Overview of the track on Sentiment Analysis for Dravidian Languages in Code-Mixed Text

Bharathi Raja Chakravarthi
Insight SFI Research Centre for Data Analytics,
National University of Ireland Galway
Galway, Ireland
bharathi.raja@insight-centre.org

Ruba Priyadharshini
ULTRA Arts and Science College,
Madurai, India
rubapriyadharshini.a@gmail.com

Vigneshwaran Muralidaran
School of Computer Science and Informatics,
Cardiff University,
Cardiff, United Kingdom
vigneshwar.18@gmail.com

Shardul Suryawanshi
Insight SFI Research Centre for Data Analytics,
National University of Ireland Galway
Galway, Ireland
shardul.suryawanshi@insight-centre.org

Navya Jose
Indian Institute of Information Technology and Management-Kerala
Thiruvananthapuram, India
navya.mi3@iiitmk.ac.in

Elizabeth Sherly
Indian Institute of Information Technology and Management-Kerala
Thiruvananthapuram, India
sherly@iiitmk.ac.in

John P. McCrae
Insight SFI Research Centre for Data Analytics,
National University of Ireland Galway
Galway, Ireland
john.mccrae@insight-centre.org

ABSTRACT

Sentiment analysis of Dravidian languages has received attention in recent years. However, most social media text is code-mixed and there is no research available on sentiment analysis of code-mixed Dravidian languages. The Dravidian-CodeMix-FIRE 2020, a track on Sentiment Analysis for Dravidian Languages in Code-Mixed Text, focused on creating a platform for researchers to come together and investigate the problem. There were two languages for this track: (i) Tamil, and (ii) Malayalam. The participants were given a dataset of YouTube comments and the goal of the shared task submissions was to recognise the sentiment of each comment by classifying them into positive, negative, neutral, mixed-feeling classes or by recognising whether the comment is not in the intended language. The performance of the systems was evaluated by weighted-F1 score.

CCS CONCEPTS

• **Information systems** → **Clustering and classification**; • **Computing methodologies** → *Machine learning algorithms.*

FIRE '20:, December 16-20, 2020, Hyderabad, India

ACM ISBN 978-1-4503-8978-5/20/12…$15.00
https://doi.org/10.1145/3441501.3441515

KEYWORDS

Sentiment analysis; Dravidian languages; Tamil; Malayalam; code-mixing; text classification; deep learning

ACM Reference Format:
Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Shardul Suryawanshi, Navya Jose, Elizabeth Sherly, and John P. McCrae. 2020. Overview of the track on Sentiment Analysis for Dravidian Languages in Code-Mixed Text. In *Forum for Information Retrieval Evaluation (FIRE '20), December 16–20, 2020, Hyderabad, India.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3441501.3441515

1 INTRODUCTION

Sentiment analysis is the task of identifying subjective opinions or responses about a given topic. It has been an active area of research in the past two decades in both academia and industry. There is an increasing demand for sentiment analysis on social media texts which are largely code-mixed. Code-mixing is a prevalent phenomenon in a multilingual community where the words, morphemes and phrases from two or more languages are mixed in speech or writing. Code-mixed texts are often written in non-native scripts particularly on social media. Systems trained on monolingual data fail on code-mixed data due to the complexity of code-switching at different linguistic levels in the text. This shared task presents a new gold standard corpus for sentiment analysis of code-mixed text in Dravidian languages (Tamil-English and Malayalam-English).

Tamil is one of the Dravidian languages spoken by Tamil people in India, Sri Lanka and by the Tamil diaspora around the world, with official recognition in India, Sri Lanka and Singapore. Malayalam is another Dravidian language spoken in the southern region

of India with official recognition in the Indian state of Kerala and the Union Territories of Lakshadweep and Puducherry. There are nearly 75 million Tamil speakers[1] and 45 million Malayalam speakers[2] in India and other countries. Tamil and Malayalam are highly agglutinative languages.

Tamil script evolved from the Tamili script[3], Vatteluttu alphabet, and Chola-Pallava script. The modern Tamil script descended from the Chola-Pallava script. It has 12 vowels, 18 consonants, and 1 āytam (voiceless velar fricative). Minority languages such as Saurashtra, Badaga, Irula, and Paniya are also written in the Tamil script. The Malayalam script is the Vatteluttu alphabet extended with symbols from the Grantha script. Both Tamil and Malayalam scripts are alpha-syllabic, belonging to a family of the abugida writing systems that are partially alphabetic and partially syllable-based. However, social media users often adopt Roman script for typing because it is easy to input. Hence, the majority of the data available in social media for these under-resourced languages are code-mixed.

The goal of this task is to identify the sentiment polarity of the code-mixed dataset of comments/posts in Dravidian Languages (Malayalam-English and Tamil-English) collected from social media. The comment/post may contain more than one sentence but the average sentence length of the corpora is 1. Each comment/post is annotated with sentiment polarity at the comment/post level. This dataset also has class imbalance problems which depicts the real-world scenarios. The dataset provided for training and development contains 11,335 and 1,260 sentences for Tamil, 4,851 and 541 sentences for Malayalam. More details about the annotation of the dataset can be found in [6] and [5].

Our shared task aimed to encourage research that will reveal how sentiment is expressed in code-mixed scenarios on Dravidian social media text. The participants were provided with development, training and test datasets.

2 TASK DESCRIPTION

This is a message-level polarity classification task. Given a YouTube comment, the goal of the systems submitted to the shared task was to classify the comment into positive, negative, neutral, mixed feeling classes or recognise if the comment is not in the intended languages. The dataset contains all the three types of code-mixed sentences - Inter-Sentential switch, Intra-Sentential switch and Tag switching. All comments were written in Roman script with either the grammar of native language with English lexicon or English grammar with native lexicon. The following examples from Tamil dataset illustrates the point.

- **Intha padam vantha piragu yellarum Thala ya kondaduvanga.**- *After the movie release, everybody will celebrate the hero.* Tamil words written in Roman script with no English switch.
- **Trailer late ah parthavanga like podunga.** - *Those who watched the trailer late, please like it.* Tag switching with English words.

[1]Between 2011- 2015 Source https://en.wikipedia.org/wiki/Tamil_language
[2]Between 2011- 2019 Source https://en.wikipedia.org/wiki/Malayalam
[3]This was also called Damili or Tamil-Brahmi script

No.	TeamName	Precision	Recall	F1-Score	Rank
01	SRJ[24]	0.64	0.67	0.65	1
02	DT	0.62	0.68	0.64	2
03	YUN111 [26]	0.63	0.67	0.64	2
04	codemixed_umsnh [17]	0.61	0.68	0.63	3
05	LucasHub [9]	0.61	0.68	0.63	3
06	YNU [18]	0.61	0.67	0.63	3
07	MUCS[3]	0.60	0.66	0.62	4
08	PITS [12]	0.62	0.69	0.62	4
09	datamafia	0.60	0.65	0.62	4
10	gauravarora [2]	0.65	0.69	0.62	4
11	jiaming gao	0.61	0.64	0.62	4
12	Theedhum [15]	0.64	0.67	0.62	4
13	HRS-TECHIE [25]	0.59	0.65	0.61	5
14	NITP-AI-NLP [14]	0.59	0.64	0.61	5
15	SSNCSE_NLP[1]	0.60	0.65	0.61	5
16	zyy1510 [27]	0.59	0.66	0.61	5
17	bits2020 [23]	0.62	0.66	0.61	5
18	SSN_NLP_MLRG [11]	0.60	0.68	0.60	6
19	Siva [21]	0.59	0.63	0.60	6
20	IRLab@IITV	0.59	0.61	0.59	7
21	CMSAOne [8]	0.58	0.67	0.58	8
22	CodeMixedNLP	0.56	0.68	0.58	8
23	ComMA	0.58	0.66	0.58	8
24	IRLab@IITBHU [7]	0.57	0.61	0.58	8
25	JUNLP[16]	0.59	0.66	0.58	8
26	TADS [22]	0.57	0.67	0.56	9
27	Parameswari [13]	0.55	0.66	0.55	10
28	Judith Jeyafreeda [10]	0.57	0.66	0.54	11
29	Thirumurugan R	0.67	0.66	0.54	11
30	DLRG	0.62	0.49	0.53	12
31	NUIG_Shubhanker [4]	0.52	0.52	0.51	13
32	Anbukkarasi [20]	0.33	0.07	0.10	14

Table 1: Rank list based weighted average F1-score along with other evaluation matrics (Precision and Recall) for Tamil Subtask

- **Omg .. use head phones. Enna bgm da saami ..** - *OMG! Use your headphones. Good Lord, What a background score!* Inter-sentential switch
- **I think sivakarthickku hero getup set aagala.** - *I think the hero role does not suit Sivakarthick.* Intra-sentential switch between clauses.

3 EVALUATION

The distribution of the sentiment classes are imbalanced in both the datasets. In Malayalam-English code-mixed dataset, we have a class imbalance with the majority of comments belonging to positive (2,811) and neutral (1,903) classes. Similarly, Tamil-English code-mixed dataset has the class imbalance with Positive (10,559), Negative (2,037) and Mixed feelings (1,801) being majority classes. This imbalance demands to be addressed. Hence, we chose a weighted average F1-score to rank the system submission. The weighted average F1-score is calculated by averaging the support-weighted mean per-class F1 scores (i.e. weights on class distribution). This takes into account the varying degrees of importance of each class in the

No.	TeamName	Precision	Recall	F1-Score	Rank
01	SRJ [24]	0.74	0.75	0.74	1
02	datamafia	0.74	0.74	0.74	1
03	YNU [18]	0.74	0.74	0.74	1
04	YUN111 [26]	0.73	0.73	0.73	2
05	LucasHub [9]	0.73	0.73	0.73	2
06	jiaming gao	0.73	0.73	0.73	2
07	DT	0.72	0.72	0.72	3
08	CIA_NITT [19]	0.71	0.71	0.71	4
09	PITS [12]	0.70	0.71	0.71	4
10	SSNCSE_NLP [1]	0.71	0.71	0.71	4
11	NITP-AI-NLP [14]	0.69	0.69	0.69	5
12	gauravarora [2]	0.69	0.70	0.69	5
13	MUCS [3]	0.68	0.68	0.68	6
14	codemixed_umsnh [17]	0.68	0.69	0.68	6
15	TADS [22]	0.68	0.68	0.67	7
16	CMSAOne [8]	0.66	0.67	0.66	8
17	Siva [21]	0.67	0.67	0.66	8
18	Theedhum Nandrum [15]	0.67	0.66	0.65	9
19	ComMA	0.64	0.66	0.64	10
20	zyy1510 [27]	0.64	0.64	0.64	10
21	IRLab@IITBHU [7]	0.63	0.64	0.63	11
22	CodeMixedNLP	0.59	0.62	0.60	12
23	IRLab@IITV	0.68	0.60	0.60	12
24	SSN_NLP_MLRG [11]	0.61	0.61	0.60	12
25	bits2020 [23]	0.67	0.59	0.60	12
26	Judith Jeyafreeda [10]	0.68	0.62	0.58	13
27	Parameswari [13]	0.53	0.51	0.48	14
28	NUIG_Shubhanker [4]	0.48	0.50	0.46	15

Table 2: Rank list based weighted average F1-score along with other evaluation matrics (Precision and Recall) for Malayalam Subtask

dataset. We used a classification report tool from Scikit learn[4].

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{F-Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

4 RESULTS

Overall, 119 participants registered for this track. 32 teams submitted final results for Tamil and 28 teams submitted results for Malayalam. Table 1 and Table 2 shows the rank list of Tamil and Malayalam task respectively.

The runs are sorted in decreasing order of the weighted F1-scores. The best performing runs achieved weighted F1-score of 0.65 and 0.74 for Tamil and Malayalam respectively. These scores are relatively low F1-scores compared to monolingual sentiment analysis results in high-resourced languages such as English. This reflects the complexity of code-mixing and the class imbalance problem observed in the real-world setting. The top team "SRJ" used XLM-Roberta and CNN to propose a new model to extract semantic information.

[4]https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html

5 CONCLUSION

This paper overviews the first shared task on sentiment analysis in code-mixed Dravidian text from social media that aims at classifying YouTube comments. A hundred and nineteen teams participated in the task, and a total of 32 teams for Tamil and 28 teams Malayalam submitted the results. Systems have been trained on the unbalanced dataset. The methods proposed by participants ranged from traditional machine learning models with features based approaches to using state-of-the-art embedding methods in deep learning models.

ACKNOWLEDGMENTS

This publication has been supported in part by a research grant from Science Foundation Ireland (SFI) (SFI/12/RC/2289_P2 Insight_2), co-funded by the European Regional Development Fund as well as by the EU H2020 programme under grant agreements 825182 (Prêt-à-LLOD), and Irish Research Council grant IRCLA/2017/129 (CARDAMOM-Comparative Deep Models of Language for Minority and Historical Languages).

REFERENCES

[1] Nitin Nikamanth Appiah Balaji, Bharathi B, and Bhuvana J. 2020. SSNCSE_NLP@Dravidian-CodeMix-FIRE2020: Sentiment Analysis for Dravidian Languages in Code-Mixed Text. In *FIRE (Working Notes)*. CEUR, Hyderabad, India.

[2] Gaurav Arora. 2020. Gauravarora@HASOC-Dravidian-CodeMix- FIRE2020: Pretraining ULMFiT on Synthetically Generated Code-Mixed Data for Hate Speech Detection. In *FIRE (Working Notes)*. CEUR, Hyderabad, India.

[3] Fazlourrahman Balouchzahi and H L Shashirekha. 2020. MUCS@Dravidian-CodeMix-FIRE2020:SACO-SentimentsAnalysis for CodeMix Text. In *FIRE (Working Notes)*. CEUR, Hyderabad, India.

[4] Shubhanker Banerjee, Arun Jaypal, and Sajeetha Thavareesan. 2020. NUIG-Shubhanker@Dravidian-CodeMix- FIRE2020:Sentiment Analysis of Code-Mixed Dravidian text using XLNet. In *FIRE (Working Notes)*. CEUR, Hyderabad, India.

[5] Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John Philip McCrae. 2020. 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, 177–184. https://www.aclweb.org/anthology/2020.sltu-1.25

[6] Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020. 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, 202–210. https://www.aclweb.org/anthology/2020.sltu-1.28

[7] Supriya Chanda and Sukomal Pal. 2020. IRLab@IITBHU@Dravidian-CodeMix-FIRE2020: Sentiment Analysis for Dravidian Languages in Code-Mixed Text. In *FIRE (Working Notes)*. CEUR, Hyderabad, India.

[8] Suman Dowlagar and Radhika Mamidi. 2020. CMSAOne@Dravidian-CodeMix-FIRE2020: A Meta Embedding and Transformer model for Code-Mixed Sentiment Analysis on Social Media Text. In *FIRE (Working Notes)*. CEUR, Hyderabad, India.

[9] Bo Huang and Yang Bai. 2020. LucasHub@Dravidian-CodeMix-FIRE2020: Sentiment Analysis on Multilingual Code Mixing Text with M-BERT and XLM-RoBERTa. In *FIRE (Working Notes)*. CEUR, Hyderabad, India.

[10] Judith Jeyafreeda. 2020. JudithJeyafreeda@Dravidian-CodeMix-FIRE2020:Sentiment Analysis of YouTube Comments for DravidianLanguages. In *FIRE (Working Notes)*. CEUR, Hyderabad, India.

[11] A. Kalaivani and D. Thenmozhi. 2020. SSN_NLP_MLRG@Dravidian-CodeMix-FIRE2020: Sentiment Code-Mixed Text Classification in Tamil and Malayalam using ULMFiT. In *FIRE (Working Notes)*. CEUR, Hyderabad, India.

[12] Nikita Kanwar, Megha Agarwal, and Rajesh Kumar Mundotiya. 2020. PITS@Dravidian-CodeMix-FIRE2020: Traditional Approach to Noisy Code-Mixed Sentiment Analysis. In *FIRE (Working Notes)*. CEUR, Hyderabad, India.

[13] Parameswari Krishnamurthy, Faith Varghese, and Nagaraju Vuppala. 2020. Parameswari_faith_nagaraju@Dravidian-CodeMix-FIRE2020: A machine-learning approach using n-grams in sentiment analysis for code-mixed texts:A case study in Tamil and Malayalam. In *FIRE (Working Notes)*. CEUR, Hyderabad, India.

[14] Abhinav Kumar, Sunil Saumya, and Jyoti Prakash Singh. 2020. NITP-AI-NLP@Dravidian-CodeMix-FIRE2020: A Hybrid CNN and Bi-LSTM Network for Sentiment Analysis of Dravidian Code-Mixed Social Media Posts. In *FIRE (Working Notes)*. CEUR, Hyderabad, India.

[15] Balasundararaman L and Sanjeeth Kumar Ravindranath. 2020. Theedum Nandrum@Dravidian-CodeMix-FIRE2020: A sentiment polarity detection system for YouTube comments with code switching between Tamil, Malayalam and English. In *FIRE (Working Notes)*. CEUR, Hyderabad, India.

[16] Sainik Kumar Mahata, Dipankar Das, and Sivaji Bandyopadhyay. 2020. JUNLP@Dravidian-CodeMix-FIRE2020: Sentiment Classification of Code-Mixed Tweets using Bi-Directional RNN and Language Tags. In *FIRE (Working Notes)*. CEUR, Hyderabad, India.

[17] Jose Ortiz-Bejar, Jesus Ortiz-Bejar, Jaime Cerda-Jacobo, Mario Graff, and Eric S. Tellez. 2020. UMSNH-INFOTEC@Dravidian-CodeMix-FIRE2020: An ensemble approach based on a multiple text representations. In *FIRE (Working Notes)*. CEUR, Hyderabad, India.

[18] Xiaozhi Ou and Hongling Li. 2020. YNU@Dravidian-CodeMix-FIRE2020: XLM-RoBERTa for Multi-language Sentiment Analysis. In *FIRE (Working Notes)*. CEUR, Hyderabad, India.

[19] Yandrapati Prakash Babu, Rajagopal Eswari, and K Nimmi. 2020. CIA_NITT@Dravidian-CodeMix-FIRE2020: Malayalam-English Code Mixed Sentiment Analysis Using Sentence BERT And Sentiment Features. In *FIRE (Working Notes)*. CEUR, Hyderabad, India.

[20] Anbukkarasi S and Varadhaganapathy S. 2020. SA_SVG@Dravidian-CodeMix-FIRE2020: Deep Learning Based Sentiment Analysis in Code-mixed Tamil-English Text. In *FIRE (Working Notes)*. CEUR, Hyderabad, India.

[21] Siva Sai and Yashvardhan Sharma. 2020. Siva@HASOC-Dravidian-CodeMix-FIRE-2020: Multilingual Offensive Speech Detection in Code-mixed and Romanized Text. In *FIRE (Working Notes)*. CEUR, Hyderabad, India.

[22] Deepesh Sharma. 2020. TADS@Dravidian-CodeMix-FIRE2020: SentimentAnalysisonCodeMixDravidianLanguage. In *FIRE (Working Notes)*. CEUR, Hyderabad, India.

[23] Yashvardhan Sharma and Asrita Venkata Mandalam. 2020. bits2020@Dravidian-CodeMix-FIRE2020: Sub-Word Level Sentiment Analysis of Dravidian Code Mixed Data. In *FIRE (Working Notes)*. CEUR, Hyderabad, India.

[24] Ruijie Sun and Xiaobin Zhou. 2020. SRJ @ Dravidian-CodeMix-FIRE2020:Automatic Classification and Identification Sentiment in Code-mixed Text. In *FIRE (Working Notes)*. CEUR, Hyderabad, India.

[25] Sridhar Swaminathan, Hari Krishnan Ganesan, and Radhakrishnan Pandiyarajan. 2020. HRS-TECHIE@Dravidian-CodeMix-FIRE2020 Social Media Sentiment Analysis for Dravidian Languages using Machine Learning, Deep Learning and Ensemble Approaches. In *FIRE (Working Notes)*. CEUR, Hyderabad, India.

[26] Yueying Zhu and Kunjie Dong. 2020. YUN111@Dravidian-CodeMix-FIRE2020: Sentiment Analysis of Dravidian Code Mixed Text. In *FIRE (Working Notes)*. CEUR, Hyderabad, India.

[27] Yueying Zhu and Xiaobing Zhou. 2020. Zyy1510@HASOC-Dravidian-CodeMix-FIRE2020: An Ensemble Model for Offensive Language Identification. In *FIRE (Working Notes)*. CEUR, Hyderabad, India.