IIITT@LT-EDI-EACL2021-Hope Speech Detection: There is always Hope in Transformers

Karthik Puranik[1], **Adeep Hande**[1], **Ruba Priyadarshini**[2],
Sajeetha Thavareesan[3], **Bharathi Raja Chakravarthi**[4]
[1]Indian Institute of Information Technology Tiruchirappalli
[2]ULTRA Arts and Science College, India,[3]Eastern University, Sri Lanka
[4]National University of Ireland Galway
karthikp18c@iiitt.ac.in

Abstract

In a world filled with serious challenges like climate change, religious and political conflicts, global pandemics, terrorism, and racial discrimination, an internet full of hate speech, abusive and offensive content is the last thing we desire for. In this paper, we work to identify and promote positive and supportive content on these platforms. We work with several transformer-based models to classify social media comments as hope speech or not-hope speech in English, Malayalam and Tamil languages. This paper portrays our work for the Shared Task on Hope Speech Detection for Equality, Diversity, and Inclusion at LT-EDI 2021- EACL 2021. The codes for our best submission can be viewed[1].

1 Introduction

Social Media has inherently changed the way people interact and carry on with their everyday lives as people using the internet (Jose et al., 2020; Priyadharshini et al., 2020). Due to the vast amount of data being available on social media applications such as YouTube, Facebook, an Twitter it has resulted in people stating their opinions in the form of comments that could imply hate or negative sentiment towards an individual or a community (Chakravarthi et al., 2020c; Mandl et al., 2020). This results in people feeling hostile about certain posts and thus feeling very hurt (Bhardwaj et al., 2020).

Being a free platform, social media runs on user-generated content. With people from multifarious backgrounds present, it creates a rich social structure (Kapoor et al., 2018) and has become an exceptional source of information. It has laid it's roots so deeply into the lives of people that they count on it for their every need. Regardless, this tends to mislead people in search of credible information. Certain individuals or ethnic groups also fall prey to people utilizing these platforms to foster destructive or harmful behaviour which is a common scenario in cyberbullying (Abaido, 2020).

The earliest inscription in India dated from 580 BCE was the Tamil inscription in pottery and then, the Asoka inscription in Prakrit, Greek and Aramaic dating from 260 BCE. Thunchaththu Ramanujan Ezhuthachan split Malayalam from Tamil after the 15th century CE by using Pallava Grantha script to write religious texts. Pallava Grantha was used in South India to write Sanskrit and foreign words in Tamil literature. Modern Tamil and Malayalam have their own script. However, people use the Latin script to write on social media (Chakravarthi et al., 2018, 2019; Chakravarthi, 2020b).

The automatic detection of hateful, offensive, and unwanted language related to events and subjects on gender, religion, race or ethnicity in social media posts is very much necessary (Rizwan et al., 2020; Ghanghor et al., 2021a,b). Such harmful content could spread, stimulate, and vindicate hatred, outrage, and prejudice against the targeted users. Removing such comments was never an option as it suppresses the freedom of speech of the user and it is highly unlikely to stop the person from posting more. In fact, he/she/they would be prompted to post more of such comments[2] (Yasaswini et al., 2021; Hegde et al., 2021). This brings us to our goal to spread positivism and hope and identify such posts to strengthen an open-minded, tolerant, and unprejudiced society.

[1]https://github.com/karthikpuranik11/Hope-Speech-Detection-

[2]https://www.qs.com/negative-comments-on-social-media/

Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion, pages 98–106
April 19, 2021.

Text	Language	Label
God gave us a choice my choice is to love, I would die for that kid	English	Hope
The Democrats are.Backed powerful rich people like Soros	English	Not hope
ESTE PSICÃ“PATA MASÃ“N LUCIFERIANO ES HOMBRE TRANS	English	Not English
Neega podara vedio nalla iruku ana subtitle vainthuchu ahh yella language papaga	Tamil	Hope
Avan matum enkita maatunan... Avana kolla paniduven	Tamil	Not hope
I can't uninstall mY Pubg	Tamil	Not Tamil
ooororutharum avarude ishtam pole jeevikatte . k.	Malayalam	Hope
Etraem aduthu nilkallae Arunae	Malayalam	Not hope
Phoenix contact me give you're mail I'd I hope I can support you sure!	Malayalam	Not Malayalam

Table 1: Examples of hope speech or not hope speech

2 Related Works

The need for the segregation of toxic comments from social media platforms has been identified back in the day. Founta et al. (2018) has tried to study the textual properties and behaviour of abusive postings on Twitter using a Unified Deep Learning Architecture. Hate speech can be classified into various categories like hatred against an individual or group belonging to a race, religion, skin colour, ethnicity, gender, disability, or nation[3] and there have been studies to observe it's evolution in social media over the past thirty years (Tontodimamma et al., 2021). Deep Learning methods were used to classify hate speech into racist, sexist or neither in Badjatiya et al. (2017).

Hope is support, reassurance or any kind of positive reinforcement at the time of crisis (Chakravarthi, 2020a). Palakodety et al. (2020) identifies the need for the automatic detection of content that can eliminate hostility and bring about a sense of hope during times of wrangling and brink of a war between nations. There have also been works to identify hate speech in multilingual (Aluru et al., 2020) and code-mixed data in Tamil, Malayalam, and Kannada language (Chakravarthi et al., 2020b,a; Hande et al., 2020). However, there have been very fewer works in Hope speech detection for Indian languages.

3 Dataset

The dataset is provided by (Chakravarthi, 2020a) (Chakravarthi and Muralidaran, 2021) and contains 59,354 comments from the famous online video sharing platform YouTube out of which 28,451 are in English, 20,198 in Tamil, and 10,705 comments are in Malayalam (Table 2) which can be classified as Hope speech, not hope speech and other languages. This dataset is split into train (80%), development (10%) and test (10%) dataset (Table 3).

Subjects like hope speech might raise confusions and disagreements between annotators belonging to different groups. The dataset was annotated by a minimum of three annotators and the inter-annotator agreement was determined using Krippendorff's alpha (krippendorff, 2011). Refer table 1 for examples of hope speech, not hope speech and other languages for English, Tamil and Malayalam datasets respectively.

Class	English	Tamil	Malayalam
Hope	2,484	7,899	2,052
Not Hope	25,940	9,816	7,765
Other lang	27	2,483	888
Total	28,451	20,198	10,705

Table 2: Classwise Data Distribution

Split	English	Tamil	Malayalam
Training	22,762	16,160	8564
Development	2,843	2,018	1070
Test	2,846	2,020	1071
Total	28,451	20,198	10,705

Table 3: Train-Development-Test Data Distribution

4 Experiment Setup

In this section, we give a detailed explanation of the experimental conditions upon which the models are developed.

[3] http://www.ala.org/advocacy/intfreedom/hate (Accessed January 16, 2021)

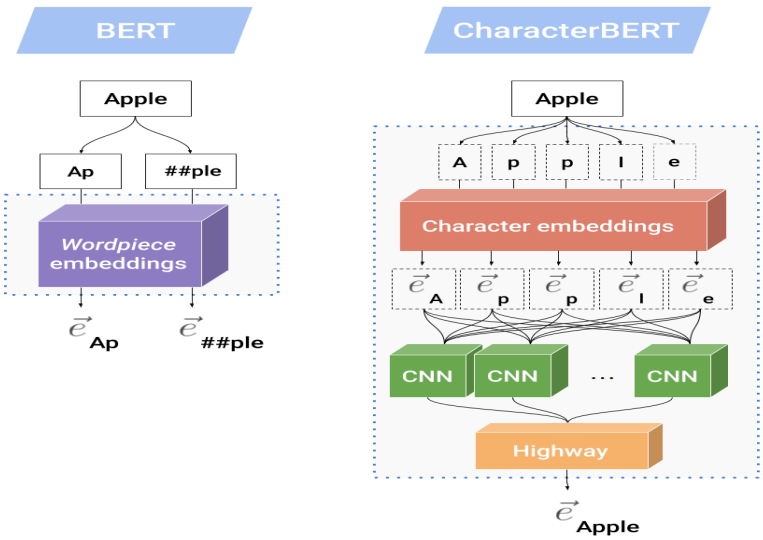

Figure 1: Context-independent representations in BERT and CharacterBERT (Source: El Boukkouri et al. (2020))

4.1 Architecture

4.1.1 Dense

The dense layers used in CNN (convolutional neural networks) connects all layers in the next layer with each other in a feed-forward fashion (Huang et al., 2018). Though they have the same formulae as the linear layers i.e. wx+b, the output is passed through an activation function which is a non-linear function. We implemented our models with 2 dense layers, rectified linear units (ReLU) (Agarap, 2019) as the activation function and dropout of 0.4.

4.1.2 Bidirectional LSTM

Bidirectional LSTM or biLSTM is a sequence processing model (Schuster and Paliwal, 1997). It uses both the future and past input features at a time as it contains two LSTM's, one taking input in the forward direction and another in the backward direction (Schuster and Paliwal, 1997). The backward and forward pass through the unfolded network just like any regular network. However, BiLSTM requires us to unfold the hidden states for every time step. It produces a drastic increase in the size of information being fed thus, improving the context available (Huang et al., 2015). Refer Table 4 for the parameters used in the BiLSTM model.

Parameter	Value
Number of LSTM units	256
Dropout	0.4
Activation Function	ReLU
Max Len	128
Batch Size	32
Optimizer	AdamW
Learning Rate	2e-5
Loss Function	cross-entropy
Number of epochs	5

Table 4: Parameters for the BiLSTM model

4.2 Embeddings

4.2.1 BERT

Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). The multilingual base model is pretrained on the top 104 languages of the world on Wikipedia (2.5B words) with 110 thousand shared wordpiece vocabulary. The input is encoded into vectors with BERT's innovation of bidirectionally training the language model which catches a deeper context and flow of the language. Furthermore, novel tasks like Next Sentence Prediction (NSP) and Masked Language Modelling (MLM) are used to train the model.

The pretrained BERT Multilingual model ***bert-base-multilingual-uncased*** (Pires et al., 2019)

from Huggingface[4] (Wolf et al., 2020) is executed in PyTorch (Paszke et al., 2019). It consists of 12-layers, 768 hidden, 12 attention heads and 110M parameters which are fine-tuned by concatenating with bidirectional LSTM layers. The BiLSTM layers take the embeddings from the transformer encoder as the input which increases the information being fed, which in turn betters the context and accuracy. Adam algorithm with weight decay fix is used as an optimizer. We train our models with the default learning rate of $2e-5$. We use the cross-entropy loss as it is a multilabel classification task.

4.2.2 ALBERT

It has a similar architecture as that of BERT but due to memory limitations and longer training periods, ALBERT or A Lite BERT introduces two parameter reduction techniques (Chiang et al., 2020). ALBERT distinguishes itself from BERT with features like factorization of the embedding matrix, cross-layer parameter sharing and inter-sentence coherence prediction. We implemented ***albert-base-v2*** pretrained model with 12 repeating layers, 768 hidden, 12 attention heads, and 12M parameters for the English dataset.

4.2.3 DistilBERT

DistilBERT is a distilled version of BERT to make it smaller, cheaper, faster, and lighter (Sanh et al., 2019). With up to 40% less number of parameters than ***bert-base-uncased***, it promises to run 60% faster while preserving 97% of it's performance. We employ ***distilbert-base-uncased*** for the English dataset and ***distilbert-base-multilingual-cased*** for the Tamil and Malayalam datasets. Both models have 6-layers, 768-hidden, 12-heads and while the former has 66M parameters, the latter has 134M parameters.

4.2.4 RoBERTa

A Robustly optimized BERT Pretraining Approach (RoBERTa) is a modification of BERT (Liu et al., 2020). RoBERTa is trained for longer, with larger batches on 1000% more data than BERT. The Next Sentence Prediction (NSP) task employed in BERT's pre-training is removed and dynamic masking during training is introduced. It's additionally trained on a 76 GB large new dataset (CC-NEWS). ***roberta-base*** follows the BERT architecture but has 125M parameters and is used for the English dataset.

4.2.5 CharacterBERT

CharacterBERT (CharBERT) (El Boukkouri et al., 2020) is a variant of BERT (Devlin et al., 2019) which uses CharacterCNN (Zhang et al., 2015) like ELMo (Peters et al., 2018), instead of relying on WordPieces (Wu et al., 2016). CharacterBERT is highly desired as it produces a single embedding for any input token which is more suitable than having an inconstant number of WordPiece vectors for each token. It furthermore replaces BERT from domain-specific wordpiece vocabulary and enables it to be more robust to noisy inputs.

We use the pretrained model ***general-character-bert***[5] which was pretrained on the same corpus of that of BERT, but with a different tokenization approach. A CharacterCNN module is used that produces word-level contextual representations and it can be re-adapted to any domain without needing to worry about the suitability of any wordpieces (Figure 1). This approach helps for superior robustness by approaching the character of the inputs.

4.2.6 ULMFiT

Universal Language Model Fine-tuning, or ULMFiT, was a transfer learning method introduced to perform various NLP tasks (Howard and Ruder, 2018). Training of ULMFiT involves pretraining the general language model on a Wikipedia-based corpus, fine-tuning the language model on a target text, and finally, fine-tuning the classifier on the target task. Discriminative fine-tuning is applied to fine-tune the model as different layers capture the different extent of information. It is then trained using the learning rate scheduling strategy, Slanted triangular learning rates (STLR), where the learning rate increases initially and then drops. Gradual unfreezing is used to fine-tune the target classifier rather than training all layers at once, which might lead to catastrophic forgetting.

Pretrained model, **AWD-LSTM** (Merity et al., 2017) with 3 layers and 1150 hidden activation per layer and an embedding size of 400 is used as the language model for the English dataset. Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$ is implemented. Later, the start and end learning rates are

[4] https://huggingface.co/transformers/pretrained_models.html

[5] https://github.com/helboukkouri/character-bert

Architecture	Embeddings	F1-Score validation	F1-Score test
BiLSTM	bert-base-uncased	0.9112	0.9241
Dense	bert-base-uncased	0.9164	0.9240
	albert-base	0.9143	0.9210
	distilbert-base-uncased	0.9238	0.9283
	roberta-base	0.9141	0.9235
	character-bert	**0.9264**	0.9220
	ULMFiT	0.9252	0.9356

Table 5: Weighted F1-scores of hope speech detection classifier models on English dataset

Architecture	Embeddings	F1-Score validation	F1-Score test
BiLSTM	mbert-uncased	**0.8436**	0.8545
	mbert-cased	0.8280	0.8482
	xlm-roberta-base	0.8271	0.8233
	MuRIL	0.8089	0.8212
Dense	mbert-uncased	0.8373	0.8433
	indic-bert	0.7719	0.8264
	xlm-roberta-base	0.7757	0.7001
	distilmbert-cased	0.8312	0.8395
	MuRIL	0.8023	0.8187

Table 6: Weighted F1-scores of hope speech detection classifier model on Malayalam dataset

Architecture	Embeddings	F1-Score Validation	F1-Score test
BiLSTM	mbert-uncased	0.6124	0.5601
	mbert-cased	**0.6183**	0.5297
	xlm-roberta-base	0.5472	0.5738
	MuRIL	0.5802	0.5463
Dense	mbert-uncased	0.5916	0.4473
	mbert-cased	0.5946	0.4527
	indic-bert	0.5609	0.5785
	xlm-roberta-base	0.5481	0.3936
	distilmbert-cased	0.6034	0.5926
	MuRIL	0.5504	0.5291

Table 7: Weighted F1-scores of hope speech detection classifier models on Tamil dataset

set to *1e-8 and 1e-2* respectively and fine-tuned by gradually unfreezing the layers to produce better results. Dropouts with a multiplier of 0.5 were applied.

4.2.7 XLM-RoBERTa

XLM-RoBERTa (Ruder et al., 2019) is a pre-trained multilingual language model to execute diverse NLP transfer tasks. It's trained on over 2TB of filtered CommonCrawl data in 100 different languages. It was an update to the XLM-100 model (Lample and Conneau, 2019) but with increased training data. As it shares the same training routine with the RoBERTa model, "RoBERTa" was included in the name. ***xlm-roberta-base*** with 12 layers, 768 hidden, 12 heads, and 270M parameters were used. It is fine-tuned for classifying code-mixed Tamil and Malayalam datasets.

4.2.8 MuRIL

MuRiL[6] was introduced by Google Research India to enhance Indian NLU (Natural Language Understanding). The model has a BERT based architecture trained on 17 Indian languages with

[6] https://tfhub.dev/google/MuRIL/1

Language	Hope-Speech	Not-hope speech	Other Language	Macro Avg	Weighted Avg
Precision					
English	0.9464	0.6346	0.0000	0.5270	0.9193
Malayalam	0.6540	0.9032	0.8941	0.8171	0.572
Tamil	0.4824	0.5819	0.5709	0.5451	0.5403
Recall					
English	0.9781	0.4108	0.0000	0.4630	0.9293
Malayalam	0.7113	0.9021	0.7525	0.7886	0.8534
Tamil	0.2687	0.7812	0.6525	0.5675	0.5579
F1-Score					
English	0.9620	0.4987	0.0000	0.4869	0.9220
Malayalam	0.6815	0.9026	0.8172	0.8004	0.8545
Tamil	0.3452	0.6670	0.6090	0.5404	0.5207

Table 8: Classification report for our system models based on the results of test set

Wikipedia, Common Crawl[7], PMINDIA[8] and Dakshina[9] datasets. MuRIL is trained on translation and transliteration segment pairs which give an advantage as the transliterated text is very common in social media. It is used for the Malayalam and Tamil datasets.

4.2.9 IndicBERT

IndicBERT (Kakwani et al., 2020) is an ALBERT model pretrained on 12 major Indian languages with a corpus of over 9 billion tokens. It performs as well as other multilingual models with considerably fewer parameters for various NLP tasks. It's trained by choosing a single model for all languages to learn the relationship between languages and understand code-mixed data. ***ai4bharat/indic-bert*** model was employed for the Tamil and Malayalam task.

5 Results

In this section, we have compared the F1-scores of our transformer-based models to successfully classify social media comments/posts into hope speech or not hope speech and detect the usage of other languages if any. We have tabulated the weighted average F1-scores of our various models for validation and test dataset for English, Malayalam and Tamil languages in tables 5, 6 and 7 respectively.

Table 5 demonstrates that the character-aware model CharacterBERT performed exceptionally well for the validation dataset. It beat ULMFiT (Howard and Ruder, 2018) by a mere difference of 0.0012, but other BERT-based models like BERT (Devlin et al., 2019) with dense and BiLSTM architecture, ALBERT (Chiang et al., 2020), DistilBERT (Sanh et al., 2019) and RoBERTa (Liu et al., 2020) by about a percent. This promising result shown by character-bert for the validation dataset made it our best model. Unfortunately, few models managed to perform better than it for the test dataset. The considerable class imbalance of about 2,484 hope to 25,940 not hope comments and the interference of comments in other languages have significantly affected the results.

Similar transformer-based model trained on multilingual data was used to classify Malayalam and Tamil datasets. Models like multilingual BERT, XLM-RoBERTa (Ruder et al., 2019), MuRIL, IndicBERT [10] and DistilBERT multilingual with both BiLSTM and Dense architectures. mBERT (Multilingual BERT) uncased with BiLSTM concatenated to it outperformed the other models for the Malayalam validation dataset and continued its dominance for the test data as well.

The data distribution for the Tamil dataset seemed a bit balanced with an approximate ratio of 4:5 between hope and not-hope. mBERT cased with BiLSTM architecture appeared to be the best model with an F1-score of 0.6183 for validation but dropped drastically by 8% for the test data. We witnessed a considerable fall in the scores of other models like mBERT and XLM-RoBERTa with linear layers of up to 15%.

Multilingual comments experience an enor-

[7] http://commoncrawl.org/the-data/

[8] http://lotus.kuee.kyoto-u.ac.jp/WAT/indic-multilingual/index.html

[9] https://github.com/google-research-datasets/dakshina

[10] https://indicnlp.ai4bharat.org/indic-bert/

mous variety of text as people tend to write in code-mixed data and other non-native scripts which are inclined to be mispredicted. A variation in the concentration of such comments between train, validation and test can result in a fluctuation in the test results. The precision, recall and F1-scores of CharacterBERT, mBERT-uncased, and mBERT-cased are tabulated under English, Malayalam, and Tamil respectively, as shown in Table 8. They were the best performing models on the validation set.

6 Conclusion

During these unprecedented times, there is a need to detect positive, enjoyable content on social media in order to help people who are combating depression, anxiety, melancholy, etc. This paper presents several methodologies that can detect hope in social media comments. We have traversed through transfer learning of several state-of-the-art transformer models for languages such as English, Tamil, and Malayalam. Due to its superior fine-tuning method, ULMFiT achieves an F1-score of 0.9356 on English data. We observe that mBERT achieves 0.8545 on Malayalam test set and distilmBERT achieves 0.5926 weighted F1-score on Tamil test set.

References

Ghada M. Abaido. 2020. Cyberbullying on social media platforms among university students in the united arab emirates. *International Journal of Adolescence and Youth*, 25(1):407–420.

Abien Fred Agarap. 2019. Deep learning using rectified linear units (relu).

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. Deep learning models for multilingual hate speech detection.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*.

Mohit Bhardwaj, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2020. Hostility detection dataset in hindi.

Bharathi Raja Chakravarthi. 2020a. 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. 2020b. *Leveraging orthographic information to improve machine translation of under-resourced languages*. Ph.D. thesis, NUI Galway.

Bharathi Raja Chakravarthi, Mihael Arcan, and John P. McCrae. 2018. Improving wordnets for under-resourced languages using machine translation. In *Proceedings of the 9th Global Wordnet Conference*, pages 77–86, Nanyang Technological University (NTU), Singapore. Global Wordnet Association.

Bharathi Raja Chakravarthi, Mihael Arcan, and John P. McCrae. 2019. WordNet gloss translation for under-resourced languages using multilingual neural machine translation. In *Proceedings of the Second Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation*, pages 1–7, Dublin, Ireland. European Association for Machine Translation.

Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John Philip McCrae. 2020a. A sentiment analysis dataset for code-mixed Malayalam-English. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 177–184, Marseille, France. European Language Resources association.

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, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020b. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210, Marseille, France. European Language Resources association.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Shardul Suryawanshi, Navya Jose, Elizabeth Sherly, and John P. McCrae. 2020c. Overview of the Track on Sentiment Analysis for Dravidian Languages in Code-Mixed Text. In *Forum for Information Retrieval Evaluation*, FIRE 2020, page 21–24, New York, NY, USA. Association for Computing Machinery.

Cheng-Han Chiang, Sung-Feng Huang, and Hung-yi Lee. 2020. Pretrained language model embryology: The birth of ALBERT. In *Proceedings of the 2020*

Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6813–6828, Online. Association for Computational Linguistics.

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.

Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, Hiroshi Noji, Pierre Zweigenbaum, and Jun'ichi Tsujii. 2020. CharacterBERT: Reconciling ELMo and BERT for word-level open-vocabulary representations from characters. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6903–6915, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Antigoni-Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2018. A unified deep learning architecture for abuse detection.

Nikhil Kumar Ghanghor, Parameswari Krishnamurthy, Sajeetha Thavareesan, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2021a. IIITK@DravidianLangTech-EACL2021: Offensive Language Identification and Meme Classification in Tamil, Malayalam and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, Online. Association for Computational Linguistics.

Nikhil Kumar Ghanghor, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021b. 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. Association for Computational Linguistics.

Adeep Hande, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2020. KanCMD: Kannada CodeMixed dataset for sentiment analysis and offensive language detection. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 54–63, Barcelona, Spain (Online). Association for Computational Linguistics.

Siddhanth U Hegde, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. UVCE-IIITT@DravidianLangTech-EACL2021: Tamil Troll Meme Classification: You need to Pay more Attention. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2018. Densely connected convolutional networks.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging.

Navya Jose, Bharathi Raja Chakravarthi, Shardul Suryawanshi, Elizabeth Sherly, and John P. McCrae. 2020. A Survey of Current Datasets for Code-Switching Research. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 136–141.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*.

Kawaljeet Kapoor, Kuttimani Tamilmani, Nripendra Rana, Pushp Patil, Yogesh Dwivedi, and Sridhar Nerur. 2018. Advances in social media research: Past, present and future. *Information Systems Frontiers*, 20.

klaus krippendorff. 2011. Computing krippendorff's alpha-reliability.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

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.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2017. Regularizing and optimizing lstm language models.

Shriphani Palakodety, Ashiqur R. KhudaBukhsh, and Jaime G. Carbonell. 2020. Hope speech detection: A computational analysis of the voice of peace.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pages 8026–8037. Curran Associates, Inc.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Mani Vegupatti, and John P. McCrae. 2020. Named Entity Recognition for Code-Mixed Indian Corpus using Meta Embedding. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 68–72.

Hammad Rizwan, Muhammad Haroon Shakeel, and Asim Karim. 2020. Hate-speech and offensive language detection in Roman Urdu. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2512–2522, Online. Association for Computational Linguistics.

Sebastian Ruder, Anders Søgaard, and Ivan Vulić. 2019. Unsupervised cross-lingual representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 31–38, Florence, Italy. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

M. Schuster and K. K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Mike Schuster and K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.*, 45:2673–2681.

Alice Tontodimamma, Eugenia Nissi, Annalina Sarra, and Lara Fontanella. 2021. Thirty years of research into hate speech: topics of interest and their evolution. *Scientometrics*, 126(1):157–179.

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. Huggingface's transformers: State-of-the-art natural language processing.

Y. Wu, Mike Schuster, Z. Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, M. Krikun, Yuan Cao, Q. Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, Taku Kudo, H. Kazawa, K. Stevens, G. Kurian, Nishant Patil, W. Wang, C. Young, J. Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, G. S. Corrado, Macduff Hughes, and J. Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv*, abs/1609.08144.

Konthala Yasaswini, Karthik Puranik, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. IIITT@DravidianLangTech-EACL2021: Transfer Learning for Offensive Language Detection in Dravidian Languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28, pages 649–657. Curran Associates, Inc.