Mind Your Language: Abuse and Offense Detection for Code-Switched Languages

Raghav Kapoor
MIDAS Lab, NSIT-Delhi
raghavk.co@nsit.net.in

Yaman Kumar
MIDAS Lab, IIIT-Delhi
yaman@nsitonline.in

Kshitij Rajput
MIDAS Lab, NSIT-Delhi
rajput.kshitij97@gmail.com

Rajiv Ratn Shah
MIDAS Lab, IIIT-Delhi
rajivratn@iiitd.ac.in

Ponnurangam Kumaraguru
IIIT, Delhi
pk@iiitd.ac.in

Roger Zimmermann
NUS, Singapore
rogerz@comp.nus.edu.sg

Abstract

In multilingual societies like the Indian subcontinent, use of code-switched languages is much popular and convenient for the users. In this paper, we study offense and abuse detection in the code-switched pair of Hindi and English (i.e, Hinglish), the pair that is the most spoken. The task is made difficult due to non-fixed grammar, vocabulary, semantics and spellings of Hinglish language. We apply transfer learning and make a LSTM based model for hate speech classification. This model surpasses the performance shown by the current best models to establish itself as the state-of-the-art in the unexplored domain of Hinglish offensive text classification. We also release our model and the embeddings trained for research purposes.

Introduction

With the penetration of Internet among masses, the content being posted on social media channels has uptaken. Specifically, in the Indian subcontinent, number of Internet users has crossed 500 mi[1], and is rising rapidly due to inexpensive data[2]. With this rise, comes the problem of hate speech, offensive and abusive posts on social media. Although there are many previous works which deal with Hindi and English hate speech (the top two languages in India), but very few on the code-switched version (Hinglish) of the two (Mathur et al. 2018). This is partially due to the following reasons: (1) Hinglish consists of no-fixed grammar and vocabulary. It derives a part of its semantics from Devnagari and another part from the Roman script. (2) Hinglish speech and written text consists of a concoction of words spoken in Hindi as well as English, but written in the Roman script. This makes the spellings variable and dependent on the writer of the text. Hence, code-switched languages present tough challenges in terms of parsing and getting the meaning out of the text. For instance, the sentence, *"Modiji foreign yatra par hai"*, is in the Hinglish language. Somewhat correct translation of this would be, *"Mr. Modi is on a foriegn tour"*. However, even this translation has some flaws due to no direct translation available for the word *ji*, which is used to show respect. Verbatim translation would lead to *"Mr. Modi foreign tour*

Hinglish	English	Hinglish	English
acha	good	gunda	thug
s**la	blo*dy	ra*di	h*oker

Table 1: Examples of word-pairs in Hinglish-English dictionary

on is". Moreover, the word *yatra* here, can have phonetic variations, which would result in multiple spellings of the word as *yatra, yaatra, yaatraa*, etc. Also, the problem of hate speech has been rising in India, and according to the policies of the Indian government and the various social networks, one is not allowed to misuse his right to speech to abuse some other community or religion. Due to the various difficulties associated with the Hinglish language, it is challenging to automatically detect and ban such kind of speech.

Thus, with this in mind, we build a transfer learning based model for the code-switched language Hinglish, which outperforms the baseline model of (Mathur et al. 2018). We also release the embeddings and the model trained at `https://github.com/dhumketu/mind-your-language-aaai`.

Methodology

Our methodology primarily consists of: pre-processing of the dataset, training of word embeddings, training of the classifier model and then using that on HEOT dataset.

Pre-Processing

In this work, we use the datasets released by (Davidson et al. 2017) and HEOT dataset provided by (Mathur et al. 2018). The datasets obtained pass through these steps of processing: (i) Removal of punctuations, stopwords, URLs, numbers, emoticons, *etc.* This was then followed by transliteration using the Xlit-Crowd conversion dictionary [3] and translation of each word to English using Hindi to English dictionary[4]. To deal with the spelling variations, we manually added some common variations of popular Hinglish words. Final dictionary comprised of 7200 word pairs. Additionally, to deal with profane words, which are not present in

[1] https://bit.ly/2MCXz2Q
[2] https://bit.ly/2NUjOGh

[3] http://bit.ly/2OGL872
[4] http://bit.ly/2PYPSK8

Category	Tweet	Translation
Benign	sache sapooto aap ka balidan hamesha yaad rahega	True sons, your sacrifice would be remembered.
Hate Inducing	Bik gya Porkistan	Porkistan (Derogatory term for Pakistan) has been sold
Abusive	Kis m*darch*d ki he giri hui harkt	Which m*therf*cker has done this

Table 2: Examples of tweets in the dataset and their translations

Model	Accuracy
Davidson et al.	0.57
Our Model with embeddings trained on Glove	**0.87**
Our Model with embeddings trained on Word2Vec	0.82
Our Model with pre-trained Word2Vec embeddings	0.59
Mathur et al	0.83

Table 3: Comparison of accuracy scores on HEOT dataset

Model	Accuracy
Davidson et al.	**0.90**
Our Model with embeddings trained on Glove	0.89
Our Model with embeddings trained on Word2Vec	0.86
Mathur et al.	0.75

Table 4: Comparison of accuracy scores on (Davidson et al. 2017) dataset

Xlit-Crowd, we had to make a profanity dictionary (with 209 profane words) as well. Table 1 gives some examples from the dictionary.

Training Word Embeddings

We tried Glove (Pennington, Socher, and Manning 2014) and Twitter word2vec (Godin et al. 2015) code for training embeddings for the processed tweets. The embeddings were trained on both the datasets provided by (Davidson et al. 2017) and HEOT. These embeddings help to learn distributed representations of tweets. After experimentation, we kept the size of embeddings fixed to 100.

Classifier Model

Both the HEOT and (Davidson et al. 2017) datasets contain tweets which are annotated in three categories: offensive, abusive and none (or benign). Some examples from the dataset are shown in Table 2. We use a LSTM based classifier model for training our model to classify these tweets into these three categories. An overview of the model is given in the Figure 1. The model consists of one layer of LSTM followed by three dense layers. The LSTM layer uses a dropout value of 0.2. Categorical crossentropy loss was used for the last layer due to the presence of multiple classes. We use Adam optimizer along with L2 regularisation to prevent overfitting. As indicated by the Figure 1, the model was initially trained on the dataset provided by (Davidson et al. 2017), and then re-trained on the HEOT dataset so as to benefit from the transfer of learned features in the last stage. The model hyperparameters were experimentally selected by trying out a large number of combinations through grid search.
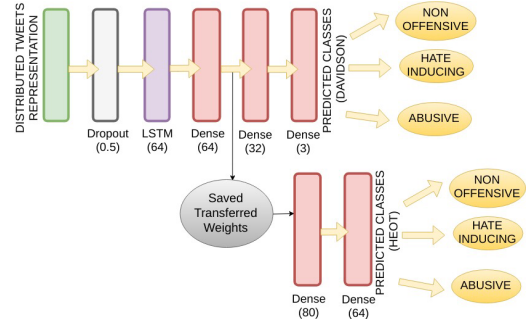

Figure 1: LSTM based model for tweet classification

Results

Table 3 shows the performance of our model (after getting trained on (Davidson et al. 2017)) with two types of embeddings in comparison to the models by (Mathur et al. 2018) and (Davidson et al. 2017) on the HEOT dataset averaged over three runs. We also compare results on pre-trained embeddings. As shown in the table, our model when given Glove embeddings performs better than all other models. For comparison purposes, in Table 4 we have also evaluated our results on the dataset by (Davidson et al. 2017).

Conclusion

In this paper, we presented a pipeline which given Hinglish text can classify it into three categories: hate-inducing, abusive and benign. This LSTM based model performs better than the other systems present. We also release the code, the dictionary made and the embeddings trained in the process. We believe this model would be useful in hate speech detection tasks for code-switched languages.

References

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. *arXiv preprint arXiv:1703.04009*.

Godin, F.; Vandersmissen, B.; De Neve, W.; and Van de Walle, R. 2015. Multimedia lab @ acl wnut ner shared task: Named entity recognition for twitter microposts using distributed word representations. In *Proceedings of the Workshop on Noisy User-generated Text*, 146–153.

Mathur, P.; Shah, R.; Sawhney, R.; and Mahata, D. 2018. Detecting offensive tweets in hindi-english code-switched language. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, 18–26.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.