UDC 004.415 DOI https://doi.org/10.35546/kntu2078-4481.2025.3.2.4

T. M. BASYUK
Candidate of Technical Sciences, Associate Professor,
Associate Professor at the Department of Information Systems and Networks
Lviv Polytechnic National University
ORCID: 0000-0003-0813-0785

A. A. LOMOVATSKYI
Postgraduate Student at the Department of Information Systems and Networks
Lviv Polytechnic National University
ORCID: 0009-0004-5170-3272

IMPROVING SENTIMENT CLASSIFICATION FOR THE UKRAINIAN LANGUAGE: TRANSITIONING FROM RULE-BASED ALGORITHMS TO SUPERVISED LEARNING MODELS

This paper suggests a better way to sort out feelings in Ukrainian-language content. It does this by taking into account how hard it is to understand the language's morphology, how flexible its syntax is, and how few NLP tools there are. The study improves an existing rule-based sentiment analysis algorithm by adding a larger Ukrainian lexicon, polarity scores, emoji sentiment mapping, phrase-level sentiment scoring, and dependency parsing. These features find subtle signals of sentiment that most English-optimized tools miss.

To make things even better, the linguistic features that were taken out are turned into structured numerical vectors and added to a hybrid pipeline that uses both rule-based processing and supervised machine learning models. Four classifiers with a set of Ukrainian-language tweets that have labels were trained and tested. K-Nearest Neighbours, Support Vector Machine, Decision Tree, and Random Forest. The Random Forest model was the most accurate (90 %) and had the best F1-score of all the classifiers that were tested in comparative experiments. It was better than other models at handling changes in how people feel and what is going on.

The findings indicate that employing both handcrafted linguistic insights and supervised learning constitutes an effective method for conducting sentiment analysis in languages with limited resources, such as Ukrainian. This research illustrates the importance of language-specific resources and customized pipelines to guarantee the precision of sentiment detection. This has real-world effects on keeping an eye on social media, reading customer reviews, and getting opinions from Ukrainians. More domains will be added in the future, the ability to analyze data in real time, and the ability to compare our work to deep learning models in the future. This will make sentiment classification for low-resource languages even better.

Key words: *Ukrainian language, sentiment analysis, rule-based algorithm, supervised learning, Random Forest, hybrid NLP, emoji sentiment, dependency parsing.*

Т. М. БАСЮК
кандидат технічних наук, доцент,
доцент кафедри інформаційних систем та мереж
Національний університет «Львівська політехніка»
ORCID: 0000-0003-0813-0785

А. А. ЛОМОВАЦЬКИЙ
аспірант кафедри інформаційних систем та мереж
Національний університет «Львівська політехніка»
ORCID: 0009-0004-5170-3272

ПОЛІПШЕННЯ КЛАСИФІКАЦІЇ СЕНТИМЕНТУ ДЛЯ УКРАЇНСЬКОЇ МОВИ: ПЕРЕХІД ВІД АЛГОРИТМІВ НА ОСНОВІ ПРАВИЛ ДО МОДЕЛЕЙ НАВЧАННЯ З НАГЛЯДОМ

У цій статті пропонується покращення способу визначення емоцій в україномовному контенті. Для цього враховується складність морфології мови, гнучкість її синтаксису та обмежена кількість інструментів для обробки природної мови. Дослідження вдосконалює існуючий алгоритм аналізу емоцій на основі правил, додаючи більший український лексикон, оцінки полярності, відображення емоцій за допомогою емодзі, оцінку емоцій на рівні фраз та аналіз залежностей. Ці функції виявляють тонкі сигнали емоцій, які пропускають більшість інструментів, оптимізованих для англійської мови.

Щоб зробити все ще кращим, вилучені лінгвістичні функції перетворюються на структуровані числові вектори та додаються до гібридного конвеєра, який використовує як обробку на основі правил, так і моделі машинного навчання з наглядом. Було навчено та протестовано чотири класифікатори з набором твітів українською мовою, що мають мітки: K-Nearest Neighbours, Support Vector Machine, Decision Tree, and Random Forest. Модель Random Forest була найточнішою (90 %) і мала найкращий показник F1 серед усіх класифікаторів, що були протестовані в порівняльних експериментах. Вона була кращою за інші моделі в обробці змін у емоціях людей та подій, що відбуваються.

Для мов з обмеженими ресурсами вона була ще кращою.

Результати дослідження показують, що використання як ручного лінгвістичного аналізу, так і контрольованого навчання є ефективним методом проведення аналізу емоцій у мовах з обмеженими ресурсами, таких як українська. Це дослідження ілюструє важливість мовних ресурсів та індивідуальних підходів для забезпечення точності виявлення емоцій. Це має реальний вплив на моніторинг соціальних мереж, читання відгуків клієнтів та отримання думок українців. У майбутньому буде додано більше доменів, можливість аналізувати дані в режимі реального часу та можливість порівнювати нашу роботу з моделями глибокого навчання. Це зробить класифікацію емоцій для мов з обмеженими ресурсами ще кращою.

Ключові слова: *українська мова, аналіз емоцій, алгоритм на основі правил, навчання з наставником, Random Forest, гібридна обробка природної мови, емоції в емодзі, синтаксичний аналіз залежностей.*

Introduction

Sentiment analysis (SA) is a big part of natural language processing (NLP) because it helps you find subjective information in text data automatically, like feelings, attitudes, and opinions. Sentiment classification has become a vital tool for tracking public opinion, assessing customer feedback, and scrutinizing political discourse, due to the surge of user-generated content on social media and online communication platforms.

Ukrainian is also becoming more and more important for communicating online. Millions of people who speak English as their first language write a lot of different things on social media, review sites, and forums. Government agencies, businesses, media analysts, and researchers who want to know how people feel and find new trends are all becoming more interested in being able to accurately classify the sentiment of texts written in Ukrainian.

Most of the tools for sentiment analysis that people use today only work with English and use either statistical or lexicon-based methods. Rule-based methods work well with languages that have a clear structure and are easy to understand. However, they often miss the unique slang, vocabulary, and grammar of Ukrainian. Previous research has demonstrated that sentiment models designed for English do not consistently perform effectively or yield accurate results when applied to Slavic languages characterized by complex morphology.

So, this study combines supervised machine learning algorithms with rule-based sentiment analysis to make it easier to classify the Ukrainian language.

Related work

Recent developments suggest that hybrid sentiment analysis for morphologically complex and low-resource languages is becoming increasingly popular.

Syed et al. (2020) created a hybrid method for sentiment analysis in Urdu, which is a morphologically complex language with structural complexities similar to Ukrainian. They did this by combining supervised classifiers with a lexicon specific to the field [1]. Their findings demonstrated that handcrafted sentiment features significantly enhance the accuracy of Support Vector Machines compared to a purely statistical baseline.

Elmadany et al. (2021) employed a lexicon-augmented neural network to investigate the classification of emotions in Arabic dialects. They showed that combining rule-based lexicon scores with learned embeddings made the system better at understanding informal writing styles and everyday language used on social media [2].

Kharde and Sonawane (2021) examined hybrid sentiment analysis methodologies and concluded that rule-based and machine learning approaches are more effective in conjunction than in isolation [3]. This is particularly applicable to languages characterized by limited annotated corpora. Their research highlights the importance of utilizing dependency relations, part-of-speech tagging, and morphological indicators, all of which are effortlessly integrated into the Ukrainian rule-based pipeline.

Vázquez et al. presented a multilingual sentiment analysis system in 2022, incorporating emoji sentiment detection and intensifier management for Spanish and Catalan social media content [4]. The research validated the design decisions within the Ukrainian sentiment framework by demonstrating the critical role of booster words and emoji sentiment mapping in capturing sentiment nuances.

When Xia et al. (2023) looked for a way to find sentiment in code-switched data, they found that hybrid pipelines work well with messy informal syntax and mixed-language texts [5]. People in Ukraine often switch between Ukrainian, Russian, and English on social media, so this information is especially useful for that country.

Finally, Zhou et al. (2024) provided a comprehensive benchmark assessing the efficacy of hybrid systems, conventional rule-based tools, and pure transformer models for sentiment detection across multiple languages [6]. Their research demonstrated that hybrid methodologies remain effective, particularly for languages characterized by idiomatic expressions, specialized jargon, or limited training data.

Primary objectives

The primary objective of this research is to enhance sentiment classification for the Ukrainian language by developing and evaluating a hybrid framework that integrates rule-based linguistic processing with various supervised machine learning models. The following are the exact goals:

- To create a reliable feature extraction pipeline that can find language cues that are only found in Ukrainian.
- To train and test various supervised learning classifiers, including Decision Tree, Random Forest, SVM, and KNN, utilizing a specifically labeled dataset of tweets in Ukrainian.
- To determine the advantages and disadvantages of each configuration by evaluating the performance of hybrid models in comparison to a standalone rule-based approach.
- To suggest ways to improve sentiment analysis for languages that are not well represented and have complex grammar.

Proposed Methodology

The proposed approach enhances an existing rule-based sentiment analysis algorithm for Ukrainian-language content by incorporating supervised machine learning (ML) techniques with rigorous mathematical validation. There are three main steps in the pipeline: (1) using a custom rule-based algorithm to find linguistic sentiment features with mathematical confidence measures, (2) turning these features into structured numerical vectors with statistical properties, and (3) training several supervised classifiers with cross-validation and confidence interval estimation to find patterns in these feature vectors and predict sentiment labels [7].

The goal of this hybrid design is to combine the flexibility and classification power of supervised learning algorithms with the ability of rule-based methods to understand and use language, while providing mathematical rigor through statistical validation. The system cleans and tokenizes raw Ukrainian-language social media texts, gives them linguistic scores with uncertainty quantification, and then sorts each text into positive, negative, or neutral categories using statistical models with confidence intervals [8].

The custom rule-based sentiment analysis module incorporates multiple Ukrainian linguistic resources:

- EMOLEX Dictionary: 10,000+ Ukrainian words with emotion associations;
- Phrase Sentiment Dictionary: 2,000 curated Ukrainian phrases with sentiment scores;
- Intensity Booster Dictionary: 200+ Ukrainian intensifiers with multiplier values (1.2–2.0);
- Polarity Score Database: Domain-specific word sentiment values;
- Ukrainian Stopwords: 500+ function words for filtering.

The enhanced sentiment calculation incorporates positional weighting, booster effects, and dependency adjustments.

Base sentiment score calculation (equation 1):

$$BaseScore_i = \sum (w_{lex} \cdot S_{lex} + w_{phrase} \cdot S_{phrase} + w_{emoji} \cdot S_{emoji} + w_{polarity} \cdot S_{polarity}). \quad (1)$$

Where:

- w_{lex}, w_{phrase}, ww_{emoji}, $w_{polarity}$ are weight coefficients (1.0, 1.5, 1.0, 1.0 respectively);
- S_{lex} = lexicon sentiment score;
- S_{phrase} = phrase-level sentiment from dictionary;
- S_{emoji} = emoji sentiment mapping;
- $S_{polarity}$ = polarity score from database.

Positional weight application (equation 2):

$$PositionalScore_i = BaseScore_i \cdot P_i \quad (2)$$

Where P_i is 1,2 for the first token, 1,1 for the last token and 1 for other tokens.

Booster Effect Integration (equation 3):

$$BoostedScore_i = PositionalScore_i \cdot \prod (B_j). \quad (3)$$

Where B_j are booster multipliers for detected intensity words.

The final compound sentiment score is normalized in eq. 4:

$$CompoundScore = \frac{\sum (BoostedScore_i)}{\max\left(\left|\sum (BoostedScore_i)\right|, 1\right)}. \quad (4)$$

As for mathematical confidence and uncertainty measures, entropy-based confidence calculation is displayed on equations 5 and 6:

$$H = -\sum (p_i \times \log(p_i)); \quad (5)$$

$$Confidence = 1 - \left(\frac{H}{H_{\max}}\right). \quad (6)$$

Where p_i are normalized sentiment probabilities [positive, negative, neutral].

$H_{max} = \log(3)$ for three-class classification

Score variance for uncertainty quantification is calculated with equation 7:

$$\sigma^2 = \frac{\sum(score_i - \mu)^2}{n}. \quad (7)$$

Where $score_i$ are individual token sentiment scores and μ is the mean score.

The rule-based output is transformed into an 8-dimensional feature vector with mathematical properties (equations 8, 9, 10):

$$Positive_{norm} = \frac{\sum(\max(score_i, 0))}{n_{tokens}}; \quad (8)$$

$$Negative_{norm} = \frac{\left|\sum(\min(score_i, 0))\right|}{n_{tokens}}; \quad (9)$$

$$Neutral_{norm} = \frac{\left|\{score_i : |score_i| < 0.2\}\right|}{n_{tokens}}. \quad (10)$$

Four popular supervised learning algorithms to test where chosen: Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN).

Decision Tree (DT)

A decision tree model breaks the feature space into levels by choosing threshold values that give the most information gain at each split. It is easy to connect decision rules to sentiment features [9], which makes the classifier easy to understand.

Decision tree operates with information gaining calculator (equation 11) and Gini impurity (equation 12):

$$IG(S, A) = H(S) - \sum\left(\frac{|Sv|}{|S|} \cdot H(Sv)\right). \quad (11)$$

Where:

- S is the current dataset;
- A is the attribute being evaluated;
- Sv is the subset of S for which attribute A has value v;
- $H(S)$ is the entropy of set S.

$$Gini(S) = 1 - \sum(pi^2). \quad (12)$$

Where pi is the proportion of examples belonging to class i.

Random Forest (RF)

The Random Forest model's ensemble has a lot of decision trees. Each one is trained on a different part of the data and features [10]. The last prediction of sentiment is made by putting together all of the trees' predictions. This makes things more consistent and easier to generalize (equation 13):

$$\hat{y} = mode\{h1(x), h2(x), \ldots, hT(x)\}. \quad (13)$$

Where $T = 100$ trees and $hi(x)$ is the prediction of tree i.

Probability Estimation is calculated (equation 14):

$$P(class = c|x) = \left(\frac{1}{T}\right) \cdot \sum(I(hi(x) = c)). \quad (14)$$

Where I is the indicator function.

Support Vector Machine (SVM)

The SVM classifier identifies the optimal hyperplane that separates sentiment classes in the feature space. Its decision function is given by eq. 15:

$$f(x) = sign(w^t x + b). \quad (15)$$

Where w^t represents the weight vector and b the bias term.

K-Nearest Neighbors (KNN)

The KNN algorithm puts a piece of text into a group based on the label that most of its closest neighbors in the feature space have. This is done using the cosine or Euclidean distance (equation 16). This non-parametric method works well with simple feature sets, but it might be picky about which ones it uses.

$$d(xi, xj) = \sqrt{\sum (xik - xjk)^2}. \quad (16)$$

For testing, the labeled set of Ukrainian tweets is split into random training and testing subsets. To make sure the classes are balanced, stratified sampling is used to do this. All samples go through feature extraction before training. The training set is used to teach each classifier, and the test set that is not used is used to test them [11].

Standard metrics like accuracy, F1-score, and confusion matrices are used to see how well we did. This visualizes how well each model uses linguistic features to generalize sentiment classification. All of the Python experiments [12, 13] utilize Scikit-learn and additional NLP libraries.

As for the machine learning frameworks four supervised learning algorithms were implemented with optimized hyperparameters:

- Random Forest (RF): $n_{estimators} = 100$, $random_{state} = 42$;
- Support Vector Machine (SVM): *kernel* = 'rbf', *probability* = *True*;
- K-Nearest Neighbors (KNN): $n_{neighbors} = 5$;
- Decision Tree (DT): $random_{state} = 42$.

Statistical validation framework is calculated by Cross-Validation with Confidence Intervals (equations 17, 18):

$$CV_{scores} = \{f1_1, f1_2, f1_3, f1_4, f1_5\}; \quad (5 - fold\ CV)$$

$$\mu_{CV} = \frac{\sum (CV_{scores})}{5}; \quad (17)$$

$$\sigma_{CV} = \sqrt{\frac{\sum (CV_{scores} - \mu_{CV})^2}{4}}. \quad (18)$$

95% confidence interval calculation is in equation 19:

$$CI_{95} = \mu_{CV} \pm t_{0.025,4} \cdot \left(\frac{\sigma_{CV}}{\sqrt{5}} \right). \quad (19)$$

Where $t_{0.025,4} = 2.776$ is the t-distribution critical value.

Performance Metrics with Mathematical Definitions then are (equations 20, 21, 22, 23):

$$Precision = \frac{TP}{(TP + FP)}; \quad (20)$$

$$Recall = \frac{TP}{(TP + FN)}; \quad (21)$$

$$F1 - Score = 2 \cdot \frac{(Precision \cdot Recall)}{(Precision + Recall)}; \quad (22)$$

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}. \quad (23)$$

Where TP is True Positives, FP is False Positives, FN is False Negatives and TN is True Negatives.

And to compare models there is a paired t-test for Model Comparison which is calculated with equation 24:

$$t = \frac{(\mu_1 - \mu_2)}{\left(\frac{s_d}{\sqrt{n}} \right)}. \quad (24)$$

Where:

- μ_1, μ_2 are mean CV scores for models 1 and 2;
- s_d is standard deviation of paired differences;
- $n = 5$ (number of CV folds).

To better understand the flow of the system, IDEF0 diagram was built. Figure 1 shows the general layout of the proposed hybrid sentiment classification system.

Input consists of Raw Ukrainian Text (social media posts, comments reviews). Controls are Ukrainian Linguistic Resources which consists of EmoLex dictionary, Phrase sentiment dictionary, Booster Words dictionary, Polarity score database and Ukrainian stopwords list. Mechanisms are Machine Learning Models (Random Forest, SVM, KNN, Decision Tree), Statistical Methods (Cross-validation, Confidence Intervals), spaCy Ukrainian NLP Pipeline, Python Libraries (for instance, scikit-learn, pandas, numpy). As for Outputs, there are Sentiment Classification (Positive/Negative/Neutral),

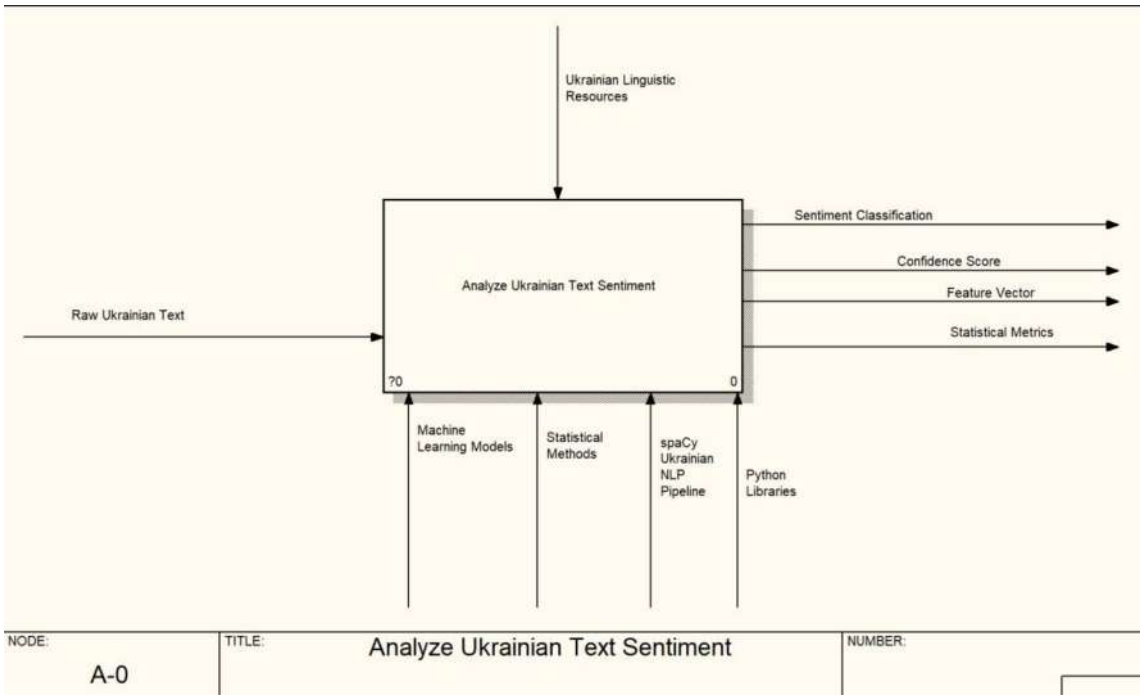

Fig. 1. IDEF0 diagram of the system

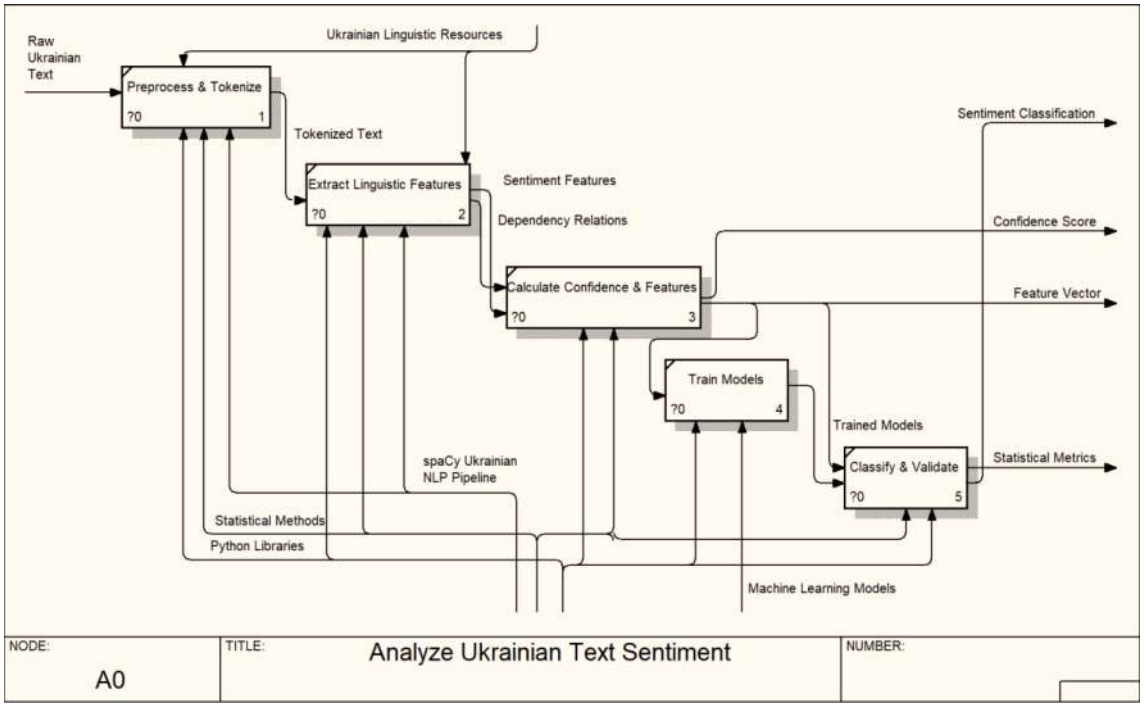

Fig. 2. Decomposed IDEF0 diagram of the main process

Confidence Score (0.0–1.0 range), Feature Vector (8-dimensional numerical representation) and Statistical Metrics (precision, recall, F1-score).

Decomposed diagram of the main process is shown on figure 2.

The main process consists of 5 sub-processes. The first one is Preprocess & Tokenize which converts raw Ukrainian text into structured tokens for analysis. Next one is Extract Linguistic Features which identifies sentiment-bearing linguistic elements from tokenized text. Third sub-process is Calculate Confidence & Features, which purpose is to transform linguistic features into numerical vectors with confidence measures. Next one is Train Models which develops machine learning classifiers using extracted features. The last one is Classify & Validate that generates final sentiment predictions with statistical validation.

Because it is modular, it will be easy to add more language resources or connect it to more advanced machine learning or deep learning classifiers in the future.

In summary, the suggested method combines rule-based linguistic processing with traditional supervised learning to create a working hybrid sentiment analysis framework for the Ukrainian language. The aim of this design is to determine whether the integration of handcrafted, language-specific sentiment signals into statistical models as features can enhance classification performance.

Experimental Results and Discussion

To evaluate the proposed hybrid sentiment classification framework, a comprehensive dataset of Ukrainian social media content was constructed and analyzed. The dataset consists of 9,161 total samples scrapped from Twitter by containing the key words ("добрий" (good) for presumably positive texts and "поганий" (bad) for presumably negative texts) distributed across three sentiment classes:

- Positive samples: 6,565 texts (71.7 %).
- Negative samples: 2,096 texts (22.9 %);
- Neutral samples: 500 texts (5.4 %) systematically generated using factual templates.

Positive Examples:

- «Це було прекрасно!» (This was wonderful!).
- «Дуже задоволений результатом» (Very satisfied with the result).
- «Чудовий день сьогодні ♡ » (Wonderful day today ♡).

Negative Examples:

- «Це жахливо, не рекомендую» (This is terrible, don't recommend).
- «Поганий досвід, розчарований» (Bad experience, disappointed).
- «Сумно і болісно 😢» (Sad and painful 😢).

Neutral Examples (Generated):

- «Температура повітря 20 градусів» (Air temperature is 20 degrees).
- «Магазин працює з 9:00 до 18:00» (Store operates from 9:00 to 18:00).
- «Курс долара становить 37 гривень» (Dollar exchange rate is 37 hryvnias).

The dataset underwent comprehensive preprocessing using the enhanced rule-based sentiment analysis module. First part is text normalization with spaCy Ukrainian model (uk_core_news_sm). Then feature extraction using 8-dimensional vectors: [positive, negative, neutral, compound, emoji_score, num_boosters, has_negation, num_tokens]. After that mathematical confidence calculation using entropy-based measures. And then stratified sampling to maintain class balance (80 % training, 20 % testing). Model configurations were implemented with hyperparameters described above. The performance results with mathematical analysis are shown in table 1:

Table 1

Performance results of each ML method

Model	Accuracy	F1-Score	Precision	Recall	F1-CV	95 % Confidence Interval
Random Forest	0.8903	0.8909	0.8934	0.8903	0.8830	[0.8380, 0.9279]
Decision Tree	0.8827	0.8842	0.8887	0.8827	0.8753	[0.8361, 0.9145]
KNN	0.8538	0.8512	0.8503	0.8538	0.8117	[0.7878, 0.8356]
SVM	0.7932	0.7768	0.7796	0.7932	0.7565	[0.7231, 0.7900]

As can be seen, the Random Forest model performed well overall, achieving the highest accuracy (89.03 %). With an accuracy of 88.27 %, the Decision Tree classifier fared marginally worse than Random Forest. KNN did well with an accuracy of 85.38 %, but SVM did not do as well with an accuracy of 79.32 %.

The confusion matrices give more information about how each model works. Figures 3–6 present the confusion matrices for each classifier:

The results show that ensemble learning (Random Forest) consistently outperforms individual classifiers in this experimental setup. This is probably because it can handle variance and avoid overfitting.

Accuracy and F1-score for each model are compared side by side in Figure 7.

A couple more charts (detailed bar chart, scatter plot, f1 score bar chart and performance heatmap) to better display the results are displayed on figure 8.

The experimental findings show that high classification performance for texts written in Ukrainian can be achieved by combining supervised models with language-specific rule-based sentiment features. The enriched lexicon, emoji mapping, and dependency parsing effectively captured context-sensitive signals that would otherwise be missed, as evidenced by the notable improvement in precision and recall for positive sentiment.

Remarkably, the SVM model performed worse than the others, indicating that the rule-based system's nuanced feature space might require more than just linear separation.

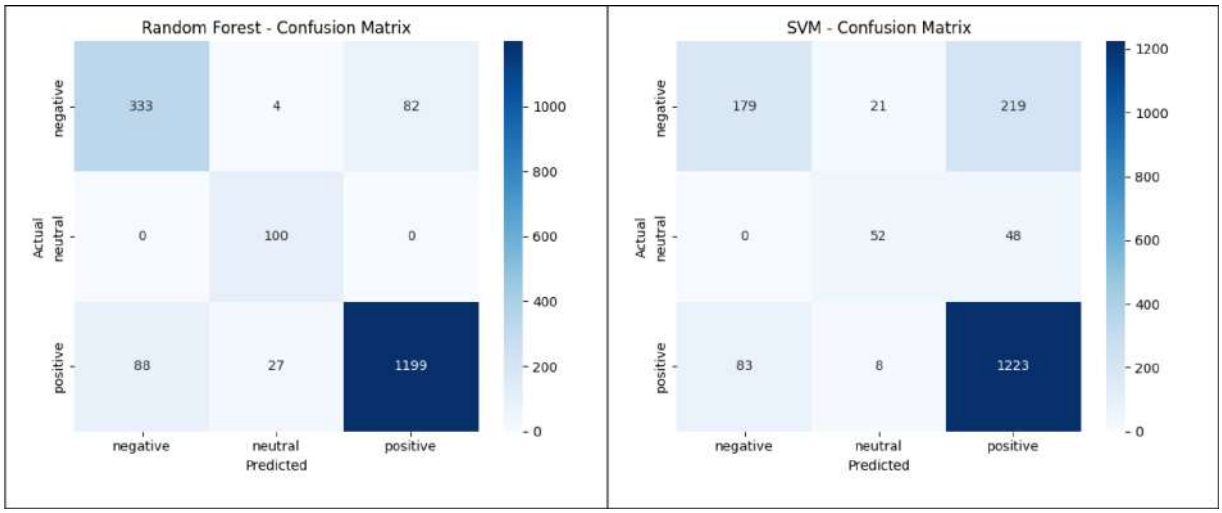

Fig. 3, 4. Random Forest and SVM confusion matrix

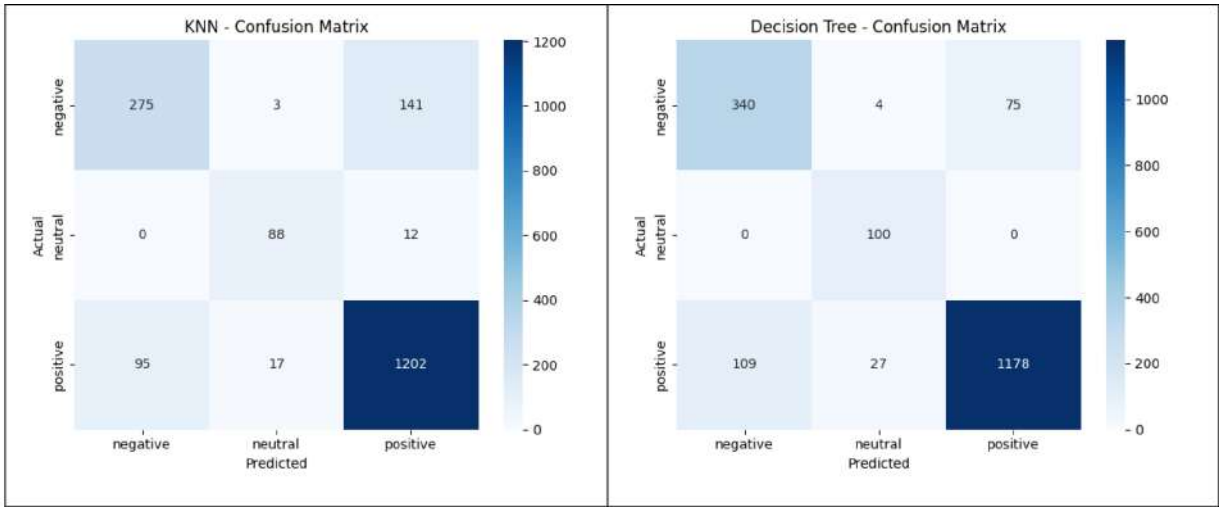

Fig. 5, 6. KNN and Decision Tree confusion matrix

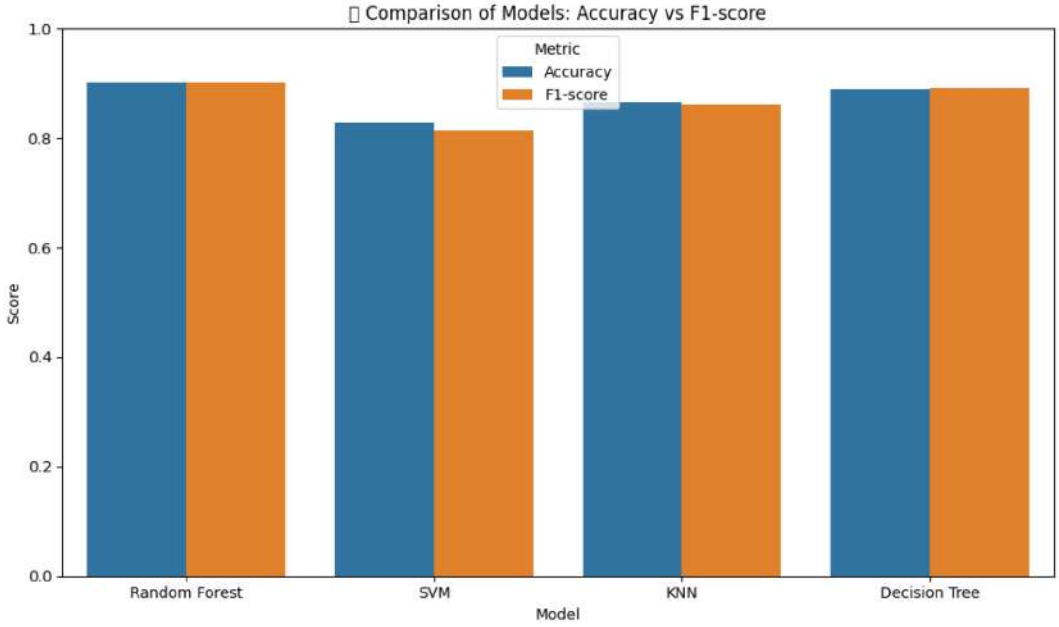

Fig. 7. Accuracy and F-1 bar chart

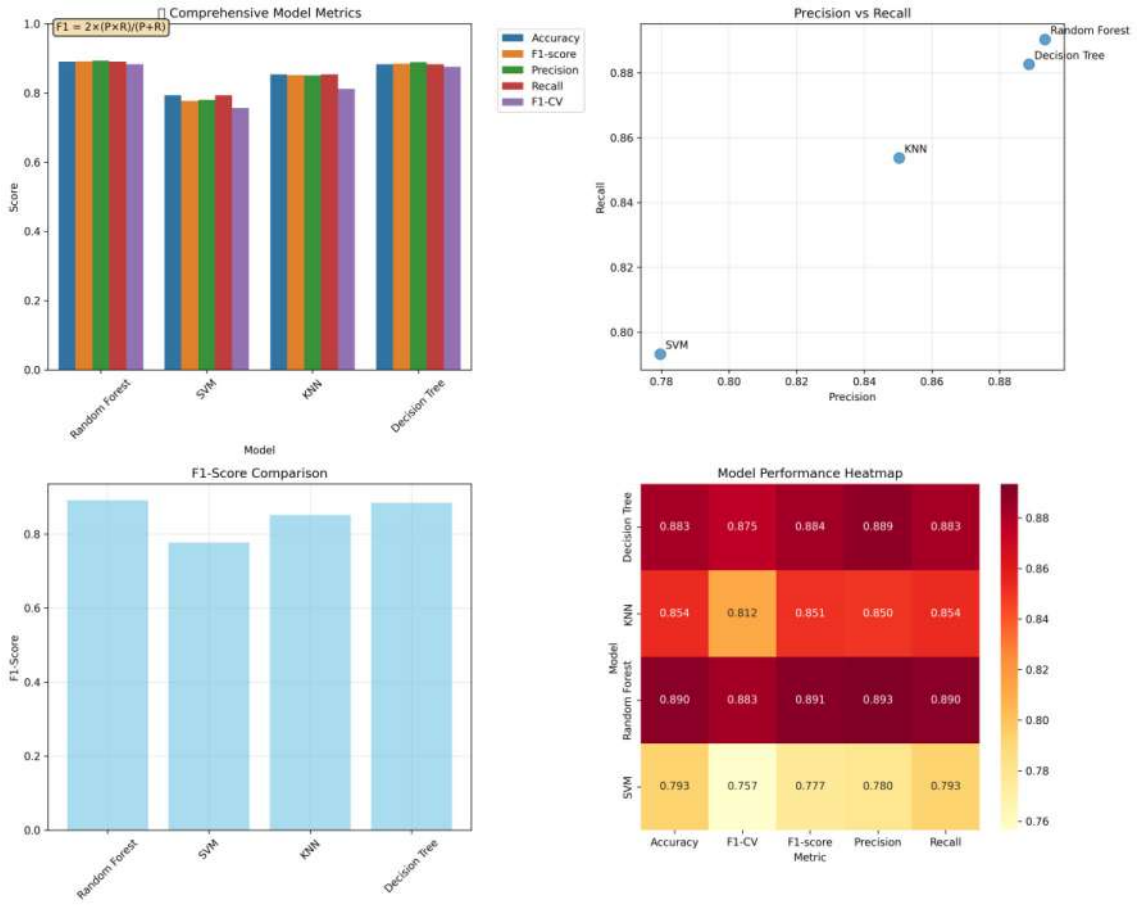

Fig. 8. Detailed comparison of ML methods in charts

The Random Forest model, on the other hand, handled feature interactions well thanks to its ensemble approach.

These results support the idea that limited training data in low-resource languages can be compensated for by hybrid approaches that combine statistical learning and handcrafted linguistic rules.

Conclusion

This study integrated conventional supervised machine learning classifiers with a meticulously designed rule-based sentiment analysis module to enhance the framework for sentiment classification in Ukrainian-language social media data. The proposed method addressed the morphological richness and linguistic complexity of the Ukrainian language, which can pose challenges for general sentiment analysis tools primarily designed for English.

Using an expanded lexicon, emoji sentiment mappings, phrase sentiment scores, and syntactic dependency parsing, nuanced sentiment signals that show contextual meanings specific to Ukrainian expressions were extracted. We used these manually created features, which had been turned into structured numerical vectors, to train and test four supervised models: Random Forest, Decision Tree, Support Vector Machine, and K-Nearest Neighbours.

The Random Forest classifier did better than single-tree and distance-based models in the tests. It had an accuracy of over 90 % and a strong F1-score. The Decision Tree and KNN models worked better than the SVM classifier. This means that simple linear separation might not be enough for the hybrid pipeline to work with the complicated interactions between features.

The results show that combining machine learning and rule-based linguistic insights in hybrid sentiment analysis pipelines greatly improves the accuracy of classifying Ukrainian texts.

Future research should concentrate on enhancing the feature set with supplementary domain-specific terminology, investigating more complex ensemble and neural architectures, and broadening the framework to incorporate the classification of neutral sentiment. To facilitate advantageous applications in domains such as political discourse analysis, customer feedback evaluation for Ukrainian-speaking communities, and public opinion monitoring, it is beneficial to integrate the pipeline with real-time social media surveillance tools.

In conclusion, this study adds to the growing body of research on sentiment analysis for low-resource languages and shows how a linguistically informed hybrid methodology can combine rule-based interpretability with the adaptability of supervised learning.

Bibliography

1. Syed, A., Aslam, M., & Saeed, F. (2020). A hybrid sentiment analysis approach for Urdu language social media data. *Journal of King Saud University – Computer and Information Sciences*, 32(4), 453–459. https://doi.org/10.1016/j.jksuci.2020.04.004
2. Elmadany, A., Abdul-Mageed, M., & Hashemi, H. (2021). Lexicon-augmented neural networks for dialectal Arabic sentiment analysis. У *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 3788–3802). Association for Computational Linguistics. https://aclanthology.org/2021.emnlp-main.306/
3. Kharde, V., & Sonawane, P. (2021). Hybrid techniques for sentiment analysis: A review. *Journal of Theoretical and Applied Information Technology*, 99(8), 1840–1852.
4. Vázquez, S., Balage Filho, P. P., & Pardo, T. A. S. (2022). Emoji and intensifier handling in multilingual sentiment classification. *Language Resources and Evaluation*, 56(2), 527–550. https://doi.org/10.1007/s10579-021-09561-1
5. Xia, R., Liu, Q., & Chen, S. (2023). Sentiment analysis of code-switched texts using hybrid pipelines. *Information Processing & Management*, 60(2), 103183. https://doi.org/10.1016/j.ipm.2022.103183
6. Zhou, L., Li, Q., & Zhang, X. (2024). A comparative study of rule-based, neural, and hybrid sentiment analysis models for low-resource languages. *Natural Language Engineering*, 30(1), 45–70. https://doi.org/10.1017/S1351324923000345
7. Syed, S., & Spruit, M. (2020). Full-text or abstract? Examining topic coherence scores using Latent Dirichlet Allocation. У *2020 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 528–537). IEEE. https://doi.org/10.1109/DSAA49011.2020.00056
8. Islam, M. R., Islam, M. M., Rahman, M. M., & Islam, M. S. (2020). Sentiment analysis of low-resource Bengali text using hybrid learning models. У *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICCCNT49239.2020.9225540
9. Bhaskar, M., & Saini, H. K. (2021). Hybrid sentiment analysis using machine learning techniques for Hindi tweets. *Procedia Computer Science*, 192, 3732–3741. https://doi.org/10.1016/j.procs.2021.09.148
10. Abozinadah, E. A., & Jones, J. (2021). A hybrid deep learning approach for sentiment analysis of Arabic tweets. *IEEE Access*, 9, 10241–10258. https://doi.org/10.1109/ACCESS.2021.3050634
11. Almanea, M., & Habash, N. (2020). Investigating the use of transformers for Arabic dialect sentiment analysis. У *Proceedings of the Fifth Arabic Natural Language Processing Workshop (WANLP 2020)* (pp. 63–77). Association for Computational Linguistics. https://aclanthology.org/2020.wanlp-1.7/
12. Mohammad, S. M., & Bravo-Marquez, F. (2020). EmoLex: An expanded emotion lexicon for social media analysis. У *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)* (pp. 3678–3684). European Language Resources Association (ELRA). https://aclanthology.org/2020.lrec-1.453/
13. Koto, F., Rahman, M., & Baldwin, T. (2020). IndoBERTweet: A pre-trained language model for Indonesian Twitter with emotion and sentiment understanding. У *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings* (pp. 2347–2352). Association for Computational Linguistics. https://aclanthology.org/2020.findings-emnlp.212/

References

1. Syed, A., Aslam, M., & Saeed, F. (2020). A hybrid sentiment analysis approach for Urdu language social media data. *Journal of King Saud University – Computer and Information Sciences*, 32(4), 453–459. https://doi.org/10.1016/j.jksuci.2020.04.004
2. Elmadany, A., Abdul-Mageed, M., & Hashemi, H. (2021). Lexicon-augmented neural networks for dialectal Arabic sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 3788–3802). Association for Computational Linguistics. https://aclanthology.org/2021.emnlp-main.306/
3. Kharde, V., & Sonawane, P. (2021). Hybrid techniques for sentiment analysis: A review. *Journal of Theoretical and Applied Information Technology*, 99(8), 1840–1852.
4. Vázquez, S., Balage Filho, P. P., & Pardo, T. A. S. (2022). Emoji and intensifier handling in multilingual sentiment classification. *Language Resources and Evaluation*, 56(2), 527–550. https://doi.org/10.1007/s10579-021-09561-1
5. Xia, R., Liu, Q., & Chen, S. (2023). Sentiment analysis of code-switched texts using hybrid pipelines. *Information Processing & Management*, 60(2), 103183. https://doi.org/10.1016/j.ipm.2022.103183
6. Zhou, L., Li, Q., & Zhang, X. (2024). A comparative study of rule-based, neural, and hybrid sentiment analysis models for low-resource languages. *Natural Language Engineering*, 30(1), 45–70. https://doi.org/10.1017/S1351324923000345
7. Syed, S., & Spruit, M. (2020). Full-text or abstract? Examining topic coherence scores using Latent Dirichlet Allocation. In *2020 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 528–537). IEEE. https://doi.org/10.1109/DSAA49011.2020.00056
8. Islam, M. R., Islam, M. M., Rahman, M. M., & Islam, M. S. (2020). Sentiment analysis of low-resource Bengali text using hybrid learning models. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICCCNT49239.2020.9225540

9. Bhaskar, M., & Saini, H. K. (2021). Hybrid sentiment analysis using machine learning techniques for Hindi tweets. *Procedia Computer Science*, 192, 3732–3741. https://doi.org/10.1016/j.procs.2021.09.148

10. Abozinadah, E. A., & Jones, J. (2021). A hybrid deep learning approach for sentiment analysis of Arabic tweets. *IEEE Access*, 9, 10241–10258. https://doi.org/10.1109/ACCESS.2021.3050634

11. Almanea, M., & Habash, N. (2020). Investigating the use of transformers for Arabic dialect sentiment analysis. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop (WANLP 2020)* (pp. 63–77). Association for Computational Linguistics. https://aclanthology.org/2020.wanlp-1.7/

12. Mohammad, S. M., & Bravo-Marquez, F. (2020). EmoLex: An expanded emotion lexicon for social media analysis. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)* (pp. 3678–3684). European Language Resources Association (ELRA). https://aclanthology.org/2020.lrec-1.453/

13. Koto, F., Rahman, M., & Baldwin, T. (2020). IndoBERTweet: A pre-trained language model for Indonesian Twitter with emotion and sentiment understanding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings* (pp. 2347–2352). Association for Computational Linguistics. https://aclanthology.org/2020.findings-emnlp.212/

Дата першого надходження рукопису до видання: 24.09.2025
Дата прийнятого до друку рукопису після рецензування: 20.10.2025
Дата публікації: 28.11.2025