КІБЕРБЕЗПЕКА ТА ЗАХИСТ ІНФОРМАЦІЇ

UDC 004.8-9 https://doi.org/10.32515/2664-262X.2025.12(43).1.90-98

Olga Lozynska, **Oksana Markiv**, Assoc. Prof., PhD tech. sci.,
Victoria Vysotska, Assoc. Prof., Doct. tech. sci.
Lviv Polytechnic National University, Lviv, Ukraine
e-mail: olha.v.lozynska@lpnu.ua, oksana.o.markiv@lpnu.ua, victoria.a.vysotska@lpnu.ua

Identifying Sources and Participants of Propaganda in TikTok Using Machine Learning

The paper presents an approach to identifying sources of disinformation, fakes, and propaganda on the TikTok social network using modern natural language processing (NLP) and artificial intelligence methods. The main goal of the research is to create a system capable of automatically analyzing comments on videos with propaganda content, as well as the source of their distribution and potential participants in propaganda. As part of the research, a corpus of comments in Ukrainian and Russian was manually collected, which were classified as propaganda or neutral. Based on the analysis of the dataset, certain criteria were identified for identifying sources of disinformation and its potential participants, in particular through the use of the Russian language, repeated propaganda narratives, as well as the fakeness of accounts. The steps of the comment preprocessing algorithm are given. Two approaches to analysis were developed: a classification model based on RandomForestClassifier and a clustering model using the KMeans algorithm. Both models use RoBERTa transformers for the respective languages, as well as an additional manually generated set of comment features. A Telegram bot and a graphical interface were built for the convenient use of the system, which allows receiving comments from TikTok, classifying them, and providing the user with analysis results. The proposed system is a relevant tool for information security and combating propaganda in the digital environment.
disinformation, propaganda sources, dataset, RoBERTa model, clustering, potential propaganda participants, set of criteria for identifying propaganda participants

Problem Statement. The modern information space is constantly transforming under the influence of technological progress, but these changes bring not only new opportunities but also unprecedented challenges. One of the most acute and socially significant problems of today is the rapid spread of disinformation, fake news, and propaganda. The methods and tools available today, designed to detect such content, often demonstrate insufficient efficiency, as they lack the time to adapt to new, increasingly complex techniques for generating disinformation. This situation creates serious risks for public trust, can destabilize political processes, and undermine the foundations of information security both at the level of individual citizens and entire states. It is to counteract these complex threats that the development of a system for identifying sources of disinformation and propaganda, as well as their potential participants, is being initiated.

Comments, as a tool for spreading propaganda, can serve as a channel for promoting a certain ideological or political narrative. For example, bots or coordinated accounts leave messages that repeat the same messages, and disinformation is placed in comments under the guise of "the opinions of ordinary people". Therefore, it is very important to develop tools to identify potential participants in the spread of propaganda and disinformation, as well as the source itself.

Analysis of recent research and publications. Given the growth of information warfare in social networks, especially in the context of military aggression and hybrid threats, there is a need for automated solutions for monitoring and detecting propaganda. Comments in the online space play a dual role in the processes of information influence: on the one hand,

they can be a channel for the spread of propaganda through bot activity and manipulative messages; on the other, a source for studying the mechanisms of information influence and a possible tool for countering disinformation. For the classification of comments and reviews in the Ukrainian language, such language models as BERT, DistilBERT, RoBERTa, XLM-RoBERTa RoBERTa and Ukr-RoBERTa are used. The authors chose the RoBERTa transformer model for research and classification.

RoBERTa is an optimized BERT model retrained with an improved training methodology, more data, and hardware resources, as proposed in [1]. It improves BERT by carefully and intelligently optimizing the training hyperparameters for BERT. RoBERTa uses the same architecture as BERT. However, unlike BERT, it only trains the generation of the missing token during pretraining (BERT was also pretrained to predict the next sentence). That is, RoBERTa without the concept of predicting the next sentence is similar to BERT and uses dynamic masking.

The scientific paper [2] presents the use of different algorithms for the multi-class and cross-lingual task of fake news detection using the RoBERTa model. The results of the macro F1-score evaluation using the pre-trained RoBERTa model are 28.60% for the monolingual task in English.

Scientists [3] conducted a study of the problem of automatic fake news detection in Urdu. The article presents the use of various word embedding models, such as RoBERTa, ALBERT, XLM-RoBERTa, and Multilingual BERT. Model RoBERTa is trained on Urdu news data from Pakistani newspapers and gives the results of accuracy 92%. It is built on BERT, trained with larger mini-batches and learning rates.

The study [4] describes the fine-tuning of the BERT, DistilBERT, XLM-RoBERTa and Ukr-RoBERTa models for sentiment analysis of Ukrainian language reviews, since transformer models demonstrate better understanding of the context and show high efficiency in solving natural language processing tasks. The dataset for the study consists of approximately 11,000 user comments in Ukrainian on various topics. The authors proposed to classify text data into two classes: positive and negative. After pre-processing the text, the dataset was divided into training and test samples in a ratio of 80:20. The effectiveness of the classification models was evaluated using metrics such as accuracy, recall, precision, and F1-score. According to the research results, the XLM-RoBERTa model achieved the highest accuracy of 91.32%. However, considering the time required to train the model and all classification metrics, Ukr-RoBERTa is a more optimal choice.

The study [5] proposes a framework that allows you to quickly create corpora for classifying Ukrainian-language news with minimal data annotation efforts. The results of testing on this set showed that the ukr-RoBERTa, ukr-ELECTRA, and XLM-R large models demonstrate the highest accuracy rates. At the same time, XLM-R large and ukr-ELECTRA perform better in processing longer texts, while ukr-RoBERTa shows an advantage in classifying short text fragments.

The scientific paper [6] presents EMOBENCH-UA, the first annotated corpus for emotion recognition in Ukrainian-language texts. The annotation scheme was adapted from English-language research in this field [7], taking into account instructions and cultural features. The data was collected through crowdsourcing on the Toloka.ai platform, which made it possible to ensure high-quality labelling. During the research, various approaches to analysing this corpus were evaluated – from basic linguistic models and translated synthetic sets to modern large language models (LLM). The results of the experiments demonstrate the complexity of the task of emotional classification for languages with limited resources, in particular Ukrainian, and indicate the need for further development of models and training materials focused specifically on the Ukrainian language context.

The article [8] addresses the problems of sentiment analysis in Ukrainian social networks, where users frequently switch encodings between Russian and other languages. Authors present COSMUS (COde-Switched MUltilingual Sentiment for Ukrainian Social media) – a corpus of 12,224 texts collected from Telegram channels, product review sites, and open datasets annotated with positive, negative, neutral, and mixed sentiment classes, as well as language tags (Ukrainian, Russian, with code-switching). Fine-tuning UkrRoBERTa using GPT-4o-based data augmentations provides a maximum accuracy of 73.6%, which exceeds the baseline levels of mBERT.

The scientific paper [9] presents LiBERTa Large – the first BERT Large model that was fully trained from scratch exclusively on the Ukrainian-language corpus. By using large-scale multilingual corpora with a significant proportion of Ukrainian texts, the model forms a solid foundation for solving natural language processing problems in Ukrainian. The results show that LiBERTa Large outperforms both existing multilingual and monolingual models pre-trained in Ukrainian, and demonstrates performance comparable to or better than models based on knowledge transfer from English.

The authors [10] propose a conceptual model of the information warfare system, identify its weaknesses, and suggest an improvement mechanism, in particular through the creation and dissemination of content on social media, expanding audience coverage on WhatsApp, and utilizing IT, advanced technologies, and AI for faster threat detection. The proposed approach aims to reduce the vulnerability of state institutions and society to disinformation and propaganda, increase the responsiveness of countermeasures, and improve the effectiveness of risk assessment in the sphere of national security.

The article [11] describes a method for detecting sources of disinformation based on ensemble machine learning models. The authors used two types of text embeddings and corresponding classification models - linear and logistic regression. At the final stage, an ensemble of models was applied, which allowed combining their predictive capabilities and increasing the accuracy from 71% to 78%.

Based on the analysis of scientific papers, the authors selected the RoBERTa language model for the study. To test the model, a dataset of comments was manually collected, which will be used for training and classification.

Task statement. The purpose of the research is to create a system to combat information aggression, disinformation, fakes, and propaganda in social networks, in particular in TikTok, using modern methods of natural language processing and artificial intelligence. Of great importance in this is the identification of potential participants in the process, as well as the source of the spread of fakes and propaganda. To solve this problem, it is necessary to: analyze methods and means of identifying sources of disinformation and propaganda; analyze the collected dataset of comments using deep learning and clustering models (such as BERT, RoBERTa), which will allow dividing comments by semantic features into propaganda and neutral ones; develop a system that automatically receives comments from videos in TikTok as a source, processes them and provides the user with information about the probable presence of propaganda.

Presentation of the Main Material. A dataset of comments was manually collected for training and testing the system. Comments were selected from videos containing trigger tags, such as #ukraine, #zelensky, #crimea, etc. Since the bot specializes in detecting russian propaganda, those comments that promote russian narratives, such as denial of Ukraine's independence and sovereignty, denial of the legitimacy of the Ukrainian government, justification of aggression, cruel treatment of the Armed Forces of Ukraine and Ukrainians, etc., were marked as propaganda. Thus, about 700 comments in Ukrainian and 1,300 comments in russian were collected and labeled for the classification model. More than 6,000

and 10,000 comments in Ukrainian and Russian, respectively, were collected for the clustering model.

The data sets have the following form (see Table 1 and Table 2).

If there is propaganda in the comment, the value of the field "Propaganda" is equal to 1, otherwise, it is 0.

Table 1 – Description of the dataset for the classification model

Column name	Data type	Description
Comment_ID	Integer	Comment identifier
Comment	String	Comment text
Propaganda	Bool	Presence or absence of propaganda features

Source: developed by the authors.

A custom scraper was used to collect comments. However, due to certain functional limitations of this scraper, the auxiliary tool TTCommentExporter was used. In addition, the generated dataset contained not only Ukrainian and russian, but also others. Therefore, a software module was created that automatically separates comments by language features. After that, an algorithm for pre-processing data from the dataset was developed.

Table 2 – Description of the dataset for the clustering model

Column name	Data type	Description
Comment_ID	Integer	Comment identifier
Comment	String	Comment text

Source: developed by the authors.

The data pre-processing algorithm contains the following steps:

Step 1. Reading the comment. The comment is sent by the user to the Telegram bot or downloaded from an Excel file (during the training stage).

Step 2. Determining the language of the comment. Using the langdetect library, it is determined whether the comment is written in Ukrainian or russian. The choice of the transformer model (UA_model or RU_model) depends on this.

Step 3. Tokenization. The AutoTokenizer from the corresponding Hugging Face model is used. The text is converted into a sequence of tokens taking into account:

• maximum size (truncation to 128 tokens);

• automatic padding;

• conversion to tensor (return_tensors='pt').

Step 4. Obtaining the embedding. From the AutoModel model (e.g. youscan/ukr-roberta-base), the CLS embedding (the first vector of the last layer) is extracted, which is considered a representation of the entire text.

Step 5. Extracting custom features.

Step 6. Scaling the features. Before merging, the custom features are normalized using StandardScaler() in order to avoid scale imbalance (CLS embedding has a different scale).

Step 7. Merging the features. The final feature is formed as a horizontal concatenation

Step 8. Feeding to the classifier / clusterer.

Step 9. Saving. Embeddings or merged vectors can be saved as .npy or .pkl for reuse; the trained model (KMeans, StandardScaler) is also saved using joblib.

The system implemented language models for comments in Ukrainian and russian:

• UA_model – for processing Ukrainian comments;

• RU_model – for processing russian comments.

The development of the models took place in two stages: experiments with classification with a teacher and the final transition to clustering without a teacher.

The main process of the system's operation includes the following stages:

• Model training. Training a machine learning model on data. This includes text vectorization, clustering, and testing the effectiveness of the model.

• User interface creation. Developing a graphical interface where users can send links to video materials that need to be analyzed and displaying the results of the analysis.

• Comment clustering. The process in which the system accepts a link provided by the user, receives comments, generates data sets, and determines which comments are propaganda and which are not. Includes vectorization, model application, and result calculation.

• Result display. Visualization of analysis results, including the percentage of propaganda. Fig. 1 shows a UML activity diagram that describes the basic process of the system.

Classification model. For the classification model, data collection and preparation were carried out, namely, a small corpus of comments was manually collected and marked, where each text was marked as:

• 1 – propaganda;

• 0 – not propaganda.

The next stage was to calculate embeddings. For this, the comments were processed through youscan/ukr-roberta-base for the Ukrainian language, and for each comment, a representation was made in the form of a CLS vector. In addition, the following manual set of features was manually extracted for each comment:

- the presence of flags RU/UA;
- the proportion of capital letters;
- the number of character repetitions;
- the number of emojis;
- the length of the text, etc.

After that, the features were scaled using StandardScaler and combined with the CLS vector to form the final feature vector.

The next stage was training the classifier. For this, the RandomForestClassifier model was selected, which was trained on the collected vectors. Parameters were obtained via GridSearchCV with cross-validation (cv=5). The model was trained to determine whether the text contains propaganda.

Results for the classification model are next:

- recall for propaganda (1) – 100%;
- recall for non-propaganda (0) – only 5%;
- due to sample imbalance, the model tended to label almost everything as "propaganda".

In addition, an analysis of the criteria and parameters that can influence the identification of sources of disinformation and propaganda, as well as their potential participants, was conducted. Among the criteria that make it possible to identify sources of disinformation, one can single out certain propaganda narratives, as well as the language in which the comment is written (the vast majority of comments are written mostly in russian). Another criterion for identifying potential participants in propaganda is the fakeness of the account itself and, accordingly, all posts and comments of this participant. These criteria can be described as a set:

$$K = \{L, F_u, F_p\},$$

where L is the language, F_u is the fakeness of the account and F_p is the fakeness of posts.

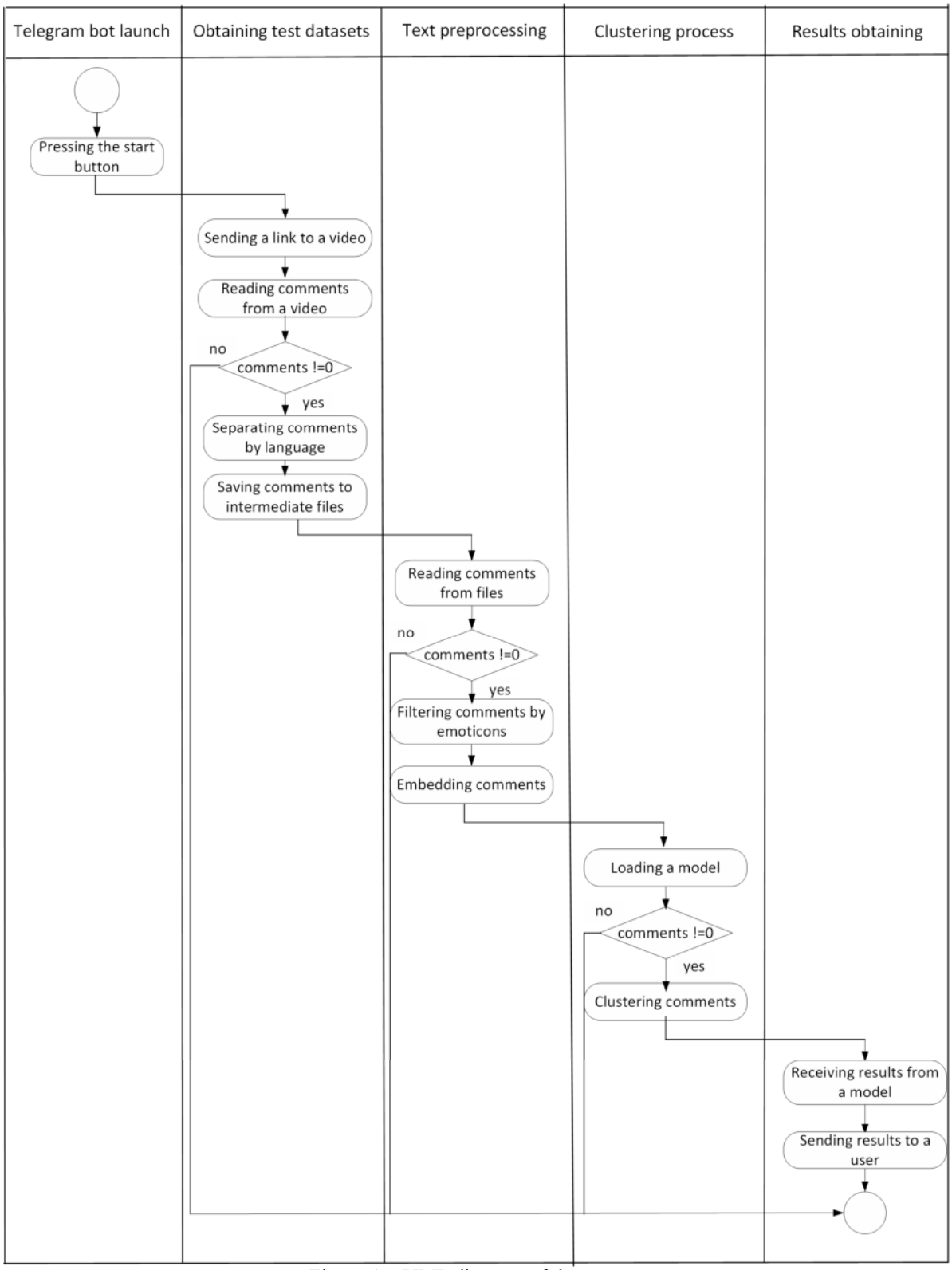

Figure 1 – UML diagram of the system

Source: developed by the authors

Clustering model. For the clustering model, the RoBERTa language model was used separately for the Ukrainian (UA_model) and russian (RU_model) language classes. During the collection of comments and their labeling, it was found that many comments are “gray”, that is, neutral or not unambiguously propaganda.

The same structure was implemented for the two models (UA_model and RU_model) using the corresponding RoBERTa language model, performing vectorization and feature extraction, as well as training the KMeans clusterer (2 clusters). The KMeans clustering algorithm divides all vectors into 2 groups without knowing the true labels. After clustering, the system analyzes the content of each cluster by the frequency of propaganda phrases. For each cluster, the following is calculated:

• the proportion of texts containing keywords;

• the proportion of propaganda emojis.

Accordingly, the cluster with the highest "propaganda score" is considered to correspond to the "Propaganda" class.

For new texts, the following occurs:

- vectorization (CLS + features);
- prediction of belonging to one of the clusters;
- determination of whether it is propaganda, based on comparison with the propaganda cluster.

Fig. 2 shows an example of using a Telegram bot to detect the presence of propaganda in comments.

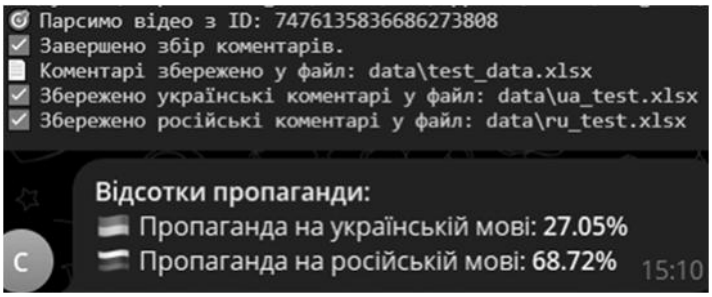

Figure 2 – The result of the clustering model

Source: developed by the authors.

In the process of research, a system was developed that allows for analyzing and classifying comments from videos on the TikTok network as propaganda or neutral. This will allow determining the source of the spread of propaganda, fakes, and disinformation.

Acknowledgements. The research was carried out with the grant support of the National Research Fund of Ukraine "Information system development for automatic detection of misinformation sources and inauthentic behavior of chat users", project registration number 33/0012 from 3/03/2025 (2023.04/0012).

Conclusions. As a result of the research, an effective system was developed for automatic detection of propaganda in comments to TikTok videos. The use of modern transformer models (RoBERTa), machine learning algorithms (classification and clustering), as well as the development of a full-fledged software package with a graphical interface, made it possible to implement a full cycle of data analysis, from collection to visualization of results. The structure of the system assumes a modular architecture: individual components are responsible for text vectorization, manual feature extraction, pre-processing, clustering, decision-making, and user interaction. Experiments showed high sensitivity of the classification model to detect propaganda messages, and accordingly, to track potential participants in disinformation and propaganda. However, due to the imbalance of the sample, limitations in accuracy were identified for neutral content. The clustering model allowed for better work with "gray zones" - ambiguous comments. Additionally, an analysis was conducted of parameters that allow us to identify not only propaganda, but also potential sources of its

distribution. These parameters include the use of characteristic propaganda narratives, the language of comments (mostly Russian), and signs of fake accounts - the same type of posts, low level of profile authenticity, etc. The created system can be integrated into platforms for monitoring the information space, which will significantly strengthen the capabilities of state and public institutions in countering hybrid threats in the information sphere.

List of References

1. Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L., Stoyanov V. Roberta: A robustly optimized Bert pretraining approach. 2019. *arXiv preprint* arXiv:1907.11692. URL: https://doi.org/10.48550/arXiv.1907.11692.
2. Arif M., Tonja A. L., Ameer I., Kolesnikova O., Gelbukh A., Sidorov G., Meque A. G. M. CIC at CheckThat! 2022: Multi-class and Cross-lingual Fake News Detection. *CEUR Workshop Proceedings* 2022. Pp. 434 – 443.
3. Kalraa S., Vermas P., Sharma Y., Chauhan G. S. Ensembling of Various Transformer Based Models for the Fake News Detection Task in the Urdu Language. *Proceedings of the Forum for Information Retrieval Evaluation*. 2021.
4. Prytula M. Fine-tuning BERT, DistilBERT, XLM-RoBERTa, and Ukr-RoBERTa models for sentiment analysis of Ukrainian language reviews. *Artificial Intelligence*. 2024. № 29(2). Pp. 85–97. URL: https://doi.org/10.15407/jai2024.02.085.
5. Panchenko D., Tytarenko S., et al. Evaluation and Analysis of the NLP Model Zoo for Ukrainian Text Classification. *In Information and Communication Technologies in Education, Research, and Industrial Applications*. Springer. 2022. Pp. 109–123. URL: https://doi.org/10.1007/978-3-031-20834-8_6.
6. Dementieva D., Babakov N., Fraser A. EmoBench-UA: A Benchmark Dataset for Emotion Detection in Ukrainian. *arXiv preprint.* 2025. URL: https://doi.org/10.48550/arXiv.2505.23297.
7. Saif M. Mohammad. Ethics Sheet for Automatic Emotion Recognition and Sentiment Analysis. *Computational Linguistics*. 2022. № 48(2). Pp. 239–278. URL: https://doi.org/10.48550/arXiv.2109.08256.
8. Shynkarov Y., Solopova V., Schmitt V. Improving Sentiment Analysis for Ukrainian Social Media Code-Switching Data. *In Proceedings of UNLP-2025 Workshop (COSMUS benchmark)*. 2025.
9. Haltiuk M., Smywiński-Pohl A. LiBERTa: Advancing Ukrainian Language Modeling through Pre-training from Scratch. *In Proceedings of the Third Ukrainian Natural Language Processing Workshop.* 2024. Pp. 120–128.
10. Доренський О.П., Улічев О.С., Задорожний К.О., Коваленко А.С., Дрєєва Г.М. Концептуальна модель системи інформаційного протиборства координаційного центру з питань національної безпеки і оборони. *Центральноукраїнський науковий вісник. Технічні науки*. 2024. Вип. 10(41). Ч. 2. С. 23-31. URL: https://doi.org/10.32515/2664-262X.2024.10(41).2.23-31.
11. Лозинська О.В., Марків О.О., Висоцька В.А. Метод виявлення джерел дезінформації на основі ансамблевих моделей машинного навчання. *Біоніка інтелекту*. 2025. № 1 (102). С. 11-19. URL: https://doi.org/10.30837/ bi.2025.1(102).02.

References

1. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized Bert pretraining approach. *arXiv preprint* arXiv:1907.11692. DOI: 10.48550/arXiv.1907.11692.
2. Arif, M., Tonja, A. L., Ameer, I., Kolesnikova, O., Gelbukh, A., Sidorov, G., & Meque, A. G. M. (2022). CIC at CheckThat! 2022: Multi-class and Cross-lingual Fake News Detection. *CEUR Workshop Proceedings* (pp. 434 – 443).
3. Kalraa, S., Verma, P., Sharma, Y., & Chauhan, G. S. (2021). Ensembling of Various Transformer Based Models for the Fake News Detection Task in the Urdu Language. *Proceedings of the Forum for Information Retrieval Evaluation*.
4. Prytula, M. (2024). Fine-tuning BERT, DistilBERT, XLM-RoBERTa, and Ukr-RoBERTa models for sentiment analysis of Ukrainian language reviews. *Artificial Intelligence*, 29(2), 85–97. https://doi.org/10.15407/jai2024.02.085
5. Panchenko, D., Tytarenko, S., et al. (2022). Evaluation and Analysis of the NLP Model Zoo for Ukrainian Text Classification. *In Information and Communication Technologies in Education, Research, and Industrial Applications* (pp. 109–123). Springer. https://doi.org/10.1007/978-3-031-20834-8_6.

6. Dementieva, D., Babakov, N., & Fraser, A. (2025, May 29). EmoBench-UA: A Benchmark Dataset for Emotion Detection in Ukrainian. *arXiv preprint*. https://doi.org/10.48550/arXiv.2505.23297.
7. Saif M. Mohammad (2022). Ethics Sheet for Automatic Emotion Recognition and Sentiment Analysis. *Computational Linguistics*, 48(2), 239–278. https://doi.org/10.48550/arXiv.2109.08256.
8. Shynkarov, Y., Solopova, V., & Schmitt, V. (2025). Improving Sentiment Analysis for Ukrainian Social Media Code-Switching Data. *In Proceedings of UNLP-2025 Workshop (COSMUS benchmark)*.
9. Haltiuk, M., & Smywiński-Pohl, A. (2024). LiBERTa: Advancing Ukrainian Language Modeling through Pre-training from Scratch. *In Proceedings of the Third Ukrainian Natural Language Processing Workshop* (pp. 120–128).
10. Dorenskyi, O.P., Ulichev, O.S., Zadorozhnyi, K.O., Kovalenko, A.S., & Dreeva, G.M. (2024). The conceptual model of the information counteraction system of the coordination center for national security and defense issues. *Central Ukrainian Scientific Bulletin. Technical Sciences*, 10(41), Part 2, 23-31. DOI: 10.32515/2664-262X.2024.10(41).2.23-31.
11. Lozynska, O.V., Markiv, O.O., & Vysotska, V.A. (2025). Method for detecting sources of disinformation based on ensemble machine learning models. *Bionics of Intelligence*, 1(102), 11–19. DOI: 10.30837/ bi.2025.1(102).02.

О. В. Лозинська, **О. О. Марків**, **В. А. Висоцька**
Національний університет «Львівська політехніка», м. Львів, Україна

Виявлення джерел та учасників пропаганди в TikTok із використанням методів машинного навчання

У роботі представлено підхід до виявлення джерел дезінформації, фейків та пропаганди в соціальній мережі TikTok з використанням сучасних методів обробки природної мови (NLP) та штучного інтелекту. Основною метою дослідження є створення системи, здатної автоматично аналізувати коментарі до відео з пропагандистським вмістом, а також джерела їх поширення та потенційних учасників пропаганди.

У межах дослідження вручну зібрано корпус коментарів українською та російською мовами, які класифікувались як пропагандистські або нейтральні. На основі проведеного аналізу датасету, визначено певні критерії для виявлення джерел дезінформації та потенційних її учасників, зокрема через такі як використання російської мови, повторювані пропагандистські наративи, а також фейковість акаунтів. Розроблено два підходи до аналізу: класифікаційна модель на основі RandomForestClassifier та кластеризаційна модель з використанням алгоритму KMeans. Обидві моделі використовують трансформери RoBERTa для відповідних мов, а також додаткову вручну сформовану множину ознак коментарів. Побудовано Telegram-бот і графічний інтерфейс для зручного використання системи, що дозволяє отримувати коментарі з TikTok, класифікувати їх і надавати користувачеві результати аналізу. Запропонована система є актуальним інструментом для інформаційної безпеки та боротьби з пропагандою в цифровому середовищі.

У результаті дослідження було розроблено ефективну систему для автоматичного виявлення російської пропаганди в коментарях до відео в TikTok. Використання сучасних моделей трансформерів (RoBERTa), алгоритмів машинного навчання (класифікації та кластеризації), а також розробка повноцінного програмного комплексу з графічним інтерфейсом дозволили реалізувати повноцінний цикл аналізу даних - від збору до візуалізації результатів. Експерименти показали високу чутливість класифікаційної моделі до виявлення пропагандистських повідомлень, а відповідно і відслідковувати потенційних учасників дезінформацію та пропаганду. Однак через дисбаланс вибірки були виявлені обмеження у точності для нейтрального контенту. Кластеризаційна модель дозволила краще працювати з "сірими зонами" - неоднозначними коментарями. Додатково проведено аналіз параметрів, що дозволяють виявляти не лише пропаганду, а й потенційні джерела її поширення. До таких параметрів відносяться використання характерних пропагандистських наративів, мови коментарів (переважно російська), та ознаки фейковості акаунтів - однотипність дописів, низький рівень автентичності профілю тощо. Створена система може бути інтегрована в платформи моніторингу інформаційного простору, що значно посилить можливості державних та громадських інституцій у протидії гібридним загрозам в інформаційній сфері.

дезінформація, джерела пропаганди, датасет, модель RoBERTa, кластеризація, потенційні учасники пропаганди, множина критеріїв для виявлення учасників пропаганди

Одержано (Received) 11.08.2025 *Прорецензовано (Reviewed) 25.08.2025*
Прийнято до друку (Approved) 27.08.2025