Journal of Theoretical and Applied Information Technology

30th November 2023. Vol.101. No 22
© 2023 Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

EFFICIENCY OF FAKE NEWS DETECTION WITH TEXT CLASSIFICATION USING NATURAL LANGUAGE PROCESSING

SALMAN AL FARISY AZHAR[1], FELIK HIDAYAT[2], MUHAMMAD HANIF AZFAREZAT, GHINAA ZAIN NABIILAH[4], ROJALI[5]

Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480

E-mail: [1]Salman.Azhar@binus.ac.id, [2]FelikHidayat@binus.ac.id, [3]muhammad.azfareza@binus.ac.id, [4]ghinaa.nabiilah@binus.ac.id, [5]rojali@binus.edu

ABSTRACT

Fake news has become a significant concern in today's information landscape, posing serious threats to society, democracy, and individual well-being. To combat the spread of fake news, effective detection mechanisms are essential. This paper investigates the efficiency of fake news detection through text classification using natural language processing (NLP) techniques. The study explores the application of various NLP algorithms, including feature extraction methods, sentiment analysis, and machine learning classifiers, to identify and classify news articles as either real or fake. The performance of different NLP approaches is evaluated using a comprehensive dataset comprising diverse news sources. In this paper we have used several methods of Algorithm such as Naïve Bayes, Random Forest, Logistic Regression, Decision Tree and, XGBoost. The results showed that certain algorithms like recurrent XGboost and Decision Tree machines performed well in detecting fake news with an accuracy score of 99.76% and 99.53%.

Keywords: *Fake News Detection, Text Classification, NLP, Machine Learning Classifiers, Dataset. Random Forest, Logistic Regression, Xgboost, Decision Tree, Naive Bayes, Preprocessing*

1. INTRODUCTION

The digital era has changed everything, we can get information easily and practically. In this era, all human beings can get and search for news easily on the internet. Currently, not many people, from the young to the elderly, read the news in the newspapers, the majority read news articles on the internet. News on the internet is usually in the form of articles uploaded by news spreaders. With so much uploaded news, there must be news that is not true with the facts. Fake news can circulate quickly so that there are many misunderstandings among the internet community who read it. As a solution, fake news can be prevented using the help of algorithms that are applied to programming languages. The algorithm in question includes a classification of sentences that will help determine whether the news is true or fake news. This classification determination has many algorithms that can detect fake news. The purpose of this research is to find the most accurate, fast and efficient method for detecting fake news using text data.

2. LITERATURE REVIEW

This Efficiency of Fake News Detection study aims to evaluate and improve the effectiveness of currently used fake news detection methods, through the use of new approaches or the development of more sophisticated detection algorithms. The focus of this research may include data analysis, artificial intelligence models, or the proliferation of fake news poses significant challenges in ensuring the accuracy and reliability of information in today's digital age. Detecting and combating fake news is crucial to prevent the spread of misinformation and its potential consequences.

Several more complex methods of text analysis. The algorithmic method we use in our project is an artificial intelligence model, or a more complex text analysis method. As we know that to make a literature review it is necessary to analyse

several journals, such as the journals below that I have found to be reviewed about our project. researchers have proposed different approaches and methodologies to address this issue. This literature review aims to analyze and compare various studies that focus on fake news detection using different techniques, such as NLP and machine learning classifiers.

This Efficiency of Fake News Detection study aims to evaluate and improve the effectiveness of currently used fake news detection methods, through the use of new approaches or the development of more sophisticated detection algorithms. The focus of this research may include data analysis, artificial intelligence models, or The proliferation of fake news In a study conducted by Shaina Raza and Chen Ding (2022), the authors propose a transformer-based approach for fake news detection. Their framework aims to overcome the challenges of early detection and the lack of labeled data. The study utilizes natural language processing techniques to achieve high accuracy in detecting fake news. It concludes that advanced machine learning techniques are necessary to tackle the challenging task of fake news detection [1].

Rakhmat Arianto, Spits Warnars Harco Leslie, Yaya Heryadi, and Dan Edi Abdurachman (2021) develop a fake news detection model based on credibility measurement for Indonesian online news. Their research incorporates a survey and literature study to address the challenges specific to the Indonesian context. The model achieves an accuracy of 87.5% in detecting fake news, making it a significant contribution to preventing the spread of false information in Indonesia [2].

Aisyah Awalina, Jibran Fawaid, Rifky Yunus Krisnabayu, and Dan Novanto Yudistira (2021) Explore Fake News Detection in Indonesia Using a Transformer Network. Their survey-based study concludes that natural language processing techniques can be effective in detecting fake news. They highlight different approaches, including rule-based, machine learning-based, and knowledge-based methods, but also acknowledge challenges like the lack of quality training data and complex language variations [3].

Shalini Pandey, Sankeerthi Prabhakaran, N V Subba Reddy, and Dinesh Acharya (2022) Focus on Fake News Detection From Online Media Using Machine Learning Classifiers. Their study employs NLP techniques and evaluates the performance of various classifiers. The results show high accuracy rates for KNN, Logistic Regression, Naïve Bayes, and SVM, highlighting the effectiveness of machine learning in detecting fake news [4].

In a study by F.A. Kulam Magdoom, Hari Narayanan N, and A. Ramachandran (2022), a Combination of The Aggressive Classifier and BERT is Utilized for Hoax Detection and Review Analysis. Their research shows that this combination achieves good accuracy in predicting fake news. They also emphasize the importance of accuracy and efficiency in fake news detection, with most classifiers exhibiting high accuracy rates, except for the Decision Tree classifier [5].

Arianna D'Ulizia, Maria Chiara Caschera, Fernando Ferri, and Patrizia Grifoni (2021) Conduct a Survey on Evaluation Datasets Used to Assess Fake News Detection Systems. They highlight the significance of fake news detection as an important task that has garnered significant attention. The survey provides insights into the evaluation data sets employed in this field [6].

Fatemeh Torabi Asr and Maite Taboada (2013) Emphasize the Challenges of Fake News Detection and The Requirement for Advanced Techniques From Natural Language Processing and Social Network Analysis. Their survey-based study underscores the importance of big data and quality data in detecting fake news and misinformation [7].

Muhammad Shahzad Faisal, Tahir Ahmad, Atif Rizwan, and Reem Alkanhel (2022) propose an efficient fake news detection mechanism using an enhanced deep learning model. It concludes that the model captures semantic and temporal information from news articles, achieving high accuracy in detecting fake news [8].

Jamal Abdul Nasir, Osama Subhani Khan, and Iraklis Varlamis (2021) present a hybrid CNN-RNN based deep learning approach for fake news detection. Their studydemonstrates that the proposed model outperforms existing models in terms of accuracy and efficiency [9].

Ihsan Ali, Mohamad Nizam Bin Ayub, Palaiahnakote Shivakumara, and Nurul Fazmidar Binti Mohd Noor (2022) survey fake news detection techniques on social media. Their study focuses on machine learning techniques and

highlights the importance of developing predictive models for cellular network service quality [10].

Rio Yunanto, Apriani Puti Pufini, and Angga Prabuwisesa (2020) propose a deep learning approach for fake news detection. Their study revolves around the development of a web-based information system using the waterfall model [11]. Another study investigates the detection of fake news on social media using the Scikit Learn data mining technique. The study emphasizes the impact of service quality, trust, and customer satisfaction on customer loyalty in the banking industry [12].

Gerardo Ernesto Rolong Agudelo, Octavio José Salcedo Parra, and Julio Barón Velandia (2018) propose a fake news detection model using machine learning in Python, specifically utilizing the Natural Language Toolkit (NLTK). The study concludes that the classification process is complex, but it can achieve a certainty level of over 95% [13].

Sajjad Ahmed, Knut Hinkelmann, and Flavio Corradini (2020) Develop a Fake News Model Using Machine Learning and NLP Techniques, including Naive Bayes, Support Vector Machines (SVM), Passive Aggressive, and Logistics Regression. It concludes that after comparing all methods, the study identifies Passive Aggressive as the most accurate, achieving a high accuracy rate of 93% [14].

Ray Oshikawa, Jing Qian, and William Yang Wang (2020) conduct a survey on NLP techniques for fake news detection. The survey concludes that NLP techniques are effective in detecting fake news, although their performance varies depending on the specific NLP technique used and the nature of the dataset. The survey identifies challenges such as the lack of standardized datasets, data bias, and difficulties in detecting subtle forms of fake news [15].

In conclusion, the literature review demonstrates that the detection of fake news is a challenging task that necessitates advanced techniques and methodologies. The studies reviewed highlight the significance of natural language processing and machine learning in fake news detection. Approaches such as transformer-based models, credibility measurement, and the utilization of various machine learning classifiers have shown promising results in identifying fake news accurately. However, challenges such as the unavailability of labeled data, complex language variations, and the efficiency of detection models remain areas that require further research and development. Future studies should focus on addressing these challenges to enhance the effectiveness of fake news detection methods and contribute to the prevention of the spread of false information.above the sub section while no space should be given below the heading and text

3. METHODOLOGY

3.1 Dataset

Our dataset is a dataset originating from kaggle.com with the title "Fake and real news dataset" which has two csv files namely True.csv and Fake.csv which will be processed through Google Colab by inputting a library that allows datasets from kaggle to be directly read by Google Colab so it can be directly accessed and there is no need to download it.

3.2 Preprocessing

In the Processing process the dataset will be changed to lower-case, then delete characters that are not letters or numbers, after that the text will be tokenized into words, then delete words which are Stop words, then change words to the basic form or lemmatization, then finally combines the words into text again.

3.3 Model/Algorithm

Based on the theory that has been presented previously, Figure 1 contains the stages of the research conducted. Where the modeling process is carried out respectively based on the five model selections that have been made, namely Naive Bayes, Random Forest, Decision Tree, Logistic Regression, and XGBoost.

Before the modeling process is carried out, data preprocessing is carried out first, which is to clean the resulting text data so that there are no capital letters, numbers, punctuation marks or hyperlinks.

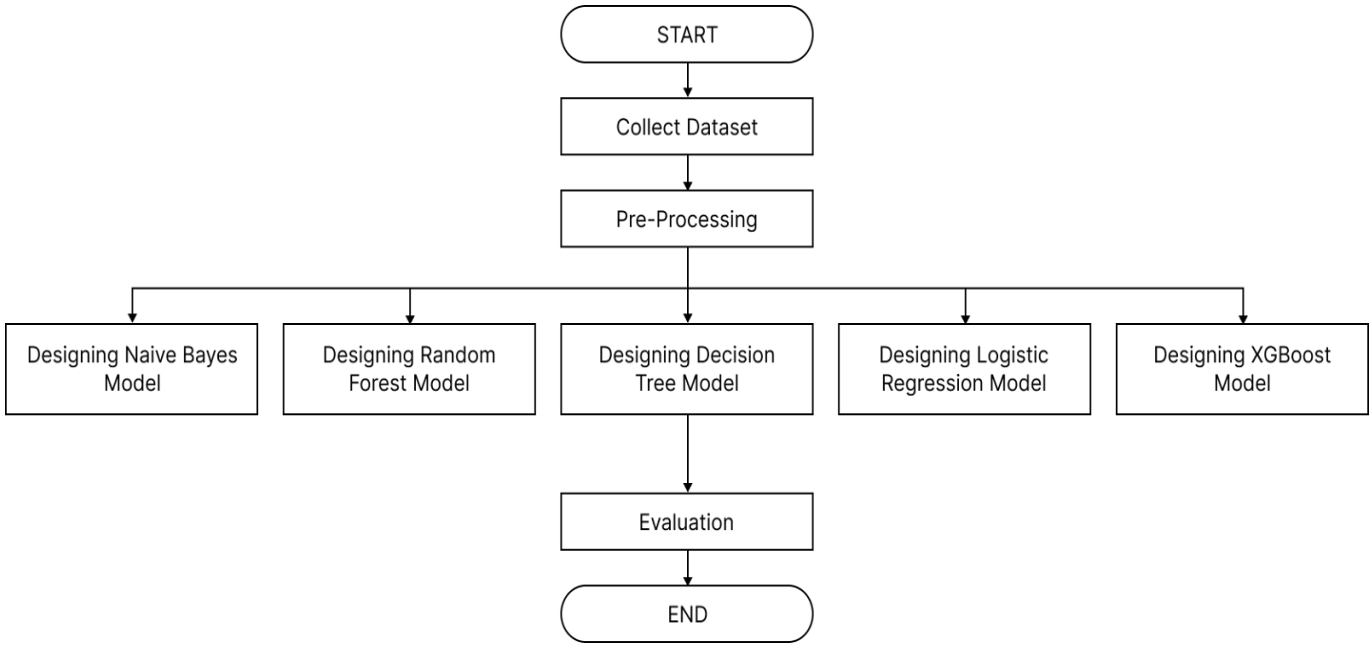

Figure 1. Model Algorithm Processing

3.3.1 Naïve Bayes

The Naive Bayes method is a classification technique based on the assumption of independence between predictors known as the Bayes theorem. The Naive Bayes classifier can assume that the presence of certain features in a class is unrelated to the presence of other features. This method is the most popular classification method used with a high degree of accuracy.

Naive Bayes is a method that can be used to detect fake news. The steps that need to be carried out can start with data collection, the tokenization process, the modeling phase, the evaluation phase, and the deployment phase. Based on the analysis of the model that has been created, we can conclude that this model is suitable for predicting fake news or facts with an F1-Score value that can be obtained with an accuracy of 81%.

3.3.2 Random Forest

Random Forest is a machine learning algorithm used for efficient fake news detection. By creating decision trees and combining their predictions, it classifies news articles as real or fake. To use Random Forest, we need a labeled dataset of news articles. We extract features and split the data into training and testing sets. The model is trained on the training set and evaluated using metrics like accuracy. To improve efficiency, we fine-tune parameters such as tree number and depth. The optimized model is then used to classify new articles as real or fake, automating the detection process. While no algorithm is perfect, Random Forest, when combined with other approaches, helps efficiently combat fake news.

3.3.3 Logistic Regression

Logistic Regression is a type of statistical analysis that is often used for predictive modeling. In this analytic approach, the dependent variable is finite or categorical, it can be either A or B (binary regression) or multiple choices such as A, B, C or D (multinomial regression). The Logistic Regression method can be used to detect hoax news by going through several stages, namely data collection, the tokenization process, the modeling phase, the evaluation phase, to the deployment phase. The Logistic Regression method can be used to detect fake news with an accuracy of up to 80%. In addition, this method can also be used to compare the accuracy of text properties with the Support Vector Machine (SVM) algorithm.

3.3.4 Decision Tree

Decision tree is one of the machine learning algorithms used for classification. Decision tree can be used to detect fake news by

using a classification model that can distinguish between fake news and genuine news. Decision trees can work by dividing data into smaller parts and then selecting the best features to divide the data. After that, the decision tree will divide the data into two parts based on these features and continue to do the same thing until it cannot divide the data anymore. In detecting fake news, a decision tree is used to build a classification model that can distinguish between fake news and genuine news.

3.3.5 XGBoost (Extreme Gradient Boosting)

When it comes to detecting fake news efficiently, the XGBoost (Extreme Gradient Boosting) method can be a valuable tool. XGBoost is a powerful machine learning algorithm commonly used for classifying text, including fake news. Fake news detection involves analyzing text to determine if it's real or fabricated. XGBoost is well-suited for this task because it can handle high-dimensional and sparse features found in text data. It has been successful in tasks like sentiment analysis and text classification. One of the advantages of XGBoost is its ability to handle complex relationships and interactions among features, which is crucial for detecting fake news. It creates a combination of weak prediction models called decision trees and uses their predictions to make a final decision. This boosting technique enhances the accuracy and reliability of the model.

4. RESULT AND DISCUSSION

Baased on the results of the experiments conducted, Table 1 contains the performance results of each classification model. evaluation is done by testing the performance of the model on the f1-score, accuracy, precision, and recall values. From the evaluation results, the overall model has good performance because the model can recognize patterns well in the data tested. but the most optimal performance is obtained using the XGBoost algorithm.

Table 1: Accuracy, Precision, Recall, F1-Score of all models

Algorithm / Model	Accuracy	Precision	Recall
Naïve Bayes	95,36%	96,10%	94,90%
Random Forest	99,12%	99,25%	99,07%
Logistic Regression	98,64%	98,79%	98,62%
Decision Tree	99,53%	99,50%	99,59%
XGBoost	99,76%	99,89%	99,66%

5. CONCLUSION

In conclusion, the results showed that certain algorithms like recurrent XGboost and Decision Tree machines performed well in detecting fake news. Developing high-quality fake news datasets presents a significant challenge due to the necessity of readily available data for training and evaluating algorithms that can identify false news. Based on studies we have read indicate that the spread of misinformation through social media has had significant consequences for both individuals and society at large. The studies reviewed highlight the significance of natural language processing and machine learning in fake news detection.

Approaches such as transformer-based models, credibility measurement, and the utilization of various machine learning classifiers have shown promising results in identifying fake news accurately. The research presented in this paper highlights the promising potential of NLP-based text classification methods for effectively and efficiently identifying fake news. The combination of advanced machine learning algorithms and additional features can enhance the accuracy of fake news detection systems. These findings contribute to the ongoing efforts to combat the spread of misinformation and improve the reliability of information in the digital age.

REFERENCES:

[1] Shaina Raza, Chen Ding. Fake News Detection Based On News Content And Social Contexts: A Transformer-Based Approach. (2022).
[2] Rakhmat Arianto, Spits Warnars Harco Leslie, Yaya Heryadi, Dan Edi Abdurachman, “Fake News Detection Model Based On Credibility Measurement For Indonesian Online News,” Natural Language Processing, 2021
[3] Aisyah Awalina, Jibran Fawaid, Rifky Yunus Krisnabayu, Dan Novanto Yudistira, “Indonesia’s Fake News Detection Using Transformer Network,” Natural Language Processing, 2021
[4] S. Pandey, S. Pabhakaran, N.V Subba Reddy And D. Acharya “Fake News Detection From Online Media Using Machine Learning Classifiers”, In Iop Conf., Vol. 2161,Pp.1742-6596, 2022, Doi: 10.1088/1742-6596/2161/1/012027.

[5] F.A Kulam Magdoom, Hari Narayanan N And A. Ramachandran, "Hoax Detection And Review Analysis Using Machine Learning". India: B.S.Abdur Rahman Crescent Institute Of Science And Technology, Department Of Computer Science And Engineering, 2022. [Online]. Available At Ssrn: Https://Ssrn.Com/Abstract=4119652 Or Http://Dx.Doi.Org/10.2139/Ssrn.4119652

[6] Arianna D'ulizia, Maria Chiara Caschera, Fernando Ferri, Patrizia Grifoni. Fake News Detection: A Survey Of Evaluation Datasets. (2021).

[7] Fatemeh Torabi Asr And Maite Taboada. Big Data And Quality Data For Fake News And Misinformation Detection. (2019).

[8] Muhammad Shahzad Faisal, Tahir Ahmad, Atif Rizwan, & Reem Alkanhel. Efficient Fake News Detection Mechanism Using Enhanced Deep Learning Model. (2022).

[9] Jamal Abdul Nasir, Osama Subhani Khan & Iraklis Varlamis. Fake News Detection: A Hybrid Cnn-Rnn Based Deep Learning Approach. (2021).

[10] Ihsan Ali, Mohamad Nizam Bin Ayub, Palaiahnakote Shivakumara, Dan Nurul Fazmidar Binti Mohd Noor, "Fake News Detection Techniques On Social Media: A Survey," Natural Language Processing, 2022

[11] Rio Yunanto, Apriani Puti Pufini, Angga Prabuwisesa, "Detect Fake News Using A Deep Learning Approach," Natural Language Processing, 2020

[12] Ir. Munawar Mmsi., M. Com, Phd, "Fake Detection System (Fake News) On Social Media Using The Scikit Learn Data Mining Technique", Data Analys, 2019

[13] G.E Rolong Agudelo, O.J. Salcedo Para And J.B. Velandia, "Raising A Model For Fake News Detection Using Machine Learning In Python". Et Al. Challenges And Opportunities In The Digital Era. I3e 2018. Lecture Notes In Computer Science(), Vol 11195. Pp 56-604, 2018, Springer, Cham. Https://Doi.Org/10.1007/978-3-030-02131-3_52.

[14] S. Ahmed, K. Hinkelmann And F. Corradini, "Development Of Fake News Model Using Machine Learning Through Natural Language Processing", Open Science Index, Vol. 14, No. 12, 2020. Doi: Https://Doi.Org/10.48550/Arxiv.2201.07489.

[15] R. Oshikawa, J. Qian And W.Y. Wang, "A Survey On Natural Processing For Fake News Detection", In Arxiv Access, Pp. 6086-6093, No. 2, 2020. Doi: Https://Doi.Org/10.48550/Arxiv.1811.00770.