Weighted Ensemble Model for Tackling Fake News

Ananya Kohli[1], Divyashree Shetti[1], Sri Lakshmi G N[1], Vaishnavi Bhat[1] and Shashank Hegde[1]
[1]*School of Computer Science Engineering, KLE Technological University, Hubballi, India*

Keywords: Fake News Detection, Ensemble Learning, BERT, Classification Models, Weighted Averaging, Transformers.

Abstract: Fake news detection has become crucial in resisting misinformation across multiple domains like social media, news outlets, and public communications. Accurate classification and sentiment analysis play a pivotal role in addressing this challenge. Although traditional machine learning models have shown moderate success, they face limitations in achieving high accuracy and adaptability when applied to diverse types of content. To address this, a fake news detection model is proposed that evaluates the authenticity of news reports by leveraging feature extraction and credibility scoring through accuracy. The proposed study presents a robust fake news detection model that combines BERT (Bidirectional Encoder Representations from Transformers) embeddings with ensemble learning techniques. Eight machine learning classifiers - Logistic Regression, SGD (Stochastic Gradient Descent), XGBoost (Extreme Gradient Boosting), SVM (Support Vector Machine), Random Forest, AdaBoost (Adaptive Boosting), KNN (K-Nearest Neighbor) and Naive Bayes were trained on an 80:20 train-validation split. Using ensemble techniques including Majority Voting, Unweighted Averaging and Weighted Averaging, the proposed work with Weighted Averaging proved to be the most accurate method, with an accuracy of 94.8317%. This is because the weights were normalized depending on the individual model approach, making the model a reliable and adaptable solution to misinformation detection.

1 INTRODUCTION

In today's digital age, misinformation spreads faster than ever before, creating new challenges in how we consume and trust the information around us. Detecting fake news is not a simple task, falsehoods often mirror the structure and tone of credible news, making the lines between fact and fiction difficult to discern, even for human readers. Researchers have long sought to develop systems that can differentiate between true and false information based on patterns in the text (Castillo et al., 2011). One such breakthrough is BERT (Bidirectional Encoder Representations from Transformers), which has revolutionized NLP (Natural Language Processing) by capturing contextual relationships in text with extraordinary accuracy (Devlin et al., 2018). Unlike earlier used models, BERT processes text bidirectionally, allowing it to consider the full context of each word in a sentence, making it exceptionally well-suited for understanding the complications of language. The proposed fake news detection approach combines BERT's robust embeddings with ensemble learning techniques.

The proposed approach is built on the premise that no single model is perfect, but by combining the predictions of multiple classifiers, we can achieve a more reliable and accurate result. This is related by "No Free Lunch Theorem" which highlights that no single algorithm can outperform all others across all types of problems, underscoring the necessity for tailored solutions. Thus this research uses BERT embeddings as the foundation, feeding them into a variety of classifiers, including Logistic Regression, XGBoost, SVM and more. Each classifier has been fine-tuned with optimized hyper-parameters, ensuring it performs at its best. Through the use of regularization techniques, over-fitting is prevented, ensuring that the developed framework generalizes well to new, unseen data(Bengio et al., 2012). Additionally, these trained classifiers are integrated into ensemble techniques such as Majority Voting, Weighted Averaging (with normalized weights assigned based on validation accuracies) and Unweighted Averaging. This method produces a system that is both highly accurate and adaptable to the varied misinformation strategies employed across different platforms, achieving an impressive accuracy of 94.8317%. The outcome is a fake news detection tool capable of evolving with new challenges, maintaining its relevance and effec-

Kohli, A., Shetti, D., Lakshmi G N, S., Bhat, V. and Hegde, S.
Weighted Ensemble Model for Tackling Fake News.
DOI: 10.5220/0013606400004664
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 3rd International Conference on Futuristic Technology (INCOFT 2025) - Volume 2*, pages 878-884
ISBN: 978-989-758-763-4
Proceedings Copyright © 2025 by SCITEPRESS – Science and Technology Publications, Lda.

tiveness in an ever-changing information ecosystem.

However, the approach does face limitations, primarily because of its resource-intensive nature of high computational demands of BERT embeddings and ensemble learning techniques along with the constraint in model's scalability when applied to multilingual datasets or specialized domains, where further refinement is necessary to ensure consistent performance. Despite these challenges, the integration of ensemble methods and BERT embeddings provides a robust framework for combating misinformation, with potential for real-world applications in media platforms, fact-checking organizations, and beyond.

The paper is organized as follows: Section 2 reviews existing fake news detection algorithms, with a focus on ensemble techniques and their applications. Section 3 delves into the architecture of previous fake news detection systems, offering insights into their strengths and weaknesses. Section 4 introduces the proposed methodology, detailing how BERT embeddings are utilized to train a diverse set of models using regularization techniques. Section 5 presents the experimental results by comparing the accuracy of various models and identifying the most successful approach. Finally, Section 6 concludes the paper by summarizing the findings and reflecting on the broader implications of the proposed approach.

2 BACKGROUND STUDY

The detection of fake news has significantly evolved, transitioning from traditional machine learning methods like Logistic Regression and SVM to advanced deep learning approaches. Earlier methods relied on linguistic features such as TF-IDF for classification, which performed well for straightforward tasks but struggled with understanding deeper contextual relationships within text. The advent of deep learning, particularly models like LSTM and BERT, revolutionized fake news detection by capturing semantic nuances and bidirectional context in language. BERT, with its robust contextual understanding, has greatly improved classification performance (Devlin et al., 2018; Vaswani et al., 2017; Yang and Cui, 2021). However, challenges such as overfitting on limited datasets and domain adaptation issues hinder their generalization (Jin et al., 2022; Wang et al., 2023). Ensemble learning methods which include Random Forest and XGBoost mitigate these limitations by combining multiple models, reducing overfitting, and enhancing robustness (Breiman, 2001; Chen et al., 2016; Friedman, 2001). These methods also facilitate improved decision-making through diverse feature combinations, which is crucial for handling complex and ambiguous fake news content. Additionally, the integration of explainable AI (XAI) techniques in ensemble models offers more transparent insights into the decision-making process, further strengthening trust in automated systems (Gilpin et al., 2018).

The increase in fake news across social media and digital platforms highlights the need for adaptable systems capable of handling rapidly evolving content types and domains. Traditional approaches relying on handcrafted linguistic features such as n-grams, bag-of-words, and syntactic structures (Joachims, 1998; Salton, 1986) often fall short in addressing the complexities of modern strategies for spreading false information. Deep learning models, including CNNs and LSTMs, brought advancements by capturing hierarchical and temporal patterns from text (Kim, 2014; Hochreiter and Schmidhuber, 1997), yet they struggle with diverse, noisy, or domain-specific datasets. Current research emphasizes hybrid methods that combine the powerful feature extraction of models like BERT with ensemble strategies. These approaches provide scalability and adaptability for misinformation detection across varied contexts (Zhang and Bao, 2020; Jiang et al., 2021; Zhou et al., 2020).

Despite the accuracy gains of deep learning models like BERT, they are computationally expensive and prone to overfitting, particularly on imbalanced datasets (Vaswani et al., 2017; Czapla et al., 2019). Additionally, their "black-box" nature raises concerns about interpretability and trust (Gilpin et al., 2018; Marco Tulio Ribeiro, 2016). To address these issues, the proposed research introduces a novel ensemble learning approach that integrates BERT with simpler classifiers such as Logistic Regression, SVM, XGBoost and others. This ensemble reduces reliance on any single model, improving both generalization and computational efficiency while enhancing transparency and robustness. Further incorporate feature importance analysis using XGBoost is incorporated to provide greater model explainability (Caruana and Niculescu-Mizil, 2006; Marco Tulio Ribeiro, 2016). By combining SGD and Naive Bayes, this approach ensures scalability and better performance in dynamic, high-dimensional and real-time environments (Freund and Schapire, 1997; McCallum and Nigam, 1998). This sets the stage for the proposed methodology, where the aim is to leverage these insights and integrate various models to build a more effective and efficient fake news detection system.

3 PROPOSED METHODOLOGY

The proposed methodology follows a structured workflow that integrates BERT embeddings with traditional machine learning models to enhance fake news detection. The process begins with cleaning the dataset for any invalid values followed by dividing it into training and validation sets, which are then passed through BERT for embedding generation. The BERT model using equation 1, processes the input data and produces context-aware embeddings, which are stored for efficient retrieval and usage. These embeddings capture the textual features of the news articles and serve as input for various machine learning classifiers, including Logistic Regression, SGD (Stochastic Gradient Descent), XGBoost (Extreme Gradient Boosting), SVM (Support Vector Machine), Random Forest, AdaBoost (Adaptive Boosting), KNN (K-Nearest Neighbor) and Naive Bayes. The BERT embedding calculation is given by the equation 1

$$\mathbf{E}\text{output} = f\text{BERT}(\text{Tokenized Input}) \quad (1)$$

Once the embeddings are ready, the individual models are trained on the dataset with an 80:20 split for training and validation. Later these trained models are given as inputs to ensemble techniques such as Majority Voting, Unweighted Averaging and Weighted Averaging. The performance of individual and ensemble models are assessed using accuracy, precision, F1 score and recall on both the training and validation sets. These performance metrics are stored for further analysis. The models' validation accuracies are used to weigh their contributions in the ensemble techniques. Specifically, for Weighted Averaging in equation 2, normalized weights are assigned based on the validation accuracies of the individual models, allowing more accurate models to have a greater influence in the final decision. The Weighted Averaging calculation is given by the equation 2

$$\hat{y}_{\text{final}} = \frac{\sum_{m=1}^{M} w_m \cdot \hat{y}_m}{\sum_{m=1}^{M} w_m} \quad (2)$$

where $\hat{y}_m$ is the prediction and w_m is the weight associated with model m and M refers to the total number of individual models used. This equation 2 calculates the final prediction by summing the weighted predictions and applying a threshold (0.5) to get a binary class.

Alongside Weighted Averaging, two additional ensemble techniques such as Majority Voting and Unweighted Averaging are applied to combine the predictions from all models, providing a more generalized output. In Majority Voting, according to the equation 3, the final predicted class is determined by the majority vote across all models and Unweighted Averaging from equation 4 assigns equal importance to each of the input models. The equations for Majority Voting 3 and Unweighted Averaging 4 are

$$\hat{y}_m = \arg\max_{c \in C} \sum_{m=1}^{M} \delta(y_m = c) \quad (3)$$

$$\hat{y}_{\text{final}} = \frac{\sum_{m=1}^{M} y_m}{M} \quad (4)$$

where δ is an indicator function for each model's prediction and 1 is the weight associated with model m.

After the models are trained, they predict labels for the test embeddings. The predicted labels from each model are stored for performance evaluation. To assess the overall effectiveness of the system, accuracy, precision, F1 score, and recall are computed not only for the individual models but also for the ensemble techniques.

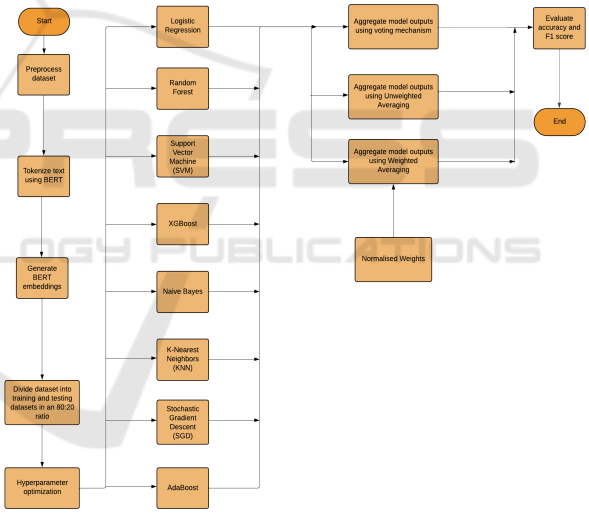

Figure 1: Proposed Ensemble Framework.

The dataset (Lifferth, 2018) consisted of primarily textual data which included a training set with labeled records and a test set with a similar structure but no ground truth labels. The training data included unique id, titles, authors, and text content for analysis. The final output incorporates ensemble predictions, specifically using the columns y_pred_majority, y_pred_unweighted and y_pred_weighted, which aggregated model outputs through Majority Voting, Unweighted Averaging and the proposed Weighted Averaging model, respectively to enhance prediction accuracy and reliability.

3.1 Proposed Architecture

The architecture shown in Figure 2 illustrates the components and workflow of the entire ensemble model. In this setup, the training and testing data are input to BERT, which generates embeddings that are then provided to traditional machine learning models. The predictions from these models are fed into ensemble classifiers. The accuracy of each ensemble classifier's predictions is calculated, and the predictions with the highest accuracy are selected as the final ensemble output.

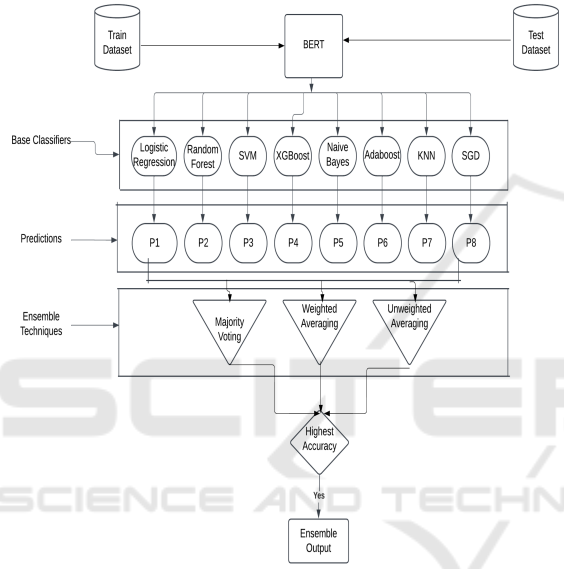

Figure 2: Fake news detection ensemble architecture.

3.2 Proposed Algorithm

The algorithm 1 shows that this methodology integrates BERT for feature extraction, followed by training a range of classifiers and finally combines their outputs using ensemble methods like Majority Voting, Unweighted Averaging and Weighted Averaging. This approach ensures a robust and accurate fake news detection system which demonstrates the strength of combining diverse model predictions based on their validation performance.

Data: Textual train dataset labeled as reliable (0) or potentially fake (1), and a test dataset.
Result: Predicted labels for the test dataset and evaluation metrics.
Initialize train and test datasets;
Preprocess the data by removing invalid values or tuples;
Generate BERT embeddings for both datasets and save them;
while *train-validation split is incomplete* **do**
  Split train dataset into 80:20 train-validation;
  Train individual machine learning models on the training data;
  Compute validation accuracies for each model;
  if *validation accuracy is acceptable* **then**
    Store the model and its accuracy;
  end
  else
    Re-adjust hyperparameters or preprocessing and re-train the models;
  end
end
Combine predictions using ensemble methods;
if *ensemble method is Majority Voting* **then**
  Assign labels based on the most frequent prediction;
end
else if *ensemble method is Unweighted Averaging* **then**
  Average predicted probabilities and assign labels;
end
else if *ensemble method is Weighted Averaging* **then**
  Use normalized validation accuracies as weights, compute weighted averages and assign labels;
end
Calculate validation accuracy for each of these ensemble methods and select the ensemble method with the highest accuracy;
Predict labels for the test dataset using the chosen ensemble method;
Evaluate results with accuracy, precision, recall, and F1-score;

Algorithm 1: Weighted Ensemble Model for Tackling Fake News

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

4 RESULTS AND ANALYSIS

The evaluation of proposed fake news detection system reveals significant insights into the performance of individual machine learning models and ensemble techniques. Leveraging BERT embeddings as feature representations, the models were assessed using metrics such as accuracy, precision, recall and F1-score. The results, summarized in Table 1, highlight the comparative strengths of different approaches, including the enhanced reliability achieved through ensemble methods like Weighted Averaging.

Table 1: Performance Metrics for Models and Ensemble Techniques

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.954087	0.954106	0.954092	0.954086
SVM	0.946154	0.946419	0.946172	0.946147
KNN	0.891106	0.893549	0.891163	0.890946
XGBoost	0.931490	0.931823	0.931511	0.931479
SGD	0.938942	0.941321	0.938996	0.938865
Random Forest	0.891106	0.892637	0.891151	0.891007
AdaBoost	0.842788	0.843168	0.842812	0.842751
Naive Bayes	0.670433	0.720533	0.670776	0.650858
Majority Voting	0.931010	0.933763	0.931067	0.930906
Unweighted Averaging	0.859856	0.883128	0.859856	0.857666
Proposed model (Weighted Averaging)	0.948317	0.949070	0.948317	0.948297

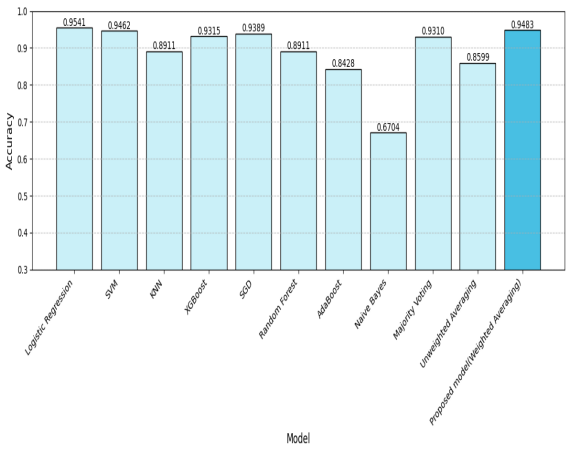

Figure 3: Model accuracy comparison.

Table 1 presents the performance metrics of individual machine learning models and ensemble techniques for fake news detection. The models were evaluated based on accuracy, precision, recall, and F1-score to determine their effectiveness. Logistic Regression achieved the highest accuracy (0.954087) among individual models, followed by SVM (0.946154) and SGD (0.938942), indicating strong classification capabilities. In contrast, Naïve Bayes had the lowest accuracy (0.670433), highlighting its limitations in this context. Figure 3 further visualizes the accuracy comparison, reinforcing the superior performance of the proposed model.

The diagrammatical accuracy comparison among different models in Figure 3 show that Logistic Regression achieves the highest accuracy, followed by Weighted Averaging and Support Vector Machine. While Logistic Regression performs the best in terms of individual model accuracy, Weighted Averaging improves on this by combining the strengths of multiple models, leading to more robust predictions than a single model.

The graph in Figure 4 compares the accuracy of ensemble techniques such as Majority Voting, Weighted Averaging, and Unweighted Averaging. Weighted Averaging achieves the highest accuracy, while Majority Voting performs better than Unweighted Averaging with an accuracy of 0.948317. Both Majority Voting and Unweighted Averaging underperform because they treat all models equally, allowing weaker models to influence the final predictions. In contrast, Weighted Averaging improves accuracy by giving more weight to stronger models and reducing the impact of weaker ones.

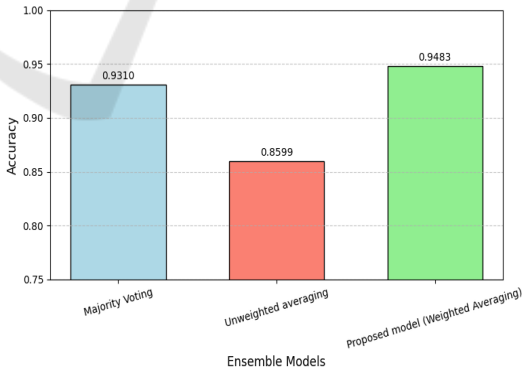

Figure 4: Ensemble techniques accuracy comparison

Although the approach enhances reliability, it comes with several limitations. The system is resource-intensive due to the high computational demands of BERT embeddings and ensemble techniques, which may limit its deployment in resource-constrained environments. Also the model primarily processes textual data, which restricts its ability to

handle multimodal fake news, such as misinformation spread through images or videos. Additionally challenges arise in deploying the model at scale for multilingual datasets or adapting it to highly specialized domains as it requires further refinement to maintain optimal performance. These limitations underscore the need for continued research to improve the system's versatility.

5 Conclusion and Future Work

This research developed a hybrid fake news detection system by integrating BERT embeddings with ensemble machine learning models. The system effectively captured the semantic meaning of news content, achieving improved accuracy and reliability through Voting, Unweighted and Weighted Averaging techniques. Weighted Averaging proved to be the most reliable, leveraging the strengths of diverse models and mitigating the impact of outliers using normalized weights for consistent performance. Furthermore, the system demonstrated scalability and adaptability across different datasets, making it suitable for real-world applications. By combining the power of deep learning with traditional classifiers, it addresses key challenges such as overfitting and model interpretability. The integration of these techniques lays the foundation for building a robust and efficient fake news detection system. Additionally, the approach's transparency helps enhance trust and accountability in automated decision-making.

The proposed approach contributes to future advancements in fake news detection by enhancing accuracy through weighted averaging in ensemble learning, making it a scalable and adaptable framework. News verification systems can leverage the model to assist journalists and media organizations in assessing the credibility of articles before publication. Search engines can incorporate the model to filter out misleading content, enhancing the integrity of online information. The model can be enhanced to prevent market manipulation through fake financial news and also detect false health claims, medical misinformation, and prevent public health crises.

Future enhancements include exploring diverse data types, fine-tuning BERT for domain-specific applications and enabling real-time detection capabilities. Expanding support for multiple languages and utilizing larger datasets will further improve system performance. Additionally, incorporating explainable AI and robust defenses against fake content can enhance transparency and reliability in detection.

REFERENCES

Bengio, Y. et al. (2012). Practical recommendations for gradient-based training of deep architectures. *Neural Networks: Tricks of the Trade*, pages 437–478.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Caruana, R. and Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd international conference on Machine learning*, pages 161–168.

Castillo, C. et al. (2011). Information credibility on twitter. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 675–684. ACM.

Chen, T. et al. (2016). Xgboost: A scalable tree boosting system. *ACM SIGKDD*, pages 785–794.

Czapla, P., Gugger, S., Howard, J., and Kardas, M. (2019). Universal language model fine-tuning for polish hate speech detection. In *Proceedings of the PolEval2019 Workshop*, page 149.

Devlin, J. et al. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Gilpin, L. H. et al. (2018). Explaining explanations: An overview of interpretability of machine learning. *ACM Computing Surveys (CSUR)*, 51(5):93.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Jiang, T., Yu, X., Li, C., Song, Y., and Zhan, Y. (2021). A novel stacking approach for accurate detection of fake news. *IEEE Access*, 9:22626–22639.

Jin, Y. et al. (2022). Towards fine-grained reasoning for fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6339–6346.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *European Conference on Machine Learning (ECML)*.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Lifferth, W. (2018). Fake news. https://kaggle.com/competitions/fake-news. Kaggle.

Marco Tulio Ribeiro, Sameer Singh, C. G. (2016). "why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144.

McCallum, A. and Nigam, K. (1998). A comparison of event models for naive bayes text classification. In *Proceedings of the AAAI-98 Workshop on Learning for Text Categorization*, pages 41–48. AAAI Press.

Salton, G. (1986). Another look at automatic text-retrieval systems. *Communications of the ACM*, 29(7):648–656.

Vaswani, A. et al. (2017). Attention is all you need. In *Proceedings of NIPS*, pages 5998–6008.

Wang, J. et al. (2023). Tlfnd: A multimodal fusion model based on three-level feature matching distance for fake news detection. *Entropy*, 25(11):1533.

Yang, Y. and Cui, X. (2021). Bert-enhanced text graph neural network for classification. *Entropy*, 23(11):1536.

Zhang, X. and Bao, L. (2020). Fake news detection via nlp techniques: A review. *Journal of Computer Science and Technology*.

Zhou, L. et al. (2020). Stacked ensemble learning for fake news detection. *IEEE Access*, 8:21390–21401.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS