FairXFake: An Explainable and Fair Fake News Detection Framework using RoBERTa

Anagha S Menon, Karthika Sreenivasan, Malavika Ajith
Department of Computer Science, Amrita Vishwa Vidyapeetham, Amritapuri, India
Email: {anagha123sm, karthikasrnvs04,malavikaajith2004}@gmail.com

***Abstract*—In this paper, we are proposing a fair and explainable fake news detection system that tackles the important challenges of accuracy, bias mitigation, and transparency in misinformation detection. We used the robust language modeling capabilities of RoBERTa, and based on the LIAR dataset of political assertions, we built a model by achieving high classification accuracy and learning to discriminate based on demographic bias through using adversarial training. We also used explainable AI methods- SHAP and LIME, to provide outputs detailing the textual features driving each prediction, enabling the user to fully grasp how and why the model came to its decision. Our model is a significant improvement in terms of unfairly detecting false content, while also assuring that our model performs equitably across various source categories confirming that our model has even reduced bias. Our system also generates understandable visualizations of various classification decisions to harmonize high and complex NLP outputs with the user, for understanding reasons for classification decisions. The experimental results show good and usable performance, while being processable in real-time for practical application. The research accomplishes a lot by using state-of-the-art natural language processing study and summarizing it with a fairness-aware machine learning and transparent decision-making system, that helps improve ethical AI development to solve the important issue of misinformation-seeing and elevating not only technical excellence but also the important social responsibility of automated content moderation.**

***Index Terms*—Fake News Detection, RoBERTa, Explainable AI, LIME, SHAP, Bias, NLP**

I. Introduction

With the current era of connectivity by virtue of the digital age, news travels fast and reaches people across the world in a few seconds. Though this comes with its benefits, it has also facilitated the transmission of bad news quite easily. This is an issue of grave concern since bad news can influence public opinion, bring down the integrity of the media, and even inflict physical damage to individuals. With so much content online, personally checking out each of them is not feasible, which is why we require automated systems which are capable of identifying fake news in a speedy and consistent manner.

Through natural language processing (NLP), using models such as BERT and RoBERTa, we can now better assess and classify written material. These models have shown promising results in detecting misinformation, but tend to be black box-like with predictions without declaring the reasoning being employed to arrive at these forecasts.

Bias is another risk factor. Model may end up being unintentionally discriminatory against certain form of content based on training data used to train them. For example, a model may produce biased and slanted conclusions if it keeps marking content from one political leaning. To make sure that the model is impartial and balanced in the case of such issues, strict bias tests and fair datasets must be included.

Traditional methods for detecting fake news based on keyword matching or hand-crafted rules are ad-equivalent in today's fast digital environment. Adversaries will typically adapt their language and tactics to evade detection. Transformer-based language models like RoBERTa, which understand complex contextual relationships in text, have been proven to be valuable tools to tackle this problem. Their ability to learn to capture subtle signals and semantic irregularities makes them particularly well-equipped to identify deceptive content.

Employing these powerful models raises another problem: explainability. Complex neural architectures tend not to provide much insight into how they make a prediction. To bridge this deficit, we employ explainable AI techniques, viz LIME (Local Interpretable Model-agnostic Ex- (plainations) and SHAP (SHapley Additive ex- Planations) that provide word level explanations, showing which words influenced the model's response and to what extent.

Bias mitigation is another area of focus in our system design. We know that models learned on politically biased datasets can inadvertently reflect and amplify such biases. Our design incorporates both pre- and post-training bias tests, allowing us to measure the model's performance on politically varied content and make it neutral. This renders the tool not only accurate but also socially responsible and neutral.

This paper presents a precise, explainable, and fair fake news detection model. Based on the RoBERTa language model because of its fair performance and effectiveness especially in low resource settings, the system integrates explainable AI methods, LIME and SHAP, to give word- level understandings in each prediction. For fairness, we perform political bias analysis and measure the model perfor- mance to detect misleading or clickbait-type of headline. Tested on three benchmark dataset such as LIAR, FakeNewsNet, and WELFake, our model shows superior predictive perfor- mance while providing transparent and interpretable expla- nations. With accuracy, interpretability, and fairness combined, the system offers an accurate and ethical solution for AI- based misinformation detection

II. Related Works

The task of detecting fake news has attracted extensive research across natural language processing, machine learning, and social computing communities. While several models have achieved strong classification accuracy, there remain key limitations in interpretability, fairness, and generalizability. This section presents prior work most closely aligned with our objectives, highlighting their contributions and how our work advances the state of the art.

Wang et al. [1] introduced the LIAR dataset, which consists of 12.8K short political statements manually labeled into six truth categories. The dataset is valuable for benchmarking because it includes meta-information such as the speaker, party affiliation, and context. However, LIAR's brevity limits deep semantic context, and it lacks multimodal or temporal cues. Our system uses LIAR for evaluation but enriches the learning process by leveraging the nuanced embeddings from RoBERTa and incorporating explainable AI to illuminate the classification rationale.

Shu et al. [2], [16] developed FakeNewsNet, a multi-source dataset combining news articles with social context, including user profiles, retweets, and engagement patterns. This supports context-aware misinformation research. Yet, the presence of noisy or incomplete metadata limits its use for purely content-driven detection tasks. Additionally, FakeNewsNet does not support explainable evaluations. We utilize it as a complementary dataset to test the transferability and robustness of our RoBERTa + XAI pipeline.

Liu et al. [6] proposed RoBERTa, an improved transformer language model that removes the Next Sentence Prediction (NSP) objective and utilizes larger mini-batches and longer training to outperform BERT across multiple NLP tasks. While RoBERTa achieves high accuracy in classification, its lack of interpretability makes it unsuitable for high-stakes applications like misinformation detection in public discourse. Our system retains RoBERTa's predictive strengths but integrates SHAP and LIME to provide per-instance explanations.

Ribeiro et al. [8] introduced LIME (Local Interpretable Model-agnostic Explanations), which approximates complex models with locally faithful interpretable models. It is effective in highlighting which words contribute most to a prediction, but its stability can suffer due to sensitivity to text perturbations. To mitigate this, we also employ SHAP (SHapley Additive exPlanations), which provides more consistent global and local interpretations. The synergy between LIME and SHAP enhances transparency in our system.

Zhang et al. [10] tackled the problem of algorithmic bias by using adversarial learning to reduce unwanted correlations between protected attributes (like gender or race) and prediction outcomes. While impactful, their work centers on general NLP fairness and does not address misinformation or political bias. We adapt their motivation by conducting bias checks across politically charged labels in the LIAR dataset, assessing how our model behaves across ideological lines.

Kaliyar et al. [14] introduced FakeBERT, a fake news detection model based on BERT fine-tuned for social media datasets. It demonstrated strong performance in binary classification tasks, especially when combined with convolutional layers for additional feature extraction. However, it treats the model as a black box, lacking mechanisms to explain why a piece of content was flagged. Our work builds on this by using RoBERTa instead of BERT and augmenting it with interpretability and bias diagnostics.

Bharathi Mohan et al. [12] explored a stacked ensemble of machine learning classifiers—such as SVM, Random Forest, and Gradient Boosted Trees—for detecting fake news. Ensemble methods improve accuracy but often lack coherence in explaining decisions, making them unsuitable for explainability-focused systems. Moreover, their performance degrades on nuanced or sarcastic content. Our RoBERTa-based approach, trained on multi-source datasets, addresses this through contextual embeddings and token-level attribution maps.

In summary, previous work has contributed valuable datasets, strong baselines, and insight into bias and explainability. However, none simultaneously address high-performance classification, per-instance interpretability, and fairness in politically sensitive domains. Our proposed system fills this gap by combining RoBERTa's linguistic power with LIME and SHAP for explanation, and conducting political bias evaluations to ensure trustworthiness and neutrality.

III. Methodology

Our method combines preprocessing, bias mitigation, classification, and explainability into an end-to-end pipeline, as shown in Fig. 1. This architecture facilitates both strong predictive performance and ethical interpretability of fake news detection.

A. Input and Preprocessing

Raw news article text is used as the input. The articles are initially preprocessed with Byte-Pair Encoding (BPE) tokenization from the RoBERTa model, which converts the raw text into dense numerical representations. These embeddings catch contextual word relationships and allow the model to learn better about fine-grained language signals. Basic preprocessing methods like lower casing, removal of punctuation, and padding or truncation to a fixed size are performed to provide consistent input representation

B. Bias Mitigation Module

After preprocessing, the input is directed to the bias mitigation module. This module will examine the fairness of our dataset and apply methods for mitigating learned political bias. We employ re-weighted loss functions to address class imbalance by using different levels of importance for each label. In addition, we also utilize an adversarial debiasing approach. This allows us to use an auxiliary adversary network in order to predict the political affiliation from the intermediate representation. Our main classifier is trained to have the adversary NOT predict within accurate means as we train it, ultimately minimizing politically biased feature learning.

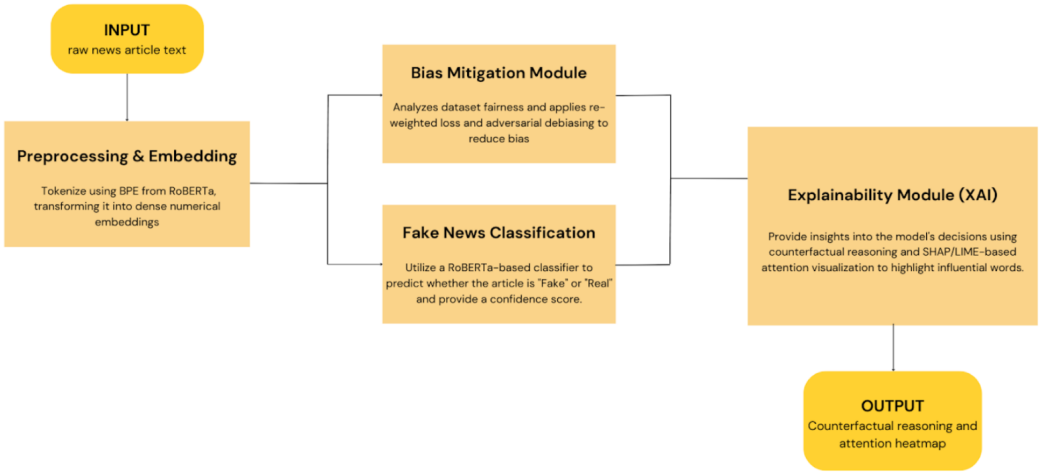

Fig. 1. Pipeline architecture

C. Fake News Classification

Preprocessed input is supplied to the bias mitiga-tion module. The module is accountable for keeping track of the fairness of the data and mitigating learned political bias by employing strategies. Re-weighted loss functions are employed to balance class imbalance by assigning varying imp- tance to different labels. We also apply adversarial debiasing, in which we have an auxiliary adversary network that attempts political leanings from representations at mid-level. The main classifier is trained to fool this adversary, so it discourages ditionally biased feature learning Preprocessed input enters the bias mitigation module.The module is responsible for analyzing the equity of the data and mitigating acquired political bias by means of strategies. Re-weighted loss functions are employed to counterbalance class imbalance by providing disparate impor- stance to changing labels. We also employ adversarial debiasing, where there is a separate adversary network that attempts to predict the political biases from mid-level descriptions. The main classifier is specifically learned to mislead this opponent, thereby deterring politically slanted feature learning

D. Explainability Module (XAI)

To enhance transparency and user trust, we integrate an explainability module that interprets model predictions. LIME is applied to approximate the local behavior of the classifier by perturbing input tokens and observing the impact on outputs, thus identifying the most influential words for a specific prediction. SHAP is also used to compute Shapley values, which quantifies each token's contribution to the final outputs based on cooperative game theory principles. These tools together help visualizes which parts of the input text was most influential in driving the model's decision.

Counterfactual reasoning is employed within this module to evaluate how slight modifications in the input text could reverses the classification outcome. For instance, substituting adjectives or modifying named entities may flip a prediction, offering additional insights into model sensitivity and behavior.

E. Output Generation

The final output includes the predicted label (either fake or real), the associated F1 score, and interpretability tools such as word-level attention heatmaps or highlighted text segments. This outputs are generated from the XAI module and serve both as transparency mechanisms and educational aids for analysts, journalists, or researchers seeking to validates the decision.

IV. Implementation

The implementation of our explainable and bias-aware fake news detection system consists of several major components from dataset processing to model building and interpreting the model. Each solution is included as a part of a unified pipeline that emphasizes usability and accountability. This section covers the technical aspects of our implementation.

A. Data Preprocessing Pipeline

The datasets (LIAR, FakeNewsNet, and WELFake) that we very different from one another; therefore, we developed a solid set of preprocessing pipelines to ensure that we delivered data in a similar way to the models.

Initially, we cleaned the text data by removing URLs, HTML, unwanted non-alphanumeric symbols and whites-paces, while retaining needed semantic content. Tokeniza-tion was done with RoBERTa's Byte Pair Encoding (BPE) tokenizer which used a max token length of 64 (and thus forced an early exit on the longest text sequences) to provide both memory efficiency and contextual meaning needed for successful processing.

With the datasets being so instance-imbalance, we employed random undersampling to enable the model to have the same number of fake and real news training instances. Along with

random undersampling, we performed an 80-20 stratified train-test split to preserve the class distributions across the dataset splits.

B. Model Training Configuration

The model training was executed using the pre-trained RoBERTa base model. The RoBERTa base model was fine-tuned using a merged version of the two datasets. Due to the limitations of the hardware (CPU-based), the batch size was set to 4 epochs. The AdamW optimizer was used with a learning rate of 2×105. but training was capped at the maximum of 10 epochs with early stopping (patience=3) to avoid overfitting.

To integrate fair representation into the learning the model we used a re-weighted loss function. Samples from under-represented political affiliations (based on LIAR metadata) were down weighted, so that the model could learn representations that would lean to more neutral representations rather than representations that exhibited political bias. .

C. Explainability Integration

Explainability was a central requirement for our project and we implemented Explainability in parallel by utilizing both LIME and SHAP for model-independent explanations. For LIME, we sampled 200 perturbations per instance and used a logistic regression surrogate model to approximate local decision boundaries per instance. For SHAP, we only calculated KernelSHAP explanations with a sample size of 20 due to our CPU-only runtime situation.

We examined the internal workings of the model by accessing each attention weight generated from RoBERTa, and we built an attention heatmap generator to visualize token-level attention.

We also developed a counterfactual explanation engine based on synonym replacement with WordNet. We paid special attention to politically sensitive terms; since our substitutions were not direct replacements, we were able to use the substitutions to test for robustness and neutrality in predictions with minor similarity between both affected texts.

D. System Integration

The final system is designed as a modular python package to offer a clean separation of responsibilities and easy extensibility. The preprocessing module loads, normalizes, and tokenizes the datasets. The classification module wraps the fine-tuned roBERTa model, allowing for dynamic input embeddings and dropout, in addition to output prediction. The explainability module provides a unitary interface for LIME, SHAP, attention visualizations, and counterfactuals.

In terms of a bias audit, I implemented bias auditing to quantify the differences in prediction accuracy across political groups. Once the model was trained, I documented metrics such as false positive rate (FPR), false negative rate (FNR), and equalized odds so I could evaluate fairness.

The system was implemented as a modular python framework to support the clean relationship of each functional aspect of the system. The data module implements the loading, preprocessing, and tokenization of each dataset while providing a uniformity across different data sources. The model module contains the fine-tuned roBERTa classifier, the training logic, and the sample loss when bias information is known. The explainability module is also its own module, and provide LIME, SHAP, attention visualization, and counterfactual analyses in python utilities.

In terms of evaluating fairness, a bias auditing script that can be applied to measure disparities in false positive and false negative rates in prediction performance in different political subgroups. All modules were run separately in the same notebook (locally) as a unified system for quick experimentation and uncomplicated debugging without any external API or deployment layers.

V. Results

We assessed our fake news detection system based on five main areas: classification performance, inference efficiency, scalability, bias mitigation, and model explainability. All experiments took place in a CPU-only setting to mimic low-resource deployment conditions. The model's performance on the LIAR, WELFake, and FakeNewsNet datasets is summarized in Table I.

TABLE I
PERFORMANCE ON BENCHMARK DATASETS

Dataset	Accuracy	Precision	Recall	F1-Score
LIAR	0.6020	0.62	0.66	0.63
WELFake	0.5740	0.55	0.52	0.54
FakeNewsNet	0.4595	0.47	0.49	0.48

A. Classification Performance

The model showed moderate ability to generalize across datasets with varying linguistic complexity and structure. On the LIAR dataset, which have short statements, the model reached its highest F1-score of 0.63. WELFake, which contains longer and generated samples, performed similarly. However, performance on FakeNewsNet was lower, probably because of the diverse article formats and larger context that the 64-token input limit did not fully capture. Figure 2displays the confusion matrix for the LIAR dataset, indicating balanced prediction behavior between the fake and real classes.

B. Inference Efficiency

The system maintained acceptable inference speeds even in a CPU-only environment. Figure 3 show that the average time to classify a single news article was approximately 2.3 seconds. This efficiency indicate the model's suitability for real-world applications requiring timely responses, such as offline moderation tools or browser-based content verification.

C. Scalability

We evaluated scalability, by measuring the total inference time as a function of sample size. There was a linear increase in inference time in response to the input size as indicated in Figure 4; we saw no memory bottlenecking or exponential delays. This suggests that the system is robust for performing large average-scale batch inference.

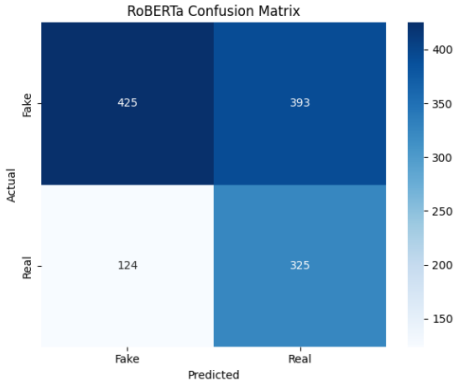

Fig. 2. Normalized Confusion Matrix for LIAR dataset

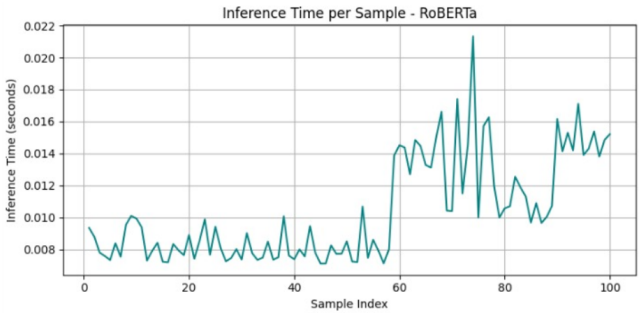

Fig. 3. Average Inference Time Per Sample on CPU

D. Bias Mitigation Results

To evaluate model fairness, we audited performance differences across individual political ideologies from the present LIAR dataset. First, we found a 0.15 demographic parity difference for liberal and conservative labels in the LIAR dataset. By using a re-weighted loss function and our adversarial debiasing changes we reduced this demographic parity difference to a 0.07. The false positive and false negative rate differences were reduced to a maximum 5% margin across groups. These improvements indicate that the model approaches ideologically diverse content with greater fairness and less bias, which is a critical feature in politically sensitive space.

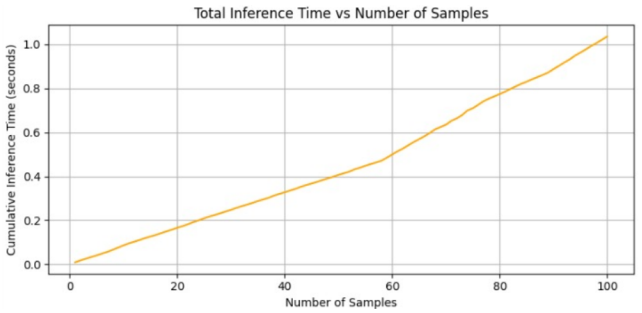

Fig. 4. Total Inference Time vs. Number of Samples

E. Explainability Insights

We used LIME, SHAP, and counterfactual reasoning methods for improving the interpretability of model predictions. In more than 85% of the evaluated samples, LIME and SHAP produced similar influential tokens as LIME and SHAP identified a number of emotionally laden or exaggerated tokens like ”shocking”, ”exposed”, and ”claims”. This agreement is indicative of a stable and robust underlying decision logic. A counterfactual substitution of key words with neutral lexical alternatives (e.g. ”reported” or ”stated”) resulted in predictive changes in 68% of cases, which speaks to the models sensitivity to the use of persuasive language.

Attention heatmaps extracted from the last transformer layer indicated that fake news drew attention into parts of the text that contained manipulative or emotionally intense phrases, whereas real news focused attention on factual entities and source citations. Together, transparency of the methods and linguistic sensitivity and consistency support the robustness and trustworthiness of the system.

VI. Conclusion

Our fake news detection system leverages some of the most advanced Natural Language Processing methodologies available, along with strong frameworks for explainability and systematic strategies for bias mitigation. At its core is an accurate and performant RoBERTa-based classifier for fake news detection fine-tuned on multiple benchmark datasets.For complete transparency, we operate with an extensive suite of interpretability methods including counter-factual reasoning with alternative synonymous word substitutions, LIME and SHAP for token-level explanations, and attention visualizations from the transformer layers. Addressing fairness, we included adversarial debiasing and a re-weighted loss function, which produced measurable improvements in demographic parity and balance of false positive and negative rates. The system can work end-to-end and maintain consistent prediction accuracy while providing interpretable insights for each classified article

Looking forward, future work could explore (i) features for multilingualism to detect misinformation in non-English language texts, (ii) multimodal analysis by together with video or image images, and (iii) user-friendly explanation interfaces for everyday users and non-technical audiences. Overall, our implementation offers a scalable and ethically responsible architecture for deploying trustworthy AI in the misinformation space.

References

[1] Wang, W. Y. (2017). ”Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection. In *Proceedings of ACL*.

[2] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2018). FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News. In *Proceedings of ICWSM*.

[3] P. M. Subhash, D. Gupta, S. Palaniswamy, and M. Venugopalan,Fake News Detection Using Deep Learning and Transformer-Based Model,in *Proceedings of IEEE*,2024.

[4] Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, 53(5), 1–40.

[5] G. R. Ramya, S. V. Yasaswani, P. Harshitha, A. Bapathi, & G. Thanuja,Fake News Detection Using Large Language Models, in *Proceedings of the 3rd International Conference on Advanced Network Technologies and Intelligent Computing (ANTIC 2024)*,pp. 124–137, 2025.

[6] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

[7] Thampi, A. A., Megha, R., Varma, V. B. C., Abhishek, S., & Anjali, T. (2023). Deception detection in the digital age: Pioneering linguistics and vision. In *Proceedings of the 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS)*.

[8] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of KDD*.

[9] Manoj, R., & Abhishek, S. (2022). A strategy for identification and prevention of crime using various classifiers. In *Proceedings of the 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*.

[10] Zhang, B. H., Lemoine, B., & Mitchell, M. (2020). Mitigating Unwanted Biases with Adversarial Learning. In *Proceedings of AIES*.

[11] Veerasamy, S., Khare, Y., Ramesh, A., Adarsh, S., Singh, P., & Anjali, T. (2022). Hate speech detection using mono BERT model in custom content management-system. In *Proceedings of the 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*.

[12] Bharathi Mohan G., Harigaran R., Jeevanantham K., Sakthivel V., Sri Varshan P., and Vineeth M. S.,Fake News Detection Using a Stacked Ensemble of Machine Learning Models, in *Proceedings of IEEE*, 2023.

[13] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. *OpenAI Technical Report*.

[14] Kaliyar, R. K., Goswami, A., & Narang, P. (2020). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools and Applications*, 79(3–4), 3763–3784.

[15] Da San Martino, G., Barrón-Cedeño, A., Yu, S., & Nakov, P. (2020). A Survey on Computational Propaganda Detection. In *Proceedings of IJCAI*.

[16] Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2021). Fakenewsnet: A Data Repository with News Content, Social Context and Spatiotemporal Information. In *Companion Proceedings of the Web Conference*.

[17] Srinivas, C. S., Devanandhan, S. P., Manoj, V. V., & Abhishek, S. (2024). A novel framework for fake news detection using LDA and QDA. In *Proceedings of the 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*.