GCDH@LT-EDI-EACL2021: XLM-RoBERTa for Hope Speech Detection in English, Malayalam, and Tamil

Stefan Ziehe
Institute of Computer Science
University of Göttingen
stefan.ziehe@cs.uni-goettingen.de

Franziska Pannach
Göttingen Centre for Digital Humanites
University of Göttingen
franziska.pannach@uni-goettingen.de

Aravind Krishnan
Department of Electronics and Communication Engineering
College of Engineering, Trivandrum
aravindhp1999@gmail.com

Abstract

This paper describes approaches to identify Hope Speech in short, informal texts in English, Malayalam and Tamil using different machine learning techniques. We demonstrate that even very simple baseline algorithms perform reasonably well on this task if provided with enough training data. However, our best performing algorithm is a cross-lingual transfer learning approach in which we fine-tune XLM-RoBERTa.

1 Introduction

In recent years, the spread of negative comments and hatred through social media has created a focus on the detection of misinformation and hate speech in the NLP community.

On the other hand, Hope Speech detection is a relatively new task. It can be employed to identify the positive aspects of large collection of social media posts, e.g. detecting pro-peace voices during politically heated situations. These detection efforts can potentially counter-act the perceived majority of hatred, and be a means to prevent harsh measures such as the disabling of internet access (Palakodety et al., 2020).

The ability to identify and therefore foster these aspects of online communication is a step towards a more positive representation of the internet. The hypothesis is: if hate-speech can incite violence, Hope Speech can ease tensions. For the first time to our knowledge, a shared task has been announced to identify Hope Speech in a multi-lingual dataset of Youtube comments. (Chakravarthi and Muralidaran, 2021)

2 Data

The dataset consists of three different subsets. It contains 10,705 comments for Malayalam, of which 8,564 are assigned for training, 1,070 for validation, and 1,071 for testing. For Tamil, the organizers provided 20,198 comments, of which 16,160 were designated for training, 2,018 for validation, and a test set of 2,020 comments. The English data consists of 28,451 comments in total, 22,762 for training, 2,843 for development, and 2,846 for testing.

The organizers provided annotations for the classes *Hope Speech*, *non Hope Speech*, and *not-"target language"*, where "target language" is either Malayalam, Tamil, or English. Figure 1 illustrates how the different classes are represented in the data sets. Notably, we can observe that the Tamil data is more balanced, whereas the Malayalam and English data sets show an over-representation of non Hope Speech data. Furthermore, only 22 comments in the English training set are labelled "not-English", but 12 % of the comments in the Tamil training set are "not-Tamil", and 8 % of the Malayalam comments are "not-Malayalam".

Annotations for the Hope Speech class include utterances that convey a generally bright prospect to the future, are supportive, insightful, and promote values such as inclusiveness and equality, among others. Some of these concepts for the English training set are illustrated in figure 2. For a full description of the annotation process, see (Chakravarthi, 2020).

3 Baseline

Using scikit-learn (Pedregosa et al., 2011), we implemented two baseline algorithms for the task of

Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion, pages 132–135
April 19, 2021.

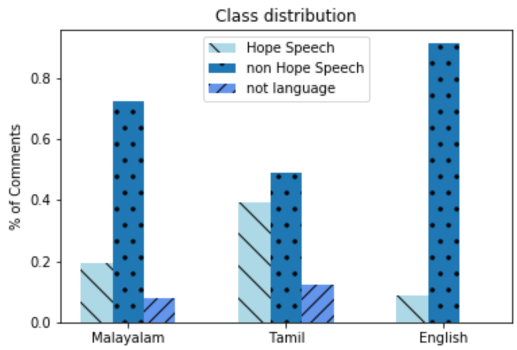

Figure 1: Comparison of Hope Speech to non Hope Speech data in the training sets

Figure 2: A selection of salient words that correspond with the Hope Speech label in the English training set.

Hope Speech detection, a Naive Bayes algorithm, and a Support Vector Machine. From the short texts, we first extracted a number of features including unigrams, bigrams, length and number of stopwords. Additionally, we assigned emotion scores for all three languages based on the NRC emotion lexicons (Mohammad et al., 2013). However, we found that solely relying on the textual data using a simple vectorizer based on tf-idf yielded better results.

As a first baseline, we implemented a Complement Naive Bayes model (Rennie et al., 2003). We iteratively found the best smoothing parameters $\alpha = 1.3$ for Malayalam, $\alpha = 10.5$ for Tamil and $\alpha = 1.4$ for English.

For the second baseline, we used an SVM with a degree of three and linear kernel. We found the best regularization parameters $c = 1.3$ for Malayalam, $c = 1.1$ for Tamil and $c = 1.4$ for English.

4 Transformer Model

Our main model is based on XLM-RoBERTa (Conneau et al., 2019), a pre-trained neural language model that uses the transformer architecture (more specifically, we use the `xlm-roberta-base` version from the huggingface transformers library (Wolf et al., 2020)). It is particularly useful for this task, because it is already pre-trained on 100 different languages, including Malayalam, Tamil and English.

We preprocess the provided datasets by tokenizing the comments using the pre-trained SentencePiece (Kudo and Richardson, 2018) tokenizer that belongs to XLM-RoBERTa. All comments are padded or truncated to a length of 128 tokens. XLM-RoBERTa is able to implicitly infer the language used in a text based on the tokens that it contains, which enables it to deal with code-switching. This also means that we can use batches that contain multiple languages at the same time during training.

The model is fine-tuned for hope speech detection on the combined training datasets for all three languages. In order to account for the different sizes of the datasets, Malayalam and Tamil samples are assigned sample weights of 2 and 3, respectively. We use a batch size of 64 and the Adam (Kingma and Ba, 2015) optimizer with a learning rate of $2 \cdot 10^{-5}$ for 4 epochs. We further employ decoupled weight decay (Loshchilov and Hutter, 2019) with a decay rate of 0.01 and shrinking learning rates for early transformer layers (Sun et al., 2019) with a decay rate of 0.95. Since the datasets only contain a small portion of all possible tokens, the token embeddings are deliberately not fine-tuned. Training takes place on an Nvidia GeForce RTX 2080 Ti in mixed precision mode and takes roughly 20 minutes.

5 Results

5.1 Baseline Results

Language	Precision	Recall	F1-Score
	SVM		
Malayalam	0.80	0.82	0.80
Tamil	0.61	0.61	0.60
English	0.80	0.82	0.80
	CNB		
Malayalam	0.78	0.76	0.77
Tamil	0.61	0.61	0.61
English	0.91	0.91	0.91

Table 1: Weighted Average Scores for SVM and Complement Naive Bayes on the development sets

We observe that even a simple algorithm such as Complement NB results in a high F1-Score of

0.91 for English. However, for Malayalam the SVM performs better with an F1-Score of 0.80. For Tamil, both baseline systems perform similarly with 0.60 (SVM) resp. 0.61 (CNB) F1-Score.

5.2 Results of the XLM-RoBERTa model

Class	Precision	Recall	F1-Score
	Malayalam		
HS	0.65	0.51	0.57
not HS	0.86	0.91	0.89
not ML	0.74	0.67	0.70
weighted avg	0.81	0.82	0.81
	Tamil		
HS	0.67	0.40	0.50
not HS	0.64	0.79	0.70
not Tamil	0.58	0.73	0.65
weighted avg	0.64	0.64	0.62
	English		
HS	0.63	0.57	0.60
not HS	0.95	0.96	0.96
not EN	0.00	0.00	0.00
weighted avg	0.92	0.92	0.92

Table 2: Results for the XLM-RoBERTa model by language and class (HS = Hope Speech) on the development sets

As seen in table 2, the fine-tuned XLM-RoBERTa model is an improvement over the baseline when applied to the development set. It performs strongest on English data, followed by Malayalam and Tamil, with F1-Scores of 0.92, 0.81 and 0.62 respectively.

Language	Precision	Recall	F1-Score
Malayalam	0.84	0.85	0.85
Tamil	0.59	0.59	0.58
English	0.93	0.93	0.93

Table 3: Official results for the XLM-RoBERTa model on the test set as provided by the task organizers

The official results on the test set (see table 3) place the model at first rank for English and Malayalam and fourth rank for Tamil in the competition.

5.3 Analysis

The improvements the fine-tuned XLM-RoBERTa model offers over the baseline models are not surprising, since large transformer models have a high capacity and are able to take the order of tokens into account. Differences between the languages are likely related to structural properties of the datasets, which is supported by the fact that the other competitors received similar results. In a similar way, imbalances in the datasets as described in section 2 can explain why the classifiers perform better on the *non Hope Speech* class than the other classes.

6 Conclusion

In this paper we presented several machine learning models trained for Hope Speech Detection in three different languages (Malayalam, Tamil and English). As baseline models we used SVM and Complement Naive Bayes classifiers for each language. The best results were achieved by a cross-lingual transfer learning approach in which we fine-tuned XLM-RoBERTa. In the competition, this model achieved the first rank for Malayalam and English and the fourth rank for Tamil. In conclusion, we provide a cross-lingual model that is both effective and relatively fast to train.

References

Bharathi Raja Chakravarthi. 2020. HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53, Barcelona, Spain (Online). Association for Computational Linguistics.

Bharathi Raja Chakravarthi and Vigneshwaran Muralidaran. 2021. Findings of the shared task on Hope Speech Detection for Equality, Diversity, and Inclusion. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

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

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, Conference Track Proceedings*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 321–327, Atlanta, Georgia, USA. Association for Computational Linguistics.

Shriphani Palakodety, Ashiqur R. KhudaBukhsh, and Jaime G. Carbonell. 2020. Hope speech detection: A computational analysis of the voice of peace. In *Proceedings of the 24th European Conference on Artificial Intelligence - ECAI 2020*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 1881–1889. IOS Press.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Jason D Rennie, Lawrence Shih, Jaime Teevan, and David R Karger. 2003. Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 616–623.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification? In *Chinese Computational Linguistics*, pages 194–206, Cham. Springer International Publishing.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.