Improving Offensive Language Detection for Low Resource Languages

Joanie Weaver
University of California
Berkeley
jcweaver@berkeley.edu

Joshua Noble
University of California
Berkeley
joshuajnoble@berkeley.edu

Abstract

Offensive language detection and other classification tasks often rely on large and well-labeled datasets. Low Resource Languages are languages for which large datasets may not exist or which may have incomplete or noisy datasets to train classification or other NLP tasks. In this project, we use pre-trained multilingual transformer models, mBERT and XLM-R, and a dataset in Amharic, a language spoken by around 50 million people in Ethiopia. In this work, we explore several techniques for utilizing and evaluating transfer learning and data-augmentation to increase performance in classifying social media posts as offensive or not in a Low Resource Language setting. The primary contributions we present in our paper are a data augmentation process and baselines that can provide the foundation for future work for researchers with the proper Amharic language background.

1 Motivation

As the popularity and influence of social media platforms grows, the importance of identifying offensive, illegal, hurtful, or otherwise undesirable posts on social media platforms has prompted a number of studies that range from automatically detecting various kinds of offensive content (e.g. aggression (Kumar et al., 2018), cyber-bullying (Rosa et al., 2019), hate speech (Kumar et al., 2020), etc). Hate speech is a specific type of offensive content that is sometimes legislated as a crime that must be identified and blocked or removed. Perhaps because of this legal imperative, it is by far the most studied phenomenon with several studies published on various languages. Competitions such as HatEval which took place at the SemEval (Rosso et al.) conference in 2019 featured datasets in English and Spanish to address hate speech against women and migrants. The 2020 edition of the same competition used data in Turkish, Danish, Arabic, and Greek. Of equal interest to researchers is offensive language, that is, language which may not be illegal but is deeply undesirable on a platform. The OffensEval (Zampieri et al., 2020) shared task introduced a taxonomy to classify multiple types of offensive content across languages for researchers to identify.

Typically once likely offensive content is identified by an online system, it is flagged for human moderation, which raises the problem of not only identifying offensive language, but also of helping moderators understand what might be offensive in a language which they may not understand well or at all. This makes the situation of a Low Resource Language even more challenging as there may not be adequate data for training models or fluent speakers to help identify and interpret offensive content.

2 Languages and Datasets

We utilized a dataset of 30,000 Amharic comments from Facebook posts on activists' pages created by Surafel Getachew at the Addis Ababa Science and Technology University (Getachew, 2020). To qualify for inclusion in this dataset, all posts must have had at least 300 comments and come from a page with at least 50,000 likes (followers). All of these posts were classified as offensive or not-offensive by a group of 10 annotators from different backgrounds, religions, cultures and ethnic groups. One challenge with this dataset is that the Fleiss' kappa score for agreement amongst the raters was 0.662 (Getachew, 2020). According to Landis and Koch (Landis and Koch, 1977),

0.41 – 0.60 indicates "Moderate agreement" and a score between 0.61 – 0.80 indicates "Substantial agreement". Given that there are only two categories, 0.662 indicates a higher level of disagreement, which may contribute to difficulty in accurately classifying posts.

For fine-tuning we used the OffensEval collection of labeled offensive comments that provided datasets in English, Arabic, Danish, Turkish, and Greek (Zampieri et al., 2020) as well as an Amharic News Text classification dataset from Israel Abebe Azime. (Azime and Mohammed, 2021).

3 Techniques

Different datasets and modeling strategies are often combined to bootstrap classification for low resource settings. We utilize a comparative approach to show how fine-tuning and synthetic data in mBERT and XLM-RoBERTa can improve performance in correctly classifying our Amharic offensive language dataset.

Similar to other transformer architectures, XLM-R transformer architecture can also be used for text classification tasks. XLM-R-base model contains approximately 250M parameters with 12-layers, 768 hidden-states, 3072 feed-forward hidden-states and 8-heads. For our XLM-R explorations, we used a version of XLM-R available on huggingface that was finetuned on 133M amharic data from the CC-100 (Adelani, b). We will refer to this model as *Davlan XLMR*.

mBERT is a multilingual version of BERT that was pre-trained from monolingual corpora in 104 languages (Devlin et al., 2019). Multilingual BERT (mBERT) (Devlin et al., 2019) has shown good cross-lingual performance on several NLP tasks, even without explicit crosslingual signals, although the majority of these have been high resource languages. Additionally, these 104 languages did not include Amharic so we used a version of mBERT available on huggingface that was finetuned on 133M amharic data from the CC-100 (Adelani, a). We will refer to this model as *Davlan mBERT*. We evaluate a few different approaches for fine tuning this model further in order to see the best results for classifying offensive text in our Amharic test set.

4 Evaluation

We will primarily use accuracy as our evaluation metric as this is standard for classification problems to evaluate the measure of all of the correctly identified cases. Since the Amharic dataset contains about 50% of the data in each of the two classes, it's not as necessary to use F1-score. Yet, many of our training and augmentation datasets contain imbalanced data, so we plan to use F1-score to evaluate the harmonic mean of precision and recall as well.

5 Baseline

To establish a baseline, we trained an LSTM model on our Amharic dataset using a 32 dimension LSTM and one 512 dimension hidden layer.

Model	Accuracy	F1	Off. Ratio
Base LSTM	0.626	0.613	0.506

Table 1: Baseline LSTM Performance

We can now compare transformer based architectures to this baseline performance.

6 Transfer Learning

One strategy for working with Low Resource Languages is to rely on transfer learning to make connections classifying offensive speech across languages by using the modeling information present in a higher resource language to a lower resource language. For instance, (Barret Zoph, 2016) used transfer learning for Machine Translation by training a high-resource language pair and then transferring the learned parameters to the low-resource pair to initialize and constrain training. In this section, we train our base mBERT and XLM-R models on the Amharic set and then see how training on the OLID set (English, Danish, Greek, Turkish, and Arabic) improves performance.

6.1 Transfer learning with mBERT

We sought first to ensure that mBERT was capable of any learning with the Amharic set. Then we sought to test whether the addition of the OLID set improved performance through transfer learning. First, we used the Davlan mBERT model and fine-tuned it using just

our offensive Amharic dataset. Then, we compared the performance of that model with the Davlan mBERT that was fine-tuned on both offensive Amharic Dataset and the full OLID dataset:

Finetune	Accuracy	F1	Off. Ratio
Amharic comments	0.633	0.645	0.506
OLID + Amharic comments	0.648	0.673	0.328

Table 2: Davlan mBERT Performance

In this case, we see higher accuracy and F1 scores for both models than the LSTM performance. We also see an additional increase of 0.015 in accuracy and 0.027 in F1-score just from including the OLID data that provides more offensive examples in other languages. This likely means that transfer learning is occurring from the non-Amharic languages in the OLID set to allow the model to use these additional labelled examples to find ways to further differentiate between offensive and not.

Overall, these scores are still quite low, especially compared to the overall ratio of offensive examples in the Amharic dataset. We discuss some analysis of this and a few other approaches we took to increasing these scores later.

6.2 Transfer learning with XLM-R

We used Davlan XLMR and fine-tuned it using just our offensive Amharic Dataset and compared the performance of that model with a version of Davlan XLMR that was fine-tuned on both offensive Amharic Dataset and the full OLID dataset:

Finetune	Accuracy	F1	Off. Ratio
Amharic comments	0.668	0.643	0.516
OLID + Amharic comments	0.660	0.663	0.492

Table 3: XLM-R Performance

As with the mBERT models, using the entire OLID dataset gives a lower accuracy score than just using the Amharic data but a higher F1 score. This indicates that transfer learning is occurring and the model is finding some ways to differentiate between offensive and not. However, the accuracy still leaves significant room for improvement.

7 Data Augmentation

Even in the optimal case, our classifier achieved 66.8% accuracy which highlighted the need for a technique to further increase accuracy. Since one of the key challenges of a Low Resource Language is the lack of data for training and fine-tuning models, we explore a technique for augmenting our existing data to better identify offensive or inoffensive comments from our dataset. Augmentation for LRLs has been explored before, in (Marivate et al., 2020) but these did not leverage a transformer based approach, relying instead on random selection and cosine distance similarity from a word2vec model.

Our approach was to identify the terms unique to comments which were correctly labeled offensive and the terms unique to comments which were correctly labeled inoffensive. We hold these key indicator terms back in order to understand what words are most crucial in a comment that our model should be able to identify correctly. We fine-tuned the Davlan XLM-R dataset a collection of Amharic news articles using a masked-language model task to create a more capable model for filling in the masked tokens. We loop through our dataset and select words which are not the identified key indicator terms to be replaced by a token indicating that our model should fill in the blank. We then generate a new comment which can take the label of the parent comment and which can then be used to augment our dataset.

1. Split the available labeled data into a held-back testing set and a train set
2. Create a list of key terms for each class which are unique to that class from the train dataset
3. For each comment, select a word that is not in key terms list for that comment label (offensive/inoffensive) and replace with the '<mask>' token

4. Use the fine-tuned XLM-R to fill in the blank

5. Re-add generated term to dataset and re-train

6. Test against the held-back dataset

In our case we split our training set of 30,000 Facebook comments into a 29,500 comment training set to be used throughout the training process, and 500 comments to be held back to test our dataset augmented with synthetic comments. We then identify terms most likely to appear in offensive comments, but not in inoffensive comments. From our dataset, we obtained a list of 151 terms that occurred in offensive comments and not in inoffensive ones, and 106 terms that were likely to appear in comments labeled inoffensive, but not offensive ones. Here is an instance of a comment labeled offensive but was incorrectly labeled by our model:

ጦርነቱ ዛሬ ወይስ ነገ ስትሉ የነበራችሁ አማራ ጠሎች እርር ድብን በሉ አሜንአሁንም ቢሆን የአማራ ከፍተኛ የመንግስት አመራሮች ተራርቃችሁም ቢሆን ተሰብስቡተመካከሩ ምክኒያቱም እንጨት እሳትን ያበዘዋልና

The enemies of Amhara, who are on the verge of the war today or tomorrow, apologize. Amen.

We select a list of terms from this comment which is not in our offensive key phrases list and replace it with a token:

ጦርነቱ ዛሬ ወይስ <mask> ስትሉ የነበራችሁ አማራ ጠሎች እርር ድብን በሉ አሜንአሁንም ቢሆን የአማራ ከፍተኛ የመንግስት አመራሮች ተራርቃችሁም ቢሆን ተሰብስቡተመካከሩ ምክኒያቱም እንጨት እሳትን ያበዘዋልና

We then pass this to a mask-fill pipeline to generate replacement tokens for the <mask> token. The original is now transformed into:

ጦርነቱ ዛሬ ወይስ ነገ ስትሉ የነበራችሁ አማራ ጠሎች እርር ድብን በሉ ባሕር ዳር፡ መስ ቢሆን የአማራ ከፍተኛ የመንግስት አመራሮች ተራርቃችሁም ቢሆን ተሰብስቡተመካከሩ ምክኒያቱም እንጨት እሳትን ያበዘዋልና

If the war is on today or tomorrow, the enemies of Amhara should apologize.

And

ጦርነቱ ዛሬ ወይስ ነገ ስትሉ የነበራችሁ አማራ ጠሎች እርር ድብን በሉ ዲስ አበባ ፣ ቢሆን የአማራ ከፍተኛ የመንግስት አመራሮች ተራርቃችሁም ቢሆን ተሰብስቡተመካከሩ ምክኒያቱም እንጨት እሳትን ያበዘዋልና

The enemies of Amhara, who were on the verge of the war today or tomorrow, apologize in Addis Ababa.

These sentences are now added to the original dataset and used to retrain the model. Then the 500 comments that were held back in step 1 are now used to test the model.

This approach runs the risk of over-fitting the dataset, however limited datasets are a necessary feature of working with Low-Resource Languages. This approach should not be seen as replacing the need for more accurate data but rather a way to bootstrap a system into enough functionality that it can function with intermittent human supervision.

7.1 Results

Finetune	Accuracy	F1	Off. Ratio
Synthetic comments	0.683	0.698	0.59
OLID + synthetic comments	0.673	0.727	0.498

Table 4: Synthetic comments with XLM-R

Finetune	Accuracy	F1	Off. Ratio
Synthetic comments	0.653	0.671	0.65
OLID + synthetic comments	0.655	0.666	0.487

Table 5: Synthetic comments with mBERT

Training on synthetic data provided us with a small boost in both accuracy and F1 score in both mBERT and XLM-R models. XLM-R has the best performance thus far, so we will discuss that more in detail. When fine-tuning with synthetic comments using a monolingual model, we see an increase of 0.015 in accuracy and 0.055 in F1 score. When fine-tuning with synthetic comments using a multi-lingual model, we see an increase in accuracy of 0.013 and an increase of 0.064 increase in F1 score. These indicate that while generating realistic synthetic data to augment a dataset is not a cure-all for a Low Resource Language model, it can provide small increases in the ability of a

Table 6: mBERT pretrained with news articles for Masked Language Modeling

Finetune	Accuracy	F1	Off. Ratio
MLM pretrain + Amharic	0.666 (+0.033) [1]	0.684 (+0.038)	0.506
MLM pretrain + OLID + Amharic	0.670 (+0.022)	0.696 (+0.022)	0.328
MLM pretrain + synthetic	0.656 (+0.003)	0.679 (+0.008)	0.65
MLM pretrain + synthetic + OLID	0.652 (-0.003)	0.672 (+0.008)	0.487

model to accurately classify comments. Moreover, since the model used to classify offensive comments is the same as the model that generates the comments, improving the capabilities for one task can improve capabilities in the other.

8 Additional Layer of Finetuning on Amharic News

Since we did not see as great of an increase in the accuracy and F1 scores from finetuning our models with synthetic training data, we tried an alternative approach of adding an additional layer of finetuning using another dataset we found in Amharic. We fine-tuned the Davlan XLM-R and Davlan mBERT using a collection of Amharic news articles (Azime and Mohammed, 2021) on a masked-language model task to create a more capable model for filling in the masked tokens. The models were saved after this finetuning and used to run the same iterations as earlier. Results are shown above in Table 6.

Adding the additional layer of training using masked language modeling on Amharic news sentences did not improve performance for the XLM-R model. For the mBERT model, it increased the scores when the synthetic data was not used in the training. It has previously been shown that XLM-R significantly outperforms mBERT, even performing well in low resource settings, (Conneau et al., 2020). XLM-R had a greater vocabulary and exposure to language data before our pretraining than mBERT, which is likely why we see no performance gains for this added layer of pretraining for XLM-R that we do see in mBERT. Since this news set is not directly related to our offensive classification set, we still don't see as great of performance from mBERT as XLM-R did when using synthetic data to finetune.

9 Language Analysis

We used the XLM-R model with the highest accuracy from finetuning on synthetic comments to output predictions for our reserved Amharic test data. We then compared these predictions to the correct outputs and identified the common words across these categories of: correctly predicted as offensive, falsely predicted as offensive, correctly predicted as inoffensive, and falsely predicted as offensive.

By more closely evaluating the text data, we identified a few challenges that likely make it difficult to form better predictions:

1. Tigrinya, is a Semitic language spoken in Tigray, an area of northern Ethiopia. Amharic, the language of our dataset, uses a version of Ge'ez script, called Fidel. Amharic has a strong influence on Tigrinya, but also differs markedly in a few ways such as having phrasal verbs, and in using a word order that places the main verb last instead of first in the sentence. There are over 40 languages that use versions of Ge'ez script (Titov, 1976). These languages may use very similar alphabets with only the addition or omission of a few letters in Ge'ez script. It's possible that some of the comments in our dataset include some letters, words, or phrases from Tigrinya or other languages using the Ge'ez script as the dataset overview page lists that the main language of pages where comments are taken from must be Amharic and the script must be Ge'ez script but it does not provide a guarantee that all comments or all words are in Amharic. We do not have access to a human or machine translator to cross-check translations in this

language.

2. Improper encoding and/or rendering support working with this data. Some words show up as punctuation or other symbols instead of Ethiopic text. This may be due to the use of punctuation in these phrases but is most likely due to lack of encoding/rendering support on our machines to view these words/characters properly to then get a translation. These were a few examples: ¤, !!, •

3. Since the dataset was pulled from comments from Facebook pages of activists, there are likely many terms, abbreviations, or other references associated with activism or the date of capture in 2020, that differ from initial model pre-training vocabulary developed of the 133M Amharic dataset used for finetuning the mBERT and XLM-R models from the web crawl data during the year of 2018 (Wenzek et al., 2020). For example, the EPRDF, Ethiopian People's Revolutionary Democratic Front, which is a phrase that our model falsely predicted as inoffensive but actually offensive, was dissolved in 2019. This dissolution may impact its use in prior phrases collected during the web crawl of 2018 from its use in Facebook comments in 2020. Additionally, many terms appear to be references to religious or cultural characters that our model may not have picked up from its training sets.

4. Single words can make a big difference in the meaning of the sentence/paragraph and can vary what they mean based on initial look up versus in context. For example, **ምርምር** translates to 'research'. When used in the training set in this paragraph: **ይኸ ቀፈታም ዞምቢ ሰው መስሎት ብዙ ሰው ተሸውዶ ነበር ያኔ በሆርን አፈር ስለመለስ ፅድቅ ሲቀደድ ነበር ዛሬም ከቤተመንግስት እየተከፈለው ከኢንሳ የሚሰጠውን ወሬ እያመላለሰ ከርሱን ይሞላል ጓደኞቹ ስንት ምርምር ያደርጋሉ ዞምቢው ወሬ እያቃጠረ ከርሱን ይሞላል አየህ ቀፈታም ሰው በሆዱ ስትመጣበት አበደ''** which translates to "This is a zombie man. Many people thought he was a jerk. At that time, he was torn to shreds by the horns of Meles Zenawi." Removing the word **ምርምር** from the paragraph results in a translation of "This is a zombie. Many people thought he was a zombie."

5. Amharic has its own punctuation, including a symbol ":" that acts as a word separator. When typing Amharic, many typists use spaces instead. Since this varies, our models may not have consistently picked up on this. It's also possible that our models composed of large amounts of Latin-based scripts have learned to treat ":" differently.

10 Conclusion

Working with the low resource language Amharic, especially in the context of social media comments on activism pages, presents a variety of challenges: access to native speakers or domain experts can be difficult, training corpuses can fit poorly to the classification tasks, and slang, abbreviations, or slurs can be particularly difficult to identify. We particularly learned the importance of having access to native speakers when struggling to understand how the removal of single words from phrases could change the translation so much or whether any words in phrases could mean different things in a slightly different language in the Ge'ez script. We explored how we could improve performance through transfer learning on offensive comments in other languages, augmenting our datasets with synthetic data, and, of particular note, using transformer based architectures which allow the same model to be used for synthetic generation and for classification. Lastly, we explored and compared how pre-training on additional language data affects different pre-trained models differently.

With this paper, we have developed baselines and provided exploratory data notebooks for offensive speech classification in Amharic for future researchers who have access to native speakers or domain knowledge to improve further with their Amharic expertise.

11 Acknowledgements

We'd like to thank the entire MIDS w266 instruction team for their help throughout the term. We'd also like to thank the Masakhane Community for their research and for making datasets widely available.

12 References

References

David Adelani. a. Davlan bert huggingface model.

David Adelani. b. Davlan xlmr huggingface model.

Israel Abebe Azime and Nebil Mohammed. 2021. An amharic news text classification dataset.

Jonathan May Kevin Knight Barret Zoph, Deniz Yuret. 2016. Transfer learning for low-resource neural machine translation.

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.

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.

Surafel Getachew. 2020. Amharic facebook dataset for hate speech detection.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2020. Evaluating aggression identification in social media. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 1–5, Marseille, France. European Language Resources Association (ELRA).

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Vukosi Marivate, Tshephisho Sefara, Vongani Chabalala, Keamogetswe Makhaya, Tumisho Mokgonyane, Rethabile Mokoena, and Abiodun Modupe. 2020. Investigating an approach for low resource language dataset creation, curation and classification: Setswana and sepedi.

Hugo Rosa, Nádia Salgado Pereira, Ricardo Ribeiro, Paula Ferreira, Joao Carvalho, Sofia Oliveira, Luisa Coheur, Paula Paulino, Ana Margarida Veiga Simão, and Isabel Trancoso. 2019. Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, 93:333–345.

Paolo Rosso, Francisco Rangel, Elisabetta Fersini, Debora Nozza, Viviana Patti, Valerio Basile, Cristina Bosco, and Manuela Sanguinetti. Multilingual detection of hate speech against immigrants and women in twitter (hateval).

Evgenij Grigor'evič Titov. 1976. The modern amharic language. In Marvin L. et al. Bender, editor, *The Modern Amharic language*. Moscow Nauka, Moscow.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020).