Sentiment Analysis of Dravidian Code Mixed Data

Asrita Venkata Mandalam
Department of CSIS
BITS Pilani, Pilani Campus
Rajasthan, India
f20171179
@pilani.bits-pilani.ac.in

Yashvardhan Sharma
Department of CSIS
BITS Pilani, Pilani Campus
Rajasthan, India
yash
@pilani.bits-pilani.ac.in

Abstract

This paper presents the methodologies implemented while classifying Dravidian code-mixed comments according to their polarity. With datasets of code-mixed Tamil and Malayalam available, three methods are proposed - a sub-word level model, a word embedding based model and a machine learning based architecture. The sub-word and word embedding based models utilized Long Short Term Memory (LSTM) network along with language-specific preprocessing while the machine learning model used term frequency–inverse document frequency (TF-IDF) vectorization along with a Logistic Regression model. The sub-word level model was submitted to the the track 'Sentiment Analysis for Dravidian Languages in Code-Mixed Text' proposed by Forum of Information Retrieval Evaluation in 2020 (FIRE 2020). Although it received a rank of 5 and 12 for the Tamil and Malayalam tasks respectively in the FIRE 2020 track, this paper improves upon the results by a margin to attain final weighted F1-scores of 0.65 for the Tamil task and 0.68 for the Malayalam task. The former score is equivalent to that attained by the highest ranked team of the Tamil track.

1 Introduction

Code-mixing usually involves two languages to create another language that consists of elements of both in a structurally understandable manner. It has been noted that bilingual and multilingual societies use multiple languages together in informal speech and text. Dravidian code-mixed languages, including but not limited to Malayalam and Tamil, are increasingly used by younger generations in advertising, entertainment and social media. The language is commonly written in Roman script.

With the rise in the number of non-English and multilingual speakers using social media, there is an interest in analysing the sentiment of the content posted by them. As code-mixed data does not belong to one language and is often written using Roman script, identifying its polarity cannot be done using traditional sentiment analysis models(Puranik et al., 2021; Hegde et al., 2021; Yasaswini et al., 2021; Ghanghor et al., 2021b,a). In social media, low-resourced languages such as Tamil and Malayalam have been increasingly used along with English (Thavareesan and Mahesan, 2019, 2020a,b). Indian government declared Tamil as classical language in 2004, meaning that it meets three criteria: its sources are old; it has an independent tradition; and it has a significant body of ancient literature. The most ancient non-Sanskritic Indian literature is found in Tamil. The earliest existing Tamil literary works and their commentaries commemorate the organisation of long-term Tamil Sangams (600 BCE to 300 CE) by the Pandiyan Rulers, which studied, established and made improvements in the language of Tamil. Identifying the sentiment of this data can prove to be useful in social media monitoring and feedback of users towards other online content.

The 'Sentiment Analysis for Dravidian Languages in Code-Mixed Text' task proposed by FIRE 2020 (Chakravarthi et al., 2020c; Chakravarthi, 2020) contains a code-mixed dataset consisting of comments from social media websites for both Tamil and Malayalam. Each team had to submit a set of predicted sentiments for the Tamil-English and Malayalam-English mixed test sets (Chakravarthi et al., 2020d). Along with language specific preprocessing techniques, the implemented model made use of sub-word level representations to incorporate features at the morpheme level, the smallest meaningful unit of any language. Evaluated by weighted average F1-score, the sub-word level approach achieved the 5th highest score

Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pages 46–54

in the Tamil task and the 12th rank in the Malayalam task (Sharma and Mandalam, 2020).

This paper presents the aforementioned model and two more models - a word embedding based model and a machine learning model in an attempt to improve upon the submitted weighted F1-scores. The highest scoring model for the Tamil dataset received a score equivalent to the highest scoring model presented at the Sentiment Analysis for Dravidian Languages in Code-Mixed Text track of FIRE 2020. Although they achieved it using a Bidirectional Encoder Representations from Transformers (BERT) based model, the model presented in this paper uses Word2Vec and FastText embeddings to achieve the same score. The implemented models are available on Github[1].

2 Related Work

Analysing the sentiment of code-mixed data is important as traditional methods fail when given such data. Barman et al. (2014) concluded that n-grams proved to be useful in their experiments that involved multiple languages with Roman script.

Bojanowski et al. (2017) used character n-grams in their skip-gram model. The lack of preprocessing resulted in a shorter training time and outperformed baselines that did not consider sub-word information. Joshi et al. (2016) outperformed existing systems as well by using a sub-word based LSTM architecture. Their dataset consisted of 15% negative, 50% neutral and 35% positive comments. As their dataset was imbalanced like the one used in this paper, the submitted approach involved morpheme extraction as it would help in identifying the polarity of the dataset. Using a hybrid architecture, Lal et al. (2019) an F1-score of 0.827 on the Hindi-English dataset released by Joshi et al. (2016). After generating the sub-word level information of the input data, they used one Bidirectional LSTM to figure out the sentiment of the given sentence and another to select the sub-words that contributed to the overall sentiment of the sentence.

In more recent work, Jose et al. (2020) surveyed publicly available code-mixed datasets. They noted statistics about each dataset such as vocabulary size and sentence length. Priyadharshini et al. (2020) used embeddings of closely related languages of the code-mixed corpus to predict Named Entities of the same corpus.Yadav and Chakraborty (2020) used multilingual embeddings to identify the sentiment of code-mixed text. As they used an unsupervised approach, they performed sentiment analysis by using the embeddings to transfer knowledge from monolingual text to their code-mixed input. Yadav et al. (2020) designed an ensemble model consisting of four classifiers - Naive Bayes, Support-Vector Machines (SVM), Linear Regression, and Stochastic Gradient Descent (SGD) classifiers. They tested it on a Hindi-English code-mix dataset and managed to surpass the scores attained by the baseline and state-of-art systems.

The proposed models test three approaches - sub-word analysis, word-level embeddings and machine learning algorithm based classification. The top few submissions submitted at the 'Sentiment Analysis for Dravidian Languages in Code-Mixed Text' task proposed by FIRE 2020 (Chakravarthi et al., 2020c) used BERT-based models. This work aims to attain the same F1-score with help of non-BERT-based approaches. As Dravidian code-mixed languages are under-resourced, this paper also describes the preprocessing methods used to aid in the sentiment classification of comments that contain both Roman and Tamil/Malayalam characters.

3 Dataset

The model has been trained, validated and tested using the Tamil (Chakravarthi et al., 2020b) and Malayalam (Chakravarthi et al., 2020a) datasets provided by the organizers of the Dravidian Code-Mix FIRE 2020 task. The Tamil code-mix dataset consists of 11,335 comments for the train set, 1,260 for the validation set and 3,149 comments for testing the model. In the Malayalam code-mix dataset, there are 4,851 comments for training, 541 for validating and 1,348 for testing the model. Table 1 gives the distribution of each sentiment in each dataset.

4 Proposed Technique

4.1 Sub-word level model

The submitted approach used a sub-word level model as it accounts for words that have a similar morpheme. For example, in the Tamil dataset, *Ivan*, *Ivanga* and *Ivana* have similar meanings due to their root word *Ivan*.

First, the dataset is preprocessed to replace all emojis with their corresponding description in English. As the dataset contains both Roman and

[1] https://github.com/avmand/SA_Dravidian

Dataset	Positive	Negative	Mixed Feelings	Unknown State	Other Languages
Tamil	10,559	2,037	1,801	850	497
Malayalam	2,811	738	403	1,903	884

Table 1: Distribution of Data in the Tamil and Malayalam Dataset

Tamil (or Malayalam) characters, the latter is replaced with its corresponding Roman script representation.

From the preprocessed data, a set of characters was obtained. The input to the model is a set of character embeddings. The sub-word level representation is generated through a 1-D convolution layer with activation as ReLU, size of convolutional window as 5 and number of output filters as 128. After getting a morpheme-level feature map, a 1-D maximum pooling layer is used to obtain its most prominent features. To obtain the connections between each of these features, LSTMs are used due to their ability to process sequences and retain information. The first and second LSTM layers have a dropout of 0.4 and 0.2 respectively. Finally, it is passed to a fully connected layer. Batch normalization has been used in the model to prevent overfitting. While training the model, early stopping has been utilized to stop training when the validation loss shows no improvement after 4 epochs. The training data was shuffled before each epoch.

4.2 Word Embedding model

This model utilizes two word embedding architectures - continuous bag of words (CBOW) and FastText. This approach was tested after the track had ended and submissions were no longer accepted.

After replacing all of the emojis with their description, non-Tamil and non-Malayalam characters were replaced with their corresponding Roman script representation. Words with multiple repeating characters were shortened to include only one of those multiple characters.

The CBOW model utilized the Word2Vec (Mikolov et al., 2013) method provided by the *gensim* library. As the average length of each comment was 50, the window size was taken as 15. The minimum frequency of words that were taken into consideration was 2 and the architecture was trained on the training data for 10 epoch cycles. The FastText model was trained for 10 iterations and had a window size of 15 as well. These word vectors were created with a dimensionality of 100 and were concatenated to create an embedding matrix. For the classification of the input data, two LSTM layers were used after the embedding layer. The dimensionality of the output space of the first and second LSTM layers were 64 and 32 repectively. Similar to the previous model, batch normalization and early stopping were used with the model automatically stopping training after 5 epoch cycles.

4.3 Machine Learning model

Till now, all of the tested models used deep learning based classifiers. As the input data is skewed and is considerably small, two different machine learning models were used to predict the sentiment of the given code-mixed comments.

Before using the classifier, the data was preprocessed and feature extraction using TF-IDF vectorization. The top 5000 features were extracted. Originally, only 300 features were extracted but increasing the number of features resulted in an increase in the weighted F1-score. The features included both unigrams and bigrams. Sub-linear term frequency scaling was applied.

Two different classifiers were tested with the help of the scikit-learn library- Linear Support Vector Machine Classification (Linear SVC) and Logistic Regression. The tolerance for the Linear SVC model was '1e-6' and the multi-class strategy used was 'crammer_singer'. For the Logistic Regression model, the 'newton-cg' algorithm was used for optimization. A binary problem was fit for each of the labels by using the 'ovr' option for the multi_class parameter. The value for the inverse of regulation strength (C) was tested with values in the range of 1 to 20 inclusive.

5 Result

The submitted run achieved a rank of 5 and 12 for the Tamil and Malayalam tasks respectively. There were a total of 32 teams that submitted their results for the Tamil task. For the Malayalam task, 28 teams submitted their solution. The final rank was evaluated based on the weighted average F1-score. The classification report is shown in Table 2. The Tamil task received Precision, Recall and an F1-

score of 0.62, 0.66 and 0.61 respectively. For the Malayalam task, the submission received scores of 0.67, 0.59 and 0.60 respectively.

Out of the models that were tested after the deadline for the FIRE 2020 track, the word embedding model performed the best for the Tamil dataset. It attained a weighted F1-score of 0.65, which is the same as the team that scored the highest in the Dravidian Code-Mix FIRE 2020 track. For the Malayalam dataset, the Logistic Regression based model scored the highest with a weighted F1-score of 0.68. For both datasets, the Linear SVC model performed worse than the Logistic Regression model by at least 1%. It was noted that the optimal value for C was 12 for the Tamil dataset and 5 for the Malayalam dataset.

6 Error Analysis

6.1 Tamil Task

6.1.1 Sub-word level model

From Table 2, one can see that the F1-score of the positive comments is the highest with a value of 0.80. The next highest score is only at 0.46, attained by the class of comments that are not in Tamil. The order of classes from the highest to the lowest F1-scores are Positive, Not Tamil, Negative, Mixed Feelings and Unknown State. The weighted F1-score is lower than the Precision and Recall as the weighted score takes into account the proportion of each class in the dataset.

Due to the higher number of positive comments in the overall dataset, it is not surprising that the model trains well and produces the best results for that class. Non-Tamil comments get the next highest score due to the different morphemes used in them. These comments are usually in a different Indian language like Hindi or Telugu and are written using the English alphabet. Some comments are written in the script of their respective language. This class does not achieve a higher score due to words that they have in common with the Tamil-English code-mixed comments such as *Rajinikanth* and *Thalaiva*. The same can be concluded for the negative label as well as it had many words that were common with those of the positive comments. Comments from the mixed feelings class were misclassified as either positive or negative. They were not misclassified as comments from the unknown state class possibly due to the relatively lower ratio of unknown comments as compared to the positive and negative classes. As these comments contained both positive and negative sentiments, there was a much higher chance of them being classified into one of those classes. The unknown state class receives the lowest F1-score. Its precision is 0.67 but its recall is very low at 0.01. This implies that there is a high false negative rate and is because all of the comments use words from the Tamil vocabulary. Most of those words are common with those of the positive class. Table 5 gives a representation of the misclassified Tamil comments.

6.1.2 Machine Learning model

The machine learning based model performed better than the sub-word level model for the Tamil dataset. Out of the tested machine learning models, the Logistic Regression model performed better, with a weighted F1-score of 0.62. The detailed result can be seen in Table 4. As the Linear SVC model received poorer results with a weighted F1-score of 0.61, it has not been represented in any of the tables. There has been an improvement in the weighted F1-scores of the negative, non-Tamil and unknown state classes. An interesting improvement is the increase in the F1-score of the unknown class as it jumped from 0.02 to 0.22. Table 7 represents the distribution of predicted classes for each class.

6.1.3 Word Embedding model

The word embedding based model performed the best out of the tested models for this dataset with a weighted F1-score of 0.65, the same score as the highest ranked model submitted to the Sentiment Analysis of Dravidian Code Mixed Data FIRE 2020 track. The result of this model can be seen in Table 3. In this model as well, the positive class performed the best with an F1-score of 0.81. Compared to the other two tested models, this model received a higher F1-score for the mixed feelings, negative, positive and unknown state classes. For the classification of comments that were not in Tamil, this class performed much better than the sub-word model but slightly worse than the machine learning model. Table 6 shows that most of the misclassified comments were classified as positive. For the positive comments, the misclassified comments were mainly misclassified into the negative class. This is because negative and positive words, although they may be opposites, are used in similar sentences with a similar word set. This might have led the word embedding models to assign close vectors for them. Another reason for the same might be the fact that the negative class had

Language	Report	Precision	Recall	F1-Score	Support
Tamil	Mixed Feelings	0.26	0.17	0.21	377
	Negative	0.42	0.22	0.29	424
	Positive	0.72	0.91	**0.80**	2075
	Not Tamil	0.71	0.34	0.46	100
	Unknown State	0.67	0.01	0.02	173
	Macro Avg	0.55	0.33	0.35	3149
	Weighted Avg	**0.62**	**0.66**	**0.61**	3149
Malayalam	Mixed Feelings	0.21	0.51	0.30	70
	Negative	0.35	0.55	0.43	138
	Positive	0.73	0.73	**0.73**	565
	Not Malayalam	0.58	0.68	0.63	177
	Unknown State	0.81	0.38	0.52	398
	Macro Avg	0.54	0.57	0.52	3149
	Weighted Avg	**0.67**	**0.59**	**0.60**	1348

Table 2: Classification report for each dataset and class for the sub-word level model.

Language	Report	Precision	Recall	F1-Score	Support
Tamil	Mixed Feelings	0.19	0.36	0.25	377
	Negative	0.31	0.49	0.38	424
	Positive	0.70	0.95	**0.81**	2075
	Not Tamil	0.58	0.64	0.61	100
	Unknown State	0.26	0.36	0.30	173
	Macro Avg	0.41	0.56	0.47	3149
	Weighted Avg	**0.56**	**0.78**	**0.65**	3149
Malayalam	Mixed Feelings	0.35	0.40	0.37	70
	Negative	0.39	0.54	0.45	138
	Positive	0.58	0.87	**0.70**	565
	Not Malayalam	0.55	0.85	0.67	177
	Unknown State	0.51	0.84	0.63	398
	Macro Avg	0.47	0.70	0.56	3149
	Weighted Avg	**0.52**	**0.80**	**0.63**	1348

Table 3: Classification report for each dataset and class for the word embedding model.

Language	Report	Precision	Recall	F1-Score	Support
Tamil	Mixed Feelings	0.25	0.10	0.14	377
	Negative	0.40	0.26	0.32	424
	Positive	0.73	0.90	**0.80**	2075
	Not Tamil	0.72	0.57	0.64	100
	Unknown State	0.35	0.16	0.22	173
	Macro Avg	0.49	0.40	0.42	3149
	Weighted Avg	**0.60**	**0.67**	**0.62**	3149
Malayalam	Mixed Feelings	0.65	0.34	0.45	70
	Negative	0.69	0.49	0.57	138
	Positive	0.70	0.81	**0.75**	565
	Not Malayalam	0.72	0.65	0.68	177
	Unknown State	0.64	0.64	0.64	398
	Macro Avg	0.68	0.59	0.62	3149
	Weighted Avg	**0.68**	**0.68**	**0.68**	1348

Table 4: Classification report for each dataset and class for the machine learning model.

the next highest count of comments.

6.2 Malayalam Task

6.2.1 Sub-word level model

The classification report of the Malayalam task can be seen in Table 2. The F1-score of the positive comments is the highest at 0.73. The next highest is at 0.63, for the class of comments that are not in Malayalam. The order of classes from the highest to the lowest F1-scores are Positive, Not Malayalam, Unknown State, Negative and Mixed Feelings.

The Malayalam dataset was more balanced as compared to the Tamil dataset with the second largest class less than 1000 comments behind the largest one. Similar to the Tamil dataset, the positive class has the highest number of comments. This led to the relatively higher F1-score and an equally low false positive and false negative rate. For the class of comments that were not in Malayalam, the classifier identified all of the comments that were not written in the Roman or Malayalam script. However, words that were commonly found in positive comments, such as names of Malayalam actors, and were used with English words were classified incorrectly. For the unknown state class, it is noted that the misclassified comments were majorly assigned a positive, negative or mixed feelings tag. Although the overall sentiment of the sentence was unknown, a portion of that sentence had similarities with one of the other classes. For the mixed feelings class, the same was deduced and most of the wrongly classified comments were assigned either a positive or negative tag. Most of the misclassified comments from the negative class were labelled as comments with mixed feelings. Sarcastic comments that used positive words but implied negative sentiments were not accounted for by the model. The distribution of misclassified comments can be seen in Table 5.

6.2.2 Word Embedding model

The Malayalam dataset received a weighted F1-score of 0.63 when tested upon the word embedding model. Although this is higher than the F1-score obtained by the sub-word level model, it did not attain a greater accuracy than that of the Tamil dataset. This is possibly due to the fact that the Tamil dataset was much larger than the Malayalam dataset. The result can be seen in Table 3. Compared to the sub-word level model, this model managed to gain a higher F1-score in all of the classes except for the positive class. Table 6 represents the distribution of misclassified comments across all of the tested classes.

6.2.3 Machine Learning model

The classification report of this model can be seen in Table 4. This model performed the best out of the three tested models with a weighted average F1-score of 0.68. It outperformed the rest of the models in terms of the F1-score for each class as well. The positive class received the highest score with a value of 0.75. For this dataset, the Logistic Regression model performed better than the Linear SVC model which attained a weighted F1-score of 0.66 and hence, the result of the Linear SVC model has not been recorded in any of the tables. Table 7 represents the distribution of misclassified comments, organized by the expected classification. It can be seen that not many comments were misclassified as not Malayalam and this is mainly due to the difference in vocabulary. Not many comments were misclassified as negative or mixed-feelings. This is partially due to the fact that the top two classes by count are the positive and unknown state classes.

7 Conclusion

This paper presents the submitted approach for the Sentiment Analysis for Dravidian Languages in Code-Mixed Text track of Forum for Information Retrieval Evaluation (FIRE) 2020 and two more models to improve upon the previously attained results. The results show that the positive class in each dataset receives the highest F1-scores, regardless of the model. This is due to the higher ratio of the same as compared to the rest of the classes. Comments that were not in the language of their dataset received the next highest score as their vocabulary included sub-words and words that were not a part of the rest of the datasets. Although, the top ranked submissions (Chakravarthi et al., 2020d) used BERT-based models for the Tamil task of the Dravidian Languages in Code-Mixed Text FIRE 2020 track, the presented work made use of word-embedding modelling and attained a weighted F1-score equal to that achieved by the highest scoring team. For future work, a sarcasm detection feature could be included to avoid misclassification of comments from the positive, negative and mixed-feelings classes.

		Predicted				
Language	Actual	Mixed-Feelings	Negative	Positive	Not Tamil/ Malayalam	Unknown State
Tamil	Mixed-Feelings	65	37	273	2	0
	Negative	61	92	268	3	0
	Positive	117	69	1880	8	1
	Not Tamil	3	12	51	34	0
	Unknown State	8	11	151	1	2
Malayalam	Mixed-Feelings	36	12	18	4	0
	Negative	39	76	12	4	7
	Positive	44	56	413	25	27
	Not Malayalam	5	8	42	121	1
	Unknown State	48	66	77	54	153

Table 5: Error Analysis for each dataset and class for the sub-word level model.

		Predicted				
Language	Actual	Mixed-Feelings	Negative	Positive	Not Tamil/ Malayalam	Unknown State
Tamil	Mixed-Feelings	14	38	311	4	10
	Negative	16	80	315	6	7
	Positive	20	75	1929	15	36
	Not Tamil	0	1	47	51	1
	Unknown State	6	13	130	1	23
Malayalam	Mixed-Feelings	21	1	21	3	24
	Negative	3	49	26	3	57
	Positive	8	14	396	20	127
	Not Malayalam	3	1	35	113	25
	Unknown State	6	9	71	21	291

Table 6: Error Analysis for each dataset and class for the word embedding model.

		Predicted				
Language	Actual	Mixed-Feelings	Negative	Positive	Not Tamil/ Malayalam	Unknown State
Tamil	Mixed-Feelings	36	46	284	5	6
	Negative	32	112	268	4	8
	Positive	63	101	1865	10	36
	Not Tamil	1	5	35	57	2
	Unknown State	12	14	116	3	28
Malayalam	Mixed-Feelings	23	3	28	2	14
	Negative	2	67	41	3	25
	Positive	1	10	463	19	72
	Not Malayalam	0	4	33	113	27
	Unknown State	7	11	101	21	258

Table 7: Error Analysis for each dataset and class for the machine learning model.

Acknowledgments

The authors would like to convey their sincere thanks to the Department of Science and Technology (ICPS Division), New Delhi, India, for providing financial assistance under the Data Science (DS) Research of Interdisciplinary Cyber Physical Systems (ICPS) Programme [DST/ICPS/CLUSTER/Data Science/2018/Proposal-16:(T-856)] at the department of computer science, Birla Institute of Technology and Science, Pilani, India.

Authors are also thankful to the authorities of Birla Institute of Technology and Science, Pilani, to provide basic infrastructure facilities during the preparation of the paper.

References

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 13–23, Doha, Qatar. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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

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, 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 Philip McCrae. 2020c. Overview of the track on Sentiment Analysis for Dravidian Languages in Code-Mixed Text. In *Proceedings of the 12th Forum for Information Retrieval Evaluation*, FIRE '20.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Shardul Suryawanshi, Navya Jose, Elizabeth Sherly, and John Philip McCrae. 2020d. Overview of the track on Sentiment Analysis for Dravidian Languages in Code-Mixed Text. In *Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2020). CEUR Workshop Proceedings. In: CEUR-WS. org, Hyderabad, India*.

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.

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.

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.

Aditya Joshi, Ameya Prabhu, Manish Shrivastava, and Vasudeva Varma. 2016. Towards sub-word level compositions for sentiment analysis of Hindi-English code mixed text. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2482–2491, Osaka, Japan. The COLING 2016 Organizing Committee.

Yash Kumar Lal, Vaibhav Kumar, Mrinal Dhar, Manish Shrivastava, and Philipp Koehn. 2019. De-mixing sentiment from code-mixed text. In *Proceedings of the 57th Annual Meeting of the Association for*

Computational Linguistics: Student Research Workshop, pages 371–377, Florence, Italy. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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.

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.

Yashvardhan Sharma and Asrita Venkata Mandalam. 2020. bits2020@dravidian-codemix-fire2020: Subword level sentiment analysis of dravidian code mixed data. In *Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2020). CEUR Workshop Proceedings. In: CEUR-WS. org, Hyderabad, India*.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2019. Sentiment Analysis in Tamil Texts: A Study on Machine Learning Techniques and Feature Representation. In *2019 14th Conference on Industrial and Information Systems (ICIIS)*, pages 320–325.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020a. Sentiment Lexicon Expansion using Word2vec and fastText for Sentiment Prediction in Tamil texts. In *2020 Moratuwa Engineering Research Conference (MERCon)*, pages 272–276.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020b. Word embedding-based Part of Speech tagging in Tamil texts. In *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*, pages 478–482.

Konark Yadav, Aashish Lamba, Dhruv Gupta, Ansh Gupta, Purnendu Karmakar, and Sandeep Saini. 2020. Bi-lstm and ensemble based bilingual sentiment analysis for a code-mixed hindi-english social media text. In *2020 IEEE 17th India Council International Conference (INDICON)*, pages 1–6.

Siddharth Yadav and Tanmoy Chakraborty. 2020. Unsupervised sentiment analysis for code-mixed data.

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.