УДК 004.8:004.9

INTELLIGENT SENTIMENT ANALYSIS OF NON-STANDARD UKRAINIAN REVIEWS

Bozhok R. V.[1]; Zhulkovskyi O. O., Candidate of technical sciences, Assoc. Prof.[2]
[1,2] Dniprovsky State Technical University
[1,2] Kamianske, Ukraine
[1] rodion6568@gmail.com; [2] olalzh@ukr.net

Abstract*. A software prototype for sentiment analysis of multidomain Ukrainian reviews with group-based data splitting and controlled annotation of 278 texts has been developed. Source ratings agreed with human labels for 98.6% of reviews with unambiguous sentiment. A TF-IDF model trained on rating-derived labels achieved a macro-F1 score of 0.87 on human labels versus 0.35 for the majority baseline, and 0.82 and 0.81 on restaurant and hotel reviews. Offloading AdamW optimizer states to system memory enabled full XLM-R fine-tuning on a 4 GB GPU. The prototype can support review monitoring on marketplaces and booking services.*
Keywords: *sentiment analysis; Ukrainian reviews; XLM-R; cross-domain transfer; lexical robustness.*

Introduction. Ukrainian-language reviews on e-commerce marketplaces and booking platforms represent a massive source of product and service evaluations; however, they frequently contain colloquialisms, Anglicisms, Russianisms, code-switching, and non-standard spellings. Models trained on normative texts from a single domain often experience performance degradation when applied to such noisy reviews and unseen domains, while data leakage of related records across dataset splits masks this decline. The current state of research in sentiment analysis is characterized by the active adoption of transfer learning methods; a systematic review of these approaches, along with the challenges of cross-domain classification and mixed-sentiment identification, is presented in [1]. Particular attention is given to the methodological rigor of experiments; specifically, [2] demonstrates that data leakage across splits distorts the performance evaluation of machine learning models. Promising results in cross-lingual transfer have been reported for the XLM-R architecture [3], whereas for the classical Naive Bayes classifier, prospective directions for improving accuracy include analyzing contextual word sequences and applying logarithmic transformations [4]. Despite existing advances, the performance of these models on multi-domain Ukrainian reviews under leak-free data splitting remains an open question. Therefore, developing tools for controlled corpus curation of Ukrainian reviews and reproducible evaluation of sentiment analysis models is a relevant task.

Research aim – to develop a software prototype for intelligent sentiment analysis of non-standard Ukrainian reviews that ensures controlled corpus preparation and reproducible model evaluation.

Main Part. The research methodology is based on the principles of a controlled experiment, where each label has a documented provenance, related reviews are not split across training and test subsets, and the model checkpoint is selected strictly using validation data. Since manually annotating a dataset large enough for training is resource-intensive, a weak supervision approach is applied: training labels are derived from source ratings, human annotations are reserved exclusively for testing, and groups containing test reviews are excluded from the training set. Two evaluation protocols are established: in-domain evaluation and leave-one-domain-out cross-domain evaluation. Text non-standardness is treated as a dedicated analytical dimension: non-standard lexical variants

are preserved in the text rather than removed, and are explicitly tagged during annotation. The novelty of this work lies in integrating group-aware splitting of related records, training on rating-derived labels while evaluating on human-annotated labels, and span-level annotation of lexical variants within a unified protocol for constructing a Ukrainian review corpus.

The language of 662,907 raw reviews was identified using the fastText lid.176 model, and the selected texts along with their source metadata were imported into a unified data schema. Related records were merged into transitive groups based on the normalized text hash and the reviewed entity key. Benchmark annotation is performed via a web interface that conceals the source rating, model predictions, and the other annotator's labels; the annotator determines whether the text is a valid review, identifies its language, domain, and sentiment (positive, negative, neutral, or mixed), and tags Anglicisms, Russianisms, slang, non-standard spellings, and typographical errors. For training, ratings of 1–2 were mapped to the negative class, and ratings of 4–5 to the positive class, while reviews with a rating of 3 were excluded. The main experiment compared a majority-class baseline against logistic regression using word (1–2-grams) and character (2–5-grams) TF-IDF features with class weighting; the pilot study additionally evaluated TF-IDF with LinearSVC and XLM-R base (278M parameters), which was fully fine-tuned using AdamW (learning rate $2 \cdot 10^{-5}$, batch size 8, maximum sequence length of 256 tokens). The primary evaluation metric is macro-averaged F1 (macro-F1), with 95% confidence intervals (CIs) estimated via bootstrap resampling with 2,000 iterations.

The implementation was developed in Python using the PyTorch, Transformers, and scikit-learn libraries.

Following language filtering, 95,420 candidate reviews were retained (14.4% of the initial dataset), comprising 94,725 product, 602 restaurant, and 93 hotel reviews, which were grouped into 14,791 clusters. For benchmark annotation, 278 reviews from distinct groups were sampled (100 product, 100 restaurant, and 78 hotel reviews) and labeled by a single annotator. Three texts were identified as non-reviews; among the remaining 275 valid reviews, 124 are positive, 102 negative, 46 mixed, and 3 neutral. Across 82 texts (29.5%), 117 lexical spans were tagged: typographical errors (48), Russianisms (22), foreign-language insertions (32), slang (8), English borrowings (4), and non-standard spellings (3); additionally, 24 texts exhibit mixed language.

Comparing the human annotations against the source ratings revealed that for reviews with unambiguous sentiment, the mapping rule "1–2 as negative, 4–5 as positive" matches the human label in 205 out of 208 cases (98.6%). At the same time, the annotator classified 15.3% of reviews with a rating of 4–5 as mixed, whereas a rating of 3 does not reliably indicate sentiment (comprising 16 mixed, 14 negative, and 4 positive reviews).

In the pilot study on 30 product reviews using stratified 2-fold cross-validation (15 training examples per fold), all four methods predicted exclusively the positive class, yielding an accuracy of 0.60 and a macro-F1 of 0.1875. Consequently, in the main experiment, the model was trained on rating-derived labels.

The training set for the main experiment comprises 84,992 reviews with rating-derived labels (84,602 product reviews; 88.9% positive), while the test set consists of 226 positive and negative reviews with human-annotated labels. The TF-IDF model achieved a macro-F1 of 0.870 (95% CI: 0.82–0.91), compared to 0.354 for the majority-class baseline. Across individual domains, macro-F1 reached 0.90 (0.82–0.96) for products, 0.82 (0.73–0.89) for restaurants, and 0.81 (0.64–0.94) for hotels. The model misclassified 27 out of 102 negative reviews and 1 out of 124 positive reviews; furthermore, 40 out of 46 mixed reviews were assigned to the positive class.

On a GPU with 4 GB of VRAM, full fine-tuning of XLM-R in FP32 with standard on-device AdamW placement resulted in out-of-memory (OOM) errors even at a batch size of 1×64 tokens. Offloading the optimizer states (≈2,07 GiB) and parameter copies (≈1,04 GiB) to system RAM

enabled full-parameter training: for a batch size of 8 × 256 peak GPU memory usage reached 2.33 GiB, process RAM usage was 7.32 GiB, and the average step duration was 2.71 s. A configuration with frozen word embeddings completed a step in 0.134 s (batch size 2 × 256), but updated only 85.6M parameters.

An analysis of the results revealed the following patterns. Source ratings serve as a reliable source of training labels for positive and negative reviews, but not for mixed ones: weak supervision makes it possible to utilize the entire candidate corpus, whereas human annotation is required for testing and for identifying the mixed class. On restaurant and hotel reviews, macro-F1 is lower than on product reviews by approximately 0.09; however, the 95% CI of the difference (from −0,03 to 0,20 for restaurants) includes zero, meaning a larger test set is required to confirm a statistically significant performance drop. After excluding the test groups, restaurant and hotel reviews account for less than 0.5% of the training set; consequently, the obtained results effectively characterize domain transfer from the product domain to the others, while the leave-one-domain-out protocol excluding the product domain is infeasible without additional data collection. The predominance of the positive class in the training set induced a systematic bias: over a quarter of negative reviews were misclassified as positive. Nearly a third of the texts contain non-standard lexical spans, confirming the need to separately evaluate model robustness to lexical variation. At an identical batch size of 2 × 256 offloading optimizer states increases step duration by a factor of approximately 17 (2.27 s vs. 0.134 s) compared to the frozen-embeddings configuration; nevertheless, among the evaluated configurations, it alone enabled full-parameter training on the available GPU.

Conclusions. A software prototype for the automated processing and sentiment analysis of Ukrainian-language reviews has been developed and validated. The proposed approach of utilizing rating-derived labels enabled high-quality training of baseline and transformer models on consumer-grade hardware through GPU memory usage optimization. The obtained results confirm the effectiveness of the developed solutions for reputation monitoring across e-commerce marketplaces and booking platforms. Future research will focus on expanding the domain diversity of the dataset, refining models for mixed-sentiment analysis, and improving their robustness to lexical variation.

Use of AI Tools. During the development of the software prototype, ChatGPT (GPT-6.1 Sol model, OpenAI), Gemini (Gemini 3.8 Flash model, Google), and Claude (Claude Opus 5.5 model, Anthropic) were used as assistants for writing and debugging source code and designing the system architecture. The Claude large language model (Claude Opus 5.5) was employed for stylistic editing, proofreading, and initial verification of the manuscript structure.

REFERENCES

[1] Chan, J. Y.-L., Bea, K. T., Leow, S. M. H., Phoong, S. W., & Cheng, W. K. (2023). State of the art: A review of sentiment analysis based on sequential transfer learning. *Artificial Intelligence Review, 56*(1), 749–780. https://doi.org/10.1007/s10462-022-10183-8

[2] Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns, 4*(9), Article 100804. https://doi.org/10.1016/j.patter.2023.100804

[3] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747

[4] Шумейко, О. О., Сотник, В. С., Жульковська, І. І., & Жульковський, О. О. (2021). Використання методів Data Mining для обробки мовної інформації. *Математичне моделювання*, 2(45), 48–57. https://doi.org/10.31319/2519-8106.2(45)2021.246944