Hate Speech Detection in Hinglish and Korean Media

Students: Jared Dec, Mili Gera, Devashish Kulkarni

Abstract

Exposure to hate speech leads to damaging psychological effects and can also be associated with physical offline victimization[1]. The majority of all published research in hate speech detection has centered on English[2]. In contrast, we analyzed Hindi and Korean: while the former is linguistically and culturally similar to English, the latter bears no linguistic relation to English and also departs from English by the way in which hate is expressed. The objective of this research was to compare the applicability of various transformer-based language models on two linguistically different languages (as opposed to maximizing a single metric like an F1 score). Furthermore, we postulated the following 1) languages which are linguistically and culturally distinct will likely see varying boosts and/or decay when comparing their mBERT (Multi-Lingual Bidirectional Encoder Representations from Transformers) baseline performance to their performance using newer, less-explored language-agnostic models 2) although the goal of a maximized F1 score is both a needed and worthwhile pursuit, in order to have real-world applicability, we must consider other metrics in order to judge what makes for a "better" model. In deviation from our first hypothesis, our results show that advanced language-agnostic frameworks produce similar or marginally higher F1 scores than more traditional models and they do so at the cost of sacrificing higher scores in recall for higher scores in precision for both Hindi and Korean (i.e languages behaved similarly). What our study does confirm is the need to clearly consider multiple metrics before designating a best model.

Introduction

Hate speech is commonly defined as any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic[4] and has meaningful negative consequences, a prominent example being the suicide of South Korean popstar Sulli[3].

Since the majority of hate speech detection research has focused on English[1], there is a need for hate speech detection techniques in languages other than English. We focus our investigation on Hindi, which is close in both cultural expression and linguistic relation to English, and Korean, which has more subtle hate expression and no linguistic relation to Western languages.

Given these differences we hypothesized that different model architectures may prove to be the "best fit" for each language. Specifically, seeing that out-of-the-box mBERT does not perform well in mapping multilingual sentences of the same meaning to the same vector space, we wanted to see if it would be advantageous to have cross-lingual sentence embeddings which "map" similar meaning sentences from two different languages. This piqued our interest in LaBSE (Language Agnostic BERT Sentence Embedding) to see whether it would better capture Korean subtleties[5]. Similarly, we wanted to see whether layering a mBERT-type model, which is able to create contextualized embeddings, with a CNN (Convolutional Neural Network) model, which extracts important fragments, gives a bigger boost to the more subtle Korean hate speech data. Hence we added a LaBSE with CNN model to our analysis.

To determine our "best fit" we focused on the question of real world applicability, versus a single maximized F1 or accuracy score. That is, to answer the question, which model performed better, we must first ask: "what do we mean by better?" We found that even in research papers which shared precision, recall in addition to F1 scores, there was no key discussion around which metric was being used to determine best performance. By contrast, we propose that the key metric of interest in our analysis is the recall score. That is, in a real world application, we would rather over-flag hate speech than miss

instances of hate speech, due to its negative psychological effects.

Background

Models employed for Korean hate speech detection tended to be variations of BERT (such as KoBERT), CNN, XLM-R, T5 and Electra. The overall best result was from a proprietary model built from scratch for this particular task from a 2021 paper which had an f1 score of 65.55 (See Appendix A for a complete list of past methodologies and scores)[8].

Various approaches have been used for Hindi hate speech detection. This includes transformer-based models such as mBERT, RoBERTa, and non-transformer based models such as CNNs. The top score is from a team at a hate speech detection competition which achieved a Macro F1 score of ~0.78 while precision and recall scores are not discussed or submitted[9]. In addition, a paper from researchers at the Pune Institute of Computer Technology and the Indian Institute of Technology, Madras used the HASOC 2021 dataset for binary classification of Hindi hate speech using various transformer and non-transformer based architectures (such as CNN, BILSTM - Bidirectional Long Short-Term Memory and RoBERTa - Robustly Optimized BERT) [10]. The researchers concluded that the RoBERTa model, with an accuracy score of .80, performed better over all of the non-transformer based models. Since we noticed that in fact it was the CNN model which had the highest recall, precision and even F1 score, we assumed that the definition of better in the case of this study was overall accuracy. None of the approaches above are incorrect, and they in fact provide a range for what may be considered close to the state-of-the-art results. However, our research is not directly comparable because firstly, due to our prior knowledge of the linguistic and cultural differences between our two chosen languages, our interest lay in whether different architectures may work better for these two languages. To our knowledge, an attempt to compare the effectiveness of standard NLP models when applied to languages with significantly different cultural and linguistic properties has not been conducted, at least with Korean as one of the studied languages. In addition, because we aim for a deeper study, we prioritized the recall score over precision and F1 scores.

Data

Korean hate speech analysis came from a dataset of 10,000 annotated comments on entertainment news websites by researchers at Seoul National University[6]. Our Hindi dataset is sourced from Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages[7].

To ensure that the models learn functions that detect meaningful features of hate speech, we pre-processed and balanced all the datasets to ensure the number of hate and non-hate comments are equal. In addition, for the scarcely available Hindi data, we combined the datasets from HASOC 2020 and 2021. Hence, our results cannot be directly compared to the literature. The Hindi dataset was annotated by two annotators, with conflicts between the annotators being resolved by a third annotator[9]. The Korean dataset was annotated by 32 annotators, with each comment being seen by three random annotators where the final designation of hate speech or not was decided by majority rule[6].

The final data used in our experiments contains 5,294 examples in Hindi, 7,292 examples in Korean and 30,588 examples in English (used for zero-shot models discussed below). Each of the datasets was split into 80:10:10 as train, development and test sets, respectively. We also conducted a qualitative analysis of both the Hindi and Korean datasets. This confirmed what we expected such as the use of “Hinglish” phrases or Hindi transliterated into the Latin alphabet rather than the traditional Devanagari script. In addition, many of the tweets in the dataset contain some English words intermixed with native Hindi words. We also noted the subtlety of Korean hate speech. An example from the Korean hate speech dataset is the following comment which was annotated as hate by the dataset’s compilers “설마 ㅈ[sic] 현정

작가 아니지??” . This approximately translates to, “There’s no way Hyeonjeong is really a writer, is there?” . The subtlety of Korean hate speech may possibly be related to the fact that the language itself is highly structured to convey different levels of respect and formality with every verb conjugation[11]. Many examples flagged as hate speech in Korean consist of swear-less observations about someone in lower formality, showing less respect. In sharp contrast, Hindi hate speech consists of more explicit insults that more resemble a Western language.

Methods

Baseline mBERT

Description: Since mBERT (or a monolingual representation of mBERT such as KoBERT) is one of the state-of-the-art techniques for both Korean and Hindi hate speech detection, we use mBERT results for both Korean and Hindi as our baseline from which we would conduct further analysis.

Architecture: Our baseline model uses the pre-trained contextualized representations of our input from mBERT. We retrieve the [CLS] representations which are subsequently propagated through multiple dense activation layers and are finally classified with sigmoid activation. We use the ADAM optimizer, and layer normalization and dropout for the dense layers.

Hyperparameter Tuning: Hyperparameter tuning was performed by adjusting learning rate, dropout for the hidden layers, freezing/unfreezing weights of mBERT layers and hidden layers size. The best performance was obtained for the highest learning rate (0.001) and for the largest hidden layers size (256, 128), suggesting that performance may be further improved by increasing the complexity of the models. A dropout of 0 performed best on the Hindi dataset and a dropout of 0.15 performed best on the Korean dataset. This may be a result of the difference in the number of examples in each dataset, with the larger Korean dataset being more susceptible to overfitting. The weights of transformer layers were frozen in the best performing models, likely pointing to the benefit of preserving the pretrained language understanding in the transformer model. Early stopping with a patience of 3 was used with the objective of maximizing the F1 score on the development datasets. The training/test configurations for the mBERT models are listed in Table 1.

Table 1-Task Training/Test Variation for mBERT-based models

Model Name	Train Data	Test Data
mBERT baseline	Hi or Ko.	Hi or Ko
mBERT zero-shot	En	Hi or Ko
mBERT task-training*2	En + Hi or Ko	Hi or Ko
mBERT translated	En	Hi or Ko translated to En
mBERT task-train*2 + translate*2	En + translated Hi or Ko	Hi or Ko translated to En

1. Hi - Hindi 2. Ko - Korean 3. En - English

Language-agnostic BERT Sentence Embedding (LaBSE)

Architecture: Our LaBSE model uses the pre-trained contextualized representations of our input from LaBSE. We retrieve the “pooled_output” which represents each input sequence as a whole. The “pooled_output” is subsequently propagated through multiple dropout layers and dense layer activations and is then finally classified with sigmoid activation. We use the ADAM optimizer, a loss function of binary cross entropy, and layer normalization for the dense layers.

Hyperparameter Tuning: Hyperparameter tuning was performed on learning rate, dropout rate for the hidden layers and hidden layers size. We kept the following two configurations for our final analysis phase.

Configuration one, which we named LaBSE baseline hyperparameters, had settings which matched our baseline hyperparameters. Namely, a learning rate of .001, two dense layers with

size 256, and 128 and 0 dropout rate. Configuration two, which we named LaBSE optimized parameters, consisted of a learning rate of .001, two dense layers with size 200 and 128 and two-dropout layers with rate 0.1 and 0.2. The training/test configurations for the LaBSE models are listed in table 2.

Table 2-Task Training/Test Variation for LaBSE-based models

Model Name	Train Data	Test Data
LaBSE baseline hyperparameters	Hi or Ko	Hi or Ko
LaBSE baseline hyperparameters + zero-shot	Eng	Hi or Ko
LaBSE optimized hyperparameters	Hi or Ko	Hi or Ko
LaBSE optimized hyperparameters + zero-shot	Eng	Hi or Ko

1. Hi - Hindi 2. Ko - Korean 3. En - English

Hybrid LaBSE/Convolutional Neural Network (CNN)

Our hybrid LaBSE/CNN model uses pre-trained contextualized representations of our input from LaBSE (version 2). Unlike the LaBSE-only model, the hybrid model retrieves the "sequence_output" which retrieves each input token in the context. The embeddings are then passed into convolutional filters of five different sizes: 768 X 1, 768 X 2, 768 X 3,768 X 4, 768 X 5 . There are 32 filters for each size. The output is passed through a ReLU activation function, and then is global max-pooled. The pooling is concatenated and passed through a dropout and a dense layer and then a sigmoid activation is performed for the activation.

Hyperparameter Tuning: Our filter number and size settings as well as our learning rate settings comes from a successful experiment on hate speech detection in Arabic using a hybrid mBERT/CNN model. The learning rate we used was .00002 and the dropout rate was 0.7[12]. The training/test configurations for the LaBSE+CNN models are listed in table 3.

Table 3-Task Training/Test Variation for LabSE+CNN models

Model Name	Train Data	Test Data
LaBSE with CNN	Hi or Ko	Hi or Ko
LaBSE with CNN + zero-shot	En	Hi or Ko

1. Hi - Hindi 2. Ko - Korean 3. En - English

Results

All zero-shot models performed quite poorly compared to non-zero shot models, hence the high-level results for those models are included in Appendix B only. This makes intuitive sense as we expect that neither Hindi nor Korean models would achieve enough transfer-learning through English task-trained baseline mBERT or LaBSE in order to outperform the models specifically tasked-trained in the language of interest. This was true even though the English dataset was three times larger than both the Hindi and Korean datasets. We examined preprocessing of Hindi tweets by replacing usernames with a special token, removing urls and emojis, however, this resulted in a reduction of the model performance. This suggests the information contained in these is useful for the model for detecting hate speech. We also explored translations of Hindi and Korean to English and testing these on a model trained on a much larger English data set. However, the model performance did not improve upon the baseline on untranslated text and we chose not to explore these methods further. Translations of the Hindi and Korean datasets to English were obtained through two systems, Google Translate and Helsinki NLP transformer-based models[13]. Our review of the translations determined Google Translate was the more reliable system.

Listed in Table 4 are the confusion matrices for both Hindi and Korean for the models which had the highest F1 scores. As stated earlier, having the singular goal of maximum F1 scores was not our desire. Therefore, one obvious pattern we

noticed right away was that even though the Hindi LaBSE optimized hyperparameter model (with highest F1 score) outperformed the Hindi mBERT baseline model (with the lowest F1 score) by a ~2 percent gain in the F1 score, it did so at a great expense to the recall score. In fact the recall dropped by ~13% in the LaBSE model even though the precision went up by ~14%. Similarly for the Korean language, even though there was a ~ 6% difference in the highest and lowest F1 scores, there is a ~9% percent drop in recall when we compared the LaBSE with CNN model (highest F1 score) to the LaBSE baseline hyperparameter model (lowest F1 score).

The scores below suggest that the mBERT baseline models outperformed other models for both Hindi and Korean. For Korean, the mBERT model had a .8% lower F1 score than the highest F1 scoring Korean model (LaBSE with CNN), but a 13% higher recall. For Hindi, the mBERT model had an average of ~14% recall gain as compared to all other models while losing only ~2% from its F1 score. This was a departure from our initial hypothesis that culturally and linguistically different languages will have different "best-fit" models.

Hindi False Negative Analysis

The mBERT model shared 30% of its predicted false negatives with all other models (i.e. all models also predicted a false negative for the same examples). 13% of the ones predicted by the mBERT model as false negatives were predicted as true positives by all the LaBSE models. In the end, however, the baseline mBERT model had a ~50% better performance on the detection of false negatives as compared to all other models.

All of the LaBSE-based models predicted the same examples as false negatives greater than 45% of the time (39 out of the max of 86 instances of false negatives for any given LaBSE model). Furthermore, approximately 71% of these LaBSE-shared false negatives seem to be subtle hate speech with some with more of a sarcastic lean. For example, one tweet reads: "RT @ColdCigar: "इंजीनियर लोग शायद सड़क पे डिवाइडर बनाना भूल गए, देश के लिये मैं खुद ही खड़ा हो जाता हूँ।" #DividerInChief https://t.co/zFb…". This is roughly translated as, "Engineers probably forgot to make dividers on the road. I will stand up for my country myself". Roughly 28% of these shared false negatives, however, seem to contain pretty direct hate speech content. For example, a hate-filled tweet reads: "हथियार चलाना और मारना सीख लो हिन्दुओं, हर मोड पर दुशमन वैठा है, कहां तक भागोगे #BengalBurning #HinduGenocideBengal #HindusLivesMatter". The translation for this is, "Hindus, learn to use weapons and kill. Every turn has enemies. How far will you run". For the rest of the 55% of examples where the LaBSE models predicted either false negative or true positive, there were no obvious patterns as to why certain models turn up true positives while others predict false negatives. The tweet, "@simon robin @yadavakhilesh अरे यह बस सैफई में रंडी ही नचा सकता है [translated as, this guy can only make a prostitute dance in his defense]", was predicted as a false negative by the LaBSE baseline hyperparameter model but was caught as a true positive by both the LaBSE optimized hyperparameter model and the LaBSE with CNN model.

From the high-performance of the mBERT model and the varied performance among the LaBSE models, we could not conclude that the LaBSE models caught subtle Hindi hate speech data any better.

Korean False Negative Analysis

The mBERT model shared 42% of its predicted false negatives with all other models (i.e. all models also predicted a false negative for the same examples). 18% of the ones predicted by the mBERT model as false negatives are predicted as true positives by all LaBSE models. In the end, however, the baseline mBERT model had a 39% or greater performance on the detection of false negatives as compared to all other models.

For Korean data, all of the LaBSE-based models predict the same examples as false negatives greater than 39% of the time (56 out of the max of 142 instances of false negatives for any given LaBSE model). Although we did not observe a

clear breakdown of subtle versus non-subtle hate speech, we did notice that LaBSE missed many blatantly obvious hate speech examples. For example, the following comment was labeled as a false negative: "판사 음주운전 차량에 치어 뒤지길 빕니다" which roughly means "I hope the judge gets run over by a drunk driver". A possible reason this sentence is misclassified could be because the writer is using the highest possible formality when talking about the judge in question. In that sense, LaBSE could be capturing subtleties in the Korean language which in fact increase false negatives rather than decrease them. Seeing the high-performance of the mBERT model above and the varied performance among the LaBSE models, we could not conclude that the LaBSE models caught subtle Korean hate speech data any better.

Table 4-Confusion Matrices

Hindi	**TP**	**FP**	**TN**	**FN**	**Pr[1]**	**Re[2]**	**F1[3]**
mBERT baseline	229	151	104	46	0.60	0.83	0.70
LabSE baseline hyper-parameters	189	63	192	86	0.75	0.69	0.72
LaBSE optimized hyper-parameters	192	65	190	83	0.75	0.70	0.72
LaBSE with CNN	189	70	185	86	0.73	0.69	0.71
Korean	**TP**	**FP**	**TN**	**FN**	**Pr[1]**	**Re [2]**	**F1[3]**
mBERT baseline	299	194	175	62	0.61	0.83	0.70
LaBSE baseline hyper-parameters	219	100	269	142	0.69	0.61	0.64
LaBSE optimized hyper-parameters	260	128	241	101	0.67	0.72	0.69
LaBSE with CNN	252	99	270	109	0.72	0.70	0.71

1. Pr - Precision Rate, 2. Re - Recall Rate, 3. F1 - F1 Score (test)

Conclusions

This project investigated various techniques for hate speech detection in Hindi and Korean. We tested transformer-based models such as mBERT and LaBSE and focused on recall as our primary metric. We found that a strong baseline can be achieved using the mBERT model for both Korean and Hindi. While the LaBSE frameworks performed equally or marginally better in aggregate with both Hindi and LaBSE, they did so at the expense of worsening the recall rate for gains in the precision rate. The Hindi LaBSE models had great variation amongst each other in terms of the examples predicted as false negatives and we therefore could not deduce any meaningful linguistic patterns which led to this result. Text that used formal speech in a sarcastic or deriding manner was still labeled as not hate speech in some cases with the Korean dataset with a LaBSE framework whereas the mBERT framework correctly classified these examples as hate speech. Notably, both Korean and Hindi achieved the optimal balance in recall and F1 scores using the mBERT model. In addition, the language-agnostic framework did not produce noticeably better results with the Korean data. This meant that the original hypothesis that these languages may have varying benefits due to LaBSE frameworks was not in fact, true. However, based on the hyperparameter tuning process, it is possible that neither of these frameworks were optimized as more complex models seemed to perform better at these tasks. Furthermore, access to more labeled data likely would have contributed to better performance. In conclusion, mBERT-based frameworks still appear to be the superior framework for hate speech detection if recall is the primary metric under consideration. Future work might include investigating character-based models, such as ByT5, to have an approach that is less dependent on the use of formal vocabulary in the training data.

References

[1] Saha, K., Chandrasekharan, E. & De Choudhury, M. “Prevalence and psychological effects of hateful speech in online college communities.” In *Proceedings of the 10th ACM Conference on Web Science*, 255–264 (2019), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7500692/

[2] Jahan, Md Saroar & Oussalah, Mourad. “A systematic review of Hate Speech automatic detection using Natural Language Processing.” In *ArXiv* (2021), https://arxiv.org/pdf/2106.00742.pdf

[3] “Korean Pop Star Sulli Dies at 25.” *NBCNews.com*, NBCUniversal News Group, 14 Oct. 2019, https://www.nbcnews.com/pop-culture/music/korean-pop-star-sulli-dies-25-n1065651.

[4] Oksanen, Atte & Hawdon, James & Holkeri, Emma & Näsi, Matti & Räsänen, Pekka. “Exposure to Online Hate among Young Social Media Users.” In *Soul of Society: A Focus on the Lives of Children & Youth (Sociological Studies of Children and Youth, Volume 18)* 253 - 273 (2014), https://www.researchgate.net/publication/266392546_Exposure_to_Online_Hate_among_Young_Social_Media_Users

[5] Yang, Yinfei, and Fangxiaoyu Feng. “Language-Agnostic Bert Sentence Embedding.” *Google AI Blog*, Google Research, 18 Aug. 2020,

[6] Moon, Jihyung, Won Ik Cho, and Junbum Lee. "BEEP! Korean corpus of online news comments for toxic speech detection." In *ArXiv* (2020) preprint arXiv:2005.12503. https://arxiv.org/pdf/2005.12503.pdf

[7] “HASOC (2021).” *HASOC*, Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages, https://hasocfire.github.io/hasoc/2021/dataset.html.

[8] Lee C, Yang K, Whang T, Park C, Matteson A, Lim H. “Exploring the Data Efficiency of Cross-Lingual Post-Training in Pretrained Language Models.” *Applied Sciences*. 2021; 11(5):1974. https://doi.org/10.3390/app11051974

[9] Mandl, Thomas, et al. "Overview of the HASOC subtrack at FIRE 2021: Hate speech and offensive content identification in English and Indo-Aryan languages." In *arXiv* (2021) preprint arXiv:2112.09301, https://arxiv.org/pdf/2112.09301.pdf

[10] Velankar, Abhishek, et al. “Hate and Offensive Speech Detection in Hindi and Marathi.” *ArXiv.org*, ArXiv, 1 Nov. 2021, https://arxiv.org/abs/2110.12200.

[11] Brown, Lucien & Whitman, John. (2015). “Honorifics and politeness in Korean.” *Korean Linguistics.* 17. 127-131. 10.1075/kl.17.2.001int.

[12] Safaya, A., Abdullatif, M. & Yuret D. “KUISAIL at SemEval-2020 Task12: BERT-CNN for Offensive Speech Identification in Social Media.”*arXiv:2007.13184* (2020), https://arxiv.org/abs/2007.13184

[13] “Helsinki-NLP/Opus-Mt: Open Neural Machine Translation Models and Web Services.” *GitHub*, Helsinki-NLP, 21 Aug. 2020, https://github.com/Helsinki-NLP/Opus-MT.

Appendix A

A complete list of the prior published papers that cited the Korean hate speech data used for our study is listed below. There were an additional 21 papers that cited the dataset but did not attempt to predict hate speech labels on the data. Please note that due to early tendencies in our model to assign all of one label to the Korean data, we found it necessary to make the allocation of hate and non-hate speech more balanced between the training and validation datasets compared to the original allocation. Thai means that our results are not directly comparable with the results listed here.

Name of Paper	**Methodologies Used**	**Best Score**
BEEP! Korean Corpus of Online News Comments for Toxic Speech Detection (2020)	KoBERT, CharCNN, Bidirectional LSTM	52.5 (KoBERT)
Exploring the Data Efficiency of Cross-Lingual Post-Training in Pretrained Language Models (2021)	mBERT, XLM-R, KoBERT	64.68 (KoBERT)
Exploring the Data Efficiency of Cross-Lingual Post-Training in Pretrained Language Models (2021)	3 different models created by team for this task specifically	65.55 (proprietary model trained from scratch for this task)
A Study on the Toxic Comments Classification Using CNN Modeling with Highway Network and OOV Process (2020)	CNN	62.25
딥러닝 기술을 활용한 악성댓글 분류: Highway Network 기반 CNN 모델링 연구 (2020)	CNN	62.25
Hate Speech Classification Using Ordinal Regression (2021)	LSTM-sigmoid, LSTM-cor	59 (LSTM-cor)
The Korean Morphologically Tight-Fitting Tokenizer for Noisy User-Generated Texts (2021)	3-layer CNN with different tokenizer methodologies	56.69 (3-layer CNN with Mecab-ko tokenizer)
A Model of Cross-Lingual Knowledge-Grounded Response Generation for Open-Domain Dialogue Systems (2021)	KE-T5 (proprietary T5 model made for translating Korean to English)	64.14 (KE-T5 base)
Hate Speech Detection in Chatbot Data Using KoELECTRA (2021)	Koelectra	62.7
Characterization and mechanical properties of offensive language taxonomy and detection techniques (2021)	CharCNN, BiLSTM, KoBERT	63.3

Appendix B

Our initial modeling attempts with various model configurations and training procedures are tabulated below.

Hindi	Train	$F1^3$ Test Hindi	$F1^3$ Test Korean
mBERT baseline	Hi or Ko.	**0.70**	**0.70**
mBERT zero-shot	En	0.41	0.58
mBERT task-training*2	En + Hi or Ko	0.72	0.70
mBERT translated (translate test)	En	0.38	0.54
mBERT task-train*2 + translate*2 (translate train and test)	En + translated Hi or Ko	0.71	0.69
LabSE baseline hyper- parameters	Hi or Ko	0.72	0.64
LaBSE baseline hyperparameters + zero-shot	Eng	0.57	0.53
LaBSE optimized hyper- parameters	Hi or Ko	0.72	0.69
LaBSE optimized hyperparameters + zero-shot	Eng	0.62	0.54
LaBSE with CNN	Hi or Ko	0.71	0.71
LaBSE With CNN + zero-shot	Eng	0.66	0.64

1. Hi - Hindi 2. Ko - Korean 3. En - English

We observe that the baseline mBERT model achieves strong F1 scores for both the languages of ~0.7. LaBSE with optimized hyperparameters provides a slight improvement over the Hindi dataset (0.7218). Interestingly, the performance of LaBSE on Korean is slightly worse (~0.69). The zero shot performance of all the models is significantly worse than the baselines, suggesting that the information in the English data set isn't sufficient for detecting toxic speech in Hindi/Korean. However, training the mBERT models on English data and subsequently training on the Hindi data set improves the model performance on the Hindi test set(0.721). This suggests that the model does learn useful information from the English data set, but needs to see the Hindi data as well to perform well on the Hindi test set. A similar procedure does not improve upon the baseline score on the Korean data set. This suggests that the information learnt from the English data set is not useful for detecting hate speech in Korean, possibly because Hindi and English are a part of the same language family, while Korean is a language isolate and there exist cultural differences in how hate speech is expressed in these languages.

We explore another approach to utilize the information in the much larger English data set by testing the models on translations of the original Hindi/Korean data to English. We find that simply training the models on the English data set and testing on translated test sets leads to a very poor performance. In contrast, training the models on the translated training sets after being trained on English gives a better than baseline performance on Hindi, and slightly worse than baseline performance on Korean. Note, that the translations themselves are not perfect and may not capture the entire meaning og the original sentence. Again, we find that the information from the English data set improves performance on the Hindi data set than Korean.