Technical sciences

UDC 005.160

Izhevska Yelyzaveta
Student of the
Taras Shevchenko National University of Kyiv

Haina Heorhiy
Candidate of Technical Sciences,
Associate Professor of the Department of Intellectual Technologies
Taras Shevchenko National University of Kyiv

CROSS-LINGUAL SENTIMENT ANALYSIS FOR UKRAINIAN SOCIAL MEDIA SENTIMENT CLASSIFICATION

Summary. *Understanding the sentiment derived from user-generated content on social media platforms plays a crucial part in today's digital world. However, multilingual datasets create a challenging landscape for sentiment analysis, especially when two or more languages are used interchangeably. This research explores the use of cross-lingual sentiment analysis on Ukrainian social media platforms, where content is expressed in either Ukrainian or russian languages or a mix of them. Using a combination of machine learning and natural language processing techniques, this study established an effective model for cross-lingual sentiment classification.*

Key words: *cross-lingual sentiment analysis, opinion mining, transfer learning, machine learning, word embedding.*

Analyzing sentiments from social media platforms is a crucial endeavor in today's data-driven world. It enlightens businesses, governments, and research institutions about public opinion, helping shape their strategies and operations

appropriately. However, accurately detecting and classifying sentiments can become overwhelmingly challenging when dealing with multilingual data, where different languages are interchangeably used in the same landscape.

A key showcase of such linguistic interchangeability is observed also in Ukraine, where Ukrainian and russian languages, unfortunately, can be combined on social media platforms. It brings unique complexities into sentiment analysis, as traditional sentiment analysis methodologies, developed predominantly for monolingual datasets, struggle to offer reliable results in such cross-lingual cases. Grasping the nuances and sentiment-related contexts of different languages requires an enriched understanding of the languages concerned and a formidable model that can handle the said complexities effectively.

Owing to the burgeoning needs of the digital age, the adaptation to cross-lingual sentiment analysis methodologies has become imperative. These advanced methods, often based on machine learning and natural language processing algorithms, have shown a promising edge in the landscape of sentiment analysis within multilingual data. The primary objective of this research essentially lies in exploring and crafting an effective cross-lingual sentiment analysis model, specifically catered to handle Ukrainian sentiments on social media platforms.

One of the ways for Ukrainian and Russian sentiment analysis is by leveraging cross-lingual word embeddings and machine learning algorithms. The data, like posts and comments, can be collected from various Ukrainian social media platforms, which feature content in both Ukrainian and russian languages.

The dataset includes both Ukrainian and russian social media posts. It should be preprocessed by removing noise such as URLs, special characters, and numbers. After that, texts go through tokenization, which involves breaking down the text into individual words or tokens. A morphological analysis must also be performed for both languages to bring the words to their basic form.

Cross-lingual word embeddings are a powerful tool in overcoming language barriers in NLP. At the same time, the Supervised Cross-Lingual Word Embedding (CLWE) model relies on many bilingual parallel texts. Most existing CLSA models only cover rough sentiment analysis, such as sentence-level or document-level sentiment analysis [1]. They map equivalent words from related languages to a similar region in the vector space. This helps in conveying semantic and syntactic similarities across languages. Given the close linguistic relations between the two languages, such embeddings can be pretty effective in the context of studied languages.

For example, a Supervised Word Translation includes a linear mapping function that learns to project word embeddings from one language to the source language space, preserving semantic and syntactical information [2].

Mathematically, if E_1 and E_2 are the word embeddings of language $L_1$and L_2, our goal is to learn a transformation matrix W that can map E_1 to E_2 or vice versa. If $E_1$is a word in language L_1, and $E_2$is its equivalent in language L_2, the transformation can be expressed as:

$$E_2 = W * E_1 ... \quad (1)$$

The transformation matrix W can be learned by minimizing the mean squared error in equation between the predicted E_2 and actual E_2 over all words in our training set.

Cross-lingual word embeddings can be used as input to Machine Learning algorithms for sentiment classification. Different models can be applied, including Naive Bayes, Support Vector Machines, and ensemble methods, and the best performance is generally obtained using ensemble methods. Ensemble methods combine the decisions from multiple models to improve the overall performance [3].

In the context of an Ensemble Model with n base models, if we denote the prediction of each base model as P_1, P_2, ..., P_n, and the ultimate ensemble model prediction as P, the ensemble can be represented as:

$$P = f(P_1, P_2, ..., P_2)... \quad (2)$$

In this ensemble method, a majority voting system can be used where the final sentiment is determined by the majority of the base models' sentiment predictions

Model performance can be evaluated using common matrix like accuracy, precision, recall, and F1-score, in addition to confusion matrices for a detailed view of the model's proficiency to classify the sentiments correctly.

In conclusion, this research establishes the potential for cross-lingual sentiment analysis to assess sentiments in multilingual contexts, crucial in countries like Ukraine, where multiple languages are used in the social media landscape. Though this approach to sentiment classification compared to traditional monolingual models can significantly improve, some issues, like the accurate identification of sarcasm or neutrality, still will be difficult for the development. It indicates the need for further research and refinement better to leverage the cross-lingual model for Ukrainian sentiment analysis.

References

1. Yuemei Xu, Han Cao, Wanze Du, Wenqing Wang. A Survey of Cross-lingual Sentiment Analysis: Methodologies, Models and Evaluations. *Data Science and Engineering.* 2022. Vol. 7. P. 287-289. URL: https://link.springer.com/article/10.1007/s41019-022-00187-3 (date of access: 09.04.2024)

2. Conneau A. Word translation without parallel data. 2017. 8 p. URL: https://arxiv.org/format/1710.04087 (date of access: 10.04.2024).

3. Mikolov T., Le Quoc V., Sutskever I. Exploiting similarities among languages for machine translation. 2013. URL: https://arxiv.org/abs/1309.4168 (date of access: 10.04.2024).