Identifying fake news in political discourse behavior using machine learning methods

Victoria Vysotska[1,†], Nadiia Babkova[2,*,†], Dina Huliieva [2,†], Zoia Kochuieva[2,†], Nataliia Ugolnikova[2,†] and Mariia Kozulia[2,†]

[1] *Lviv Polytechnic National University, 12 Bandera str., 79013, Lviv, Ukraine*
[2] *National Technical University «Kharkiv Polytechnic Institute», 2, Kyrpychova str., Kharkiv, 61002, Ukraine*

Abstract

The paper examines the current problem of the rapid spread of political fake news, which can significantly influence public opinion, electoral processes, and intergovernmental relations. It highlights how the growth of online content and the dominance of social media as a source of information are speeding up the spread of misinformation and making it harder to spot it in time. On a theoretical level, the paper is based on the classification of approaches to fake detection (linguistic, topic-agnostic, machine learning, knowledge-based, and hybrid), emphasizing that only 28% of studies today focus specifically on ML methods, which determines the scientific and practical relevance of the proposed approach. The BERT model demonstrated the highest quality (Accuracy = 0.997; F1 = 0.998; ROC AUC = 0.997), significantly outperforming Random Forest (0.909; 0.930; 0.892) and Naive Bayes (0.881; 0.905; 0.870). Thus, modern transformer architectures provide not only maximum accuracy but also a balance between the completeness and specificity of predictions. The results obtained create the basis for the formation of highly accurate systems to counter disinformation, capable of responding quickly to the challenges of the modern media space.

Keywords

fake news, political discourse, artificial intelligence, machine learning, transformers, BERT, natural language processing

1. Introduction

During the last decade, the digital transformation of the media environment has led to an unprecedented growth in the volume of information circulating in the global public space. Social networks and news aggregators are increasingly serving as the primary source of political information and, at the same time, becoming the main channel for the dissemination of disinformation. “Fake news” – deliberately false reports disguised as credible media content with the aim of misleading the audience – is now seen not just as information noise, but as a factor that can selectively and systematically change political narratives, influence election

PhD Workshop on Artificial Intelligence in Computer Science at 9th International Conference on Computational Linguistics and Intelligent Systems (CoLInS-2025), May 15–16, 2025, Kharkiv, Ukraine
* Corresponding author.
† These authors contributed equally.
victoria.a.vysotska@lpnu.ua (V. Vysotska); nadiia.babkova@khpi.edu.ua (N. Babkova); dina.guliieva@khpi.edu.ua (D. Huliieva); zoia.kochuieva@khpi.edu.ua (Z. Kochuieva); nataliia.ugolnikova@khpi.edu.ua (N. Ugolnikova); mariia.kozulia@khpi.edu.ua (M. Kozulia);
0000-0001-6417-3689 (V. Vysotska); 0000-0002-5385-5761 (N. Babkova); 0000-0001-8310-745X (D. Huliieva) ; 0000-0002-4300-3370 (Z. Kochuieva) ; 0000-0003-2322-0922 (N. Ugolnikova); 0000-0002-4090-8481 (M. Kozulia)

results, and destabilize international relations [1]. Research confirms that the scale and speed of disinformation campaigns have never before reached today's level of threat, and empirical cases, such as the 2016 US presidential election, demonstrate the potential of fake news to become an "information weapon" in highly competitive political environments.

The issue of fake news is the focus of interdisciplinary research – from cognitive psychology, which studies the mechanisms of perception and dissemination of unreliable messages, to applied computer science, which develops algorithms for their automatic detection. Conceptually, approaches to detecting fakes are usually classified into five groups (**Figure 1**): language-based, topic-agnostic, machine learning, knowledge-based, and hybrid. Although each approach has its own advantages, a meta-analysis of the current literature shows that only about 28% of works are devoted to the application of machine learning (ML) methods — an area that can provide the greatest scalability and adaptability of monitoring systems.

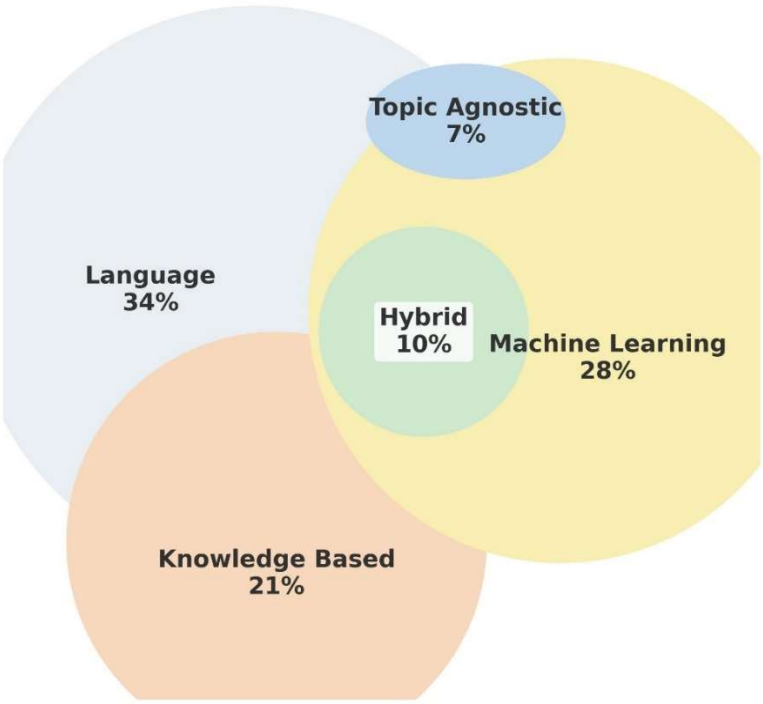

Figure 2: Categories of approaches used to identify fake news

At the same time, there are a number of gaps that hinder the effective implementation of ML solutions in the real media space. First, most studies use a broad thematic cross-section or adjust models on "mixed" corpora, where political texts constitute only a fraction of the total data volume. Such heterogeneity leads to a decrease in the quality of classification in domain subsets. Second, a significant portion of the work relies on classical algorithms with superficial features (Bag-of-Words, TF-IDF), ignoring the contextual connections between lexical units, which are critical for political discourse saturated with allusions, sarcasm, and ideological codes. Third, the low proportion of research devoted to model explainability complicates the integration of automated systems into data journalism and fact-checking workflows, as editors need transparent grounds for making decisions about content reliability.

Given these challenges, this article proposes a specialized software module for automated detection of political fake news, built on modern transformer architectures. The specific objectives of the study are formulated as follows:

1. Formation of a domain-specific corpus of political news (2021–2023), balanced between "real" and "fake" information.
2. Multilevel text preprocessing, including normalization, lemmatization, and noise filtering to improve the semantic purity of features.
3. Comparison of the effectiveness of classical algorithms (Naive Bayes, Random Forest) and the BERT contextual model in the task of binary classification.
4. Evaluation of models using a set of metrics (Accuracy, Precision, Recall, F1-Score, ROC AUC) and visual analysis of errors (confusion matrices) to identify systemic weaknesses.

The methodological basis is the assumption that transformer language models are capable not only of effectively capturing intra-contextual dependencies, but also of reconciling the discursive characteristics of political texts with markers of fakeness — from semantic inconsistencies to stylistic deviations. At the same time, it is advisable to assess whether classical algorithms, enhanced with Sentence-BERT embeddings, can compete with modern end-to-end solutions at lower computational costs — an aspect that is important for editorial offices with limited resources.

The scientific novelty of the article lies in the combination of two key components: domain specialization of the corpus, which minimizes thematic noise and allows relevant linguistic patterns to be derived, and the integration of deep contextual embeddings at the stage of both classical and neural network modeling. This approach is expected to significantly improve accuracy and compensate for the shortcomings of previous works that used superficial or generalized features.

The practical significance of the research is determined by the possibility of implementing the developed module in real-time systems — in particular, in the internal fact-checking tools of media companies, social media platforms for preliminary content filtering, as well as in state information security monitoring services. In addition, the modular architecture allows the solution to be scaled to multilingual corpora and adapted to other domains (e.g., medical or economic news) by retraining the language model on relevant data.

2. Related works

The first automated approaches to detecting fake news in political discourse relied mainly on superficial linguistic features — lexical frequency, readability indicators, and syntactic complexity. However, the rapid development of representation learning and the emergence of transformer architecture have led to a gradual transition to context-dependent methods. A comparative study of BERT-like encoders and large language models (LLMs) showed that even "classic" transformers configured with a softmax classifier are capable of delivering consistently high results on political news corpora, while LLM approaches require a more complex label extractor, but under certain conditions outperform BERT in terms of the flexibility of generative responses.

Of particular interest are studies focused on low-resource languages. For the Spanish-language corpus, it has been demonstrated that combining XLM-R and domain-specific training

on regional media increases detection accuracy by more than 8% compared to a multilingual model without adaptation [2]. This confirms the importance of the linguistic and cultural sensitivity of algorithms.

Political discourse is characterized by complex text structures containing thesis statements, quotations, and multi-level narratives. Hierarchical models such as the Three-level Hierarchical Attention Network and subsequent modifications (eHCAN, TR-HGAN) integrate attention at the word, sentence, and document levels and also take into account the topology of relationships between publications. Compared to flat transformers, they not only increase accuracy by 2–3 percentage points, but also provide interpretability through attention weights, which is critical for fact-checkers and journalists.

A growing body of research shows that text content alone is not a sufficient predictor of fakeness. A number of studies propose modeling user interactions, reposts, and temporal dynamics through graph neural networks (GNN). Ensemble Graph Neural Network (EGNN) aligns text vectors with social interaction features, showing a 3–5 percentage point increase in accuracy on political subsets of datasets such as GossipCop and PolitiFact. Similar results are reported by the authors of TR-HGAN, who incorporate the semantics of "user-tweet" relationships, minimizing classification errors in manipulative citations.

Political fakes are often accompanied by memes, edited images, or video clips. In response, multimodal architectures have emerged that synchronize BERT representations of text with CNN embeddings of images or audio codecs [3]. A systematic review in 2025 emphasizes that the multimodal approach demonstrates an accuracy gain of up to 6 percentage points compared to text-only methods. Specific implementations — MCOT (Contrastive Learning + Optimal Transport) and Event-Radar (event-driven multi-view learning) — introduce contrastive loss to bring together matched "text-image" pairs and push apart unmatched ones, which increases robustness to noisy data. Multimodal Fusion Network with BERT and VGG-19 demonstrates effectiveness even on small samples thanks to a high-level attention-fusion layer.

Identifying fake news is an important task in today's world, which is why a lot of research has been conducted on this topic recently.

In [4], the authors presented a basic approach to detecting fake news using a Naive Bayes classifier. The study analyzed Facebook data, including texts from three different sources (Politico, CNN, ABC News). This model was designed to detect not only fake news but also spam messages, as it was assumed that the methods of combating fake news and spam were similar. It showed a classification accuracy of about 74–75% for spam, but the results for detecting fake news were slightly lower, which the authors attribute to an imbalance in the dataset: only 4.9% of the materials were marked as fake.

The study [5] proposes methods for automating the detection of fake news on the social network Twitter. The work is based on the analysis of two open datasets (CREDBANK and PHEME) with ready-made accuracy estimates, and the tests were conducted on a dataset with fake news from BuzzFeed. Despite the positive results, which are consistent with the conclusions of previous works, this study has one key drawback. The developed algorithm focuses on analyzing only popular discussion threads. Because of this, the proposed approach is only effective for a limited selection of topics, which greatly narrows the range of its possible applications.

The most notable trend of the last five years has been the widespread adoption of BERT/Roberta/XLM-R models for binary classification of fake news. Researchers adapt them

through fine-tuning on domain-specific corpora of political content, which allows them to achieve an accuracy of over 97% [6]. At the same time, several areas for improvement have emerged: introducing external knowledge through knowledge graphs; integrating metadata about the source, date, or tone; and multilingual and multi-source scenarios. For example, the recent Dual-Stream Graph Augmented Transformer combines BERT text encoding with the topology of news propagation on social networks, achieving a significant increase in F1-score compared to baseline models. Similarly, KAHAN adds a layer of knowledge-based attention that takes user comments into account and enhances the detection of latent signs of fakeness.

There are also a number of studies, such as [7, 8], that place a strong emphasis on comparing algorithms and classification models, such as the aforementioned Naive Bayes, Decision Tree, Random Forest, Support Vector Machine (SVM), and KNN (k-Nearest Neighbors). However, an important aspect of this task is the vectorization of texts before further work with them, and these studies paid relatively little attention to this step and used methods that were not the most modern, namely Bag of Words and Term Frequency-Inverse Document Frequency Vectorizer (TF-IDF).

Generally speaking, the vast majority of studies focus on analyzing fake news in general, which can be considered a truly difficult task. Texts containing information from different areas of life may have their own unique key features that indicate fakeness, so even high accuracy rates cannot be considered a complete success. In study [9], the field of world politics will be considered separately, so it is expected that specific textual features characteristic of the selected field will be found.

An analysis of various studies and scientific publications on the topic of automatic identification of fake news has shown that most existing solutions mainly use classical machine learning models, namely Naive Bayes, Decision Tree, Random Forest, and SVM. These methods demonstrate acceptable accuracy while requiring a small amount of computing resources, especially on small datasets [10, 11]. However, it is important to note that the effectiveness of such methods is limited by the use of outdated vectorization methods that do not take into account the context and semantics of the text.

In addition, a number of studies show limited topics and sources, as well as a lack of specialization in the subject area of the text. This reduces the versatility and accuracy of the models when attempting to apply them to news of a specific focus, such as politics.

Although most contemporary work focuses on maximizing classification metrics, explainable AI is receiving increasing attention. Hierarchical attention networks and post-hoc methods (LIME, SHAP) demonstrate the potential to identify key phrases or argumentative structures that lead the model to a “fake” decision [12]. This is critical for political content, where issues of transparency and trust go beyond technical efficiency. A separate block of research analyzes the ethical risks of automatic labeling, including possible censorship implications and model biases against marginalized groups.

A review of the literature highlights several unresolved issues. First, despite significant progress, most models remain vulnerable to adversarial attacks, where minor changes in wording cause the system to produce false decisions. Second, universal approaches still show a noticeable drop in accuracy when transferred to new regions or electoral cycles. Third, the lack of multimodal political domain corpora significantly limits work with visual and audio manipulations [13]. Accordingly, promising areas include (a) adaptive retraining of LLM during

deployment, (b) strengthening models through contrastive learning on "mixed" and adversarial data, and (c) integrating deepfake content detectors into general fact-checking pipelines [14].

Thus, there is a need for more in-depth work on both the vectorization stage and the specialization of classifiers for a specific field — in this case, world politics. The application of modern language models and training on thematically homogeneous corpora appears to be the most promising direction for improving the accuracy of fake news detection.

3. Description of dataset

The dataset contains two types of news articles: fake and real. The real articles were obtained from the Reuters.com news website. Examples of fake articles were taken from various sources marked as unreliable by Politifact (a fact-checking organization in the US) and Wikipedia. Initially, the dataset contained various types of articles on different topics, but only those related to world politics were retained.

As a result, the dataset consists of two datasets, the first of which contains more than 11,000 real records, and the second contains more than 7,000 fake records. In addition to the text of the article itself, additional information is also stored: the title of the article, the type, and the date of publication. Most of the collected data relates to the period from 2021 to 2023.

4. Data preprocessing and Data analysis

First, all records that did not contain text or contained incorrect values were removed from the combined dataset. Then, the following steps were performed sequentially:

1. Conversion to a single case: all text characters were converted to lowercase in order to unify and eliminate differences between words that differed only in case.
2. Stop word removal: a list of English stop words was obtained from the nltk.corpus library. These words do not carry significant semantic meaning and, as a rule, do not contribute to improving the quality of classification, especially when using models based on frequency or embedding features.
3. Lemmatization: using the spaCy library and its English model en_core_web_sm, words were lemmatized — each word was reduced to its base form. This reduces the size of the dictionary and increases the generalizability of the model.
4. Token filtering: all non-alphabetic tokens (e.g., numbers, punctuation marks, and special characters) were excluded from the text. This further reduces the level of "noise" in the data and improves the quality of the features.
5. Removing redundant records: after completing all stages of cleaning, records in which the text became empty or consisted of one or two insignificant tokens were removed from the dataset.

The final cleaned dataset is saved in a separate file for further use without the need for re-preprocessing.

After cleaning the texts, they were converted into vector representations suitable for submission to machine learning models. Two different approaches to vectorization were implemented in this software module.

For further training of the Naive Bayes and Random Forest models, embeddings were generated using the pre-trained all-MiniLM-L6-v2 language model from the sentence-transformers library. This model allows each text to be converted into a vector of fixed dimension while preserving its semantic meaning [15-18]. Unlike classical vectorizers, Sentence-BERT is trained on semantic matching tasks and provides better quality when comparing the meanings of different sentences. The generated embeddings were saved to the embeddings.npy file, and the corresponding class labels were saved to labels.npy.

To train a binary classifier based on BERT, text tokenization was performed using BertTokenizer. This tokenization takes into account the specifics of the BERT architecture, including the division of words into sub-tokens, sequence length alignment, and attention mask formation.

5. Experiments and analysis of the results obtained

Three classification models were trained: Naive Bayes, Random Forest, and BERT Classifier. All models were trained on the same sample, which was previously divided into training and testing parts in a ratio of 80:20.

The Naive Bayes classifier was trained on previously obtained embeddings. The training and test samples were obtained using the train_test_split function from sklearn.model_selection. The GaussianNB class from the sklearn.naive_bayes library was used for training. Training was performed on the X_train array containing Sentence-BERT vectors and the corresponding y_train labels. After training on the test part (X_test), predictions y_pred_nb were obtained.

Random Forest was trained using a similar scheme. The classifier was created using the RandomForestClassifier class from the sklearn.ensemble library. When creating the model, the parameters of the number of trees in the ensemble, a fixed source of randomness, and the number of available processor cores for parallel processing were also explicitly specified. Predictions on the test sample were stored in the y_pred_rf array.

The BertForSequenceClassification architecture with two output classes (fake and real news) was used to train the BERT model. The training parameters also specified the number of epochs, batch size, and number of steps for logging. Predictions on the test sample were saved in the y_pred_bert array.

The quality of the classification models was evaluated on the test sample, and the following metrics were used for analysis:

1. Accuracy — the proportion of correctly classified examples.
2. Precision — accuracy: the proportion of genuinely fake news among those predicted as fake.
3. Recall — completeness: the proportion of correctly predicted fake news among all fake news.
4. F1-Score — harmonic mean between precision and recall.
5. ROC AUC — area under the ROC curve, reflecting the overall quality of the model.

The evaluation results for each of the three models are presented in Table 1.

Table 1
Evaluation of classification models

Metric	Naive Bayes	Random Forest	BERT Classifier
Accuracy	0.8808	0.9092	0.9970
Precision	0.8995	0.9018	0.9982
Recall	0.9114	0.9595	0.9969
F1-Score	0.9054	0.9297	0.9976
ROC AUC	0.8704	0.8922	0.9970

All metrics show that the BERT model significantly outperforms classical algorithms. It demonstrates a particularly noticeable advantage in terms of F1-Score and ROC AUC, which indicates not only high accuracy but also balanced predictions. The Naive Bayes model shows a bias towards precision, while Random Forest shows a bias towards recall, whereas BERT demonstrates high values for both.

For a more visual comparison, a histogram was constructed showing the comparative quality of the models according to the main metrics, which is presented in Figure 2.

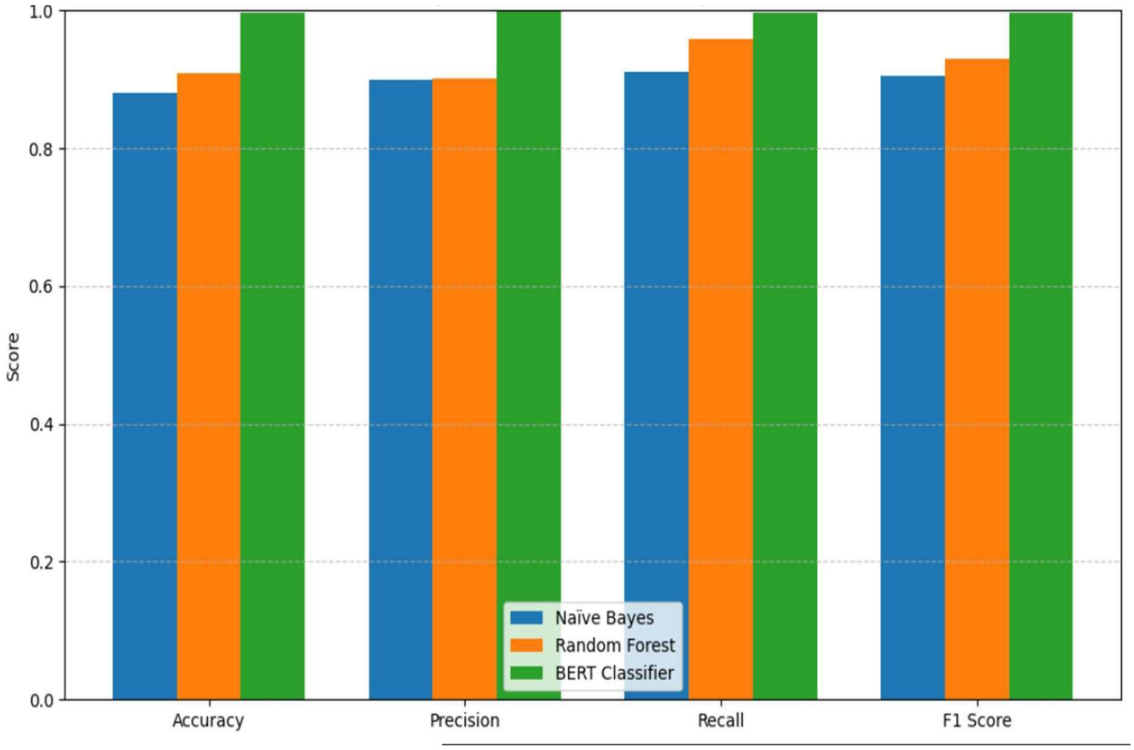

Figure 2: Metrics histogram

Error matrices were also constructed for each model, showing the nature of the errors made. These are presented in Figure 3. Based on these, we can conclude that Naive Bayes has a high proportion of false positives and false negatives, Random Forest demonstrates the best FN value, and BERT has virtually no errors — neither false positives nor false negatives.

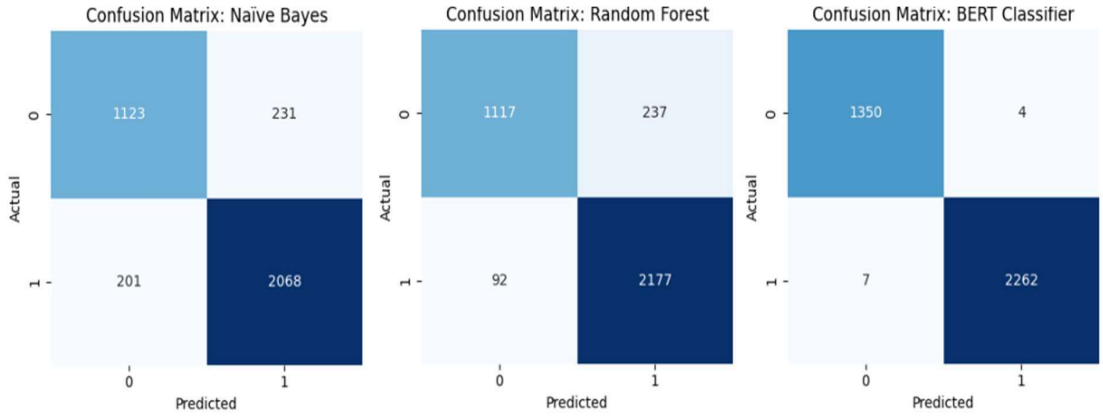

Figure 3: Error matrices

In addition to evaluating the models on the test sample, an analysis of the classification results for individual articles was performed. Several examples were taken from the PolitiFact website, pre-labeled as true and fake. After loading the module and preprocessing the text, the system used the saved retrained BERT model to predict the class label. It was found that the software module makes correct predictions on examples of true and fake news.

To better understand the reason for a particular conclusion, the text of the article can be analyzed: it contained emotionally charged expressions, references to controversial sources, and repeated mentions of political figures in an aggressive context. The BERT model, trained on a large corpus of news articles, compared these characteristics with examples of fake texts from the training set. Although the model itself works as a "black box," the high confidence of the prediction and the match with the observed patterns in the text confirm the correctness of the conclusion. This functionality of the software module allows it to be used not only within the framework of the experiment, but also as a full-fledged tool for the initial analysis of news content.

6. Conclusions and further research

The research proposes and empirically tests automated identification of political fake news based on modern transformer architectures. The experiment allowed comparing three different classification approaches: Naive Bayes, Random Forest, and BERT Base. The results demonstrate the clear superiority of the transformer model, with BERT providing not only the highest overall accuracy but also a balance between completeness and specificity of predictions, which is particularly important in conditions of high social risks of misclassification.

The selection of thematically homogeneous texts on world politics minimized information noise, allowing models to focus on specific rhetorical constructions characteristic of political discourse (sarcasm, epithets, references to authoritative sources, allusions). This confirms the feasibility of using domain-specific pre-training and thematic corpora when deploying news detectors in real editorial practices.

Sentence-BERT contextual embeddings significantly enhance classical algorithms, but even with the latest generation of vectorization, they are inferior to end-to-end transformers. This means that for editorial offices with limited computing resources, the Sentence-BERT + Random Forest model stack may be a compromise solution, but further improvements should be directed

toward optimizing and distilling large language models rather than increasing the complexity of traditional algorithms.

Transformer architectures are suitable for rapid integration into fact-checking systems, but the issue of explainability needs to be addressed. Despite its exceptional metrics, the "black box" nature of BERT makes it difficult for journalists to accept its conclusions without additional explanations. Further research should focus on combining internal hierarchical attention (which we partially tested) with post-hoc methods (LIME/SHAP) and developing user-friendly visualizers for editors.

The results highlight the vulnerability of traditional ML solutions to discursive manipulation. Analysis of error matrices showed that Naive Bayes and Random Forest more often misclassify texts with aggressive vocabulary or quotes taken from authoritative sources. In contrast, BERT, using deeper semantic dependencies, demonstrates resistance to such "masking" strategies. This is particularly valuable in the context of the growth of "adversarial" techniques for concealing propaganda under the guise of legitimate material.

In addition, thanks to a clear division into stages (pre-processing → vectorization → classification → evaluation), the proposed pipeline can be extended to other topics (economics, healthcare) or languages by simply retraining the language model on new data. This makes the approach relevant for international media platforms, where politically charged disinformation is multilingual in nature.

Despite the convincing results, the experiment has a number of limitations. First, although the corpus is balanced in terms of "fake/real," it is geographically biased toward English-language sources from the US and the UK; this may reduce the quality of generalization to texts from other information ecosystems (e.g., countries of the Global South). Second, the study used only text content, while modern political fakes are increasingly accompanied by images, videos, or audio clips, which were left out of the modeling field. Third, the system was not tested for resistance to deliberate adversarial attacks, where an attacker adjusts lexical patterns to bypass the detector. Finally, we did not consider the regulatory and ethical aspects of automatic filtering: the risk of over-moderation and possible model biases towards marginal groups require separate regulatory analysis.

As for further development, the next logical step is to integrate visual and audio features into the BERT text core. Combining language encoding with CNN representations of images (e.g., ResNet50) and video clips (I3D/ViT) will allow us to detect fakes that manipulate not only words but also graphic evidence (memes, deep-fake videos) aimed at emotionally influencing the audience.

In the dynamics of election campaigns, fake narratives change every month, so it is important to deploy an online learning mechanism with the ability to retrain the model on fresh examples. Active learning with minimal involvement of analysts will allow the detector to remain relevant at a limited cost of manual labeling.

A critical requirement remains the development of interactive dashboards with interpretations of decisions made: highlighting trigger phrases, assessing the impact of thematic and tonal indicators, and a graph of source cross-references. This will strengthen the trust of journalists, fact-checkers, and regulators in automatic labeling.

Given the global nature of political disinformation, it is necessary to test the model in other languages (Spanish, Ukrainian, Arabic) by retraining XLM-R or mBERT. The experience of the

Spanish-language corpus, where accuracy increased by 8%, shows that linguistic and cultural sensitivity is crucial for interregional universality.

We plan to include adversarial training strategies (adding grammatical errors, rearranging words, synonym replacement) in the training process, as well as contrastive learning, which increases robustness to lexical mutations. This minimizes the risk of bypassing the detector through automated "paraphrasing" services.

Successful implementation requires compliance with regulatory acts (Digital Services Act, AI Act, GDPR) and the development of "soft moderation" protocols that will limit excessive content blocking. Further research should include bias audits and assessments of the social consequences of automatic censorship.

The study demonstrates that transformer language models, adjusted for the domain-specific corpus of world politics, are currently the most effective tool for automatic detection of fake news. The metrics obtained outperform traditional ML algorithms, confirming the hypothesis of the fundamental advantage of contextual learning of representations. At the same time, a complex set of technological, ethical, and regulatory challenges means that further progress in this area will require a synergistic approach: a combination of multimodal analytics, explainable decision-making mechanisms, adaptive learning, and deep consideration of the sociocultural factors of disinformation dissemination.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] M. Raza, E. Acheampong, X. Chen, "Towards reliable fake news detection: Enhanced attention-based transformer model," *Information*, vol. 5, no. 3, p. 43, 2024. doi:10.3390/jcp5030043.

[2] P. Qi, Y. Bu, J. Cao, W. Ji, R. Shui, J. Xiao, T.-S. Chua, "Fakesv: A multimodal benchmark with rich social context for fake news detection on short-video platforms," in: *Proc. AAAI Conf. Artif. Intell.*, vol. 37, pp. 14444–14452, 2023. doi:10.48550/arXiv.2211.10973.

[3] X. Chen, Y. Li, "Ensemble graph neural networks for fake news detection using user stance and text features," *Data & Knowledge Engineering*, vol. 158, p. 102121, 2024. doi:10.1016/j.rineng.2024.103081.

[4] T. Suárez, P. Martínez, R. García, "Fake news detection: Comparative evaluation of BERT-like models and large language models," *arXiv preprint*, arXiv:2412.14276, 2024. URL: https://arxiv.org/abs/2412.14276.

[5] M. Liu, K. Yan, Y. Liu, R. Fu, Z. Wen, C. Li, "Exploring modality disruption in multimodal fake news detection," *arXiv preprint*, arXiv:2504.09154, 2025. URL: https://arxiv.org/abs/2504.09154.

[6] Y. Xuan, L. Wang, J. Li, "Exploiting stance similarity and graph neural networks for fake news detection," *Pattern Recognition Letters*, vol. 166, pp. 44–51, 2023. doi:10.1016/j.patrec.2023.10.013.

[7] H. Li, M. Zhou, T. Zhang, "Decision-based heterogeneous graph attention network for fake news detection," *Knowledge-Based Systems*, vol. 285, p. 110433, 2025. doi:10.48550/arXiv.2501.03290.

[8] R. Fernandes, P. Patel, K. Sharma, "Ensemble techniques for robust fake news detection: Integrating transformers, NLP and machine learning," *Sensors*, vol. 24, no. 18, p. 6062, 2024. doi:10.3390/s24186062.

[9] P. Li, J. Zhang, C. Song, "CroMe: Cross-modal tri-transformer and metric learning for multimodal fake news detection," *arXiv preprint*, arXiv:2501.12422, 2025. URL: https://arxiv.org/abs/2501.12422.

[10] S. Choudhary, A. Verma, N. Bhatia, "A systematic review of multimodal fake news detection on social media using deep learning models," *Information Processing & Management*, vol. 62, p. 104213, 2025. doi:10.1016/j.rineng.2025.104752.

[11] Q. Yang, B. Sun, Y. Liu, "Hybrid optimisation-driven fake news detection using a reinforced modified transformer model," *Scientific Reports*, vol. 15, p. 99936, 2025. doi:10.1038/s41598-025-99936-3.

[12] X. Zhou, J. Wu, R. Zafarani, "SAFE: Similarity-aware multimodal fake news detection," in: *Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining*, Springer, pp. 354–367, 2020. URL: https://arxiv.org/abs/2003.04981.

[13] J. Xiong, C. Wang, X. Guo, "Similarity-aware multimodal prompt learning for fake news detection," *Information Sciences*, vol. 665, p. 119446, 2023. doi:10.1016/j.ins.2023.119446.

[14] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020. doi:10.1016/j.inffus.2019.12.012.

[15] M. Nyzova, V. Vysotska, L. Chyrun, Z. Hu, Y. Ushenko, D. Uhryn, "Smart tool for text content analysis to identify key propaganda narratives and disinformation in news based on NLP and machine learning," *Int. J. Comput. Netw. Inf. Security (IJCNIS)*, vol. 17, no. 4, pp. 113–175, 2025. doi:10.5815/ijcnis.2025.04.08.

[16] R. Lynnyk, V. Vysotska, Z. Hu, D. Uhryn, L. Diachenko, K. Smelyakov, "Information technology for modelling social trends in Telegram using E5 vectors and hybrid cluster analysis," *Int. J. Inf. Technol. Comput. Sci. (IJITCS)*, vol. 17, no. 4, pp. 80–119, 2025. doi:10.5815/ijitcs.2025.04.07.

[17] P. Meel, D. K. Vishwakarma, "Multi-modal fusion using fine-tuned self-attention and transfer learning for veracity analysis of web information," *arXiv preprint*, arXiv:2109.12547, 2021. URL: https://arxiv.org/abs/2109.12547.

[18] D. Levkivskyi, V. Vysotska, L. Chyrun, Y. Ushenko, D. Uhryn, and C. Hu, "Agile methodology of information engineering for semantic annotations categorization and creation in scientific articles based on NLP and machine learning methods," *Int. J. Inf. Eng. Electron. Bus. (IJIEEB)*, vol. 17, no. 2, pp. 1–50, 2025. doi: 10.5815/ijieeb.2025.02.014.