Fake News Detection using Classification Techniques

BITS Pilani Hyderabad Campus

CS F415 Data Mining Project

By

Achyut Dedania (F20212807@hyderabad.bits-pilani.ac.in)

Teerth Patel (F20212090@hyderabad.bits-pilani.ac.in)

Vansh Rastogi (F20210407@hyderabad.bits-pilani.ac.in)

Abstract: The proliferation of misinformation, colloquially referred to as "fake news," across social media platforms and other media outlets presents a substantial and pressing concern, given its potential to instigate significant societal and national repercussions. Considerable scholarly attention has been directed towards the detection of such misinformation. This paper undertakes an examination of existing research endeavors pertaining to the identification of fake news, with a particular focus on traditional machine learning methodologies. The aim is to discern the most efficacious approach, culminating in the development of a supervised machine learning model capable of discerning the veracity of news articles. Leveraging tools such as Python's scikit-learn library and natural language processing (NLP) techniques for textual analysis, the proposed model will undertake feature extraction and vectorization processes. Specifically, we advocate for the utilization of scikit-learn's functionalities for tokenization and feature extraction, harnessing tools like CountVectorizer and TfidfVectorizer for this purpose. Then we performed GridSearchCV technique for finding the optimal parameter values from a given set of parameters. Through this methodological framework, we endeavor to construct a robust classifier capable of distinguishing between genuine and fabricated news items. By plotting the learning curve for each algorithm we found that Random Forest outperformed the rest.

Keywords: Misinformation, Natural Language Processing, Supervised Classification, Scikit-learn

I. INTRODUCTION

The problem at hand involves the pervasive dissemination of fake news via social media platforms, necessitating the development of effective detection mechanisms. The primary objective of this research is to investigate the efficacy of machine learning algorithms, namely Naïve Bayes classifier, Logistic Regression, Random Forest, Decision Tree, and Support Vector Machine in discerning fake news from genuine ones. This study seeks to evaluate the performance of these algorithms using a manually labeled dataset, and assess their potential utility in addressing the challenge of fake news detection.

The phenomenon of fake news proliferating through online channels presents significant societal and informational challenges. As individuals increasingly rely on social media for news consumption, the unchecked dissemination of misinformation threatens the integrity of public discourse and decision-making processes [1]. Given the detrimental impact of fake news on individual beliefs, societal cohesion, and democratic processes, the development of reliable detection mechanisms assumes paramount importance. This research endeavors to contribute to mitigating the deleterious effects of fake news by exploring the potential of machine learning algorithms in identifying and combating misinformation.

Detecting fake news presents a formidable challenge due to several inherent complexities. Unlike traditional journalism, social media platforms offer a rapid and cost-effective means of news dissemination, facilitating the swift propagation of misinformation. Furthermore, the nuanced nature of fake news necessitates discerning subtle cues and patterns that distinguish it from genuine content. Naive approaches often falter in this endeavor due to their inability to account for the multifaceted and dynamic characteristics of fake news, underscoring the need for more sophisticated computational techniques.

Prior attempts to address the fake news problem have encountered various limitations. While some studies have explored deception detection techniques using machine learning algorithms, their efficacy remains constrained by factors such as dataset quality and algorithmic complexity. Previous solutions have struggled to achieve high levels of accuracy and generalizability, impeding their practical utility in real-world contexts. Additionally, the evolving nature of fake news presents an ongoing challenge, as new tactics and strategies continually emerge, necessitating adaptive and robust detection mechanisms.

The key components of our approach encompass the utilization of various machine learning algorithms, including Naïve Bayes, Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM). Each algorithm is deployed to discern fake news from genuine content, leveraging their unique methodologies for classification. Additionally, we employ the GridSearchCV technique for parameter tuning, optimizing the performance of each algorithm. Our results indicate that the accuracy rates, in descending order, are as follows: Random Forest outperformed Decision Tree, Logistic Regression, SVM, and Naïve Bayes. Through this comprehensive approach, we aim to provide insights into the comparative performance of different machine learning techniques in tackling the challenge of fake news detection.

II. RELATED WORK

Mykhailo Granik et al. introduced a simple method for detecting fake news using a naive Bayes classifier [2]. They tested this approach on Facebook news posts from various ideological perspectives, as well as mainstream political news pages like Politico, CNN, and ABC News. The classification accuracy reached approximately 74%, although detecting fake news was slightly less accurate, possibly due to the dataset's skewed distribution, with only 4.9% of instances being fake news.

In their study, Himank Gupta et al. [3] proposed a comprehensive framework employing various machine learning approaches to tackle multiple challenges, including accuracy limitations and time delays. They collected a dataset of 400,000 tweets from HSpam14, categorized into 150,000 spam tweets and 250,000 non-spam tweets. Additionally, they identified lightweight features and the top 30 words with the highest information gain using a Bag-of-Words model. Through their framework, Gupta et al. achieved an impressive accuracy of 91.65%, surpassing existing solutions by approximately 18%.

In the study conducted by [4] [5], the utilization of fake news during the 2012 Dutch elections on Twitter was examined. The study evaluated the performance of eight supervised machine learning classifiers using a Twitter dataset [6]. It was determined that the decision tree algorithm yielded the highest performance for the dataset, achieving an F1-score of 88%. The dataset comprised 613,033 tweets, with 328,897 classified as genuine and 284,136 as false. Through qualitative analysis of false tweets disseminated during the election, features and characteristics of misinformation were identified and categorized into six distinct groups [7].

[8] introduced a counterfeit detection model utilizing N-gram analysis alongside diverse characteristic extraction techniques. Additionally, the study investigated various feature extraction techniques and six different machine learning methods. The proposed model achieves its highest accuracy using a combination of unigram and linear SVM, reaching an accuracy of 92%.

Shivam B. Parikh et al. [9] aim to understand the characterization of news stories in modern society, exploring different content types and their impact on readers. The paper discusses methods for detecting fake news, primarily through text-based analyses, and highlights key datasets used in this field. Additionally, the authors outline four research challenges for future consideration. This theoretical approach examines psychological factors in fake news detection, providing illustrative insights into the process.

III. METHODOLOGY

The primary objective of our work is to address the pervasive issue of identifying fake news articles by implementing a robust classification model. To tackle this challenge effectively, we rely on a methodology grounded in supervised classification machine learning techniques. Our approach begins with the critical phase of dataset collection, drawing from reputable sources such as GFG and Kaggle, which provide diverse sets of news articles for analysis. Following dataset acquisition, we embark on a comprehensive pre-processing stage to ensure data quality and consistency. This involves tasks such as text normalization, cleaning, and tokenization to prepare the dataset for further analysis. Additionally, we employ advanced feature selection methods to extract relevant information and optimize model performance. Once the dataset is refined, we proceed with the splitting of dataset into training and testing set . Finally we utilize a range of classification algorithms to develop a robust model capable of discerning between genuine and fake news articles.

In our research endeavour, we employed a dataset comprising six distinct columns, namely index, title, text, subject, date, and class. This dataset encompassed a total of 44,919 rows, each representing a unique article. However, during the initial exploratory analysis, it was identified that 21 rows contained missing values across certain columns. Consequently, to maintain data integrity and ensure robust analysis, these rows were removed from the dataset. Following data cleansing, it was observed that approximately 23,000 rows were associated with genuine news articles, while the remaining 21,000 rows pertained to articles deemed as fake news. Moreover, the balanced distribution between genuine and fake news instances within the dataset provided an equitable basis for model training and evaluation.

After splitting the data into training and testing sets, we proceeded to apply Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, and Support Vector Machine algorithms. Logistic Regression is employed for binary classification tasks, modeling the probability of an event using a logit function and multiple independent variables. Naïve Bayes operates on probabilistic principles, assuming strong independence between features to calculate the probability of an outcome given an event occurrence. Decision Tree analysis involves classifying data by constructing a flowchart where each internal node represents a test condition, branches denote outcomes, and leaf nodes indicate class labels. Random Forest, an ensemble learning technique, concurrently constructs multiple trees for classification or regression tasks. Support Vector Machine (SVM) is utilized to potentially enhance model accuracies by creating an optimal decision boundary, or hyperplane, in n-dimensional space to segregate classes effectively. SVM identifies support vectors, extreme points aiding in hyperplane creation, thereby optimizing classification performance.

Building upon prior research indicating suboptimal accuracies with similar algorithms, our study sought to improve performance by employing different preprocessing techniques and optimizing hyperparameters. Through systematic exploration using the GridSearchCV technique, we aimed to enhance the accuracy of our algorithms.

IV. EXPERIMENT

i. Dataset

We undertook several data preprocessing steps to facilitate efficient model training and enhance accuracy. Initially, we removed redundant columns such as index, date, and title, the latter deemed redundant as it was a subset of the text. Subsequently, we merged the text and subject columns, eliminating the latter to streamline the dataset. Identification of 21 rows containing missing values prompted their removal to uphold data integrity. We then employed stratified sampling to expedite training while preserving dataset representativeness. Standardization procedures involved converting all words to lowercase, removing special characters and punctuation, as well as eliminating links from the text. Tokenization was applied to split the text column into individual words, followed

by lemmatization and stop word removal to refine textual data. Visual representation of word frequency via word cloud analysis provided insight into data characteristics. Subsequently, we transformed the textual data into numerical vectors using TF-IDF and Bag-of-Words (BOW) techniques. Principal Component Analysis (PCA) was utilized to assess the separability of these techniques, informing further processing decisions. The resultant processed data comprised 1000 numerical columns and one target column for subsequent analysis.

ii. Evaluation method/ Metrics

The evaluation of proposed methodologies entails the utilization of various performance metrics, including accuracy, precision, recall, F1-score, and support. These metrics are comprehensively integrated into a classification report generated using the Sci-kit learn library. Accuracy represents the proportion of correctly classified instances among the total number of instances, providing an overall measure of model correctness. Precision measures the proportion of true positive predictions among all positive predictions, indicating the model's ability to avoid false positives. Recall, also known as sensitivity, assesses the proportion of true positive predictions among all actual positive instances, highlighting the model's capacity to identify relevant instances. The F1-score, a harmonic mean of precision and recall, offers a balanced measure of a model's performance, especially in scenarios with imbalanced class distributions. Lastly, support denotes the number of occurrences of each class in the dataset, providing insight into class imbalance issues. Additionally, learning curves have been constructed for each algorithm across different datasets, notably training and testing sets. These learning curves offer valuable insights into the variance and bias present in the data. Analysis of these curves aids in understanding whether the model exhibits overfitting (high variance) or underfitting (high bias), thus guiding further optimization efforts to enhance model robustness and predictive accuracy.

Total	**Class 1 (Predicted)**	**Class 2 (Predicted)**
Class 1 (Actual)	TP	FN
Class 2 (Actual)	FP	TN

Table 1: Confusion Matrix

By formulating this as a classification problem we can define the following metrices:

$$\text{Precision} = \frac{TP}{TP+FP} \qquad \text{F1-Score} = 2\,\frac{Precision*Recall}{Precision+Recall}$$

$$\text{Recall} = \frac{TP}{TP+FN} \qquad \text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

iii. Experimental Setup

After preprocessing the final dataset, we partitioned it into training, testing, and validation sets in an 80:10:10 ratio. Subsequently, we employed various classification algorithms, including Naïve Bayes, Logistic Regression, Random Forest, Decision Trees, and Support Vector Machine. For each algorithm, we implemented 3-fold cross-validation and hyperparameter tuning using the GridSearchCV technique available in the Sci-kit learn library.

For Naïve Bayes, we optimized the hyperparameter var_smoothing (=1e-9), which introduces a user-defined value to the distribution's variance. For Logistic Regression, hyperparameters C (=10) and tol (=0.0001) were tuned, where "C" regulates the strength of regularization, and "tol" denotes the tolerance for the stopping criteria. In Decision Trees, hyperparameters criterion (=gini index) and max-depth (=10) were adjusted. "Criterion" determines the metric for selecting the feature as a splitting node, while "max-depth" defines the maximum depth allowed for a tree. Similarly, for Decision Trees, with criterion set to entropy and max-depth to 50, hyperparameters were tuned. Support Vector Machine hyperparameters were optimized by adjusting C (=1) and kernel (=linear). Here, "C" mirrors its role in logistic regression, and "kernel" refers to the function utilized for transforming input data into a higher-dimensional space.

V. RESULTS

In this section, we present the evaluation metrics and performance analysis of our proposed models. We employ various evaluation metrics including accuracy, F1 score, precision, and recall to comprehensively assess the efficacy of our models. The results are summarized in Table 2.

Models	Validation Accuracy	Test Accuracy	F1-Score
Naïve Bayes	96.88%	96.43%	0.97
Logistic Regression	99.33%	97.55%	0.98
Decision Tree	99.77%	99.55%	1.00
Random Forest	99.97%	99.77%	1.00
SVM	99.55%	97.77%	0.98

Table 2: Performance metrics of proposed models.

Furthermore, the learning curves of the proposed models are illustrated from Figure 1-5, showcasing their convergence behavior and performance trends over the course of training.

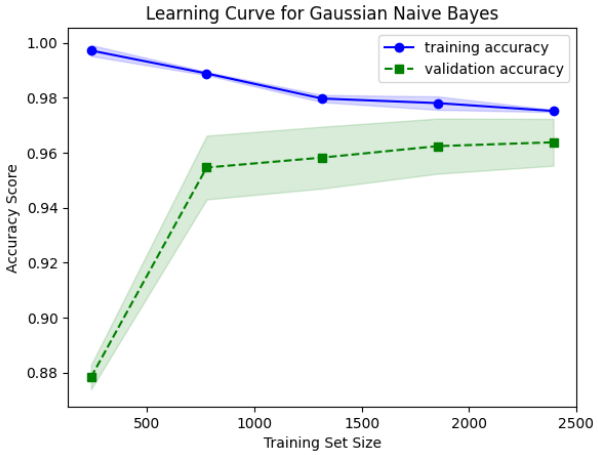

Figure 1

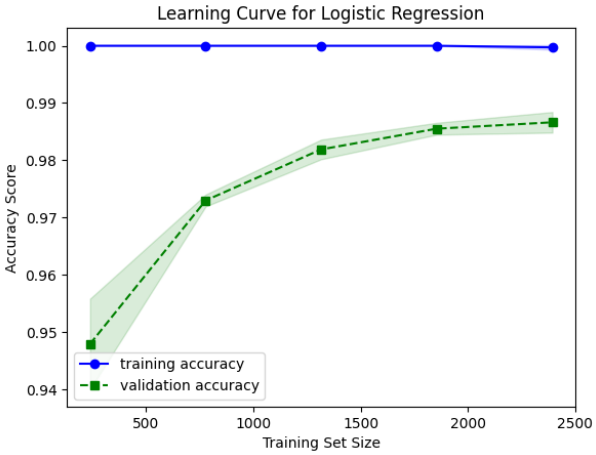

Figure 2

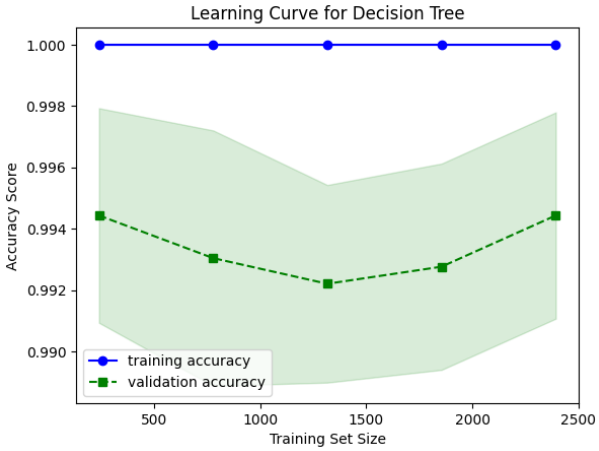

Figure 3

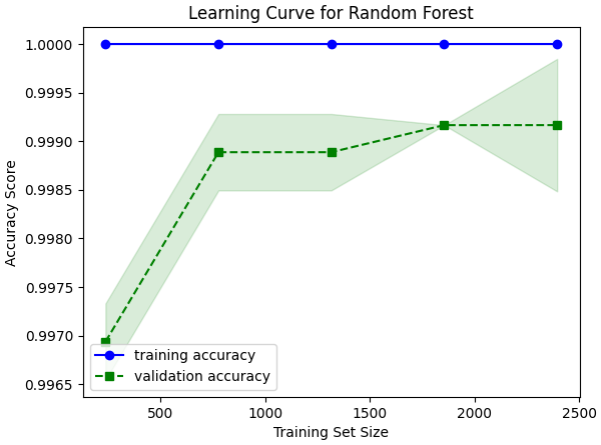

Figure 4

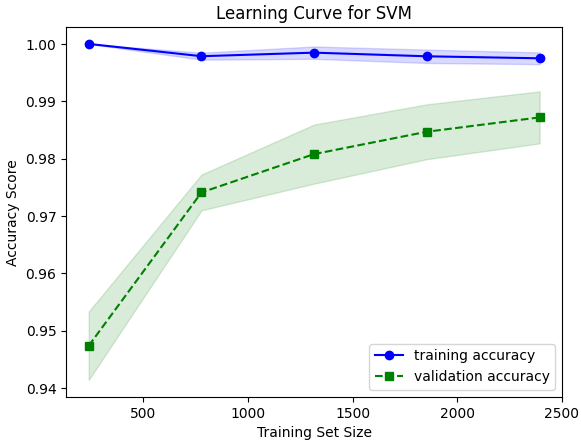

Figure 5

Our experiments reveal significant improvements over existing approaches, particularly highlighted by the substantial increase in accuracy. For instance, while the highest accuracy reported in related work was 92% for SVM, our proposed SVM model achieved a remarkable accuracy of 97.77%, signifying a notable enhancement in predictive performance. Moreover, the proposed random forest (RF) model exhibited outstanding accuracy, reaching 99.77%. These results underscore the effectiveness of our proposed methodologies in surpassing the state-of-the-art performance benchmarks. Additionally, we assessed the generalization capability of our models by evaluating their performance on unseen data. As depicted in Table 1, the models demonstrated robust performance on the testing dataset, reaffirming their efficacy in real-world applications.

VI. CONCLUSION

In today's digital age, the landscape of news consumption has shifted significantly, with traditional print newspapers being gradually replaced by online platforms such as social media networks, news websites, and messaging applications like WhatsApp. However, alongside this transition, there has been a concerning rise in the dissemination of fake news – misinformation deliberately crafted to manipulate public opinion and attitudes towards digital technology.

To address this challenge, we have developed a Fake News Detection system that aims to discern between genuine and fabricated news articles. Leveraging a combination of Natural Language Processing (NLP) and Machine Learning techniques, our system analyzes textual data extracted from news headlines and articles. Through rigorous training on curated datasets and comprehensive performance evaluations using various metrics, we strive to identify the most effective approach for distinguishing between authentic and deceptive news content.

Initial findings from our research indicate promising results, particularly with the implementation of the Random Forest model, which achieved an impressive accuracy rate of 99.77% following parameter tuning using GridSearchCV. Looking ahead, we are committed to further enhancing the capabilities of our system by integrating advanced neural network architectures and expanding its scope to analyze visual and audio news formats. Additionally, we aim to establish and maintain a dynamic dataset, continually updated to reflect the latest news developments, thereby ensuring the ongoing relevance and accuracy of our Fake News Detection system in combating misinformation on digital platforms. Through these endeavors, we endeavor to contribute to the safeguarding of informed discourse and the preservation of trust in online information sources.

VII. References

[1] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, Huan Liu, "Fake News Detection on Social Media: A Data Mining Perspective," 3 Sep 2017.

[2] M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," in *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, 2017.

[3] H. Gupta, M. S. Jamal, S. Madisetty and M. S. Desarkar, "A framework for real-time spam detection in Twitter," in *2018 10th International Conference on Communication Systems & Networks (COMSNETS)*, Bengaluru, 2018.

[4] I. Traore, "Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques.," in *International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, 2017.

[5] Khanam Z., Alkhaldi S., "An Intelligent Recommendation Engine for Selecting the University for Graduate Courses in KSA: SARS Student Admission Recommender System.," in *Smys S., Bestak R., Rocha Á. (eds) Inventive Computation Technologies*, vol. 98, Springer, Cham., 2019.

[6] Khanam Z. and Ahsan M.N., "Implementation of the pHash algorithm for face recognition in secured remote online examination system.," *International Journal of Advances in Scientific Research and Engineering*, vol. 4, no. 11, November 2018.

[7] Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., & Liu, Y., "Combating fake news: A survey on identification and mitigation techniques. ACM Transactions on Intelligent Systems and Technology," 2019.

[8] M. S. Looijenga, "The Detection of Fake Messages using Machine Learning.," in *29 Twente Student Conference on IT*, Enschede, The Netherlands, Jun. 6th, 2018.

[9] S. B. Parikh and P. K. Atrey, "Media-Rich Fake News Detection: A Survey," in *2018 IEEE Conference on Multimedia Information Processing and Retrieval*, Miami, FL, 2018.