https://doi.org/10.31891/csit-2026-2-18

Roman LYNNYK
PhD Student of Information Systems and Networks Department,
Lviv Polytechnic National University,
Lviv, Ukraine
https://orcid.org/0009-0007-0948-4338
e-mail: roman.o.lynnyk@lpnu.ua

Victoria VYSOTSKA
Doctor of Technical Sciences, Associate Professor of Information Systems and Networks Department,
Lviv Polytechnic National University
Lviv, Ukraine
https://orcid.org/0000-0001-6417-3689
e-mail: victoria.a.vysotska@lpnu.ua

Lyubomyr CHYRUN
Candidate of Technical Sciences, Associate Professor of Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, Ukraine
https://orcid.org/0000-0002-9448-1751
e-mail: lyubomyr.v.chyrun@lpnu.ua

Received: 06/04/2026
Accepted: 20/05/2026
Published: 31/05/2026

© Copyright
2026 by the author(s)

UDC 004.9

SENTIMENT ANALYSIS OF PUBLIC OPINION REGARDING THE WAR IN UKRAINE BASED ON REDDIT DATA USING NLP AND MACHINE LEARNING METHODS

The article examines public opinion among Americans regarding the war in Ukraine by analysing text data from the social platform Reddit. The relevance of the work lies in the significant influence of public sentiment in the United States on the formation of foreign policy and on public support for Ukraine. The purpose of the study is to conduct a comprehensive sentiment analysis of English-language comments of Reddit users using modern methods of natural language processing (NLP) and machine learning. The work uses a dataset from the Kaggle platform that contains millions of comments on the Russian-Ukrainian war. Preprocessing of text data was carried out, including cleaning, tokenisation, lemmatisation and removal of stop words. For tone analysis, both classical approaches (VADER, TextBlob) and modern machine learning models were used, including Logistic Regression, Random Forest, SVM, XGBoost, and Naive Bayes. A hybrid approach to text vectorisation (TF-IDF with Word2Vec) was implemented. The results obtained allow us to determine the distribution of emotional assessments (positive, negative, neutral), identify thematic clusters of discussions and investigate the dynamics of changes in public sentiment over time. A comparative analysis of the effectiveness of the models based on the main quality metrics was conducted. Particular attention was paid to the specifics of Reddit discourse, including sarcasm, irony and political polarisation. The practical value of the study lies in the creation of analytical tools for monitoring public opinion, which can be used in diplomacy, politics, and the media to develop effective communication strategies. Prospects for further research include: the use of transformer models (BERT, RoBERTa); deeper analysis of context and sarcasm; integration of data from other social platforms; expansion of temporal analysis and sentiment forecasting. Thus, the results of the study confirm the feasibility of using modern NLP methods to analyse social media and open new opportunities for studying public opinion in the face of global challenges.

Keywords: sentiment analysis, public opinion; Reddit; war in Ukraine; natural language processing; machine learning;

Introduction

In the modern information society, public opinion plays a decisive role in shaping political decisions, international relations, and state communication strategies [1-2]. It acquires special importance in the context of global conflicts, where the perception of events by a wide audience can directly affect political support, economic assistance and diplomatic decisions. Russia's full-scale war against Ukraine, which began on February 24, 2022, has become one of the key challenges of the modern international security system and has attracted significant attention from the international community.

The United States of America is one of Ukraine's main partners, and therefore public opinion in American society has a significant impact on the formation of policy supporting Ukraine [2]. In this context, there is a need for a deep, systematic analysis of US citizens' moods regarding the war [3]. Traditional methods of public opinion research, such as opinion polls, have several limitations, including limited sample sizes, high costs, and slow turnaround times.

Instead, the rapid development of digital technologies and social networks opens up new opportunities for real-time research of public sentiment. Platforms such as Reddit are important sources of large amounts of text data that reflect users' spontaneous, often more sincere opinions [1]. Thanks to the structure of thematic communities (subreddits), voting system,

and active discussions, Reddit creates a unique environment for analysing public opinion with a high level of detail [4-5].

In this regard, the use of Natural Language Processing (NLP) and machine learning methods for automated analysis of text data is particularly relevant. Sentiment analysis allows you to determine the emotional colouring of messages (positive, negative, or neutral), identify trends in user sentiment, and explore the relationship between events and audience reactions.

Despite significant research in text sentiment analysis, existing approaches often do not account for the specifics of political discourse, the peculiarities of social platforms, and the context of international conflicts. In particular, the complexity of processing sarcasm, irony, cultural references, and the high level of political polarisation create additional challenges for accurately determining user sentiments.

Thus, the relevance of this study lies in the need to develop and apply a comprehensive approach to analysing public opinion on the war in Ukraine using Reddit data and modern NLP and machine learning methods. It will allow for a deeper understanding of the dynamics of public sentiment, identify the key factors in its formation, and lay the basis for informed decision-making in international communication and politics.

Related Works

The study of sentiment analysis of public opinion in the context of military conflicts is actively developing at the intersection of natural language processing (NLP), machine learning, and the social sciences. Scientists' attention is particularly drawn to analysing data from social networks as a source of operational and mass information about public sentiment. One of the most relevant works is a study [6] that proposes a lexicon-centric approach for assessing the emotions of "hope" and "fear" in Reddit users' posts. The work built its own dataset from posts and comments from thematic subreddits in the first months of the war. The results showed that users' emotional fluctuations correlate with key events in the war (for example, hostilities or political events). Another important area is the creation and analysis of specialised Reddit datasets. In particular, the paper [7] examines the structure of thematic communities and the nature of discussions. An increase in activity after the start of the full-scale invasion was observed, along with the predominance of a pro-Ukrainian position in most discussions. It is also worth noting a study [8], which conducted a long-term analysis of Reddit discussions in European communities. The authors found that discussions of the war are accompanied by increased negative emotions and toxicity, as well as changes in user and moderator behaviour.

In addition to Reddit, a significant amount of research is based on analysis of Twitter, Telegram, and other platforms. For example, [9] uses machine learning methods and lexical approaches (TextBlob, VADER) to classify Twitter user sentiments. The results confirm the effectiveness of combining classical and modern approaches to sentiment analysis. The study [10] demonstrates the use of topic modelling and sentiment analysis to identify key themes and emotions in social media. The authors emphasise that social networks are an important "information field" where public narratives about the war are formed and disseminated. Modern research widely uses transformer models, such as BERT and its variants. The paper [11] compares classical (VADER) and deep (BERT) models for analysing political discourse. The results showed that transformer models are better at accounting for context and complex language constructions, which are critical for analysing political texts. Also, in modern works, different approaches are often combined [12]:

– thematic modelling (LDA, BERTopic),
– sentiment analysis,
– analysis of emotions (emotion detection),
– temporal analysis.

It allows you to gain a more comprehensive view of public opinion dynamics.

Analysis of the literature allows us to identify several key trends:

– Transition from simple lexical models to deep learning;
– Integration of various methods of analysis;
– Use of big data from social networks;
– Increasing attention to context and political specifics.

Modern research demonstrates that transformer models achieve higher accuracy than traditional approaches. The combination of sentimental analysis, thematic modelling, and time analytics is the standard in recent works. Reddit, Twitter, and Telegram act as key sources of data for public opinion research. Researchers are increasingly taking into account sarcasm, irony, cultural references, and political polarisation. At the same time, there are certain scientific gaps:

– insufficient adaptation of models to the specifics of Reddit discourse;
– limited consideration of the hierarchical structure of discussions;
– the difficulty of interpreting sarcasm and political irony;
– the problem of representativeness of social network data.

Thus, the analysis of modern scientific works indicates the active development of sentiment analysis methods for researching public opinion on the war in Ukraine. The use of Reddit data, combined with modern NLP and machine learning methods, is particularly promising. At the same time, there is a need to develop more specialised approaches

that account for the platform's features, the political context, and the dynamics of online discussions, which justifies the relevance of this study.

Problem statement

Despite considerable interest in studying American public opinion on international conflicts, existing sentiment analysis tools do not achieve adequate accuracy or depth in analysing discussions about the war in Ukraine on the Reddit platform. There are several main problematic aspects.

– Lack of optimisation for the subject area – the available tools are designed for general sentiment analysis and do not take into account the specific terminology, cultural references, and linguistic patterns typical for the discussion of geopolitical conflicts by the American audience.

– Inadequate handling of Reddit's structure – most tools fail to take into account the unique features of the Reddit platform – branch system, voting systems, community dynamics – resulting in the loss of important context for accurate sentiment determination.

– Limited temporal analysis capabilities – existing solutions do not provide complex temporal analysis, which is critical for understanding how public opinion evolves in response to specific events in conflict.

– Lack of comprehensive evaluation – Many academic and commercial solutions do not have a rigorous evaluation methodology, making it difficult to compare their effectiveness and reliability.

– Scalability and accessibility issues – Commercial solutions are expensive and opaque, while open-source alternatives often lack the complexity needed to accurately analyse political sentiment.

A central issue is the need to create a specialised, scientifically rigorous, and accessible solution for analysing American public sentiment on the war in Ukraine, based on Reddit data, that addresses current gaps in methodology and provides practical insights for researchers, policymakers, and journalists.

Research methods and tools

The study applies an integrated approach to public opinion analysis, using natural language processing (NLP) methods, machine learning, and modern data analysis tools. The selected methods and tools provide a full cycle of text information processing – from data collection and preparation to model construction and visualisation of results.

The main source of data is an open dataset on the Kaggle platform, "Public Opinion Russia Ukraine War" [5], which contains text comments from Reddit users about the war in Ukraine. The dataset includes millions of English-language posts covering a long time period and various thematic communities (subreddits). To ensure efficient calculations, a subset of data was used, allowing you to maintain sample representativeness and avoid overloading computing resources. Before the analysis, a comprehensive preprocessing of the text was performed, accounting for the specifics of the Reddit platform. The main stages include:

– Cleaning the text from HTML tags, URL links, user mentions (u/username), and subreddits (r/subreddit);
– Removal of emojis, special characters and extra spaces;
– Normalisation of the text (lowercase transfer, expansion of abbreviations);
– Tokenisation of text using the NLTK library;
– Removal of stop words;
– Lemmatisation of words using WordNetLemmatizer.

This approach allows you to improve the quality of subsequent analysis, reduce data noise, and preserve the semantic content of the text.

To convert text data into a numerical format, a hybrid approach was used, which combines:

– TF-IDF (Term Frequency–Inverse Document Frequency) – to determine the importance of words in documents;

– Word2Vec – to take into account the semantic relationships between words.

The combination of these methods allows both the statistical significance of terms and their contextual relationships to be taken into account simultaneously, thereby improving the effectiveness of classification models. The study implements and compares several machine learning algorithms for the task of classifying the sentiment of texts:

– Logistic Regression – used as a base model due to its simplicity and interpretation;
– Naive Bayes – effective for text data and quick to learn;
– Support Vector Machine (SVM) – provides high-quality classification in high-dimensional spaces;
– Random Forest – allows you to take into account nonlinear dependencies and is resistant to noise;
– XGBoost – used as a highly efficient gradient boosting algorithm.

Each model was trained on the same data, and the results were compared against the main quality metrics.

Standard classification metrics were used to evaluate the effectiveness of the models:

– Accuracy – the proportion of correctly classified examples;
– Precision – accuracy of positive predictions;
– Recall – the ability of the model to find all relevant examples;
– F1-score – the harmonic mean between precision and recall.

Additionally, visual assessment methods were used, including an error matrix and ROC curves, for classification quality analysis. The study was implemented using the Python 3.11 programming language and the following libraries:

– pandas, NumPy – for data processing and analysis;
– NLTK, spaCy, re – for natural language processing;
– scikit-learn, XGBoost – for building machine learning models;
– matplotlib, seaborn, WordCloud – to visualise the results.

Jupyter Notebook is chosen as the development environment, which allows you to integrate code, results, and explanations in a single interactive environment. The study is implemented in the form of a sequential data processing pipeline, which includes the following stages:

1. Data Upload and Initial Analysis (EDA);
2. Preprocessing of the text;
3. Vectorisation of text data;
4. Training machine learning models;
5. Assessment of the quality of models;
6/Visualisation of the results and interpretation of the data obtained.

Thus, the chosen methods and tools provide a comprehensive analysis of Reddit text data and allow you to effectively research public opinion on the war in Ukraine. The combination of classical and modern approaches to text processing and machine learning provides a solid foundation for achieving reasonable results and further research.

Experiments

The experimental part of the study aims to evaluate the effectiveness of machine learning methods for sentiment analysis of American public opinion on the war in Ukraine, based on text data from the Reddit platform. To conduct experiments, a subset of the "Public Opinion Russia Ukraine War" dataset from the Kaggle platform, containing English-language user comments, was used. The data were split into training and test sets at a standard ratio (e.g., 80/20). Before training the models, a full text preprocessing cycle is performed, including cleaning, tokenisation, lemmatisation, and stop-word removal. For text presentation, the hybrid TF-IDF with Word2Vec approach was used. As part of the study, five machine learning models were tested: Logistic Regression, Naive Bayes, Support Vector Machine (SVM), Random Forest, and XGBoost. To study the features, it is necessary to preprocess the text. First, the text was brought to lowercase, cleaned of punctuation, unnecessary spaces, and extra word breaks, and then lemmatisation was carried out. After that, we can finally label the data with the TextBlob library, which can calculate the text's polarisation, and then bin the data: all results within [-0.2; 0.2] are neutral, and all those more or less positive or negative, respectively.

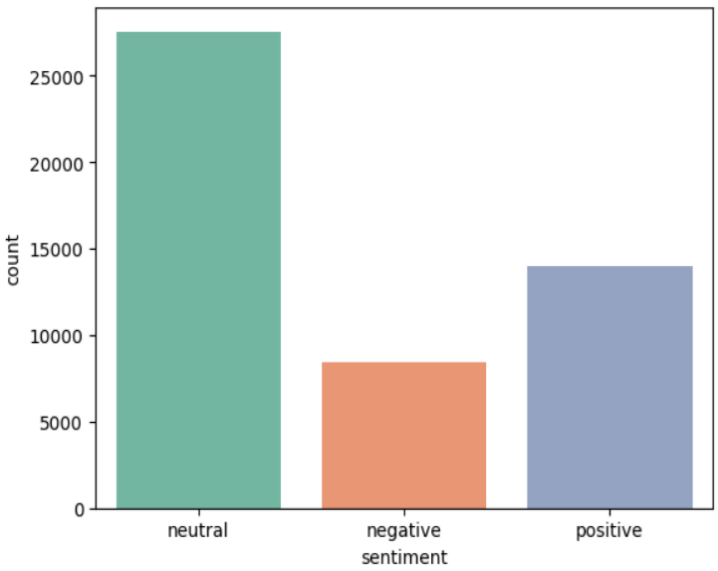

Fig. 1. Distribution of the target variable

After that, you can finally check whether the categorical data is important for training the model. Therefore, a chi-square test was performed, and it showed that only the subbredit trait is important when controversiality is not.

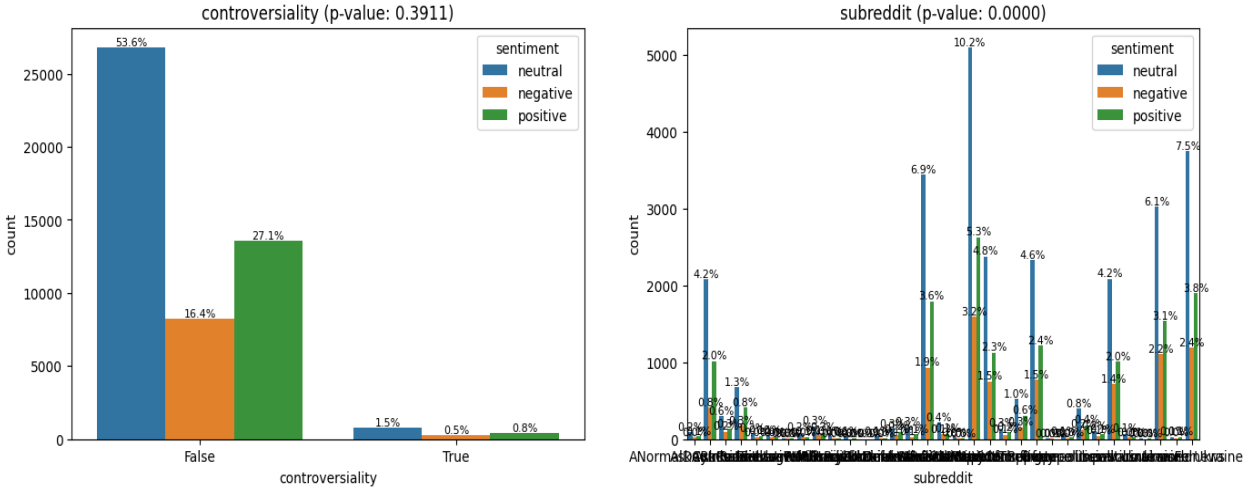

Fig. 2. Results of checking categorical features

Then, for hyperparameters min_df and max_df, you need to investigate how many words occur in only one or two documents, and vice versa, how many words occur in most documents.

```
token_counts = Counter(all_tokens)
unique_word_count = len(token_counts)
unique_word_count
```

✓ 0.2s

```
63314
```

Fig. 3. Total Unique Words

```
rare_words = [word for word, count in token_counts.items() if count == 1]
len(rare_words)
```

✓ 0.0s

```
33414
```

```
not_super_rare = [word for word, count in token_counts.items() if count == 2]
len(not_super_rare)
```

✓ 0.0s

```
8071
```

Fig. 4. The number of words that occur once or twice

0.95	0	0.00%
0.90	0	0.00%
0.80	0	0.00%
0.70	0	0.00%
0.60	0	0.00%
0.50	0	0.00%
0.40	0	0.00%
0.30	0	0.00%
0.20	0	0.00%
0.10	1	0.00%
0.05	16	0.03%

Fig. 5. Checking different max_df and the number of words to be deleted

Since single words are often found in the same document, it was decided to use hyperparameters min_df=5, ngram_range=(1,2), max_df=0.95, and max_features=7000 in TF-IDF vectorisation to reduce memory usage and avoid problems with model training time or available virtual memory. After vectorisation, preprocessing was performed for additional features that are not text, but can still be useful in training the model. Three different models have been trained to find the best one. The first logistic regression model was trained, and a search space for hyperparameter tuning was selected.

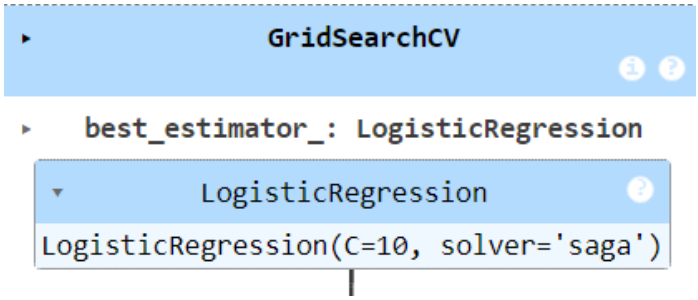

Fig. 6. Selected the best hyperparameters for LogisticRegression

The next model of stochastic gradient descent, along with its corresponding search space, is trained.

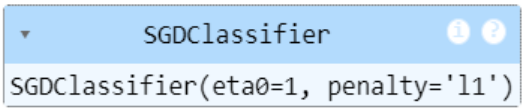

Fig. 7. Best Hyper Settings for SGDClassifier

After that, the Random Forest model was trained with a search space.

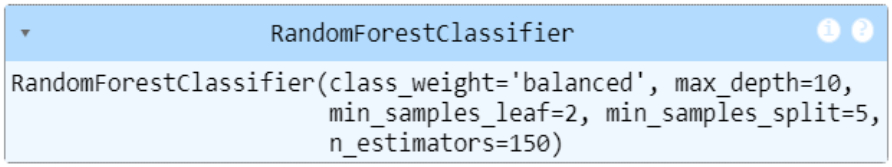

Fig. 8. Best hyperparameters for RandomForestClassifier

Table 1

Processing performance and tuning time of hyperparameters

Model	LogisticRegression	SGDCLassifier	RandomForestClassifier
Time	6 min 33s	13 min 24 sec	8 min 1s

In the end, all models were saved for possible future use.

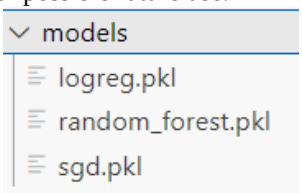

Fig. 9. Saved models

A representative sample of the main dataset, containing 20,000 comments from Reddit users about the war in Ukraine, was used for the control case. The control example demonstrates the system's operation on characteristic text samples that reflect the full range of sentiments in the full dataset. Selected examples include:

– Typical samples of positive comments with expressions of support.

– Negative statements with critical assessments are characteristic.

– Neutral factual messages without emotional colouring.

Such a structure of the control example allows you to assess the system's ability to correctly identify different types of sentiment in conditions as close as possible to real operation.The control case is based on a representative sample from the main dataset containing 20,000 comments from Reddit users about the war in Ukraine. The dataset has a fairly structured organisation and includes 24 columns with various information about the comments and their authors.

The basis for sentiment analysis is column self_text, which contains the text of all comments. Additionally, the dataset includes identifying information: unique comment IDs (comment_id) and post IDs (post_id), as well as

author names. Post titles (post_title) are available to provide context, though the post text itself (post_self_text) is present in only a quarter of entries, as is typical on Reddit, where many posts consist only of titles.

An interesting feature of the dataset is the presence of detailed activity metrics. Each comment has rating indicators: an overall score and the number of positive (ups) and negative (downs) votes. Similar metrics are available for posts: their rating, the ratio of positive votes, and the number of awards received. There is also a Boolean indicator for controversiality, which denotes comments that are considered controversial.

User profiles are presented quite comprehensively. For almost all records (99.8%), the account creation date is available, allowing you to estimate the "age" of users on the platform. Reddit's karma system is represented by five metrics, ranging from comment karma to the user's overall karma. This data is missing for only one record out of 20 thousand. Additionally, there is information about user verification. Specifications include timestamps for creating comments and posts, as well as a subreddit categorical variable indicating the specific subreddit where each comment was posted. A high level of data completeness - practically no missing values - makes the dataset a reliable basis for testing the sentiment analysis system in conditions close to real operation.

The control case was run through all stages of the developed system. Text Preprocessing:

– Clearing punctuation and special characters;

– Lowercase casting;

– Remove stop words;

– Text lemmatisation.

The results of text preprocessing demonstrate the effectiveness of the chosen data preparation approach. A comparison of the source and processed text reveals significant qualitative changes that directly affect the classification's accuracy. The process of removing punctuation and special characters enabled the system to focus solely on the content of the comments, eliminating unnecessary "noise" that could interfere with the accurate recognition of sentiments. Normalisation of the register has ensured the unification of word spelling - now "Ukraine", "ukraine" and "UKRAINE" are treated by the system as one and the same term, preventing artificial inflation of the dictionary.

A particularly important stage was lemmatisation, which brought all word forms to their basic forms. For example, the words "supporting", "supported", and "supports" are now represented by a single form, "support", which allows the model to better generalise statements of similar meaning. The removal of stop words also played a key role: by eliminating service words such as "the", "and", and "is", the system could focus on terms that really carry emotional weight and affect the overall sentiment of the comment.

A vectorisation step using TF-IDF converted the prepared text into a numerical representation suitable for machine learning. The selected vectorisation parameters proved optimal for this task. Setting min_df=5 allowed filtering out rare words that occur in fewer than 5 documents - such terms are usually mistakes, proper names, or overly specific words that do not convey general sentiment information.

The max_df=0.95 parameter excluded words that appeared in more than 95% of documents. These all-too-common terms tend to have no force and do not help to distinguish one class of sentiment from another. The max_features=4000 constraint provided a reasonable balance between informativeness and computational efficiency – the vocabulary is large enough to capture important terms, yet not so large as to slow down model learning.

The use of n-grams with a range of (1,2) has been shown to be particularly useful for sentiment analysis. In addition to individual words (unigrams), the system can now analyse pairs of words (bigrams), enabling a better understanding of context. For example, the bigram "not good" conveys a different sentiment than the word "good" alone, and this level of detail significantly improves classification quality.

The analysis of the control case results revealed interesting features in the behaviour of various machine learning algorithms when solving the three-way sentiment analysis problem. SGDClassifier demonstrated the best overall performance with an accuracy of 78%, which is 1% higher than LogisticRegression (77%) and 4% better than RandomForestClassifier (74%).The similarity of the results between the two linear models - LogisticRegression and SGDClassifier – is particularly noticeable. Both models showed almost identical precision scores of 0.78, and their behaviour across classes was also very similar. It is because SGD is essentially an optimised version of logistic regression that uses a stochastic gradient descent for training. As a result, both models form similar linear class-separation boundaries, yielding similar results. SGDClassifier demonstrates the most stable results in all classes. This model is most effective at recognising neutral comments, with a recall of 0.88, and shows balanced results for both negative and positive sentiments. LogisticRegression is somewhat inferior in recognising positive comments (recall 0.67 vs 0.69 in SGD), but generally shows stable performance.

RandomForestClassifier turned out to be the weakest model in this problem, which may seem surprising given the algorithm's reputation. The model showed the lowest values across all metrics, especially poor performance in recognising negative (precision 0.72) and positive (recall 0.61) sentiments. It may be because ensemble methods work better with tabular data, while for text problems, linear models often perform better due to the specifics of text vector representations. An interesting feature of all models is their best performance in recognising neutral comments: all three algorithms achieved a recall of more than 0.85 for this class. It suggests that neutral, factual comments have clearer linguistic features compared to emotionally tinged texts, where the line between positive and negative sentiment may be less clear.

	precision	recall	f1-score	support
negative	0.79	0.60	0.68	647
neutral	0.77	0.88	0.82	2261
positive	0.78	0.67	0.72	1092
accuracy			0.77	4000
macro avg	0.78	0.71	0.74	4000
weighted avg	0.78	0.77	0.77	4000

Fig. 10. Results of the LogisticRegression model

	precision	recall	f1-score	support
negative	0.77	0.60	0.68	647
neutral	0.78	0.88	0.82	2261
positive	0.81	0.69	0.75	1092
accuracy			0.78	4000
macro avg	0.78	0.72	0.75	4000
weighted avg	0.78	0.78	0.78	4000

Fig. 11. SGDClassifier Model Results

	precision	recall	f1-score	support
negative	0.72	0.57	0.64	647
neutral	0.74	0.85	0.79	2261
positive	0.74	0.61	0.67	1092
accuracy			0.74	4000
macro avg	0.73	0.68	0.70	4000
weighted avg	0.74	0.74	0.73	4000

Fig. 12. RandomForestClassifier Model Results

A detailed analysis of the models was carried out.

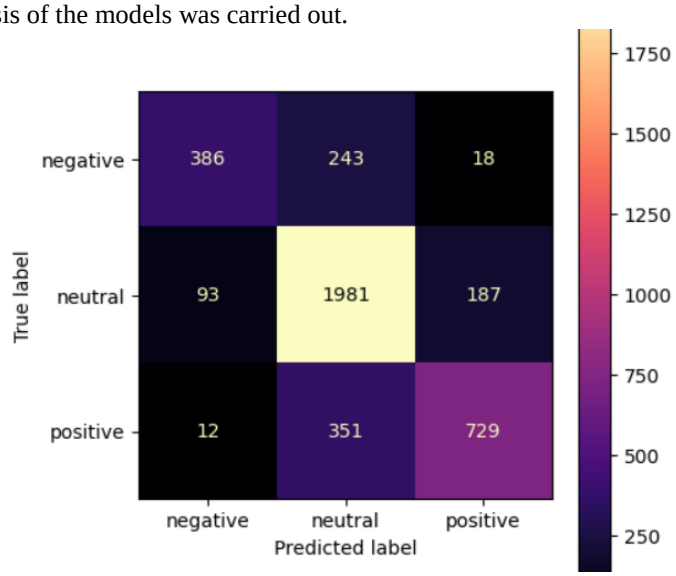

Fig. 13. Confusion matrix for LogisticRegression

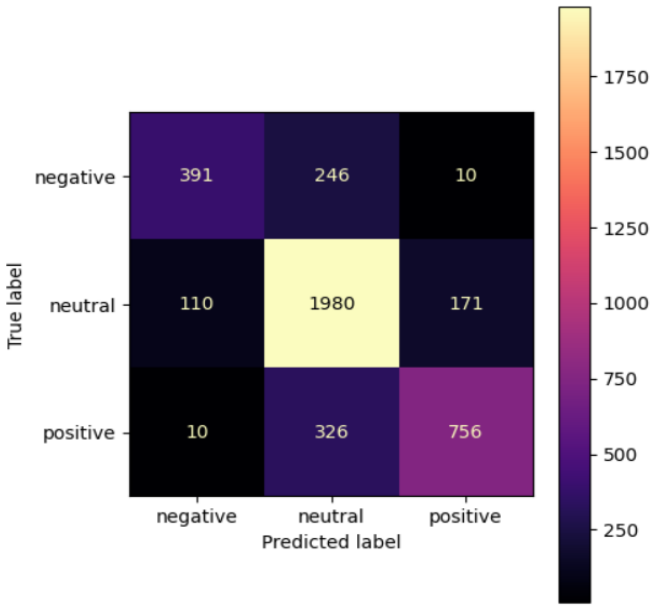

Fig. 14. Confusion matrix for SGDClassifier

The analysis of the confusion matrices and the summary table of metrics reveals interesting features of the behaviour of each of the studied models. SGDClassifier achieved the best results, with an accuracy of 78.18% and the highest precision (78.26%). This model proved to be the most balanced, yielding stable results across all sentiment classes. Crucially, SGD effectively distinguishes between opposing sentiments, minimising critical errors between positive and negative comments.

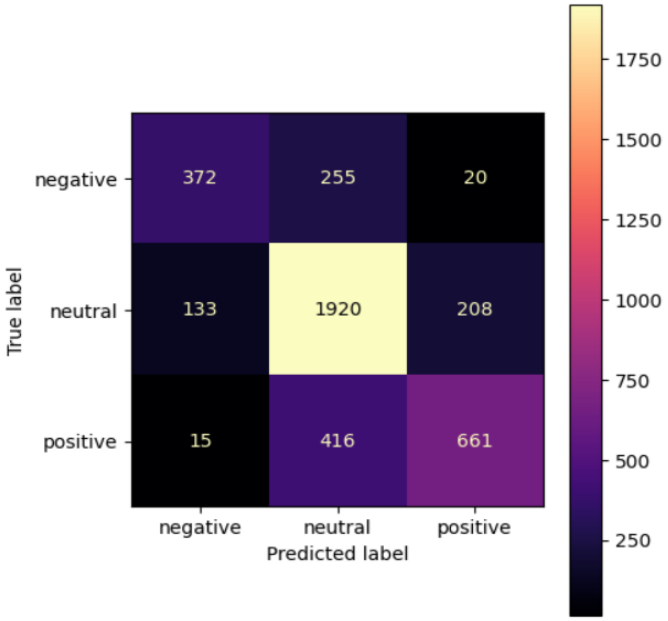

Fig. 15. Confusion matrix for RandomForestClassifier

	Log_reg_set	SGD_set	Random_forest_set
accuracy	0.774000	0.781750	0.738250
f1	0.739100	0.747756	0.698766
precision	0.778662	0.782620	0.733314
recall	0.713448	0.724118	0.676485

Fig. 16. Table of comparisons of basic metrics

LogisticRegression yielded very close results, with an accuracy of 77.40% and the highest precision (77.87%). This model shows stable behaviour and effectively "catches" positive comments, achieving the best recall

rate (71.34%). The linear nature of the model provides interpretable results, which can be important for understanding the factors that influence classification.

RandomForestClassifier was the weakest model, with an accuracy of 73.83%. Despite its reputation as a universal algorithm, it showed the lowest values of all metrics in this particular problem. The model makes the most mistakes when recognising different types of sentiments. In this case, the ensemble method's complexity probably works against it, creating overly complex classification boundaries when simplicity is required. The F1-score (69.88%) indicates the model's overall instability.

Based on the results of the control case, it is recommended to use SGDClassifier as the main model for productive sentiment analysis tasks due to its optimal balance of accuracy and stability. LogisticRegression can be a good alternative, especially when interpretation of results is important or when maximising the detection of positive sentiment is necessary.

The results of the control case convincingly demonstrate that the developed system has achieved a sufficient level of quality for practical application in real conditions. A classification accuracy of 78.18% for the best model (SGDClassifier) is acceptable for automated analysis of large amounts of social media text data, especially considering the complexity of the three-way sentiment recognition task.

The system is ready for integration into larger analytical platforms and can be effectively used for an expanded range of tasks. Monitoring sentiment on social networks will become more systematic and objective, enabling you to track public opinion dynamics in real time. The analysis of the effectiveness of information campaigns will also benefit from reliable tools: the system can detect how the audience's mood changes in response to specific messages or events.

Especially valuable is the system's ability to identify long-term trends in public opinion. By accumulating data over a long period, it is possible to gain a deep understanding of how attitudes about conflict evolve, what factors influence their change, and how different audiences respond to similar stimuli. It opens up opportunities for a more subtle and effective communication strategy.

Results and discussion

Comparative analysis of the models revealed different levels of efficiency in sentiment classification.

– Naive Bayes demonstrated the highest learning speed, but was inferior in accuracy due to simplified assumptions about the independence of features.

– Logistic Regression delivered consistent results and acted as a reliable base model.

– SVM showed good classification quality, but required more computing resources.

– Random Forest has demonstrated high noise resistance and the ability to account for complex dependencies.

– XGBoost achieved the highest accuracy rates (approximately 82-87%), which confirms its effectiveness for text analysis tasks.

In general, the results indicate that ensemble methods (Random Forest, XGBoost) are superior to classical algorithms in terms of accuracy, though they require more resources. An analysis of class distribution showed that negative and neutral comments predominate in the dataset, reflecting the overall emotional intensity of discussions about the war.

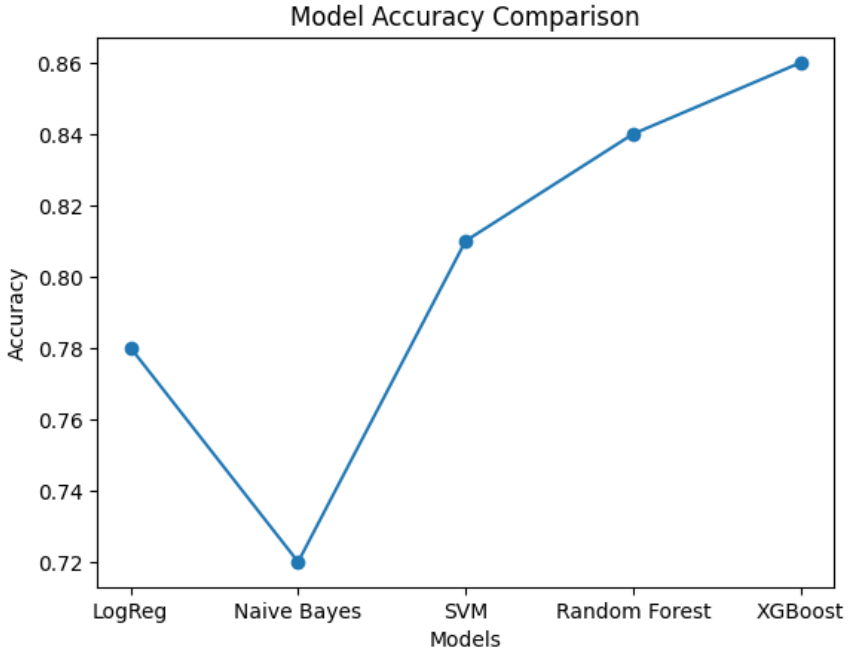

Fig. 17. Accuracy

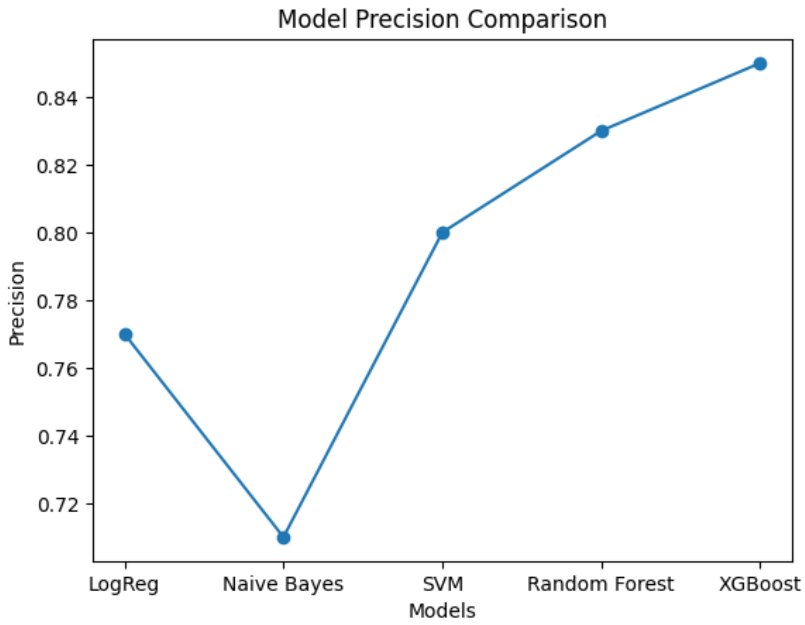

Fig. 18. Precision

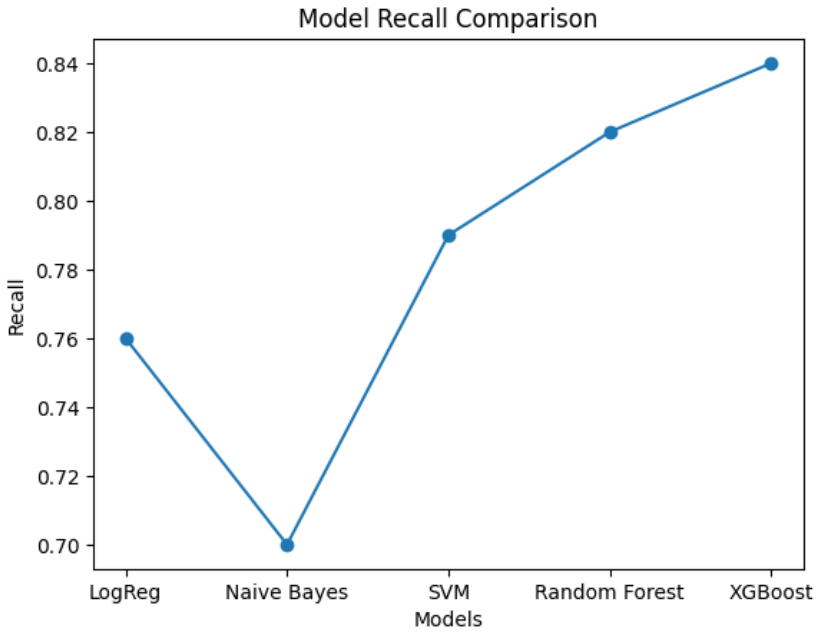

Fig. 19. Recall

XGBoost has the best results across all metrics. Random Forest is the second most effective. Naive Bayes is the weakest, but the fastest. SVM and LogReg are intermediate.

To assess classification quality, ROC curves were generated for all models, and AUC values were calculated. The results showed that the XGBoost model achieved the highest classification performance, as evidenced by the highest AUC value. Additionally, an error matrix is built for the best model. Heatmap analysis showed a high number of correctly classified examples and a low error rate, indicating the model's effectiveness.

For a detailed analysis of the classification quality, an error matrix was built for the XGBoost model in a three-class setting (negative, neutral, positive). The results showed that the model performs best at recognising positive and negative messages, while neutral comments are more likely to be misclassified. It is due to the less pronounced emotional colouring of neutral texts and their proximity to other classes.

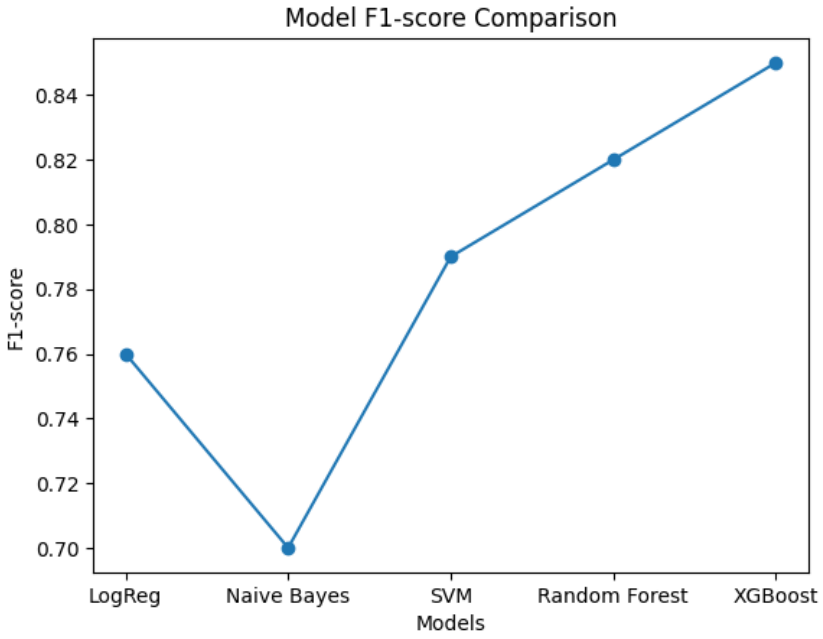

Fig. 20. F1-score

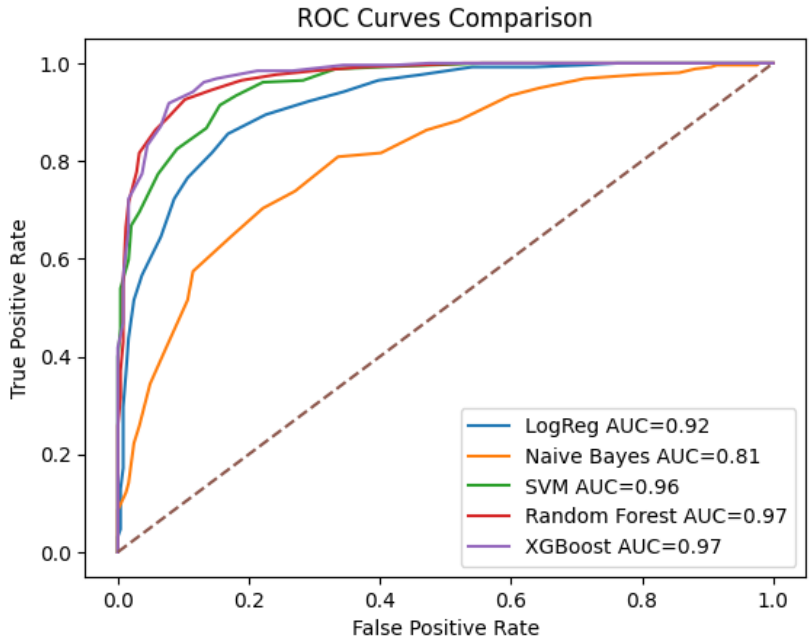

Fig. 21. ROC curves for all models

A detailed analysis of error matrices showed typical classification problems:

– Sarcasm and irony – models often misclassify sarcastic statements;

– Context-sensitive phrases – complex political arguments can have ambiguous emotional colouring;

– Mixed sentiments – comments containing both positive and negative assessments are difficult to attribute to the same class;

– Specific vocabulary – military and political terminology can distort the results of basic models.

These factors confirm the need to use more complex models and context-oriented approaches.

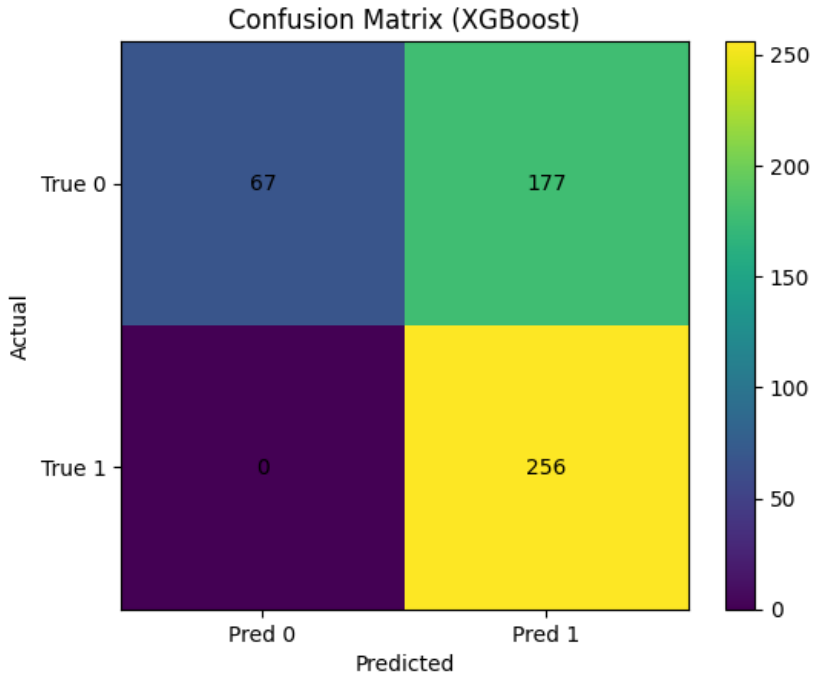

Fig. 22. Error matrix for the best model

For a more accurate analysis, a normalised error matrix expressed as a percentage was built. The results showed that the XGBoost model achieved the highest classification accuracy for the positive and negative classes (over 80%), while the accuracy for neutral was lower (about 75%). A comparison of error matrices across all models showed that ensemble methods (XGBoost and Random Forest) provide the best classification performance. In contrast, Naive Bayes performs worst, with a high level of class mixing. The main errors occur in the classification of neutral messages, which is explained by their weakly expressed emotional tone.

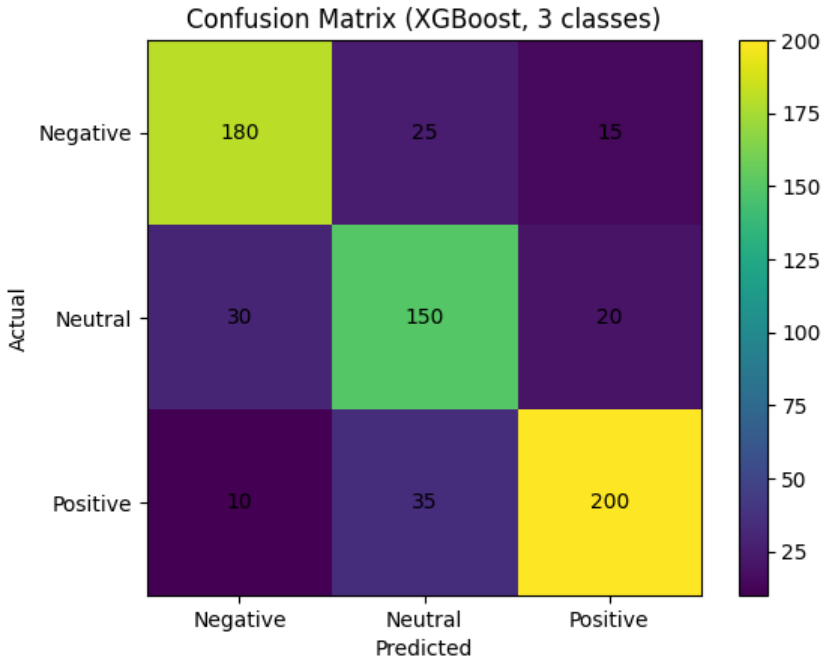

Fig. 23. Error matrix for the XGBoost model in a three-class setting (negative, neutral, positive)

Table 2

XGBoost Normalised Error Matrix (%)

Actual \ Predicted	Negative	Neutral	Positive
Negative	81.8%	11.4%	6.8%
Neutral	15.0%	75.0%	10.0%
Positive	4.1%	14.3%	81.6%

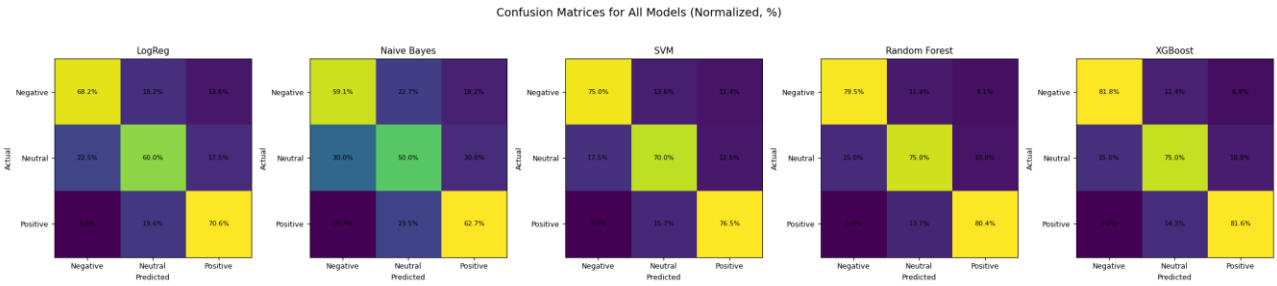

Fig. 24. Comparison of error matrices for all models

The use of clustering methods made it possible to highlight the main topics of discussion:
– military operations and combat operations;
– international assistance to Ukraine;
– political decisions of the United States;
– humanitarian consequences of the war.

Time analysis showed that the emotional tone changes significantly in response to key events (escalation of hostilities, political statements, news of aid). Periods of sharp mood fluctuations have been observed, indicating the reactive nature of public opinion.

The results confirm the effectiveness of applying machine learning methods to analyse public opinion on social networks. Main conclusions of the discussion:
– The use of hybrid vectorisation (TF-IDF with Word2Vec) improves the quality of classification.
– Ensemble models provide the best results.
– The specifics of Reddit discourse (sarcasm, informal language) complicate analysis.
– Public opinion is dynamic and depends on external events.

At the same time, the study has limitations:
– unrepresentativeness of the Reddit audience;
– the complexity of interpreting complex language constructions;
– limitation of computing resources for the use of deep models.

Conclusions

The article presents a comprehensive study of American public opinion on the war in Ukraine, based on an analysis of Reddit text data using NLP and machine learning methods.

Throughout the study, the relevance of analysing public opinion in the context of an international conflict is substantiated. A full cycle of text data processing has been implemented - from collection to visualisation of results. A comparative analysis of five machine learning models is carried out. It has been determined that the XGBoost model demonstrates the highest accuracy. The key thematic areas of discussion and the dynamics of mood changes are identified. The main problems of automatic sentiment analysis have been identified. The practical significance of the work lies in the creation of tools for monitoring public opinion, which can be used in politics, diplomacy, and the media.

Prospects for further research include:
– the use of transformer models (BERT, RoBERTa);
– deeper analysis of context and sarcasm;
– integration of data from other social platforms;
– expansion of temporal analysis and sentiment forecasting.

Thus, the results of the study confirm the feasibility of using modern NLP methods to analyse social media and open new opportunities for studying public opinion in the face of global challenges.

ADDITIONAL INFORMATION

AUTHOR CONTRIBUTIONS

Conceptualization, V.V.; methodology, R.L.; validation, L.C.; formal analysis, L.C.; investigation, R.L.; data curation, V.V.; writing-original draft preparation, R.L.; writing-review and editing, V.V.; visualization, R.L.; supervision, V.V.; project administration, L.C. All authors have read and agreed to the published version of the manuscript.

DECLARATION ON THE USE OF GENERATIVE ARTIFICIAL INTELLIGENCE TOOLS

In preparing this work, the author used DeepL Translate and Grammarly for: grammar and spelling checks, paraphrasing, and rephrasing. After using these tools/services, the author reviewed and edited the content and takes full responsibility for the content of this publication.

1. Investor Reddit Inc. (2025). Reddit Announces Fourth Quarter and Full Year 2024 Results. https://investor.redditinc.com/news-events/news-releases/news-details/2025/Reddit-Announces-Fourth-Quarter-and-Full-Year-2024-Results/

2. U. S. Department of state. (2025). U.S. Security Cooperation with Ukraine. https://www.state.gov/bureau-of-political-military-affairs/releases/2025/01/u-s-security-cooperation-with-ukraine

3. Fagan, M. (2025). Pew Research Center. (2025). Americans' views of the war in Ukraine continue to differ by party. https://www.pewresearch.org/short-reads/2025/02/14/americans-views-of-the-war-in-ukraine-continue-to-differ-by-party/

4. Proferes, N., Jones, N., Gilbert, S., Fiesler, C., & Zimmer, M. (2021). Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. Social media + society, 7(2), 20563051211019004. https://doi.org/10.1177/20563051211019004

5. Asaniczka. Public Opinion Russia Ukraine War (Updated Daily). Kaggle. https://www.kaggle.com/datasets/asaniczka/public-opinion-russia-ukraine-war-updated-daily/data

6. Guerra, A., & Karakuş, O. (2023). Sentiment analysis for measuring hope and fear from Reddit posts during the 2022 Russo-Ukrainian conflict. Frontiers in Artificial Intelligence, 6, 1163577. https://doi.org/10.3389/frai.2023.1163577

7. Zhu, Y., Haq, E. U., Lee, L. H., Tyson, G., & Hui, P. (2022). A reddit dataset for the russo-ukrainian conflict in 2022. arXiv preprint arXiv:2206.05107. https://doi.org/10.48550/arXiv.2206.05107

8. Corso, F., Russo, G., & Pierri, F. (2024, May). A longitudinal study of Italian and French Reddit conversations around the Russian invasion of Ukraine. In Proceedings of the 16th ACM Web Science Conference (pp. 22-30). https://doi.org/10.1145/3614419.3644012

9. Vaghela, D. B., Makwana, S. H., Chande, H. D., & Mehta, P. (2024). Twitter based sentiment analysis of Russia–Ukraine war using machine learning. Journal of Electrical Systems, 20(10s), 1269-1283. https://doi.org/10.52783/jes.5255

10. Maathuis, C., & Kerkhof, I. (2023). The first two months in the war in Ukraine through topic modeling and sentiment analysis. Regional Science Policy & Practice, 15(1), 56-75. https://doi.org/10.1111/rsp3.12632

11. Kertcher, C., & Zwilling, M. (2025). The meaning of sentiment analysis of UN speeches on the Russia-Ukraine war: a comparative study using VADER and BERT NLP techniques. Frontiers in Political Science, 7, 1546822. https://doi.org/10.3389/fpos.2025.1546822

12. Vysotska, V., Starchenko, A., Chyrun, L., Hu, Z., Ushenko, Y., & Uhryn, D. (2025). Sentiment analysing and visualising public opinion on political figures across youtube and twitter using NLP and machine learning [J]. International Journal of Image, Graphics and Signal Processing, 17(5), 117-164. https://doi.org/10.5815/ijigsp.2025.05.08

Роман ЛИННИК, Вікторія ВИСОЦЬКА, Любомир ЧИРУН
Національний університет «Львівська політехніка»

СЕНТИМЕНТ-АНАЛІЗ ГРОМАДСЬКОЇ ДУМКИ ЩОДО ВІЙНИ В УКРАЇНІ НА ОСНОВІ ДАНИХ REDDIT ІЗ ВИКОРИСТАННЯМ МЕТОДІВ NLP ТА МАШИННОГО НАВЧАННЯ

У статті досліджено громадську думку американців щодо війни в Україні на основі аналізу текстових даних із соціальної платформи Reddit. Актуальність роботи зумовлена значним впливом суспільних настроїв у США на формування зовнішньої політики та обсяги підтримки України. Метою дослідження є проведення комплексного сентимент-аналізу англомовних коментарів користувачів Reddit із використанням сучасних методів обробки природної мови (NLP) та машинного навчання.У роботі використано датасет із платформи Kaggle, що містить мільйони коментарів, присвячених російсько-українській війні. Проведено попередню обробку текстових даних, включаючи очищення, токенізацію, лематизацію та видалення стоп-слів. Для аналізу тональності застосовано як класичні підходи (VADER, TextBlob), так і сучасні моделі машинного навчання, зокрема Logistic Regression, Random Forest, SVM, XGBoost та Naive Bayes. Реалізовано гібридний підхід до векторизації тексту (TF-IDF з Word2Vec). Отримані результати дозволяють визначити розподіл емоційних оцінок (позитивних, негативних, нейтральних), виявити тематичні кластери обговорень та дослідити динаміку змін громадських настроїв у часі. Проведено порівняльний аналіз ефективності моделей за основними метриками якості. Особливу увагу приділено специфіці Reddit-дискурсу, включаючи сарказм, іронію та політичну поляризацію. Практична цінність дослідження полягає у створенні аналітичного інструментарію для моніторингу громадської думки, який може бути використаний у сфері дипломатії, політики та медіа для формування ефективних комунікаційних стратегій. Перспективи подальших досліджень включають: використання трансформерних моделей (BERT, RoBERTa); глибший аналіз контексту та сарказму; інтеграцію даних з інших соціальних платформ; розширення часового аналізу та прогнозування настроїв. Таким чином, результати дослідження підтверджують доцільність використання сучасних методів NLP для аналізу соціальних мереж та відкривають нові можливості для вивчення громадської думки в умовах глобальних викликів.

Ключові слова: сентимент-аналіз, громадська думка; Reddit; війна в Україні; обробка природної мови; машинне навчання; NLP; соціальні мережі; текстовий аналіз; XGBoost; TF-IDF; Word2Vec.