EPiC Series in Computing

Volume 93, 2023, Pages 203–215

Proceedings of Society 5.0 Conference 2023

Combating Hate: How Multilingual Transformers Can Help Detect Topical Hate Speech

Trishanta Srikissoon and Vukosi Marivate
Department of Computer Science, University of Pretoria
u28105967@tuks.co.za, vukosi.marivate@cs.up.ac.za

Abstract

Automated hate speech detection is important to protecting people's dignity, online experiences, and physical safety in Society 5.0. Transformers are sophisticated pre-trained language models that can be fine-tuned for multilingual hate speech detection. Many studies consider this application as a binary classification problem. Additionally, research on topical hate speech detection use target-specific datasets containing assertions about a particular group. In this paper we investigate multi-class hate speech detection using target-generic datasets. We assess the performance of mBERT and XLM-RoBERTA on high and low resource languages, with limited sample sizes and class imbalance. We find that our fine-tuned mBERT models are performant in detecting gender-targeted hate speech. Our Urdu classifier produces a 31% lift on the baseline model. We also present a pipeline for processing multilingual datasets for multi-class hate speech detection. Our approach could be used in future works on topically focused hate speech detection for other low resource languages, particularly African languages which remain under-explored in this domain.

1 Introduction

Social media platforms (SMP) are effective tools for networking and collaboration, but they can also be misused to share harmful content. Ease of use and anonymity allows hate speech to spread rapidly and with little consequence (Pamungkas, Basile, & Patti, 2021). Hate speech is any spoken, written, or behavioural communication that causes harm or prejudice against individuals or groups based on innate characteristics (United Nations, 2023). The circulation of hate speech on SMPs not only desensitises users but also manifests as real-world crimes (Aluru, Mathew, Saha, & Mukherjee, 2020; Bhatia, et al., 2021). Instances of hate speech grew manifold following changes to Twitter's leadership and policies on censorship and account verification (Ray & Anyanwu, 2022; Frankel & Conger, 2022). In a separate incident, a member of an online community, that provokes gender-based hate crimes, was incarcerated for a planned mass shooting of women (Department of Justice, 2022). These examples illustrate the social importance of safeguarding people's dignity and security against hate speech as we move closer to Society 5.0.

Society 5.0 seeks to create a cyber-physical environment where the interaction of humans and technology leads to improved economic and social well-being (Society 5.0 Conference, 2023). From a human-centred perspective, social media companies should aid in protecting users from online and physical attacks attributed to hate speech. From a technological perspective, they could leverage sophisticated language models to flag and suppress hateful content. Furthermore, they should account for language differences. Two-thirds of the world's population are non-English speaking (Lyne, 2022) and the definition of hate speech varies culturally and geographically (Desphande, Kumar, & Farris, 2022).

Studies on multilingual hate speech detection have investigated coarse and fine-grained classification. With fine-grained classification, posts exhibiting hate speech are further categorised by the type of language used or the target group being attacked (Bhatia, et al., 2021; Dowlagar & Mamidi, 2021; Fortuna, Soler-Company, & Wanner, 2021; Basile, et al., 2019; Chiril, Zitoune, Moriceau, Coulomb-Gully, & Kumar, 2019; Ishman, 2020). However, both applications have been treated as binary classification tasks. Some studies have considered topical hate speech detection as a multi-label task (Chiril, Pamungkas, Benamara, Moriceau, & Patti, 2021). However, these experiments were performed on English datasets. Studies have also used target-specific datasets where samples were collected for the assertions made on a target group (Basile, et al., 2019; Karayiğit, Akdagli, & Aci, 2022; Ishman, 2020). We believe that topically focused hate speech detection should solve multi-class outcomes. A live implementation on target-specific data would also introduce some complexity. Solutions would need to filter content by topic before passing texts through a hate speech classifier. Target-specific datasets are also not representative of the real-world given the low incident rate on SMPs (Madukwe, Gao, & Xue, 2020).

We evaluate the performance of Multilingual BERT (mBERT) and XLM-RoBERTA for topical hate speech detection using target-generic datasets. We use high and low resource languages to fine-tune the models. We note that transformers are more performant in detecting gender-targeted hate speech than our baseline. Our Urdu mBERT classifier produces an F1 score of 0.799 - a 31% lift on the baseline. We observe moderate F1 scores on the Portuguese and Korean classifiers. Class imbalance in the Arabic and French datasets result in weak model fit. We recommend LASER + SVM as an alternative for smaller, imbalanced datasets, while mBERT is suitable for larger samples where class distribution might not be a concern.

We present a pipeline for processing multilingual datasets. Our pipeline could be used to fine-tune transformer models for multi-class topical hate speech detection. To the best of our knowledge, this is an early work on multi-class detection using low resource languages. Our processing and implementation codes can be released on GitHub. We encourage further work on topical hate speech detection for other low resource languages, specifically African languages. We believe our findings are also relevant to real-world implementations and potential challenges to collecting new data. Twitter's APIs will be concealed behind paywalls going forward (Barnes, 2023). This could impede future research due to access and financial constraints. However, transformer models have joint and cross-lingual learning mechanisms that could address some of the challenges.

2 Related Work

Embedding systems and language models have spurred the research on hate speech detection. MUSE and LASER are efficient embedding systems for downstream multilingual tasks. These systems remove the need for language-specific text representations owing to cross-lingual and joint learning mechanisms (Conneau, Lample, Denoyer, Ranzato, & Jegou, 2017; Schwenk, 2019). Some works have demonstrated the effectiveness of MUSE and LASER in multilingual hate speech detection (Aluru, Mathew, Saha, & Mukherjee, 2020; Dowlagar & Mamidi, 2021; Desphande, Kumar, & Farris, 2022). The authors use the embeddings with traditional classifiers as their baseline models to evaluate more sophisticated architectures.

Convolution Neural Networks, Gated Recurrent Networks, and Bidirectional Long Short-Term Memory (BiLSTM) models have been shown to outperform traditional classifiers in both monolingual and multilingual settings (Zhang, Robinson, & Tepper, 2018; Desphande, Kumar, & Farris, 2022). However, these models must be trained for each task. Transformers minimise the time and cost complexities of doing so. Transformers are pre-trained masked language models that use transfer learning to generalise text representations (Devlin, Chang, Lee, & Toutanova, 2019). Early models like BERT base and RoBERTa were trained on English corpora (Casola, Lauriola, & Lavelli, 2022). Multilingual versions such as mBERT and XLM-RoBERTA have since been released. Developers have also released monolingual BERT models, such as German BERT and MahaBERT for Marathi (Velankar, Patil, & Joshi, 2022). However, the multilingual versions are still widely used in literature. The works above performed coarse and fine-grained hate speech detection using neural and transformer models. The models were trained on binary outcomes to distinguish hate speech from neutral speech, and in some cases further identify the type of language.

Our work is concerned with topical hate speech detection. Frenda, Ghanem, Montes, Gomez, & Rosso (2019) used simple text representations and linear models to identify sexist and misogynistic hate speech in separate binary classification tasks. Chiril, Pamungkas, Benamara, Moriceau, & Patti (2021) conducted a comprehensive evaluation of hate speech detection using generic and topic-specific datasets. Their work considered several categories of topical hate speech as a multi-label task. However, both works used English datasets in their models.

Chiril, Zitoune, Moriceau, Coulomb-Gully, & Kumar (2019) examined sexist hate speech on a French corpus. They trained a BiLSTM on GloVe embeddings and FastText vectors. However, severe class imbalance and lack of contextual information resulted in low F1 scores and high misclassification (Chiril, Zitoune, Moriceau, Coulomb-Gully, & Kumar, 2019). SemEval 2019 used a Spanish corpus to detect hate speech against immigrants and women (Basile, et al., 2019). A linear SVM and a BERT model were among the top performing classifiers (Basile, et al., 2019). Participants were given target-specific datasets and both tasks were treated as binary classification.

Karayiğit, Akdagli, & Aci (2022) investigated binary and multi-class classification on a Turkish dataset. They collected posts that contained homophobic remarks (Karayiğit, Akdagli, & Aci, 2022). The authors applied over and under-sampling to correct class imbalances and evaluated various architectures, including ensemble learning. The fine-tuned mBERT model produced an F1 score of 0.90. The authors proposed that the model was performant was due to the size of the Turkish corpus that mBERT was originally trained on (Karayiğit, Akdagli, & Aci, 2022).

We investigate the use of topic-generic datasets for gender-targeted hate speech. While we perform coarse-grained hate speech detection, we are primarily concerned with model performance in multi-class detection. We do not restrict the target variables to a standard definition of hate speech. Rather, we accommodate language differences, annotator judgement, and different manifestations of gender-targeted hate speech. Sexism and misogyny are sometimes used interchangeably when viewing hate speech directed at women (Frenda, Ghanem, Montes, Gomez, & Rosso, 2019).

3 Dataset Descriptions

We used publicly available data from **hatespeechdata.com**. The website is a catalogue of annotated texts for hate speech and offensive language detection (Vidgen & Derczynski, 2020).

3.1 English Datasets

We combine two English datasets with a 5% sampling rate on each. The **Measuring Hate Speech Dataset** contains annotated posts from Twitter, Facebook, and Gab (Kennedy, Bacon, Sahn, & Vacano, 2020). There are 10 ordinal labels for different sentiments and 42 target group attributes. The posts were annotated through crowdsourcing, and a hate speech severity score is calculated from supervised multitask transformer-based deep learning and nonlinear post-processing (Kennedy, Bacon, Sahn, & Vacano, 2020). A severity score of 0.5 or higher represents hate speech (Kennedy, Bacon, Sahn, & Vacano, 2020). We applied this condition to encode our target variables.

The **HateXplain** dataset is extracted from Twitter and Gab (Mathew, et al., 2021). The corpus has fine-grained labels to classify posts as hateful, offensive, or normal, and has 5 target group attributes. The posts were also annotated through crowdsourcing (Mathew, et al., 2021). We used label count and majority vote to encode the outcomes. We tagged a post as hate speech if more annotators labelled it as hate speech than those who labelled it as normal.

3.2 MLMA Arabic and French Datasets

The **Multilingual and Multi-Aspect Hate Speech (MLMA)** dataset was sourced from Twitter in Arabic, French, and English (Ousidhoum, Lin, Zhang, Song, & Yeung, 2019). Three annotators labelled the posts resulting in 5 attributes representing the sentiment, directness, reaction, target, and sub-target group (Ousidhoum, Lin, Zhang, Song, & Yeung, 2019). We used the Arabic and French texts, and the sentiment and target group attributes to encode our outcomes. The sentiment attribute labels texts as normal, hateful, abusive, offensive or a combination thereof. We determined that a post does not have hate speech if any part of the string contained the word ‘normal’.

3.3 ToLD-BR Portuguese Dataset

The **Toxic Language Dataset in Brazilian Portuguese (ToLD-BR)** was retrieved through GATE Cloud’s Twitter Collector (João, Leite, Silva, Bontcheva, & Scarton, 2020). Volunteers were recruited to annotate the posts and labels were aggregated into a toxic count for different target groups (João, Leite, Silva, Bontcheva, & Scarton, 2020). The value in each attribute ranges from 0 to 3 and indicates the number of annotators who flagged the post as toxic (João, Leite, Silva, Bontcheva, & Scarton, 2020). We added the toxic count for each topical attribute. We tagged a post as hate speech where the final toxic count was greater than 1.

3.4 BEEP! Korean Corpus

The **BEEP! Korean Corpus** was collected from Naver (Moon, Cho, & Lee, 2020). Thirty-two annotators were recruited to determine the category of hate speech and the targeted group (Moon, Cho, & Lee, 2020). The authors provided aggregated labels based on inter-annotator agreement using Krippendorff’s alpha (Moon, Cho, & Lee, 2020). We used the hate and bias attributes to encode our target variables. We flagged a post as ‘not hate speech’ if the hate attribute had a value of ‘none’. We also considered a post as gender-targeted if the bias attribute contained the word ‘gender’.

3.5 Romanised Urdu Dataset

The **Hate-Speech and Offensive Language Detection in Roman Urdu** dataset is a collection of Romanised tweets (Rizwan, Shakeel, & Karim, 2020). The dataset can be used for coarse-grained and fine-grained classification. Three independent annotators provided labels, which were numerically encoded by the authors (Rizwan, Shakeel, & Karim, 2020). Values of 1 and 4 refer to normal and profane-untargeted speech, respectively (Rizwan, Shakeel, & Karim, 2020). A value of 3 indicates sexist remarks. We used these guidelines to code our target variables.

Dataset Name	Language	Number of Items	% Hate Speech
Measuring Hate Speech	English	135556	36.2
HateXplain	English	20148	37.8
MLMA Arabic	Arabic	3353	64.3
MLMA French	French	4014	72.0
ToLD-BR Portuguese	Portuguese	21000	44.1
BEEP! Korean Corpus	Korean	7896	55.9
Romanised Urdu	Urdu	7209	40.2

Table 1: Dataset descriptions and statistics before pre-processing.

4 Methodology

Our models were fine-tuned in a monolingual-train and monolingual-test setting. In this setup, each model is trained, validated, and tested on the same language (Aluru, Mathew, Saha, & Mukherjee, 2020; Desphande, Kumar, & Farris, 2022). The processed datasets were partitioned into training and test samples using an 80-20 ratio. We used 10% of the training sample in the validation step for the transformer models.

4.1 Pre-Processing Methodology

We Romanised the MLMA Arabic dataset using Buckwalter transliteration from the ‘lang-detect’ package in Python. We used Academic transliteration from the ‘hangul_romanize’ package for the Korean dataset. We removed special characters, digits, non-ASCII words, and emojis using the ‘clean_text’ package. The English dataset was lemmatised using ‘spacy’. We also removed any words that were fewer than 2 characters. Whitespaces were compressed, and items with less than 3 words or more than 512 characters were deleted due to the maximum character limit in the transformers. Some datasets had masked usernames and URLs, which were removed to avoid any distortion in the embeddings and classification tokens.

4.2 Target Variable Encoding

The definition of hate speech varies across languages due to cultural and geographic factors. Additionally, annotator judgement introduces a degree of error as people may interpret texts differently (Nguyen, et al., 2022). We chose not to relabel the texts on a standard definition of hate speech. Instead, we applied the usage guidelines provided with the datasets. For example, Kennedy, Bacon, Sahn, & Vacano (2020) recommend that texts in the **Measuring Hate Speech Dataset** should be categorised as hate speech if the severity score is greater than 0.5. The usage guidelines for each dataset are briefly mentioned in section 3. We believe that conforming to the data owners’ guidelines captures the various interpretations of hate speech, whether targeted or untargeted. We also observe class distributions that are still representative of the original datasets after encoding the labels across the datasets.

We use a binary outcome for the coarse-grained classification task. We fine-tune our models to distinguish between hate speech and neutral speech. Our target variable is encoded with a 1-0 indicator, with 1 representing hate speech. Referring to the **Measuring Hate Speech Dataset** again as an example, texts with a severity score greater than 0.50 would be assigned a value of 1 in our encoding method.

We use a three-class outcome for topical hate speech detection. We note that gender-related attributes are available across the datasets. While there are different manifestations of gender-targeted hate speech, we do not account for hate speech directed at any specific gender identity. This is largely due to differences in how gender-targeted hate speech is labelled in the datasets. For example, annotators marked texts that were specifically directed at women when labelling the **HateXplain Dataset**, while annotators of the **ToLD-BR Portuguese Dataset** voted on whether a text was misogynistic in nature. The remaining datasets only indicated whether there was gender bias or not. We use values 0 to 2 to represent the outcomes. Texts that were tagged as hate speech in the first round of encoding were further classified as untargeted (value of 1) or gender-targeted (value of 2) hate speech.

Table 2 shows low occurrences of gender-targeted hate speech. This may be due to decisions taken by the data owners to compile the datasets. However, high occurrences of gender-targeted hate speech would not be observed in the real-world. Posts on SMPs are more often neutral than hateful or toxic, constraining sample sizes and class distributions (Song, Huang, & Xiao, 2021). Some authors posit that the natural occurrence of the outcome should not be changed (Madukwe, Gao, & Xue, 2020; Madukwe & Gao, 2019; Fortuna, et al., 2019; Zhang, Robinson, & Tepper, 2018; Davidson, Warmsley, Macy, & Weber, 2017). This approach allows researchers to measure how a model generalises on smaller samples and during a live implementation (Madukwe, Gao, & Xue, 2020). We agree with this rationale and do not augment the class distribution.

Dataset Name	Language	Number of Items	% Hate Speech	% Gender Targeted
English Combined	English	7717	36.4	8.4
MLMA Arabic	Arabic	3286	64.2	7.5
MLMA French	French	3911	71.5	0.5
ToLD-BR Portuguese	Portuguese	20359	44.1	2.3
BEEP! Korean Corpus	Korean	7679	56.0	14.6
Romanised Urdu	Urdu	7125	49.8	8.3

Table 2: Dataset descriptions and statistics after pre-processing.

4.3 Evaluation Criteria

Consistent with other studies, model performance was evaluated on accuracy and the macro-average F1, precision, and recall scores. Accuracy as a single measure of model performance can be misleading when there is class imbalance (Song, Huang, & Xiao, 2021). The macro-average F1, precision, and recall scores overcome this limitation by maximising the true positive rate. The F1 score is the harmonic mean of precision and recall and balances the false negatives and false positives. We give the F1 score precedence when identifying the performant model.

4.4 LASER + SVM Baseline

The LASER embedding system was developed for zero-shot transfer learning and maps sentences to a single vector space regardless of language differences (Schwenk, 2019). Sentences with similar meanings from different languages are embedded close together. LASER, thus, captures various ways that hate speech might manifest in multilingual applications. A further advantage of LASER is that it was designed for joint learning, which makes it a highly efficient system (Schwenk, 2019). We use the ‘laserembeddings’ package to generate sentence-level embeddings.

We use a linear kernel Support Vector Machine (SVM) for classification with the default parameters in ‘scikit-learn’. Each language classifier is trained with 5-fold cross-validation and assessed on the test samples. We set the decision function shape to one-over-one when training our baseline for multi-class classification. Previous studies have used a combination of LASER embeddings with logistic regression (Aluru, Mathew, Saha, & Mukherjee, 2020) or ELMo embeddings with SVMs (Dowlagar & Mamidi, 2021) as baselines. We draw from these works and use a LASER + SVM architecture for our baseline model to evaluate the performance of fine-tuned BERT models.

4.5 mBERT and XLM-RoBERTA Implementation

BERT models are built on a transformer encoder architecture and use a self-attention mechanism to learn word representations (Malik, Pang, & Hengel, 2022). The models can perform classification and sentence prediction on downstream tasks, using the corresponding [CLS] and [SEQ] tokens (Casola, Lauriola, & Lavelli, 2022). We use the [CLS] tokens to fine-tune the models. We access mBERT and XLM-RoBERTA via the HuggingFace transformers library. mBERT is trained on 104 languages from Wikipedia pages, while XLM-RoBERTA is trained on 100 languages from CommonCrawl (Bhatia, et al., 2021). The six languages that we have selected comprise both models’ training corpora.

The transformers were fine-tuned on 5 epochs with a batch size of 64 and a maximum token length of 128. The models were optimised on cross-entropy loss, and the AdamW optimiser parameters were set to the default values - we used a learning rate of 4e-5 and an epsilon of 1e-8. We ran our experiments in Google Colab Pro with an A100-SXM4-40GB GPU and a High-RAM runtime. We set the output dimensions to 2 and 3 for the binary and multi-class classification tasks, respectively.

5 Results

5.1 Coarse-grained Hate Speech Detection

We fine-tuned the models to distinguish hate speech from neutral speech. Our results are presented in Table 3. The evaluation metrics of the best-performing models are highlighted in bold.

We observe that XLM-RoBERTA and mBERT produce higher F1 scores than the LASER + SVM baseline. XLM-RoBERTa is the performant classifier in four languages – English, Arabic, Portuguese, and Urdu. The Urdu XLM-RoBERTa classifier has an F1 score of 0.844, surpassing the baseline by 17%. When fine-tuned on the English and Portuguese datasets, XLM-RoBERTa produces an F1 score of 0.838 and 0.762, respectively. The English and Urdu datasets have similar sample sizes, while the Portuguese dataset has the largest number of items. All three languages have balanced classes and the transformer models show a marginal decrease in the F1 score against accuracy.

We also observe poor performance on all three French classifiers. The F1 scores are lower than the accuracy scores. We attribute the loss in accuracy to the class distribution toward the target outcome – the MLMA French dataset has a 71% occurrence of hate speech. Similarly, the difference between the F1 score and accuracy ranges from 0.01 to 0.10 points on the Arabic classifiers. The MLMA Arabic dataset also exhibits class imbalance, though to a lesser degree than the French dataset.

Model	Dataset Name	F1 Score	Precision	Recall	Accuracy
LASER + SVM Baseline	English Combined	0.788	0.785	0.790	0.805
	MLMA Arabic	0.606	0.613	0.730	0.709
	MLMA French	0.462	0.521	0.779	0.725
	ToLD-BR Portuguese	0.669	0.668	0.675	0.679
	BEEP! Korean Corpus	0.592	0.608	0.661	0.642
	Romanised Urdu	0.721	0.714	0.749	0.748
XLM-RoBERTa	English Combined	**0.838**	**0.838**	**0.837**	**0.850**
	MLMA Arabic	**0.706**	**0.703**	**0.715**	**0.720**
	MLMA French	0.417	0.358	0.500	0.715
	ToLD-BR Portuguese	**0.762**	**0.762**	**0.766**	**0.763**
	BEEP! Korean Corpus	0.629	0.629	0.631	0.632
	Romanised Urdu	**0.844**	**0.843**	**0.845**	**0.850**
mBERT	English Combined	0.835	0.838	0.833	0.848
	MLMA Arabic	0.702	0.709	0.697	0.733
	MLMA French	**0.660**	**0.660**	**0.661**	**0.723**
	ToLD-BR Portuguese	0.745	0.744	0.745	0.748
	BEEP! Korean Corpus	**0.675**	**0.676**	**0.674**	**0.682**
	Romanised Urdu	0.834	0.833	0.835	0.840

Table 3: Evaluation metrics for coarse-grained hate speech detection.

5.2 Topically Focused Hate Speech Detection

The models are fine-tuned to distinguish between neutral speech, untargeted hate speech, and gender-targeted hate speech. Our results are presented in Table 4. Again, we observe that the transformer models outperform the LASER + SVM baseline. However, our fine-tuned mBERT models are more performant in detecting gender-targeted hate speech than XLM-RoBERTa. We observe this on four of the datasets. The best result is produced by the Urdu mBERT classifier across all configurations. The model produces an F1 score of 0.799, which is a 31% lift on the baseline F1 score (0.610).

We also notice a decrease in performance across the board when compared to coarse-grained hate speech detection. The F1 score on the Urdu mBERT classifier decreases by 4% when fine-tuned for multi-class outcomes - the F1 score on the binary classifier was 0.834. In contrast, the F1 score on the French mBERT multi-class classifier (0.448) decreases by 32% against the binary classifier. We note that both transformer models produce weaker outcomes than LASER + SVM on the French dataset.

Model	Dataset Name	F1 Score	Precision	Recall	Accuracy
LASER + SVM Baseline	English Combined	0.654	0.650	0.661	0.773
	MLMA Arabic	0.519	0.509	0.544	0.651
	MLMA French	**0.513**	**0.573**	**0.483**	**0.690**
	ToLD-BR Portuguese	0.517	0.496	0.570	0.661
	BEEP! Korean Corpus	0.478	0.471	0.498	0.523
	Romanised Urdu	0.610	0.590	0.642	0.712
XLM-RoBERTa	English Combined	0.682	0.685	0.682	0.804
	MLMA Arabic	0.493	0.473	0.519	0.720
	MLMA French	0.444	0.443	0.445	0.725
	ToLD-BR Portuguese	**0.656**	**0.672**	**0.646**	**0.750**
	BEEP! Korean Corpus	0.501	0.520	0.518	0.509
	Romanised Urdu	0.774	0.783	0.766	0.825
mBERT	English Combined	**0.708**	**0.715**	**0.702**	**0.817**
	MLMA Arabic	**0.570**	**0.601**	**0.559**	**0.714**
	MLMA French	0.448	0.456	0.444	0.745
	ToLD-BR Portuguese	0.650	0.665	0.637	0.745
	BEEP! Korean Corpus	**0.645**	**0.645**	**0.646**	**0.635**
	Romanised Urdu	**0.799**	**0.792**	**0.808**	**0.829**

Table 4: Evaluation metrics for topically focused hate speech detection.

5.3 Final Remarks and Future Work

Our results suggest that the transformer models produce similar levels of accuracy on high and low resource languages when sample sizes are comparable. We observe this from the models' performance on the Urdu and English datasets. A larger sample size, in the case of the Portuguese classifiers, does not necessarily result in high F1 scores. Yet, XLM-RoBERTa still produces a satisfactory fit on the Portuguese dataset for coarse-grained classification. Overall, we notice that XLM-RoBERTa handles two-class hate speech detection better than the LASER + SVM baseline model on four of the languages. On the other hand, our fine-tuned mBERT model is performant in gender-targeted hate speech detection.

The LASER + SVM baseline model may be a reasonable alternative for smaller, imbalanced datasets. Ousidhoum, Lin, Zhang, Song, and Yeung (2019) may have oversampled the Arabic and French datasets. Their experiments also produced weak F1 scores. The authors could have also introduced selection bias by mainly searching for derogatory comments and controversial topics (Ousidhoum, Lin, Zhang, Song, & Yeung, 2019). Oversampling can result in an overfitted model if it is conducted before partitioning data into training and test samples (Arango, Pérez, & Poblete, 2019). It is recommended that oversampling is performed after sample partitioning (Arango, Pérez, & Poblete, 2019). This closely simulates a live implementation where a model can be fine-tuned on a balanced training set and validated on test data that represents the natural occurrence of hate speech on SMPs.

Future work would consider cross-lingual zero-shot learning for topical hate speech detection in low resource languages. Future work on hate speech detection should also consider African languages, which is under-explored. African languages are under-represented in models like mBERT and XLM-RoBERTa due to challenges in compiling large samples (Ogueji, Zhu, & Lin, 2021; Alabi, Adelani, Mosbach, & Klakow, 2022). Recent state-of-the-art models, like AfriBERTa and AfroXLMR, address these challenges and have exhibited superior performance on downstream NLP tasks (Ogueji, Zhu, & Lin, 2021; Alabi, Adelani, Mosbach, & Klakow, 2022). However, these models have mainly been evaluated on Named Entity Recognition, sentiment analysis, and news classification. Extant research on hate speech detection in African languages continue to experiment with earlier model architectures (Mossie & Wang, 2018; Oriola & Kotzé, 2020; Demilie & Salau, 2022). Thus, a gap in the literature could be addressed by comparing the performance of AfriBERTa and AfroXLMR against mBERT and XLM-RoBERTa in hate speech detection.

6 Conclusion

We evaluated the performance of mBERT and XLM-RoBERTA for topical hate speech detection using target-generic datasets. We observed that the transformer models outperformed the baseline. However, smaller datasets with class imbalance produced a weak fit. In these cases, we recommend LASER + SVM as a suitable alternative. Our results also suggest that mBERT is a superior model for gender-targeted hate speech detection. Our fine-tuned Urdu mBERT classifier produces an F1 score that exceeds the baseline by 31%. The performance is also comparable to the English mBERT mult-class classifier. Our work could be used to improve the pipeline for processing low-resource language datasets for multi-class hate speech detection. We encourage researchers to consider topical hate speech detection for African languages using AfriBERTa and AfroXLMR.

Acknowledgements

We would like to acknowledge funding from the ABSA Chair of Data Science, the TensorFlow Award for Machine Learning Grant. Many thanks to all the anonymous reviewers for their valuable feedback. We also want to acknowledge the assistance of different members of the Data Science for Social Impact research group.

References

Alabi, J., Adelani, D., Mosbach, M., & Klakow, D. (2022). Multilingual Language Model Adaptive Fine-Tuning: A Study on African Languages. *arXiv*.

Aluru, S., Mathew, B., Saha, P., & Mukherjee, A. (2020). Deep Learning Models for Multilingual Hate Speech Detection. *ECML-PKDD 2020*.

Arango, A., Pérez, J., & Poblete, B. (2019). Hate Speech Detection is Not as Easy as You May Think: A Closer Look at Model Validation. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19)* (pp. 45-53). Paris: ACM.

Barnes, J. (2023, February 3). *Twitter Ends Its Free API: Here's Who Will Be Affected*. Retrieved from Forbes: https://www.forbes.com/sites/jenaebarnes/2023/02/03/twitter-ends-its-free-api-heres-who-will-be-affected/

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F., . . . Sanguinetti, M. (2019). SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. *Proceedings of the 13th International Workshop on Semantic Evaluation.* Minnesota, USA.

Bhatia, M., Bhotia, T., Agarwal, A., Ramesh, P., Gupta, S., Shridhar, K., . . . Dash, A. (2021). One to Rule Them All: Towards Joint Indic Language Hate Speech Detection. *ArXiv, abs/2109.13711*.

Casola, A., Lauriola, I., & Lavelli, A. (2022). Pre-trained transformers: an empirical comparison. *Machine Learning with Applications*.

Chiril, P., Pamungkas, E., Benamara, F., Moriceau, V., & Patti, V. (2021). Emotionally Informed Hate Speech Detection: A Multi-target Perspective. *Cognitive Computation*, 322–352.

Chiril, P., Zitoune, F., Moriceau, V., Coulomb-Gully, M., & Kumar, A. (2019). Multilingual and Multitarget Hate Speech Detection in Tweets. *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume II : Articles courts*, (pp. 351–360). Toulouse, France.

Conneau, A., Lample, G., Denoyer, L., Ranzato, M., & Jegou, H. (2017). Word Translation Without Parallel Data. *arXiv preprint arXiv:1710.04087*.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. *International Conference on Web and Social Media.* (pp. 512–515). ICWSM.

Demilie, W., & Salau, A. (2022). Detection of fake news and hate speech for Ethiopian languages: a systematic review of the approaches. *Journal of Big Data*.

Department of Justice. (2022, October 11). *Ohio Man Pleads Guilty to Attempting Hate Crime*. Retrieved from The United States Department of Justice: https://www.justice.gov/opa/pr/ohio-man-pleads-guilty-attempting-hate-crime

Desphande, N., Kumar, V., & Farris, N. (2022). Highly Generalizable Models for Multilingual Hate Speech Detection. *Proceedings of ACM Conference (Conference'17).* New York, NYC, USA.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv, abs/1810.04805*.

Dowlagar, S., & Mamidi, R. (2021). HASOCOne@FIRE-HASOC2020: Using BERT and Multilingual BERT models for Hate Speech Detection. *Forum for Information Retrieval and Evaluation (FIRE).* Hyderabad, India.

Fortuna, A., Chatzakou, D., Kourtellis, N., Blackburn, J., Vakali, A., & Leontiadis, I. (2019). A Unified Deep Learning Architecture for Abuse Detection. *Proceedings of the 10th ACM Conference on Web Science* (pp. 105–114). New York: Association for Computing Machinery.

Fortuna, P., Soler-Company, J., & Wanner, L. (2021). How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*.

Frankel, S., & Conger, K. (2022, December 2). *Hate Speech's Rise on Twitter Is Unprecedented, Researchers Find*. Retrieved from The New York Times: https://www.nytimes.com/2022/12/02/technology/twitter-hate-speech.html

Frenda, S., Ghanem, B., Montes, Y., Gomez, M., & Rosso, P. (2019). Online Hate Speech against Women: Automatic Identification of Misogyny and Sexism on Twitter. *Journal ofIntelligent & Fuzzy Systems*, 4743-4752.

Ishman, A. (2020). Towards Interpretable Multilingual Detection of Hate Speech against Immigrants and Women in Twitter at SemEval-2019 Task 5. *arXiv:2011.13238*.

João, A., Leite, D., Silva, F., Bontcheva, K., & Scarton, C. (2020). Toxic Language Detection in Social Media for Brazilian Portuguese: New Dataset and Multilingual Analysis. *AACL-IJCNLP 2020*.

Karayiğit, H., Akdagli, A., & Aci, C. (2022). Homophobic and Hate Speech Detection Using Multilingual-BERT Model on Turkish Social Media. *Information Technology and Control*, 356-375.

Kennedy, C., Bacon, G., Sahn, A., & Vacano, C. (2020). Constructing interval variables via faceted Rasch measurement and multitask deep learning: a hate speech application. *ArXiv, abs/2009.10277*.

Lyne, C. (2022). *Everyone speaks English, don't they?* Retrieved from Sheffield Hallam University: https://www.shu.ac.uk/about-us/academic-departments/sheffield-business-school/alumni/articles/everyone-speaks-english-dont-they

Madukwe, J., & Gao, X. (2019). The Thin Line Between Hate and Profanity. In J. Liu, & J. Bailey, *AI 2019: Advances in Artificial Intelligence. AI 2019. Lecture Notes in Computer Science()* (pp. 344–356). Springer.

Madukwe, K., Gao, X., & Xue, B. (2020). In Data We Trust: A Critical Analysis of Hate Speech Detection Datasets. *Proceedings of the Fourth Workshop on Online Abuse and Harms* (pp. 151-161). Online: Association for Computational Linguistics.

Malik, J., Pang, G., & Hengel, A. (2022). Deep Learning for Hate Speech Detection: A Comparative Study. *ArXiv, abs/2202.09517.*

Mathew, B., Saha, P., Yimam, S., Biemann, C., Goyal, P., & Mukherjee, A. (2021). HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, (pp. 14867-14875).

Moon, J., Cho, W., & Lee, J. (2020). BEEP! Korean Corpus of Online News Comments for Toxic Speech Detection. *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media* (pp. 25-31). Association for Computational Linguistics.

Mossie, Z., & Wang, J. (2018). Social Network Hate Speech Detection for Amharic Language. *Conference: 4th International Conference on Natural Language Computing (NATL 2018)*, (pp. 41-55).

Nguyen, V., Shi, P., Ramakrishnan, J., Torabi, N., Arora, N., Weinsberg, U., & Tingley, M. (2022). Crowdsourcing with Contextual Uncertainty. *KDD '22: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 3645–3655). New York: Association for Computing Machinery.

Ogueji, K., Zhu, Y., & Lin, J. (2021). Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-Resource Languages. *Proceedings of the 1st Workshop on Multilingual Representation Learning* (pp. 116–126). Association for Computational Linguistics.

Oriola, O., & Kotzé, E. (2020). Evaluating Machine Learning Techniques for Detecting Offensive and Hate Speech in South African Tweets. *IEEE Access*.

Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., & Yeung, D. (2019). Multilingual and Multi-Aspect Hate Speech Analysis. *arXiv:1908.11049*.

Pamungkas, E., Basile, V., & Patti, V. (2021). A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection. *Information Processing & Management*.

Ray, R., & Anyanwu, J. (2022, November 23). *Why is Elon Musk's Twitter takeover increasing hate speech?* Retrieved from Brookings: https://www.brookings.edu/blog/how-we-rise/2022/11/23/why-is-elon-musks-twitter-takeover-increasing-hate-speech/

Rizwan, H., Shakeel, M., & Karim, A. (2020). Hate-speech and offensive language detection in Roman Urdu. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics.

Schwenk, H. (2019). *Zero-shot transfer across 93 languages: Open-sourcing enhanced LASER library.* Retrieved from https://engineering.fb.com/2019/01/22/ai-research/laser-multilingual-sentence-embeddings/

Society 5.0 Conference. (2023). *Society 5.0. Human centeredness in a cyber-physical society*. Retrieved from Society 5.0 Conference: https://www.conference-society5.org/home

Song, G., Huang, D., & Xiao, Z. (2021). A Study of Multilingual Toxic Text Detection Approaches under Imbalanced Sample Distribution. *Information*.

United Nations. (2023). *Understanding Hate Speech*. Retrieved from United Nations: https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech?

Velankar, A., Patil, H., & Joshi, R. (2022). Mono vs Multilingual BERT for Hate Speech Detection and Text Classification: A Case Study in Marathi. *IAPR International Workshop on Artificial Neural Networks in Pattern Recognition*.

Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLoS ONE*.

Zhang, Z., Robinson, D., & Tepper, J. (2018). Detecting hate speech on twitter using a convolution-GRU based deep neural network. *The Semantic Web: Proceedings of the 15th European Semantic Web Conference (ESWC 2018).* Crete, Greece.