

# Hybrid method for building a balanced Ukrainian-language news corpus for fake news detection<sup>\*</sup>

Yuliia Sobchuk<sup>1,†</sup>, Sviatoslav Krushelnytskyi<sup>1,†</sup>, Khrystyna Lipianina-Honcharenko<sup>1,†</sup>, Andrii Ivasechko<sup>1,†</sup> and Tetiana Drakokhrust<sup>1,†</sup>

<sup>1</sup> Faculty of Computer Information Technologies, West Ukrainian National University, 46000 Ternopil, Ukraine

## Abstract

This paper presents a method for constructing a balanced Ukrainian-language news corpus for fake-news detection that combines LLM-based controlled generation with editorial verification. The truthful subset is collected from authoritative media using topic- and year-stratified sampling (2022–2025), while fake examples are produced via a parameterized LLM prompt controlling tone, style, manipulation types, and topics. The pipeline comprises multi-stage normalization, stop-word removal, lemmatization (Stanza), language identification, near-duplicate filtering (hybrid cosine/Jaccard-trigram similarity), and human moderation of borderline cases. The resulting corpus contains ~40k texts (~20k “Trusted” and ~20k “Fake”) with an average length of ~250 tokens and a bimodal length distribution. Reproducibility is ensured by publishing data schemas and fixed 80/20 splits. A BiLSTM baseline with FastText (300d) achieves 99.25% accuracy, 0.9925 macro-F1, and 0.985 MCC, with false-positive/false-negative rates  $\leq 0.9\%$ . These results indicate strong class separability and validate the corpus as a benchmark for future studies, including transformer-based models, ablation of synthetic components, robustness assessment, and probability calibration.

## Keywords

fake news detection, Ukrainian corpus, large language models, LLM-based generation, editorial verification, text preprocessing, BiLSTM, FastText, reproducibility. <sup>1</sup>

## 1. Introduction

The intensive dynamics of the Ukrainian media space in 2022–2025 have been accompanied by a significant increase in the volume of disinformation, which complicates the verification of news content and necessitates reliable data for training fake news detectors. Despite substantial progress in global research, Ukrainian-language annotated corpora remain limited in both scale and thematic-stylistic coverage, while the transfer of models trained on English-language data often leads to a loss of accuracy due to linguistic and domain shifts. The absence of reproducible corpus construction pipelines and transparent quality control procedures further complicates the comparability of results.

The aim of this study is to develop a method for creating a balanced corpus of Ukrainian-language news for the task of fake news detection by combining controlled generation using large language models with editorial verification. The proposed approach integrates stratified selection of trustworthy materials from reputable sources with synthetic generation of fake examples under controlled parameters of tone, style, type of manipulation, and topic. The methodology is designed to ensure diversity, minimize bias, and comply with ethical requirements for the use of generative AI.

The pipeline involves multi-stage normalization and linguistic processing of texts, language verification, removal of duplicates and stylistic anomalies, as well as manual moderation of

<sup>\*</sup>BAITmp’2025: The 2nd International Workshop on “Bioinformatics and Applied Information Technologies for medical purpose”, November 12–13, 2025, Ben Guerir, Morocco

<sup>1</sup> Corresponding author.

<sup>†</sup> These authors contributed equally.

✉ yuliia.sob4uk@gmail.com (Y. Sobchuk); sviatoslav.kru@gmail.com (S. Krushelnytskyi); kh.lipianina@wunu.edu.ua (K. Lipianina-Honcharenko); andrewivasechko@gmail.com (A. Ivasechko); t.drakokhrust@wunu.edu.ua (T. Drakokhrust)

ORCID 0009-0005-9542-0028 (Y. Sobchuk); 0009-0001-1257-9051 (S. Krushelnytskyi); 0000-0002-2441-6292 (K. Lipianina-Honcharenko); 0009-0002-8623-7828 (A. Ivasechko); 0000-0002-4761-7943 (T. Drakokhrust)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

borderline cases. To ensure reproducibility, stochastic parameters are fixed, data schemas are published, and train-validation split lists are provided. As a result, a balanced corpus of approximately 40,000 news texts was obtained (around 20,000 in each of the “Trusted” and “Fake” classes), with representative thematic and stylistic coverage.

This paper presents a method for constructing a balanced corpus of Ukrainian-language news for fake news detection that combines controlled LLM generation with editorial verification. Section 2 summarizes related work; Section 3 describes the methodology and data collection pipeline, including stratified selection of reliable materials, parameterized generation of synthetic examples, preprocessing, and quality control; Section 4 provides the corpus composition and statistics, as well as baseline benchmarks (including BiLSTM+FastText); Section 5 presents experimental results with a confusion matrix and analysis of training dynamics; Section 6 formulates conclusions, highlights scientific significance, and outlines directions for future work, including testing transformer models, conducting ablation studies, assessing robustness, and addressing ethical considerations in the use of generative AI.

## 2. Related work

Early attempts at fake news detection were primarily based on classical machine learning algorithms with manual feature engineering. In particular, the study presented in [1] highlights the use of methods such as Support Vector Machines (SVM), Naive Bayes classifiers (NB), and Random Forests (RF). The effectiveness of these algorithms largely depended on the quality of the selected linguistic and meta-feature characteristics.

The research conducted in [2] focuses on classifying fake news in social media based on textual content. This work applied four traditional text feature extraction methods (TF-IDF, Count Vector, Character-level Vector, N-Gram Level Vector) and ten different machine learning and deep learning classifiers. The obtained results demonstrated that textual fake news can be effectively classified, with classification accuracy ranging from 81% to 100% depending on the classifier used, with convolutional neural networks (CNN) showing particularly high effectiveness.

With the rise of deep learning, fake news detection methods have gained new momentum. Recurrent neural networks, particularly bidirectional LSTMs (Bi-LSTMs), enable models to learn context in both forward and backward directions.

In [3], a combination of BERT and LSTM was used for fake news classification based on headlines. The model was trained on the FakeNewsNet dataset (PolitiFact, GossipCop), achieving accuracy improvements of 2.5% and 1.1%, respectively, compared to the baseline BERT model, confirming the effectiveness of combining transformers and recurrent networks.

The study in [4] proposed a fake news detection framework that combines analysis of news content and social context. The model is built on a Transformer architecture with an encoder for feature extraction and a decoder for predicting the subsequent behavior of the news. To address the lack of labeled data, the authors applied a custom automatic labeling technique. Experiments with real-world data showed that the model provides higher accuracy in early detection (within minutes of dissemination) compared to baseline methods.

In [5], the authors explored the potential of fine-tuning the modern language model GPT-3 for the task of fake news detection. The model was adapted on the ISOT dataset and demonstrated high effectiveness, achieving an accuracy of 99.90%, precision of 99.81%, recall of 99.99%, and an F1-score of 99.90%, significantly outperforming existing solutions. These results confirm the promise of using GPT-3 to combat disinformation in social media and news outlets.

The study in [6] presents the Generative Bidirectional Encoder Representations from Transformers (GBERT) framework, which combines BERT’s deep contextual understanding with GPT’s generative capabilities for fake news classification. Both models were fine-tuned on two real-world benchmark datasets, achieving an accuracy of 95.30%, precision of 95.13%, recall of 97.35%, and an F1-score of 96.23%. The obtained results demonstrate the high efficiency of GBERT

and the potential of this approach in countering the spread of disinformation in the digital environment.

In [7], a review of machine learning algorithms and datasets used for fake news detection was conducted. Among the most effective models identified were the Stacking Method with 99.9% accuracy, BiRNN, and CNN — both at 99.8%. Most studies relied on data from controlled environments (e.g., Kaggle) or from sources without real-time updates, which limits their practical applicability in social media, where disinformation spreads most actively. The most frequently used datasets included Kaggle, Weibo, FNC-1, COVID-19 Fake News, and Twitter. The authors emphasize the need to expand topics beyond political news and to apply hybrid methods in future research.

The study in [8] presents the OLTW-TEC method (Online Learning with Sliding Windows for Text Classifier Ensembles), developed for detecting disinformation in the Ukrainian-language information space. The approach combines an ensemble of classifiers with a “sliding window” mechanism for dynamically updating the model to incorporate new data, thereby increasing its adaptability to changing fake news dissemination tactics. The method was tested on a specially constructed dataset of authentic and fake news, achieving an accuracy of 93%. The results confirm the effectiveness of OLTW-TEC and its suitability for operating under information warfare conditions, as well as its potential for adaptation to other languages and regions.

A comparative study [9] showed that RNN, LSTM, and Bi-LSTM models achieve similar results with around 91% accuracy, although LSTM outperformed in recall and RNN in precision. This highlights the importance of selecting an architecture that aligns with specific objectives.

Further research has focused on transformer architectures. For instance, the authors of [10] evaluated the performance of BERT, CNN, Bi-LSTM, and their ensemble combination. The latter achieved the highest accuracy — 98.24% — demonstrating the effectiveness of hybrid solutions.

In [11], transformers (BERT, RoBERTa, GPT-2) were compared with graph neural networks (GNN). Transformers showed significantly better results: RoBERTa reached 99.99% on ISOT, and GPT-2 achieved 99.72% on WELFake, highlighting their ability to work with contextually rich data.

The study in [12] analyzed the effectiveness of various machine learning methods for detecting disinformation in Ukrainian-language news collected during the military conflict. Evaluated models included logistic regression, SVM, random forest, gradient boosting, KNN, decision trees, XGBoost, and AdaBoost. The random forest demonstrated the best results. The authors emphasize the importance of adapting models to the specifics of the task and the need for further research in this area.

The authors of [13] argue that combining transformers with text summarization further increases accuracy. RoBERTa fine-tuned on summarized content achieved 98.39%, which is among the highest metrics among modern models.

Special attention should be paid to hybrid models that integrate Word2Vec vectors with CNN and LSTM. In [14], the authors focused on fake news classification using a combination of machine learning (ML) and natural language processing (NLP) methods based on textual content. They compared several modern ML models and neural networks. Experiments showed that all traditional ML models achieved over 85% accuracy, while neural networks outperformed them, reaching over 90% accuracy.

Research (Table 1) on fake news detection has evolved from classical machine learning approaches with manual feature engineering to deep and transformer-based architectures (see [1–14]). In the early stages, SVM, NB, and RF were applied with textual representations such as TF-IDF or Bag-of-Words, where accuracy largely depended on feature selection and data domain. Subsequent studies focused on neural models: RNN/LSTM/Bi-LSTM achieved results around ~91% [9], while hybrids combining CNN with vector representations (e.g., FastText) improved performance to 0.99 in Accuracy and 0.97–0.99 in F1-score [13–14]. Transformers, particularly BERT/RoBERTa and their ensembles, reached 98.24% and higher [10], with some well-known datasets (ISOT, WELFake) reporting values up to 99.99% [5]. However, these metrics vary

significantly across datasets and experimental setups, complicating the generalization of conclusions and accurate comparison of approaches.

**Table 1**

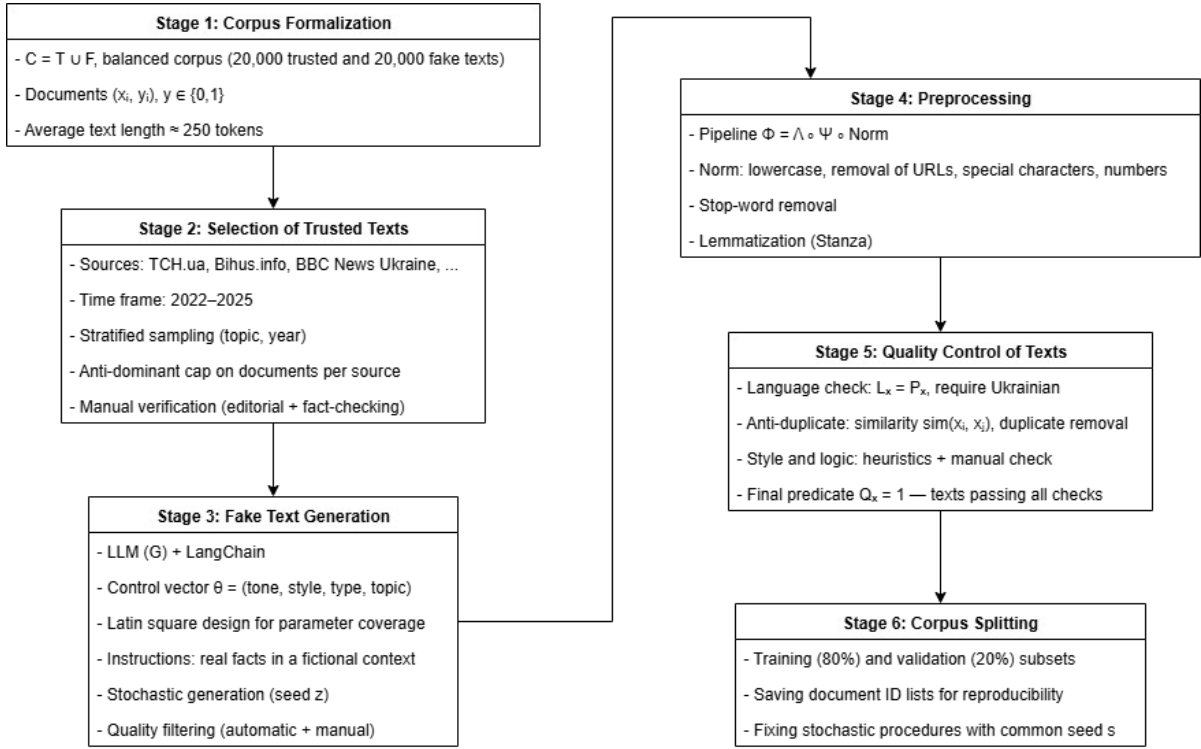
Comparison of existing approaches

№	Author(s), year	Type of approach	Technologies, model	Accuracy\results
1	Alluri et al., 2023 [1]	Review	Systematic analysis of methods	—
2	Abdulrahman, 2020 [2]	Classical ML/DL	SVM, NB, RF, CNN	81% – 100%
3	Dhiman et al., 2024 [6]	Transformers	GBERT	95,30%
4	Airlangga, 2024 [9]	Deep learning	RNN, LSTM, Bi-LSTM	~91%
5	Kuntur et al., 2024 [10]	Transformers/GNN	BERT, RoBERTa, GPT-2, GNN	99.99%
6	Saadi et al., 2025 [11]	Transformers + Summary	RoBERTa with text summarization	98.39%
7	Hashmi et al., 2024 [13]	Hybrid (DL + NLP)	FastText + CNN + LSTM	Acc: 0.99, F1: 0.97–0.99
8	Lai et al., 2022 [14]	Neural networks	CNN, LSTM	~90%

Existing studies indicate progress in methods but reveal several research gaps, particularly for the Ukrainian-language segment: (i) a lack of large public annotated corpora with transparent preparation pipelines, fixed splits, and detailed documentation; (ii) limited evaluation under temporal and domain shift scenarios, with insufficient attention to robustness against paraphrasing/adversarial attacks and probability calibration; (iii) inadequate analysis of bias across topics/genres, as well as the impact of synthetic examples on model generalization and stability; (iv) incomplete reproducibility due to the absence of publicly available code, fixed seeds, and detailed preprocessing protocols. The scientific significance of this research lies in addressing these gaps by creating a reproducible Ukrainian-language corpus with balanced classes, a clearly specified construction and quality control methodology, and establishing benchmark standards that allow accurate comparison of modern architectures and investigation of their robustness under realistic conditions.

### 3. Method

The following outlines the sequential stages of constructing a balanced corpus of Ukrainian-language news for fake news detection, combining verified texts from reputable media with controlled LLM-generated examples (stages 1–6, Fig. 1).



**Figure 1:** Stages of constructing the Ukrainian-language news corpus.

*Stage 1. Corpus formalization.*

Let  $C = T \cup F$  be the final corpus of Ukrainian-language news for the two-class fake news classification task, where  $F$  is the set of trusted texts and  $T$  is the set of fake texts. Each document is a pair  $(x_i, y_i)$ , where  $x_i \in \Sigma^*$  is the text, and  $y_i \in [0, 1]$  is the class label ( $y = 1$  – «Fake»,  $y = 0$  – «Trusted»). The corpus is balanced:  $|T| = |F| = 20\,000$ , so the prior class probabilities are  $\pi_0 = \pi_1 = \frac{1}{2}$ .

Let the empirical distribution of text lengths in tokens be denoted as  $\hat{p}_L(l)$ . Following preprocessing, the average length is  $\hat{\mu}_L = E_{\hat{p}_L}[L] \approx 250$  tokens.

*Stage 2. Selection of trusted texts.*

Let  $S = [TCH.ua, Bihus.info, BBC News Україна, \dots]$  be the set of sources. Over the time interval  $t \in [2022, 2025]$ , the initial sample is formed as

$$U_T = \{x : x \text{ was published on } s \in S, t \in [2022, 2025]\}.$$

To avoid dominance of individual sites or topics, stratified sampling is applied by topic  $\tau \in [politics, economy, society, defense, health, \dots]$  and publication year. Within each stratum  $(\tau, year)$  random sampling without replacement is performed with an upper limit  $m_s$  of documents per source  $s$  (anti-dominance cap). Each candidate undergoes manual verification of editorial standards and fact-checking; acceptance is denoted by the predicate  $RT(x) \in [0, 1]$ . The final set is:

$$T = \{x \in U_T : RT(x) = 1\}$$

*Stage 3. Fake text generation.*

Fake texts are generated by a large language model  $G$  via LangChain using a parameterized prompt template  $\tau(\theta)$ . The control vector is

$$\theta = (tone, style, type, topic)$$

where

–  $tone \in [neutral, alarming, reassuring]$ ;

- $style \in \{analytical, populist, ironic, factual\}$ ;
- $type \in \{disinformation, manipulation, emotional influence, propaganda\}$ ;
- $topic \in \{politics, economy, education, defense, infrastructure\}$ .

To ensure diversity,  $\theta$  is covered almost uniformly (Latin square / combinatorial sweep), and  $G$  is instructed to use real facts/persons in a fictional context while avoiding fantastical events or clichés. Generation occurs as

$$x = G(\tau(\theta), z),$$

where  $z$  is the stochastic seed/model temperature. Each synthetic text undergoes both automatic and manual quality filtering.

*Stage 4. Preprocessing.*

Let

$$\Phi = \Lambda \circ \Psi \circ Norm$$

be the preprocessing pipeline,

where

$Norm(x)$ : lowercase conversion, removal of URLs, hashtags, special characters, and numeric markers;

$\Psi(x)$ : stop-word removal;

$\Lambda(x)$ : lemmatization (Stanza for Ukrainian).

The resulting text  $x' = \Phi(x)$  is fed into the validation and statistical analysis modules.

*Stage 5. Quality control of texts.*

Three independent acceptance predicates are applied:

1. Language identification.

$$L(x) = P(x), \text{ requiring } L(x) \geq \tau_{lang}.$$

2. Anti-duplicate check. Let

$$\hat{c}(x_i, x_j) = \alpha \cdot \cos \cos(tfidf(x_i), tfidf(x_j)) + (1 - \alpha) \cdot J_3(x_i, x_j),$$

where  $J_3$  is the Jaccard similarity of trigrams. A text  $x$  is discarded if  $\max_{j < i} \hat{c}(x, x_j) \geq \tau_{dup}$  (practically implemented via MinHash/LSH).

3. Style/logic check. Automatic heuristics (length, anomalous n-gram repetition) plus manual review  $R(x)$ .

The final filter is

$$Q(x) = 1 \{ L(x) \geq \tau_{lang}, \sim(x, x_j) < \tau_{dup} \} \wedge R(x),$$

Where  $\tau_{lang}$  and  $\tau_{dup}$  denote threshold parameters for language and duplication filtering, respectively.

The final sets  $T, F$  consist only of texts that satisfy  $Q(x) = 1$ .

*Stage 6. Corpus splitting.*

The corpus is split into training and validation subsets while preserving class balance:

$$C^{train} \cup C^{val} = C, / C^{train} / \approx 0.8 / C / , / C^{val} / \approx 0.2 / C / .$$

Lists of document IDs for each split are stored separately to ensure reproducibility, and all stochastic procedures are fixed using a common seed  $s$ .

## 4. Result

To construct the trusted news class, we used materials from reputable Ukrainian information sources, including TCH.ua, Bihus.info, BBC News Україна, and others. The full list of trusted sources is available online [15]. The news covers the period 2022–2025 and topics related to politics, economy, society, military events, healthcare, etc. In total, approximately 20,000 texts were selected, each undergoing manual verification for accuracy and compliance with contemporary



journalistic style. Manual verification was conducted by two authors, with the dataset evenly split between them for independent assessment.

Fake news was generated using the large language model Gemini 2.0. Generation was performed via the LangChain interface using a pre-designed prompt template that specified the parameters of the resulting text.

Each fake news item was created taking into account the following characteristics::

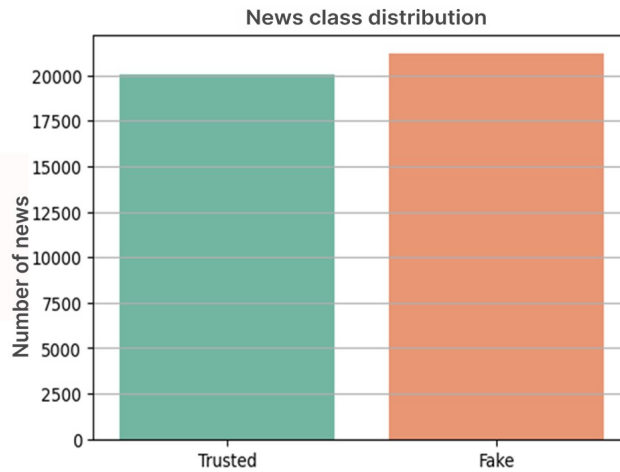
- Tone: neutral, alarming, reassuring;
- Writing style: analytical, populist, ironic, factual;
- Type of fake: disinformation, manipulation, emotional influence, propaganda;
- Topic: politics, economy, education, defense, infrastructure.

The prompt template also instructed the model to incorporate real facts, institutions, and persons within a fictional context, making the texts as close as possible to authentic media content. Generation was accompanied by guidelines to avoid fantastical events or obvious clichés.

Despite automated generation, all fake news items underwent manual verification. Texts with low plausibility, artificial language, logical inconsistencies, or violations of style guidelines were filtered out. As a result, a balanced corpus was created, consisting of 20,000 fake and 20,000 trusted news articles.

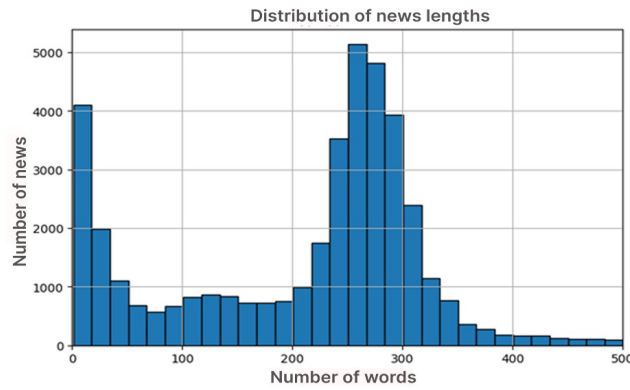
The news corpus [16] was cleaned of noisy elements: hyperlinks, special characters, numeric markers, and hashtags were removed. Texts were converted to lowercase, stripped of stop-words, and lemmatized using the Stanza library, which supports Ukrainian morphology.

The combined corpus contains approximately 40,000 Ukrainian-language news articles with nearly equal class representation (Fig. 2): Trusted  $\approx 20,000$ , Fake  $\approx 20\text{--}21,000$ ; the deviation from a 50/50 split does not exceed  $\approx 10\%$ . Such balance reduces the risk of metric bias toward the larger class.



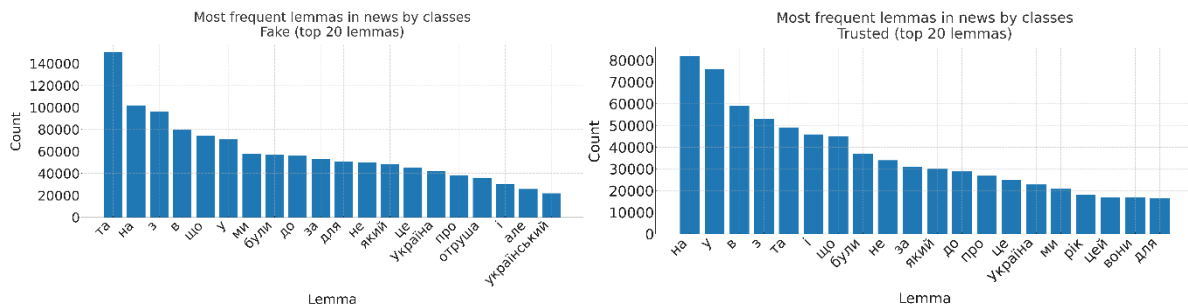
**Figure 2:** Histogram of class distribution.

The length distribution exhibits a pronounced bimodality (Fig. 3): the first local peak corresponds to short notes ( $\sim 20\text{--}60$  words), while the second corresponds to full-length articles of  $\approx 250\text{--}300$  words. The mean length is approximately 250 tokens/words, which matches a typical news item and provides sufficient context for linguistic features.



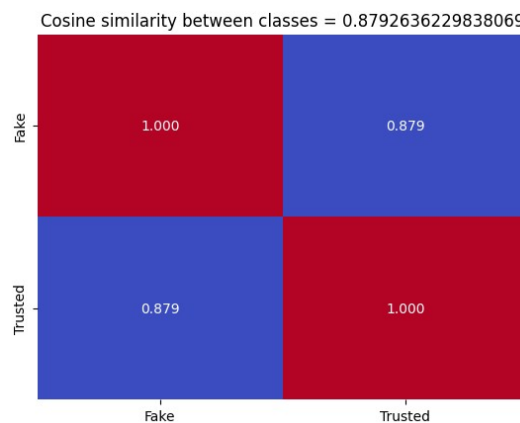
**Figure 3:** Histogram of news length distribution.

The top-20 lemmas by class (Fig. 4) show the expected dominance of function words (conjunctions, prepositions), indicating a homogeneous underlying syntactic structure across both classes. At the same time, differences in content lemmas are noticeable: in the Fake class, terms like “український” (Ukrainian), “ситуація” (situation), “про” (about), “але” (but) appear more frequently, whereas in Trusted, “Україна” (Ukraine), “рік” (year), “вони” (they), “для” (for) are more common. This reflects stylistic distinctions: fake texts tend to use generalizing and evaluative formulations, while trusted texts feature nominative references to institutions/country and temporal markers.



**Figure 4:** Frequency plot of the most common lemmas.

The cosine similarity between the sets of key lemmas was 0.879, indicating a high lexical overlap. This suggests that fake and trusted news often share the same topical vocabulary, which makes the classification task realistic and shifts the discriminative power towards stylistic and contextual features rather than mere word occurrence. The heatmap of cosine similarities (Fig. 5) further illustrates this overlap, showing the strong lexical proximity between the two classes.



**Figure 5:** Heatmap of cosine similarity between Fake and Trusted classes.



The bimodality in text lengths reflects two dominant forms of news presentation (short “notes” and full-length articles), which is useful for building robust models: the classifier is exposed to different styles and text volumes. High model metrics on the balanced corpus confirm strong class separability and the quality of data preparation (cleaning, language and duplicate control). Differences in content lemmas illustrate stylistic signals that can be used as interpretable features or for further bias analysis.

For vector representation, FastText in skip-gram mode was applied. The vectorizer was trained on the preprocessed corpus with the following hyperparameters: vector size – 300, number of epochs – 15, context window width – 5, minimum word frequency – 10.

News vectorization was performed by truncating or padding with zero vectors to a fixed length of 100 tokens. The classifier architecture is based on a bidirectional LSTM network with additional Dropout and Dense layers. Optimization was performed using Adam with an initial learning rate of 0.001.

After training on 80% of the dataset and validating on the remaining 20%, the model achieved an accuracy of 99.25% (Table 2). Precision and recall coefficients exceed 0.99 for both classes, which is also confirmed by the confusion matrix (Fig. 5). The training dynamics are shown in Figure 6, illustrating a gradual decrease in the loss function without signs of overfitting.

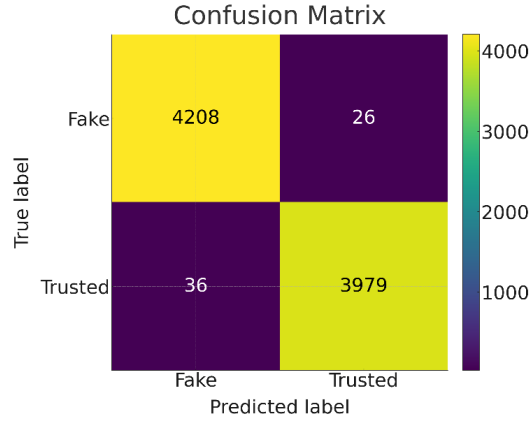
**Table 2**  
Frequency of Special Characters

Class	Precision	Recall	F1-score	Support
Fake	0.9915	0.9939	0.9927	4234
Trusted	0.9935	0.9910	0.9923	4015
Average	0.9925	0.9924	0.9925	4124

The column “Support” indicates the number of instances belonging to each class in the test dataset. Overall classification performance (see Fig. 5): Accuracy = 0.9925 (99.25%), macro-averaged F1 = 0.9925, Matthews correlation coefficient (MCC) = 0.985 (see Table 2). From the confusion matrix (Fig. 6):

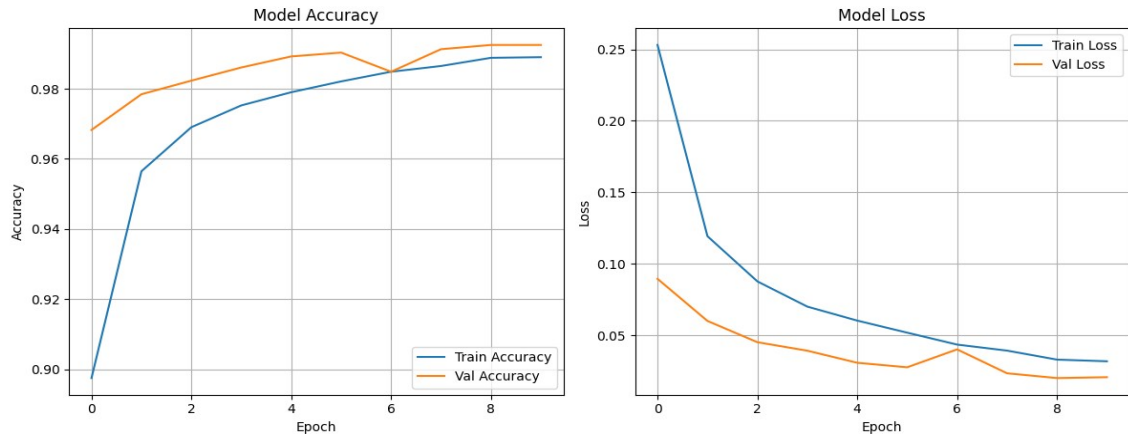
- Fake class: TP = 4208, FN = 26, FP = 36, TN = 3979; Precision = 0.9915, Recall = 0.9939, F1 = 0.9927; TPR = 0.9939, TNR = 0.9910, FNR = 0.0061, FPR = 0.0090;
- Trusted class: TP = 3979, FN = 36, FP = 26, TN = 4208; Precision = 0.9935, Recall = 0.9910, F1 = 0.9923; TPR = 0.9910, TNR = 0.9939, FNR = 0.0090, FPR = 0.0061.

The average balanced accuracy equals  $\frac{TPR_{Fake} + TPR_{Trusted}}{2} = 0.99245$ , corresponding to a BER = 0.00755. The low false positive and false negative rates ( $\leq 0.9\%$ ) in each class confirm strong class separability and the absence of bias toward any label.



**Figure 6:** Confusion Matrix.

The training dynamics (Fig. 7) show a monotonic increase in accuracy on both the training and validation sets, reaching  $\approx 0.99$  and plateauing after approximately the 7th epoch. The loss function decreases steadily across both subsets without divergence. The absence of rising validation error and the minimal generalization gap indicate no signs of overfitting under the chosen hyperparameters (FastText 300d, window = 5, epochs = 15, min\_count = 10; BiLSTM + Dropout + Dense, Adam optimizer,  $\eta = 10^{-3}$ ).



**Figure 7:** Accuracy and loss dynamics on training and validation.

## 5. Conclusion

This study introduces a balanced corpus of Ukrainian-language news for fake news detection, comprising  $\sim 40,000$  texts ( $\sim 20k$  “Trusted” and  $\sim 20k$  “Fake”) from 2022–2025. Trusted data were sourced from verified media, while synthetic fakes were generated with LLMs under controlled prompts, followed by normalization, filtering, and lemmatization. The corpus shows clear stylistic differences between classes and an average length of  $\sim 250$  tokens, making it suitable for machine learning.

Evaluation with a BiLSTM + FastText model achieved accuracy of 99.25% and macro-F1 of 0.9925, confirming both the quality of the dataset and the feasibility of automated fake news detection. Misclassification rates remained below 1%, with stable learning dynamics and no overfitting.

The dataset and approach can be applied in practice for media monitoring and early detection of disinformation in Ukraine. Future work will include benchmarking transformer models, robustness testing, and releasing artifacts to support reproducible research and regular updates of the corpus.

## Declaration on Generative AI

The authors have not employed any Generative AI tools.

## References

- [1] Alluri, C. R., Reddy, K. S., Sarma, B. M., & Kumar, D. S. (2023). Fake news detection: a systematic review and knowledge mapping. *The Journal of Supercomputing*, 79(2), 1735–1770.
- [2] A. Abdulrahman and M. Baykara, "Fake News Detection Using Machine Learning and Deep Learning Algorithms," 2020 International Conference on Advanced Science and Engineering (ICOASE), Duhok, Iraq, 2020, pp. 18-23, doi: 10.1109/ICOASE51841.2020.9436605.
- [3] Rai, N., Kumar, D., Kaushik, N., Raj, C., & Ali, A. (2022). Fake News Classification using transformer based enhanced LSTM and BERT. *International Journal of Cognitive Computing in Engineering*, 3, 98-105.
- [4] Raza, S., & 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), 335–362. <https://doi.org/10.1007/s41060-021-00302-z>.
- [5] Hemina, K., Boumahdi, F., Madani, A., Remmide, M.A. (2023). A Cross-Validated Fine-Tuned GPT-3 as a Novel Approach to Fake News Detection. In: Zantout, H., Ragab Hassen, H. (eds) *Proceedings of the International Conference on Applied Cybersecurity (ACS) 2023*. ACS 2023. *Lecture Notes in Networks and Systems*, vol 760. Springer, Cham. .
- [6] Dhiman, P., Kaur, A., Gupta, D., Juneja, S., Nauman, A., & Muhammad, G. (2024). GBERT: A hybrid deep learning model based on GPT-BERT for fake news detection. *Heliyon*, 10(16).
- [7] Villela, H. F., Corrêa, F., Ribeiro, J. S. D. A. N., Rabelo, A., & Carvalho, D. B. F. (2023). Fake news detection: a systematic literature review of machine learning algorithms and datasets. *Journal on Interactive Systems*, 14(1), 47-58.
- [8] Lipianina-Honcharenko, K., Soia, M., Yurkiv, K., & Ivasechko, A. (2024, May). Evaluation of the effectiveness of machine learning methods for detecting disinformation in Ukrainian text data. In *Proceedings of the Seventh International Workshop on Computer Modeling and Intelligent Systems (CMIS-2024)*, Zaporizhzhia, Ukraine.
- [9] Airlangga, G. (2024). Advancing fake news detection: a comparative study of RNN, LSTM, and Bidirectional LSTM Architectures. *Jurnal Teknik Informatika C.I.T Medicom*, 16(1), 13–23.
- [10] Kuntur, S., Krzywda, M., Wróblewska, A., Paprzycki, M., & Ganzha, M. (2024). Comparative Analysis of Graph Neural Networks and Transformers for Robust Fake News Detection: A Verification and Reimplementation Study. *Electronics*, 13(23), 4784.
- [11] Saadi, A., Belhadeh, H., Guessas, A., and Hafirassou, O. 2025. Enhancing Fake News Detection with Transformer Models and Summarization. *Engineering, Technology & Applied Science Research*. 15, 3 (Jun. 2025), 23253–23259. DOI:<https://doi.org/10.48084/etasr.10678>.
- [12] Lipianina-Honcharenko, K., Bodyanskiy, Y., Kustra, N., & Ivasechko, A. (2024). OLTW-TEC: online learning with sliding windows for text classifier ensembles. *Frontiers in Artificial Intelligence*, 7, 1401126.
- [13] Hashmi, E., Yayilgan, S. Y., Yamin, M. M., Ali, S., & Abomhara, M. (2024). Advancing fake news detection: Hybrid deep learning with fasttext and explainable ai. *IEEE Access*.
- [14] Lai, C.-M., Chen, M.-H., Kristiani, E., Verma, V. K., & Yang, C.-T. (2022). Fake News Classification Based on Content Level Features. *Applied Sciences*, 12(3), 1116. <https://doi.org/10.3390/app12031116>.
- [15] Dataset. (2025). Trusted sources for fake news detection. Google Drive. [https://drive.google.com/drive/folders/13c\\_QRvuMuXTByYZkzbJOq4VcXzURniVx?usp=drive\\_link](https://drive.google.com/drive/folders/13c_QRvuMuXTByYZkzbJOq4VcXzURniVx?usp=drive_link).
- [16] Sobchuk, Yulia (2025). fake\_true\_ukrainian\_news\_dataset.csv. figshare. Dataset. <https://doi.org/10.6084/m9.figshare.29257568.v1>.