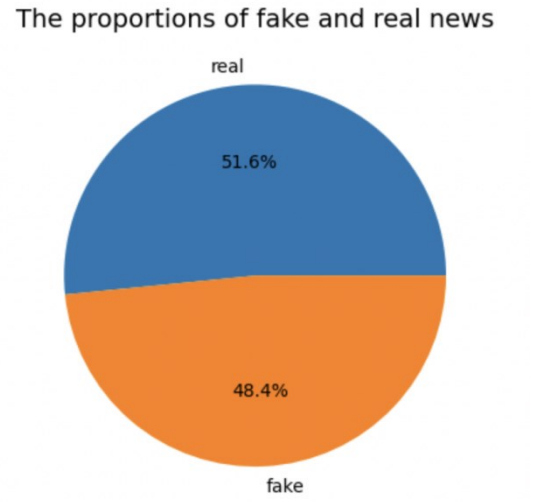

Figure 4. The proportions of fake and real news.

Figure 5. Word cloud of fake news.

DOI: 10.54254/2755-2721/40/20230661

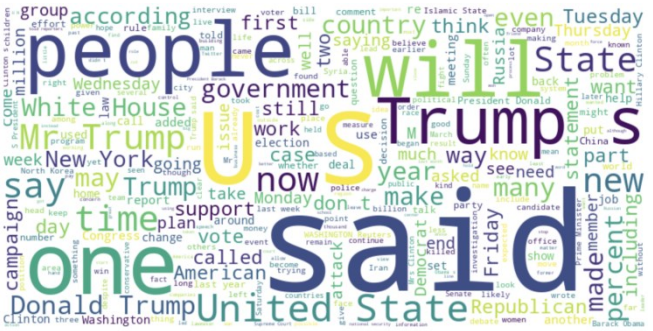

Figure 6. Word cloud of real news.

Then, it comes to the second main part---- the model constructing part. In this part, XGBoost displays a great performance in the identification of real and fake news. Before the author uses the XGBoost to construct the model, the author tries several other machine learning methods. In Table 1, there are performances of different machine learning methods, and compared with other methods, the author found out the XGBoost shows the best performance for the aspects of accuracy and efficiency.

Table 1. The result of the models.

Algorithm /Method	Accuracy (/1)	Time-costing(/s)
XGBoost	0.95	199.078
Naive Bayes	0.87	151.490
Random Forest	0.93	720.978
Support Vector Machine	0.91	142.904

Compared with other traditional machine learning methods, XGBoost, though sometimes poses lower speed than some algorithms, showcases the best accuracy and better performance when taking both accuracy and efficiency into account under the same data preprocessing condition. XGBoost, based on the thinking of ensemble learning, combines the efforts of some weaker models through the process of iteration and modification. Through the process of iteration and modification, multiple features are taken into consideration to improve the performance of the model, which helps XGBoost to capture the important features automatically and cope with the complex relationships between features. This makes the XGBoost display an efficient and accurate performance in the problem of identification of real and fake news. Finally, the XGBoost illustrates an accurate prediction and identification of real and fake news with 94.9% accuracy. This model, crafted for discriminating between real and fabricated news, has extensive practical implications in society. It aids in enhancing the public's discerning capabilities amidst the vast sea of unverified content inundating the internet, ultimately facilitating the identification of authentic and spurious news. Up next is Figure 7, a displayed heatmap of the project model's confusion matrix, offering a visual representation of the model's outcome and precision.

DOI: 10.54254/2755-2721/40/20230661

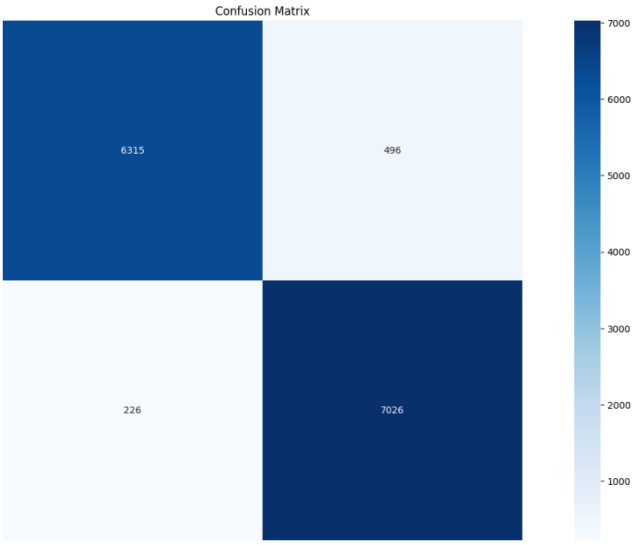

Figure 7. Heatmap of the confusion matrix of the model in this project.

4. Conclusion

This paper unveils a model meticulously crafted to adeptly differentiate between authentic and counterfeit news, thereby guiding the public to pinpoint and circumnavigate the pitfalls of misinformation. The utilization of the XGBoost algorithm plays a key role in analyzing, modeling, and identifying true and false news. The algorithm demonstrates remarkable accuracy and efficiency when handling extensive datasets, enabling robust identification of genuine and fabricated news. The research process encompasses data preprocessing, instantiation of the XGBoost classifier, model training (including iterative refinement), testing, and accuracy evaluation. The model achieves an accuracy rate of nearly 95% in recognizing true and false news, thereby offering significant practical utility in combating fake news in real-world scenarios. Future investigations can focus on enhancing the accuracy and efficiency of the model's capability to discern between authentic and fabricated news. Moreover, extending the model's scope beyond fake news to encompass other forms of misinformation and its application across various domains represent promising avenues for future research.

References

[1] Keyes R 2004 The post-truth era: Dishonesty and deception in contemporary life Macmillan
[2] Tian F 2022 Public Opinion on Education in the Post-truth Era Journal of East China Normal University (Educational Sciences) 40(3): p 30
[3] Silverman C Singer-Vine J 2016 Most Americans who see fake news believe it new survey says BuzzFeed news 6(2)
[4] Weidner K Beuk F Bal A 2020 Fake news and the willingness to share: a schemer schema and confirmatory bias perspective Journal of Product & Brand Management 29(2): pp 180-187
[5] Howell L 2013 Digital wildfires in a hyper connected world WEF Report 3: pp 15-94
[6] World Economic Forum. The rapid spread of misinformation online 2014.
[7] Islam M R Liu S Wang X et al 2020 Deep learning for misinformation detection on online social networks: a survey and new perspectives Social Network Analysis and Mining 10: pp 1-20
[8] Thota A Tilak P Ahluwalia S et al 2018 Fake news detection: a deep learning approach SMU Data Science Review 1(3): p 10.
[9] Girgis S Amer E Gadallah M 2018 Deep learning algorithms for detecting fake news in online text international conference on computer engineering and systems (ICCES) IEEE pp 93-97

DOI: 10.54254/2755-2721/40/20230661

[10] Aphiwongsophon S Chongstitvatana P 2018 Detecting fake news with machine learning method international conference on electrical engineering/electronics computer telecommunications and information technology (ECTI-CON) IEEE pp 528-531
[11] Ahmad I Yousaf M Yousaf S et al 2020 Fake news detection using machine learning ensemble methods Complexity 2020: pp 1-11
[12] Dataset https://www.kaggle.com/code/adityarahul/fakenewsdetection/input