Секція 2. Міждисциплінарність як свідчення продуктивної гібридності в наукових розвідках з германістики та поза ними (штучний інтелект включно).

Sektion 2. Interdisziplinarität als Zeichen einer produktiven Hybridität innerhalb und außerhalb germanistischer Forschung, KI inbegriffen.

DOI: https://doi.org/10.20535/IWPOK3.2025.art.21
UDK 004.89:81'322.5

Hiloviants, K. D.
Ivanenko, S. M.
Shapoval, N. V.
Nationale Technische Universität der Ukraine
"Kyjiwer Igor-Sikorsky Polytechnisches Institut"

ANWENDUNG MODERNER SPRACHMODELLE FÜR MULTICLASS-SENTIMENT-ANALYSE BEI CODE-SWITCHING

Einleitung. Die Sentiment-Analyse ist eine Methode der Verarbeitung natürlicher Sprache (Natural Language Processing, NLP), die zur automatisierten Erkennung emotional gefärbter Lexik in Texten und zur Bestimmung der emotionalen Bewertung des Autors in Bezug auf die im Text beschriebenen Objekte entwickelt wurde. Die Sentiment-Analyse gehört zu den Kernaufgaben des NLP.

Die Sentiment-Analyse findet breite Anwendung in der Wirtschaft (Analyse von Kundenbewertungen), in den Medien (das Monitoring der öffentlichen Meinung), in den Sozialwissenschaften (Untersuchung gesellschaftlicher Stimmungen) und in der Informationssicherheit (Erkennung toxischer Inhalte). Dank der Einführung von Transformer-Architekturen – einer modernen neuronalen Netzarchitektur mit Aufmerksamkeitsmechanismus (Attention Mechanism) zur effizienten Verarbeitung von Datensequenzen – erreichen Sentiment-Klassifikatoren über 90 % Genauigkeit auf

englischsprachigen Datensätzen. Diese Erfolge sind jedoch ungleichmäßig auf Sprachen verteilt.

Ressourcenreiche Sprachen (High-Resource Languages) verfügen über große Textdatenmengen (Hunderte Milliarden Token), annotierte Datensätze und eine etablierte Forschungsinfrastruktur. Dazu gehören Englisch, Chinesisch, Spanisch, Deutsch und Französisch. Low-Resource-Sprachen sind hingegen durch einen Mangel an annotierten Korpora (normalerweise weniger als 100 Milliarden Token) und eine begrenzte Anzahl von NLP-Tools gekennzeichnet.

Relevanz. Zu den Low-Resource-Sprachen gehören zahlreiche afrikanische, asiatische und europäische Sprachen, darunter Sanskrit, Suaheli, dravidische Sprachen sowie viele Sprachen Mittel- und Osteuropas. Die ukrainische Sprache ist trotz erheblicher Fortschritte bei der Ressourcenerstellung – insbesondere des Kobza-Datensatzes (Haltiuk & Smywiński-Pohl, 2025), der etwa 60 Milliarden Token umfasst – im Vergleich zu ressourcenreichen Sprachen im Kontext spezialisierter NLP-Aufgaben, insbesondere für die Sentiment-Analyse gemischtsprachiger Inhalte, noch immer unterrepräsentiert.

Die verbreitete Verwendung von Code-Switching (Sprachwechsel innerhalb eines Textes, im Ukrainischen als Suržyk bezeichnet) in der digitalen Kommunikation ukrainischer Sprecher verkompliziert diese Aufgabe zusätzlich. Die Entwicklung effizienter Systeme zur Sentiment-Analyse ukrainischsprachiger Inhalte ist von großer Bedeutung für das Monitoring der öffentlichen Meinung, die Erkennung von Desinformation, die Verbesserung von Business Analytics und die Gewährleistung digitaler Sicherheit, insbesondere unter den Bedingungen des Krieges und der informationellen Konfrontation.

Sentiment-Analyse für Low-Resource-Sprachen. Die Forschungsgemeinschaft hat mehrere Schlüsselstrategien zur Lösung des Datenproblems bei Low-Resource-Sprachen entwickelt:

1. Sprachübergreifendes Lernen (Cross-Lingual Learning) und Transfer-Learning sind Ansätze, bei denen das von einem Modell in einer ressourcenreichen Sprache erworbene Wissen zur Verbesserung der Leistung in einer ressourcenarmen Sprache übertragen wird. Chen et al. (2025) präsentierten einen umfassenden Überblick über Strategien des Wissenstransfers in der sprachübergreifenden Sentiment-Analyse und betonten das Potenzial dieser Techniken zur Überwindung von Sprachbarrieren. Gladys & Vetriselvi (2024) untersuchten die Integration multimodalen Repräsentationslernens mit sprachübergreifendem Transfer-Learning und demonstrierten, wie multimodale Merkmale (Text, Bild, Audio) die Leistung auf Datensätzen von Low-Resource-Sprachen verbessern können.

2. Zero-Shot-Lernen (ohne spezifische Trainingsbeispiele) unter Verwendung multilingualer Lexika. Koto et al. (2024) schlugen einen Ansatz vor, der die Abhängigkeit von Texten in Low-Resource-Sprachen durch die Verwendung multilingualer Sentiment-Lexika (Wörterbücher mit expressiv markierten Wörtern in verschiedenen Sprachen) im Vortraining reduziert. Ihre Methode zeigte eine überlegene Zero-Shot- Leistung im Vergleich zu Modellen, die auf englischsprachigen Sentiment-Datensätzen nachtrainiert wurden, sowie zu großen Sprachmodellen (GPT-3.5, BLOOMZ, XGLM) für 34 Sprachen, darunter 25 Low-Resource-Sprachen.

3. Ansätze für Code-Switching-Inhalte – Aryal et al. (2022) schlugen einen mehrstufigen NLP-Algorithmus vor, der Code-Switching-Punkte in gemischtem Text (Stellen, an denen ein Sprachwechsel stattfindet) identifiziert und eine Sentiment-Analyse um diese Punkte herum durchführt. Ihr Algorithmus nutzt semantische Ähnlichkeit aus großen vortrainierten multilingualen Modellen und übertrifft das Basismodell um 11,2 % in der Accuracy und 11,64 % beim F1-Score auf einem spanisch-englischen Datensatz.

Sentiment-Analyse für die ukrainische Sprache. Die Erforschung der Sentiment-Analyse für die ukrainische Sprache begann mit lexikon-orientierten und regelbasierten Ansätzen. Romanyshyn (2013) präsentierte eine regelbasierte Methode zur Analyse ukrainischsprachiger Bewertungen, und Bobichev et al. (2017) untersuchten Sentiment-Trends in ukrainischen und russischen Nachrichtenartikeln unter Anwendung traditioneller Methoden des maschinellen Lernens (Naive Bayes, SVM).

Prytula (2024) verglich die Effektivität des Fine-Tunings der Modelle BERT, DistilBERT, XLM-RoBERTa und UkrRoBERTa für die binäre Klassifikation ukrainischsprachiger Bewertungen, wobei XLM-RoBERTa (Conneau et al., 2020) die besten Ergebnisse zeigte. Lashyn et al. (2025) wendeten Transformer-Architektur für die Dreiklassen-Klassifikation ukrainischsprachiger Texte an.

Am wichtigsten für die aktuelle Forschung ist die Arbeit von Shynkarov et al. (2025), die den COSMUS-Datensatz (COde-Switched MUltilingual Sentiment for Ukrainian Social media) vorstellten – das erste öffentlich verfügbare Korpus mit 12.224 Texten aus Telegram-Kanälen, Produktbewertungs-Websites und offenen Datensätzen. Die Texte sind nach zwei Parametern annotiert: Sentiment (positiv, negativ, neutral, gemischt) und Sprache (Ukrainisch, Russisch, Code-Switching bzw. Suržyk). Die Autoren führten ein Benchmarking dreier Ansätze durch: Few-Shot-Lernen mit GPT-4o und DeepSeek V2-chat, Fine-Tuning des multilingualen mBERT und des ukrainischzentrierten UkrRoberta. Das beste Ergebnis zeigte das feinjustierte UkrRoberta mit GPT-4o-gesteuerter Datenanreicherung (73,6 % Genauigkeit, F1-Score 0,64), das mBERT und Few-Shot GPT-4o übertraf.

Aufgabenstellung. Trotz bedeutender Fortschritte in der Sentiment-Analyse für die ukrainische Sprache fehlt noch immer eine vergleichende Bewertung der neuesten ukrainischzentrierten Architekturen für code-gemischte Inhalte. Die vorliegende Untersuchung zielt darauf ab, diese Lücke durch einen systematischen Vergleich von decoder-only LLM und encoder-basierten

Modellen für die Aufgabe der Multiclass-Sentiment-Analyse (4 Klassen) von Texten aus sozialen Netzwerken zu schließen. Für die Untersuchung wurden drei Modelltypen ausgewählt, von denen jeder spezifische Vorteile aufweist.

Erstens MamayLM 1.0 (basierend auf Gemma 3) – das erste ukrainischzentrierte LLM, das auf vielfältigen ukrainischsprachigen Daten trainiert wurde. Es ist geplant, verschiedene Strategien zu testen: Zero-Shot-Lernen (ohne Demonstrationsbeispiele), Few-Shot-Lernen (mit Beispielen) und Fine-Tuning mit LoRA/QLoRA-Methoden (Hu et al., 2021; Dettmers et al., 2023). Theoretisch sollte die Exposition gegenüber vielfältigen ukrainischsprachigen Daten während des Vortrainings eine bessere Anpassung an gemischtsprachige Texte gewährleisten. Zweitens Modern-LiBERTa (trainiert auf dem Kobza-Datensatz) – das modernste ukrainischzentrierte Encoder- Modell. Es bietet den Vorteil punktgenauer Fokussierung auf die ukrainische Sprache und einer effizienten Architektur für Klassifikationsaufgaben. Die Erfahrung von Shynkarov et al. (2025) zeigte, dass sprachspezifisches Vortraining (UkrRoberta) erhebliche Vorteile für die ukrainische Sprache bietet. Drittens XLM-RoBERTa – ein multilinguales Modell, das zuvor gute Ergebnisse für die binäre und Dreiklassen-Klassifikation des Sentiments ukrainischsprachiger Texte demonstrierte. Es wurde jedoch noch nicht auf dem Vierklassen-Schema von COSMUS mit der Klasse "gemischt" bewertet.

Es wird der standardmäßige Train/Validation/Test-Split des COSMUS-Datensatzes verwendet. Die Bewertung umfasst Accuracy, F1-Score, Precision und Recall auf verschiedenen sprachlichen Teilmengen zur Identifizierung potenzieller sprachlicher Verzerrungen. Es ist auch geplant, Methoden der erklärbaren künstlichen Intelligenz (Explainable AI: LIME, SHAP) zur Analyse lexikalischer und grammatischer Merkmale anzuwenden, die die Modelle zur Klassifikation verwenden, sowie eine Analyse der Modellkalibrierung (Expected Calibration Error) zur Bewertung der Zuverlässigkeit von Vorhersagen, wie dies Shynkarov et al. (2025) durchführten. Zukünftige Forschungen können die Sammlung zusätzlicher Daten aus Telegram und Reddit zur Erhöhung der

Repräsentanz von Suržyk umfassen sowie die Erweiterung der Aufgabe auf die Erkennung spezifischer Emotionen (Emotion Extraction) und ausgebaute Sarkasmuserkennung.

Schlussfolgerungen. Zur Lösung der Aufgabe der Multiclass-Sentiment-Analyse ukrainischsprachiger Inhalte aus sozialen Netzwerken wird vorgeschlagen, eine umfassende vergleichende Analyse dreier architektonisch unterschiedlicher Modelle (MamayLM, Modern-LiBERTa, XLM-RoBERTa) auf dem spezialisierten COSMUS-Datensatz durchzuführen. Die Untersuchung soll die Vorteile und Einschränkungen von Decoder-only-LLMs im Vergleich zu Encoder-Modellen aufzeigen sowie optimale Strategien zur Modellanpassung für ukrainischsprachige Texte identifizieren, einschließlich Fällen von Code-Switching.

Die praktische Bedeutung der Arbeit liegt in der Entwicklung effizienter Systeme zur Sentiment-Analyse ukrainischsprachiger Inhalte – ein wichtiges Instrument für die Überwachung der öffentlichen Meinung, die Erkennung von Desinformation und die Schaffung eines sichereren digitalen Raums. Diese Untersuchung verbindet Methoden der Computerlinguistik, des maschinellen Lernens und der Soziolinguistik, was ihren interdisziplinären Charakter unterstreicht.

Literatur

1. Aryal, S. K., Prioleau, H., & Washington, G. (2022). *Sentiment Classification of Code-Switched Text using Pre-trained Multilingual Embeddings and Segmentation*. ArXiv.org. https://arxiv.org/abs/2210.16461
2. Bobichev, V., Kanishcheva, O., & Cherednichenko, O. (2017, May 1). *Sentiment analysis in the Ukrainian and Russian news*. IEEE Xplore. https://doi.org/10.1109/UKRCON.2017.8100410
3. Chen, L., Shang, S., & Wang, Y. (2025). Bridging resource gaps in cross-lingual sentiment analysis: adaptive self-alignment with data

augmentation and transfer learning. *PeerJ Computer Science*, *11*, e2851. https://doi.org/10.7717/peerj-cs.2851

4. 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. *ArXiv:1911.02116 [Cs]*. https://arxiv.org/abs/1911.02116
5. Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). *QLoRA: Efficient Finetuning of Quantized LLMs*. https://doi.org/10.48550/arxiv.2305.14314
6. Gladys, A.A., & Vetriselvi, V. (2024). Sentiment analysis on a low-resource language dataset using multimodal representation learning and cross-lingual transfer learning. *Applied Soft Computing*, *157*, 111553. https://doi.org/10.1016/j.asoc.2024.111553
7. Haltiuk, M., & Smywiński-Pohl, A. (2025). On the Path to Make Ukrainian a High-Resource Language. *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, 120–130. https://doi.org/10.18653/v1/2025.unlp-1.14
8. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. *ArXiv:2106.09685 [Cs]*. https://arxiv.org/abs/2106.09685
9. Koto, F., Beck, T., Talat, Z., Gurevych, I., & Baldwin, T. (2024). *Zero- shot Sentiment Analysis in Low-Resource Languages Using a Multilingual Sentiment Lexicon*. ArXiv.org. https://arxiv.org/abs/2402.02113
10. Lashyn, Y., Trofymchuk, O., Zabolotnyi, S., Voitko, O., & Seabra, E. (2025). SENTIMENT ANALYSIS OF TEXTS USING RECURRENT NEURAL NETWORKS OF THE TRANSFORMER ARCHITECTURE. *Advanced Information Systems*, *9*(3), 91–101. https://doi.org/10.20998/2522-9052.2025.3.11

11. Prytula, M. (2024). *Fine-tuning BERT, DistilBERT, XLM-RoBERTa and Ukr-RoBERTa models for sentiment analysis of ukrainian language reviews*. Jai.in.ua. https://jai.in.ua/index.php/en/issues?paper_num=1623
12. Romanyshyn, M. (2013). Rule-Based Sentiment Analysis of Ukrainian Reviews. *International Journal of Artificial Intelligence & Applications*, *4*(4), 103–111. https://doi.org/10.5121/ijaia.2013.4410
13. Shynkarov, Y., Solopova, V., & Schmitt, V. (2025). Improving Sentiment Analysis for Ukrainian Social Media Code-Switching Data. *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, 179–193. https://doi.org/10.18653/v1/2025.unlp-1.18