IDIAP Submission@LT-EDI-ACL2022 : Hope Speech Detection for Equality, Diversity and Inclusion

Muskaan Singh, Petr Motlicek
IDIAP Research Institute,
Martigny, Swizerland
(msingh,petr.motlicek)@idiap.ch

Abstract

Social media platforms have been provoking masses of people. The individual comments affect a prevalent way of thinking by moving away from preoccupation with discrimination, loneliness, or influence in building confidence, support, and good qualities. This paper aims to identify hope in these social media posts. Hope significantly impacts the well-being of people, as suggested by health professionals. It reflects the belief to achieve an objective, discovers a new path, or become motivated to formulate pathways. In this paper we classify given a social media post, *hope speech or not hope speech*, using ensembled voting of BERT, ERNIE 2.0 and RoBERTa for English language with 0.54 macro F1-score (2^{st} rank). For non-English languages Malayalam, Spanish and Tamil we utilized XLM RoBERTA with 0.50, 0.81, 0.3 macro F1 score (1^{st}, 1^{st},3^{rd} rank) respectively. For Kannada, we use Multilingual BERT with 0.32 F1 score(5^{th})position. We release our code-base here https://github.com/Muskaan-Singh/Hate-Speech-detection.git

1 Introduction and Related Work

Hope plays a significant role in well-being, (Milk, 1997), recuperation, and restoration of human life by health professionals. Hope provides a belief for an individual to discover and utilize their pathways (Chang, 1998). It gives the problem-solving ability and coping with various challenges to one objective (Snyder et al., 1991; Cover, 2013; Youssef and Luthans, 2007). In this work, we aim to identify this hope through social media comments by individuals as these comments promote confidence, support, good qualities, shifting the vision of thinking from preoccupation with discrimination or loneliness. Social media has influenced hate-related crimes or spread hatred. Social media platforms such as Facebook, YouTube, Twitter are working tirelessly to detect and bring down such hateful content from their platforms. Since hate content must not be confused with Freedom of speech and expression, thus it becomes quite challenging to reduce the number of false positives.

Earlier attempts for hope speech detection, in LT-EDI-2021 workshop (Huang and Bai, 2021) involves best-performing model uses a combination of XLM and RoBERTa, XLM-RoBERTa language model (Conneau et al., 2019a). It also addressed non-English language comments by using TF-IDF to filter out the error due to multilingualism and code-mixing after extracting the weighted output of the final layer of the XLM-RoBERTa model. Another attempt by (Gundapu and Mamidi, 2021) with language identification model to detect non-English hope speech. The classification architecture presented a transformer-based ensembled architecture consisting of a BERT pre-trained model and a language identification model. Further (Rajput et al., 2021), presented a simple classification model which initially created the static BERT (Devlin et al., 2018) embeddings matrix of the data to extract the contextual information of the data and then experimented with various Deep Neural Networks (DNN) to train a binary classifier. Motivated from the last year's best performing submission in LT-EDI-2021 using the transformers, we ensemble various transformers and utilize the predicted labels with voting.

2 Shared Task Description

The shared task comprised of Hope Speech Detection for Equality, Diversity, and Inclusion (HopeEDI) (Chakravarthi, 2020; Chakravarthi and Muralidaran, 2021). We are provided with social media comments for English, Kannada, Malayalam, Spanish and Tamil languages. We participated in all languages. The dataset consists of annotation with *Hope Speech, Not Hope Speech* for training development sets, respectively. We have reported the dataset statistics in detail in Table3.

Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pages 350 - 355

Label	Language-wise distribution (Train + Dev)				
	English	**Kannada**	**Malayalam**	**Spanish**	**Tamil**
Hope Speech	2234	1909	1858	660	7084
Not Hope Speech	23347	3649	6989	660	8870

Table 1: Data distribution for the HopeEDI database.

Comment	Label
all lives matter .without that we never have peace so to me forever all lives matter.	Hope Speech
Only one race the Human Race	Hope Speech
She saves lives with her music.	Not Hope Speech
Police are already killing people	Not Hope Speech

Table 2: Examples for Hope Speech, Not Hope Speech in the HopeEDI dataset.

Table 3: Dataset Statistics for training, development and test sets for English, Kannada, Malayalam, Spanish and Tamil

	Train	**Dev**	**Test**
English	22739	2840	388
Kannada	4939	617	617
Malayalam	7872	973	1070
Spanish	990	330	330
Tamil	14198	1754	1760

We also did report the hope speech and not speech labels data distribution for all the languages in Table 1. Some examples for the hope speech and not hope speech comments are presented in Table 10. Baseline code with machine learning algorithms (Multinomial Naive Bayes, SVM, KNN, Logistic Regression, and Decision trees) are also provided to the participants.

3 System Description

In this section, we provide a detailed explanation of our system submission. In this paper, we have proposed a pipeline architecture with data pre-processing Section: 3.1, feature extraction in Section: 3.2 and proposed ensembled voting model in Section:3.3.

3.1 Data Pre-processing

Social media comments are usually unstructured data with special characters. We apply preliminary pre-processing removed stopwords, emoji, and punctuation removal with NLTK library (Loper and Bird, 2002).

3.2 Feature Extraction

We tokenize all the sentences and map the tokens to their word IDs to extract features. For every sentence in the dataset, we follow a series of steps (i) tokenize the sentences (ii) prepend the [CLS] token to the start (iii) append the [SEP] token to the end (iv) map the token to their IDs (v) pad or truncate the sentenced to max length (vi) mapping of attention masks for [PAD] tokens. We padded and truncated the max_length=30. The generated sequence sentences are passed for encoding with its attention mask (differentiating padding from non-padding).

3.3 Proposed Methodology

For English language, we formulate an ensembled voting classifier with BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019) and ERNIE (Sun et al., 2020). Firstly, we began encoding comments with specific pre-trained embeddings for formulating the matrix.

3.3.1 BERT

Bidirectional Encoder Representations from Transformers (BERT) involves pre-training deep bi-directional transformers for language understanding. It utilizes unlabeled text by jointly training the left and proper context in all layers. BERT takes input as a concatenation of two segments (sequences of tokens),$x\_1, ..., x\_Nandy\_1, ..., y\_M$. Segments usually consist of more than one natural sentence. The two segments are presented as a single input sequence to BERT with special tokens delimiting them: $[CLS], x\_1, ..., x\_N, [SEP], y\_1, ..., y\_M, [EOS]$. M and N are constrained such that $M + N < T$, where T is a parameter that controls the maximum

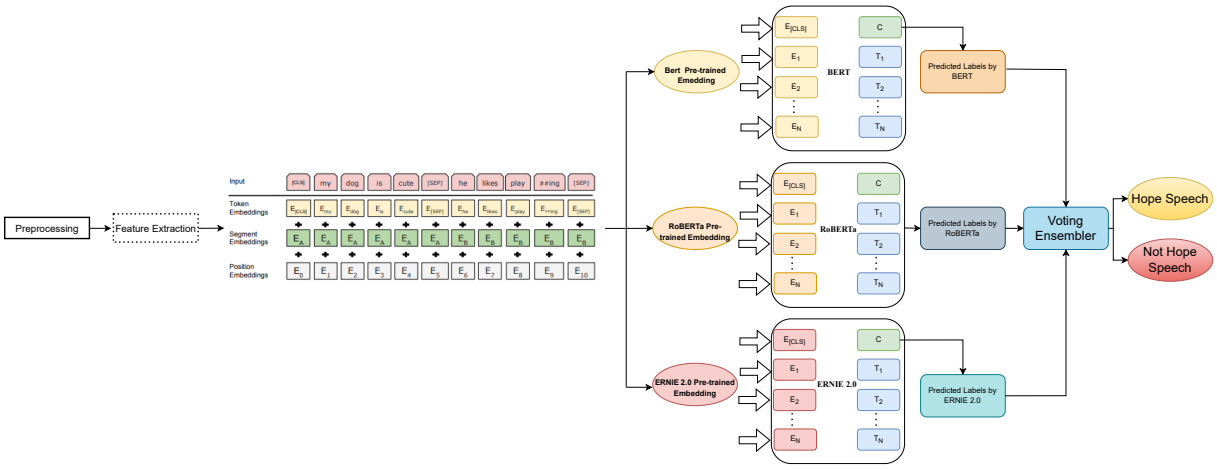

Figure 1: Proposed Methodology for Hope Speech Detection

sequence length during training. Fine-tuning of the pre-trained model can be easily handled by adding the output layer to create state-of-art models for various NLP tasks without substantial task-specific architecture modification.

3.3.2 RoBERTa

Robustly Optimized BERT approach has emphasized data being used for pre-training and the number of passes for training. The BERT model is optimized with dynamic masking, more extended training with big batches over more data, removing the next prediction objective, and dynamically changing masking patterns for training data. The model achieved state-of-art results on GLUE, RACE, and SQuAD without multi-task finetuning for GLUE or additional data for SQuAD.

3.3.3 ERNIE 2.0

ERNIE 2.0 is another continual pre-training framework that efficiently supports customized training tasks in multi-task learning incrementally. The pre-trained model is finetuned to adapt to various language understanding tasks. The framework has demonstrated significant improvement over BERT and XLNET on approximately 16 tasks, including GLUE.

Further, due to limited models for multilingual, we restricted our experiment for Malayalam, Spanish and Tamil languages to XLM ROBERTA (Conneau et al., 2019b). It significantly aims at cross-lingual transfer tasks for pre-trained multilingual language models. The model performs exceptionally well on low resource languages at a scale. The empirical analysis presents positive transfer and capacity delusion. Further, the model also allows multilingual modeling without sacrificing per-language performance. It has shown competitive results with strong monolingual models on GLUE.

For Kannada, we utilize Multilingual BERT (M-BERT) (Pires et al., 2019), released by Devlin et al. (2019). It is a language model trained with monolingual corpora in 104 languages. It reports exceptional results on zero-shot cross-lingual model transfer. Task-specific annotations for a language are used to finetune evaluation on others—the multilingual representation exhibits systematic deficiencies affecting some language pairs.

3.3.4 Experimental Setup

We use V1 100 GPU with 53GB RAM alongside 8 CPU cores for the experimental setup. We divide the entire dataset with a 90:10 train and validation split of eight batches, with a learning rate (1e-5) and Adam optimizer (Kingma and Ba, 2014) with epsilon (1e-8). We feed a seed_val of 42. For calculating the training loss over all the batches, we use gradient descents (Andrychowicz et al., 2016) with clipping the norm to 1.0 to avoid exploding gradient problem.

3.4 Comparative Approaches explored

We explore a couple of other methods as presented in Table: 11 and 9 for system submission for detecting hope and not-hope speech from social media comments. We experimented with ERNIE 2.0, RoBERTa, XLNET, and BERT and ensembled best-performing approaches i.e., BERT, ERNIE 2.0, and RoBERTa. The results depict ensemble results are outperforming all other experimented models for English. While for Tamil, Malayalam, and Spanish, we see XLR-RoBERTa performs exceptionally bet-

	Precision		Recall		F1-Score	
Model	**Marco**	**Weighted**	**Macro**	**Weighted**	**Macro**	**Weighted**
Top performing	0.56	0.87	0.54	0.89	0.55	0.88
Proposed model	0.55	0.87	0.54	0.88	0.54	0.87
Average score	0.47	0.85	0.46	0.80	0.43	0.80

Table 4: Comparison with the top-performing model results for English

	Precision		Recall		F1-Score	
Model	**Macro**	**Weighted**	**Macro**	**Weighted**	**Macro**	**Weighted**
Top performing	0.30	0.39	0.34	0.46	0.32	0.42
Proposed model	0.29	0.38	0.33	0.44	0.30	0.40
Average score	0.28	0.375	0.33	0.438	0.303	0.39

Table 5: Comparison with the top-performing model results for Tamil

	Precision		Recall		F1-Score	
Model	**Macro**	**Weighted**	**Macro**	**Weighted**	**Macro**	**Weighted**
Proposed model	0.64	0.76	0.53	0.79	0.50	0.75
Average score	0.45	0.67	0.45	0.73	0.44	0.69

Table 6: Comparison with the top-performing model results for Malayalam

	Precision		Recall		F1-Score	
Model	**Macro**	**Weighted**	**Macro**	**Weighted**	**Macro**	**Weighted**
Proposed model	0.81	0.81	0.81	0.81	0.81	0.81
Average score	0.79	0.79	0.79	0.79	0.79	0.79

Table 7: Comparison with the top-performing model results for Spanish

	Precision		Recall		F1-Score	
Model	**Macro**	**Weighted**	**Macro**	**Weighted**	**Macro**	**Weighted**
Top performing	0.49	0.74	0.48	0.76	0.48	0.75
Proposed model	0.31	0.53	0.32	0.54	0.32	0.54
Average score	0.41	0.65	0.41	0.64	0.40	0.64

Table 8: Comparison with the top-performing model results for Kannada

	Tamil			**Malayalam**			**Spanish**			**Kannada**		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
M-BERT	0.64	0.61	0.60	0.72	0.62	0.64	0.75	0.75	0.75	**0.72**	**0.70**	**0.71**
XLM-R	**0.65**	**0.63**	**0.63**	**0.76**	**0.63**	**0.66**	**0.82**	**0.81**	**0.81**	0.70	0.69	0.69

Table 9: Comparative approaches explored for the system submission to classify hate and non-hate speech for Tamil, Malayalam, Spanish and Kannada

Comment	**Predicted Label**
9.20 To never give hope - to never give up ! She said it with conviction .	Hope Speech
how can you disrespect your own body? It is YOURS!	Not Hope Speech
Maddona saved my Soul in 1999	Not Hope Speech

Table 10: Qualitative Results for Hope Speech, Not Hope Speech

	P	R	F1
ERNIE 2.0	0.8	**0.73**	0.76
BERT	0.81	0.7	0.75
RoBERTa	0.8	0.71	0.75
XLNET	0.8	0.72	0.74
Ensemble	**0.81**	0.72	**0.76**

Table 11: We explored comparative analysis for the system submission to classify hate and non-hate speech for the English language. In the ensemble approach, we choose the best of all the models (ERNIE+BERT+RoBERTa).

ter than M-BERT. For Kannada, M-BERT performs distinctly well.

4 Results and Analysis

We evaluate our model quantitatively and qualitatively for the HopeEDI dataset. The classification report for our proposed model with average and best submission among all the teams is reported in Table: 8. The proposed model has shown progressive results with 0.55, 0.54, 0.54 F1 for English, Tamil, Malayalam, Kannada, Spanish on the leaderboardhttps://competitions.codalab.org/competitions/36393#learn_the_details-result with (2^{st},1^{st},1^{st},3^{rd},5^{th}) rank respectively.

- For the English language, 0.55, 0.54, 0.54 are the reported precision, Recall, and F1-score, which is relatively 0.08, 0.08, 0.11 more than the average and 0.01, 0, 0.01 less for best-performing submission, respectively.
- For the Tamil language, 0.29, 0.33, 0.30 are the reported precision, Recall, and F1-score, which is relatively 0.01, 0, 0.003 more than the average and 0.01, 0.01, 0.02 less for best-performing submission, respectively.
- For the Malayalam language, 0.64, 0.53, and 0.50 are the reported precision, Recall, and F1-score, relative, 0.19, 0.08, and 0.06 more than the average. As we were the best-performing submission, we did not report the scores and differences from our submission.
- For the Spanish language, 0.81, 0.81, and 0.81 are the reported precision, Recall, and F1-score, relative, 0.02, 0.02, and 0.02 more than the average. As we were the best-performing submission, we did not report the scores and differences from our submission.
- For the Kannada language, 0.31, 0.32, 0.32 are the reported precision, Recall, and F1-score, which is relatively 0.01, 0.09, 0.08 less than the average, and 0.18, 0.16, 0.16 less for best-performing submission, respectively.

Additionally, we also evaluate our prediction results qualitatively in Table 10. The results display useful predictions; for hope speech, see Instance 1, *"never give up hope,"* portrays a sense of hope in the person writing it. While for non-hope speech, the terms *"how can you disrespect your own body? It is YOURS."* show that the model focuses on the negative expressions and can successfully understand the context of the statement. The last example we have presented, *"Maddona saved my Soul in 1999,"* is classified as non-hope speech, which indicates that the model fails to understand the context of the entire statement and focuses more on the sentiments of the words. As it is clearly understood, the person who wrote this got a sense of hope from Maddona; this statement can be classified as a hope speech. However, the model has predicted it as not hope speech, which is a false positive case.

5 Conclusion

In this paper, we classify given a social media post, *hope speech or not hope speech*, using ensembled voting of BERT, ERNIE 2.0, and RoBERTa for the English language with 0.54 macro F1-score (2^{st} rank). For non-English languages Malayalam, Spanish and Tamil we utilized XLM RoBERTA with 0.50, 0.81, 0.3 macro F1 score (1^{st}, 1^{st},3^{rd} rank) respectively. For Kannada, we use Multilingual BERT with 0.32 F1 score(5^{th})position. We also performed a qualitative analysis. The system performs quite well to recognize the comments for hope speech; In the future, we intend to work on a multi-task learning framework to handle all kinds of hate speech (aggression, misogyny, racism). We also aim to detect multilingual hope speech in the code-mixing scenarios.

Acknowledgements

This work was supported by the European Union's Horizon 2020 research and innovation program under grant agreement No. 833635 (project ROXANNE: Real-time network, text, and speaker analytics for combating organized crime, 2019-2022).

References

Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. 2016. Learning to learn by gradient descent by gradient descent. *Advances in neural information processing systems*, 29.

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*, pages 61–72, Kyiv. Association for Computational Linguistics.

Edward C Chang. 1998. Hope, problem-solving ability, and coping in a college student population: Some implications for theory and practice. *Journal of clinical psychology*, 54(7):953–962.

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

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

Rob Cover. 2013. Queer youth resilience: Critiquing the discourse of hope and hopelessness in lgbt suicide representation. *M/C Journal*, 16(5).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Sunil Gundapu and Radhika Mamidi. 2021. Autobots@LT-EDI-EACL2021: One world, one family: Hope speech detection with BERT transformer model. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 143–148, Kyiv. Association for Computational Linguistics.

Bo Huang and Yang Bai. 2021. Team hub@LT-EDI-EACL2021: Hope speech detection based on pretrained language model. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 122–127, Kyiv. Association for Computational Linguistics.

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

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*.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

Harvey Milk. 1997. The hope speech. *We are everywhere: A historical sourcebook of gay and lesbian politics*, pages 51–53.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Gaurav Rajput, Narinder Singh Punn, Sanjay Kumar Sonbhadra, and Sonali Agarwal. 2021. Hate speech detection using static bert embeddings. In *Big Data Analytics: 9th International Conference, BDA 2021, Virtual Event, December 15-18, 2021, Proceedings*, page 67–77, Berlin, Heidelberg. Springer-Verlag.

CR Snyder, Ch Harris, JR Anderson, and SA Holleran. 1991. Irving. *LM, Sigmon, ST, Yoshinobu, L., Gibb, J., Langelle, C., & Harney, P*, pages 570–585.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8968–8975.

Carolyn M Youssef and Fred Luthans. 2007. Positive organizational behavior in the workplace: The impact of hope, optimism, and resilience. *Journal of management*, 33(5):774–800.