

# IRLab@IITBHU@Dravidian-CodeMix-FIRE2020: Sentiment Analysis for Dravidian Languages in Code-Mixed Text

Supriya Chanda<sup>a</sup>, Sukomal Pal<sup>b</sup>

<sup>a</sup>Indian Institute of Technology (BHU), Varanasi, INDIA

<sup>b</sup>Indian Institute of Technology (BHU), Varanasi, INDIA

## Abstract

This paper describes the IRLab@IITBHU system for the Dravidian-CodeMix - FIRE 2020: Sentiment Analysis for Dravidian Languages pairs Tamil-English (TA-EN) and Malayalam-English (ML-EN) in Code-Mixed text. We submitted three models for sentiment analysis of code-mixed TA-EN and MA-EN datasets. Run-1 was obtained from the BERT and Logistic regression classifier, Run-2 used the DistilBERT and Logistic regression classifier, and Run-3 used the fastText model for producing the results. Run-3 outperformed Run-1 and Run-2 for both the datasets. We obtained an  $F_1$ -score of 0.58, rank 8/14 in TA-EN language pair and for ML-EN, an  $F_1$ -score of 0.63 with rank 11/15.

## Keywords

Code Mixed, Malayalam, Tamil, BERT, fastText, Sentiment Analysis,

## 1. Introduction

*Internet and digitization* enabled people express their views, sentiments, opinions through blog posts, online forums, product review websites, and different social media. Millions of people from different linguistic and cultural backgrounds use social networking sites like Facebook, Twitter, LinkedIn, and YouTube to express their emotions, opinions, and share views on different issues that matter in their lives. As a large number of Indian users can speak multiple languages proficiently (at least two: native languages like Malayalam, Tamil, Hindi, and English), an unplanned switching between languages often happens unconsciously. Even though many languages have their own scripts, social media users often use non-native scripts, usually Roman script, because of socio-linguistics reasons. This phenomenon is called code-mixing and is defined as “the embedding of linguistic units such as phrases, words and morphemes of one language into an utterance of another language” (Myers-Scotton[1]). Code-mixed data is generally observed in a place of informal communication like social media. The data can be easily extracted from social media sources using different APIs. Sentiment analysis (SA) on social media text has become an important research task in academia and industry in the past two decades. SA helps understand people’s opinion from movie/product reviews, and thus help take decision to improve customer satisfaction through advertisement and marketing.

The shared task [2, 3] here aims to identify sentiment polarity of the code-mixed data of YouTube comments in Dravidian Language pairs (Malayalam-English [4] and Tamil-English [5]) collected from social media. In the past few years, there have been multiple attempts to process code-mixed data, and

*FIRE 2020: Forum for Information Retrieval Evaluation, December 16-20, 2020, Hyderabad, India*

EMAIL: supriyachanda.rs.cse18@itbhu.ac.in (S. Chanda); spal.cse@itbhu.ac.in (S. Pal)


URL: <https://cse-iitbhu.github.io/irlab/supriya.html> (S. Chanda); <https://cse-iitbhu.github.io/irlab/spal.html> (S. Pal)

ORCID: 0000-0001-8743-9830 (S. Pal)



© 2020 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

**Table 1**  
Data Distribution

Class	Tamil - English				Malayalam - English			
	Training	Development	Test	Total	Training	Development	Test	Total
Mixed_feelings	1283	141	377	1801	289	44	70	403
Negative	1448	165	424	2037	549	51	138	738
Positive	7627	857	2075	10559	2022	224	565	2811
not-Tamil	368	29	100	497	-	-	-	-
not-malayalam	-	-	-	-	647	60	177	884
unknown_state	609	68	173	850	1344	161	398	1903
Total	11335	1260	3149	15744	4851	540	1348	6739

a shared task on sentiment analysis of code-mixed Indian languages[6] was organized in ICON 2017. However, the freely available data apart from Hindi-English and Bengali-English are still limited in Indian languages, although some other languages like English-Spanish and Chinese-English datasets are available for research.

The rest of the paper is organized as follows. Section 2 describes the dataset, pre-processing and processing techniques. In Section 3, we report our results and analysis. Finally we conclude in Section 4.

## 2. System Description

### 2.1. Datasets

The Dravidian-CodeMix shared task<sup>1</sup> organizers provided a dataset that consists of 15,744 Tamil-English and 6,739 Malayalam-English YouTube video comments. The statistics of training, development, and test data corpus collection and their class distribution are shown in Table 1. Here, each comment is annotated by six (for ML-EN) and eleven (for TA-EN) independent annotators. An inter-annotator agreement score of 0.6 with Krippendorff’s alpha is obtained for the Tamil-English dataset, and score of 0.8 with Krippendorff’s alpha for the Malayalam-English dataset. Some comment examples from the training dataset (Tamil-English) are shown in Table 2. The dataset provided suffers from general problems of social media data, particularly code-mixed data. The sentences are short with lack of well-defined grammatical structures, and many spelling mistakes.

### 2.2. Data Pre-processing

The YouTube comment dataset used in this work is already labelled into five categories: Positive, Negative, Mixed\_feelings, unknown\_state and not-Tamil or not-Malayalam. Our pre-processing of comments includes the following steps:

- Removal of extended words: number of words which have one or more contiguous repeating characters<sup>2</sup>
- Removal of exclamations and other punctuation
- Removal of non-ASCII characters, all the emoticons, symbols, numbers, special characters.

<sup>1</sup><https://dravidian-codemix.github.io/2020/index.html>

<sup>2</sup><https://github.com/SupriyaChanda/Dravidian-CodeMix-FIRE2020>

**Table 2**

Example YouTube comments from the Dravidian-CodeMix dataset for all clases

Sample comments from dataset(Tamil-English)	Category
Ena da bgm ithu yuvannnnnnnnnn rockssssss	Positive
Kola gaadula iruka... Thalaivaaaaaaaaa waiting layea veri aaguthey	Negative
Wow wow wow... Thalaivaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa..... proud to be every Indian...	Mixed_feelings
<3 thanks to shankar sir and holl team...	not-Tamil
Nenu ee movie chusanu super movie	unknown_state
Super. 1 like is equivalent to 100 likes.	

### 2.3. Word Embedding

Word embedding is arguably the most widely known technology in the recent history of NLP. It captures the semantic property of a word. We use `bert-base-uncased` and `distilbert-base-uncased` pre-trained models<sup>3</sup> to get a vector as an embedding for the sentence that we can use for classification. Apart from these two pre-trained models, we experiment with other pre-trained models like `bert-base-multilingual-uncased`, `bert-base-multilingual-cased`.

- **BERT:** Bidirectional Encoder Representations from Transformers (BERT)[7] is a technique for NLP pre-training developed by Google. BERT is pre-trained on a large corpus of unlabelled text, including the entire Wikipedia (that is 2,500 million words!) and the Book Corpus (800 million words). BERT-Base uncased has 12 layers (transformer blocks), 12 attention heads, and 110 million parameters.
- **DistilBERT:** DistilBERT[8] is a smaller version of BERT developed and open-sourced by the team at HuggingFace. It is a lighter and faster version of BERT that roughly matches its performance. DistilBERT also compares surprisingly well to BERT on downstream tasks while having about half and one third the number of parameters.
- **fastText:** fastText, developed by Facebook, combines certain concepts introduced by the NLP and ML communities, representing sentences with a bag-of-words and n-grams using subword information and sharing them across classes through a hidden representation. fastText[9] can learn vector representations of out-of-vocabulary words, which is useful for our dataset that contains Malayalam and Tamil words in Roman script.

After pre-processing our data and transforming all the comments into vector, we implement our classification algorithms and construct our training models. We used the multinomial logistic regression<sup>4</sup> with the fastText embeddings for unigrams, bigrams, and trigrams present along with different learning rates and epochs. we got the maximum  $F_1$  score on fastText text classification model with `-wordNgrams= 1`, `learning rate = 0.1` and `epochs = 10`.

## 3. Results and Analysis

We use `scikit-learn`<sup>5</sup> machine learning package for the implementation. A Macro  $F_1$  score was used to evaluate every system. Macro  $F_1$  score of the overall system was the average of  $F_1$  scores of

<sup>3</sup>[https://huggingface.co/transformers/pretrained\\_models.html](https://huggingface.co/transformers/pretrained_models.html)

<sup>4</sup><https://fasttext.cc/docs/en/supervised-tutorial.html>

<sup>5</sup><http://scikit-learn.org>

**Table 3**

Evaluation results on test data and rank list

Team Name	Tamil - English				Malayalam - English			
	Precision	Recall	$F_1$ score	Rank	Precision	Recall	$F_1$ score	Rank
SRJ	0.64	0.67	0.65	1/14	0.74	0.75	0.74	1/15
IRLab@IITBHU	0.57	0.61	0.58	8/14	0.63	0.64	0.63	11/15

**Table 4**Precision, recall,  $F_1$ -score, and support for all experiment on Tamil-English test data

	BERT			DistilBERT			FastText			support
	Precision	Recall	$F_1$ -score	Precision	Recall	$F_1$ -score	Precision	Recall	$F_1$ -score	
Mixed_feelings	0.17	0.02	0.04	0.12	0.01	0.01	0.23	0.07	0.11	377
Negative	0.40	0.12	0.18	0.42	0.08	0.13	0.34	0.28	0.31	424
Positive	0.69	0.95	0.80	0.68	0.97	0.80	0.72	0.82	0.76	2075
not-Tamil	0.66	0.57	0.61	0.67	0.53	0.59	0.27	0.61	0.37	100
unknown_state	0.19	0.04	0.07	0.29	0.02	0.04	0.23	0.13	0.17	173
macro avg	0.42	0.34	0.34	0.44	0.32	0.32	0.36	0.38	0.34	3149
weighted avg	0.56	0.66	<b>0.58</b>	0.56	0.67	0.57	0.57	0.61	<b>0.58</b>	3149
Accuracy	0.66			0.67			0.61			

**Table 5**Precision, recall,  $F_1$ -scores, and support for all experiment on Malayalam-English test data

	BERT			DistilBERT			FastText			support
	Precision	Recall	$F_1$ -score	Precision	Recall	$F_1$ -score	Precision	Recall	$F_1$ -score	
Mixed_feelings	0.31	0.14	0.20	0.50	0.09	0.15	0.39	0.27	0.32	70
Negative	0.47	0.35	0.40	0.52	0.41	0.46	0.50	0.46	0.48	138
Positive	0.63	0.75	0.68	0.64	0.77	0.70	0.73	0.70	0.72	565
not-malayalam	0.67	0.71	0.69	0.68	0.72	0.70	0.61	0.64	0.62	177
unknown_state	0.63	0.57	0.59	0.60	0.54	0.57	0.60	0.66	0.63	398
macro avg	0.54	0.50	0.51	0.59	0.51	0.52	0.56	0.55	0.55	1348
weighted avg	0.60	0.62	0.60	0.61	0.62	0.61	0.63	0.64	<b>0.63</b>	1348
Accuracy	0.62			0.62			0.64			

the individual classes. Table 3 shows our official performances as shared by the organizers vis-a-vis the best performing team. Table 4 and Table 5 report our results on Tamil-English and Malayalam-English dataset respectively. We select three models that performed well during the validation phase and submit them for final prediction of the test dataset. We observe that fastText gives better  $F_1$  scores over others which was also in the official results (shown in Table 3).

In the training data, there are some ambiguous samples. Some examples are given below.

- The Tamil-English sentence *Srk fan plz dislike tha video* is labeled as Positive, when the sentence has negative sentiment word like *dislike*.
- The Tamil-English sentence *Wow wow wow... Thalaivaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa..... proud to be every Indian... <3 thanks to shankar sir and holl team...* is labeled as mixed\_feelings, when there is many positive words like *wow, proud, thanks*.

Our models were trained on this ambiguous data, and we could not verify the correctness of labelling as we do not have knowledge of Tamil or Malayalam languages. Inconsistency of the labellings, if any, might have worsened the results on test data. Another aspect is very small sentence length. That might also be the reason why fastText unigram gave better results than n-grams, word n-grams were not able to capture the sentiment of a sentence.

**Table 6**  
Error analysis

Language pair	Sample comments from dataset	Given	Predicted
TA-EN	Just amazing thalaivaaaaaaaaa ARR sir u r that Best ( BGM )	Negative	Positive
TA-EN	Next year national award Competition DHANUSH ( asuran) Karthi( kaithi) @rya (makamuni)	Negative	positive
TA-EN	Petta paraak.rajin sir is still young	Negative	Positive

Some of the examples that were marked as incorrect predictions by our best model are shown in Table 6. The ‘Given’ column in the table denotes the expected sentiment, as available in the gold standard dataset against the ones predicted by our system. It seems that our predicted sentiment was correct.

## 4. Conclusion

This study reports performance of our system for the shared task on Sentiment Analysis for Dravidian Languages in Code-Mixed Text in Dravidian-CodeMix - FIRE 2020. We conducted a number of experiments on a real-world code-mixed YouTube comments dataset involving a few embedding techniques: fastText, BERT, and DistilBERT. We find that fastText outperforms other techniques on this task. However, there are room for improvement. In the future, we plan to use other pre-trained models with necessary fine-tuning. We also plan to explore multilingual embeddings for the languages.

## References

- [1] C. Myers-Scotton, Common and Uncommon Ground: Social and Structural Factors in Codeswitching, *Language in Society* 22 (1993) 475–503. URL: <http://www.jstor.org/stable/4168471>.
- [2] B. R. Chakravarthi, R. Priyadharshini, V. Muralidaran, S. Suryawanshi, N. Jose, J. P. Sherly, Elizabeth McCrae, Overview of the track on Sentiment Analysis for Davidian Languages in Code-Mixed Text, in: *Proceedings of the 12th Forum for Information Retrieval Evaluation, FIRE '20, 2020*.
- [3] B. R. Chakravarthi, R. Priyadharshini, V. Muralidaran, S. Suryawanshi, N. Jose, J. P. Sherly, Elizabeth McCrae, Overview of the track on Sentiment Analysis for Davidian 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, 2020.
- [4] B. R. Chakravarthi, N. Jose, S. Suryawanshi, E. Sherly, J. P. McCrae, 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), European Language Resources association, Marseille, France, 2020*, pp. 177–184. URL: <https://www.aclweb.org/anthology/2020.sltu-1.25>.
- [5] B. R. Chakravarthi, V. Muralidaran, R. Priyadharshini, J. P. McCrae, 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), European Language Resources association, Marseille, France, 2020*, pp. 202–210. URL: <https://www.aclweb.org/anthology/2020.sltu-1.28>.
- [6] B. G. Patra, D. Das, A. Das, Sentiment Analysis of Code-Mixed Indian Languages: An Overview of SAIL Code-Mixed Shared Task @ICON-2017, 2018. *arXiv:1803.06745*.

- [7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Proceedings of the 2019 Conference of the North (2019). doi:10.18653/v1/n19-1423.
- [8] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, 2019. arXiv:1910.01108.
- [9] T. Mikolov, E. Grave, P. Bojanowski, C. Puhersch, A. Joulin, Advances in Pre-Training Distributed Word Representations, in: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018), 2018.