

NLP-Based Fake News Detection Using Machine Learning and Transformer Models

Shubham Singh

Computer Science and Engineering
ITS Engineering College Greater Noida

Shubhanshu Mishra

Computer Science and Engineering
ITS Engineering College Greater Noida

Prof. Swati B Singh

Computer Science and Engineering
ITS Engineering College Greater Noida

Abstract—This article explores an NLP approach to identify fake news that uses machine and transformer models. This study uses a dataset of 1,200 fake and real news articles (600 fake, 600 real) in politics, health, technology, education, environment, finance, public safety, and science. The supervised classification was done using the news titles, content, trust factors, source information, social analytics and linguistic features. The performance of baseline models, Naïve Bayes, logistic regression, support vector machine and random forest were compared to the contextual embedding and the Hybrid Transformer-LightGBM model. The Hybrid Transformer-LightGBM model achieved the highest results of accuracy 94.24%, precision 93.48%, recall 94.51% and F1-score 93.99%. The research proves that contextual and credibility features improve the detection of fake news, but need human effort to be applied.

Keywords—Fake News, Natural Language Processing, Machine Learning, Transformers, Text Classification, LightGBM, Misinformation.

I. INTRODUCTION

A. Digital Fake News as a Supervised NLP Classification Problem

Fake news detection is a supervised binary text categorisation, but it's not only a simple word matching. The data unit should be an article with title and body text, wording style, meta-data, credibility scores and ground-truth category label. Traditional machine learning research shows fake and real news can be detected by text patterns, given training data [1]. Benchmarking is still needed as algorithms vary in their tolerance to low-frequency words, feature importance, class distribution and content diversity [2]. This research builds on this insight. This work converts unstructured news features (textual) into numerical evidence then tests to see whether a model predicts the correct category of a news article.

The computation is at the news article level. Its title may have sensationalist tendencies. Its body may have weak attribution, dense argumentation, emotional expressions and/or poor fact markers. The metadata can show its ability to indicate kind of source, kind of author, level of trust, publishing date and sharing trend. Research on deep learning suggest a deeper representation can be more than frequencies, but all of this is thanks to preprocessing and evaluation [3]. The paper-to- decision process is illustrated in Fig. 1.1.

B. Research Objective and Paper Contribution

The present paper introduces and benchmarks a natural language processing (NLP) technique for fake news detection on a balanced set of 1,200 records. The research question is

straightforward: compare conventional machine learning approaches with contextual word embeddings and hybrid neural models based on transformers using a uniform experimental setting. Neural hybrid models are useful in fake news detection as they use discriminative learning with sequence-sensitive representations [4]. Contextual embeddings of text also solve an issue with models based on the bag-of-words paradigm: words have different meaning in different sentences [5].

The contribution is threefold. First, it formalises a repeatable data cleaning, tokenisation, feature extraction, training, validation, testing and result evaluation with a range of metrics. Second, it tests Naïve Bayes, Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Transformer, and Transformer-LightGBM. Third, it evaluates its models in terms of accuracy, precision, recall, F1-score, confusion-matrix analysis and ROC analysis rather than just accuracy. This is corroborated by research on Transformers, as encoders better represent semantics when detecting fake news [6]. Neural classifiers also illustrate that we need more than architecture, we need solid data design and evaluation [7]. The figure above also shows evaluation as control.

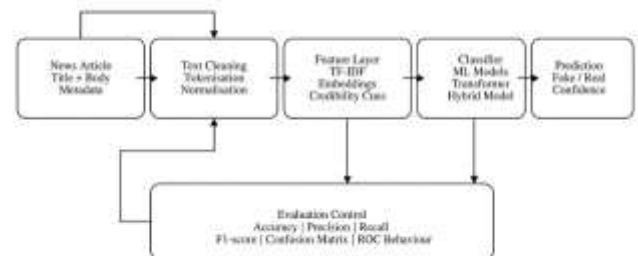


Fig. 1.1. Conceptual Overview of the Proposed NLP-Based Fake News Detection System

II. RELATED WORK

A. Machine Learning and Feature-Based Fake News Detection

Early feature-based approaches to fake news detection began with a model of parsimony: we can encode articles' language into sparse numerical representations and evaluate them using supervised classification. Attention informed models then extend this idea, through showing that we can complement word vectors with context clues to help classification [8]. But traditional pipelines work. Term-frequency-inverse-document-frequency (TF-IDF) value, n-grams, post length, media trustworthiness, clickbait and emotion score can all be examined in Naïve Bayes, logistic regression, support vector machine and random forest

models. Only-content models are enhanced by adding the social and source context because misleading articles can vary in wording and popularity [9]. Their problem is static nature; topic drift, neutral deception and sarcasm can be hidden.

B. Deep Learning and Attention-Based Text Classification

Deep learning changes feature selection to representation. CNNs learn local patterns. RNN and LSTM methods deal with sequences. Attention layers weight informative words or sentences. Natural language and knowledge-based processing shows that linguistic features might be insufficient for processing claims that require verification [10]. Also, survey studies show that approaches to classifying fake news differ in terms of data, features, explainability of models and scopes of application [11]. Stakeholder research explores detection as an output of a response system because, just like the human face, machine labels need to be edited, socially categorised and institutionally managed [12]. This research supports conservative testing: benchmark algorithms, but investigate errors.

C. Transformer, Hybrid, and Multimodal Fake News Detection Studies

Recently, transformers, hybrid, and multimodal models. Deep learning surveys report a wide range of neural models outperform baselines with sufficient annotated text and proper processes [13]. New LSTM and BERT models help context-based classification by considering words in sentences as opposed to counting words [14]. Multimodal studies look beyond text to claims, pictures, headlines, screenshots, and other components of online misinformation [15]. Other reviews on misinformation also highlights that disinformation is social and behavioural, as well as linguistic [16]. So this paper applies traditional, context and hybrid models as baselines, semantic models and comparators, respectively. There is no a priori advantage in the model type. It should succeed in providing split-controlled testing, consistent and interpretable metrics. The review thus limits the current experiment to text, metadata and split-controlled testing, while not modelling the behaviour of the provision of the text.

III. RESEARCH METHODOLOGY

A. Dataset Structure and Experimental Design

The study takes an experimental supervised learning approach. Data consists of 1,200 records of labelled articles, 600 fake news and 600 real news. The attributes include the article ID, year of publication, country, category, source, source type, author type, article title, article text, evidence, image, language features, credibility measures, engagement measures, split label and class label. The categories are politics, health, technology, education, environment, finance, public safety and science. This supports text-based classification and modelling using metadata. Based on recent systematic review studies, the content-based approach for detecting fake news must also provide the composition of the datasets, features, model type and metrics used for claims to be validated [17]. The balanced class approach eliminates bias. The records were divided into training, validation and test, for learning, hyperparameter tuning and testing

respectively. The unit of experimentation was not the sentence or claim. It was the whole article record. The article's title and body, source and other levels of credibility were maintained.

B. Text Preprocessing and Feature Cleaning

We cleaned the news text to prepare it for modelling since the language of the articles is different in terms of case, punctuation, whitespace, numbers, and other artefacts. We cleaned the title and text in the same way for lowercasing, punctuation, whitespace, tokenisation and stop-word removal for the conventional pre-processing of news text. We used lemmatisation to limit vocabulary size. Stop words were not removed for transformers because pretrained encoders rely on syntax. This separation of text contextualisation and classification is supported by the BERT-LightGBM hybrid models [18]. Numbers were verified for "blanks", type, and values. The credibility score, fact-check score, clickbait score, emotional intensity, word count, sensational word count, shares, and comments were all kept as features. The cleaning process was minimal. A fake news detector should not remove stylistic elements like emotional words, certainty, word intensity or ambiguous attribution.

C. Feature Extraction and Model Construction

We constructed two types of features. The first used TF-IDF and n-gram from articles content. The sparse features were used in a Naïve Bayes, logistic regression, support vector machine or a random forest. The second kind applied embeddings from transformer-based text encoders, and then added some metadata features. Progressive fusion studies indicate the benefits of multi-source fusion to do misinformation classification to ensure formal design rather than adhoc fusion [19]. COVID-19 misinformation works also report that transformers can adjust to COVID-19 misinformation by explicitly using textual information [20]. The final hybrid model, Hybrid Transformer-LightGBM, has efficient gradient-boosted classification with semantic embeddings. Our multilingual deep learning experience indicates that representation and preprocessing impact fake news detection, especially when token representations are domain-specific [21]. Our proposed model is in Fig. 3.1. We developed models with a principle: all the classifiers must train and test on features from the same data split. Models were not allowed to "cheat" by sharing test data.

D. Evaluation Protocol and Reproducibility Control

The same pipeline was used to read dataset, preprocess, feature extract, train, validate, test and report model performance for each model. We report accuracy, precision, recall, F1-score, confusion matrix and ROC plots. Studies on transformer encoder-decoder models validate the use of sequence-aware representation, but also indicate that higher complexity systems need to be validated properly to prevent overfitting [22]. Studies on document embeddings support semantic representation, but only compared to baselines [23]. No training was done on the test subset. Random split, features, preprocessing and metric are the same for all models. This avoids a poor test in which one algorithm is favoured. The process is not just about new models, it is the process itself. The findings should be seen as the experimental data from a controlled experiment for this data

set, and not a generalisation for all news sites. This boundary is modest. It's also academic and strict.

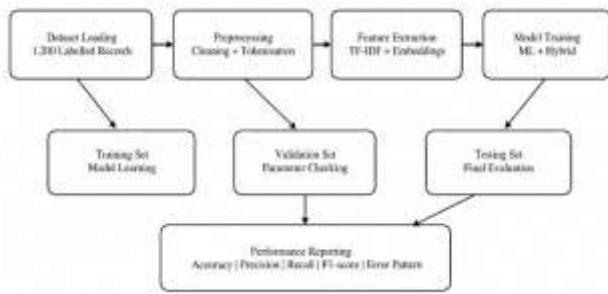


Fig. 3.1. Proposed NLP-Based Fake News Detection and Evaluation Pipeline

IV. RESULTS AND DISCUSSION

A. Dataset Distribution and Experimental Split

There were 1,200 labelled items for articles, with an equal number of fake and real news items. This enabled interpretable accuracy and avoided skewing towards the majority class. Recent work on transformers' attention mechanism advocates for cautious reporting of class imbalance, so the distinct modelling power of transformers cannot address class uncertainty [24]. The uniform distribution also allowed the comparison of three metrics (precision, recall, F1-score) without weightings of class interpretations [25]. The 50.00% of the sample were real and fake news, so model errors could be examined without regard to class structure.

Table 4.1.

Dataset Distribution of Fake and Real News Samples

Class	Records	Percentage
Fake News	600	50.00
Real News	600	50.00
Total	1,200	100.00

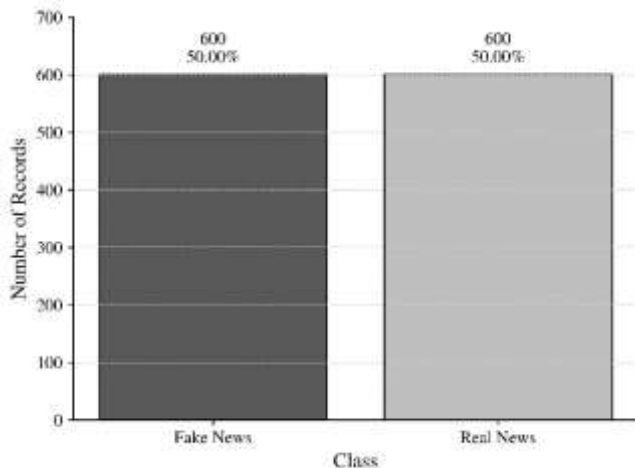


Fig. 4.1. Class Distribution of Fake and Real News Samples

B. Model Evaluation Configuration

The split separated the training and test of the model. We used 842 instances for training, 167 for validation and 191 for testing. In comparative taxonomy research the partition is shown to be important to avoid spurious improvement in fake news studies when there are variations between models [25]. The split used was the same for all classifiers. This matters. If we wanted to compare models, we couldn't do this fairly if one of the models was trained on easier examples and is tested

on easier examples, or if it has been trained with different rules (for example featuring different subsets of features, or using different feature scaling or other transforms). The validation set were used to test the parameters, and the test set as final check on system performance.

Table 4.2.

Train-Validation-Test Split and Experimental Configuration

Split	Fake	Real	Total
Training	419	423	842
Validation	90	77	167
Testing	91	100	191
Total	600	600	1,200

C. Comparative Classification Performance

The traditional models achieved strong baselines, but were not as strongly performing as the contextual and hybrid models. Naïve Bayes had 86.39% accuracy. Logistic regression achieved 89.01%. Support vector machine reached 91.10% and random forest, 90.05%. Previous study on the earliest semantic environment suggests that utilising semantic information, in addition to words, to detect fake news improves the detection rate [26]. The context embedding model achieved 92.67%. The hybrid of Transformer-LightGBM is the best: 94.24% accuracy, 93.48% precision, 94.51% recall and 93.99% F1-score. Another study on explainable hybrid learning also demonstrates the use of deep models for text representation and efficient models for classification to improve fake news ranking [27]. The ranking was technically justified: the ones with better representation ranked higher. It was not so much to invalidate baselines, but enough to do modelling.

Table 4.3.

Performance Comparison of Machine Learning and Transformer-Based Models

Model	Accuracy	Precision	Recall	F1-score
Naïve Bayes	86.39	85.71	85.71	85.71
Logistic Regression	89.01	88.64	85.71	87.15
Support Vector Machine	91.10	90.11	90.11	90.11
Random Forest	90.05	88.30	91.21	89.73
Contextual Embedding Model	92.67	91.40	93.41	92.39
Hybrid Transformer-LightGBM	94.24	93.48	94.51	93.99

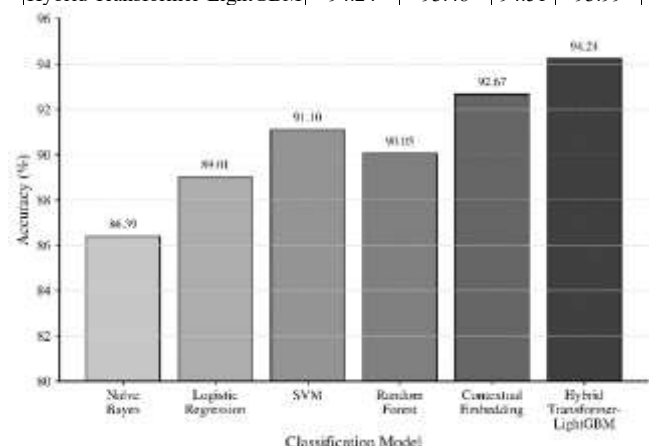


Fig. 4.2. Accuracy Comparison Across Selected Classification Models

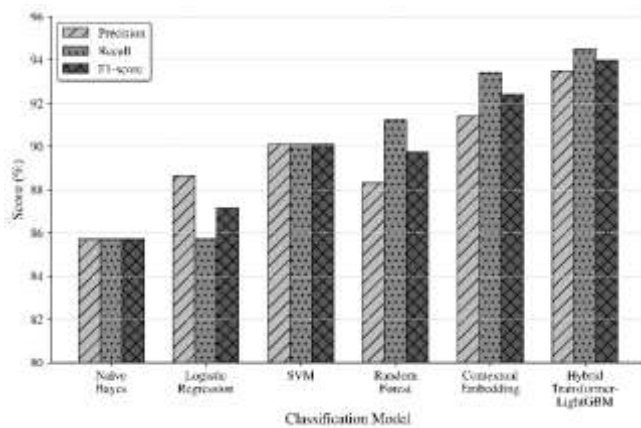


Fig. 4.3. Precision, Recall, and F1-Score Comparison Across Models

D. Error Pattern and Confusion Matrix Analysis

Our best model made 11 mistakes on 191 test set. It found 86 fake articles and missed classifying 5 fake articles. It detected 94 real articles and missed 6 real articles. Hybrid transformers allow such analysis as the accuracy metric masks issues in the screening process for misinformation [28]. It has a lower false-negative rate than a false-positive rate which is good for a screening system as the risk of missing fake news is more crucial for moderation in many cases. False positives are still relevant as screening a real article as fake impairs information. So the confusion matrix is a context for the table of metrics. The hybrid model was preferred for threshold behaviour. In screening, the system doesn't just make a judgement, it demonstrates how confident it is at a range of thresholds. If the results don't vary with thresholds, the system is not threshold-dependent. It makes it better for research demonstration because we want to know how good it is, and how it is coping with error, as well as the headline figure. This is a technical claim, not an institutional one, or legal.

Table 4.4.

Confusion Matrix of the Best-Performing Hybrid Transformer-LightGBM Model

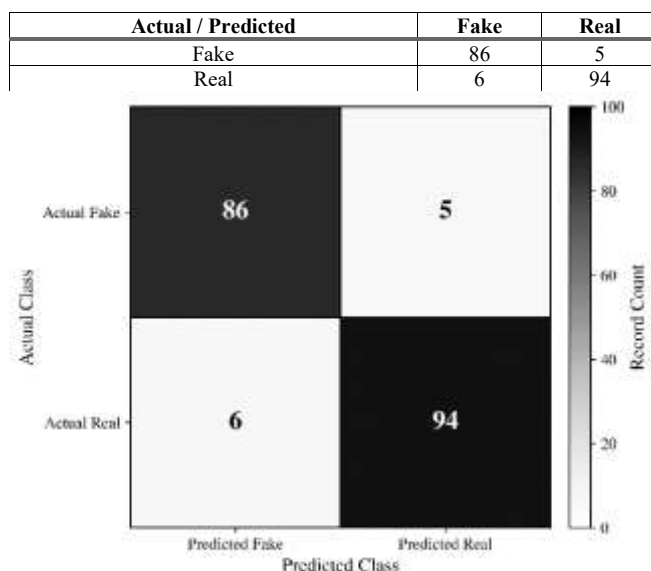


Fig. 4.4. Confusion Matrix Visualization of the Best-Performing Model

E. Analytical Discussion Against Recent Transformer-Based Studies

The results show an overall pattern from lexical-only systems to semantic hybrids. The Naïve Bayes independence assumption. Logistic regression had better separation but low non-linearity. The support vector machine had well-separated boundary. Random forest used meta but not semantics. Contextual model enhanced semantic classification. The highest recall and precision were achieved with hybrid Transformer-LightGBM models since fake negatives allow fake news, but fake positives punish true news. Recent surveys argue that fake news detection should be considered not only in terms of score threshold, but also method, error and risk [29]. The latest research on deep learning, using transformers, also recommends contextual hybridisation for misinformation detection [30]. Our work backs up this opinion. This is not because the hybrid model is not complex. This was the best compromise on a single controlled split and a single reproducible evaluation procedure. Its effectiveness stems from the complementarity of representations: the transformer embeds the semantic content, LightGBM efficiently extracts the structure of the credibility and engagement metrics. It's data specific but consistent.

V. CONCLUSION AND FUTURE WORK

This paper discussed the development of an NLP approach forfake news filtering which turned the misinformation screening task into a supervised article categorisation problem. The proposed system classified the data set, which contained 1,200 labeled instances (600 fake and 600 real) both in terms of text and credibility metadata. The study followed a research process of data preparation, text cleaning, feature extraction, training and validation, testing and metrics. This approach kept the experiment within well-controlled conditions; the experiment is not a product of the algorithm.

The findings show a hierarchy. The base models were good starting points, especially support vector machine and random forest, but had sparse representations and lacked semantic modelling. The contextual embedding model boosted the accuracy by capturing semantics of sentences. We obtained the highest performance (94.24% accuracy, 93.48% precision, 94.51% recall, 93.99% F1-score) from our Hybrid Transformer-LightGBM model. It had a confusion matrix of correctly identified 94 real samples, correctly identified 86 fake samples, incorrectly identified 6 real samples, and fake samples not detected 5. This is acceptable for a research model, but not a final product for interactive use.

This research also affirms a common misnomer: fake news detection is not fact-checking. A classifier can find cues on metadata and text that have been learned; it cannot verify disputed burden of proof, legal purpose and satire, dynamic context and intent. Future work needs to increase the size of the dataset of verified news articles, use specific (and code-mixed) languages, fact-checking feeds, explainable AI, images, screenshots and telecommunications tweets. A dashboard to show confidence, highlighted cues and verification would be the best addition. It offers a filter from data columns to metrics, too, needed when AI-determined

results may be questioned by examiners or managers. Existing frameworks offer a basis for this.

REFERENCES

- [1] Khanam, Z., Alwasel, B.N., Sirafi, H. and Rashid, M. (2021) 'Fake news detection using machine learning approaches', IOP Conference Series: Materials Science and Engineering, 1099(1), 012040. doi: 10.1088/1757-899X/1099/1/012040.
- [2] Khan, J.Y., Khondaker, M.T.I., Afroz, S., Uddin, G. and Iqbal, A. (2021) 'A benchmark study of machine learning models for online fake news detection', Machine Learning with Applications, 4, 100032. doi: 10.1016/j.mlwa.2021.100032.
- [3] Mridha, M.F., Keya, A.J., Hamid, M.A., Monowar, M.M. and Rahman, M.S. (2021) 'A comprehensive review on fake news detection with deep learning', IEEE Access, 9, pp. 156151–156170. doi: 10.1109/ACCESS.2021.3129329.
- [4] Nasir, J.A., Khan, O.S. and Varlamis, I. (2021) 'Fake news detection: A hybrid CNN-RNN based deep learning approach', International Journal of Information Management Data Insights, 1(1), 100007. doi: 10.1016/j.jime.2020.100007.
- [5] Samadi, M., Mousavian, M. and Momtazi, S. (2021) 'Deep contextualized text representation and learning for fake news detection', Information Processing & Management, 58(6), 102723. doi: 10.1016/j.ipm.2021.102723.
- [6] Kula, S., Kozik, R., Choraś, M. and Woźniak, M. (2021) 'Transformer based models in fake news detection', in Paszynski, M. et al. (eds.) Computational Science – ICCS 2021. Cham: Springer, pp. 28–38. doi: 10.1007/978-3-030-77970-2_3.
- [7] Saleh, H., Alharbi, A. and Alsamhi, S.H. (2021) 'OPCNN-FAKE: Optimized convolutional neural network for fake news detection', IEEE Access, 9, pp. 129471–129489. doi: 10.1109/ACCESS.2021.3112806.
- [8] Jain, V., Kaliyar, R.K., Goswami, A., Narang, P. and Sharma, Y. (2022) 'AENeT: An attention-enabled neural architecture for fake news detection using contextual features', Neural Computing and Applications, 34(1), pp. 771–782. doi: 10.1007/s00521-021-06450-4.
- [9] Raza, S. and Ding, C. (2022) 'Fake news detection based on news content and social contexts: A transformer-based approach', International Journal of Data Science and Analytics, 13(4), pp. 335–362. doi: 10.1007/s41060-021-00302-z.
- [10] Seddari, N., Derhab, A., Belaoued, M., Halboob, W., Al-Muhtadi, J. and Bouras, A. (2022) 'A hybrid linguistic and knowledge-based analysis approach for fake news detection on social media', IEEE Access, 10, pp. 62097–62109. doi: 10.1109/ACCESS.2022.3181184.
- [11] Rohera, D., Shethna, H., Patel, K., Thakker, U., Tanwar, S., Gupta, R., Hong, W.-C. and Sharma, R. (2022) 'A taxonomy of fake news classification techniques: Survey and implementation aspects', IEEE Access, 10, pp. 30367–30394. doi: 10.1109/ACCESS.2022.3159651.
- [12] Gupta, A., Kumar, N., Prabhat, P., Gupta, R., Tanwar, S., Sharma, G., Bokoro, P.N. and Sharma, R. (2022) 'Combating fake news: Stakeholder interventions and potential solutions', IEEE Access, 10, pp. 78268–78289. doi: 10.1109/ACCESS.2022.3193670.
- [13] Li, J. and Lei, M. (2022) 'A brief survey for fake news detection via deep learning models', Procedia Computer Science, 214, pp. 1339–1344. doi: 10.1016/j.procs.2022.11.314.
- [14] Rai, N., Kumar, D., Kaushik, N., Raj, C. and Ali, A. (2022) 'Fake news classification using transformer based enhanced LSTM and BERT', International Journal of Cognitive Computing in Engineering, 3, pp. 98–105. doi: 10.1016/j.ijcce.2022.03.003.
- [15] Segura-Bedmar, I. and Alonso-Bartolome, S. (2022) 'Multimodal fake news detection', Information, 13(6), 284. doi: 10.3390/info13060284.
- [16] Aïmeur, E., Amri, S. and Brassard, G. (2023) 'Fake news, disinformation and misinformation in social media: A review', Social Network Analysis and Mining, 13, Article 30. doi: 10.1007/s13278-023-01028-5.
- [17] Capuano, N., Fenza, G., Loia, V. and Nota, F.D. (2023) 'Content-based fake news detection with machine and deep learning: A systematic review', Neurocomputing, 530, pp. 91–103. doi: 10.1016/j.neucom.2023.02.005.
- [18] Essa, E., Omar, K. and Alqahtani, A. (2023) 'Fake news detection based on a hybrid BERT and LightGBM models', Complex & Intelligent Systems, 9(6), pp. 6581–6592. doi: 10.1007/s40747-023-01098-0.
- [19] Jing, J., Wu, H., Sun, J., Fang, X. and Zhang, H. (2023) 'Multimodal fake news detection via progressive fusion networks', Information Processing & Management, 60(1), 103120. doi: 10.1016/j.ipm.2022.103120.
- [20] Alghamdi, J., Lin, Y. and Luo, S. (2023) 'Towards COVID-19 fake news detection using transformer-based models', Knowledge-Based Systems, 274, 110642. doi: 10.1016/j.knsys.2023.110642.
- [21] Mohawesh, R., Maqsood, S. and Althebyan, Q. (2023) 'Multilingual deep learning framework for fake news detection using capsule neural network', Journal of Intelligent Information Systems, 60(3), pp. 655–671. doi: 10.1007/s10844-023-00788-y.
- [22] Narayanan, M.B., Ramesh, A.K., Gayathri, K.S. and Shahina, A. (2023) 'Fake news detection using a deep learning transformer based encoder-decoder architecture', Journal of Intelligent & Fuzzy Systems, 45(5), pp. 8001–8013. doi: 10.3233/JIFS-223980.
- [23] Truică, C.-O. and Apostol, E.-S. (2023) 'It's all in the embedding! Fake news detection using document embeddings', Mathematics, 11(3), 508. doi: 10.3390/math11030508.
- [24] Yadav, A., Gaba, S., Khan, H., Budhiraja, I., Singh, A. and Singh, K.K. (2024) 'ETMA: Efficient transformer-based multilevel attention framework for multimodal fake news detection', IEEE Transactions on Computational Social Systems, 11(4), pp. 5015–5027. doi: 10.1109/TCSS.2023.3255242.
- [25] Farhangian, F., Cruz, R.M.O. and Cavalcanti, G.D.C. (2024) 'Fake news detection: Taxonomy and comparative study', Information Fusion, 103, 102140. doi: 10.1016/j.inffus.2023.102140.
- [26] Fang, X., Wu, H., Jing, J., Meng, Y., Yu, B., Yu, H. and Zhang, H. (2024) 'NSEP: Early fake news detection via news semantic environment perception', Information Processing & Management, 61(2), 103594. doi: 10.1016/j.ipm.2023.103594.
- [27] Hashmi, E., Yayilgan, S.Y., Yamin, M.M., Ali, S. and Abomhara, M. (2024) 'Advancing fake news detection: Hybrid deep learning with FastText and explainable AI', IEEE Access, 12, pp. 44462–44480. doi: 10.1109/ACCESS.2024.3381038.
- [28] Al-Quayed, F., Javed, D., Jhanjhi, N.Z., Humayun, M. and Alnusairi, T.S. (2024) 'A hybrid transformer-based model for optimizing fake news detection', IEEE Access, 12, pp. 160822–160834. doi: 10.1109/ACCESS.2024.3476432.
- [29] Hu, B., Mao, Z. and Zhang, Y. (2025) 'An overview of fake news detection: From a new perspective', Fundamental Research, 5(1), pp. 332–346. doi: 10.1016/j.fmre.2024.01.017.
- [30] Raza, N., Abdulkadir, S.J., Abid, Y.A., Albouq, S.S., Alwadain, A., Rehman, A.U., Sumiea, E.H. and Farhan, M. (2025) 'Enhancing fake news detection with transformer-based deep learning: A multidisciplinary approach', PLOS ONE, 20(9), e0330954. doi: 10.1371/journal.pone.0330954.