Scientific Journal of the Ternopil National Technical University
2026, vol. 121, no. 1, pp. 46–55,
ISSN: 2522-4433 Print, ISSN: 2522-4441 Online
https://doi.org/10.33108/visnyk_tntu2026.01.046

Submitted: 2026-01-04 | Revised: 2026-03-20 | Accepted: 2026-03-24

Application of large language models in generating Ukrainian corpora for training text classification systems

Vyacheslav Nykytyuk *, Andrii Dolinskyi

Ternopil Ivan Puluj National Technical University, Ternopil, Department computer science, Ukraine,
e-mail address kafkn@tntu.edu.ua

*Corresponding author e-mail slavikvv89@gmail.com

Abstract. *This study deals with the application of modern Artificial Intelligence technologies, mainly deep learning methods based on recurrent neural networks of the Long Short-Term Memory (LSTM) type, for the construction and analysis of Ukrainian-language corpora. The research attention is directed to the automated classification of text sentiment (tonality), which involves distributing texts into positive, neutral, and negative categories. This, sequentially, testified the effectiveness of the proposed algorithmic solutions and their suitability for the tasks of analyzing the emotional tone of Ukrainian-language content.*

Keywords: *Artificial intelligence, deep learning, LSTM models, text sentiment, language dataset, corpus, text classification, natural language processing, Ukrainian language, automation, message analysis.*

1. INTRODUCTION

The Ukrainian language, as one of the leading languages of the Eastern European region, plays an important role in determining the cultural, educational, and information environment. However, the implementation of modern Natural Language Processing (NLP) technologies for the Ukrainian language faces a number of significant challenges, the most pressing of which is the lack of large-scale, high-quality, and representative corpora [1].

In particular, most available Ukrainian language corpora offer limited coverage and fail to adequately reflect the current state of the language, its dialectal variability, stylistic diversity, and the characteristics of various communicative domains – official, informal, scientific, and others. Furthermore, there is a noticeable imbalance in the development of language resources: while English, Chinese, and other global languages receive substantial investment in NLP infrastructure, the Ukrainian language often remains excluded from such initiatives, thereby limiting its competitiveness in the digital sphere.

The development of high-quality datasets necessitates substantial human and financial resources, as data annotation, verification, and curation procedures require the involvement of qualified domain experts. Concurrently, existing linguistic resources frequently exhibit structural inconsistencies, semantic distortions, or data incompleteness, thereby adversely impacting the performance accuracy of natural language processing models. These challenges are further compounded by the intricate grammatical architecture of the Ukrainian language, its extensive morphological paradigm, and numerous orthographic and syntactic exceptions, which collectively impede automated text processing operations [1].

Within this framework, artificial intelligence technologies, particularly Large Language Models (LLMs), present promising avenues for addressing these limitations [2, 3]. By virtue of their capacity to generate, augment, and enrich linguistic data, LLMs constitute a potentially effective instrument for constructing novel Ukrainian language corpora. Nevertheless, their deployment necessitates rigorous empirical investigation, specifically concerning the assurance

of representativeness and reliability of generated data, the mitigation of algorithmic biases, and the identification of optimal methodologies for annotation and validation procedures. The resolution of these methodological challenges is of paramount importance for advancing Ukrainian natural language processing technologies and facilitating the integration of the Ukrainian language into global digital ecosystems, research infrastructures, and communication platforms.

Addressing these issues is critically important for the development of Ukrainian NLP technologies and the integration of the Ukrainian language into global digital services, research platforms, and communication systems.

Contemporary scholarly written works demonstrates substantial progress in the application of Artificial Intelligence (AI) technologies to corpora development, with particular emphasis on under-resourced languages such as Ukrainian. The integration of AI methodologies facilitates sophisticated linguistic analysis that transcends the limitations inherent in conventional rule-based text processing approaches. Transformer-based neural architectures, which constitute the foundation of state-of-the-art language models, exhibit exceptional efficiency in this domain, demonstrating robust capacity for large-scale textual data processing and yielding high-performance outcomes across typologically diverse languages [4, 5].

Among the most prominent architectures are the Generative Pre-trained Transformer (GPT) and Bidirectional Encoder Representations from Transformers (BERT) models, which have been extensively deployed across numerous natural language processing applications. Their advanced contextual representation mechanisms substantially enhance performance in machine translation, sentiment analysis, text classification, and other core NLP tasks [1, 6]. These models demonstrate considerable cross-lingual transferability and domain adaptability, rendering them particularly promising for deployment in Ukrainian language processing contexts [7, 8].

The growing interest in linguistic diversity has stimulated the development of universal language models capable of processing multiple languages simultaneously. This approach is particularly relevant for multilingual countries such as Ukraine, where there is a pressing need to integrate the Ukrainian language into multilingual digital platforms. AI-based models can automate complex linguistic tasks, including sentiment analysis and the creation of specialized datasets for content filtering – for instance, identifying positive or negative messages on social media.

However, the effectiveness of such technologies is directly dependent on the availability of large-scale, high-quality corpora that are adapted to the specific linguistic characteristics of each language. In the case of Ukrainian, this entails the development of dedicated resources that account for its grammatical, stylistic, and communicative features. Access to such data is essential to ensure the full integration of the Ukrainian language into contemporary NLP systems and digital services.

2. RESULTS AND DISCUSSION

Large Language Models (LLMs), such as GPT-4, demonstrate substantial potential for generating high-quality texts in the Ukrainian language [2, 3, 7]. Training these systems on extensive multilingual corpora enables them to produce content with diverse thematic coverage, stylistic variation, and sentiment [2].

Peculiarities of using LLMs for Ukrainian-language content:

- Model adaptation is required through fine-tuning on specialized Ukrainian-language corpora.
- The capability to create texts that account for the complex linguistic features of the language (e.g., morphological changes, stylistic nuances, and word polysemy) [1].

The process of applying LLMs covers several key stages:

- Formulating prompts to the model that contain clear instructions regarding the intended topic, style, and desired sentiment.
- Post-processing the generated output to correct potential errors and enhance overall quality.
- Sentiment labeling, which may be performed automatically or validated by experts [9, 10].

The development of a data corpus for analyzing the sentiment of Ukrainian texts is grounded in the integration of modern text generation methods with data quality control mechanisms [9, 11]. At the initial stage, a clear structure for the corpus was defined: it must include texts representing three sentiment categories – positive, negative, and neutral. In addition, it was essential to ensure coverage across various communicative domains, including social media, reviews, news, and official documents.

The ChatGPT Large Language Model was selected for content generation in accordance with these criteria [2], demonstrating its ability to effectively adapt to the stylistic, thematic, and sentiment-related characteristics of the Ukrainian language. The automation of the data generation process was achieved through integration with the OpenAI API, which enabled the submission of explicit and structured prompts.

The proposed research methodology comprises the following sequential steps: constructing prompts for generating texts with predefined characteristics, performing post-processing of the obtained results, and storing the finalized data in a standardized format.

For example, the following prompt can be used to create messages: «*Generate 10 comments in the Ukrainian language. In the format comment – sentiment. (0 – negative, 1 – neutral, – positive)».* The generated texts undergo post-processing refinement to correct potential grammatical inconsistencies and ensure strict adherence to the specified sentiment parameters (tonality). The result of this prompt will be the list of 10 comments:

```
Цей товар жахливої якості, не рекомендую його купувати.          -0
На вулиці йде дощ, а завтра прогнозують сніг.                     -1
Дуже вдячний за швидку доставку! Все ідеально.                    -2
Обслуговування у цьому ресторані залишає бажати кращого.         -0
Курс долара сьогодні становить 37,5 гривень.                      -1
Отримав свою посилку раніше, ніж очікував, усе супер!             -2
Комп’ютер постійно зависає, працювати неможливо.                  -0
Зустріч заплановано на п’ятницю о 14:00 у залі.                   -1
Мені дуже сподобалася ваша презентація, це було чудово!           -2
Пакунок прийшов пошкодженим, дуже засмучений.                     -0
```

The translations of these comments are provided below for clarity:

```
This product quality is awful, I don't recommend buying it.       -0
It's raining outside, and they're forecasting snow for tomorrow. -1
Really grateful for the fast delivery! Everything's perfect.      -2
The service at this restaurant leaves a lot to be desired.        -0
The dollar exchange rate today is 37.5 hryvnias.                  -1
Got my package earlier than expected, everything's awesome!       -2
The computer keeps freezing, it's impossible to work.             -0
The meeting is scheduled for Friday at 2:00 PM in the hall.       -1
I really loved your presentation, it was amazing!                 -2
The package arrived damaged, very disappointed.                   -0
```

A critical component of the methodology involves the verification and validation of the generated data. To enhance result reliability, a hybrid quality assurance framework is employed, incorporating both automated assessment algorithms and manual expert evaluation of a control subset [9, 10]. This dual-methodological approach ensures objective evaluation of automated sentiment classification accuracy while facilitating the identification of potential algorithmic biases or systematic errors within the dataset.

To demonstrate practical implementation, an illustrative Application Programming Interface (API) request is provided, configured to generate 100 Ukrainian-language comments, each explicitly annotated with its corresponding sentiment category. The output data are preserved in a standardized JavaScript Object Notation (JSON) format, where each tuple comprises a key-value pair structure: the comment text string and its associated sentiment label.

```
const requestBody = {
    model: 'gpt-4', // Using the GPT-4 model
    messages: [
        {
            role: 'system',
            content: 'You are a helpful assistant.'
        },
        {
            role: 'user',
            content: 'Generate 100 comments in the ' +
                     'Ukrainian language in the format ' +
                     'comment – tonality (0 – negative, 1 – ' +
                     'neutral, 2 – positive).'
        }
    ],
    max_tokens: 2000, // Maximum number of tokens
                      // in the response
    temperature: 0.7 // Model creativity
};

fetch(endpoint, {
    method: 'POST',
    headers: {
        'Content-Type': 'application/json',
        'Authorization': `Bearer ${apiKey}`
    },
    body: JSON.stringify(requestBody)
})
.then(response => {
    const result = JSON.parse(
        await fs.readFile('data.json', 'utf-8') || '[]'
    ),
    comments = response.json()
        .choices[0]
        .message.content.split('\n')
        .filter(comment => comment.trim() !== '');

    comments.forEach(row => {
        const column = row.split('-');
        if (column.length === 2) {
            result.push([column[0], column[1]]);
        }
    });
```

```
    await fs.writeFile('data.json',
        JSON.stringify(result, null, 2),
        () => console.log('Saved file - data.json')
    );
})
.catch(error => {
    console.error('Error:', error);
});
```

Textual data retrieved via API is stored in JSON format, providing a structured and machine-readable representation suitable for downstream tasks such as sentiment analysis, machine learning model training, and text classification. The generated comments encompass the full spectrum of sentiment polarities, thereby confirming the model's capacity to produce text aligned with predefined emotional tone and stylistic parameters. Empirical evaluation validated the methodological approach, demonstrating a high degree of correspondence between the semantic content of each generated text and its assigned sentiment category.

These findings are of critical importance for the development of high-quality Ukrainian-language datasets intended for automated sentiment detection. In addition, the use of JSON ensures consistency and facilitates seamless integration into external analytical pipelines, including social media monitoring platforms and marketing intelligence systems.

```
[
  ["Цей товар жахливої якості, не рекомендую його купувати.", "0"],
  ["На вулиці йде дощ, а завтра прогнозують сніг.", "1"],
  ["Дуже вдячний за швидку доставку! Все ідеально.", "2"],
  ["Обслуговування у цьому ресторані залишає бажати кращого.", "0"],
  ["Курс долара сьогодні становить 37,5 гривень.", "1"],
  ["Отримав свою посилку раніше, ніж очікував, усе супер!", "2"],
  ["Комп'ютер постійно зависає, працювати неможливо.", "0"],
  ["Зустріч запланована на п'ятницю о 14:00 у залі.", "1"],
  ["Мені дуже сподобалася ваша презентація, це було чудово!", "2"],
  ["Пакунок прийшов пошкодженим, дуже засмучений.", "0"]
]
```

The translations of these comments are presented above and have not been altered, consistent with the original versions.

In the constructed dataset, each entry is represented as a tuple comprising the comment text and a numerical sentiment label (0 – negative, 1 – neutral, 2 – positive). This standardized format ensures consistency and facilitates downstream processing, enabling its integration into various Natural Language Processing (NLP) and machine learning workflows, particularly for training sentiment classification models.

To enhance the efficiency of the model training phase, preliminary data preprocessing and structural organization are essential. Specifically, the dataset is partitioned into two distinct logical components: the textual content of the comments and their corresponding sentiment labels. This separation allows the model to more accurately capture and interpret the affective characteristics of each utterance.

The resulting data structure, stored as «comment-sentiment» pairs, is loaded from file prior to training and split into two independent arrays: one containing the comment texts and the other containing the sentiment labels. This approach ensures a coherent and machine-readable representation of the input data, which is critical for the effective training of sentiment prediction models.

Below is a source code example illustrating the procedure for loading the constructed dataset and separating it into two arrays (comments and sentiments) in preparation for model training:

```
async function loadData() {
    const rows = JSON.parse(await fs.readFile('data.json', 'utf-8'));
    const comments = [];
    const sentiments = [];
    rows.forEach(row => {
        const comment = cols[0];
        const sentiment = row[1];
        comments.push(comment);
        sentiments.push(sentiment);
    });

    return { comments, sentiments };
}
```

This code segment implements two primary functions: extracting the generated data from the source file and partitioning it into discrete arrays containing comment texts and their corresponding sentiment labels, thereby facilitating efficient data utilization during the model training pipeline. This architectural design streamlines subsequent integration of the corpus into training datasets for deep learning architectures.

It is imperative to emphasize that rigorous data preprocessing constitutes a fundamental prerequisite for achieving optimal model performance. Leveraging the curated and preprocessed corpus, the model can acquire the capacity to predict textual sentiment with substantial accuracy, thereby enabling large-scale sentiment analysis of Ukrainian-language textual data. The segregation of data into two distinct arrays is methodologically significant, as it preserves the explicit mapping between individual comments and their annotated sentiment categories throughout the training process. This structural organization enables the model to effectively discriminate among diverse emotional valences within textual data, which represents a critical determinant of classification accuracy.

Following preprocessing procedures, the comment and sentiment arrays can be directly integrated into the training workflow of Long Short-Term Memory (LSTM) networks. LSTM architectures constitute one of the predominant recurrent neural network variants extensively employed in natural language processing tasks [6, 12]. Their computational efficacy derives from the inherent capacity to process sequential data structures while maintaining long-range dependencies among sequence elements (i.e., words within sentences or paragraphs) [13].

The primary objective of this experimental investigation was to evaluate the impact of a high-quality dataset, synthesized via a large language model, on the model's sentiment classification performance for Ukrainian textual data [1, 8].

Principal parameters of the LSTM architecture:

- Input Layer: The model receives tokenized sequences represented as dense vector embeddings (word embeddings). A pre-trained word embedding space for Ukrainian, constructed from the generated dataset, was employed for this purpose [14]. Tokenization and subsequent vectorization enable the model to capture lexical semantics within context, thereby preserving semantic information for downstream processing.

- Hidden layer: A single LSTM layer comprising 128 memory units was implemented to process sequential data. This configuration facilitates the retention of long-range dependencies among textual elements, which is essential for accurate sentiment classification.
- Output layer: The final layer consists of a fully connected (dense) layer with three output neurons corresponding to negative (0), neutral (1), and positive (2) sentiment classes. The softmax activation function is applied to generate normalized probability distributions across all classes, transforming model outputs into interpretable probabilistic estimates for each sentiment category.
- Optimizer: The Adaptive Moment Estimation (Adam) optimizer was employed for weight optimization, demonstrating robust performance with high-dimensional data and complex parameter interdependencies. This optimizer represents one of the most effective and widely adopted algorithms for tasks requiring efficient and stable convergence.
- Loss function: Categorical cross-entropy was utilized as the objective function for this multiclass classification task, constituting the standard choice for such applications. This loss function enables the model to optimize class prediction probabilities through maximum likelihood estimation.
- Batch size: A mini-batch size of 32 samples was maintained throughout training, achieving an optimal trade-off between computational efficiency and gradient estimation accuracy.
- Training duration (number of epochs): The model underwent training for 15 epochs, providing sufficient exposure to the training corpus to converge toward optimal weight configurations for sentiment classification.

The dataset synthesized via ChatGPT, comprising textual comments with corresponding sentiment labels, was employed for model training. The corpus was partitioned into three subsets according to the following distribution:

- Training set (70%): The primary dataset utilized for learning the mapping between textual inputs and their associated sentiment labels. This subset enables the model to optimize its internal parameters through iterative exposure to labeled examples.
- Validation set (20%): A held-out subset employed for monitoring model accuracy during the training phase, facilitating hyperparameter tuning and early stopping criteria to prevent overfitting.
- Test set (10%): An independent evaluation subset reserved for post-training assessment, measuring the model's generalization capacity on previously unseen data and providing an unbiased estimate of real-world performance.

Metrics used to evaluate the model's effectiveness:

- Accuracy: Determines the percentage of correct predictions made by the model.
- Precision: Shows how many of the positively classified samples actually belonged to the positive category.
- Recall: Determines how many of all truly positive samples were correctly classified.
- F1-score: The harmonic mean of precision and recall, providing a balanced measure of model performance that is particularly valuable when both precision and recall are equally important for the task.

The values of these metrics for training and test datasets are shown in the Table 1.

Table 1. Metrics of LSTM model for training and test datasets

Metric	Training set	Test set
Accuracy	92,4%	88,7%
Precision	91,8%	87,9%
Recall	92,1%	88,3%
F1-score	91,9%	88,1%

The experimental evaluation yielded high performance scores across all assessment metrics, substantiating the effectiveness of the LLM-generated dataset for training LSTM-based sentiment classifiers on Ukrainian textual data. These findings validate that a systematically constructed corpus with well-defined sentiment labels facilitates accurate emotion-based text classification, consequently expanding the potential for large-scale, real-time sentiment analysis of Ukrainian-language content in various digital environments [8, 15].

3. CONCLUSIONS

The findings of this investigation demonstrate the substantial efficacy of contemporary artificial intelligence methodologies in Ukrainian language corpora development. Specifically, the deployment of Large Language Models (LLMs), exemplified by ChatGPT, for generating sentiment-annotated textual data facilitated the construction of a high-quality, representative corpus, thereby considerably streamlining the training pipeline for automated sentiment classification systems [2, 9, 11].

The proposed framework for automated text generation, incorporating subsequent post-processing procedures to ensure grammatical accuracy and adherence to predefined sentiment parameters, enabled the acquisition of large-scale datasets. These corpora proved adequate for training deep learning architectures, particularly LSTM networks, which exhibited robust performance across multiple evaluation metrics, including Accuracy, Precision, Recall, and F1-score [6, 13, 15].

The investigation further underscores the critical importance of rigorous validation protocols for synthetically generated data. The integration of automated verification mechanisms with expert human evaluation enhances result reliability while mitigating the risk of algorithmic bias in classification tasks [9, 10]. This hybrid validation approach substantiates the applicability of LLM-generated corpora across diverse NLP applications, ranging from sentiment analysis to misinformation detection.

Consequently, these results establish the considerable potential of the proposed methodology for advancing Ukrainian natural language processing technologies. Progressive refinement of model architectures and algorithmic frameworks can facilitate increasingly precise and efficient automation of textual analysis workflows, representing a significant advancement in the development of contemporary AI systems and their practical deployment within computational linguistics domains [5, 12].

Author Contributions: Conceptualization: A. Dolinskyi and V. Nykytyuk; Methodology: A. Dolinskyi; Software: A. Dolinskyi; Validation: A. Dolinskyi and V. Nykytyuk; Formal analysis: A. Dolinskyi; Writing – original draft preparation: A. Dolinskyi; Writing – review and editing: V. Nykytyuk; Supervision: V. Nykytyuk.

Acknowledgments: The authors would like to thank the Department of Computer Systems and Networks of Ternopil National Technical University for the provided technical support and environment for conducting this research.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflicts of Interest: The authors declare no conflict of interest.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declaration of Generative AI and AI-assisted technologies in the writing process: During the preparation of this work, the authors used ChatGPT (OpenAI) in order to generate the synthetic Ukrainian-language corpus, assist in the structural organization of the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

REFERENCE

[1] O. Zalutska, Method for analyzing the Ukrainian language texts sentiment using natural language processing, in: Information Control Systems and Intelligent Technologies. Advances and Applications, Liha-Pres, Lviv, 2022, pp. 122–137. https://doi.org/10.36059/978-966-397-538-2-7. (In Ukrainian).

[2] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language Models are Few-Shot Learners, (2020). https://doi.org/10.48550/arXiv.2005.14165 (Accessed 15 October 2025).

[3] T. Chen, Y. Chen, T. Gui, Q. Huang, L. Xue, Qi. Zhang, A Survey on Large Language Models in Natural Language Processing, (2023). https://doi.org/10.48550/arXiv.2303.18223 (Accessed: 15 October 2025).

[4] R. Khrabatyn, V. Zaiats, Technologies for designing the structure of the information system for monitoring the technical condition of bridge structures, Scientific Journal of the Ternopil National Technical University. 109 (2023) 72–79. https://doi.org/10.33108/visnyk_tntu2023.01.072. (In Ukrainian).

[5] O.A. Pastukh, O.V. Tkach, Brain-computer interaction based on motor imagery using machine learning, Scientific Journal of the Ternopil National Technical University,. 112 (2023), 26–31. https://doi.org/10.33108/visnyk_tntu2023.04.026. (In Ukrainian).

[6] F. Li, C. Cui, Y. Hu, L. Wang, Sentiment Analysis of User Comment Text based on LSTM, WSEAS Transactions on Signal Processing. 19 (2023) 19–31. https://doi.org/10.37394/232014. 2023.19.3.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need, (2017). https://doi.org/10.48550/arXiv.1706.03762 (Accessed: 15 October 2025).

[8] M. Prytula, Fine-tuning BERT, DistilBERT, XLM-RoBERTa and Ukr-RoBERTa models for sentiment analysis of Ukrainian language reviews, Artificial Intelligence, 2 (2024), 85–97. https://doi.org/10.15407/jai2024.02.085. (In Ukrainian).

[9] B. Ding, J. Zhou, Z. Li, X. Long, X. Li, Z. Wu, S. Gao, Data Augmentation using LLMs: Data Perspectives, Learning Paradigms and Challenges, (2024). https://doi.org/10.48550/arXiv.2403.02990 (Accessed: 15 October 2025).

[10] H. Dai, Z. Liu, W. Liao, X. Huang, Y. Cao, Z. Wu, S. Zhao, W. Zhu, S. Wei, T. Liu, N.S. Peng, AugGPT: Leveraging ChatGPT for Text Data Augmentation, (2023). https://doi.org/10.48550/arXiv.2302.13007 (Accessed: 15 October 2025).

[11] M. Bayer, M.A. Kaufhold, C. Reuter, A Survey on Data Augmentation for Text Classification, ACM Computing Surveys., 55 (2023), 1–39. https://doi.org/10.1145/3544558.

[12] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, Cambridge, 2016. https://www.deeplearningbook.org.

[13] S. Hochreiter, J. Schmidhuber, Long Short-Term Memory, Neural Computation,. 9 (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

[14] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, (2013). https://doi.org/10.48550/arXiv.1301.3781 (Accessed: 15 October 2025).

[15] A. Simarmata, Anthony, Tiffany, M. Phanie, Sentiment Analysis On Twitter Posts About The Russia and Ukraine War With Long Short-Term Memory, Sinkron, 8 (2023), 762–772. https://doi.org/10.33395/sinkron.v8i2.12235.