UDC 004.852

V. Oliinyk, I. Matviichuk

LOW-RESOURCE TEXT CLASSIFICATION USING CROSS-LINGUAL MODELS FOR BULLYING DETECTION IN THE UKRAINIAN LANGUAGE

Annotation: This paper aims on building bullying detection model for Ukrainian language. Considering absence of labeled datasets for bullying detection and classification in Ukrainian, small Ukrainian dataset (4k samples) was gathered and used for testing models in this research. Taking into account very small number of Ukrainian datasets in general this dataset is publicly available for testing and benchmarking other text classification models.

Modern approaches to text class classification in low-resource languages are studied in the paper. We apply zero-shot technique and evaluate performance of modern multilingual, cross-lingual state-of-the-art models and embeddings for text classification in Ukrainian language, including mBERT, XLM-R, LASER and MUSE. Experimental results shows that zero-shot approaches for classification task allow to achieve F1 score of 67-69% for multilingual models trained on English dataset only, having 88-91% test accuracy on English data. We also show that machine translation of English data can be used for estimating model performance in other languages, i.e. only 0-2% difference in test accuracy compared to natural data was received for best models XLM-R and LASER. Zero-shot approach for binary detection task showed even better results 81% compared to 91,59% on original English data.

We then enhance the best XLM-R model by training it on our natural Ukrainian dataset and confirm benefits of augmenting low-resource language dataset with machine translations from resource-rich English data.

Finally, the model for bullying detection in the Ukrainian language is built achieving F1 score of 91,59% with only 12k samples dataset in different languages.

Keywords: multilingual models, zero-shot classification, bullying detection, XLM-RoBERTa, mBERT, LASER, MUSE.

Introduction

As of January 2022, the number of Internet users worldwide reached almost five billion, accounting for 62.5% of the global population, and this figure is expected to continue to grow [1]. Social networks and other tools for online communication have become an essential part of our lives bringing new possibilities and threats. For instance, 88% of teenagers who use social networks have witnessed angry or cruel behaviour by other users [2]. The anonymity provided by the Internet often leads people to write things to others that they would never say to their faces. Given that people spend an average of 7 hours per day on the Internet [1],

such interactions can be detrimental to mental health. According to recent studies, 50% of individuals who have experienced cyberbullying in the past 12 months have reported feeling depressed, 45% reported feeling anxious, and 34% had suicidal thoughts [3].

Training machine learning models requires vast amounts of data, and in the case of languages such as English, this is not a challenge. There are numerous datasets available in English, including those for bullying and hate speech. However, the same cannot be said for low-resource languages. These languages lack the necessary data to train conversational AI models, which is a significant challenge. In practice, detecting hate speech and other similar tasks for low-resource languages is just as essential as for high-resource languages. Nonetheless, such datasets may be scarce or absent entirely for low-resource languages, posing significant challenges for training machine learning models for these languages.

Generating a dataset for machine learning models is a time-consuming and costly process that involves data collection and labeling. Transfer learning has emerged as an effective approach to training models using relatively small amounts of data while achieving good metrics. However, even small datasets are often unavailable for low-resource languages. In such cases, the zero-shot training approach becomes a powerful tool for training models without using any data in the target language during the training phase.

This study aimed to create models that detect and classify bullying in the Ukrainian language based on six available English datasets and a small Ukrainian dataset collected by authors. We checked effectiveness of modern techniques for low-resource data including zero-shot learning, data augmentation technique, transfer-learning as well as state-of-the art language-agnostic embeddings and models for cross-lingual representation to find effective model and concept for such tasks

Related Works

The majority of research on the detection of hate speech, offensive language, or bullying has primarily been conducted in English, largely due to the availability of numerous pre-existing datasets in this language.

Several studies have used simpler machine learning techniques [4, 5, 6]. However, with the emergence of transformer models, such as BERT [7], these approaches have become increasingly widespread, and have been applied to a variety of tasks, including hate speech detection [8].

Despite significant advancements in the detection of hate speech, the challenge of detecting such speech in low-resource languages remains problematic. However, the development of multilingual models, such as multilingual BERT (mBERT) [9] that exhibits cross-lingual generalization, has partially addressed this issue. Although the high lexical overlap between languages contributes to improved transfer, mBERT has also exhibited the ability to

transfer between languages written in distinct scripts with no lexical overlap, suggesting that it can effectively capture multilingual representations.

Subsequent studies have led to the development of cross-lingual models, such as XLM-R, which have been shown to outperform mBERT in cross-lingual classification for low-resource languages [10].

Meanwhile, improvements to cross-lingual learning techniques based on multilingual embeddings have also been made. Conneau et al. (2018) presented a method for creating multilingual embeddings that are aligned in a common space [11]. This approach has demonstrated remarkable performance even for distant low-resource language pairs with limited parallel data available.

Artetxe et al. (2019) proposed an architecture for learning joint multilingual sentence representations for 93 languages based on a classifier on top of the resulting embeddings by utilising solely English-annotated data and subsequently transferring it to any of the supported 93 languages without any modifications [12].

In a study by Tanase et al. (2020), pre-trained transformer-based models, such as BETO, XLM-R, and mBERT, were applied for detecting aggressiveness in Mexican Spanish social media. The models were trained on various task-specific datasets, including a dataset created with machine translation [13]. Benefits of machine translation for augmenting small text datasets was also shown in [14].

In another study by Pant et al. (2020), a cross-lingual inductive approach was introduced for identifying the offensive language in tweets using the contextual word embedding XLM-RoBERTa. The created model exhibited competitive performance for five languages and worked in a zero-shot learning approach [15].

Finally, El-Alami et al. (2021) presented a study exploring multilingual offensive language detection based on both monolingual (BERT, AraBERT) and multilingual models (mBERT) [16]. The study also tackled the multilingualism problem using joint-multilingual and translation-based methods.

Predictably, we found no solutions for Ukrainian language as well as no comparative analysis of effectiveness of various models and techniques for low resource datasets.

Dataset

There is no public dataset for bullying classification available in Ukrainian, which was one of the main reasons for this paper.

English dataset

This research aims both detection and classification of offense speech and bullying therefore we required a dataset labeled with typical classes of hate speech, including ‘sexism’, ‘racism’, ‘homophobia’, ‘ableism’, and ‘none’. Due to the lack of a comprehensive dataset con-

taining all necessary classes we utilize six publicly available datasets in English for this study. The datasets consist of varying classes, with the ‘ableism’ class being represented in only two datasets, whereas the ‘sexism’ and ‘racism’ classes are present in all six datasets. Detailed information about each dataset, including their source and size, can be found in Table 1.

Datasets with the classes that match those listed above were included in the resultant dataset without changes while some of them required preprocessing. Some dataset-specific classes were either renamed and combined (i.e. ‘MUSLIMS’, ‘MIGRANTS’ from ‘Multitarget CONAN’ dataset [17] was moved to generic ‘racism’ class) or removed, i.e. ‘other’ class. In addition, we decided to move ‘religious discrimination’ to ‘racism’ class. Actual label for each class may be found in parentheses in Table 1. The ‘ETHOS’ dataset is multi-labelled, and only data samples with a 50% or higher probability for one label were included in the resultant dataset for this study.

Table 1.

Datasets details

Name of dataset	Source	Distribution of classes	Total number of data samples
Waseem [5]	Twitter	none - 7661 sexism - 2675 racism - 9	10345
Cyberbullying datasets [18]	Twitter	none - 11501 racism - 1970 sexism - 3377	16848
ETHOS [19]	Twitter	racism- 137 sexism - 83 religious_discrimination (racism)- 77 homophobia - 68 ableism - 52	417
Detect Hate Speech [20]	Twitter	none – 16844 sexism - 3963 homophobia - 87	20894
MLMA hate speech [21]	Twitter	origin (racism) - 2448 disability (ableism) - 1089 other (not used) - 890 gender (sexism) - 638 sexual_orientation (homophobia) - 514 religion (racism) - 68	5647
Multitarget CONAN [17]	Written by experts	MUSLIMS (racism)- 1335 MIGRANTS (racism)- 957 WOMEN (sexism)- 662 LGBT+ (homophobia) - 617 JEWS (racism)- 594 POC (racism)- 352 DISABLED (ableism)- 220 other (not used)- 266	5003

The resulting dataset [22], obtained by merging the previously described datasets, was unbalanced, with a disproportionately large number of samples labelled as ‘none’ and a very limited amount of data in classes such as ‘ableism’ and ‘homophobia’. A downsampling technique was applied to the dataset to address this issue, reducing the total number of samples and improving class balance. Consequently, the final dataset contained a total of 5500 samples, with 1100 samples in each class. The number of samples per class was chosen based on the class with the smallest number of samples to ensure balance across all classes. This approach resulted in an improved dataset that was well-suited for effective model training and evaluation.

Ukrainian dataset

There is currently no publicly available Ukrainian-language dataset that contains the necessary classes, or even one that is focused on a similar topic. However, as the Ukrainian-language dataset will only be used for testing, it may be smaller in size. To address this, we created the Ukrainian-language dataset using two methods:

1. The first part of the dataset comprises randomly selected samples from the English-language dataset, which were then translated into Ukrainian and verified by human translators, as opposed to machine translation.
2. The second part of the dataset was collected through a form in which students provided examples of tweets for each class.

The Ukrainian-language dataset (ULD) contains a total of 2,500 tweets, with 500 samples in each class.

Ukrainian machine translation dataset

To evaluate the quality of our collected Ukrainian-language dataset and to explore the possibility of using machine translation for similar tasks, we also created a machine-translated Ukrainian-language dataset (MTULD) using the Google Cloud Translation API.

For this dataset, we selected 500 random samples from each class of the English-language dataset, which were then translated using the Google Cloud Translation API. In total, the resulting dataset contains 2,500 tweets, with 500 samples in each class.

Zero-shot approach to classification of bullying type

Recent works show that modern multilingual models can be used in zero-shot learning approach for other languages. Therefore, in this research we test this technique for bullying classification task in Ukrainian language and evaluate performance of modern multilingual models and embeddings in a ‘no local data’ scenario.

The first essential part of any NLP task is creation of word embeddings [23]. We chose two modern multilingual word embeddings to study their effectiveness compared to special multilingual models:

LASER (Language-Agnostic SEntence Representations) are multilingual sentence embeddings trained in 147 languages including Ukrainian [12].

MUSE (Multilingual Unsupervised and Supervised Embeddings) are multilingual word embeddings with large-scale high-quality bilingual dictionaries in 30 languages, including Ukrainian, for training and evaluation. MUSE embeddings align word embedding spaces without any cross-lingual supervision, i.e. solely based on unaligned datasets for each language [11].

Models for multilingual zero-shot text classification

Finally, we created six models for the experiment: three special BERT based models and three models based on multilingual word embeddings described earlier:

1. **BERT** (Bidirectional Encoder Representations from Transformers) is a machine-learning model designed to pre-train deep bidirectional representations from the unlabeled text by jointly conditioning on both the left and right context in all layers [7]. The pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks [7].

In the experiments, a small version of BERT was used. It includes 4 encoder layers with a size of 512 and 8 attention heads. The attention head outputs are concatenated and passed through a dropout layer that deactivates 50% of the neurons. The output layer is fully connected with the size of 5 neurons, utilizing the softmax activation function.

It is important to note that BERT is exclusively trained in the English language and is employed in the experiments solely to evaluate the performance of other models on an English dataset.

2. **LASER+MLP**. The model converts sentences into 1024-dimensional vectors utilizing LASER embeddings. The embeddings are then passed through a multi-layer perceptron [24], comprising three layers with 512, 128, and 32 neurons, respectively, each utilizing a ReLU activation function. Following each fully connected layer, a dropout layer is applied, deactivating 50% of the neurons. The output layer of the model is also fully connected, with a size of 5 neurons and utilizes the softmax activation function.

3. **MUSE+MLP**. The model transforms sentences into 300-dimensional vectors using MUSE embeddings. These embeddings are then passed through a multilayer perceptron consisting of four layers, with 512, 128, 64, and 32 neurons, respectively, each utilizing a ReLU activation function. Following each fully connected layer, a dropout layer is employed, deactivating 50% of the neurons. The output layer of the model is fully connected with 5 neurons and employs the softmax activation function.

4. **MUSE+CNN.** The model transforms each word in the sentence into 300-dimensional vectors using MUSE embeddings. The sentences are truncated to a maximum of 200 words and 300x100 vectors are generated for each sentence. These embeddings are passed through three parallel one-dimensional convolutional layers, each containing 100 filters and using steps 2, 3, and 4, respectively. All three layers employ the ReLU activation

function, followed by a one-dimensional global max pooling layer. The outputs of the three convolutional layers are concatenated, followed by a dropout layer that deactivates 50% of the neurons. The output layer of the model is fully connected, containing 5 neurons, and utilizes the softmax activation function.

5. **mBERT** (Multilingual BERT) is a multilingual version of the BERT model that was pre-trained on large amounts of multilingual text data from Wikipedia, resulting in a model that supports 104 languages, including Ukrainian. By using a shared multilingual vocabulary and training in multiple languages, mBERT is able to learn a common representation among different languages, which allows it to be used for cross-language training without requiring target language samples. However, it should be noted that the model was not specifically trained to have common representations across languages, and the quality of its representations may vary across languages.

Like the original BERT model, mBERT is comprised of 12 encoder layers each with a size of 768, as well as 12 attention heads. The outputs of the attention heads are merged and fed through a dropout layer, which excludes 50% of the neurons. The output layer is fully connected and consists of 5 neurons, which are activated using the softmax function.

6. XLM-R (XLM-RoBERTa) is an enhanced version of the original model, XLM-100 (supports over 100 languages, including Ukrainian), which incorporates transformers and utilizes the masked-language modeling approach. XLM-R enables the utilization of a single large model for all languages, without sacrificing per-language performance [10].

Experiment setup

Experiments for this research were conducted in the Google Colab environment using Tensorflow and Keras libraries. All models were trained with Adam optimizer, batch size of 32 and categorical cross-entropy loss function. The values of learning rate and number of training epochs were chosen individually for each model.

In this scenario, we trained all models on ‘English dataset’ that contains five classes: 'none', 'ableism', 'homophobia', 'racism', and 'sexism' and is described above in Dataset section. This dataset was split into training, validation, and testing part in a ratio of 70%, 20%, and 10%, respectively.

After training, we evaluated performance of the models using data of all types: the English dataset, the Ukrainian dataset, and the Ukrainian machine translation dataset.

We selected the F1-score metric for each class individually and the macro F1-score to represent the model's overall performance treating all classes equally:

$$F1macro = \frac{F_1 + F_2 + \ldots + F_n}{n} \quad (1)$$

Each experiment was conducted three times for each model, and the arithmetic mean of the F1 score metric was calculated from the results to get a more robust evaluation performance.

Experimental results

Table 2 presents the results of multiclass classification, that is, bullying classification. The rows indicate the dataset on which the models were tested. The 'BERT' column shows the result of monolingual BERT trained on the English dataset, which is used as a benchmark to compare the results of multilingual models. The column names indicate the model for which the results are presented. Each model in this table was trained on an English dataset and tested on a Ukrainian dataset and a Ukrainian machine translation dataset.

Table 2.

Results of multiclass classification experiments

Test dataset	BERT	LASER	mBERT	MUSE + CNN	MUSE + MLP	XLM-R
English	92	88	92	87	80	91
Ukrainian	-	69	56	68	53	67
Ukrainian machine translation	-	67	56	58	49	67

As shown in Table 2, the LASER and XLM-R models exhibit superior performance overall and for natural Ukrainian data in particular.

Specifically, the XLM-R model achieved an F1-score of 91% for English, 67% for Ukrainian, and 67% for Ukrainian machine translation, which represents the highest result among all models. High performance was expected for XLM-R, which is specifically designed for cross-lingual learning, encompassing zero-shot learning. The model is pre-trained on an extensive range of languages with massive amounts of data, which significantly enhances its metrics for zero-shot training and training on a small English-language dataset.

The LASER model, on the other hand, scored 88% F1-score for English, 69% for Ukrainian (the highest score among all models), and 67% for Ukrainian machine translation, which is also the best result obtained by any model. Unlike XLM-R, this model was not pre-trained on vast amounts of data, which could have adversely impacted the results. However, owing to the LASER embeddings, the model still exhibits substantial performance despite a relatively small training dataset. LASER embeddings are built as universal language-agnostic sentence embeddings and have vector representations of sentences that are general concerning two dimensions: the input language and the NLP task [12]. This approach ensures a high level of generalization across languages.

The MUSE+CNN model demonstrates one of the most favourable outcomes for the Ukrainian dataset, with an F1-score of 68%, but poor results on the Ukrainian machine translation dataset, with an F1-score of 58%, which suggests its susceptibility to the translated data.

 ISSN 1560-8956

The mBERT model exhibits the best outcome for the English-language dataset, with an F1-score of 92%, but poor 56% for both Ukrainian datasets. It is completely outperformed by more novel and advanced XLM-R that is specially designed for cross-lingual learning while mBERT encodings for various languages are not aligned in a common space.

The MUSE+MLP model demonstrated the poorest performance for all datasets.

In general MUSE embeddings were outperformed by LASER. One possible explanation for this difference could be attributed to the fact that MUSE is embeddings for words rather than sentences, as is the case with LASER. This difference may have negatively impacted the models' understanding of sentence meaning and made them more sensitive to translations.

Another important result is small difference in accuracy between natural and translated date. Therefore, translated data may be used for estimating the performance of the models in the zero-shot approach.

Our next step was to compare confusion matrices of the models. Fig. 1, 2 and 3 illustrate that most of LASER, XLM-R, MUSE+CNN errors on natural Ukrainian dataset are false classification of none-bullying samples. In real-world scenarios, we will avoid combining detection and classification tasks and this class is the one where we can get as match data as we need to improve this error. Fine-tuning of the models was out of the scope of this research but we discovered an easy option to do it in future.

The mBERT and MUSE+MLP models (Fig. 4 and 5) have errors that are more diverse and demonstrate lower results in general compared to previous models.

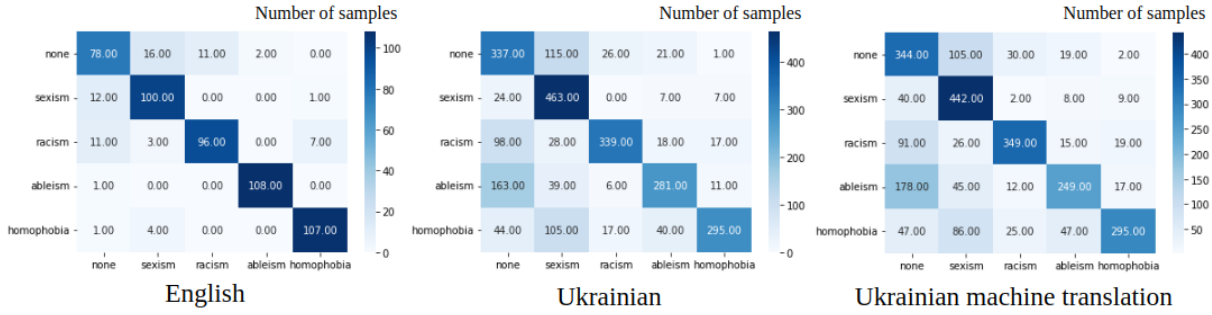

Figure 1. Confusion matrices for LASER

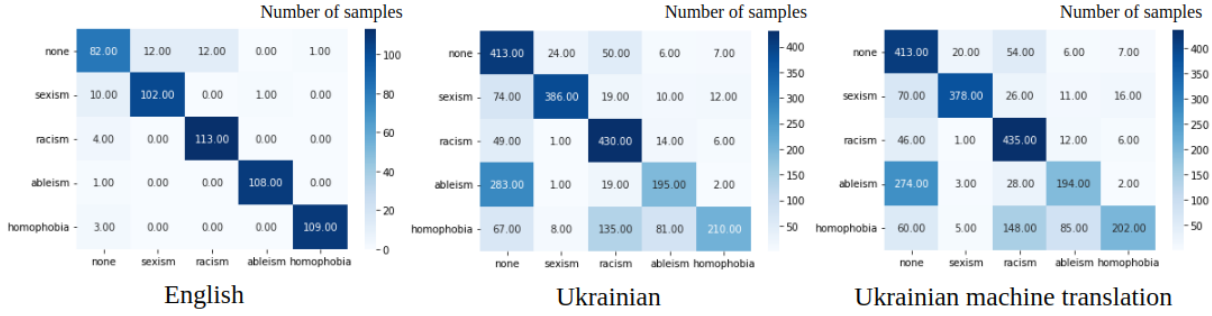

Figure 2. Confusion matrices for XLM-R

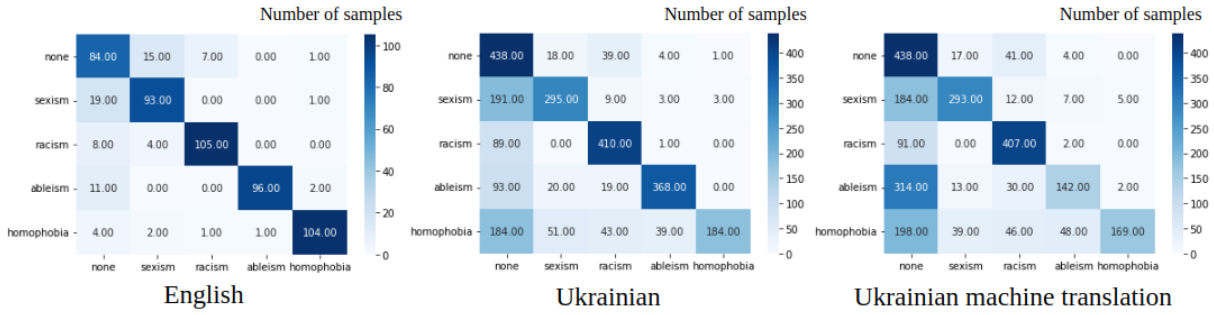

Figure 3. Confusion matrices for MUSE+CNN

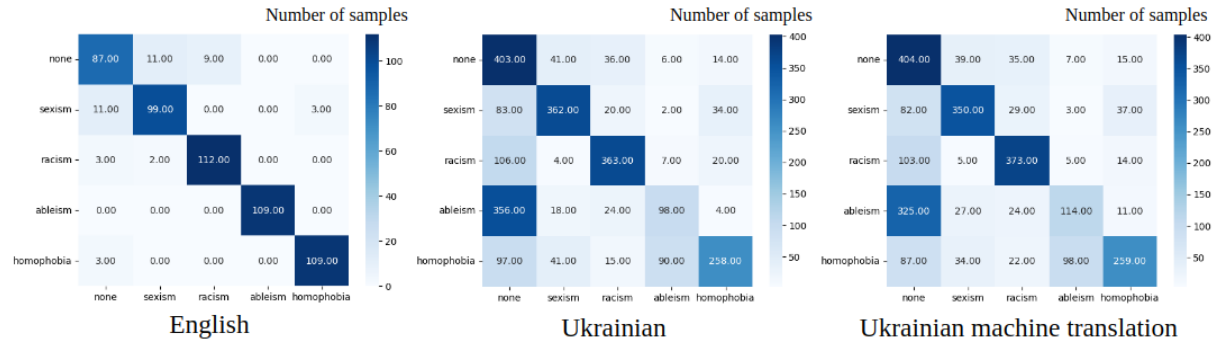

Figure 4. Confusion matrices for mBERT

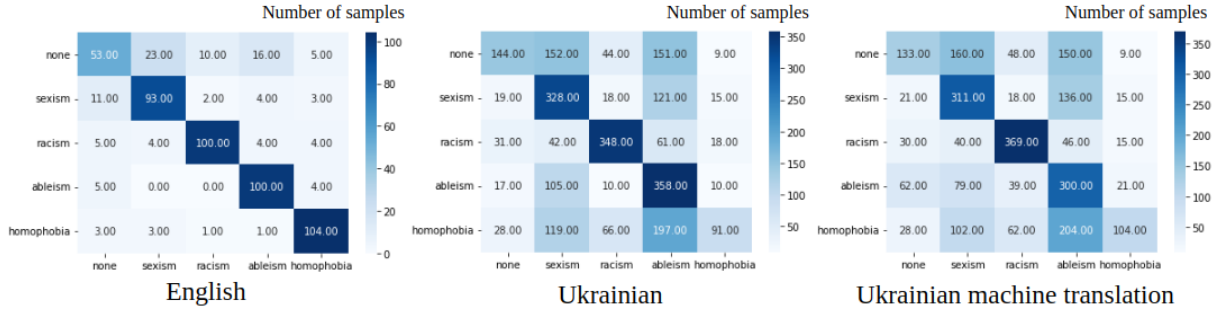

Figure 5. Confusion matrices for MUSE+MLP

Building a model for Bullying detection

To facilitate the task of detecting bullying through binary classification, three separate datasets were prepared in this study (English, Ukrainian, and Ukrainian machine translation) by consolidating data samples from all classes except ‘none’ into a single ‘aggression’ class. To balance the number of data samples in the ‘none’ class, this class was augmented with other tweets with the same statistics to match the number of data samples in the 'aggression' class.

As a result, the following binary datasets were obtained:

1. English - 8,800 data samples.
2. Ukrainian - 4,000 data samples.
3. Ukrainian machine translation - 8,800 data samples.

 ISSN 1560-8956

Based on the results of classification experiment we chose XLM-R as a base model for further fine-tuning by supervised training on available datasets. Monolingual BERT trained on the English dataset was used as a baseline for comparing the results of multilingual models trained on different datasets or their combinations.

Table 3 presents the results of binary classification, specifically bullying detection, with the rows representing the datasets on which the models were evaluated and columns representing the datasets on which the models were trained. The table presents the F1 metrics. Underlined and bold values correspond to the matching language of test and training dataset.

Table 3.

Results of binary classification experiments

Test data-set	BERT	XLM-R				
	English	**English**	**Ukrainian (4k samples)**	**Translation (4k samples)**	**Translation (8k samples)**	**Ukrainian + Translation**
English	**94,05**	**93,43**	-	-	-	-
Ukrainian	-	80,91	**87,54**	86,72	91,35	**91,59**
Ukrainian machine translation	-	79,92	85,08	**87,34**	**92,56**	92,83

First, we check zero-shot learning approach (“English” column and “Ukrainian” rows). XLM-R model achieves F1 score of 80.91% that is far from perfect 94,05% of specially trained monolingual model but can already satisfy some aggression detection tasks in Ukrainian texts.

We trained the same multilingual model on the natural Ukrainian-language dataset described earlier, which was smaller with only 4000 samples, and received an F1 score of 87.54% for the Ukrainian language. This accuracy drops quickly when decreasing number of samples for training so direct training on natural language data shows high results but requires many samples we do not have in low-resource languages.

In order to check the impact of dataset volume we trained two models were on machine translation datasets (4k and 8k samples) as we do not have enough natural data. As expected, F1 score for the same model and same sample quantity was a bit lower for translated dataset compared to natural (86.72% to 87.54%) but greatly outperformed natural Ukrainian when the quantity of training samples was doubled (91.35% to 87.54%).

These results shows importance of adding more training data for model performance even if translated data only is used. Small difference between test metrics on translated and natural data confirms closeness of their distributions and that machine translation can be effectively used for augmenting training dataset, as was also shown in [14].

As expected, the model trained on the combined dataset consisting of natural Ukrainian language and machine translation achieved the best result, with an F1 score of 91.59% for Ukrainian language.

Conclusion

This research showed that cross-lingual zero-shot learning approach might be successfully used for binary and multiclass classification in scenarios without any data in target language as was demonstrated in bullying classification task tweets written in the Ukrainian language.

Obviously, received results are lower compared to those obtained by models trained or fine-tuned on dataset in the specific language, but may still be useful. Additionally, it is feasible to use machine translation as a test dataset, providing an opportunity for quick evaluation of various models trained on resource-rich English data.

The XLM-R and LASER-based model demonstrated best results among other tested models. Moreover, we did not fine-tune those models and found obvious directions for this. We also plan to further improve performance of zero-shot approach by utilizing large language models that become available these days.

Another important result of the research is a new dataset in low-resource Ukrainian language that we used for evaluating our models in bullying detection task. It will be used for benchmarking all other users and will be publicly available.

In binary classification task besides zero-shot approach we studied the impact of augmenting natural training dataset with machine translations English and confirmed its positive effect. Indeed, for low-resource languages accuracy of classification rises with increase of samples quantity regardless of its origin. So augmenting the dataset by combining machine translation-generated data with natural language datasets can be recommended in general.

Our final detection model showed good metrics of 91.59% and is suitable for real-world applications, particularly in the context of detecting instances of bullying in the Ukrainian language on the Twitter social network etc.

REFERENCES

1. Digital 2022: Global Overview Report – DataReportal – Global Digital Insights. DataReportal – Global Digital Insights. URL: https://datareportal.com/reports/digital-2022-global-overview-report (date of access: 31.03.2023).

2. Teens, kindness and cruelty on social network sites. Pew Research Center: Internet, Science & Tech. URL: https://www.pewresearch.org/internet/2011/11/09/teens-kindness-and-cruelty-on-social-network-sites/ (date of access: 31.03.2023).

3. The Annual Bullying Survey 2018. Ditch the Label. URL: https://www.ditchthelabel.org/research-papers/the-annual-bullying-survey-2018/ (date of access: 31.03.2023).

4. Automated Hate Speech Detection and the Problem of Offensive Language / T. Davidson та ін. // Eleventh International AAAI Conference on Web and Social Media. Montreal, 2017. С.512-515. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14955/14805

5. Waseem Z., Hovy D. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter // NAACL-HLT. San Diego, 2016. С.88-93. URL: https://aclanthology.org/N16-2013.pdf

6. Hate Speech Dataset from a White Supremacy Forum / O. de Gibert та ін. // Second Workshop on Abusive Language Online. Brussels, 2018. С.11-20. URL: https://aclanthology.org/W18-5102.pdf

7. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin та ін. // NAACL-HLT. Minneapolis, 2019. С.4171-4186. URL: https://aclanthology.org/N19-1423.pdf

8. Mozafari M., Farahbakhsh R., Crespi N. A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media // Complex Networks and Their Applications VIII. Lisbon, 2019. С.928-940. URL: https://arxiv.org/pdf/1910.12574.pdf

9. Pires T., Schlinger E., Garrette D. How Multilingual is Multilingual BERT? // 57th Annual Meeting of the Association for Computational Linguistics. Florence, 2019. С.4996-5001. URL: http://aclanthology.lst.uni-saarland.de/P19-1493.pdf

10. Unsupervised Cross-lingual Representation Learning at Scale / N. Goyal та ін. // 58th Annual Meeting of the Association for Computational Linguistics. Online, 2020. С.8440-8451. URL: https://aclanthology.org/2020.acl-main.747.pdf

11. Word translation without parallel data / A. Conneau та ін. // International Conference on Learning Representations. Vancouver, 2018. С.1-14. URL: https://arxiv.org/pdf/1710.04087.pdf

12. Artetxe M., Schwenk H. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond // Transactions of the Association for Computational Linguistics. 2019. № 7. С.597-610 URL: https://aclanthology.org/Q19-1038.pdf

13. Detecting Aggressiveness in Mexican Spanish Social Media Content by Fine-Tuning Transformer-Based Models / M. Tanase та ін. // Iberian Languages Evaluation Forum. Málaga, 2020. С.236-245. URL: https://ceur-ws.org/Vol-2664/mexa3t_paper1.pdf

14. Oliinyk V. Data augmentation with foreign language content in text classification using machine learning / V. Oliinyk, K. Osadcha // Adaptive systems of automatic control, 2020. Vol. 1, №36. – P. 51-59.

15. Pant P., Dadu T. Cross-lingual Inductive Transfer to Detect Offensive Language // Fourteenth Workshop on Semantic Evaluation. Barcelona, 2020. С.2183-2189. URL: https://aclanthology.org/2020.semeval-1.290.pdf

16. El-Alami F., Ouatik El Alaoui S., En Nahnahi N. A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model //

Journal of King Saud University - Computer and Information Sciences. 2022. № 34. С.6048-6056 URL: https://reader.elsevier.com/reader/sd/pii/S1319157821001804?token=400DBBEFDEB3C197C92C7220F40176C5D6E7BAB85578FD1B27E72D9BFB24B397E250E1843203F41A492475C14D38FADC&originRegion=eu-west-1&originCreation=20230304173708

17. Human-in-the-Loop for Data Collection: a Multi-Target Counter Narrative Dataset to Fight Online Hate Speech / M. Fanton та ін. // 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Online, 2021. С.3226-3240. URL: https://aclanthology.org/2021.acl-long.250.pdf

18. Mendeley Data. URL: https://data.mendeley.com/datasets/jf4pzyvnpj/1 (date of access: 31.03.2023).

19. ETHOS: a multi-label hate speech detection dataset / I. Mollas та ін. // Complex & Intelligent Systems. 2022. № 8. С.4663-4678 URL: https://link.springer.com/epdf/10.1007/s40747-021-00608-2?sharing_token=vAEM1mW2d-Ov8Qdn4X6uFfe4RwlQNchNByi7wbcMAY4kxySlPkGvlPpyqs8pkWizk22W8[18WApb4bq9YEpB6o_dp_uF_cfCSRpZSR_xPUofDiNQmT43lsSYH5mzYRYF11IwRWLatury5RR-7JXJ_a8NWrqecOCkk14s_qgCPn_2i0%3D

20. GitHub - DataforGoodIsrael/DetectHateSpeech: A small solution for targeting Homophobic and Sexist Tweets to be reported to Twitter by Data For Good, Israel. GitHub. URL: https://github.com/DataforGoodIsrael/DetectHateSpeech (date of access: 31.03.2023).

21. Multilingual and Multi-Aspect Hate Speech Analysis / N. Ousidhoum та ін. // Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing. Hong Kong, 2019. С.4675-4684. URL: https://aclanthology.org/D19-1474.pdf

22. GitHub - IrynaMatviichuk/bullying-datasets. GitHub. URL: https://github.com/IrynaMatviichuk/bullying-datasets (date of access: 31.03.2023).

23. Almeida F., Xexeo G. Word Embeddings: A Survey / arXiv:1901.09069. - 2019.

24. Ямпольський Л.С. Нейротехнології та нейрокомп’ютерні ситеми / Л.С. Ямпольський, О.І. Лісовиченко, В.В. Олійник // Д К.: «Дорадо-Друк» – 2016, 571 с.