Automated CEFR-Level Assessment for Ukrainian Texts

Olha Kanishcheva
Heidelberg University
SET University
kanichshevaolga@gmail.com

Mikhail Kopotev
Stockholm University
University of Helsinki
mihail.kopotev@helsinki.fi

Abstract

The present study evaluates CEFR-based text complexity for Ukrainian using a new dataset compiled from textbooks, designed for language learners. We compare traditional machine learning, transformer-based models, and LLM-based evaluation across A1–B2 language proficiency levels. Results show that explicit linguistic features remain highly effective: a Random Forest classifier achieves the highest macro-F1 (0.576), slightly outperforming fine-tuned XLM-RoBERTa (0.574). While GPT-5.5 shows strong performance (macro-F1 0.564), marking a significant advancement over GPT-4.1, supervised models achieve slightly better scores in this experiment for the proficiency-level assessment. These findings suggest that structured linguistic analysis is a robust alternative to purely neural approaches for Ukrainian CEFR classification.

1 Introduction

The Common European Framework of Reference for Languages[1] (CEFR) is widely used to describe second language proficiency through six ascending levels, from A1 to C2. These levels are based on "can-do" statements that explain what learners are able to perform at each level. However, these descriptors are often too general and subjective to provide an operational definition of the linguistic features that distinguish one level from another. Therefore, more objective, data-driven methods for identifying proficiency levels are required.

Recent developments in NLP offer new possibilities for automated language assessment. Traditional approaches have relied on manually selected features within the Complexity-Accuracy-Fluency framework (Michel, 2017). Although useful, average text-level measures often hide important variation within learner texts. Modern machine learning models, including transformer-based approaches such as BERT and generative transformers, allow more accurate classification by analyzing linguistic patterns in large learner corpora (Schmalz and Brutti, 2021).

Automatic assessment is particularly important for educational purposes. It helps create appropriate learning materials and makes information accessible to different groups of learners. While widely spoken languages have many resources and tools for this task (Misgna et al., 2025; Li and Ng, 2024), Ukrainian still lacks modern systems for text classification. Existing approaches often rely on traditional readability formulas developed for English, which do not account for Ukrainian's rich morphology and flexible syntax.

This paper presents a comparative study of automated text classification for Ukrainian across CEFR levels A1 to B2. Using a curated dataset of approximately 400 educational texts, we evaluate the performance of various classification methods, ranging from classical machine learning with handcrafted linguistic features to modern LLMs. The study examines how lexical diversity, morphology, and syntactic complexity contribute to level prediction, providing the first systematic benchmarks for Ukrainian readability assessment.

2 Related Work

A feature-driven approach to CEFR classification relies on machine learning classifiers, such as support vector machines or random forests, trained on linguistically annotated data with lexical, syntactic, and morphological features (Kurdi, 2020; Sung et al., 2015; Balyan et al., 2018). In the extreme case, Hancke and Meurers utilize 3 821 features, which are grouped according to lexical, morphological, and syntactic complexity. Their study found that the strongest signals are syntactic and lexical, and performance comes from combinations of the

[1] https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions

Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), pages 209–222

features. The best model achieved 64.5% accuracy.

In practice, researchers typically select only those linguistic features that are available for automatic processing. They are grouped into four categories.

- **Lexical features**: word frequency, lexical diversity, word length, and lexical density. Word frequency is computed by matching lemmas against a reference corpus; lexical diversity is measured as the proportion of types/tokens in a text; word length is calculated as the mean number of characters per token; and lexical density is calculated as the proportion of content words (nouns, verbs, adjectives, and adverbs) among all tokens. The hypothesis underlying these metrics is that lower-level texts typically use high-frequency, simple vocabulary, whereas higher-level production includes specialized and low-frequency terms (Kurdi, 2020; Kate et al., 2010).

- **Syntactic features**: sentence and clause length, tree depth, subordination ratios, and complex constructions. These measures can be extracted from Universal Dependencies (UD) relations (Fig. 1): sentence length in tokens; clause-based measures that rely on verbal heads; subordination operationalized through dependent-clause relations; and syntactic depth calculated as the mean dependency-path in a sentence. The presupposition here is that the lower-level texts tend to use simpler syntactic structures, whereas advanced texts show greater embedding and more varied syntax (Kate et al., 2010; Dascălu et al., 2012).

- **Morphological features**: inflectional diversity, grammatical paradigms, and derivational complexity (Seiffe et al., 2022). These are computed from UD morphological feature bundles, including case, number, gender, tense, aspect, mood, person, and converbal forms. Such measures are particularly important for morphologically rich languages like Ukrainian, where proficiency is reflected not only in lexical choice but also in control over inflectional paradigms.

- **Semantic and cohesion features**: latent semantic analysis, embedding-based similarity, cohesion metrics, and contextual representations (Liu and Lee, 2023). These features are computed as lexical overlap and/or cosine similarity between adjacent sentences or larger textual units using vector representations of words, sentences, or paragraphs.

The selection of these features is theoretically motivated rather than arbitrary. They represent the main linguistic domains that have repeatedly been associated with proficiency development: lexical sophistication and diversity, syntactic elaboration, morphological control, and discourse-level cohesion. Handcrafted features are especially useful in this study because they can be automatically extracted and vary across proficiency levels. They also allow for the identification of linguistic dimensions that contribute to classification rather than treating proficiency prediction as a black-box task. Thus, the feature set is intended not to be exhaustive, but to provide a linguistically motivated and empirically testable baseline for automatic proficiency assessment.

While these features provide a comprehensive framework for analysis, language proficiency assessment has historically relied mainly on graded word lists and a limited set of grammatical features. This approach is often criticized for its inherent subjectivity, as the selection often depends too heavily on the individual experience and pedagogical intuition of the compilers (Kisselev et al., 2024).

Recent research has shifted toward more holistic methods that integrate a broader range of linguistic indicators. In parallel with these developments, data-driven deep learning methods, including fine-tuned Transformer models such as BERT in its multilingual versions, have emerged as powerful tools that can learn contextualized patterns without the need for explicit feature engineering (Imperial et al., 2025; Lee et al., 2021).

Schmalz and Brutti employed pre-trained BERT-base models to classify written exams from the EFCAMDAT and CLC-FCE corpora. Their approach achieved remarkably high accuracy, reaching nearly 98% when trained on large labeled datasets. Furthermore, they discovered that augmenting learner texts with corrections—whether provided by human evaluators or automated tools like LanguageTool—further improved the system's ability to accurately assign CEFR levels.

Hybrid architectures combine transformer embeddings with handcrafted features, achieving strong performance, especially on small- and medium-sized datasets (Lee et al., 2021). Re-

cently, descriptor-based prompting of instruction-tuned LLMs has emerged as a few-shot approach that uses CEFR level descriptors directly (Imperial et al., 2025).

Cross-lingual and multilingual strategies address the scarcity of CEFR-annotated corpora for many languages. Multilingual models can approximate monolingual performance, and large resources such as the UniversalCEFR corpus facilitate standardized benchmarking across languages (Imperial et al., 2025; Vajjala and Rama, 2018).

Research on CEFR classification for Ukrainian is currently limited, largely because of a shortage of annotated corpora. However, methodologies adapted from other morphologically rich or high-resource languages, such as multilingual transformer fine-tuning and hybrid feature-based approaches, provide a robust framework for developing effective Ukrainian proficiency classifiers.

3 Data Collection

The dataset used for training and evaluation was curated from diverse educational resources that remain copyright-protected and therefore cannot be distributed directly. This study specifically targets A1 to B2 proficiency levels, as they represent the most critical stages for language learners. A list of the textbooks used in this study is provided in Appendix A.

3.1 Initial Dataset

At the outset of this research, we faced a significant challenge: Ukrainian lacked labeled datasets based on CEFR levels A1–B2. Most of the necessary materials were only available in printed format, such as textbooks and teaching aids for Ukrainian as a foreign language. To solve this problem, we digitized these materials and labeled the texts according to the respective CEFR levels, as defined in the sources. Selected statistics for the data collected from these training materials are presented in Table 1.

Examples of texts for the selected CEFR levels are presented in Appendix B.

3.2 Linguistic Feature Extraction and Analysis

To evaluate the complexity of the Ukrainian dataset, we extracted a comprehensive set of linguistic features, categorized into four primary dimensions:

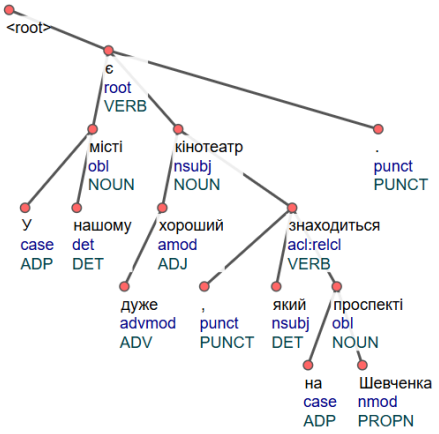


Figure 1: Example of a dependency syntactic tree for a Ukrainian sentence parsed using the UD framework.

- *Descriptive character-based metrics*: average token length (in characters), average syllables per token, and total token count.
- *Lexical diversity* : number of unique tokens and lemmas, number of hapax legomena, and Moving-Average Type-Token Ratio (MATTR) calculated for both tokens and lemmas (Covington and McFall, 2010).
- *Morphological diversity*: part-of-speech (POS) distribution frequencies and the proportion of functional words.
- *Syntactic complexity*: sentence length (mean, median, min, max in both tokens and characters), clause counts per sentence, and dependency tree depth. For each dependency tree, we identify the longest chain of syntactic dependencies from the root node to any dependent node and record its length as the tree's maximum depth. For each text, we then calculate the arithmetic mean of these maximum-depth values in all dependency trees in the text (see the example in Figure 1).

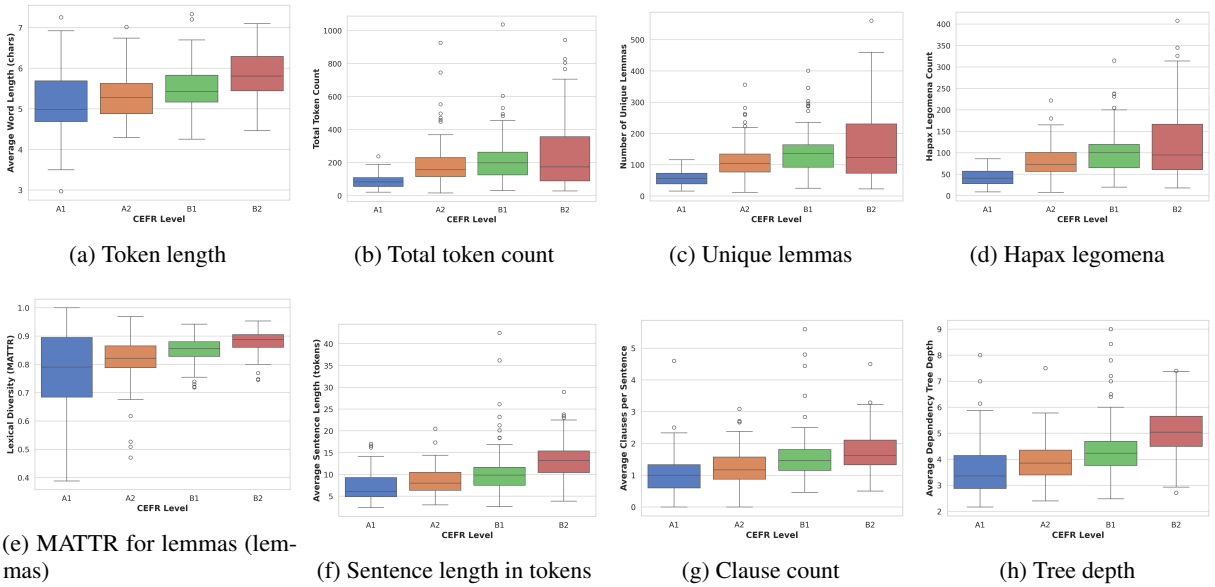
Figure 2 illustrates the distribution of selected metrics. We used boxplots to visualize median values, interquartile ranges, and potential overlaps between the levels. Detailed visualizations for all linguistic features across all CEFR levels, along with the source code for the statistical analysis, are publicly available on the project repository.[2]

The charts in (Figure 2) illustrate the main linguistic characteristics of texts across different lev-

[2] https://github.com/kopotev/fluencymeter/tree/main/UNLP_2026_paper_materials

	A1	A2	B1	B2
Number of files	89	123	110	115
Average text length (tokens)	89	188	215	242
Min/max text length (tokens)	21 / 238	16 / 926	30 / 1037	28 / 943
Total number of tokens	8,270	23,879	24,140	28,224

Table 1: Descriptive statistics for each CEFR level in the initial dataset (Ukrainian educational texts).



(a) Token length (b) Total token count (c) Unique lemmas (d) Hapax legomena

(e) MATTR for lemmas (lemmas) (f) Sentence length in tokens (g) Clause count (h) Tree depth

Figure 2: Distribution of selected linguistic features across CEFR levels A1-B2.

els of Ukrainian language proficiency (A1–B2). Overall, the results show a consistent increase in textual complexity as the proficiency level rises. Lower-level texts (A1–A2) are characterized by shorter lengths, lower average word and sentence lengths, and reduced lexical diversity. At intermediate levels (B1–B2), these indicators gradually increase, reflecting a broader vocabulary and more complex syntactic structures. Thus, the observed trends align with the expected progression of linguistic complexity across CEFR levels.

All feature values for all texts from our dataset are publicly available on GitHub.[3]

3.3 Quantitative Analysis of Text Complexity

To ensure the validity of our corpus and identify the most discriminative linguistic markers for Ukrainian, we performed a comprehensive statistical analysis using one-way analysis of variance (ANOVA).[4] This approach allows us to determine whether the complexity of texts differs significantly between CEFR levels and identify which features most effectively distinguish proficiency stages.

Furthermore, we conducted a linear trend analysis using polynomial contrasts to verify the monotonic progression of complexity from A1 to B2. This dual-method approach confirms that the observed differences are unlikely to be random variation but represent a consistent stepwise increase in linguistic difficulty. The results for the most impactful features are summarized in Table 2.

The statistical analysis revealed that 27 of 30 investigated metrics exhibited significant differences across levels ($p < 0.001$) (Appendix C). The most prominent marker of language proficiency in our study is the *average dependency tree depth* ($F = 41.60, p < 0.001$), which shows a highly significant linear trend. The mean depth increases steadily from 3.68 (A1) to 5.06 (B2), indicating that syntactic structures in Ukrainian become more hierarchical and deeply nested as the level rises.

Syntactic complexity is further evidenced by *sentence length* (mean tokens) and the *average number of clauses*. The average sentence length increases from 7.35 to 13.07 tokens, while the number of clauses nearly doubles.

[3] https://github.com/kopotev/fluencymeter/tree/main/UNLP_2026_paper_materials/figures

[4] https://en.wikipedia.org/wiki/Analysis_of_variance

Linguistic Feature	F-statistics	ANOVA p	Trend p	Mean (A1)	Mean (B2)
Avg. Dependency Tree Depth	41.60	< 0.001	< 0.001	3.68	5.06
Hapax Legomena	39.78	< 0.001	< 0.001	43.11	120.86
Avg. Sentence Length (tok)	36.92	< 0.001	< 0.001	7.35	13.07
Unique Lemmas	34.33	< 0.001	< 0.001	57.85	158.66
MATTR (Lemmas)	31.74	< 0.001	< 0.001	0.77	0.88
Avg. Clauses per Sentence	24.03	< 0.001	< 0.001	1.04	1.75
Noun Frequency	21.66	< 0.001	< 0.001	35.13	99.28
Avg. Token Length (chars)	21.56	< 0.001	< 0.001	5.22	5.85

Table 2: ANOVA results and mean values for selected linguistic features across CEFR levels (A1–B2).

Post hoc comparisons using Tukey's HSD confirmed that these increases are statistically significant at almost every level transition, reflecting a shift from simple sentences to complex multi-clause constructions. Lexical richness metrics also demonstrated robust growth. *Unique Lemmas* ($F = 34.33$) and *hapax legomena* ($F = 39.78$) show substantial increases, particularly during the transition from A1 to A2, where the number of unique words more than doubles.

The *MATTR score for lemmas* score also shows a significant linear trend ($p_{trend} < 0.001$), confirming a more diverse and less repetitive vocabulary at higher proficiency levels. Interestingly, while certain categories such as *gerunds* ($p = 0.094$) and *numerals* ($p = 0.134$) did not show significant differences across the entire range, morphological markers such as *noun frequency* ($F = 21.66$) and *function-word count* ($F = 14.04$) emerged as reliable indicators of complexity. The growth in function words aligns with increased syntactic demands, as they provide the necessary logical and structural cohesion for advanced discourse.

3.4 Thematic Consistency and Data Validation

To ensure that the classification models are driven by linguistic complexity markers rather than specific domain vocabulary, we conducted a thematic distribution analysis using the BERTopic framework.[5] This validation step is crucial for ruling out topic bias, where a model might associate a specific level with a particular subject rather than its grammatical or structural features.

3.4.1 Preprocessing and Model Configuration

Before topic modeling, the text data underwent specialized preprocessing for Ukrainian. This included lemmatization using the `Spacy` (uk_core_news_sm) library to unify inflected word forms and the removal of stop words.

[5] https://huggingface.co/docs/hub/bertopic

The BERTopic pipeline was configured with the following parameters to ensure robust cluster formation:

- **embeddings:** We used the *multilingual-e5-base* model to generate high-dimensional document representations.
- **dimensionality reduction:** UMAP was employed with $n\_neighbors = 30$ and $n\_components = 10$ to preserve local structures.
- **clustering:** HDBSCAN was configured with $min\_cluster\_size = 15$ and $min\_samples = 2$ to minimize noise while identifying distinct thematic groups.
- **vectorization:** A `CountVectorizer` with $ngram\_range = (1, 2)$ was used to capture both individual terms and common phrases.

3.4.2 Analysis of Results

The thematic analysis shows a clear progression in lexical content across CEFR levels. By applying lemmatization, we achieved higher semantic density within clusters, allowing for the identification of specific linguistic markers for different proficiency stages. The key topics identified by the BERTopic model are visualized in Figure 3.

The distribution of these topics across CEFR levels (Figure 4) shows that our data follow a natural path of language learning. Although some subjects appear at all levels, others serve as clear markers of specific proficiency stages.

- **Topic 3**, (related to *classroom/student/room* "аудиторія/ студент/кімната"), is a dominant marker of the **A1-level texts (32.6%)**, representing the basic concrete vocabulary typical of beginners.
- **Topic 6**, (focusing on *culture/technology/profession* "культура/

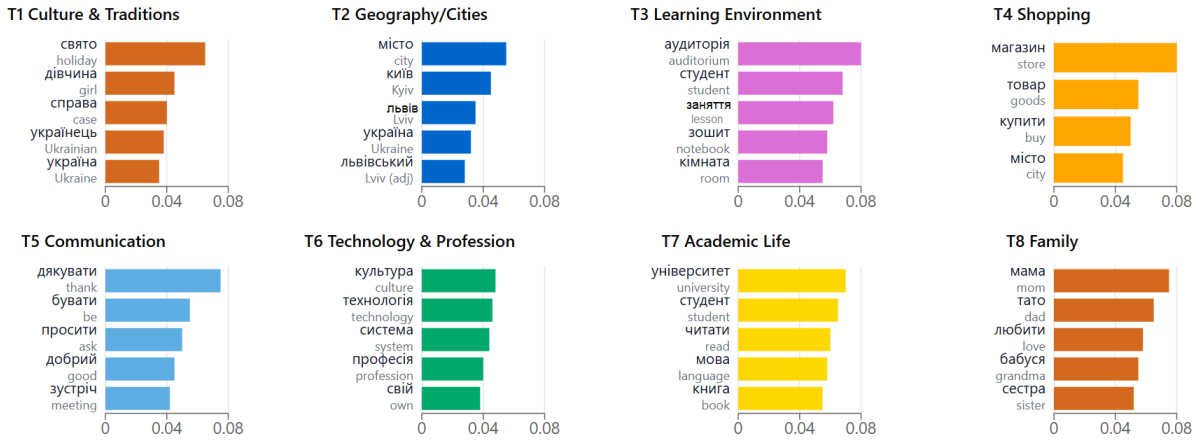


Figure 3: Top words associated with the identified latent topics (Topic Word Scores).

технологія/професія″), becomes significant at the **B2 level (25.2%)**, reflecting the shift toward abstract and professional discourse.

Although the outlier rate for Topic 0 (Unclassified) is relatively high due to the lexical diversity of the corpus, the remaining clusters show clear separation. The consistent presence of Topic 1 and Topic 2 across multiple levels suggests a shared thematic foundation, ensuring that classification performance is driven by structural linguistic complexity and syntactic markers rather than purely topic-specific vocabulary.

4 Model Training and Evaluation

In this section, we describe our experiments to evaluate the effectiveness of various computational methods for the automated CEFR classification of Ukrainian texts. To provide a comprehensive assessment, our experiments are structured around three distinct approaches: (i) feature-driven machine learning, which uses the handcrafted linguistic features analyzed in the previous section to train classical classifiers; (ii) transformer-based deep learning, which leverages state-of-the-art pretrained language models to capture deep contextual representations without manual feature engineering; and (iii) LLM inference, which uses few-shot prompting techniques.

The objective of these experiments is to determine which methodology best captures the morphological and syntactic nuances of Ukrainian while maintaining high classification accuracy across the A1–B2 proficiency levels.

Linguistic features capturing the linguistic properties of the texts were extracted and used as input variables for classification. The feature set included lexical complexity measures (e.g., average token length, average syllables per token, lexical diversity metrics such as MATTR, and hapax legomena counts), vocabulary size indicators (e.g., total tokens, unique tokens, and unique lemmas), part-of-speech frequency distributions (e.g., frequencies of nouns, verbs, adjectives, adverbs, and function words), sentence-level statistics (e.g., mean, median, minimum, and maximum sentence length in tokens and characters), as well as syntactic complexity measures derived from dependency parsing, such as the average number of clauses and the average depth of syntactic trees.

To investigate the contribution of different linguistic aspects, features were organized into several groups: lexical features, part-of-speech distribution features, sentence-level statistics, and syntactic features. In addition to evaluating these groups separately, a combined feature set containing all extracted features was also tested. Before training, numerical features were standardized using z-score normalization, and missing values were handled using median imputation to ensure model robustness.

4.1 Supervised Machine Learning Algorithms

Four supervised machine learning algorithms were evaluated: Random Forest, Gradient Boosting, Support Vector Machines (SVM), and Logistic Regression. All models were implemented using the Scikit-learn library[6] (Pedregosa et al., 2011).

To ensure optimal performance, each model was integrated into a pipeline that included median imputation for missing values and standard scaling of

[6] https://scikit-learn.org/stable/

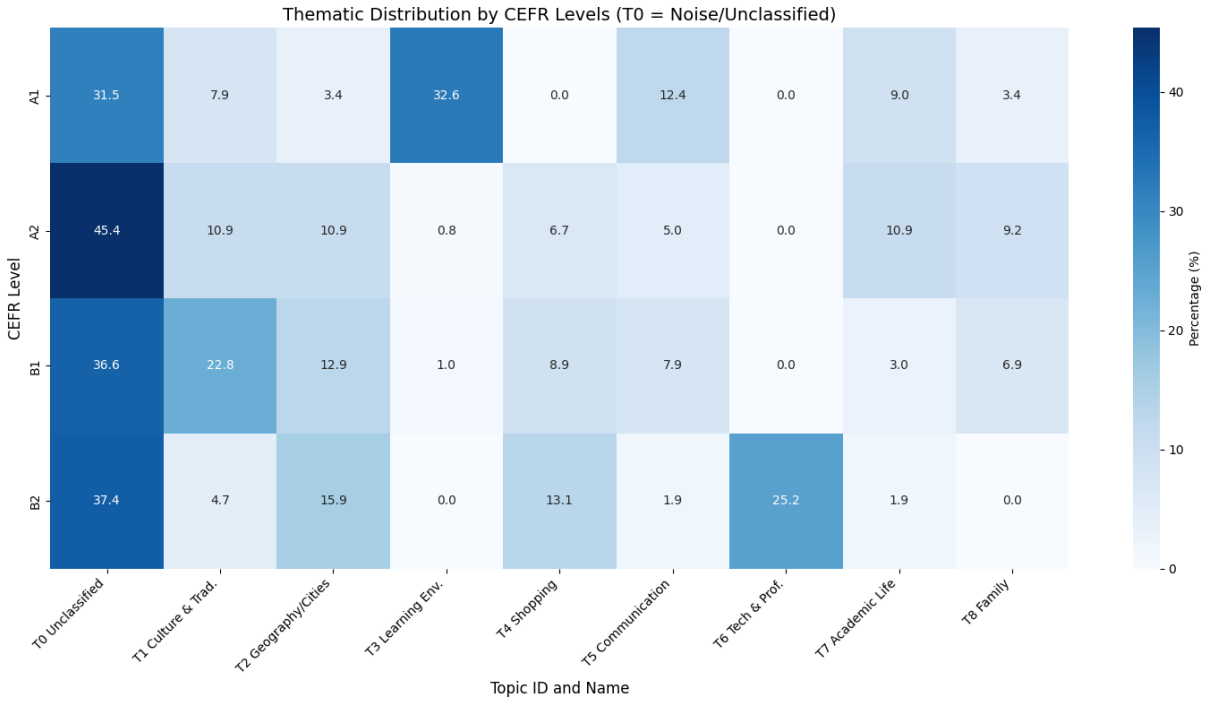


Figure 4: Heatmap of normalized latent topic distribution across CEFR proficiency levels.

linguistic features. Random Forest and Gradient Boosting models were implemented with 300 estimators, providing strong baselines for capturing nonlinear feature interactions. Logistic Regression was trained with a maximum of 2,000 optimization iterations to ensure convergence, while the SVM classifier used a radial basis function (RBF) kernel to address nonlinear decision boundaries.

Model performance was evaluated using stratified five-fold cross-validation, ensuring a balanced distribution of CEFR levels across all folds. For each model and feature configuration, we computed accuracy and macro-averaged precision, recall, and F1-score (Table 3). Macro-averaging ensures that each CEFR level is treated with equal importance, regardless of its frequency in the dataset.

4.2 Transformer-based Deep Learning Methods

For the CEFR classification task, we fine-tuned two state-of-the-art multilingual transformer models: XLM-RoBERTa-base[7] and mBERT (BERT-base-multilingual-cased)[8].

Both architectures were integrated with a sequence classification head and trained on a stratified 80/20 dataset split to ensure consistent class representation across training and testing sets. We used the Hugging Face framework with a learning rate of 2×10^{-5} and weight decay of 0.01. Training was conducted for 6 epochs for XLM-RoBERTa (with a maximum sequence length of 512 tokens) and 4 epochs for mBERT (with a 256-token limit and a batch size of 4). Model performance was evaluated using accuracy and macro-averaged precision, recall, and F1-score to provide a balanced assessment across all CEFR levels (Table 3).

4.3 LLMs and Few-shot Prompting Techniques

In addition to traditional machine learning approaches, we extend our evaluation to modern LLMs to assess their performance on Ukrainian CEFR classification. Specifically, we conducted experiments using the GPT-4.1 and GPT-5 model families via their respective APIs through a few-shot prompting strategy.

The LLM-based classification framework is structured as follows: (i) few-shot prompting is employed by providing the models with a curated set of labeled examples for each CEFR level (A1–B2); these prompts instruct the model to prioritize linguistic complexity – including lexical diversity, syntactic structures, and discourse features – while disregarding text length. This approach ensures

[7] https://huggingface.co/docs/transformers/model_doc/xlm-roberta

[8] https://huggingface.co/docs/transformers/model_doc/bert

consistency, particularly when distinguishing between borderline levels like A2 and B1; (ii) to account for the inherent stochasticity of LLM outputs, we implemented a self-consistency decoding strategy. Each input text is processed through five independent iterations with stochastic sampling. The final proficiency label is then determined by a majority vote across these runs, intended to reduce variability and improve overall prediction reliability.

The configurations for these prompts and the experimental setup are documented in the project repository.[9] Comparative results, showcasing how these LLMs perform alongside the other evaluated models, are presented in Table 3.

4.4 Evaluation Metrics and Results

To ensure a fair comparison, all models, including the feature-based classifiers and the fine-tuned XML-RoBERTa, were evaluated using an identical stratified five-fold cross-validation protocol. This ensures that the performance metrics (macro-F1, accuracy) and the resulting confusion matrices reflect the models' behavior on the same data partitions, eliminating evaluation bias.

The performance of the Random Forest model is visualized through the confusion matrix in Fig. 5. The matrix is presented in row-normalized percentages to facilitate comparison across CEFR levels. Overall, the model demonstrates a robust ability to distinguish between proficiency levels, particularly at the two ends of the scale. The highest class-wise recall rates are observed for A1 and B2 and for the A1 (70.8%) and B2 (68.7%) levels, while the intermediate levels (A2 and B1) show more substantial overlap, which is consistent with the gradual nature of language acquisition.

The transformer-based models demonstrated competitive performance. Fine-tuning XLM-RoBERTa on the CEFR-annotated dataset achieved an overall accuracy of 0.591 and a macro-F1 score of 0.574 (Table 3). While the model performed well on distinct proficiency levels, the intermediate B1 level remained a challenge for neural architectures as well, reflecting the transitional nature of these texts.

Overall, while XLM-RoBERTa achieved the highest accuracy, the feature-based Random Forest model demonstrated nearly identical macro-F1

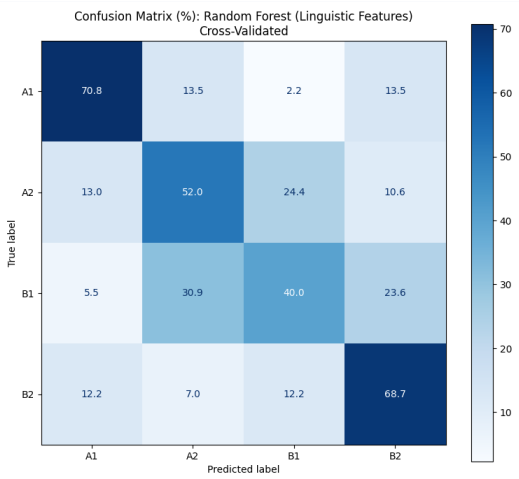


Figure 5: Row-normalized confusion matrix for the Random Forest model, %.

performance, suggesting that explicit linguistic features remain highly effective for Ukrainian CEFR classification.

The evaluation of large language models using prompting shows that few-shot learning provides reasonably strong results but does not always outperform traditional approaches. For GPT-4.1, the standard few-shot setup achieved a macro-F1 score of 0.498 with an accuracy of 0.510. Adding self-consistency led to almost no improvement (macro-F1 = 0.499).

GPT-5.5 performed better overall. Its few-shot configuration reached a macro-F1 score of 0.564 and an accuracy of 0.560, outperforming GPT-4.1 and approaching the performance of the transformer-based models. However, as with GPT-4.1, self-consistency did not bring clear benefits (macro-F1 = 0.561), suggesting that the model already produced stable predictions.

At the same time, supervised transformer models remained slightly stronger. XLM-RoBERTa achieved the highest accuracy (macro-F1 = 0.574, accuracy = 0.591). Among traditional classification methods, Random Forest also performed well (macro-F1 = 0.576).

Overall, the results show that modern LLMs such as GPT-5.5 can handle Ukrainian CEFR classification quite well with few-shot prompting, but supervised models still have a small advantage, especially for more precise distinctions between levels.

[9] https://github.com/kopotev/fluencymeter/tree/main/UNLP_2026_paper_materials/prompts

Model / Approach	Precision	Recall	Accuracy	macro-F1
Baselines (Feature-based ML)				
Random Forest (All linguistic features)	**0.575**	**0.579**	**0.572**	**0.576**
Gradient Boosting (All linguistic features)	0.531	0.533	0.529	0.527
Logistic Regression (All linguistic features)	0.506	0.499	0.501	0.492
SVM (All linguistic features)	0.522	0.507	0.508	0.495
Transformer-based Models				
mBERT (fine-tuned)	0.5299	0.5686	0.550	0.5371
XLM-RoBERTa (fine-tuned)	**0.5648**	**0.5930**	**0.5909**	**0.5742**
GPT-4.1				
Standard Few-Shot	0.620	0.500	0.510	0.4980
With Self-Consistency	0.630	0.500	0.510	0.4994
GPT-5.5				
Standard Few-Shot	**0.620**	**0.550**	**0.560**	**0.5644**
With Self-Consistency	0.620	0.550	0.560	0.5608

Table 3: Performance comparison of different classification methods for Ukrainian CEFR levels.

5 Conclusion and Future Work

This study evaluated computational approaches for the automated classification of Ukrainian texts across four CEFR proficiency levels (A1–B2). Our experimental results demonstrate that while fine-tuned transformer models like XLM-RoBERTa achieve strong performance, with a macro-F1 score of 0.574, traditional classifiers such as Random Forest remain highly competitive (macro-F1 of 0.576) when supplied with robust handcrafted features. In fact, the Random Forest model achieved the highest macro-F1 score, slightly outperforming XLM-RoBERTa in this respect.

The analysis of large language models shows that few-shot prompting achieves moderate results, but LLMs still struggle with fine distinctions between intermediate levels, especially A2 and B1. We also observed that longer input texts may bias the models toward higher-level predictions, likely because of the presence of more complex vocabulary and structures.

While LLMs are flexible and easy to apply, supervised models still offer advantages in this task. Models trained on explicit linguistic features or fine-tuned on annotated data provide more stable and interpretable results for CEFR classification.

Future work will focus on three key areas:

1. **Cross-lingual transfer.** We plan to explore the use of datasets from closely related Slavic languages through translation and adaptation to increase the volume of training data and evaluate cross-lingual feature stability.
2. **Corpus expansion.** We aim to expand our current corpus by collaborating with professional instructors of Ukrainian as a foreign language to ensure high-quality manual annotation and consistency across a wider range of text genres.
3. **Application development.** The long-term goal is to develop a web-based application that provides researchers and educators with a practical interface for the automated analysis and classification of Ukrainian texts according to language proficiency levels.

Limitations

The findings of this study should be considered in light of several limitations:

- Dataset size. The corpus used for this research is relatively small, containing approximately 400 texts. However, this reflects a broader challenge in Ukrainian NLP, where annotated pedagogical datasets are scarce. This study serves as a pilot evaluation to establish baseline benchmarks for larger-scale data collection.
- Level imbalance and coverage. The research focused on levels A1 through B2. While these represent the most active stages of language learning, the lack of C1 and C2 materials in the current dataset limits the model's ability to assess advanced proficiency, where linguistic nuances are more complex.

- Interpretability vs. performance. Although transformer-based models, including modern LLMs, demonstrated strong classification accuracy, they remain black boxes compared to handcrafted linguistic features. This limits their practical pedagogical adoption. The trade-off between the high performance of neural models and the pedagogical interpretability of linguistic features remains a key challenge for automated assessment tools.

Ethical Considerations

We followed ethical guidelines for data use and writing throughout this study. The data came from multiple sources and are protected by copyright, so we cannot share them directly. Instead, we have included a detailed description of all data sources and instructions on how to access them in Appendix A. All data were used under fair-use principles for academic research only. We also used AI tools, including ChatGPT, Gemini, and Grammarly, to edit the manuscript. We carefully reviewed all AI-generated suggestions and take full responsibility for the final content of this paper.

Acknowledgments

We would like to thank the reviewers for their time and effort in reviewing this manuscript. We sincerely appreciate their valuable comments and suggestions, which greatly helped us improve the quality of the work. The authors would like to thank Yurii Prokopenko for valuable assistance in locating relevant educational materials and for helpful consultations during the preparation of this work. This research was partially funded by the Research Council of Finland.

References

Renu Balyan, Kathryn S McCarthy, and Danielle S McNamara. 2018. Comparing machine learning classification approaches for predicting expository text difficulty. In *The Thirty-First International Florida Artificial Intelligence Research Society Conference (FLAIRS-31)*.

Michael A Covington and Joe D McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (MATTR). *Journal of quantitative linguistics*, 17(2):94–100.

Mihai Dascălu, Stefan Trausan-Matu, and Philippe Dessus. 2012. Towards an integrated approach for evaluating textual complexity for learning purposes. In *Advances in Web-Based Learning - ICWL 2012*, pages 268–278, Berlin, Heidelberg. Springer Berlin Heidelberg.

Julia Hancke and Detmar Meurers. 2013. Exploring CEFR classification for German based on rich linguistic modeling. In *Learner Corpus Research*, pages 54–56.

Joseph Marvin Imperial, Abdullah Barayan, Regina Stodden, Rodrigo Wilkens, Ricardo Muñoz Sánchez, Lingyun Gao, Melissa Torgbi, Dawn Knight, Gail Forey, Reka R. Jablonkai, Ekaterina Kochmar, Robert Joshua Reynolds, Eugénio Ribeiro, Horacio Saggion, Elena Volodina, Sowmya Vajjala, Thomas François, Fernando Alva-Manchego, and Harish Tayyar Madabushi. 2025. UniversalCEFR: Enabling open multilingual research on language proficiency assessment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9703–9755, Suzhou, China. Association for Computational Linguistics.

Rohit Kate, Xiaoqiang Luo, Siddharth Patwardhan, Martin Franz, Radu Florian, Raymond Mooney, Salim Roukos, and Chris Welty. 2010. Learning to predict readability using diverse linguistic features. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 546–554.

Olesya Kisselev, Mikhail Kopotev, and Anton Vakhranev. 2024. Measuring lexical knowledge in russian as a second language: Exploring the potential of lexical lists for language proficiency assessment. In Gillian Lord and Lara Lomicka, editors, *The Routledge Handbook of Second Language Acquisition and Technology*, chapter 5, pages 65–81. Routledge.

M Zakaria Kurdi. 2020. Text complexity classification based on linguistic information: Application to intelligent tutoring of esl. *Journal of Data Mining & Digital Humanities*.

Bruce W. Lee, Yoo Sung Jang, and Jason Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10669–10686, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shengjie Li and Vincent Ng. 2024. Automated essay scoring: Recent successes and future directions. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI)*, pages 8114–8122. ijcai.org.

Fengkai Liu and John Lee. 2023. Hybrid models for sentence readability assessment. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 448–454, Toronto, Canada. Association for Computational Linguistics.

Marije Michel. 2017. Complexity, accuracy, and fluency in L2 production. In Shawn Loewen and Masatoshi Sato, editors, *The Routledge Handbook of Instructed Second Language Acquisition*, pages 50–68. Routledge. Available via Lancaster EPrints.

Hiwot Misgna, Byung-Won On, Ingyu Lee, Geunho Choi, and Ha-Young Kim. 2025. A survey on deep learning-based automated essay scoring and feedback generation. *Artificial Intelligence Review*, 58(2):36.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, and David Cournapeau. 2011. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830.

Veronica Juliana Schmalz and Alessio Brutti. 2021. Automatic assessment of English CEFR levels using BERT embeddings. In *Proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-it 2021)*, pages 295–301.

Laura Seiffe, Fares Kallel, Sebastian Möller, Babak Naderi, and Roland Roller. 2022. Subjective text complexity assessment for German. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 707–714, Marseille, France. European Language Resources Association.

Yao-Ting Sung, Wei-Chun Lin, Scott Benjamin Dyson, Kuo-En Chang, and Yu-Chia Chen. 2015. Leveling L2 texts through readability: Combining multilevel linguistic features with the CEFR. *The Modern Language Journal*, 99(2):371–391.

Sowmya Vajjala and Taraka Rama. 2018. Experiments with universal CEFR classification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 147–153, New Orleans, Louisiana. Association for Computational Linguistics.

A Appendix

The following textbooks were used in the dataset construction:
Палінська О. М. Крок-1 (рівень А1-А2). Українська мова як іноземна: книга для студента / Олеся Палінська, Оксана Туркевич; за ред. Ірини Ключковської. — Львів, 2010. — 102 с.

Українська мова як іноземна. Тексти для читання. Практикум для студентів підготовчого відділення / С. Д. Карпенко, Т. М. Рудакова, О. Д. Будугай та ін.; за ред. С. Д. Карпенко. — Київ: Видавничий дім Дмитра Бураго, 2019. — 256 с.

Українська мова як іноземна. Змістовий модуль «Читання». Рівні складності А2, В1. Методичні вказівки для аудиторної та самостійної роботи студентів підготовчого відділення / уклад. С. Д. Карпенко. — Біла Церква: ВПЦ БНАУ, 2019. — 260 с.

Бакум З. П. Калейдоскоп культур (рівень В1): навчальний посібник / З. П. Бакум, О. О. Пальчикова. — Кривий Ріг, 2014. — 101 с.

Дерба С. М. Українська мова як іноземна: навчальний посібник для студентів-магістрів / С. М. Дерба, Н. С. Ніколаєва. — Київ: Видавництво «Фенікс», 2021. — 136 с.

Тестові завдання до сертифікаційного іспиту з української мови (рівень А1–А2) / укл. Н. М. Малюга, В. А. Городецька. — Кривий Ріг: Видавець Роман Козлов, 2019. — 119 с.

B Appendix

Level	Example (Ukrainian)	Translation (English)
A1	– Привіт! Підеш сьогодні з нами кататися на велосипеді до парку? Я б залюбки, та не можу! Річ у тім, що в мене сестричка захворіла. Зараз до аптеки поспішаю по ліки. А потім буду наглядати за нею, доки батьки з роботи повернуться. – Дуже шкода!	– Hi! Will you go cycling in the park with us today? – I'd love to, but I can't! The thing is, my sister is sick. I'm rushing to the pharmacy for medicine now. And then I'll be looking after her until my parents return from work. – That's a pity!
A2	Мене звати Мухаммед. Я з Камеруну. Моє рідне місто Яунде. Це політична столиця Камеруну. Моя рідна мова французька. Я хочу бути програмістом. Люблю футбол, комп'ютерні ігри, комп'ютерне моделювання. На фото моя родина. ...	My name is Muhammed. I am from Cameroon. My hometown is Yaoundé. It is the political capital of Cameroon. My native language is French. I want to be a programmer. I love football, computer games, and computer modeling. My family is in the photo. ...
B1	Народні звичаї та обряди є давніми формами духовної культури народу. Вони, як і рідна мова, об'єднують людей в один народ. Розрізняють два види обрядів – календарно-обрядові та сімейні. Календарно-обрядові звичаї та обряди пов'язані з календарними циклами (взимку, навесні, влітку, восени). ...	Folk customs and rituals are ancient forms of a nation's spiritual culture. Much like the native language, they unite people into a single nation. There are two main types of rituals: calendar-ritual and family-based. Calendar customs and rituals are associated with seasonal cycles (winter, spring, summer, and autumn). ...
B2	Британська компанія Citymapper запускає нову транспортну послугу Smart Ride. Розробники обіцяють об'єднати в ній переваги трьох видів транспорту: фіксовані зупинки, як у автобуса, можливість замовити поїздку, як в таксі, та єдину транспортну мережу, як в метро. ...	The British company Citymapper is launching a new transport service called Smart Ride. The developers promise to combine the advantages of three modes of transport: fixed stops like a bus, the ability to book a trip like a taxi, and a unified transport network like a subway system. ...

Table 4: Examples of Ukrainian texts by CEFR level.

C Appendix

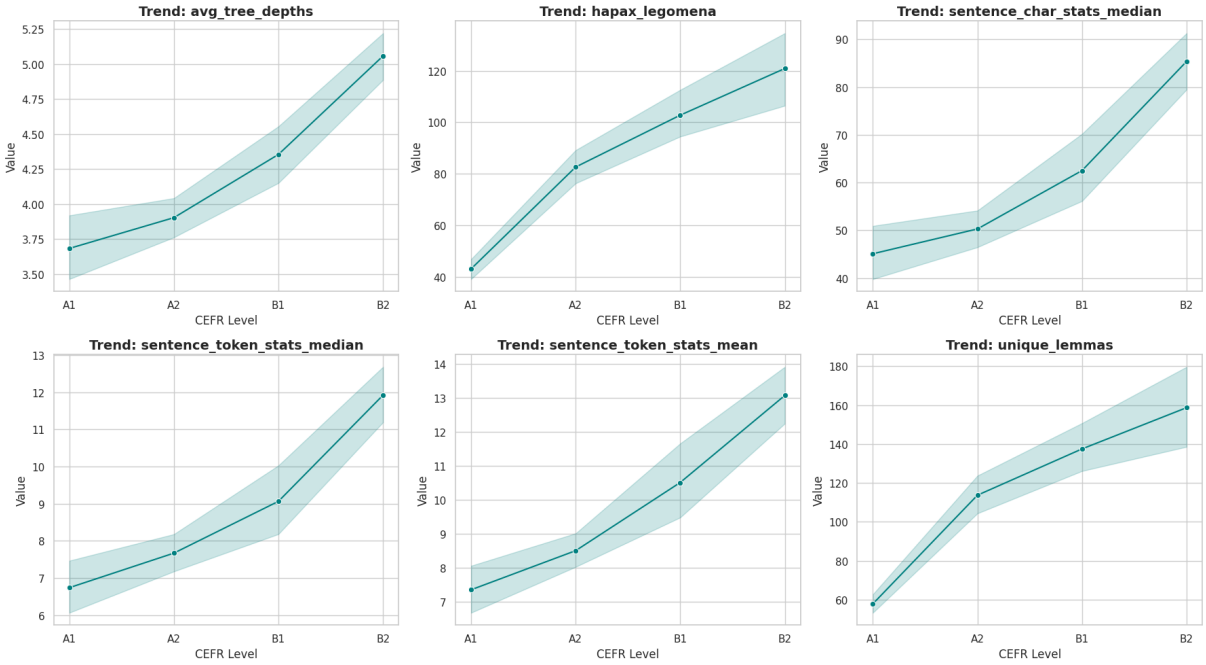


Figure 6: ANOVA trend analysis for key linguistic features across CEFR levels.