Proceedings of the 5th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/115/2025.18485

Comparative Study of Fake News Detection Using Sentiment-Integrated Logistic Regression, LSTM, and Hybrid Models

Zihao Nie[1,a,*]

[1]*School of AI and Advanced Computing, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, China*
a. Zihao.Nie23@student.xjtlu.edu.cn
**corresponding author*

Abstract: With the advent of the digital age, fake news spreads faster and faster on social media, causing a major adverse impact on social public opinion. Therefore, effective information detection technology is very important to mitigate the negative impact of false information and protect the health and stability of society. This research aims to improve fake news detection by incorporating sentiment analysis into traditional machine learning and deep learning models. The dataset used ISOT dataset contains more than 40,000 news articles and is used to compare the performance of logistic regression, Long Short-Term Memory (LSTM), and hybrid ensemble models. The study results show that the accuracy of the integrated model is the highest, reaching 99.24%, and the F1 value is 0.9922. The accuracy of logistic regression also reached 99.12%. Although sentiment analysis can add some value, it has a limited impact on model performance. This means that combining the traditional learning model with the deep learning model can enhance the fake news detection effect.

Keywords: LSTM, machine learning, logistic regression, fake news.

1. Introduction

With the wide use of digital products in human life, social media has completely changed the way people get information. While people have more convenient access to instant information, there is also a fatal and serious disadvantage: the rampant spread of fake news. Fake news exerts a significant effect on the credibility of information and public opinion. The US presidential election in 2016 is a typical example. The proliferation of fake news on social media may directly affect public opinion polls and even the final election results [1]. Similarly, during COVID-19, misinformation about the source of the virus, vaccine safety, and potential treatments spread rapidly on online platforms, leading to widespread public confusion and resistance to health measures [2]. These examples show that fake news seriously affects public trust, social stability, and even global human life and health safety. Therefore, there is an urgent need for more advanced technical solutions, especially using big data to effectively identify and combat fake news [3].

Given the widespread dangers of fake news, the topic has attracted considerable attention in various disciplines, including information science, computer science, and sociology. The main body of existing research highlights the complexity of the complicated propagation mode, diverse content characteristics, and far-reaching social impact of fake news [4]. For example, research by Vosoughi et al. shows that fake news spreads faster than real news on social media, often through key nodes in

DOI: 10.54254/2755-2721/115/2025.18485

social networks [5]. This finding highlights the unique challenges that fake news faces in the modern information ecosystem.

To tackle the challenges of detecting fake news, researchers have examined a range of automated techniques. Conroy et al. reviewed key methods, including those focused on text feature analysis and machine learning models designed to study user interactions [6]. Although these strategies show potential, they also encounter obstacles, particularly with the large and diverse datasets common in digital media. Furthermore, adapting to the constantly evolving forms of fake news remains a significant challenge, as fake news often involves language changes, hidden content, and implicit indicators. Fake news may appear in many forms, such as changes in storytelling style, concealed backgrounds, shifts in tone, and a variety of channels through which information is shared. The difficulty in distinguishing fully fake news from partially fake content adds further complexity, challenging models to effectively handle the dynamic and complex nature of fake news [7].

Big data technologies bring new possibilities for addressing these challenges. By enabling extensive data collection, storage, and analysis, big data offers valuable see on the patterns clearly and propagation of fake news across various platforms and contexts. Such insights are essential for building more accurate and resilient detection models. For instance, Shu et al. introduced a framework rooted in big data that combines text analysis, social network analysis, and user behavior patterns to improve fake news detection [8]. This approach marks a significant advancement, offering a comprehensive perspective on the factors driving misinformation spread.

Despite these advancements, a significant research gap remains. As tactics for spreading fake news evolve, detection models need continuous refinement to address emerging challenges [9]. This study contributes to these efforts by utilizing big data to develop models that are more adaptive and comprehensive in detecting fake news, ultimately aiming to address its complexity and varied nature more effectively.

This study investigates the potential of big data to enhance fake news detection. By integrating text and sentiment analysis of articles, it aims to build a more effective model for identifying fake news. The approach combines deep learning with traditional machine learning techniques to analyze extensive datasets.

The primary goal of this research is to improve fake news detection using big data technology. This involves efficiently collecting and processing large datasets, integrating sentiment analysis into big data processes, and comparing the performance of logistic regression, Long Short-Term Memory (LSTM) models, and their hybrid combinations. By leveraging multiple analytical techniques, the study aims to develop a model that adapts well across various types of fake news and digital platforms. Continuous evaluation and refinement will be crucial to maintaining the model's effectiveness in practical applications.

2. Methodology and Data

2.1. Dataset

This study used the ISOT Fake News Detection Dataset from the University of Victoria, Canada. It contains more than 40,000 short news claims, 23,481 of which have been verified to be false, and 21,417 of which are true news. These data cover many fields including politics, entertainment, health, and other daily news content, which is of practical significance for research [10,11]. Table 1, Figure 1, and Figure 2 show information such as the type of distribution of the dataset and the length distribution of the text. In particular, Figure 2 shows that the data whether real news or fake news in the ISOT dataset are the short texts.

DOI: 10.54254/2755-2721/115/2025.18485

Table 1: Data type table [10,11]

News	Size (number of articles)	Subjects	
Real-news	21417	**Type**	**Articles size**
		World-news	*10145*
		Politics-news	*11272*
Fake-news	23418	**Type**	**Articles size**
		Government-news	*1570*
		Middle-east	*778*
		Us news	*783*
		Left-news	*4459*
		Politics	*6841*
		News	9050

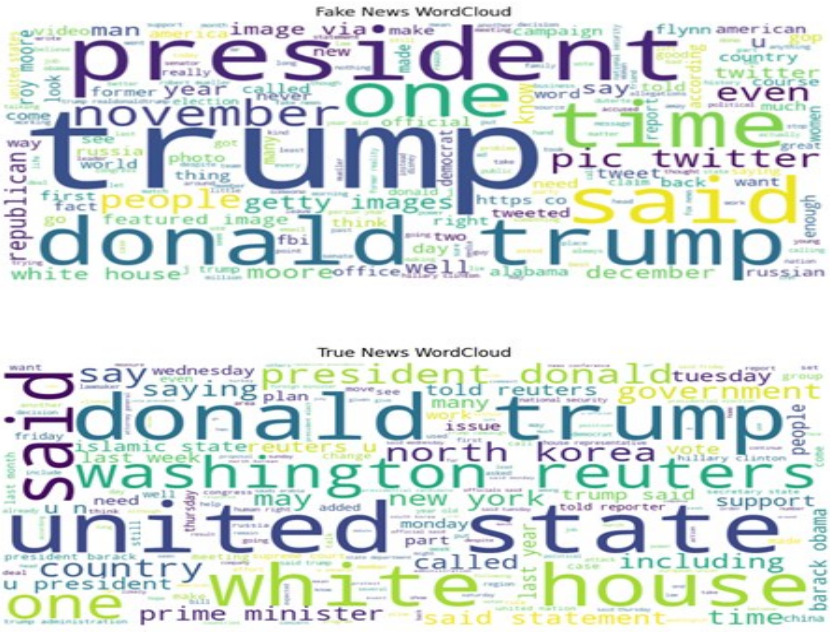

Figure 1: Word cloud of text data

DOI: 10.54254/2755-2721/115/2025.18485

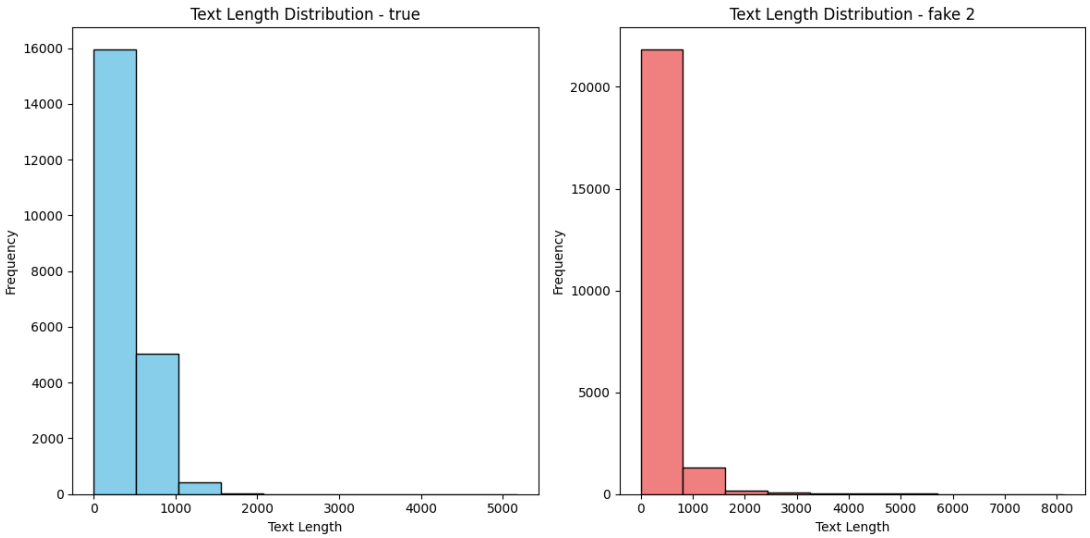

Figure 2: Histogram of text length distribution

To prepare the data, the study applied some basic text processing methods. First, all text was converted to lowercase, and unnecessary words and punctuation were removed. Then, the study split the text into individual words and standardized their forms to make it easier for the model to analyze.

2.2. Research Process

In this study, three types of models are used for comparison: traditional machine learning, deep learning, and their hybrid model. The logistic regression model is selected for the traditional machine learning model, and the LSTM model is selected for the deep learning model. The features of text sentiment analysis are integrated into each model. After training, a better performance detection model can be obtained by comparing their various indicators.

The dataset was split into 80% for training and 20% for testing. This means the paper used most of the data to train the models and the remaining portions to test how well the models performed. Additionally, the paper applied cross-validation to ensure the results were reliable by testing the models multiple times. Figure 3 shows the diagram of the entire process in the study.

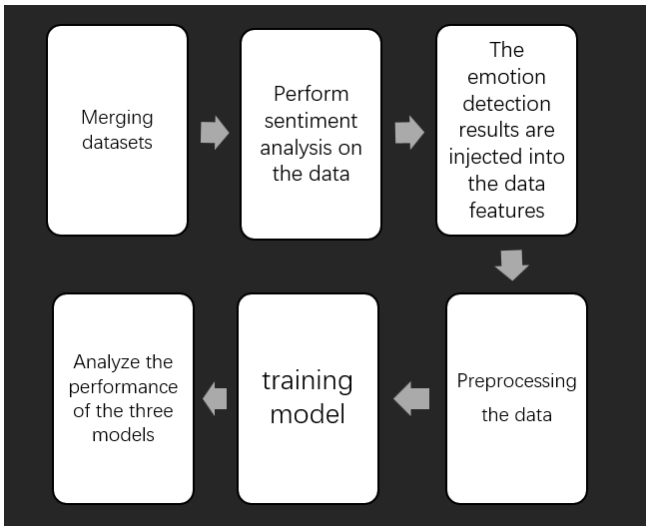

Figure 3: Flowchart of the method

DOI: 10.54254/2755-2721/115/2025.18485

2.3. The Model

2.3.1. Model Introduction

In machine learning and deep learning, choosing the right model is critical for achieving effective and accurate results in tasks like fake news detection. Traditional models, such as logistic regression, are popular for their simplicity and efficiency, while deep learning models like LSTM are particularly adept at managing sequential data with time-based dependencies and contextual relationships [12,13]. This study leverages a hybrid ensemble approach that combines the advantages of both types of models, aiming to reduce the error rate of false news detection.

2.3.2. Logistic Regression Overview

Logistic regression is a commonly used statistical model for binary classification tasks. It aims to predict the probability that an input belongs to a specific class, typically producing an output between 0 and 1 [14]. The model's underlying principle is to maximize the likelihood of a binary outcome based on the input features. Specifically, logistic regression assumes a linear relationship between the input features and the log odds of the outcome:

$$P(y = 1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\cdots+\beta_n X_n)}} \quad (1)$$

Where X1, X2,…, Xn represent the input features, and $\beta 0, \beta 1, \ldots \beta n$ are the parameters learned during training. Due to its simplicity and interpretability, logistic regression allows for straightforward insights: the sign and size of each coefficient reveal the effect of each characteristic on the outcome's probability [15]. This model is also computationally efficient, making it suitable for large datasets and high-dimensional feature spaces [16]. However, while logistic regression works well with linearly separable data, it may struggle to capture complex nonlinear relationships, limiting its effectiveness for tasks with rich contextual or semantic details, particularly in text processing [17].

2.3.3. LSTM Overview

Long Short-Term Memory (LSTM) networks are a type of Recurrent Neural Network (RNN) created to manage the challenge of long-term dependencies that traditional RNNs struggle with. LSTM addresses the vanishing and exploding gradient problems by using a memory cell along with gating mechanisms [18]. It features three critical gates: the forget gate, which removes irrelevant information; the input gate, which determines what latest information is stored; and the output gate, which shapes the output based on the memory cell's content. These mechanisms allow LSTM networks to effectively manage long-range dependencies in sequential data [19].

LSTM's benefits include its capacity to keep valuable information over extended sequences, making it suitable for tasks involving text and time-series data [20]. Unlike standard RNNs, which often struggle with vanishing gradients in long sequences, LSTM can successfully learn long-term patterns [21]. However, its complex structure means that LSTM models require more computational resources and longer training times, particularly with large datasets [22].

3. Result and Discussion

3.1. The Result

The performance of the three models—Logistic Regression, LSTM, and Ensemble—was evaluated using accuracy and F1 score metrics, as shown in Figure 4 and Table 2. Logistic Regression achieved an accuracy of 99.12% with an F1 score of 0.9909, while the LSTM model followed closely with an accuracy of 98.96% and an F1 score of 0.9892. The Ensemble model outperformed both, reaching an

DOI: 10.54254/2755-2721/115/2025.18485

accuracy of 99.24% and an F1 score of 0.9922. These results demonstrate that although all models performed well, the Ensemble model provided a slight improvement over the individual models in terms of both accuracy and F1 score

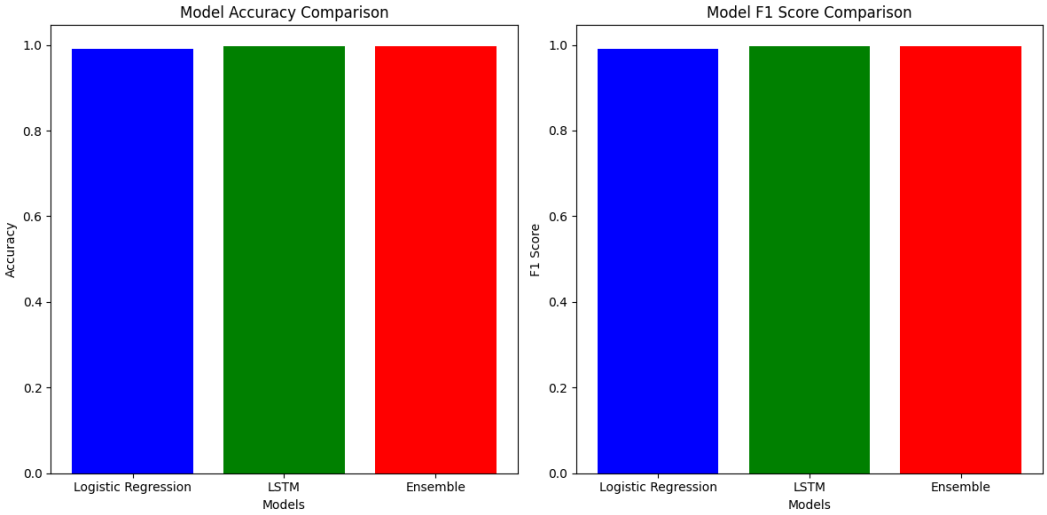

Figure 4: Comparison of accuracy of each model with f1 score histogram

Table 2: Specific data sheet of model performance

model	accuracy rate	F1 score
Logistic regression	0.9912	0.9909
LSTM	0.9896	0.9892
Hybrid model	0.9924	0.9922

3.2. Discussion

The results show that all three models—Logistic Regression, LSTM, and the Ensemble model—performed similarly, with the Ensemble model displaying a slight advantage in both accuracy and F1 score. Logistic Regression proved robust and efficient for this task, while LSTM did not deliver significant improvement, indicating that the dataset may not fully benefit from sequence-based modeling.

Sentiment analysis contributed some additional insight, but its effect was limited, given the strong baseline performance of all models. Future studies could investigate more complex datasets and advanced sentiment analysis techniques to assess the potential of hybrid models more effectively such as the Ensemble model.

In summary, while the Ensemble model led in performance, the simplicity and reliability of Logistic Regression make it a competitive choice for large-scale fake news detection.

4. Conclusion

This study compared the performance of Logistic Regression, LSTM, and an Ensemble model for fake news detection, incorporating both text features and sentiment analysis. All models achieved high accuracy and F1 scores, with the Ensemble model slightly outperforming the others. The results indicate that while deep learning models like LSTM may offer some advantages in handling sequential data, their performance gain on this dataset was minimal.

DOI: 10.54254/2755-2721/115/2025.18485

Logistic Regression proved to be a strong contender due to its simplicity and efficiency, making it suitable for large-scale applications. The Ensemble model demonstrated the value of combining different model architectures, providing a slight improvement in classification performance.

In summary, while hybrid models like the Ensemble offer potential, simpler models like Logistic Regression remain highly competitive for tasks such as fake news detection. Future work could explore more diverse datasets and advanced techniques to further evaluate model effectiveness. With the increasing development of the network today, false news detection should also be developed, to better govern the network society and develop a rational network environment.

References

[1] *Alcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. Journal of Economic Perspectives, 31(2), 211–236.*

[2] *Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... & Zittrain, J. L. (2018). The science of fake news. Science, 359(6380), 1094-1096.*

[3] *Zhao, Z., Zhang, J. (2020). Research on the Spread of Fake News in the Context of Big Data. Information Science, 38(8), 123-129.*

[4] *Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2018). Automatic detection of fake news. Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018), 3391-3401.*

[5] *Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146-1151.*

[6] *Conroy, N. K., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology, 52(1), 1-4.*

[7] *Zhou, X., & Zafarani, R. (2020). A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. ACM Computing Surveys (CSUR), 53(5),*

[8] *Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22-36.*

[9] *Wang, Y., Chen, G. (2021). Research on Fake News Detection Models Based on Big Data. Modern Information, 41(4), 56-63.*

[10] *Ahmed H, Traore I, Saad S. Detecting opinion spams and fake news using text classification,"Journal of Security and Privacy, 1, I1, January/February 2018.*

[11] *Ahmed H, Traore I, Saad S. (2017) Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In: Traore I., Woungang I., Awad A. (eds) Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. ISDDC 2017. Lecture Notes in Computer Science, 10618. Springer, Cham, 127 138.*

[12] *Kleinbaum, D.G., & Klein, M. (2010). Logistic Regression: A Self-Learning Text. Springer.*

[13] *Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735-1780.*

[14] *Hosmer, D.W., Lemeshow, S., & Sturdivant, R.X. (2013). Applied Logistic Regression (3rd ed.). Wiley.*

[15] *Kleinbaum, D.G., & Klein, M. (2010). Logistic Regression: A Self-Learning Text. Springer.*

[16] *Ng, A.Y., & Jordan, M.I. (2001). On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes. NIPS, 14, 841-848.*

[17] *Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.*

[18] *Graves, A. (2012). Supervised Sequence Labelling with Recurrent Neural Networks. Springer.*

[19] *Gers, F.A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: Continual prediction with LSTM. Neural Computation, 12(10), 2451-2471.*

[20] *Greff, K., Srivastava, R.K., Koutnik, J., Steunebrink, B.R., & Schmidhuber, J. (2017). LSTM: A Search Space Odyssey. IEEE Transactions on Neural Networks and Learning Systems, 28(10), 2222-2232.*

[21] *Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.*

[22] *Jozefowicz, R., Zaremba, W., & Sutskever, I. (2015). An Empirical Exploration of Recurrent Network Architectures. Proceedings of the 32nd International Conference on Machine Learning, 37, 2342-2350.*