Fake News Detection using Python and Machine Learning

Jumana Jouhar, Neharin Tijo, Meenakshi Mony, Dr. Anju Pratap
Department of Computer Science and Engineering
Saintgits College of Engineering, Kerala

July 14, 2023

Abstract

This report tackles the critical issue of false information proliferation in the digital era by employing machine learning and Python for fake news detection. Through rigorous evaluation, this research aims to significantly enhance information credibility and combats the growing threat of misinformation.

1 Introduction

The uncontrolled spread of false information in the digital age poses an alarming threat to society. To address this issue, this study utilizes machine learning algorithms and Python for fake news detection. By exploring and evaluating various machine learning models, this study aims to empower decision-making, fortifies information integrity, and mitigates the adverse impacts of misinformation. Its implications extend to bolstering trust in media, democratic processes, and content authenticity, benefitting news organizations, social media platforms, and government entities alike.

2 Literature Survey and Related Works

A comprehensive literature review was conducted to understand the concept of fake news detection using Python and machine learning.

In related research, Sharma et al. [1] developed a Fake News Detection system using NLP and Machine Learning techniques, exploring classifiers like Logistic Regression, Random Forest, and Passive Aggressive Classifier. Khanam et al. [2] employed XGBoost, while Pandey et al. [3] achieved high accuracy and F1-scores using Logistic Regression, Decision Tree, and other classifiers for news classification.

These findings offered insights into various classifiers and their performance metrics. This knowledge was used to implement fake news detection using logistic regression, decision tree, random forest, gradient boosting, XGBoost, and passive aggressive classifiers. The performance of the models were evaluated using accuracy, precision, and F1-score, along with other relevant metrics to ensure a comprehensive assessment.

Figure 1: Related works based on research papers on fake news detection

Author(s)	**Classifiers**	**Model Performance**			**Issue date**	**Reference**
		Accuracy	**Precision**	**F1-Score**		
Uma Sharma, Sidarth Saran, Shankar M. Patil	Naïve Bayes	0.60	0.59	0.72	2020	[1]
	Random Forest	0.59	0.62	0.67		
	Logistic Regression	0.65	0.69	0.75		
	PAC	0.9273	0.93	0.9257		

Z Khanam, B N Alwasel, H Sirafi and M Rashid	XGBOOST	> 75%	NA	NA	2020	[2]
	SVM	Approx 73%	NA	NA		
	Random Forest	Approx 73%	NA	NA		
Shalini Pandey, Sankeerthi Prabhakaran , N V Subba Reddy and Dinesh Acharya	KNN	89.98%	NA	NA	2021	[3]
	Logistic Regression	90.46%	NA	0.92		
	Naïve Bayes	86.89%	NA	0.87		
	Decision Tree	73.33%	NA	0.73		
	SVM	89.33%	NA	0.92		

Note: Accuracy – the ratio of correctly predicted observations to the total of observations.
F1-Score – performance metrics.

3 Contributions

The exploration of various models and optimization techniques, along with analysis and findings, has yielded valuable insights that contribute to the fight against misinformation.

4 Dataset Selection and Exploratory Data Analysis (EDA)

The ISOT Fake News Dataset was chosen for its diverse topics and balanced composition of articles from reputable and unreliable sources. The dataset includes dimensions such as title, text, type, and publication date. Through Exploratory Data Analysis (EDA) techniques, insights were gained into the dataset. Visualizations, such as bar graphs, were used to examine subject distribution, identify key topics, and potential features for distinguishing between true and fake news articles. It was checked that there was no data imbalance to ensure a fair and unbiased training dataset. Preprocessing steps, such as removing numbers, punctuation, and stopwords, were applied to improve the quality of the text data for higher accuracy.

News	**Size (Number of articles)**	***Subjects***	
Real-News	21417	**Type**	**Articles size**
		World-News	*10145*
		Politics-News	*11272*
Fake-News	23481	**Type**	**Articles size**
		Government-News	*1570*
		Middle-east	*778*
		US News	*783*
		left-news	*4459*
		politics	*6841*
		News	*9050*

Figure 2: Number of Articles Per Category

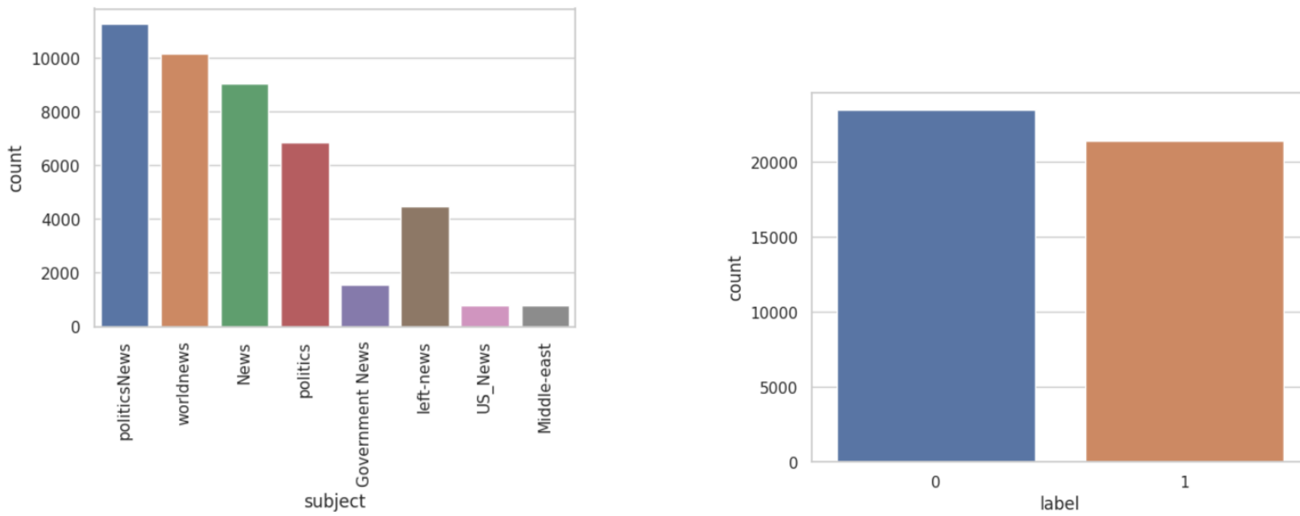


Figure 3: a) Visualization Based on Subject Column b) Comparison of Number of True and Fake Articles

5 Metric and Model Selection

The chosen metrics were: 1. Accuracy: Measures overall correctness of each model's predictions, which is crucial in determining the reliability of news classification. Computed as (TP + TN). 2. Precision: Ensures that the model accurately identifies true positives (real news) and minimizes false positives (classifying fake news as real), as avoiding the spread of false information is critical. Computed as TP / (TP + FP). 3. Recall: It is a measure of how well our model correctly classifies real news articles and avoids false negatives (misclassifying real news as fake), which is important for maintaining the credibility of legitimate news sources. Computed as TP / (TP + FN). 4. F1 Score: Provides a balanced measure, considering both false positives and false negatives, making it valuable in situations where both types of errors are equally important. 5. AUC-ROC Score: Evaluates the model's ability to distinguish between true and fake news articles by ranking true positives higher than false positives. 6. Confusion Matrix: Provides a visual representation of the model's performance, showing the counts of true positives(TP), true negatives(TN), false positives(FP), and false negatives(FN). It helps to analyze the model's classification performance across different categories.

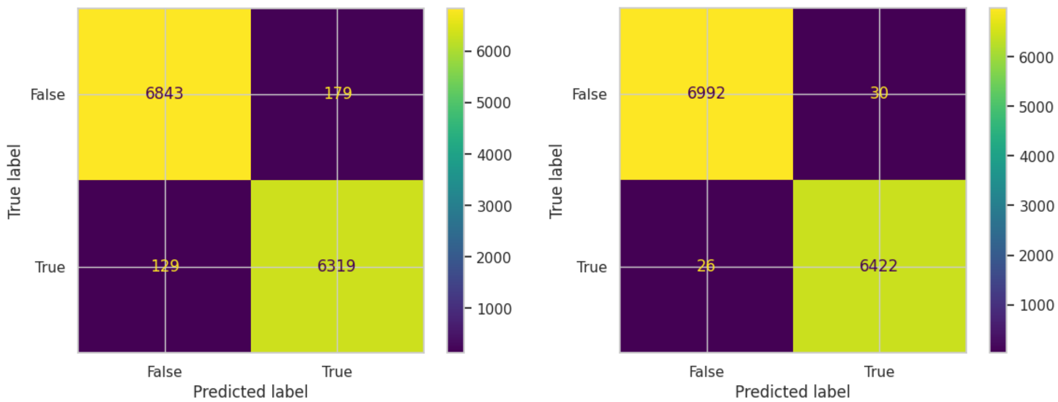
The models chosen were logistic regression, decision tree, random forest, gradient boosting, XGBoost, and passive aggressive classifier.Logistic regression was selected as the baseline model due to its interpretability and suitability for binary classification tasks. Intel Extension for Scikit-learn was used with logistic regression to shorten code runtime. The subsequent models were chosen as they offered advantages such as capturing non-linear relationships, handling complex interactions, and improving predictive performance. By utilizing a diverse set of models, we ensure a robust analysis of fake news detection.

To encode the features, the TF-IDF vectorization technique, which converts text data into numerical representations, was used. It captures the importance of words in each document and helps to improve model performance.

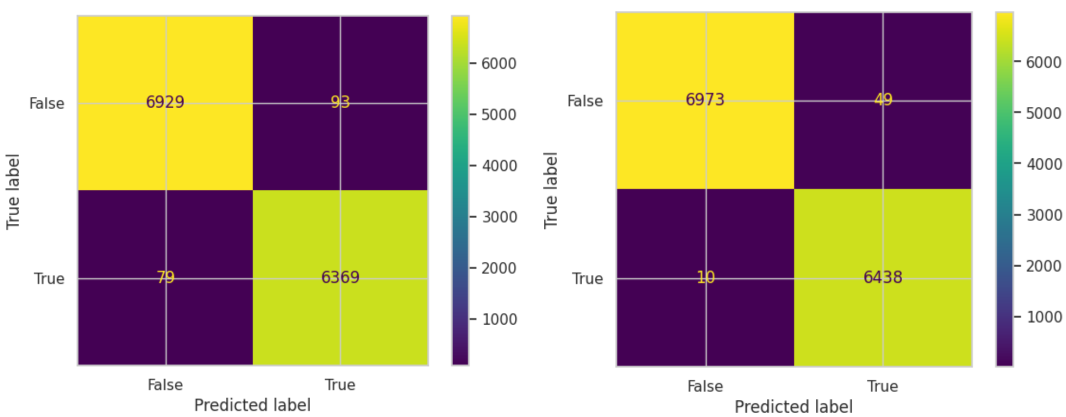
6 Model Evaluation

Contrary to the literature review, the decision tree, gradient boosting and XGBoost outperformed the expected Passive Aggressive Classifier. Factors such as complexity, the ability to handle non-linear relationships, ensemble learning, regularization techniques, and dataset characteristics may have contributed to their superior performance.

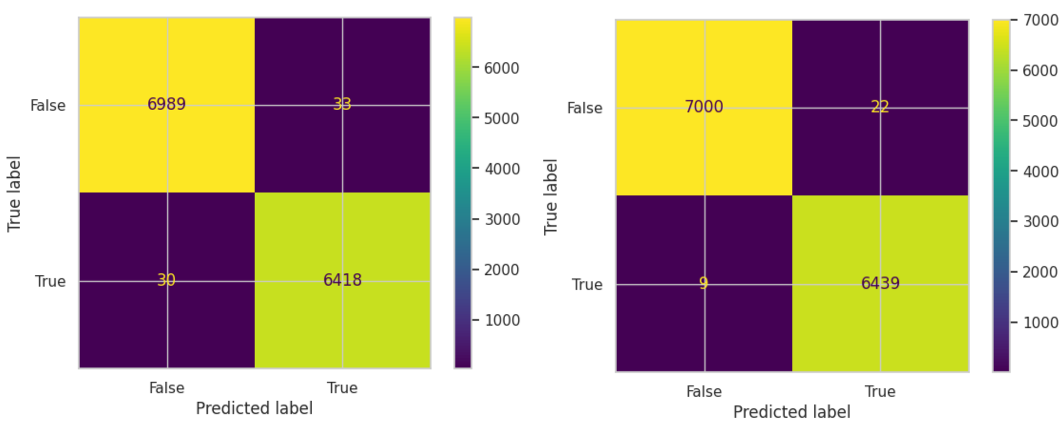
The confusion matrices obtained from test data were as follows:



(i) Logistic Regression (ii) Decision Tree



(iii) Random Forest (iv) Gradient Boosting



(v) Passive Aggressive Classifier (vi) XGBoost

Figure 4: Confusion Matrices

The results were as follows:

Figure 5: Algorithm Performance Metrics

Machine Learning Algorithm	Training Accuracy	Testing Accuracy	Testing Precision	Testing Recall	Testing F1-Score	ROC AUC Score
Logistic Regression	0.979954	0.977134	0.972453	0.979994	0.976209	0.977251
Decision Tree	1.0	0.995843	0.995350	0.995968	0.995659	0.995848
Random Forest	1.0	0.987231	0.985608	0.987748	0.986677	0.987252
Gradient Boosting	0.996786	0.995620	0.992446	0.998449	0.995439	0.995736
Passive Aggressive Classifier	1.0	0.995323	0.994885	0.995347	0.995116	0.995324
XGBoost	1.0	0.997699	0.996595	0.998604	0.997599	0.997736

From the confusion matrices, it is clear that XGBoost displayed the lowest number of false positives and true negatives. It also achieved the highest training and testing accuracy, indicating minimal overfitting and good generalization. Additionally, XGBoost showcased top precision, recall, F1 score, and ROC AUC score, indicating excellent separability. Consequently, XGBoost emerged as the best model for this specific dataset.

7 Conclusion and Future Works

In conclusion, XGBoost stood out as the most accurate machine learning algorithm. It was found that utilizing Intel optimized libraries for logistic regression resulted in reduced code runtime to 1.8x.

Future research should focus on exploring diverse data cleaning methods (e.g., stemming), testing various algorithms on different datasets, and utilizing alternative vectorizers (e.g., CountVectorizer, Word2Vec, BERT, Gensim). Additionally, optimizing parameters through techniques like grid search and random search, investigating alternative supervised (e.g., Naïve Bayes, Support Vector Machine), unsupervised (e.g., K-Nearest Neighbors), and deep learning models (e.g., Multilayer Perceptron), and pursuing other Python-based projects employing machine learning techniques will provide valuable insights into algorithm strengths and limitations. These endeavors will contribute to the enhancement of machine learning models across diverse applications.

References

[1] Sharma, U., Saran, S., & Patil, S. M. (2021). Fake News Detection using Machine Learning Algorithms. *International Journal of Engineering Research & Technology (IJERT)*, NTASU - 2020 Conference Proceedings, Special Issue - 2021.

[2] Khanam, Z., Alwasel, B. N., Sirafi, H., & Rashid, M. (n.d.). Fake News Detection Using Machine Learning Approaches. *Journal of Physics: Conference Series*, 1099(1), 012040.

[3] Pandey, S., Prabhakaran, S., Reddy, N. V. S., & Acharya, D. (n.d.). Fake News Detection from Online Media Using Machine Learning Classifiers. *Journal of Physics: Conference Series*, 2161(1), 012027.

[4] Ahmed, H., Traore, I., & Saad, S. (2018). Detecting opinion spams and fake news using text classification. *Journal of Security and Privacy*, 1(1).

[5] Ahmed, H., Traore, I., & Saad, S. (2017). Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In I. Traore, I. Woungang, & A. Awad (Eds.), *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments* (pp. 127-138). Springer.