2023 7th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech) | 979-8-3503-0551-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/IEMENTech60402.2023.10423496

Enhancing Fake News Detection with Sentiment Analysis Using Machine Learning

1st Shreea Bose
Department of Post Graduate and Research Computer Science
St Xavier's College
Kolkata, India
boseshreea.official@gmail.com

2nd Reek Roy
Department of Computer Science
Belda College
Paschim Medinipur, India
reek.beldacollege@gmail.com

***Abstract*—In the digital age we live in today, fake news has become a serious problem because it has the power to sway public opinion, affect policy decisions, and upend social harmony. Sentiment analysis, which evaluates the content's emotional tone, is essential for spotting and stopping the spread of false information. In this work, we leverage machine learning techniques for sentiment analysis to increase the detection accuracy of fake news. To scrutinize the affective milieu present in news pieces and headlines, we investigate the amalgamation of various sentiment analysis instruments, such as VADER, TextBlob, and NLTK. By translating textual input into numerical properties, TF-IDF vectorization is utilized in conjunction with sentiment analysis to increase the predictive power of machine learning models. Enhancing the feature set through the use of TF-IDF vectorization is essential for teaching machine learning models to discern between authentic and fraudulent news. As classification models, Decision trees, SVM, and Logistic regression are used; each has advantages when it comes to processing tasks involving sentiment-based classification. In this study, we try to classify the emotions that the fake news mainly shows after the prediction. We have seen the trend that mainly Negative emotion is conveyed with fake news. The models have been compared and we're getting 98% accuracy.**

***Index Terms*—SVM, classifier, Logistic regression, Decision Tree, TF-IDF, Sentiment Analysis**

I. Introduction

In the modern digital age, disseminating information via internet platforms has brought benefits and challenges. One of the most pressing issues is the extensive spread of fake news. Fake news is information purposefully manufactured or misrepresented as true news or information. It frequently seeks to mislead viewers or readers to accomplish a variety of goals, including circulating propaganda, swaying public opinion, or making money through clicks. It spreads quickly in the digital era thanks to social media, online news sources, and other digital platforms, which poses a serious obstacle to distributing trustworthy and correct information. The requirement for sentiment analysis in fake news detection stems from the fact that fake news frequently focuses on eliciting strong emotions and prejudices in its audience. In the fight against misinformation, it is imperative to comprehend the text's mood, or emotional tone, for several emotional tones, of the text for a number of reasons. Fake news frequently aims to influence its viewers' feelings. Information that causes shock, rage, or dread is more likely to be spread and taken seriously. Fake news can be philosophically, politically, or culturally discriminatory. Sentiment analysis facilitates the identification and addressing of underlying biases in information by helping to uncover them. Fake news that elicits strong emotional responses might increase participation. Sentiment analysis assists platforms and fact-checkers in identifying and regulating such information. Sentiment analysis enables the creation of customized algorithms and models for detecting fake news. It allows for the development of models that precisely target the sentiment patterns associated with misleading information. Sentiment analysis can be used in content filtering and prioritizing, assisting platforms and users in focusing on content that is more likely to be trustworthy and informative. Comprehending and assessing the effective context of information enhances our capacity to detect, classify, and impede the proliferation of false information, leading to a more knowledgeable and resilient populace confronting the challenges posed by disinformation in the digital era.

II. Related Works

Beyond theoretical discussion, fake news identification is important. It directly affects the standard of information people consume, their capacity for making wise judgments, and the general well-being of public dialogue. [1] The use of ML models contributes to broader initiatives to promote media literacy and ethical journalism in addition to helping to identify and stop the spread of misleading information. [2] Khanam Z et al. [3] has used various machine learning algorithms out of which XGBOOST has shown the highest accuracy with more than 75%, followed by SVM and Random Forest having 73% accuracy. Pandey S et al. [4] classifiers have an accuracy of 89.98% for KNN, 90.46% for Logistic Regression, 86.89% for Naive Bayes, 73.33% for Decision Tree, and 89.33% for SVM to address this challenging problem. Similarly, Shaikh J et al. [5] used SVM and found an accuracy of 95%. Baarir NF et al. [6] have used the concept of a Bag of Words with the help of the parameters meters Cost C, gamma γ, and epsilon, they have increased the accuracy of the classifier SVM. Abdulrahman A et al. [7] have compared a lot of machine learning and deep learning classifiers and combined them to get accuracy. The most important classifiers covered by them were AdaBoost, XGboost, and RNN+LSTM. Huang J [8] has used three different ways to feed the text into the model a Term Frequency-Inverse Document Frequency Vectorizer (TF-IDF), a CountVectorizer (Bag of Words), and a One-Hot representation into an embedding layer. Paper [9] has done a detailed analysis using various Machine Learning techniques like XGBoost, and SVM and deep learning techniques like LSTM. For Sentiment Analysis, they have created a Rumour Classifier Algorithm that predicts the negative emotions and tries to find if it is fake news.

III. Proposed Model

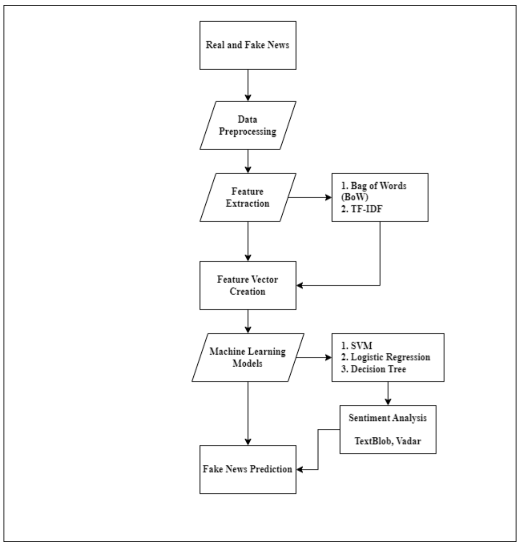


Fig. 1. Flowcahrt of Fake News Detection and Sentiment Analysis

A. Data Collection

In this study, the models are trained and assessed using a predefined dataset of tagged news articles [10]. The database in question contains data on news stories. The 'Title' of the news piece is displayed in this column. It gives a brief summary of the article's content and topic. The primary substance of the news piece is found in the "Text" column. It represents the news story's main points, details, and background. It is crucial to remember that depending on the source of the dataset, this text may be incomplete. The "Label" column acts as a binary indication, categorizing the news item. Here 1 stands for fake news and 0 stands for real news.

B. Data Preprocessing

To facilitate further analysis at the word level, tokenization first involves breaking down the text of each news story into individual words or phrases. All words are changed to lowercase to preserve consistency and prevent repetition because this technique sees terms with different capitalization as equivalent. Punctuation, special letters, and numerals are then removed to reduce noise that could otherwise impair the precision of adhering to studies. The emphasis on the linguistic substance and links between words is improved by this method. Stopwords like "the," "and," and "is" are eliminated to further refine the text because while they are frequently used, they don't have any semantic value on their own. The third phase, stemming or lemmatization, reduces words to their base forms. By normalizing word variances, this transformation enhances the model's capacity to identify the same term in several contexts. These features represent actual language variations between correct and incorrect news.

C. Feature Extraction and Feature Vector Creation

In natural language processing, feature extraction with Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) is a commonly used method, especially for text data analysis and classification of fake news. Text is converted into numerical vectors for machine learning. To begin, a vocabulary of unique terms for the dataset is created. Each document in BoW is represented as a vector with members representing term counts, resulting in a matrix. TF-IDF, on the other hand, evaluates word relevance inside a text about its frequency over the full dataset, prioritizing unique terms. Following tokenization and cleaning, BoW or TF-IDF vectors are generated and utilized as features in machine learning models to determine if news is real or false.

IV. Methodology

A. Logistic Regression

In logistic regression, the sigmoid function is used to show the probability that an input falls into a particular class. The sigmoid function is utilized in the classification of real and fake news to ascertain the probability that a news item is genuine or fraudulent based on its attributes.

$$\sigma(z) = 1/(1 + e^{-z}) \tag{1}$$

The result of the sigmoid function, denoted by $\sigma(z)$, is the expected probability. Input feature weights and their related linear combinations, along with a bias component, are represented by z. The mathematical notation for these vectors is w for weight, X for feature, and b for bias.

$$z = w * X + b. \tag{2}$$

If $\sigma(z)$ is not far from 0, there is little chance that the news is false. If it is very near to 1, there is a significant likelihood that the news is false.

B. Decision Tree

A decision tree is an arrangement like a tree that depicts options and the results of those options. Decision Trees are used to forecast fake news by leveraging characteristics extracted from news items and recursively separating the data based on the values of the features. This makes the tree a highly interpretable model for feature importance analysis, allowing it to learn decision rules that distinguish between true and false news.

C. Support Vector Machine

Every news item is represented as a point in a multidimensional space, where a different word or phrase is represented by each dimension. SVM looks for a hyperplane that maximizes the margin between the two classes (false and authentic news) in this high-dimensional space. Using SVM, the best hyperplane to split the two classes is identified.

$$w * x + b = 0 \tag{3}$$

Here w is the weight vector and b is the bias term. A data point is categorized as fake or true news depending on which side of the hyperplane it lies on.

$$y_i(w * x_i + b) >= 1 \tag{4}$$

D. Sentiment Analysis

Sentiment analysis uses programs such as Vader, TextBlob, and NLTK in Python to measure the emotional content (positive, negative, or neutral) of text using different methodologies. Vader uses lexicons and rules to award polarity scores to words and phrases depending on context and intrinsic attitude, even picking up on subtleties like sarcasm. It generates neutral, negative, and positive sentiment scores, which are aggregated to establish the overall sentiment of a document. TextBlob, on the other hand, provides an easy-to-use sentiment analysis tool that employs a pre-trained algorithm that combines machine learning and pattern-based techniques to provide subjectivity and polarity values. Tokenization, word-level sentiment scoring, and sentiment scoring using various lexicons or machine-learning models are typical features of NLTK sentiment analysis.

V. Result and Analysis

We have taken a dataset of the shape of (20800, 6) and applied the classifiers Logistic Regression, Decision Tree, and SVM. A random set of 20% of the data is taken for testing.

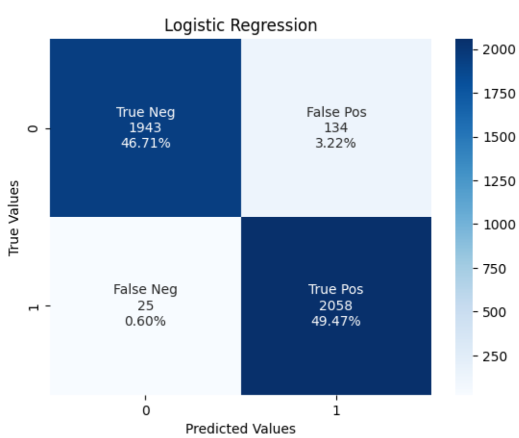


Fig. 2. Confusion Matrix of Logistic Regression

The Logistic regression model was evaluated on a dataset of 4,160 occurrences, discriminating between true news (label 0) and fake news (label 1). The results showed that the model

```
print(classification_report(Y_test, X_test_pre))

              precision    recall  f1-score   support

           0       0.99      0.94      0.96      2077
           1       0.94      0.99      0.96      2083

    accuracy                           0.96      4160
   macro avg       0.96      0.96      0.96      4160
weighted avg       0.96      0.96      0.96      4160
```

Fig. 3. Confusion Matrix of Logistic Regression

had an amazing precision of 0.99 for class 0, implying that it correctly identified real news 99% of the time. The precision for class 1 was 0.94, indicating a 94% accuracy in spotting bogus news. The model had strong recall rates of 0.94 for class 0 and 0.99 for class 1, detecting 94% of true news and 99% of fake news cases. For both courses, the F1 score of 0.96 implies a balanced trade-off between precision and recall. The model's total **accuracy of 96%** demonstrates its effectiveness in making accurate predictions on this dataset.

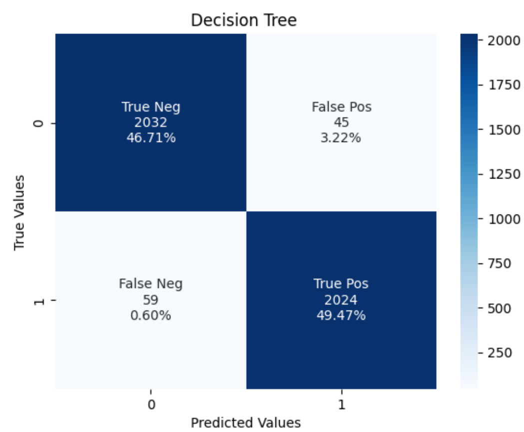


Fig. 4. Confusion Matrix of Decision Regression

```
              precision    recall  f1-score   support

           0       0.97      0.98      0.98      2077
           1       0.98      0.97      0.98      2083

    accuracy                           0.98      4160
   macro avg       0.98      0.98      0.98      4160
weighted avg       0.98      0.98      0.98      4160
```

Fig. 5. Confusion Matrix of Decision Regression

The algorithm has a high level of precision, with 97% accuracy when classifying actual news and 98% accuracy when classifying fake news. It also has an outstanding recall, identifying 98% of authentic news and 97% of fake news cases correctly. Both classes have a high F1 score of 0.98, indicating a balanced trade-off between precision and recall. The dataset contains 2,077 instances of factual news and 2,083 instances of

fraudulent news. The model correctly predicts roughly 98% o
the cases with an overall **accuracy of 98%**. The F1-scores o
0.98 on the macro and weighted average confirm its balance
and accurate performance.

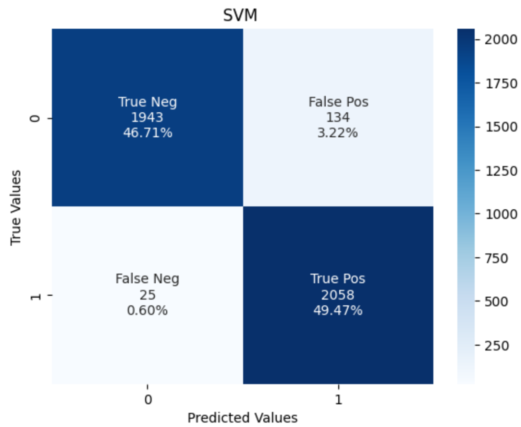
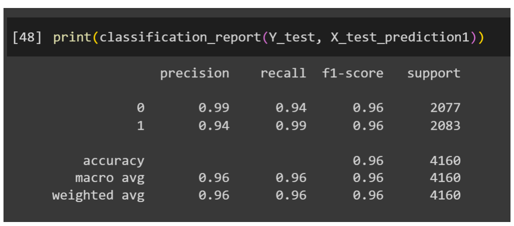


Fig. 6. Confusion Matrix of SVM



```
[48] print(classification_report(Y_test, X_test_prediction1))

              precision    recall  f1-score   support

           0       0.99      0.94      0.96      2077
           1       0.94      0.99      0.96      2083

    accuracy                           0.96      4160
   macro avg       0.96      0.96      0.96      4160
weighted avg       0.96      0.96      0.96      4160
```

Fig. 7. Confusion Matrix of SVM

The SVM model accurately predicts Class 0, which repre-sents real news, with a precision of 0.99. It almost alway properly recognizes authentic news. For class 1 (fake news the model has a high precision of 0.94, suggesting that correctly predicts bogus news 94% of the time. The SVM model has recall rates of 0.94 for class 0 and 0.99 for class collecting 94% of genuine news and nearly 99% of fake news events. Based on the dataset, the SVM model produces correct predictions with an overall **accuracy of 96%**. The weighted and macro average F1-scores are both 0.96, indicating a well balanced and efficient performance.

The false news sentiment graph reveals emotional ten-dencies within fraudulent information. It graphically depic the prevalence of negative emotion over positive and neutra that fake news frequently aims to evoke, emphasizing th necessity of sentiment analysis in recognizing and counterin disinformation.

The sentiment analysis WordCloud visualizes the mo frequently occurring words in the studied text, highlightin phrases linked with specific sentiments. The word size in th striking graphic depiction corresponds to the frequency of th term, providing immediate insight into the prominent topics and emotional tones inside the text.

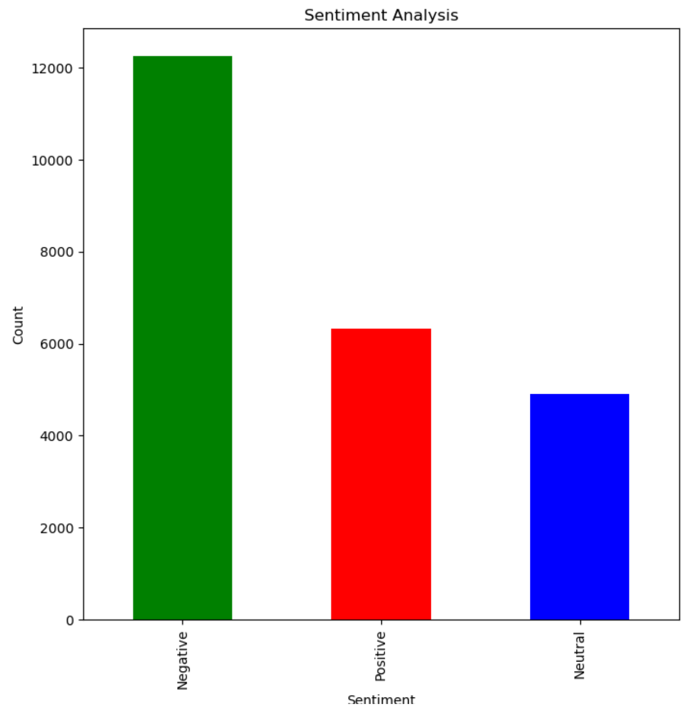


Fig. 8. Sentiment Analysis of Fake News Dataset

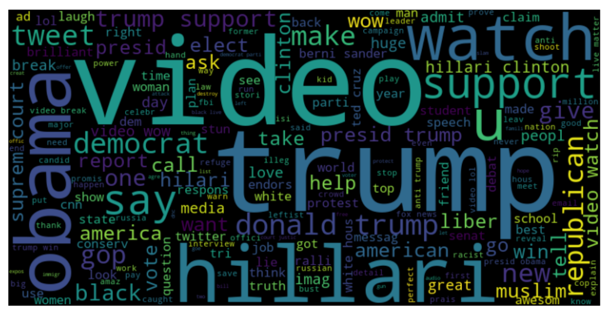


Fig. 9. Word Cloud formed by Positive Sentiments

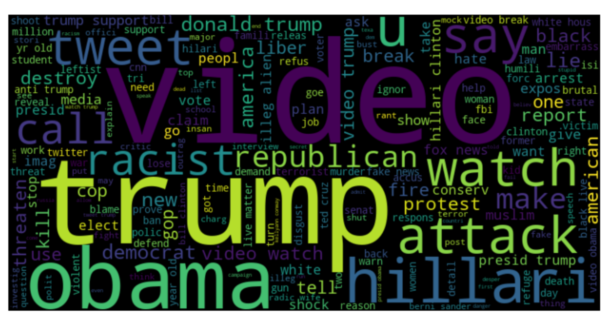


Fig. 10. Word Cloud formed by Negative Sentiments

CONCLUSION

Finally, this study emphasizes the importance of sentiment analysis as a critical instrument in the fight against the spread of fake news in the digital age. We acquire a key advantage in spotting and categorizing misleading information by comprehending and evaluating the emotional components of content. We can facilitate the distribution of correct and reliable information by being able to identify emotionally manipulative content, reveal underlying biases, and construct customized detection algorithms. In an age when disinformation is a formidable adversary, sentiment analysis emerges as a helpful ally, promising a more informed and resilient society. This study also emphasizes the importance of machine learning approaches in preventing the spread of fake news, a significant issue in the contemporary information environment.

REFERENCES

[1] M. F. Mridha, A. J. Keya, M. A. Hamid, M. M. Monowar, and M. S. Rahman, "A comprehensive review on fake news detection with deep learning," *IEEE Access*, vol. 9, p. 156151–156170, 2021.

[2] A. Jain, A. Shakya, H. Khatter, and A. K. Gupta, "A smart system for fake news detection using machine learning," *2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*, 2019.

[3] Z. Khanam, B. N. Alwasel, H. Sirafi, and M. Rashid, "Fake news detection using machine learning approaches," *IOP Conference Series: Materials Science and Engineering*, vol. 1099, no. 1, p. 012040, 2021.

[4] S. Pandey, S. Prabhakaran, N. V. Subba Reddy, and D. Acharya, "Fake news detection from online media using machine learning classifiers," *Journal of Physics: Conference Series*, vol. 2161, no. 1, p. 012027, 2022.

[5] J. Shaikh and R. Patil, "Fake news detection using machine learning," *2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC)*, 2020.

[6] N. F. Baarir and A. Djeffal, "Fake news detection using machine learning," *2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH)*, 2021.

[7] A. Abdulrahman and M. Baykara, "Fake news detection using machine learning and deep learning algorithms," *2020 International Conference on Advanced Science and Engineering (ICOASE)*, 2020.

[8] J. Huang, "Detecting fake news with machine learning," *Journal of Physics: Conference Series*, vol. 1693, no. 1, p. 012158, 2020.

[9] O. Ajao, D. Bhowmik, and S. Zargari, "Sentiment aware fake news detection on online social networks," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2507–2511.

[10] I. Kadek Sastrawan, "Fake news dataset," 2021.

.