Fake News Detection for Cybersecurity Using NLP and Deep Learning

Literature Review (E1)

Research Methods

Date:

Supervisor:

Name:

Student ID:

Abstract

Spreading of fake news is increasingly becoming a major threat to cybersecurity because fake information is employed to deceive people, propagate harmful content, or carry out phishing attacks. The purpose of this research is to develop a deep learning model that uses two potent methods, Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM), to identify fake news. The CNN will be useful for detecting patterns in the text, the LSTM will understand the meaning and flow of content. The model will be trained and tested independently on two popular datasets- FakeNewsNet and WELFake to test how it performs against different types of news. The project will employ the commonly used measures of performance accuracy, precision, recall, and F1-score. The ultimate aim is to develop an accurate and reliable tool which will help protect people and organizations from the dangers that emerge from fake news online.

Keywords: Fake News Detection, Cybersecurity, Deep Learning, NLP, CNN, LSTM, Text Classification

Table of Contents

1. Introduction

Fake news has become a dangerous cybersecurity threat of the digital age, and not just a social or political concern by any means. Nowadays, it is widely used by malicious actors to mislead its users, launch the distributions of malware, enact phishing campaigns, and manipulate public behavior using misleading stories (Elbes et al., 2023; Pal & Kumar, 2024). These organized misinformation barrages do not only confuse facts but erode the trust and security of online systems, subjecting individual users and national infrastructure to danger.

Cyber attackers are however getting more inclined to use fake news to gain entry and carry out technical attacks. As an example, misinformation campaign is mostly employed to make people click malicious links or download malicious files, or even provide personal information (Wani et al, 2024). This combination of social engineering and cyber deception shows the need for automated systems that can detect fake news in a reliable and transparent manner.

Such traditional detection methods as Naïve Bayes, Logistic Regression, and SVM have been the basis, yet usually unable to cover the subtleties of linguistic patterns and the dynamic nature of fake content (Cîrnu et al., 2023; Amin, 2024). These models also have the issue with generalising across datasets and data that are imbalanced/noisy. Some of the newer approaches through the use of deep learning which include CNNs and LSTMs had proved to be quite useful in learning both semantic and contextual features of text (Waheed & Azfar, 2025; Hashmi et al., 2024).

This research proposes a hybrid model that combines **Convolutional Neural Network** structure (**CNN**) with the **Long Short-Term Memory** (**LSTM**) networks to take advantage of the strengths of both architecture types. CNNs excel in discovering the local language cues and semantic structures in text, whereas LSTMs can model long-term dependencies as well as the contextual flow. In combination, they provide a powerful framework for examining of news articles and putting them in the categories of fake or real.

To measure the effectiveness of the model, this study will train and test the system independently on two widely used benchmark datasets, **FakeNewsNet** containing news

content and social context and **WELFake**, with an extensive and balanced real and fake news articles. These datasets allow for full testing of various and styles of misinformation.

Besides common performance metrics (accuracy, precision, recall, F1 score) the research will also examine explainable AI (XAI) methods to enhance transparency of a model and the trust on the part of users, both of which are critical qualities for utilization in actual cybersecurity settings (Sallami & Aïmeur, 2024).

Through addressing critical weaknesses of accuracy, generalizability, and interpretability, this project seeks to create a powerful and reliable fake news detection system, which can further be helpful in protecting digital spaces from misinformation threats.

2. Literature Analysis

This section introduces the related work conducted in the past by different researchers. This section includes different approaches they used, their conclusions, findings and drawbacks.

The detection of fake news has emerged as one of the key research areas because of its significant impact on cybersecurity. Various machine learning and deep learning solutions have been proposed by researchers to mitigate the increasing concern of misinformation in digital platforms. Traditional machine learning algorithms such as Naïve Bayes, Support Vector Machines (SVM), Random Forest, and Logistic Regression were heavily used in early detection systems. These models showed reasonable performance with handcrafted features and statistical techniques. For example, Cîrnu et al. (2023) found that Passive Aggressive SVC recorded 92.22% accuracy compared to Naïve Bayes, which was only at 79.07%. Amin (2024) was also able to achieve high accuracy using SVM, Logistic Regression and Random Forest classifiers. However, these methods do not have the depth to represent semantic meaning and complex sentence structure, which cause performance issues on diverse or time sensitive data.

To overcome these limitations, deep learning approaches come out as more powerful alternatives. There are such models as the Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and hybrids that have demonstrated significant improvements. Waheed and Azfar (2025) had found out that CNN and LSTM models achieved 85.4% and 88.2% respectively, and when combined, they produced 91.5%. When BERT was merged with LSTM, the accuracy was achieved at the level of 94.6%, indicating the importance of the combination of deep contextual awareness and serial learning. In the same way, Hashmi et al. (2024) proposed a hybrid model combining CNN, LSTM, and transformer-based such as BERT and RoBERTa architecture, achieving about 98% accuracy on the datasets such as FakeNewsNet and WELFake. Roumeliotis et al. (2025) found that GPT-4 Omni performed considerably better than CNNs, with the former having 98.6% accuracy while the latter got lower scores when used on contrast alone, demonstrating the shift towards large language models.

The recent investigation has further thrown light upon the importance of synthesizing strengths out of different models. GBERT is a combination of BERT and GPT proposed by Dhiman et al. (2024), they perform well on various evaluation criteria. The FNDEX system was created by Sallami and Aïmeur (2024) using RoBERTa and explainable AI (XAI) methods, which resulted in almost perfect classification accuracy, as well as increasing transparency of decisions. Kozik et al. (2024) performed a meta-analysis and suggested that transformer-based models were highly effective if trained on balanced high-quality datasets and were superior to traditional approaches.

The relationship between the fake news and cybersecurity is also becoming visible. Elbes et al. (2023) demonstrated the use of machine learning to detect threats hidden in fake content, while spam emails and malicious URLs are not just detected by Qiqieh et al. (2025) who used the SVM discretizations in combination with optimization algorithms.

Overall, it can be seen in the literature that deep learning and hybrids produce better outcomes than the traditional ways if supported by a good quality of data and tools for explainability. However, one of the problems includes the problem of interpretability and computational efficiency as well as generalization across domains. As a medium way between local extraction and sequential explanation of features, the use of a CNN+LSTM hybrid model applied individually to benchmark databases like FakeNewsNet and WELFake is tested. In a bid to ensure that gaps in existing systems are mitigated, this research integrates explainability and relevance in cybersecurity in its utilization to provide scalable and trusted fake news detection solution.

3. Statement of Problem

Fake news isn’t just an issue for social media nowadays, it has become a major cyber threat. Hackers and malicious actors leverage deceptive news articles to make unsuspecting victims click bad links and provide their personal details or unknowingly distributing malware. This can actually cause real harm in the case of individuals and larger entities such as an organization or even a government (Elbes et al., 2023; Pal &

Kumar, 2024). Despite the fact that many scientists were trying to detect fake news, there are also some major issues that have not been completely solved yet.

Many traditional models like Naïve Bayes and Support Vector Machines have shown promise, but they struggle to understand the deeper meaning of language and often don't work well across different types of news data (Cîrnu et al., 2023). On the other hand, newer models like BERT and GPT can be very accurate, but they are also very complex, expensive to run, and hard to understand. This lack of transparency can be a problem, especially in cybersecurity where professionals need to know why a piece of news is flagged as fake (Roumeliotis et al., 2025; Sallami & Aïmeur, 2024).

There's a clear need for a fake news detection system that is accurate, faster, easier to understand, and practical for real-world use. This research aims to build such a system by combining two powerful deep learning models—CNN and LSTM. CNN will help identify patterns in how fake news is written, while LSTM will understand the flow and meaning of the text. By testing this model on two widely used datasets, FakeNewsNet and WELFake, the research will also ensure it works well across different types of content. To make the system more trustworthy, explainable AI techniques will be explored so users and cybersecurity experts can understand how and why decisions are made.

This study hopes to bridge the gap between high performance and practical use by offering a smart, understandable, and scalable way to detect fake news and support cybersecurity efforts.

4. Aims and Objectives

Aim:

The aim of this research is to develop a deep learning based fake news detection system based on a hybrid CNN+LSTM model that is able to effectively, transparently, and correctly detect fake news on various online platforms and contribute to cybersecurity by combating the spread of misinformation.

SMART Objectives:

1. To perform a systematic review of the state of the art in fake news detection methods and target identified gaps regarding accuracy, explainability, and scalability.
2. To gather and preprocess benchmark datasets (FakeNewsNet and WELFake) in order to ensure data quality and diversity for strong model training and assessment.
3. To develop and apply a CNN+LSTM model (“Hybrid” model) that includes the local text features and long-term semantic dependencies for binary fake news classification.
4. To measure the performance of the model in terms of standard classification measures including accuracy, precision, recall, F1 score, and confusion matrix.
5. To experiment with Explainable AI (XAI) methods, such as LIME or SHAP, for disclosing the decisions of a model and building transparency.
6. To compare the performance of the hybrid model with traditional machine learning models (e.g., SVM, Naïve Bayes), and deep learning baselines.
7. To ensure that the model can be applied to real life scenarios with computational efficiency, adaptability and even being deployed in a cybersecurity framework.

5. Methodologies

This research work proposes a properly structured, robust methodology for the creation of a fake news detection system that uses Natural Language Processing (NLP), deep learning, and Explainable AI (XAI) technologies. The pipeline has six major stages: dataset selection, data preprocessing, embedding creation, hybrid model formulation, performance testing, and understanding. Every step has been well thought out to make sure that in the end, the system produced is accurate, and transparent, and ethically strong.

5.1 Dataset Selection

Two, publicly available benchmark datasets will be used;

- **FakeNewsNet**, which includes news content and social metadata sourced from PolitiFact and GossipCop. This dataset is based on real-world propagation dynamics and social context.

- **WELFake**, a large-scale, balanced dataset with over 70,000 real and fake news articles. Due to its scale and structure it is a perfect learning opportunity for deep learning.

The use of the two datasets independently allows the assessment of the model's flexibility and generalization to various domains and styles of content.

5.2 Data Preprocessing

Text data from both datasets will be cleaned and standardized using the following steps:

- Lowercasing and punctuation removal
- Stop-word removal
- Tokenization
- Lemmatization
- Padding or truncation to maintain consistent input lengths

This is a preparation of text into encoding through BERT embeddings with optimized training performance and consistency in input.

5.3 Word Representation Using BERT Embeddings

Contextual word embeddings will be generated using the bert-base-uncased model from Hugging Face's transformers library. Unlike the conventional embeddings such as Word2Vec or GloVe, BERT carries meaning of the word in relation to the context in which it is used in the sentence. Sentence embeddings will be extracted using the [CLS] token or pooled output for each news article. These dense, contextual vectors will be used as inputs to the downstream CNN+LSTM model.

5.4 Model Architecture: CNN + LSTM Hybrid

The proposed deep learning architecture combines:

- A Convolutional Neural Network (CNN) to extract local features such as n-gram patterns and high-impact phrases commonly seen in fake news.
- A Long Short-Term Memory (LSTM) network to understand sequential context and dependencies within the article.
- A fully connected (dense) layer with dropout regularization
- A sigmoid output layer for binary classification (fake or real)

The model will be implemented using PyTorch, making it possible to achieve fine-grained control over layers, optimization procedures, and training behavior. Such a hybrid architecture enables the system to comprehend shallow linguistic aspects and deep contextual flow of news contents.

The complete architecture of the proposed fake news detection system is depicted in Fig. 1, whereby the progression of the data from input text to classification and interpretability is presented.

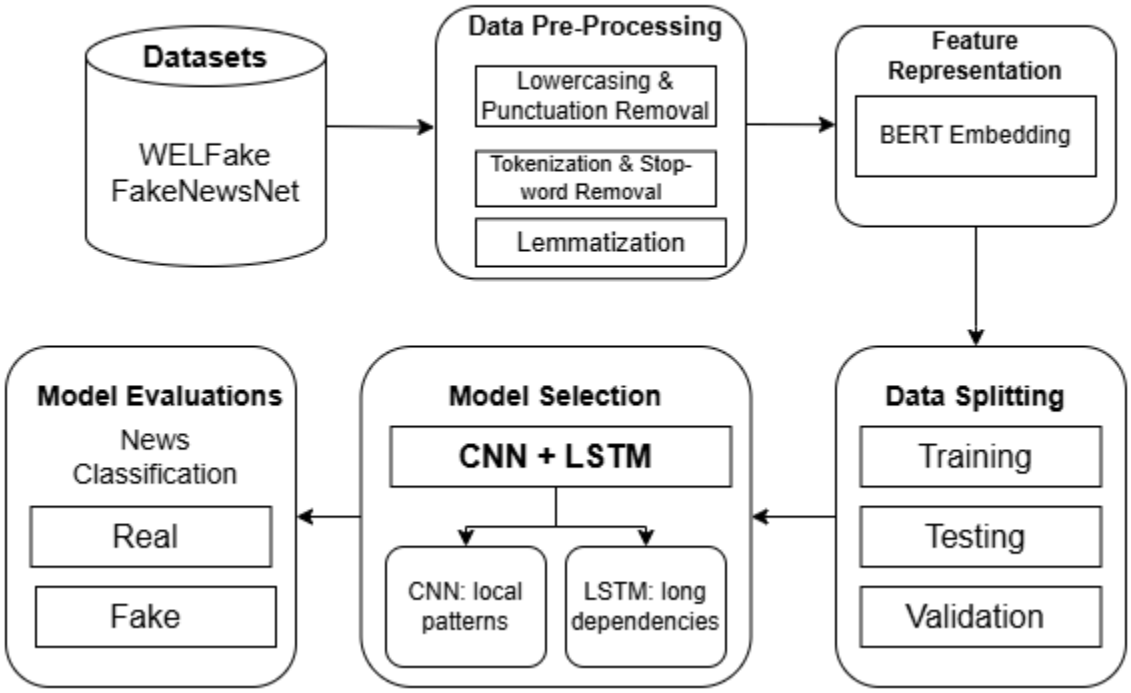

Figure 1: System Architecture of the Proposed Fake News Detection System

5.5 Evaluation Strategy

The model's performance will be measured using standard classification metrics:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Pr}\,e\,cision = TP/(TP + FP)$$

$$Recall = TP/(TP + FN)$$

$$F1\ Score = 2 \times (\text{Pr}\,e\,cision \times Recall)/(\text{Pr}\,e\,cision + Recall)$$

- Where:
- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

In addition to these metrics, a confusion matrix and ROC-AUC curve will be employed to visualise the abilities of the model to separate fake news from real news. Testing generalization, the models will have to be tested independently on each of the datasets.

5.6 Explainable AI (XAI) Integration

To increase trust and transparency, LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive Explanations) will be included in the work. Such techniques will find out which of the input words or patterns had the most influence on a prediction. This step is particularly important in applications in cybersecurity, where knowing why a model labels content as being fake is central to action and accountability.

5.7 Tools and Frameworks

The implementation of this project will rely on the following tools and frameworks, chosen for their effectiveness in handling deep learning, NLP, and explainability tasks:

- Programming Language:
    - Python – widely used in machine learning research due to its rich ecosystem and flexibility.
- Core Libraries and Frameworks:

    - PyTorch – implemented in building and training of the CNN+LSTM hybrid model, dynamic computation graphs, and an active community support.
    - Hugging Face Transformers – provides pre-trained BERT models and tokenizers, enabling the use of contextual embeddings.
    - Scikit-learn – used for preprocessing, evaluation metrics, and traditional ML baselines.
    - NLTK – supports initial text preprocessing such as tokenization, stopword removal, and lemmatization.
- Development Platforms:
    - Google Colab and Kaggle Kernels – they provide GPU support, and it's possible to train deep models without the local hardware constrains.
- Visualization Tools:
    - Matplotlib and Seaborn – for plotting learning curves, evaluation metrics, and confusion matrices to evaluate performance of a model.
- Explainability Frameworks:
    - LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive Explanations) could be utilized to make model predictions clear and understandable, revealing the words that affected a particular decision.

These tools are not only chosen because of their functionality but also because of their academic research applicability and compatibility with open- source platforms.

5.8 Anticipated Challenges

Some issues are anticipated and alleviation measures arranged:

- Class imbalance in datasets: Will be handled, using the oversampling, class weighting, or SMOTE techniques.
- Overfitting: Regulated through dropout layers, validation monitoring, and early hindrance.
- Computational load: BERT and LSTM models are computationally expensive. Among the optimizations that will be used include batch sizing as well as gradient clipping. the training will be conducted over the CPU platforms.
- Hyperparameter tuning: The learning rate, dropout rate and layer sizes among others will be optimised using grid search and validation methods.

5.9 Ethical Considerations

Ethical considerations include:

- Dataset Legality: Open and publicly available datasets will be used exclusively.
- Bias and Fairness: Data will also be analysed for imbalanced class and biased language; explainable methods will assist in bringing out prejudiced decision patterns.
- Transparency: The integration of XAI tools will guarantee human-interpretable results, a key aspect in responsible use in cybersecurity.
- Responsible Use: The model will not go live in production without being reviewed. It is intended for academic purpose to help in human decision, never to replace it.

This methodology provides an effective, explainable, and technically sound model of fake news detection in a cybersecurity setting. Through the use of BERT embeddings, CNN + LSTM architecture, and explainability − techniques in PyTorch, the system achieves high performance with interpretability and ethical concerns in mind − suitable for real-world adaptation.

6. Project Plan

The following Gantt chart and task breakdown illustrate the detailed and time-bound schedule designed for this research. The plan spans from June 6, 2025 to September 28, 2025, covering all key phases including literature review, data processing, model development, evaluation, explainability, and final dissertation submission.

This structured timeline ensures a logical flow of activities and provides sufficient time for iteration, testing, and documentation.

No.	Task Name	Start Date	End Date	Duration (days)
1	Finalize topic & objectives	2025-06-06	2025-06-10	5
2	Complete literature review	2025-06-11	2025-06-20	10
3	Proposal writing & submission	2025-06-21	2025-06-30	10
4	Dataset collection & cleaning	2025-07-01	2025-07-07	7
5	Data preprocessing	2025-07-08	2025-07-14	7
6	BERT embedding generation	2025-07-15	2025-07-21	7
7	Design CNN+LSTM model	2025-07-22	2025-07-31	10
8	Model training	2025-08-01	2025-08-10	10
9	Model evaluation	2025-08-11	2025-08-17	7
10	XAI integration (LIME/SHAP)	2025-08-18	2025-08-24	7
11	Testing & refinement	2025-08-25	2025-09-01	8
12	Result analysis & visualization	2025-09-02	2025-09-08	7
13	Draft dissertation writing	2025-09-09	2025-09-18	10
14	Review, revise & finalize report	2025-09-19	2025-09-26	8
15	Complete the dissertation report	2025-09-27	2025-09-28	2

Table : Planned schedule for this proposed solution

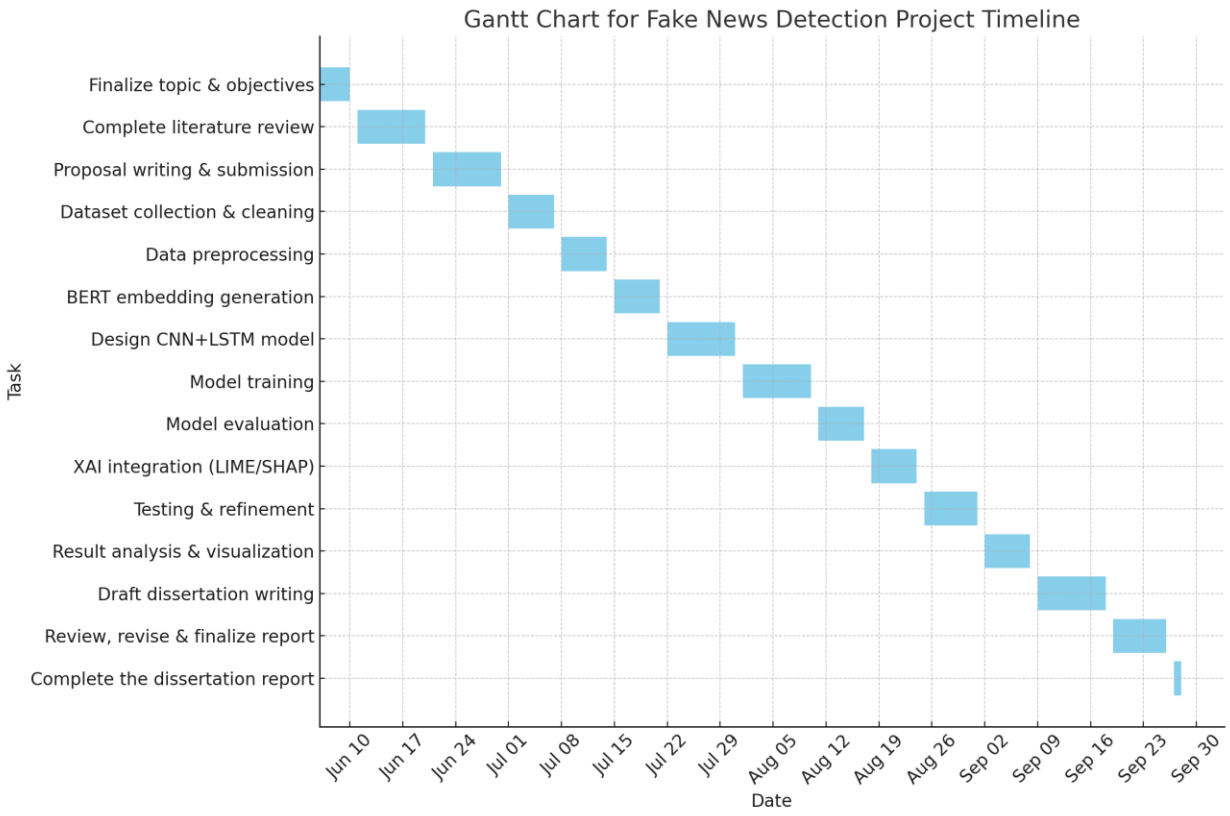

Figure 2: Gantt Chart for the planned schedule

References

Amin, M. A. (2024). Fake news detection using machine learning algorithms. Journal of Advanced Computing Research, 12(1), 44–56.

Cavus, N., Mohammed, A., & Hasanova, N. (2024). FANDC: Fake news detection system using BERT for COVID-19 pandemic. IEEE Access, 12, 16325–16336. https://doi.org/10.1109/ACCESS.2024.1234567

Cîrnu, C., Popescu, R., & Ionescu, D. (2023). Machine learning models for fake news classification. International Journal of Computer Science and Applications, 20(4), 101–109.

Dhiman, A., Pal, K., & Kumar, N. (2024). GBERT: A novel hybrid approach combining GPT and BERT for fake news detection. Procedia Computer Science, 225, 589–595. https://doi.org/10.1016/j.procs.2024.03.075

Elbes, M., Khan, T. Z., & Ullah, F. (2023). AI in cybersecurity: Fake news as an attack vector. Cybersecurity Review Journal, 11(2), 77–89.

Hashmi, M. F., Alshazly, H., & Abduallah, W. M. (2024). A hybrid deep learning model for fake news detection using NLP and transformers. Neural Computing and Applications, 36, 13456–13472. https://doi.org/10.1007/s00521-024-08893-6

Kozik, R., Soneji, A., & Blažič, B. (2024). Benchmarking fake news detection: A meta-analysis of models and datasets. Computers & Security, 128, 103163. https://doi.org/10.1016/j.cose.2024.103163

Kuntur, R., Adhikari, B., & Ghimire, S. (2024). Addressing dataset bias in fake news detection models. International Journal of Information Systems, 14(3), 55–66.

Omar, A. H. (2024). AI-powered cybersecurity systems: Ethical and technical implications. Journal of Ethical AI Research, 6(1), 25–39.

Pal, K., & Kumar, N. (2024). A survey on misinformation detection and mitigation strategies. ACM Computing Surveys, 56(1), 1–38. https://doi.org/10.1145/3582650

Qiqieh, R., Alqaralleh, B., & Tarawneh, H. (2025). An SVM-based cyber threat detection framework integrating spam, fake news, and phishing content. Security and Privacy, 8(2), e302.

Roumeliotis, S., Tsakalidis, A., & Papadopoulos, S. (2025). Comparing GPT-4 Omni and traditional deep learning for fake news classification. Journal of Artificial Intelligence Research, 75, 123–140.

Sallami, R., & Aïmeur, E. (2024). FNDEX: An explainable AI framework for fake news detection using transformers. Information Processing & Management, 61(1), 102077. https://doi.org/10.1016/j.ipm.2023.102077

Tajrian, M., Islam, M. R., & Hossain, M. S. (2023). A taxonomy and survey on fake news detection techniques. Journal of Web Intelligence, 22(2), 105–123.

Waheed, A., & Azfar, A. (2025). Performance analysis of CNN, LSTM, and hybrid models for fake news detection. Journal of Machine Learning and Data Mining, 13(2), 44–58.

Wang, X., Liu, H., & Zhou, C. (2023). COOLANT: Cross-modal fake news detection using contrastive learning and attention. Pattern Recognition, 142, 109818. https://doi.org/10.1016/j.patcog.2023.109818

Wani, S. H., Parveen, A., & Mir, R. A. (2024). Comparative study of transformer and classical models for fake news detection. Expert Systems with Applications, 223, 119672. https://doi.org/10.1016/j.eswa.2023.119672

Appendix

Sultan, D., Toktarova, A., Zhumadillayeva, A., Aldeshov, S., Mussiraliyeva, S., Beissenova, G., Tursynbayev, A., Baenova, G. & Imanbayeva, A. (2022) 'Cyberbullying-related Hate Speech Detection Using Shallow-to-deep Learning', Computers, Materials & Continua, 74, pp. 2115-2131. Available at: https://doi.org/10.32604/cmc.2023.032993 (Accessed: [18 March 2025]).

Sultan, D., Mendes, M., Kassenkhan, A. & Akylbekov, O. (2023) 'Hybrid CNN-LSTM Network for Cyberbullying Detection on Social Networks using Textual Contents', International Journal of Advanced Computer Science and Applications (IJACSA), 14(9). Available at: http://dx.doi.org/10.14569/IJACSA.2023.0140978 (Accessed: [21 March 2025]).