TeamUNCC@LT-EDI-EACL2021:
Hope Speech Detection using Transfer Learning with Transformers

Khyati Mahajan, Erfan Al-Hossami, and Samira Shaikh

Department of Computer Science
University of North Carolina at Charlotte
Charlotte, NC 28223
{kmahaja2, ealhossa, sshaikh2}@uncc.edu

Abstract

In this paper, we describe our approach towards the task of hope speech detection. We participated in Task 2: Hope Speech Detection for Equality, Diversity and Inclusion at LT-EDI-2021 @ EACL2021. The goal of this task is to predict the presence of hope speech, along with the presence of samples that do not belong to the same language in the dataset. We describe our approach to fine-tuning RoBERTa for Hope Speech detection in English and our approach to fine-tuning XLM-RoBERTa for Hope Speech detection in Tamil and Malayalam, two low resource Indic languages. We demonstrate the performance of our approach on classifying text into `hope-speech`, `non-hope` and `not-language`. Our approach ranked 1st in English ($F1 = 0.93$), 1st in Tamil ($F1 = 0.61$) and 3rd in Malayalam ($F1 = 0.83$). We make our code available on Github.[1]

1 Introduction

Hate speech and the need for its moderation on social media platforms has recently become the focus of research (Zhang and Luo, 2019; Das et al., 2020). The need for detection of hate speech is clear, and worthy of the efforts. However, detecting and removing hate speech, while providing important benefits to conversations online, should be supplemented by other efforts to improve human connection and communication. There is a clear need to find common themes in community and build bridges to reduce the recent wave of polarization which has gripped many social media platforms (Prasetya and Murata, 2020; Levy, 2020).

There has been extensive work including proposed tasks and datasets on hate speech detection online. There have been multiple *SemEval*[2] tasks to detect hate speech, including binary (2 classes, (Basile et al., 2019)), and multi-label (>2 classes with overlaps (Mollas et al., 2020)) annotations. Hate speech detection has also further been expanded to cover explainability for the approaches used by Mathew et al. (2020), who propose *HateXplain* in which the span of text constituting the hate speech must also be detected. However, hate speech is just one facet of human behavior, especially on social media (Chakravarthi et al., 2020; Mandl et al., 2020; Chakravarthi et al., 2021; Suryawanshi and Chakravarthi, 2021). There are equally interesting studies on prosocial behaviors online including solidarity (Herrera-Viedma et al., 2015; Santhanam et al., 2021 (in press) and altruism (Althoff et al., 2014), as well as hope speech.

The Hope Speech Detection for Equality, Diversity and Inclusion task at LT-EDI-2021 @ EACL2021 (Chakravarthi and Muralidaran, 2021) marks a step towards helping push for more positive, uplifting speech on social media. Exacerbated by the pandemic, and the subsequent need of communication to move to online communities, research in the detection of "hope speech", which promotes positive, uplifting discussion to build support, is an important step (Puranik et al., 2021; Ghanghor et al., 2021). Findings from this effort could have an overall positive effect in the real world.

The task mainly focuses on multilingual classification for identifying speech associated with promise, potential, support, reassurance, suggestions or inspiration provided to participants by their peers during periods of illness, stress, loneliness and depression (Snyder et al., 2005). In addition to being a multi-class problem, each language (English, Tamil and Malayalam) also had differing amounts of class imbalance. We describe our approach in detail in Section 2 and present our results in Section 3.

[1] https://tinyurl.com/teamuncc-hope

[2] https://semeval.github.io/

Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion, pages 136–142
April 19, 2021.

2 Method

RoBERTa is an improved BERT (Devlin et al., 2019) model by Facebook AI which achieves state-of-the-art results on several natural language understanding (NLU) tasks including GLUE (Wang et al., 2018) and SQuAD (Rajpurkar et al., 2016). RoBERTa is improved through training BERT for a longer duration on longer sequences, increasing the data quantity, removing the sentence prediction objective during pre-training, and changing the masking pattern applied during pre-training. With the aim of improving cross-lingual language understanding (XLU), XLM-RoBERTa was developed using a Transformer-based masked language model (MLM). XLM-RoBERTa was pre-trained using 2 terabytes of CommonCrawl data (Wenzek et al., 2020) containing one hundred languages. XLM-RoBERTa outperforms its multilingual MLMs mBERT (Devlin et al., 2019) and XLM (Lample and Conneau, 2019).

We thus do not propose a novel system, and instead rely on fine-tuning these two large transformers for the task of hope speech detection. We use the RoBERTa transformer (Liu et al., 2019) for English. We select XLM-RoBERTa (Conneau et al., 2020) for Tamil and Malayalam tasks due to its robustness and pre-training on low-resources languages including Tamil and Malayalam. For our implementation, we use the Simple Transformers[3] library which is built upon the transformers library by huggingface (Wolf et al., 2020). We use Adam (Kingma and Ba, 2014) as our optimizer. Our hyperparameters are presented in Table 1.

Hyperparameter	Value
epochs	6
bacth_size	8
α (English)	0.00002
α (Malayalam & Tamil)	0.00001
max_length	256
decay (L2)	0

Table 1: Hyperparameter values

3 Results

We present detailed results for each language below. We also compare with baselines provided in the original dataset paper (Chakravarthi, 2020), and include them in the results in Tables 2, 3, and 4. We report results to 2 significant digits, since the baseline provided by the task is also limited to 2 significant digits. We also provide the dataset distribution in these tables, in the "Support" column. We discuss our findings further in Section 4.

Evaluation for English. Table 2 overviews the results of detecting `hope-speech`, `non-hope`, and `not-english`. The baseline for English was weighted average $F1 = 0.90$ (Chakravarthi, 2020). Our approach scored a weighted average $F1 = 0.93$ on both the dev (N = 2843) and test (N = 2846) sets, achieving the 1st place in the task among 31 team submissions. We note the class imbalance between `hope-speech` ($N = 272$) and `non-hope` (N = 2569) denoted in the support column. While our approach does achieve high numbers overall, there is clearly room for improvement. Incorrectly labeled `non-hope` utterances (N = 59) tend to be fewer than incorrectly labeled `hope-speech` utterances (N = 118), hence the precision (0.69) is greater than the recall (0.53) on the test set.

Evaluation for Tamil. Table 3 overviews the results of detecting `hope-speech`, `non-hope`, and `not-tamil`. The baseline performance for Tamil was weighted average $F1$ = 0.56 (Chakravarthi, 2020). Our approach scored a weighted average $F1$ = 0.61 on the dev set (N = 2018) and a weighted average $F1 = 0.60$ on the test set (N = 2020) giving us the 1st place in the task as well against 30 team submissions. Similar to English, we observe higher precision (0.59) over recall (0.49) suggesting a tendency for false positives to be lower than false negatives.

Evaluation for Malayalam. Table 4 overviews the results of detecting `hope-speech`, `non-hope`, and `not-malayalam`. The baseline weighted F1 performance for Malayalam was $F1$ = 0.73 (Chakravarthi, 2020). At the time of submission, our submission scored weighted average $F1$ = 0.82 on the dev set (N = 1070) and weighted average $F1 = 0.83$ on the test set (N = 1071), ranking us 3rd for the task among 31 teams. The approach for Malayalam was incomplete - being trained only on 2 epochs instead of 6. After complete training ($epochs$ = 6), our approach scores a weighted average $F1$ = 0.87 on the test set. We observe very little difference between precision and recall suggesting a balance between missed `hope-speech` and false alarm `hope-speech` utterances.

[3] https://simpletransformers.ai/

Class	Dev Set				Test Set				Baseline
	Support	Precision	Recall	F1	Support	Precision	Recall	F1	Test F1
Hope	272	0.69	0.55	**0.61**	250	0.69	0.53	**0.60**	0.42
Non-hope	2569	0.95	0.97	**0.96**	2593	0.95	0.98	**0.97**	0.95
Not-English	2	0.00	0.00	0.00	3	0.00	0.00	0.00	0.00
Accuracy	2843	-	-	**0.93**	2846	-	-	**0.94**	-
Macro	2843	0.55	0.51	**0.52**	2846	0.55	0.50	**0.52**	0.46
Weighted Avg	2843	0.93	0.93	**0.93**	2846	0.93	0.94	**0.93**	0.90

Table 2: Classification Results - English

Class	Dev Set				Test Set				Baseline
	Support	Precision	Recall	F1	Support	Precision	Recall	F1	Test F1
Hope	757	0.58	0.45	**0.51**	815	0.59	0.49	**0.54**	0.46
Non-hope	998	0.65	0.70	**0.67**	946	0.64	0.68	**0.66**	0.65
Not-Tamil	263	0.59	0.8	**0.68**	259	0.55	0.72	**0.63**	0.55
Accuracy	2018	-	-	**0.62**	2020	-	-	**0.61**	-
Macro	2018	0.61	0.65	**0.62**	2020	0.59	0.63	**0.61**	0.55
Weighted Avg	2018	0.62	0.62	**0.61**	2020	0.61	0.61	**0.60**	0.56

Table 3: Classification Results - Tamil

Class	Dev Set				Test Set				Baseline
	Support	Precision	Recall	F1	Support	Precision	Recall	F1	Test F1
Hope	190	0.69	0.61	**0.64**	194	0.70	0.72	**0.71**	0.36
Non-hope	784	0.89	0.91	**0.90**	776	0.91	0.91	**0.91**	0.86
Not-Malayalam	96	0.75	0.80	**0.77**	101	0.83	0.80	**0.81**	0.45
Accuracy	1070	-	-	**0.84**	1071	-	-	**0.87**	-
Macro	1070	0.78	0.77	**0.77**	1071	0.81	0.81	**0.81**	0.56
Weighted Avg	1070	0.84	0.84	**0.84**	1071	0.87	0.87	**0.87**	0.73

Table 4: Classification Results - Malayalam

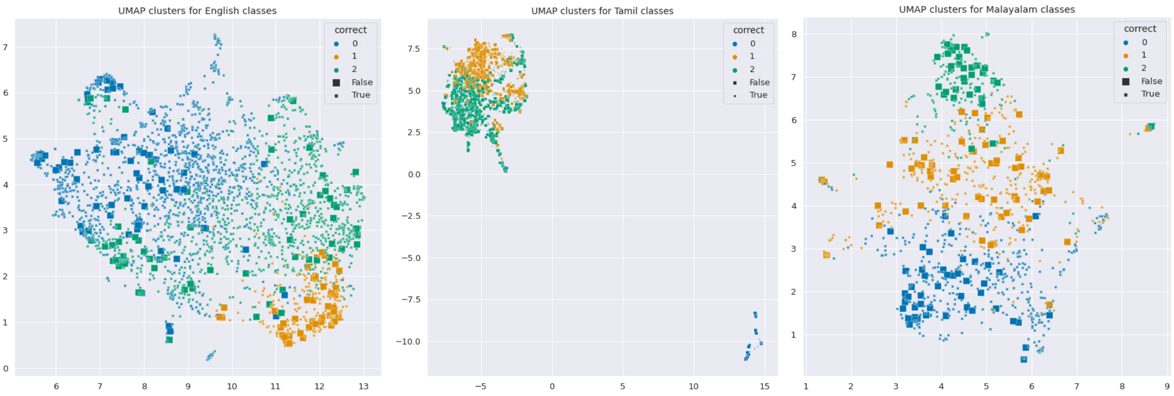
3.1 Negative Results

Since there is class imbalance between the 3 classes (non-hope >> hope-speech >> not-language), we attempted to modify the class weights during our fine-tuning process weighing the minority classes more than the majority class. The class weights for hope-speech and non-hope for the English, Tamil, and Malayalam languages are computed using Equation 1 inspired by (King et al., 2001) and implemented by the scikit-learn library (Pedregosa et al., 2011). N represents the number of samples in the dataset and x is a class label.

$$weight_x = \frac{N}{(count(classes) \times count(x))} \quad (1)$$

To address the larger imbalance between the not-language and other classes, we manually assign weights for these classes, so that they weigh less than the hope-speech class. The non-english class was weighed as 1 and the non-tamil and non-malayalam classes were weighed as 0.5. In future work, these weights will be empirically chosen. However, we did not submit these as our final models since the performance on weighted average F1 was lower.

Results. The weighted average F1 using the class-weights approach described above were a bit lower than the results described in Tables 2, 3, and 4 across all languages. On the English dev set, we scored weighted average $F1$ = 0.92, with $F1$ = 0.60 for hope-speech, $F1$ = 0.95 for non-hope, and $F1$ = 0.00 for non-english. We note that all of the results are a bit lower than the



(a) English

(b) Tamil

(c) Malayalam

Figure 1: UMAP scatter plots visualizing clusters (color code) and prediction correctness (mark shape) using sentence transformers on the test set for an analytical evaluation of our approach.

our submission results. On the Tamil dev set, we scored weighted average $F1 = 0.61$, $F1 = 0.57$ for `hope-speech`, $F1 = 0.65$ for `non-hope`, and $F1 = 0.60$ for `non-tamil`. We note that all of the results are lower than the our original submission results with the exception of the `hope-speech` class. On the Malayalam dev set, we scored weighted average $F1 = 0.84$, $F1 = 0.64$ for `hope-speech`, $F1 = 0.89$ for `non-hope`, and $F1 = 0.75$ for `non-malayalam`. We note that all of the individual class results are slightly lower than the our submission results.

4 Qualitative Evaluation

In this section, we analytically go through sample records in the English test set to better understand the dataset, and to evaluate the strengths and weaknesses of our approach.

Analytical Evaluation. We use sentence-transformers (Reimers and Gurevych, 2019, 2020) to further evaluate our approach. Sentence transformers were created by utilizing word-level representation models such as BERT and RoBERTa for better downstream computational performance on sentence-level tasks, since utilizing word-level representations for tasks such as determining the similarity of 2 sentences takes much longer than computing sentence-level representations. We use the `paraphrase-xlm-r-multilingual-v1` pretrained model to get our sentence embeddings, and then use K-means clustering using scikit-learn (Pedregosa et al., 2011) to cluster the test set into 3 clusters, the same number of classes as the classification task. We provide clustering results, showing cluster assignments of data samples and whether our classifiers labelled them correctly or not in Figure 1.

For English (Figure 1a), there are fewer misclassified samples, but most of them lie in cluster 1, shown on the bottom right of the UMAP scatter plot 1. This cluster, upon further observation, mostly consists of Youtube comments related to the Black Lives Matter (BLM) movement. We conclude that the our approach specifically struggles with correctly classifying such utterances, and elaborate with illustrative examples in the next section.

For Tamil (Figure 1b), we observe closely clustered data points, a property that is also reflected in our classification results as seen by the misclassified points (shown by square markers in the scatter plot). We see a similar trend as English in Malayalam (Figure 1c), however an observational analysis is not possible for Malayalam and Tamil since we do not speak the language, and wish not to rely on automatic translation systems for informal, code-mixed text (Chakravarthi, 2020).

Observations. In Table 5, we present a qualitative evaluation using select examples from the English test set. First, we examine misclassifications in the English test set and report on our observations. The model predicts 3 samples as `non-hope` when the ground-truth label is `not-english`. However upon examination, 2 of these 3 samples predominantly contain English words that seem to be `non-hope` sentences. We demonstrate one of the `non-english` sentences in the test set in the sixth row of Table 5). The third `not-english` sample contains a latin-alphabet utterance of a `non-english` sentence with some recognizable English words. Furthermore, upon

#	Sentence	Actual	Predicted	Cluster
1	Speak for yourself	**non-hope**	**non-hope**	**0**
2	What do you mean by the word sniped?	**non-hope**	**non-hope**	**0**
3	Everyone matters stop racism	**hope-speech**	**hope-speech**	**2**
4	Realize Black lives matter is designed to cause division. All lives matter is to state unity. We are in this together. The tactic of divide and conquer is ancient... But still works!	hope-speech	non-hope	1
5	It's one thing to ideally believe that all lives matter	non-hope	hope-speech	1
6	We have and they know we are Israelites the real jews and got jealous...	not-english	non-hope	2
7	Trying to end racism by supporting everyone equally. Blm mob	non-hope	hope-speech	2

Table 5: Illustrative examples of successful (bold) and failure cases of the TeamUNCC approach for English.

examining misclassifications of non-hope and hope-speech, we observe a trend that BLM related samples tend to compose quite a bit of our misclassified samples. Over 46 out of 177 samples seem to be BLM related. We observe that it can be ambiguous and tricky to determine whether it expresses hope or not. The ambiguity observation for some of the samples is inline with the findings of (Chakravarthi, 2020). In the fifth row, we note that the utterance expresses that all lives matter but the ground truth is non-hope. The fourth and seventh rows are similar, both utterances express support for equality but it also express negative sentiments towards the BLM movement; yet the ground truth labels for both these utterances are different. The first 3 rows showcase examples of where the classifications are correct. We observe that the model might have learned to look for linguistic markers such as questions and inquiries, and thus tended to label them as non-hope.

5 Conclusion

We found that fine-tuning RoBERTa (English) and XLM-RoBERTa (Tamil and Malayalam) for classification performed well. Providing class weights helped handle the data imbalance, leading to a balanced performance over the baselines. We submitted the results of our approach using the RoBERTa and XLM-RoBERTa models without class weights as our final submission for the Hope Speech Detection Task. Admittedly, our approach does not break new ground in terms of novelty of models or architecture; we instead rely upon fine-tuning pre-trained models. However, we present it as a first step in this direction, and aim to build on our success in future iterations. The Hope Speech Detection task was an important and necessary step towards promoting positive speech online. Our goal is to keep participating in the effort to detect and promote more positive speech online. Additionally, we hope to extend this task to COVID19 data, and work towards understanding community support during the pandemic.

References

Tim Althoff, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2014. How to ask for a favor: A case study on the success of altruistic requests. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8.

Valerio Basile, C. Bosco, E. Fersini, Debora Nozza, V. Patti, F. Pardo, P. Rosso, and M. Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *SemEval@NAACL-HLT*.

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, M Anand Kumar, John Philip McCrae, Premjith B, Soman KP, and Thomas Mandl. 2020. Overview of the track on HASOC-Offensive Language Identification-DravidianCodeMix. In *Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2020). CEUR Workshop Proceedings. In: CEUR-WS. org, Hyderabad, India.*

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.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Anand Kumar M, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan V, Elizabeth Sherly, and John Philip McCrae. 2021. Findings of the shared task on Offensive Language Identification in Tamil, Malayalam, and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Mithun Das, Binny Mathew, Punyajoy Saha, P. Goyal, and Animesh Mukherjee. 2020. Hate speech in online social media. *ACM SIGWEB Newsletter*, pages 1 – 8.

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.

Nikhil Kumar Ghanghor, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. IIITK@LT-EDI-EACL2021: Hope Speech Detection for Equality, Diversity, and Inclusion in Tamil, Malayalam and English. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, Online.

Enrique Herrera-Viedma, Juan Bernabe-Moreno, Carlos Porcel Gallego, and Maria de los Angeles Martinez Sanchez. 2015. Solidarity in social media: when users abandon their comfort zone-the charlie hebdo case. *ICONO 14, Revista de comunicación y tecnologías emergentes*, 13(2):6–22.

Gary King, Langche Zeng, et al. 2001. Logistic regression in rare events data. *Political Analysis*, 9(2):137–163.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *NeurIPS*.

Roee Levy. 2020. Social media, news consumption, and polarization: evidence from a field experiment. *News Consumption, and Polarization: Evidence from a Field Experiment (July 16, 2020)*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2020. Overview of the HASOC Track at FIRE 2020: Hate Speech and Offensive Language Identification in Tamil, Malayalam, Hindi, English and German. In *Forum for Information Retrieval Evaluation*, FIRE 2020, page 29–32, New York, NY, USA. Association for Computing Machinery.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. Hatexplain: A benchmark dataset for explainable hate speech detection. *arXiv preprint arXiv:2012.10289*.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2020. Ethos: an online hate speech detection dataset. *arXiv preprint arXiv:2006.08328*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

H. Prasetya and T. Murata. 2020. A model of opinion and propagation structure polarization in social media. *Computational Social Networks*, 7:1–35.

Karthik Puranik, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. IIITT@LT-EDI-EACL2021-Hope Speech Detection: There is always hope in Transformers. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Sashank Santhanam, Vidhushini Srinivasan, Khyati Mahajan, and Samira Shaikh. 2021 (in press). Towards understanding how emojis express solidarity in crisis events. In *International Conference on Applied Human Factors and Ergonomics*. Springer.

CR Snyder, Kevin L Rand, and David R Sigmon. 2005. Hope theory: A member of the positive psychology.

Shardul Suryawanshi and Bharathi Raja Chakravarthi. 2021. Findings of the shared task on Troll Meme Classification in Tamil. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Édouard 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.

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.

Z. Zhang and Lei Luo. 2019. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, 10:925–945.