TrelBERT: A pre-trained encoder for Polish Twitter

Wojciech Szmyd, Alicja Kotyla, Michał Zobniów
Piotr Falkiewicz, Jakub Bartczuk, Artur Zygadło
deepsense.ai
research@deepsense.ai

Abstract

Pre-trained Transformer-based models have become immensely popular amongst NLP practitioners. We present TrelBERT – the first Polish language model suited for application in the social media domain. TrelBERT is based on an existing general-domain model and adapted to the language of social media by pre-training it further on a large collection of Twitter data. We demonstrate its usefulness by evaluating it in the downstream task of cyberbullying detection, in which it achieves state-of-the-art results, outperforming larger monolingual models trained on general-domain corpora, as well as multilingual in-domain models, by a large margin. We make the model publicly available. We also release a new dataset for the problem of harmful speech detection.

1 Introduction

Pre-trained language models based on the Transformer architecture (Vaswani et al., 2017) have dominated the field of NLP. The models vary in size, with the largest ones reaching hundreds of billions of parameters, and are trained with different objectives, such as causal language modeling (Radford et al., 2019; Brown et al., 2020) or masked language modeling (Devlin et al., 2019; Liu et al., 2019). By processing large amounts of text, they learn to capture general knowledge about the language and can then be fine-tuned to perform domain-specific tasks.

Regardless of the neural network architecture design choices, an important factor is the domain of the training data. For years, research in the field of NLP was mostly focused on the English language, but models and resources for many other languages have also been published recently. Multilingual models have also been developed (Conneau et al., 2020; Xue et al., 2021) which are capable of understanding more than 100 languages at once. The corpora used for pre-training are typically mixtures of general-domain data sources, such as crawled websites, books or Wikipedia articles.

The language can vary significantly across domains, not only in terms of the vocabulary, but also syntax, semantics and pragmatics. While the language of the aforementioned general-domain sources conforms to the linguistic norms, there is a large and important domain where the language is distinctly different and rapidly changing, namely social media. Apart from the obvious differences, such as the occurrence of hashtags or emojis, people have figured out how to shout using capital letters, or that ending a message with a period might be perceived as sarcastic. In order to properly represent such characteristics in the language models, it is necessary for them to be exposed to domain-specific texts not only during the supervised fine-tuning, but also in the pre-training phase.

In this work, we introduce TrelBERT, an encoder-only language model initialized with existing general-domain weights and adapted to the social media domain by pre-training it on over 40 million Polish tweets with the masked language modeling objective. TrelBERT proves to be well-suited for application of NLP in social media, achieving state-of-the-art results for downstream tasks operating on Polish Twitter data. We make the model publicly available[1].

The main contribution of our research is the introduction of the first Polish language representation model pre-trained on Twitter data. Our model achieves state-of-the-art results in the cyberbullying detection task (part of the Polish NLP benchmark), outperforming all existing solutions, including larger general-domain Polish models, as well as multilingual in-domain models. We also release a dataset of 1000 tweet IDs labeled for the problem of harmful speech detection which is a less biased (randomly sampled using the streaming API) and more up-to-date alternative to the existing one.

[1] https://huggingface.co/deepsense-ai/trelbert

2 Related work

In this section, we review the current state of Polish NLP and provide an overview of language models trained in the social media domain.

2.1 NLP for Polish language

Polish is a language spoken by over 40 million people who constitute a large population of potential beneficiaries of high-quality NLP systems. In early 2020, according to the six-level taxonomy proposed by (Joshi et al., 2020), Polish was considered one of the "Underdogs" – languages that "have a large amount of unlabeled data [. . .] and are only challenged by lesser amount of labeled data". In the following years, taking advantage of self-supervised learning, several Transformer-based models for Polish have been released, including encoder-only models such as PolBERT (Kłeczek, 2020), Polish RoBERTa (Dadas et al., 2020) and HerBERT (Mroczkowski et al., 2021), as well as decoder-only papuGaPT2 (Wojczulis and Kłeczek, 2021) and encoder-decoder plT5 (Chrabrowa et al., 2022). All of these were trained on general-domain corpora, i.e. collections of texts extracted from sources such as Wikipedia, books, newspapers, crawled websites or movie subtitles.

To compare the performance of Polish language models across a set of downstream tasks, (Rybak et al., 2020) have designed the KLEJ benchmark. It consists of 9 datasets for classification and regression, with data sources ranging from customer reviews to news summaries to Twitter messages. In KLEJ, the current state-of-the-art results (averaged across 9 tasks) are those of (Mroczkowski et al., 2021), whose HerBERT-large ranks first in the leaderboard[2], and HerBERT-base is the best performing among the base models.

Recently, (Augustyniak et al., 2022) have proposed a newer benchmark, called LEPISZCZE[3], in which they decided to keep 5 datasets from KLEJ and introduce 8 new ones, including corpora of transcribed call center conversations, legal documents and political tweets.

2.2 Language models for social media

In recent years, language models trained specifically on Twitter data have been a topic of interest for many NLP researchers, motivated by their applicability in tasks such as sentiment analysis, hate speech detection, or named entity recognition. As confirmation of this statement, 4 out of 12 tasks in the *SemEval 2023* competition[4] were based on Twitter data. Monolingual models have been trained on tweets in languages such as English (Nguyen et al., 2020), Arabic (Antoun et al., 2020; Abdelali et al., 2021), French (Guo et al., 2021), Hebrew (Seker et al., 2022), Indonesian (Koto et al., 2021), Italian (Polignano et al., 2019) and Spanish (González et al., 2021; Pérez et al., 2022). Some of them were initialized with weights of existing general-domain models and adapted to Twitter data by continued pre-training, while others were trained on Twitter data from scratch.

Recently, following the success of multilingual models such as XLM-R (Conneau et al., 2020), analogous Twitter-specific models have also been released. XLM-T (Barbieri et al., 2022) is initialized with XLM-R weights and pre-trained on 198M tweets (1.7B tokens) reflecting the distribution of over 30 languages in Twitter data, including around 1M tweets in Polish. TwHIN-BERT (Zhang et al., 2022) is trained from scratch on 7B tweets covering over 100 languages (around 100M tweets in Polish), with a contrastive social objective in addition to masked language modeling. Both XLM-T and TwHIN-BERT use the XLM-R tokenizer. The authors of Bernice (DeLucia et al., 2022), on the other hand, create a Twitter-specific tokenizer, and use it to train a masked language model on 2.5B tweets (56B tokens) in 66 languages (including more than 10M tweets in Polish) from scratch.

3 TrelBERT

We introduce a language model trained on Polish tweets which we call TrelBERT[5]. It is a Transformer encoder model trained with the masked language modeling objective. Rather than training TrelBERT from scratch, we take advantage of existing weights and adapt them to the social media domain.

As our starting point, we use HerBERT-base (Mroczkowski et al., 2021), the best performing one among Polish *base* models[6]. HerBERT was initialized with weights from XLM-R (Conneau et al., 2020) and further pre-trained on a mixture of general-domain Polish corpora with 8.6B tokens in total. Its tokenizer is a variant of Byte-Pair Encod-

[2] https://klejbenchmark.com/leaderboard

[3] https://lepiszcze.ml

[4] https://semeval.github.io/SemEval2023/tasks

[5] *trel* is the Polish word for *trill*, as a reference to *tweet*

[6] we did not consider training *large* models in this work

ing (BPE-Dropout; Provilkov et al., 2019) and has a vocabulary of 50k tokens.

3.1 Training data

We collected a random sample of 90 million tweets in Polish using the official Twitter API. The language of tweets was determined based on information provided in Twitter metadata. Only tweets created between November 2017 (when the limit of 280 characters per tweet was introduced) and July 2022 were taken into consideration.

Similar to (Nguyen et al., 2020), we preprocessed the tweets by replacing all user mentions and URLs with special tokens: *@anonymized_account* and *@URL*. We also merged multiple user mentions at the beginning of tweets into a single token as we discovered they are not part of the tweet text content but only reflect who the user is replying to in a discussion thread. We did not preprocess hashtags or emojis.

We used the pre-computed HerBERT tokenizer extended with the two additional tokens mentioned above. To best align our model with the maximum tweet length limit, we set *max_length* for truncation of tokenized tweets to 128. We filtered out tweets that have fewer than 5 tokens after tokenization from the dataset. The resulting corpus consisted of 90M tweets (2B tokens) with an average of 23 and a median of 18 tokens per tweet.

3.2 Model pre-training

We initialized our model with HerBERT-base and trained it using AdamW optimizer with a linear learning rate schedule (peak value 5e-5, warm-up for 6% steps) and the masked language modeling objective. During our experiments, we set the batch size to 2184. We trained TrelBERT for 41,208 steps (1 epoch). As we later evaluated the predictions of several model checkpoints, we noticed a visible degradation in performance on non-Twitter downstream tasks as pre-training progressed. The publicly available TrelBERT checkpoint is one that we obtained after 20k training steps, i.e. after being trained on around 44M tweets.

4 Evaluation

To compare the performance of TrelBERT with other Polish language models and Twitter-specific multilingual models, we used the KLEJ benchmark (Rybak et al., 2020) and the Political Advertising Detection task (Augustyniak et al., 2020).

4.1 KLEJ – fine-tuning

We fine-tuned the models on KLEJ tasks using Polish RoBERTa scripts[7] which we adapted to the *transformers* library. All models were trained for 10 epochs, except for models fine-tuned on the cyberbullying detection task, which were trained for 1 epoch. We used AdamW optimizer with the following hyperparameters: $\epsilon = 10^{-6}, \beta_1 = 0.9, \beta_2 = 0.98$ and a polynomial decay learning rate schedule with a peak value of 1e-5. The batch size was set to 16. The warm-up stage was set to the first 6% of the training steps.

4.2 KLEJ – cyberbullying detection

Among the tasks available in KLEJ, the one which is most relevant to our research is called cyberbullying detection (**CBD**) (Ptaszynski et al., 2019), formulated as a binary classification of harmful Twitter messages. It was originally introduced as part of the *PolEval2019* competition[8], and then included in KLEJ.

The dataset consists of 10,041 training and 1000 test examples. It is highly imbalanced, with only 851 positive class examples in the training set and 134 examples in the test set. The F1 score is used to measure the performance of models in this task.

We repeated the fine-tuning of several pretrained models to the CBD dataset five times and evaluated them on the test set. The scores reported in Table 1 are the mean values of the five fine-tuning runs. Additionally, the score for Polbert-CB (Ptaszynski et al., 2022), the Polish BERT trained for Automatic Cyberbullying Detection, is given.

The tweets included in the CBD dataset were created in late 2018 and obtained by processing answers to tweets posted by the most popular accounts, followed by further data selection and filtering according to the procedure provided in (Ptaszynski et al., 2019). To measure how our solution generalizes to the broader Twitter data distribution, we also checked the results on another test dataset which we prepared, entitled `harmful_tweets_1k`[9]. It consists of 1000 tweets in Polish randomly sampled from the years 2019 to 2022, which were then labeled by the three of us following annotation guidelines used during the creation of cyberbullying detection task (Ptaszynski et al., 2019), achieving a Fleiss' kappa value of

[7] https://github.com/sdadas/polish-roberta

[8] http://2019.poleval.pl/index.php/tasks/task6

[9] https://github.com/deepsense-ai/trelbert

Model	F1 score	Accuracy	Recall	Precision
		CBD test dataset		
HerBERT base	66.0	90.5	68.6	63.6
HerBERT large	71.4	92.3	71.6	**71.4**
Polbert-CB	67.2	91.5	64.9	69.6
TrelBERT (ours)	**74.5**	**92.7**	**79.1**	70.4
XLM-T	66.5	90.8	68.1	65.4
TwHIN-BERT base	66.2	90.6	68.5	64.1
TwHIN-BERT large	68.8	91.8	68.3	70.1
Bernice	69.1	92.7	68.5	69.8
		`harmful_tweets_1k` dataset		
HerBERT base	58.3	90.6	62.3	55.1
HerBERT large	62.8	92.0	64.2	62.0
Polbert-CB	56.5	91.7	50.9	63.5
TrelBERT (ours)	**66.3**	**92.3**	**68.9**	64.1
XLM-T	53.1	87.6	66.2	44.5
TwHIN-BERT base	49.3	89.4	48.8	50.2
TwHIN-BERT large	59.9	92.0	56.6	**64.5**
Bernice	60.7	91.8	59.2	62.2

Table 1: Results on the cyberbullying detection task.

$\kappa = 0.74$. By doing so, we obtained the test dataset, 10.6% of which were harmful Twitter messages.

TrelBERT achieves the best average results for both datasets, significantly outperforming all existing models for Polish, as well as multilingual models trained on Twitter data. In particular, it performs much better not only than HerBERT-base (which it was initialized with), but also than the *large* models. The difference between TrelBERT and all other models is especially visible in the recall value, with precision remaining more or less on par with other best-performing models. The results indicate that, if applied in a real-world scenario, TrelBERT would be able to capture more harmful content than its competitors. For one of the fine-tuned checkpoints, we submitted the predictions to the KLEJ leaderboard, officially setting the new state-of-the-art in the CBD task (F1 score = 76.1).

4.3 KLEJ – other tasks

Apart from cyberbullying detection, the KLEJ benchmark consists of 8 other tasks:

- **CDSC-E** – natural language inference; the task is to determine the logical relationship between a pair of sentences as one of entailment, contradiction or neutral
- **CDSC-R** – a semantic relatedness task, the goal of which is to predict the relatedness (ranging from 0 to 5) between a pair of sentences
- **AR** – prediction of ratings (range 1-5) for product reviews from an e-commerce platform
- **PolEmo2.0** – sentiment analysis of online consumer reviews; the training dataset consists of reviews from two domains: medicine and hotels; in **PolEmo2.0-IN** the test set consists of reviews from the same domains, while in **PolEmo2.0-OUT** the test set comes from the product and school domains
- **DYK** – a binary classification task devised based on a question-answer dataset "Did you know" (Marcińczuk et al., 2013)
- **PSC** – a text similarity task formulated as binary classification of news article-summary pairs
- **NKJP-NER** – a named entity classification task, the goal of which is to predict the presence and type of a named entity from six categories: `persName`, `orgName`, `geogName`, `placeName`, `date` and `time`

We measured how TrelBERT and other Twitter-specific models perform in these out-of-domain tasks. In this set of experiments, we fine-tuned each model once. The scores reported in Table 2

Model	**NKJP**	**CDSC-E**	**CDSC-R**	**PE2-I**	**PE2-O**	**DYK**	**PSC**	**AR**
TwHIN-BERT base	87.0	92.0	90.8	86.0	69.4	51.3	84.8	84.4
TwHIN-BERT large	89.4	92.2	91.4	88.8	75.3	52.8	82.0	85.7
Bernice	89.0	92.2	91.1	84.8	68.2	44.9	88.2	85.1
XLM-T	90.9	93.9	91.8	86.0	76.3	41.1	82.4	85.5
TrelBERT (ours)	94.4	93.9	93.6	89.3	78.1	67.4	95.7	86.1
HerBERT base	94.5	**94.5**	94.0	90.9	80.4	68.1	**98.9**	87.7
HerBERT large	**96.4**	94.1	**94.9**	**92.2**	**81.8**	**75.8**	**98.9**	**89.1**

Table 2: Results on the KLEJ benchmark (excluding CBD). For DYK and PSC tasks, the F1 score is reported. In AR, the micro-average of the mean-absolute error per class (wMAE) is used to measure performance. In CDSC-R, Spearman correlation is applied for evaluation. For the remaining tasks, accuracy is reported.

are mostly obtained by uploading predictions to the KLEJ benchmark page without publishing the results. The scores for HerBERT-base, HerBERT-large and TrelBERT are taken directly from the leaderboard. Unsurprisingly, due to being adapted towards the language of social media, TrelBERT achieves slightly worse results than HerBERT-base on all 8 tasks operating on data out of its domain. As expected, the Twitter-specific multilingual models perform worse than Polish-only ones, although the differences for some of the tasks are not vast. The discrepancy in performance metrics between Twitter-only based models and general-domain models in general-domain tasks (particularly noticeable in tasks DYK, PSC and PE2-O) shows how the language of social media is different from linguistic norms. This might also suggest that general knowledge about the world and language (which a model can learn from general-domain corpora) is relevant to domain-specific tasks such as harmful speech detection.

4.4 Political advertising detection

We also conducted experiments on another Twitter-based downstream task, Political Advertising Detection, proposed in (Augustyniak et al., 2020). The related dataset consists of 1701 human-annotated tweets (1020 for training, 340 for validation and 341 in the test set) collected by searching for specific hashtags and keywords related to the Polish presidential elections in 2020. The goal of the task is to perform token-level sequence labeling with 9 categories (healthcare, welfare, defense, legal, education, infrastructure, society, foreign policy and immigration) with an imbalanced number of examples. The task is included in the LEPISZCZE benchmark (Augustyniak et al., 2022).

The results reported in Table 3 are macro F1 scores achieved by selected models averaged over 5 fine-tuning runs. The fine-tuning process was similar to that described in 4.1, the only difference being the learning rate which we set to 1e-5. All the evaluated Polish-only models perform better than multilingual Twitter-specific ones, but there is no significant difference between TrelBERT and the two HerBERT variants. However, taking into account the rather small size of the dataset (for a sequence labeling problem with 9 categories, some of them with very few examples) and its collection and annotation procedures (bias towards certain keywords), we do not draw any general conclusions about the capabilities of the model.

Model	Macro F1
Bernice	62.62 ± 4.28
XLM-T	64.42 ± 0.90
TwHIN-BERT large	67.20 ± 1.60
TwHIN-BERT base	67.63 ± 1.54
HerBERT base	69.23 ± 1.87
TrelBERT (ours)	70.08 ± 0.50
HerBERT large	71.32 ± 1.38

Table 3: Results on the Political Advertising Detection test set for selected models.

5 Conclusion

In this paper, we have introduced TrelBERT, the first Polish language representation model pretrained on Twitter data. It achieves state-of-the-art results in a cyberbullying detection task, outperforming all existing solutions, including larger general-domain Polish models, as well as multilingual in-domain models. Additionally, we contribute by releasing a harmful speech dataset with labeled tweet IDs which could be used as an alternative test set for cyberbullying detection.

Limitations

By taking the characteristics of the language used by the social media community into consideration, we are aware that applying a general-purpose tokenizer has some major limitations. Emojis, emoticons, user mentions, hashtags, and URLs are inseparable elements of Twitter language and their existence should not be unnoticed or treated as noise in a good-quality corpus. Emojis and emoticons could be interpreted as digital gestures or face expressions. By replacing all user mentions and URLs with *@anonymized_account* and *@URL* tokens, we lose the meaning they convey. On the other hand, doing so was necessary for ethical (user mentions) or pragmatic reasons (preprocessing and tokenizing URLs would be difficult).

Also, measuring the performance of the model on just two downstream tasks with data from Twitter does not seem to be a sufficiently fair benchmark to prove the superiority of our model. Unfortunately, the vast majority of languages (including Polish) suffer from a lack of high-quality labeled datasets.

Last but not least, the language of social media is changing rapidly. TrelBERT outperforms other models in the cyberbullying detection task, but we expect it to degrade performance on future data. Thus, updating the weights of the model by means of further pre-training on latest tweets is necessary to keep the model effective.

Ethics Statement

Due to the nature of our data, there were several ethical issues to consider. First, we anonymized all the usernames mentioned in tweets by replacing them with *@anonymized_account* token. Despite the fact that the data is publicly available, we decided to prevent the model from learning sentiment about specific users based on what the community writes about them. We did not want the model to produce harmful output tokens for specific users.

Secondly, there is a great deal of harmful content in social media, which we could possibly try to remove from the training corpus as part of data preprocessing to prevent the model from learning this kind of language. However, if we are to use such models to detect hate speech or cyberbullying, they need to know it. We believe that exposing a model to harmful content only during the fine-tuning stage may not be enough.

References

Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. 2021. Pre-training BERT on Arabic Tweets: Practical considerations. *ArXiv*, abs/2102.10684.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Lukasz Augustyniak, Krzysztof Rajda, Tomasz Kajdanowicz, and Michał Bernaczyk. 2020. Political advertising dataset: the use case of the Polish 2020 presidential elections. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, pages 110–114, Seattle, USA. Association for Computational Linguistics.

Łukasz Augustyniak, Kamil Tagowski, Albert Sawczyn, Denis Janiak, Roman Bartusiak, Adrian Szymczak, Marcin Wątroba, Arkadiusz Janz, Piotr Szymański, Mikołaj Morzy, Tomasz Kajdanowicz, and Maciej Piasecki. 2022. This is the way: designing and compiling LEPISZCZE, a comprehensive NLP benchmark for Polish.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Aleksandra Chrabrowa, Łukasz Dragan, Karol Grzegorczyk, Dariusz Kajtoch, Mikołaj Koszowski, Robert Mroczkowski, and Piotr Rybak. 2022. Evaluation of transfer learning for Polish with a text-to-text model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4374–4394, Marseille, France. European Language Resources Association.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco

Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Sławomir Dadas, Michał Perełkiewicz, and Rafał Poświata. 2020. Pre-training Polish Transformer-based language models at scale. In *Artificial Intelligence and Soft Computing: 19th International Conference, ICAISC 2020, Zakopane, Poland, October 12-14, 2020, Proceedings, Part II*, page 301–314, Berlin, Heidelberg. Springer-Verlag.

Alexandra DeLucia, Shijie Wu, Aaron Mueller, Carlos Aguirre, Philip Resnik, and Mark Dredze. 2022. Bernice: A multilingual pre-trained encoder for Twitter. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6191–6205, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

José Ángel González, Lluís-F. Hurtado, and Ferran Pla. 2021. TWilBert: Pre-trained deep bidirectional transformers for spanish twitter. *Neurocomputing*, 426:58–69.

Yanzhu Guo, Virgile Rennard, Christos Xypolopoulos, and Michalis Vazirgiannis. 2021. BERTweetFR : Domain adaptation of pre-trained language models for French tweets. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 445–450, Online. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. IndoBERTweet: A pretrained language model for Indonesian Twitter with effective domain-specific vocabulary initialization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10660–10668, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dariusz Kłeczek. 2020. Polbert: Attacking Polish NLP Tasks with Transformers. In *Proceedings of the PolEval 2020 Workshop*. Institute of Computer Science, Polish Academy of Sciences.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Michał Marcińczuk, Marcin Ptak, Adam Radziszewski, and Maciej Piasecki. 2013. Open Dataset for Development of Polish Question Answering Systems. In *Proceedings of Human Language Technologies as a Challenge for Computer Science and Linguistics'13*, pages 479–483, Poznań. Fundacja UAM.

Robert Mroczkowski, Piotr Rybak, Alina Wróblewska, and Ireneusz Gawlik. 2021. HerBERT: Efficiently pretrained transformer-based language model for Polish. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 1–10, Kiyv, Ukraine. Association for Computational Linguistics.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Juan Manuel Pérez, Damián Ariel Furman, Laura Alonso Alemany, and Franco M. Luque. 2022. RoBERTuito: a pre-trained language model for social media text in Spanish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7235–7243, Marseille, France. European Language Resources Association.

Marco Polignano, Valerio Basile, Pierpaolo Basile, Marco de Gemmis, and Giovanni Semeraro. 2019. AlBERTo: Modeling Italian social media language with bert. *Italian Journal of Computational Linguistics*, 5:11–31.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2019. Bpe-dropout: Simple and effective subword regularization. *arXiv preprint arXiv:1910.13267*.

Michal Ptaszynski, Agata Pieciukiewicz, and Paweł Dybała. 2019. Results of the poleval 2019 shared task 6: First dataset and open shared task for automatic cyberbullying detection in polish twitter.

Michal Ptaszynski, Agata Pieciukiewicz, Pawel Dybala, Pawel Skrzek, Kamil Soliwoda, Marcin Fortuna, Gniewosz Leliwa, and Michal Wroczynski. 2022. Polish bert trained for automatic cyberbullying detection.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Piotr Rybak, Robert Mroczkowski, Janusz Tracz, and Ireneusz Gawlik. 2020. KLEJ: Comprehensive benchmark for Polish language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1191–1201, Online. Association for Computational Linguistics.

Amit Seker, Elron Bandel, Dan Bareket, Idan Brusilovsky, Refael Greenfeld, and Reut Tsarfaty. 2022. AlephBERT: Language model pre-training and evaluation from sub-word to sentence level. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46–56, Dublin, Ireland. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Michał Wojczulis and Dariusz Kłeczek. 2021. papugapt2 - polish gpt2 language model.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2022. TwHIN-BERT: A socially-enriched pre-trained language model for multilingual tweet representations. *arXiv preprint arXiv:2209.07562*.