Steps towards Addressing Text Classification in Low-Resource Languages

Maximilian Weissenbacher
Information Science
University of Regensburg
Maximilian.Weissenbacher@ur.de

Udo Kruschwitz
Information Science
University of Regensburg
Udo.Kruschwitz@ur.de

Abstract

Text classification is an area of NLP in which major improvements have been observed in recent years, primarily via pre-training and fine-tuning of large language models (LLMs). However, low-resource languages still face major challenges. We explore how to address this problem using different text classification tasks across two low-resource languages. Our focus is on adopting multilingual LLMs using data expansion techniques (with and without machine translation). Results indicate that pre-trained, fine-tuned models of the resource-poor language appear more promising than multilingual models, we also find that translating into a resource-poor language is not beneficial in our experimental settings.

1 Introduction

Few languages can be considered resource-rich, the vast majority are not despite a possibly very large pool of speakers. For example, 83 million people speak Marathi (only outnumbered in India by Hindi and Bengali). Malayalam is another Indian language with a sizeable population of speakers (37 million). However, in the context of NLP both languages are considered resource-poor, and more research has been done on more prominent Indian languages like Hindi (Joshi et al., 2016) or Bengali (Patra et al., 2018). In general, resource-poor languages lack annotated training data because there are often no trained linguistic annotators for these languages, and the markets may be too small or premature to invest in such training (Ruder et al., 2019). But many people speak such languages and the amount of textual content on online platforms such as Twitter keeps growing. We adopt both languages as *exemplars* for other low-resource languages. We look at three different classification tasks (sentiment analysis, hate speech detection and claim detection) comparing language-specific fine-tuning with multilingual LLMs. We also look at data expansion by adding training data available from a high-resource language (either with or without first translating into our language of interest). This approach has some similarity with (but is different from) *data augmentation* that focuses on adding synthetic data such as via generating new data samples using autoregressive models (Wullach et al., 2021; Whitfield, 2021). We see our contribution as exploratory work into the problem which offers some interesting insights that can serve as a starting point for more work. To support reproducibility we also make all code available.[1]

2 Related Work

LLMs like BERT (Devlin et al., 2019) have established a new state of the art for text classification tasks, e.g. (Chouikhi et al., 2021; Chan et al., 2020) outperforming traditional ML approaches using Naive Bayes or Support Vector Machines (SVM) (Schmidt et al., 2022; Geetha and Karthika Renuka, 2021). Among a wide range of text classification tasks, sentiment analysis, hate speech detection and claim detection can be seen as typical classification problems (Medhat et al., 2014; Schmidt and Wiegand, 2017; Levy et al., 2014; Konstantinovskiy et al., 2021). However, research is lacking for resource-poor languages. Nevertheless, numerous test collections have been created for low-resource language, e.g. for sentiment analysis (Kulkarni et al., 2021), hate speech detection (Pitenis et al., 2020; Çöltekin, 2020; Mandl et al.,

[1] https://github.com/MaxiWeissenbacher/exploratory_bert_v2/tree/main

Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023), pages 69–76
September 18–22, 2023.

2021), and claim detection (Kazemi et al., 2021). Snæbjarnarson et al. 2023 demonstrated that the transfer learning performance of low-resource languages (Faroese in their case) could substantially improve by exploiting data and models of closely-related high-resource languages (other Scandinavian languages). That is a direction we consider promising and we explore how incorporating additional datasets will affect a transformer model. This is an important research topic to establish generalizability and transferability (Mandl et al., 2021; Fortuna et al., 2021).

3 Methodology

We explore **five different approaches**. The first approach focuses on whether **fine-tuned models** of a resource-poor language can perform better than multilingual models like mBERT and XLM-RoBERTa. The second, third, and fourth approach investigate whether it is beneficial if additional data gets **translated into the resource-poor language** and added for training. The fifth approach takes the inverse view: the dataset of the **resource-poor language gets translated into a resource-rich language** (English). After the translation process, fine-tuned English models are used to see if performance increases can be observed.

3.1 Datasets

As the availability of (even high-resource) language resources varies from one task to another we tap into different languages, such as German, Hindi, and English, in addition to the baseline datasets in this work.

Sentiment Analysis. We consider the L3-Cube-MahaSent dataset as our baseline dataset for the sentiment analysis domain, as it is one of the best-known resources in Marathi language. It contains tweets classified as positive, negative, and neutral. It has 12,114 train, 2,250 test, and 1,500 validation examples (Kulkarni et al., 2021). For approaches with data expansion, four additional datasets with the same labels but different annotation guidelines were used (see Appendix A.1) and added:

- GFES Dataset (DE), (Schmidt et al., 2022)
- SB10k Dataset (DE), (Cieliebak et al., 2017)
- Kaggle Covid Dataset (EN), (Miglani, 2020)
- Sentiment Analysis Dataset (HI)

Hate Speech Detection. For this task, the datasets of HASOC2021 Sub-task 1A were used, consisting of datasets in three different languages. The task is a binary classification in which participating systems are required to classify tweets into two classes, namely: Hate or Offensive (HOF) vs. Non-Hate and Non-Offensive (NOT). The Marathi dataset contains 1,874 tweets, the English dataset 3,843 tweets, and the Hindi dataset 4,594 tweets. The annotation quality of this dataset is considered to be reliable (Modha et al., 2021).

Claim Detection. The dataset from Kazemi et al. 2021 was used here. It contains content in high-resource (English, Hindi) and lower-resource (Bengali, Malayalam, Tamil) languages. We used Malayalam as our low-resource baseline language, added texts from the remaining languages and only used texts which were labeled as "Claim" and "No Claim". Therefore a binary classification task was conducted. With this, 4,017 texts remain in the dataset, with 730 texts in Malayalam. Three different annotators worked on this dataset, and the annotation quality is also considered reliable (Kazemi et al., 2021).

3.2 Experimental Setup and Implementation

We use Huggingface for all models and their library "Simpletransformers" (Wolf et al., 2020). We used an "NVIDIA Tesla K80" GPU server to train the different text classification models. All notebooks run on the freely available version of Google Colaboratory (all codebooks in our GitHub repository).

For translating the datasets to Marathi or translating the Marathi datasets to English, the Python library "Googletrans"[2] was used.

The project also investigates how preprocessing the data influences the performance of transformer models. For this, preprocessing steps like removing links, square brackets, punctuation, words containing numbers, and lowercasing the text were used (we also tried over- and undersampling with inconclusive results, so they were not considered further).

We computed Accuracy and weighted F1 if the distribution of the labels is not balanced.

Each model is fine-tuned for three epochs, a train and evaluation batch size of 32, the learning rate of 2e-5, the default epsilon of 1e-8 to find a better minimum for the loss function and Adam

[2] https://pypi.org/project/googletrans/

	Model Name	BASELINE	+ Translation to MR				+ NON-TRANSLATED DATA				+ ALL DATA TRANS-LATED To MR	+ ALL DATA NOT TRANS-LATED To MR
Dataset		L3-Cube-Maha-Sent	+ Kaggle-Covid Data (EN)	+ Kaggle-Tweets (HI)	+ GFES-Tweets (DE)	+ SB10K-Tweets (DE)	+ Kaggle-Covid Data (EN)	+ Kaggle-Tweets (HI)	+ GFES-Tweets (DE)	+ SB10K-Tweets (DE)		
Multilingual Models	mBERT	81.9%	80.9%	81.8%	81.6%	81.7%	82.0%	81.9%	81.9%	81.9%	80.5%	80.6%
	XLM-RoBERTa	83.4%	83.0%	83.0%	83.4%	83.3%	83.5%	83.2%	83.5%	83.6%	82.9%	83.1%
Marathi Models	IndicBERT	84.1%	83.6%	83.7%	84.2%	83.5%						
	MahaBERT	83.8%	82.9%	83.1%	83.5%	83.4%						
	MahaAlBERT	84.0%	83.6%	83.0%	83.6%	83.7%						
	MahaRoBERTa	84.7%	84.4%	84.1%	84.6%	84.3%						

Figure 1: Sentiment Analysis Results (F1 scores)

Model Name	BASELINE	+ TRANSLATION TO MR		+ NON-TRANSLATED DATA		+ ALL DATA TRANSLATED (MR)	+ ALL DATA NOT TRANSLATED
	HASOC 2021	+ EN	+ Hi	+EN	+ Hi		
mBERT	83.7%	85.6%	86.8%	85.8%	87.5%	85.0%	87.2%
XLM-RoBERTa	82.9%	85.5%	85.4%	86.8%	87.4%	85.3%	87.0%
IndicBERT	79.5%	83.2%	85.3%	/	/	/	/
MahaBERT	88.5%	86.6%	88.2%	/	/	/	/
MahaAlBERT	82.7%	83.5%	85.3%	/	/	/	/
MahaRoBERTa	87.7%	86.9%	88.0%	/	/	/	/

Figure 2: Hate Speech Detection Results (F1 scores)

Model Name	BASELINE	+ TRANSLATION TO ML				+ NON TRANSLATED DATA				+ ALL DATA TRANSLATED (ML)	+ ALL DATA NOT TRANSLATED
	Malayalam	+En	+Hi	+Ta	+Bn	+En	+Hi	+Ta	+Bn		
mBERT	82.2%	84.3%	81.8%	83.2%	80.6%	84.1%	83.1%	83.2%	83.8%	84.1%	84.2%
XLM-RoBERTa	73.9%	84.9%	84.1%	83.7%	82.2%	85.1%	83.8%	83.7%	83.1%	85.1%	86.6%

Figure 3: Claim Detection Results (F1 scores)

optimizer for stochastic gradient descent (Tato and Nkambou, 2018).

The models are trained and evaluated in a 5x5 cross-evaluation setting, and the average score over five runs gets reported. For evaluation, we compare against baseline approaches using two-tailed t-tests (with $p < 0.05$).

3.3 Model Selection

In this work we focus on both multilingual and monolingual BERT models, as they count as strong baselines for text classification tasks. The following multilingual models are used (details in Appendix A.2): mBERT-Cased, XLM-RoBERTa. And following monolingual models are used: IndicBERT, MahaBERT, MahaAlBERT, MahaRoBERTa, BERTweet, TimeLMs (Cardiff RoBERTa).

4 Results

4.1 Fine-tuning on resource-poor language

The first approach compares fine-tuned models of a resource-poor language with multilingual models like mBERT and XLM-RoBERTa. It focuses on sentiment analysis and hate speech detection datasets in Marathi. The models are trained on baseline Marathi datasets when no additional data is available. The results (Figures 1 and 2) show that all fine-tuned Marathi models perform better than mBERT and XLM-RoBERTa in sentiment analysis. The best model, MahaRoBERTa, achieves a statistically significant improvement over the best multilingual model. Other Marathi models also exhibit an upward trend in performance, although not statistically significant. MahaRoBERTa with the hyperparameters we have used outperforms the baseline results and results from related studies (Kulkarni et al., 2021; Ve-

lankar et al., 2022). For hate speech detection the pattern is slightly different. Here IndicBERT and MahaAlBERT had lower F1 scores. However, the best performing models are still Marathi fine-tuned MahaBERT and MahaRoBERTa, significantly better than the best multilingual model, mBERT. The MahaBERT model would have ranked 6th place for task 1A at the HASOC Subtrack at FIRE 2021 (Mandl et al., 2021).

4.2 Adding translated data

The second approach examines the impact of translating texts into a resource-poor language and adding them for training across three text classification domains. Results show that translating datasets to Marathi slightly decreases F1 for multilingual models in sentiment analysis (see Figure 1 - '+Translation to MR'). However, a slight increase is observed for IndicBERT combined with the translated GFES dataset. In hate speech detection (Figure 2), adding translated English and Hindi datasets benefits both multilingual and Marathi models, except for MahaBERT. The translated Hindi dataset contributes more to F1 improvement. In claim detection (Figure 3) for Malayalam, adding translated English data significantly improves results compared to the baseline model. The approach occasionally helps improve weighted F1-score for datasets in Hindi, Tamil, and Bengali, but not consistently.

4.3 Adding non-translated data

This third approach compares whether it is worth translating the data to a resource-poor language or if the multilingual models perform better if the data is added in its original form. Figure 1 shows the results ("+Non-TRANSLATED DATA"). The Sentiment Analysis approach shows that adding the non-translated data performs slightly better than adding translated data for training. The same pattern can be observed for hate speech detection and the claim detection. However, there is no statistical significance between the translation- and non-translation approaches. This approach has only been done for the multilingual models mBERT and XLM-R, because less accurate results are expected if English or German data is added to fine-tuned-, monolingual Marathi models.

4.4 Adding all datasets combined

For this fourth approach, it was tested to apply all available datasets combined, translated and not translated, for training and if this contributes positively to the model performance. For all three classification domains, the same pattern can be observed. Appending all the non-translated data achieves better results than appending all translated datasets. Compared to the baseline approach of the multilingual models we see significant improvements for hate speech and claim detection classification but not for sentiment analysis.

Model Name	BASELINE	TRANSLATION TO EN	
	L3-Cube-MahaSent	+ Preprocessing	+ without Preproccessing
mBERT	81.9%	80.8%	81.8%
XLM-RoBERTa	83.4%	82.2%	83.3%
BerTweet	/	82.8%	83.6%
Cardiff Roberta	/	83.9%	84.0%

Figure 4: Sentiment Analysis (F1 scores)

Model Name	BASELINE	TRANSLATION TO EN	
	HASOC2021	+ Preprocessing	+ without Preproccessing
mBERT	83.7%	80.3%	82.0%
XLM-RoBERTa	82.9%	79.7%	80.4%
BerTweet	/	82.2%	84.3%
Cardiff Roberta	/	83.2%	85.0%

Figure 5: Hate speech Detection (F1 scores)

Model Name	BASELINE	TRANSLATION TO EN	
	Malayalam Claims	+ Preprocessing	+ without Preproccessing
mBERT	82.0%	81.8%	82.1%
XLM-RoBERTa	73.9%	66.9%	73.4%
BerTweet	/	78.8%	82.2%
Cardiff Roberta	/	79.3%	82.5%

Figure 6: Claim Detection results (F1 scores)

4.5 Translating to English

The fifth and final approach involved translating resource-poor language datasets into English. This allowed the use of fine-tuned English classification models like BERTweet and TimeLMs (Cardiff Roberta). Results (Figure 4, 5 and 6) show that TimeLMs performed best across all tasks, with statistical significance in sentiment analysis and hate speech detection. Notably, the approach without preprocessing the data performed better than preprocessing before training.

5 Discussion

The first approach (fine-tuning baseline) showed that if fine-tuned models of the resource-poor language are available, it makes sense to use them, as they showed improved results on multilingual models. This is in line with Velankar et al. 2022 who compared mono vs multilingual models for text classification.

For the second approach (where we expanded the dataset by adding translated data), we saw no improvements for the L3-Cube-MahaSent dataset. The dataset is already quite big with more than 12,000 train texts and a balanced distribution of the labels. Adding more data makes the model noisy, as the label distribution is less balanced than the baseline model. For hate speech detection it was beneficial to add translated data. This could be because the HASOC2021 dataset is quite small with 1,874 tweets and more data helps the model make better decisions. Therefore, if researchers only have small datasets available, it might be useful to search for additional datasets, which can be from a different language, and translate them into the target language. This is not guaranteed though, as the claim detection task is in a similar situation with a small amount of data, and adding translations of different datasets did not help.

In general, the third and fourth approaches (expanding by adding non-translated data and expanding by combining all data, respectively) showed the pattern that, for multilingual models, it is better just to append the non-translated data. Reasons for this can be that there is some noise when translating the texts, which sometimes leads to worse model decisions (in line with Ponti et al. 2021). They argued the main limitation of the translation process is that sentences that are possibly not faithful to the original in the target language and/or not grammatical in the source language are fed to the classifier, which degrades its performance (Ponti et al., 2021). The resources for the translation process can therefore be saved.

The fifth and final approach (translating into English to tap into resource-rich resources for fine-tuning) was chosen because it is challenging to preprocess tweets in Marathi or Malayalam due to the different alphabet and there are not many open-source tools available to do so. The idea was to bring those texts to English and use the well-established English preprocessing methods. Clearer results with the preprocessing were expected, but the opposite was the case: The models performed better without preprocessing. This could be because some important information for the model gets removed here. For example, a high volume of punctuation could hint at a bad sentiment, but this information gets lost with preprocessing. Still, the results show the benefit of first translating data to English and then using fine-tuned English models like BERTweet or TimeLMs. For future research this appears to be a promising directions. Overall, the results show that the best performance was achieved by using fine-tuned language-specific models like MahaRoBERTa or MahaBERT.

6 Conclusion

We explored different approaches to enhance the performance of multilingual classification models for low-resource languages, specifically Marathi and Malayalam. Our findings suggest that appending additional datasets in their original form to multilingual models is more effective than translating them to the resource-poor language. Adding extra data is particularly beneficial for small baseline datasets. When the baseline dataset was translated to English without preprocessing, fine-tuned English models outperformed multilingual models. However, the best results were obtained by using fine-tuned models of the resource-poor language. In conclusion, researchers can consider using translation approaches to improve multilingual language models, but if fine-tuned models for the resource-poor language already exist, they tend to yield the best results.

7 Ethical Considerations

Whenever social media data is being processed ethical concerns naturally arise. This is particularly true if the data contains some personal information. We use existing test collections in our work to minimize such problems. In addition to that we operate within the strict framework imposed on any research within our organisation.

Wider issues emerge from the actual classification tasks. The balance between free speech and censorship in hate speech detection is an issue of ongoing debate that also has ethical questions at its heart (Zimmerman et al., 2018). Claim detection also gives rise to such issues (less so sentiment analysis).

8 Limitations

This work also has a number of limitations. First of all, the L3-Cube-MahaSent dataset from Kulkarni et al. 2021 is limited to tweets from political personalities and activists, which may not be representative of the entire Marathi-speaking population. The datasets for the hate speech and claim detection task are relatively small, making it more challenging to ensure that the training data is diverse and representative. It is important to be aware of these limitations and to make efforts to mitigate biases in the model's training and evaluation. Also, low-resource languages often have limited digital footprints, making it difficult to collect sufficient data for training text classification models. Another difficulty that comes with the datasets, especially with the open-source Kaggle datasets, is that it is unclear how the labeling process looked like and what the annotator agreement was. This is indeed important information, as the data quality can have a huge impact on the model performance. One last limitation of this work is that different languages for different NLP-tasks have been chosen as low-resource languages (Marathi and Malayalam), making it hard to generalize the findings. At first, we wanted to use a Marathi dataset for claim detection as well. But to the best of our knowledge, we did not find one and therefore used the Malayalam dataset to see similarities with another language.

Acknowledgments

We would like to thank the anonymous reviewers for their constructive feedback.

References

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. *arXiv preprint arXiv:2010.10906*.

Hasna Chouikhi, Hamza Chniter, and Fethi Jarray. 2021. Arabic sentiment analysis using bert model. In *International Conference on Computational Collective Intelligence*, pages 621–632. Springer.

Mark Cieliebak, Jan Milan Deriu, Dominic Egger, and Fatih Uzdilli. 2017. A twitter corpus and benchmark resources for german sentiment analysis. In *5th International Workshop on Natural Language Processing for Social Media, Boston MA, USA, 11 December 2017*, pages 45–51. Association for Computational Linguistics.

Çağrı Çöltekin. 2020. A corpus of turkish offensive language on social media. In *Proceedings of the 12th language resources and evaluation conference*, pages 6174–6184.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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.

Paula Fortuna, Juan Soler-Company, and Leo Wanner. 2021. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3):102524.

M.P. Geetha and D. Karthika Renuka. 2021. Improving the performance of aspect based sentiment analysis using fine-tuned bert base uncased model. *International Journal of Intelligent Networks*, 2:64–69.

Aditya Joshi, Ameya Prabhu, Manish Shrivastava, and Vasudeva Varma. 2016. Towards sub-word level compositions for sentiment analysis of hindi-english code mixed text. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2482–2491.

Raviraj Joshi. 2022. L3cube-mahacorpus and mahabert: Marathi monolingual corpus, marathi bert language models, and resources. *arXiv preprint arXiv:2202.01159*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961.

Ashkan Kazemi, Kiran Garimella, Devin Gaffney, and Scott A Hale. 2021. Claim matching beyond english to scale global fact-checking. *arXiv preprint arXiv:2106.00853*.

Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. 2021. Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. *Digital Threats: Research and Practice*, 2(2):1–16.

Atharva Kulkarni, Meet Mandhane, Manali Likhitkar, Gayatri Kshirsagar, and Raviraj Joshi. 2021. L3cubemahasent: A marathi tweet-based sentiment analysis dataset. *arXiv preprint arXiv:2103.11408*.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. Timelms: Diachronic language models from twitter. *arXiv preprint arXiv:2202.03829*.

Thomas Mandl, Sandip Modha, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Prasenjit Majumder, Johannes Schäfer, Tharindu Ranasinghe, Marcos Zampieri, Durgesh Nandini, et al. 2021. Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages. *arXiv preprint arXiv:2112.09301*.

Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.

Aman Miglani. 2020. Coronavirus tweets nlp - text classification.

Sandip Modha, Thomas Mandl, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Tharindu Ranasinghe, and Marcos Zampieri. 2021. Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages and conversational hate speech. In *Forum for Information Retrieval Evaluation*, pages 1–3.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200*.

Braja Gopal Patra, Dipankar Das, and Amitava Das. 2018. Sentiment analysis of code-mixed indian languages: An overview of sail_code-mixed shared task@ icon-2017. *arXiv preprint arXiv:1803.06745*.

Zeses Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in greek. *arXiv preprint arXiv:2003.07459*.

Edoardo Maria Ponti, Julia Kreutzer, Ivan Vulić, and Siva Reddy. 2021. Modelling latent translations for cross-lingual transfer. *arXiv preprint arXiv:2107.11353*.

Sebastian Ruder, Anders Søgaard, and Ivan Vulić. 2019. Unsupervised cross-lingual representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 31–38.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10.

Thomas Schmidt, Jakob Fehle, Maximilian Weissenbacher, Jonathan Richter, Philipp Gottschalk, and Christian Wolff. 2022. Sentiment analysis on twitter for the major german parties during the 2021 german federal election. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 74–87.

Vésteinn Snæbjarnarson, Annika Simonsen, Goran Glavaš, and Ivan Vulić. 2023. Transfer to a low-resource language via close relatives: The case study on faroese. *arXiv preprint arXiv:2304.08823*.

Ange Tato and Roger Nkambou. 2018. Improving adam optimizer. 2018. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018 Workshop Track)*.

Abhishek Velankar, Hrushikesh Patil, and Raviraj Joshi. 2022. Mono vs multilingual bert for hate speech detection and text classification: A case study in marathi. In *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, pages 121–128. Springer.

Dewayne Whitfield. 2021. Using GPT-2 to create synthetic data to improve the prediction performance of NLP machine learning classification models. *CoRR*, abs/2104.10658.

Tomer Wullach, Amir Adler, and Einat Minkov. 2021. Fight fire with fire: Fine-tuning hate detectors using large samples of generated hate speech. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 4699–4705. Association for Computational Linguistics.

Steven Zimmerman, Udo Kruschwitz, and Chris Fox. 2018. Improving hate speech detection with deep learning ensembles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

A Appendices

A.1 Sentiment Analysis Datasets

- German Federal Election Sentiment Dataset (GFES):
Schmidt et al. provided a German dataset of 2000 annotated tweets of German politicians during the federal election in 2021 (Schmidt et al., 2022). The annotation of the data has been done by students and employees of the University of Regensburg, and the annotation quality counts as reliable.
- SB10k Dataset:
Cieliebak et al. provided a big dataset of 10.000 annotated German tweets for Sentiment Analysis (Cieliebak et al., 2017). Researchers have done annotation, so the annotation quality counts as reliable.
- Kaggle Coronavirus Dataset:
This dataset from Kaggle[3] with 41.000 labeled English tweets was used to see if big, open-source datasets can be used to improve the accuracy of language models. Tweets with the label "Extremely Positive" or "Extremely Negative" were re-labeled as "Positive" and "Negative". There are no insights on how the data was annotated, so the annotation quality counts as questionable.
- Hindi Sentiment Analysis Dataset:
Also, one dataset with an Indian language, Hindi, was used for this project. The dataset consists of 9077 manually labeled tweets in Hindi. Unfortunately, the Kaggle link is no longer available, but as the experiments with this dataset have already been done, the dataset is still included in this work.

A.2 Models used in this work

A.) Multilingual-BERT-Cased (mBERT-Cased)[4]:
mBERT is a transformer-based model, pre-trained on a large corpus of multilingual data (104 languages) in a self-supervised fashion. The mBERT-Cased model is case-sensitive, so it makes a difference, for example, for "Hello World" and "hello world" (Devlin et al., 2019).

B.) XLM-RoBERTa (XLM-R)[5]:
XLM-R is a multilingual version of RoBERTa, pre-trained on 100 languages. Conneau et al. found that this model performs exceptionally well on low-resource languages (Conneau et al., 2019).

C.) IndicBERT[6]:
A multilingual ALBERT model released by Ai4Bharat trained on large-scale corpora. The training languages include 12 major Indian languages. The model has been proven to work better for tasks in Indic language (Kakwani et al., 2020).

D.) MahaBERT[7]:
A multilingual BERT (bert-base-multilingual-cased) model fine-tuned on L3Cube-MahaCorpus and other publicly available Marathi monolingual datasets (Joshi, 2022).

E.) MahaAlBERT[8]:
A monolingual AlBERT model, trained on L3Cube-MahaCorpus and other publicly available Marathi monolingual datasets (Joshi, 2022).

F.) MahaRoBERTa[9]:
A multilingual RoBERTa (xlm-roberta-base) model fine-tuned on L3Cube-MahaCorpus and other publicly available Marathi monolingual datasets (Joshi, 2022).

G.) BERTweet[10]:
A RoBERTa based model pre-trained on 850M English tweets. (Nguyen et al., 2020).

H.) TimeLMs[11]:
A RoBERTa based model pre-trained on English tweets and finetuned for sentiment analysis with the TweetEval benchmark (Loureiro et al., 2022).

[3] https://www.kaggle.com/datasets/datatattle/covid-19-nlp-text-classification?select=Corona_NLP_train.csv
[4] https://huggingface.co/bert-base-multilingual-cased
[5] https://huggingface.co/xlm-roberta-base
[6] https://huggingface.co/ai4bharat/indic-bert
[7] https://huggingface.co/l3cube-pune/marathi-bert
[8] https://huggingface.co/l3cube-pune/marathi-albert-v2
[9] https://huggingface.co/l3cube-pune/marathi-roberta
[10] https://huggingface.co/vinai/bertweet-base
[11] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest