Exploring Natural Language Processing for Fake News Detection and Classification: A Comparative Analysis of Naive Bayes, SVM and XGBoost

Paulo Henrique Lira Jr.
Campus Jaboatão dos Guararapes
Federal Institute of Pernambuco
Jaboatão dos Guararapes-PE, Brazil
phlj@discente.ifpe.edu.br

Luciano de Souza Cabral
Campus Jaboatão dos Guararapes
Federal Institute of Pernambuco
Jaboatão dos Guararapes-PE, Brazil
luciano.cabral@jaboatao.ifpe.edu.br
https://orcid.org/0000-0002-4235-5753

***Abstract*—In the digital age, the spread of fake news has emerged as a significant issue, demanding the need for reliable and scalable solutions to preserve the trustworthiness of news sources and safeguard public opinion. This study proposes an exploration of the application of Natural Language Processing (NLP) techniques for the detection and categorization of fake news. Our specific focus lies on three widely used machine learning algorithms: Naive Bayes, Support Vector Machines (SVM), and XGBoost. By conducting a comprehensive evaluation of these approaches, our objective is to assess their effectiveness in identifying fake news articles and contribute to the development of robust tools for countering misinformation.**

Keywords—machine, learning, natural language processing, Naive Bayes, SVM, XGBoost, fake news

I. Introduction

The pervasive impact of fake news extends across various domains, including public opinion, political processes, and social cohesion. With the exponential growth of digital platforms and social media networks, the dissemination of misinformation has become increasingly challenging to identify and address. Consequently, there is an urgent need for automated systems capable of effectively detecting and categorizing deceptive content to mitigate the potential harm caused by false information.

Natural Language Processing (NLP), a subfield of artificial intelligence, offers a robust framework for analyzing and comprehending human language. By harnessing NLP techniques, we can extract meaningful features and patterns from textual data, enabling the development of sophisticated models for the identification of fake news. This paper aims to explore the potential of NLP in combating the spread of fake news by examining the efficacy of three widely adopted machine learning algorithms: Naive Bayes, Support Vector Machines (SVM), and XGBoost. Through this investigation, we strive to contribute to the advancement of accurate and reliable tools for tackling the challenge of misinformation.

This research aims to achieve two primary objectives: firstly, to evaluate the effectiveness of Naive Bayes, Support Vector Machines (SVM), and XGBoost algorithms in detecting and classifying fake news articles, and secondly, to provide a comprehensive analysis of the strengths and weaknesses associated with each approach. Through a comparative analysis, we seek to identify the most suitable algorithm considering factors such as accuracy, efficiency, and interpretability.

To conduct this study, we utilize a diverse and representative dataset comprising labeled news articles, encompassing both genuine and fake news samples. Through meticulous preprocessing, feature engineering, and model training, we apply Naive Bayes, SVM, and XGBoost to classify news articles into either fake or genuine categories. By thoroughly evaluating the performance of each algorithm, we aim to develop a comprehensive understanding of their capabilities and limitations in the task of detecting fake news.

The subsequent sections of the paper are structured as follows: the State-of-the-Art section provides an extensive review of existing research on fake news detection. The Methodology section describes the employed methodology, including dataset description, preprocessing techniques, and feature extraction methods. The Results section outlines the experimental setup, evaluation metrics, and presents the findings of our comparative analysis. Lastly, the Conclusion summarizes the key insights obtained and suggests potential avenues for future research in this domain.

By investigating the efficacy of NLP techniques and comparing Naive Bayes, SVM, and XGBoost, this paper aims to make a significant contribution to the development of robust and accurate tools for the detection and classification of fake news.

II. Theoretical Framework

A. Fake News Detection

Naive Bayes: Naive Bayes classifiers are simple yet effective machine learning algorithms widely used for various classification tasks, including fake news detection. These classifiers are based on Bayes' theorem, assuming independence between features used for classification. Despite this assumption, Naive Bayes classifiers have demonstrated good performance in text classification domains such as fake news detection [1].

By leveraging the probabilistic relationships among features extracted from news articles, Naive Bayes classifiers calculate the likelihood of an article being real or fake based on observed feature frequencies in the training data. The classifier assigns a class label (real or fake) to new, unseen articles based on the highest likelihood [2]. In fact, even relatively simple artificial intelligence algorithms like Naive Bayes classifiers can yield promising results in addressing critical problems such as fake news classification [2]. In a study, the authors achieved a classification accuracy of approximately 74% on their test set.

The accuracy of Naive Bayes classifiers in fake news detection is typically evaluated by comparing the predicted labels with the true labels of a labeled dataset. This evaluation helps quantify the classifier's performance in accurately classifying news articles as real or fake.

Support Vector Machines (SVM): SVM is a supervised machine learning method primarily used for binary classification. It creates a decision boundary with maximum

margins, utilizing support vectors, which are critical points optimized to determine the hyperplane.

To differentiate between real and fake news articles, a study [3] proposes an SVM model that achieves 89% accuracy using only the title and 98% accuracy using the title and the first 1000 characters of the article. The study employed the "Fake vs Real News Dataset" created by Clement Bisaillon, consisting of 20,826 real news examples and 17,903 fake news examples. Each example included the title, text, subject matter, and publication date of the article.

The researchers conducted numerous experiments to optimize the accuracy of their SVM model, exploring different kernel types (Linear, Radial Basis Function, Polynomial, and Sigmoid). They found that their SVM model achieved an accuracy of 89.60% when classifying news articles as real or fake based on the headline alone. By including the first 1000 characters of text, the accuracy improved to 98.07% [3].

XGBoost: XGBoost is a gradient boosted decision tree implementation known for its speed and performance. It is widely used for both regression and classification problems, making it a suitable choice for our fake news classification task [4].

In a paper [4], the authors applied supervised machine learning using XGBoost to classify Indonesian news articles as either hoaxes or valid data. They collected news from various Indonesian websites such as kompas.com, detik.com, cnnindonesia.com, liputan6.com, and turnbackhoax.id, spanning the period of 2015-2020. The dataset consisted of 500 news articles, which were divided into training data (80%) and test data (20%) for model performance evaluation.

Through parameter tuning, the authors achieved high accuracy in hoax text classification for Indonesian news using XGBoost. Their model achieved an accuracy of 92% with an 80:20 split of the dataset, where 80% of the data was used for training and 20% for testing [4].

In conclusion, these studies demonstrate the effectiveness of Naive Bayes, SVM, and XGBoost in detecting and classifying fake news articles. By leveraging these algorithms, researchers have achieved notable accuracy rates, contributing to the development of reliable tools for fake news detection and classification.

III. METHODOLOGY

In this section, a detailed description of the dataset used will be provided, along with an explanation of the methods applied in the analyses conducted.

A. Dataset

Similar to [3], we utilized the "Fake vs Real News Dataset" created by Clement Bisaillon for our research. The dataset contains 23481 rows × 5 columns (title, text, subject, date, flag).

Table I. Dataset Info after concatenate and duplicates removal.

#	Column	Non-Null Count	Dtype
0	title	44689	object
1	text	44689	object
2	subject	44689	object
3	date	44679	datetime64[ns]
4	flag	44689	int64

Some relevant aspects of the experiment can be visualized in Google Colaboratory Script[1]. The following figure presents the dynamics of fake news over time.

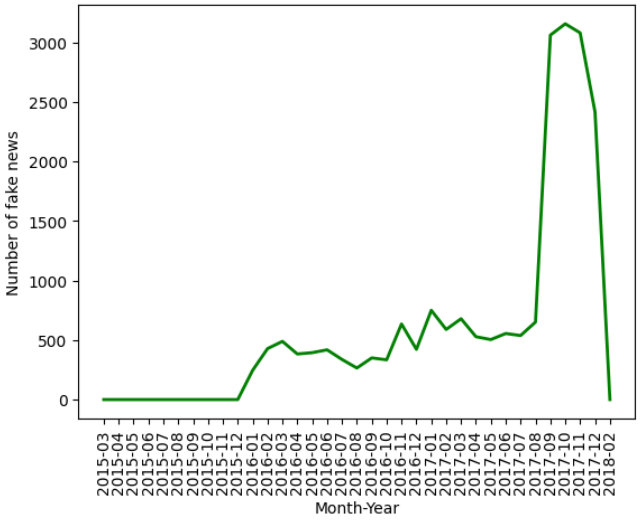

Fig. 1. Dynamics of fake news.

The peak of the dataset coincides with the period of the last US presidential election. This visualization is confirmed using a histogram per category.

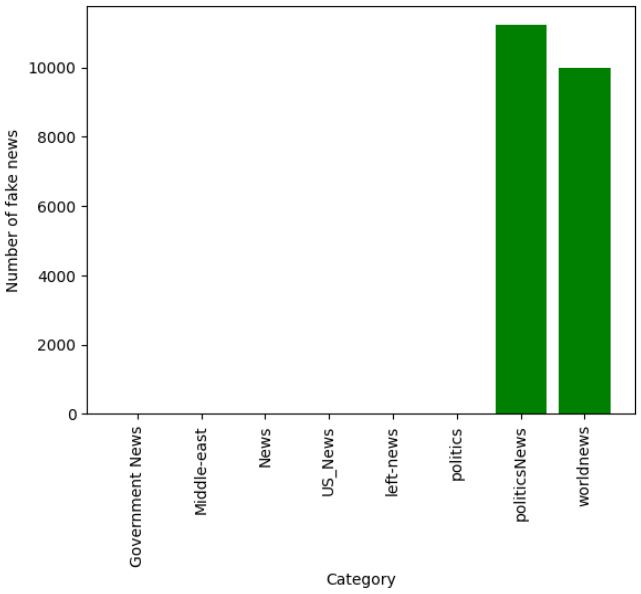

Fig. 2. Fake news categories.

The majority of the dataset instances comprehend politics and world news.

B. Preprocessing

During the preprocessing stage of the fake and real news articles, several steps were performed to transform the raw text into a suitable format for classification. The following procedures were followed:

1. Combining Titles and Text:

The titles and text of both fake and real news articles were merged into a single file. This consolidation enabled a comprehensive analysis of the entire content.

2. Selecting Relevant Text:

To capture the essence of each article while considering computational constraints, the first 1000 characters of the text were chosen. This selection provided a concise representation of the article.

[1] https://acesse.one/fVENv

3. Label Creation:

A separate file was created with numeric labels corresponding to each article. A value of 1 was assigned if the article was identified as real and 0 if it was identified as fake. These labels facilitated the supervised learning process, allowing the model to learn from labeled examples.

4. Lowercasing and Tokenization:

The titles and text were transformed to lowercase characters to ensure consistency and avoid discrepancies due to case sensitivity. The text was then tokenized, breaking it down into individual words or tokens. This step formed the foundation for further text processing.

5. Punctuation Removal:

Non-alphanumeric characters, such as punctuation marks, were removed from the text. By eliminating punctuation, the focus shifted to essential words and reduced noise in the data.

6. Stop Word Removal:

Stop words, such as "and," "the," or "is," which carry little semantic meaning, were removed from the text. These words are typically common across various articles and may not significantly contribute to the classification task.

Fig. 3. Word cloud visualization after 1-6 pre-processing stages.

7. Lemmatization:

Each word in the text was lemmatized, reducing it to its base or root form. This process helped unify similar words and reduced the dimensionality of the feature space. For example, words like "running," "runs," and "ran" would be transformed to their common root form, "run."

8. Word Vectorization and N-grams:

To represent the text data as numerical features, a word vectorizer was employed. Specifically, N-grams, particularly bigrams (pairs of consecutive words), were used to capture contextual relationships between words. This approach allowed the model to consider not only individual words but also co-occurrence patterns within the text.

V. *Results and Discussion*

To carry out the work, some classic models available in the Sklearn library [5] were chosen, using Python as a programming language in notebooks at Google Collaboratory.

After completing the pre-processing procedures, during the analysis of the methods, unexpected results were obtained. The Naive Bayes model showed an accuracy of approximately 98%, while the SVM and XGBoost models achieved a perfect accuracy of 100%. These findings indicate a potential error in the data analysis and treatment process, as such high accuracy levels are uncommon in practice.

Therefore, it is necessary to conduct a more detailed examination of the data and investigate the potential causes of these inflated accuracies. The presence of data leakage, overfitting, or other anomalies must be thoroughly examined to ensure the validity of the results.

It is evident that a more comprehensive approach is required to draw conclusive insights regarding the effectiveness of the Naive Bayes, SVM, and XGBoost models for fake news detection. Further investigations and refinements in the methodology are necessary to address the discrepancies and obtain reliable and meaningful results.

Table II. Test accuracy summary.

Dataset	Teste
Rede Neural	Acurácia Top-5 (%)
LinearSVC	1.0
MultinomialNB	0.978776783 2926498
'XGB'	1.0

A. *Limitations and challenges*

Despite the advances achieved, there are some limitations and challenges to be considered. The availability of more balanced and representative databases (with data more distributed among the classes), the interpretability of the decisions made by the algorithms, and the generalization to different vectorization models (BERT for example) are aspects that require additional attention and enhance the possibility of future work.

VI. *Conclusion*

The use of machine learning models for the detection of fake news shows promise, with encouraging results in terms of sensitivity and specificity.

In the future, it is important to continue improving detection techniques with more comprehensive databases, involving different domains and greater distribution among categories.

Exploring transformers models instead of vectorization using TF-IDF can be a factor for future analysis, correlating the results with the type of vectorization. In addition, it is essential to collaborate with professionals from other institutions to validate and incorporate these approaches in more realistic environments, helping to improve the detection of untrue news, reducing any harmful power, if any.

Acknowledgments

We would like to express our gratitude to the Federal Institute of Pernambuco (IFPE) - Jaboatão dos Guararapes Campus for the support in providing the necessary infrastructure to carry out the research and for the partial financial support for the event.

References

[1] R.J., P. et al. (2019) 'International Journal of Recent Technology and Engineering (IJRTE)', Fake News Accuracy using Naive Bayes Classifier, 8(1C2), pp. 962–964.

[2] Granik, M. and Mesyura, V. (2017) 'Fake news detection using naive Bayes classifier', 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), pp. 900–903.

[3] Altman, B. et al. (2021) DETECTING FAKE NEWS USING SUPPORT VECTOR MACHINES, pp. 1–5.

[4] Haumahu, J.P., Permana, S.D.H. and Yaddarabullah, Y. (2021) 'Fake news classification for Indonesian news using Extreme Gradient Boosting (XGBoost)', IOP Conference Series: Materials Science and Engineering [Preprint].

[5] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.