Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 19 No. 11 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijim.v19i11.55007

PAPER

Enhancing Fake News Detection via Stance Analysis: Leveraging Advanced NLP Techniques and Machine Learning Models

Mërgim H. Hoti(✉), Festina Qorrolli, Fisnik Spahija

University of Prishtina, Prishtinë, Republic of Kosovo

mergim.hoti@uni-pr.edu

ABSTRACT

Fake news detection is still a field of research that is in its infancy, and this is clearly evident as it has only recently gained significant attention from society. The use of machine learning algorithms and natural language processing (NLP) techniques offers valuable problem-solving opportunities to address these complex challenges. This study explores stance detection as a method to identify misinformation by examining the connection between article headlines and their corresponding body text. Utilizing the FNC-1 and FARN datasets, we apply advanced NLP methods and machine learning (ML) models, including logistic regression, XGBoost, and DistilBERT. Key preprocessing techniques such as lemmatization, named entity recognition (NER), sentiment analysis, and semantic similarity are employed to capture both linguistic and contextual features. The experimental results show that transformer-based models such as DistilBERT achieve superior performance compared to traditional approaches, particularly in accurately classifying nuanced stances. These findings highlight the crucial role of context-aware models in improving the accuracy of misinformation detection and demonstrate their potential for scalable, real-world applications.

KEYWORDS

fake news detection, stance detection, natural language processing (NLP), machine learning (ML), transformer models, misinformation analysis

1 INTRODUCTION

As widely acknowledged, news serves as a valuable means of acquiring information on specific issues, where individuals share their experiences on both general and specialized topics. However, not all news provides accurate information, particularly in light of technological advancements and the potential for misuse. Therefore, careful consideration of the source from which one obtains news should be prioritized equally with the importance of the news content itself. Failure to verify or identify credible sources

Hoti, M.H., Qorrolli, F., Spahija, F. (2025). Enhancing Fake News Detection via Stance Analysis: Leveraging Advanced NLP Techniques and Machine Learning Models. *International Journal of Interactive Mobile Technologies (iJIM)*, 19(11), pp. 39–50. https://doi.org/10.3991/ijim.v19i11.55007

Article submitted 2025-01-18. Revision uploaded 2025-03-08. Final acceptance 2025-03-12.

can result in the spread of misinformation, impacting a significant number of readers. The spread of misinformation especially impacts society, politics, and the economy by threatening public health and disrupting democratic processes. It quickly circulates through social platforms and forums, eroding trust in credible sources and intensifying societal polarization. The use of various techniques in daily life, particularly in the field of AI, is a challenging and complex process due to the underlying frameworks it employs. Similarly, the objective of this paper is to detect fake news through stance detection by applying various algorithms from natural language processing (NLP) and machine learning (ML). One of the tasks combating this development is automating the detection of fake news [1]. The challenges that arise from the nature of this task involve contextual understanding for subjective or ambiguous reasoning. The existing implementations have a need to improve in regard to contextual, cross-document, and semantic analysis for a reliable effort of detecting fake news [2].

Fake news [3] can be a product of misrepresentation of facts by a lack of deep understanding of the context on which the claims are made. Due to its complexity and diversity, it has become difficult to distinguish between satire, objective opinions, or controversial truths because of the subjective nature of the claims [3]. There is an absence of standardized evaluation metrics, and the dependency on limited datasets has further made the task difficult to implement in an optimized setting, thus building reliable automated systems.

The paper is structured as follows: Section 2 presents a literature review where it provides an overview of existing research on fake news detection as a whole, with a specific focus on stance detection methods using NLP and ML algorithms. Whereas Section 3 presents the methodology employed during this research. Next, Section 4 discusses models used for the preparation model and dataset before the application of the paper. In Section 5 are presented results and discussions, while in Section 6 are shown, conclusions and future work.

2 LITERATURE REVIEW

The identification of fake news is an increasingly prevalent issue, driven by the advancement and widespread adoption of technology, particularly with the continuous introduction of new autonomous models. The detection and management of fake news fall within the domain of NLP, which specifically focuses on identifying patterns that produce accurate results. Furthermore, alongside NLP techniques, various ML algorithms continue to be utilized in fake news detection. Additionally, different models demonstrate unique capabilities by employing specific architectures [3], in combination with additional algorithms [4], [5], [7], which significantly improve the accuracy of fake news detection.

Considerable work has been done towards traditional methods of stance detection by relying on content-based implementations [8]. Deep analysis of linguistic features, sentiment analysis, and stylistic markers are a few approaches used in these models, which prove an insufficient automatization for handling more sophisticated misinformation. This becomes difficult, especially in cases where the data is partially true with nuances of fake indicators [9]. Fact-checking is a well-proven effective method of dealing with the problem but has a time-intensive dimensionality, which, in combination with the availability of verified databases, results in an inefficient system. These challenges portray a need for an innovative method for dealing with fake news.

Authors [10] in this study have used a dataset from the FNC-1 dataset, where, after a careful analysis and preprocessed have generated their own dataset, which they have made public. Meanwhile, their findings indicate that the evaluation parameters

tend to favor the majority class, which is easier to predict, thus inflating the perceived effectiveness of the methods. To see how it will give solutions about majority classes, they used an F1-based metric that provides a more balanced system ranking. Then they identify the features and model architectures employed by these systems and propose a new feature-rich stacked LSTM model. This model matches the top-performing systems in overall performance but excels in handling minority class predictions.

Research such as [11], [12], [13], and [5] has also explored news propagation by analyzing how misinformation spreads across social media. While effective in domain-specific cases, this approach struggles on broader datasets that require validation from external sources. Also. On this part, findings of the paper highlight the need for more precise, equitable, and practical detection models and algorithms. We also emphasize the importance of distinguishing fake news detection from related tasks and underscore the critical role of NLP in combating misinformation.

Meanwhile, in this study [14], the authors explore fake news detection using three different ML classifiers, such as passive aggressive, Naïve Bayes (NB), and the last one, support vector machine (SVM). While text classification helps extract key features for detection, challenges remain due to the lack of comprehensive corpora. Experimental results on two public datasets show promising improvements in accuracy, highlighting the potential of ML in combating misinformation.

Hence, authors [15] treat neural and statistical combining external features to show stance detection as a subtask of fake news identification. This has aimed at determining the relationship between a news article's headline and body. Such an approach integrates neural embeddings from a deep recurrent model, statistical features from a weighted n-gram model, and handcrafted external features using feature engineering techniques. These features are combined through a deep neural layer to classify headline-body pairs into categories: agree, disagree, discuss, or unrelated. Extensive experiments demonstrate that our model surpasses state-of-the-art methods, including top submissions from the Fake News Challenge.

The authors [16], in their paper, conclude that information originates from any OSN user and quickly spreads, making the task of fact-checking news both time-consuming and resource-intensive. They take this opportunity to explore different ML techniques for automating fake news detection. This paper specifically focuses on detecting the stance of content producers—whether they support or oppose the subject of the content. According to the authors, after a long period of analyzing and assessing different competition groups, they identified three key steps for achieving high accuracy: (1) a multi-stage approach that integrates classical and neural network classifiers, (2) the extraction of additional text-based meta features from headline and article body columns, and (3) the utilization of recent pre-trained embeddings and transformer models.

3 METHODOLOGY

This paper adheres to a strict and comprehensive methodology for identifying the most optimal results. To this end, datasets were selected that were initially tested in one of the largest global events of this kind [17], where, in 2016, over 50 competing teams tested datasets containing fake news data. It is important to note that after preprocessing the FNC-1 and FARN datasets, both datasets were combined into a data frame, and a stance label ("FAKE" or "REAL") was added to each file. As observed, the participants came not only from the academic community but also from industry [18], [19], [20], predominantly consisting of groups or individuals specialized in data, NLP, ML, deep learning, and related fields. The main objective of this effort was to evaluate

stance detection and determine whether it constitutes an appropriate method for further processing and classifying news articles as “agree,” “disagree,” or “irrelevant.”

Itself, the project consists of the FNC-1 (Fake News Challenge) dataset [21], a widely recognized dataset used for stance detection, together with the FARN (Fake And Real News) dataset [22] for external evaluation, as it has been shown in Figure 1.

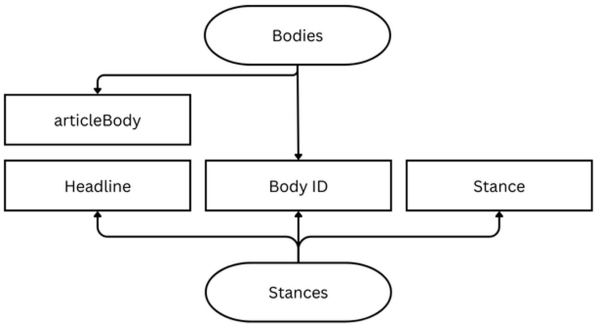

Fig. 1. The relation of attributes within the FNC-1 dataset

Feeding the model with the relationship between headlines and body texts, with the stance categorized into one of four classes: Agrees, Disagrees, Discusses, or Unrelated. Using these labels, we are able to train the model consistently and test it for fake news detection using stance detection.

We start the preprocessing step by cleaning the text to remove noise such as special characters, HTML tags, and extra spaces. Stop words were also filtered out to focus on meaningful content. We converted everything to lowercase to standardize the text and performed tokenization, which breaks the text into smaller chunks or words; then we performed lemmatization to reduce words to their base forms, which improves consistency in the data [23]. Finally, we split the dataset into training, testing, and validation sets to evaluate the model’s performance, as it is shown in Figure 2.

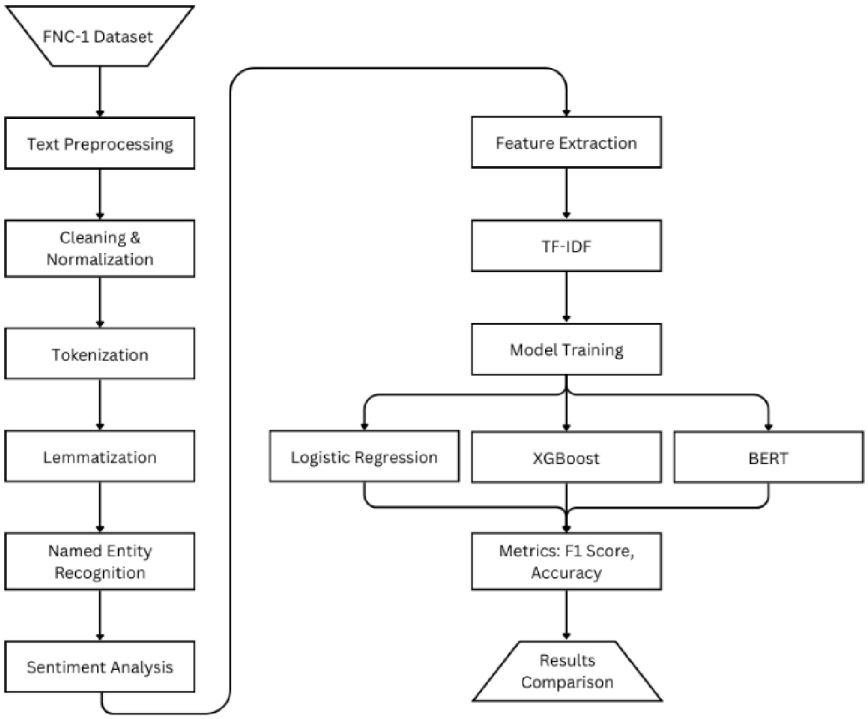

Fig. 2. Data flow diagram

A. Logistic regression: Feeding the model with the relationship between headlines and body texts is crucial, thus needing to represent the text in a format that a machine learning algorithm can process. Exploring several feature extraction techniques is made possible with traditional methods such as TF-IDF (term frequency-inverse document frequency). These methods help to quantify the importance of words in each text, allowing the model to identify which terms carried more weight when determining stance.

To capture the nuanced relationship between headlines and body texts, we have included named entity recognition (NER), sentiment analysis, and overlap features, each contributing distinct contextual insights to improve stance detection.

Named entity recognition extracts key entities such as names, locations, dates, and organizations from the headline and body text [24]. The idea is to identify whether the entities mentioned in the headline are consistently expressed in the corresponding body text. Discrepancies in entities could indicate a divergence in stance or potential misinformation. By encoding these entities as features, the model gains a more in-depth understanding of the alignment between the claim and text [25].

Sentiment analysis evaluates the tone in terms of emotional expression. Headlines often express a specific sentiment to offer reader engagement, which should align with the sentiment expressed in the body text. A headline with a strongly negative sentiment being paired with a neutral or positive sentiment in the body suggests a "Discuss" or "Unrelated" labeling. This way the model captures subtle tonal inconsistencies either as stance misalignment or incorrect framing [26], [27].

The degree of lexical, semantic, and entity overlap between the headline and the body text is also analyzed and processed. These features include:

- Word overlap: The proportion of shared words between the headline and body text, a useful metric for content alignment.
- Semantic similarity: Using word embeddings or sentence transformers to compute cosine similarity. This allows the model to capture meaningful context beyond exact word matches.
- Entity overlap: The count of common entities between the headline and body text. This is possible through the NER extraction step which reinforces context consistency.

By incorporating shallow lexical features and deep contextual insights, the model is prepared to address the complexities of stance detection in fake news scenarios.

4 USED MODELS

The experiments consist of using the following algorithms:

A. Logistic regression: Logistic regression is used for classification tasks; in our case, we use logistic regression to predict the stance of the body text towards the headline. Logistic regression uses a sigmoid function to map input features into probabilities for each class, and then the probability with the highest score is used to decide the category [28].

We decided to use logistic regression because the results gave us an initial understanding of how the dataset behaved with a traditional machine

learning approach. We standardized the dataset to make sure that all the features are on the same scale, and then we trained the model using training data. Also, there are several studies that are published and have used such an algorithm [29].

B. XGBoost: XGBoost stands for extreme gradient boosting and is known as a powerful and efficient algorithm. This algorithm builds multiple decision trees in sequence, where each one improves on the errors of the previous one. The model works with features extracted from the text to find relationships between input features and labels. This process allows it to handle complex patterns and make better decisions [30], [31]. Also, as a comparison of the same algorithms, such as in our case, we can see the research has used XGBoost and shown good results [32].

To train the model, we used standardized and preprocessed data, then tested it with testing data, and finally validated it with validation data to confirm the results.

C. DistilBERT: DistiBERT is a derived version of BERT; with its lighter and faster version, it's a good alternative to using it for fake news detection. It's efficient and a great choice when computational resources are limited. The model has a bidirectional approach, which means that it reads the text in both directions, understanding the context of the words by looking at the words before and after them. This approach helps to understand the meaning of the sentence more accurately [33].

We combine the headline and the body text into a single input by a special token to make it easier for the model to understand the relationship between those two pieces. Taking a pre-trained version of the model and training it further on our dataset to improve its accuracy, this process is called fine-tuning.

5 RESULTS AND DISCUSSIONS

To efficiently train and evaluate the stance detection models, the entire pipeline was implemented on Azure ML. The platform provides a scalable environment for training machine learning models and testing models. The environment was configured using pre-built virtual machines with GPU acceleration, ensuring that both the preprocessing steps and the training of computationally intensive models were executed efficiently.

To ensure the scalability and practical deployment of the stance detection system, the trained models and the associated scalers are serialized through Joblib. This creates easy access to the models to reuse them.

A. Logistic regression: We will measure the performance of each model, starting off with a confusion matrix. As seen in Figure 3, logistic regression has given a good performance on unrelated classification (Label 3) with 3518 true positives. However, some instances were misclassified as "Discuss" (Label 2), leading to confusion. There was a noticeable difficulty in distinguishing between "Agree" (Label 0) and "Discuss" (Label 2), as seen from the 472 samples from the "Discuss" class being misclassified as "Agree." Predictions for the "Disagree" (Label 1) class were particularly challenging, with a high number of instances incorrectly labeled as "Discuss."

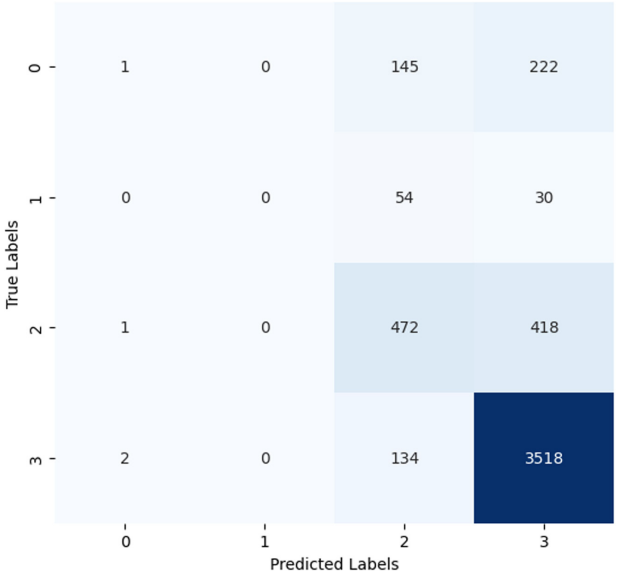

Fig. 3. Logistic regression confusion matrix for testing data

This suggests that the model is effective for separating "Unrelated" articles but struggles with more subtle distinctions such as between "Agree" and "Discuss."

B. XGBoost: The confusion matrix for XGBoost demonstrates improved performance across all classes. The "Unrelated" class (Label 3) achieved a high true positive count of 3,587, with fewer misclassifications compared to logistic regression. XGBoost showed a marked improvement in handling the "Discuss" class, correctly predicting 505 instances, although some were still misclassified as "Unrelated" (Label 3) or "Agree" (Label 0). Predictions for "Agree" (Label 0) and "Disagree" (Label 1) also showed enhanced precision, with fewer instances mislabeled in comparison, as it has been shown in Figure 4.

Overall, XGBoost outperformed logistic regression, particularly in handling nuanced distinctions between the stance classes, due to its ability to model complex, non-linear relationships in the data.

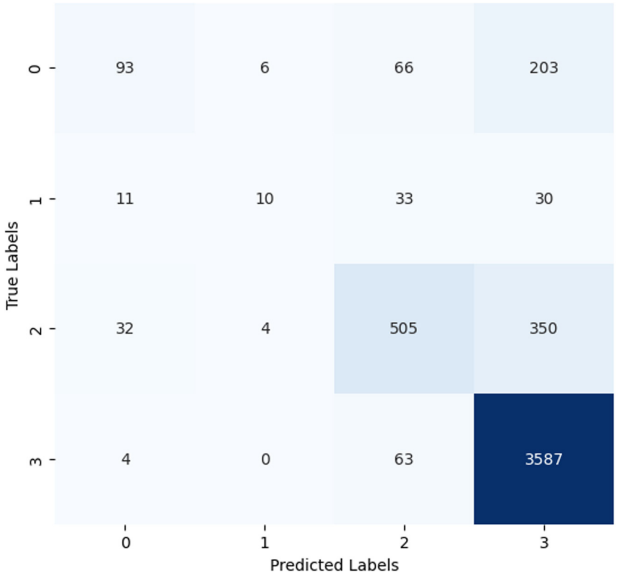

Fig. 4. XGBoost confusion matrix for testing data

C. DistilBERT: DistilBERT demonstrates superior performance across most classification metrics when compared to logistic regression and XGBoost. Even using not all the data from the dataset, its training accuracy of 93.55% and test accuracy of 90.2% indicate robust generalization to unseen data. The model excels particularly in correctly classifying "Unrelated" (Label 3) samples, achieving a substantial number of true positives. This reflects its ability to identify articles with no direct relationship to claims more effectively than the other models, as it has been shown in Figure 5.

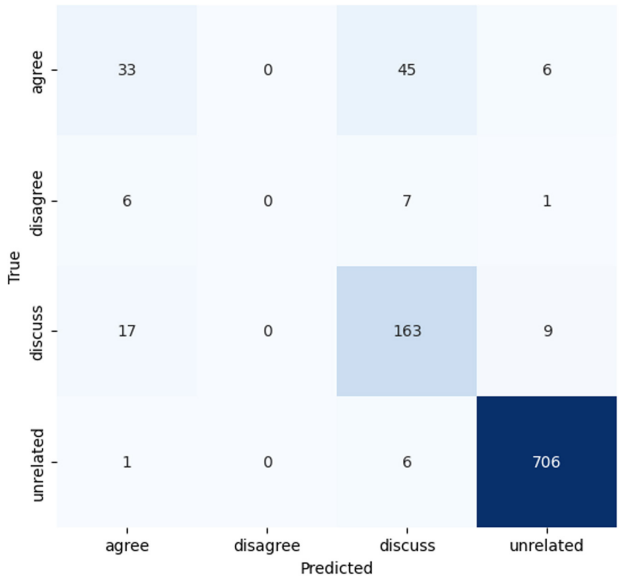

Fig. 5. DistilBERT confusion matrix for testing data

However, similar to logistic regression and XGBoost, some challenges persist in distinguishing between classes with nuanced relationships. For example, while DistilBERT performs better in separating "Agree" (Label 0) and "Discuss" (Label 2) classes, a certain degree of misclassification still exists, albeit at lower levels compared to logistic regression and XGBoost. The confusion matrix shows a notable improvement in minimizing errors across these categories, highlighting DistilBERT's ability to capture subtle contextual relationships that simpler models struggle with.

Table 1. Overall accuracy for models based on data splits

Metric	Logistic Regression	XGBoost	DistilBERT
Training Accuracy	80.13%	86.34%	93.55%
Test Accuracy	79.87%	83.95%	90.2%
Validation Accuracy	80.59%	83.61%	90.2%

In contrast to logistic regression, which exhibited considerable difficulty in predicting the "Disagree" (Label 1) class, and XGBoost, which showed moderate improvement, DistilBERT offers a more balanced performance across all classes. This improvement can be attributed to its transformer-based architecture, which effectively captures contextual dependencies in text data, as it has been shown in Table 1.

Overall, while logistic regression and XGBoost provide competitive performance in classifying "Unrelated" samples, DistilBERT's higher accuracy across all metrics underscores its superior capacity for nuanced stance detection in this dataset. This makes it a more reliable choice for applications where precise classification of relatedness and stance is critical.

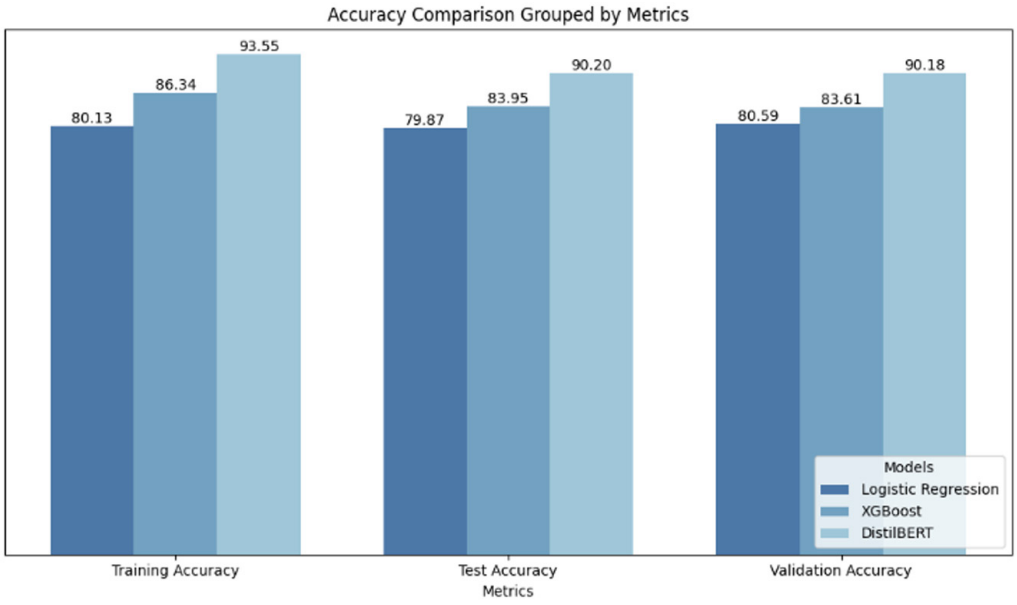

Fig. 6. Accuracy evaluation for training, testing, and validation data

Through external validation using the FARN dataset, we use the trained models with the FNC-1 dataset to evaluate the performance of these models with the corresponding dataset, as it has been shown in Figure 6. In Table 2 we see that both logistic regression and XGBoost have a great understanding of predicting accuracy.

Table 2. Overall accuracy for external validation

Metric	Logistic Regression	XGBoost
Training Accuracy	99.12%	100%
Test Accuracy	98.8%	99.72%
Validation Accuracy	98.75%	99.88%

Overall, FNC-1 proves to be a powerful source of stance detection. Through the extensive set of articles, we can expect a good percentage of models to perform exceptionally when dealing with fake news detection.

6 CONCLUSIONS

The detection of fake news through stance detection marks a significant advancement in combating misinformation within the increasingly digital information landscape. This study highlights the application of both traditional ML and transformer-based models, emphasizing their ability to process and classify relationships between headlines and body texts. Among the models examined, DistilBERT

demonstrated superior performance, achieving higher accuracy across training, testing, and validation phases due to its ability to capture nuanced contextual dependencies. In contrast, logistic regression and XGBoost, while effective in specific scenarios, showed limitations in distinguishing between closely related classes such as “Agree” and “Discuss”.

Comprehensive feature extraction techniques, including NER, sentiment analysis, and semantic similarity, substantially enhanced stance detection by providing deeper contextual insights. The primary contribution of this work lies in the integration of these feature extraction techniques with advanced algorithms to achieve greater reliability in stance detection. By refining the methodology for processing headline-body relationships and demonstrating scalable implementation using Azure ML, this study offers a significant step forward in automated fake news detection. Additionally, the study bridges the gap between traditional machine learning methods and modern NLP architectures, paving the way for further exploration in this critical domain.

Enhancing the proposed system’s robustness involves integrating natural language inference (NLI) to improve claim-evidence analysis. Additionally, experimenting with transformer models to achieve higher accuracy and refining semantic similarity measures are crucial steps. Developing a user-friendly GUI would facilitate manual testing and improve user interaction. These advancements aim to create a practical, adaptable framework for fake news detection across diverse datasets and real-world applications.

7 REFERENCES

[1] L. Yuan, H. Shen, L. Shi, N. Cheng, and H. Jiang, “An explainable fake news analysis method with stance information,” *Electronics*, vol. 12, no. 15, pp. 33–67, 2023. https://doi.org/10.3390/electronics12153367

[2] N. de Oliveira, P. Pisa, M. Lopez, D. de Medeiros, and D. Mattos, “Identifying fake news on social networks based on natural language processing: Trends and challenges,” *Information*, vol. 12, no. 1, p. 38, 2021. https://doi.org/10.3390/info12010038

[3] M. Umer *et al.*, “Fake news stance detection using deep learning architecture (CNN-LSTM),” *IEEE Access*, vol. 8, pp. 156695–156706, 2020. https://doi.org/10.1109/ACCESS.2020.3019735

[4] Z. Shahbazi and Y.-C. Byun, “Fake media detection based on natural language processing and blockchain approaches,” *IEEE Access*, vol. 9, pp. 128442–128453, 2021. https://doi.org/10.1109/ACCESS.2021.3112607

[5] A. H. Hoti, M. H. Hoti, H. Hoti, and A. Salihu, “Identifying fake news written on Albanian language in social media using Naïve Bayes, SVM, logistic regression, decision tree and random forest algorithms,” in *11th Mediterranean Conference on Embedded Computing (MECO)*, Budva, Montenegro, 2022, pp. 1–6. https://doi.org/10.1109/MECO55406.2022.9797147

[6] U. Sharma, S. Saran, and S. M. Patil, “Fake news detection using machine learning algorithms,” *International Journal of Engineering Research & Technology (IJERT), Special Issue*, vol. 9, no. 3, pp. 509–518, 2021. https://www.ijert.org/research/fake-news-detection-using-machine-learning-algorithms-IJERTCONV9IS03104.pdf

[7] H. F. Alsaif and H. D. Aldossari, “Review of stance detection for rumor verification in social media,” *Engineering Applications of Artificial Intelligence*, vol. 119, no. 4, p. 105801, 2023. https://doi.org/10.1016/j.engappai.2022.105801

[8] Bae *et al.*, "Natural language processing for assessing quality indicators in free-text colonoscopy and pathology reports: Development and usability study," *JMIR Medical Informatics*, vol. 10, no. 4, pp. 1–12, 2022. https://doi.org/10.2196/35257

[9] A. Hanselowski *et al.*, "A retrospective analysis of the fake news challenge stance detection task," in *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA, 2018.

[10] R. Oshikawa, J. Qian, and W. Y. Wang, "A survey on natural language processing for fake news detection," in *Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association*, Marseille, France, 2020, pp. 6086–6093.

[11] Z. Zhou, H. Guan, M. Bhat, and J. Hsu, "Fake news detection via NLP is vulnerable to adversarial attacks," in *Proceedings of the 11th International Conference on Agents and Artificial Intelligence*, Prague, Czech Republic, 2019, pp. 794–800. https://doi.org/10.5220/0007566307940800

[12] C. Dulhanty, J. L. Deglint, I. B. Daya, and A. Wong, "Taking a stance on fake news: Towards automatic disinformation assessment via deep bidirectional transformer language models for stance detection," *arXiv preprint arXiv:1911.11951*, 2019. https://doi.org/10.48550/arXiv.1911.11951

[13] S. Ahmed, K. Hinkelmann, and F. Corradini, "Development of fake news model using machine learning through natural language processing," *World Academy of Science, Engineering and Technology, International Journal of Computer and Information Engineering*, vol. 14, no. 12, pp. 454–461, 2020. https://arxiv.org/pdf/2201.07489

[14] G. Bhatt *et al.*, "Combining neural, statistical and external features for fake news stance identification," in *9th International Workshop on Modeling Social Media (MSM 2018)*, Lyon, France, 2018. https://doi.org/10.1145/3184558.3191577

[15] I. Alsmadi, I. Alazzam, M. Al-Ramahi, and M. Zarour, "Stance detection in the context of fake news—A new approach," *Future Internet*, vol. 16, no. 10, p. 364, 2024. https://doi.org/10.3390/fi16100364

[16] B. Sean, S. Doug, and P. Yuxi, "Talos targets disinformation with fake news challenge victory," *Cisco Talos*, 2016. [Online]. Available: https://blog.talosintelligence.com/talos-fake-news-challenge/ [Accessed: Dec. 27, 2024].

[17] X. Qiu *et al.*, "Pre-trained models for natural language processing: A survey," *Sci. China Technol. Sci.*, vol. 63, pp. 1872–1897, 2020. https://doi.org/10.1007/s11431-020-1647-3

[18] S. Naik and A. Patil, "Fake news detection using NLP," *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, vol. 9, no. 12, pp. 2022–2028, 2021. https://doi.org/10.22214/ijraset.2021.39582

[19] H. Karande *et al.*, "Stance detection with BERT embeddings for credibility analysis of information on social media," *PeerJ Comput. Sci.*, vol. 7, p. e467, 2021. https://doi.org/10.7717/peerj-cs.467

[20] Fake News Challenge, "Fake news challenge," Fake News Challenge Stage 1 (FNC-I): Stance Detection, 15 06 2017. [Online]. Available: http://www.fakenewschallenge.org/ [Accessed: Dec. 19, 2024].

[21] Clmentbisaillon, "Fake-and-real-news-dataset," *Kaggle*, 2024. [Online]. Available: https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset [Accessed: Dec. 25, 2024].

[22] N. Ghimire and S. Shrestha, "Fake news stance detection using deep neural network," *LEC Journal*, vol. 4, no. 1, pp. 49–53, 2022. https://doi.org/10.3126/lecj.v4i1.49366

[23] M. Alsafadi, "Stance classification for fake news detection with machine learning," in *International Conference on Basic Sciences, Engineering and Technology (ICBASET)*, Marmaris/Turkey, 2023, pp. 191–198. https://doi.org/10.55549/epstem.1344457

[24] O. Etzioni *et al.*, "Unsupervised named-entity extraction from the Web: An experimental study," *Artificial Intelligence*, vol. 165, no. 1, pp. 91–134, 2005. https://doi.org/10.1016/j.artint.2005.03.001

[25] Y. Mejova, *Sentiment Analysis: An Overview*. Iowa State, IA: University of Iowa, 2009.

[26] S. A. Kadam and S. T. Joglekar, "Sentiment analysis: An overview," *International Journal of Research in Engineering & Advanced Technology*, vol. 1, no. 4, pp. 1–7, 2013. http://www.ijreat.org/Papers%202013/Issue4/IJREATV1I4016.pdf

[27] B. Vimal, "Application of logistic regression in natural language processing," *International Journal of Engineering Research*, vol. 9, no. 6, pp. 1–4, 2020. https://doi.org/10.17577/IJERTV9IS060095

[28] N. N. Prachi *et al.*, "Detection of fake news using machine learning and natural language processing algorithms," *Journal of Advances in Information Technology*, vol. 13, no. 6, pp. 652–663, 2022. https://doi.org/10.12720/jait.13.6.652-661

[29] Z. Li, Q. Zhang, Y. Wang, and S. Wang, "Social media rumor refuter feature analysis and crowd identification based on XGBoost and NLP," *Applied Machine Learning*, vol. 10, no. 14, pp. 1–15, 2020. https://doi.org/10.3390/app10144711

[30] Z. Li *et al.*, "Social media rumor refuter feature analysis and crowd identification based on XGBoost and NLP," *Appl. Sci.*, vol. 10, no. 14, pp. 1–15, 2020. https://doi.org/10.3390/app10144711

[31] Z. Khanam, B. N. Alwasel, H. Sirafi, and M. Rashid, "Fake news detection using machine learning approaches," *IOP Conf. Series: Materials Science and Engineering*, vol. 1099, no. 1, p. 012040, 2021. https://doi.org/10.1088/1757-899X/1099/1/012040

[32] S. K. Akpatsa *et al.*, "Online news sentiment classification using DistilBERT," *Journal of Quantum Computing*, vol. 4, no. 1, pp. 1–11, 2022. https://doi.org/10.32604/jqc.2022.026658

8 AUTHORS

Mërgim H. Hoti has completed Dr. Sc. in Computer Science. He is currently a Teaching Assistant at the University of Prishtina "Hasan Prishtina", Faculty of Electrical and Computer Engineering. His research interests include Data Science, Machine Learning, Artificial Intelligence, and Cybersecurity, with several publications in these areas (E-mail: mergim.hoti@uni-pr.edu).

Festina Qorrolli is currently pursuing master's degree in computer and software engineering at the University of Prishtina. She works as a Data Engineer at Raiffeisen Tech Kosovo. Her areas of interest include Machine Learning, NLP, Data Science, AI, and Metaheuristics (E-mail: Festina.qorrolli@student.uni-pr.edu).

Fisnik Spahija is currently pursuing master's degree in computer and software engineering at Faculty of Electrical and Computer Engineering, University of Prishtina. He is currently working as a Full-Stack Software Developer at Swiss Code. His interests reside on topics related to Machine Learning, NLP, AI and Metaheuristics (E-mail: fisnik.spahija@student.uni-pr.edu).