Findings of the Sentiment Analysis of Dravidian Languages in Code-Mixed Text

Bharathi Raja Chakravarthi[a], Ruba Priyadharshini[b], Sajeetha Thavareesan[c], Dhivya Chinnappa[d], Durairaj Thenmozhi[e], Elizabeth Sherly[f], John P. McCrae[a], Adeep Hande[h], Rahul Ponnusamy[f], Shubhanker Banerjee[j] and Charangan Vasantharajan[k]

[a] *Insight SFI Research Centre for Data Analytics, Data Science Institute, National University of Ireland Galway*
[b] *ULTRA Arts and Science College, Madurai, Tamil Nadu, India*
[c] *Eastern University, Sri Lanka*
[d] *Thomson Reuters, USA*
[e] *Sri Sivasubramaniya Nadar College of Engineering, Tamil Nadu, India*
[f] *Indian Institute of Information Technology and Management-Kerala, India*
[h] *Indian Institute of Information Technology Tiruchirappalli*
[j] *ADAPT Centre, Data Science Institute, National University Of Ireland Galway*
[k] *Department of Computer Science and Engineering, University of Moratuwa, Sri Lanka*

Abstract

We present the results of the Dravidian-CodeMix shared task[1] held at FIRE 2021, a track on sentiment analysis for Dravidian Languages in Code-Mixed Text. We describe the task, its organization, and the submitted systems. This shared task is the continuation of last year's Dravidian-CodeMix shared task[2] held at FIRE 2020. This year's tasks included code-mixing at the intra-token and inter-token levels. Additionally, apart from Tamil and Malayalam, Kannada was also introduced. We received 22 systems for Tamil-English, 15 systems for Malayalam-English, and 15 for Kannada-English. The top system for Tamil-English, Malayalam-English and Kannada-English scored weighted average F1-score of 0.711, 0.804, and 0.630, respectively. In summary, the quality and quantity of the submission show that there is great interest in Dravidian languages in code-mixed setting and state of the art in this domain still needs more improvement.

Keywords

Sentiment analysis, Dravidian languages, Tamil, Malayalam, Kannada, Code-mixing,

[1] https://dravidian-codemix.github.io/2021/index.html/

[2] https://dravidian-codemix.github.io/2020/index.html/

FIRE 2021: Forum for Information Retrival Evaluation, December 13-17, 2021, India

bharathi.raja@insight-centre.org (B.R. Chakravarthi); rubapriyadharshini.a@gmail.com (R. Priyadharshini); sajeethas@esn.ac.lk (S. Thavareesan); dhivya.infant@gmail.com (D. Chinnappa); theni_d@ssn.edu.in (D. Thenmozhi); sherly@iiitmk.ac.in (E. Sherly); john.mccrae@insight-centre.org (J.P. McCrae); adeeph18c@iiitt.ac.in (A. Hande); rahul.mi20@iiitmk.ac.in (R. Ponnusamy); S.Banerjee3@nuigalway.ie (S. Banerjee); charangan.18@cse.mrt.ac.lk (C. Vasantharajan)

0000-0002-4575-7934 (B.R. Chakravarthi); 0000-0003-2323-1701 (R. Priyadharshini); 0000-0002-6252-5393 (S. Thavareesan); 0000-0002-8318-3337 (D. Chinnappa); 0000-0001-6508-950X (E. Sherly); 0000-0002-7227-1331 (J.P. McCrae); 0000-0002-2003-4836 (A. Hande); 0000-0001-8023-7742 (R. Ponnusamy); 0000-0002-3969-5183 (S. Banerjee); 0000-0001-7874-3881 (C. Vasantharajan)

CEUR Workshop Proceedings (CEUR-WS.org)

1. Introduction

Sentiment analysis is a text mining task that finds and extracts personal information from the source material, allowing a company/researcher to understand better the social sentiment of its brand, product, or service while monitoring online conversations [1]. In our case, we used the comments from the movie trailer, so it is about finding the viewers sentiment of the movie. The constantly increasing number of social media and user-generated comments raises the importance of finding sentiments in local languages as making these predictions is essential for local businesses. For this study, we created data in Dravidian languages, namely Tamil (ISO 639-3:tam), Malayalam (ISO 639-3:mal), and Kannada (ISO 639-3:kan). Tamil is the official language of Tamil Nadu, the Indian Union, Sri Lanka, Malaysia and is spoken in many places in South Asian countries. Malayalam and Kannada have official status in the Indian Union government and the state of Kerala and Karnataka, respectively [2, 3].

The Tamil script, the Vatteluttu alphabet, and the Chola-Pallava script all came together to form the Tamil script. The Tamil script dates back to 600 BCE, found at various archaeological sites in Tamil Nadu, Sri Lanka, Egypt, Thailand, Vietnam, Cambodia and Indonesia. The Chola-Pallava script is the ancestor of the present Tamil script. Thani Tamil Iyakkam (Pure or Independent Tamil Movement) is a Tamil linguistic purity movement that tried to avoid borrowing terms from Sanskrit, English, and other languages in 1916. Maraimalai Adigal[1], Paventhar Bharathidasan[2], Devaneya Pavanar[3], and Pavalareru Perunchitthiranaar[4] started the movement, which was spread through the Thenmozhi literary journal created by Pavalareru P. The natural continuation of this endeavour was to purge Tamil of Sanskrit influence including negative societal attitudes such discrimination based on colour and birth, central discrimination being education only for particular people which denies education for the main population that they felt kept Tamils in a condition of economic, cultural, and political slavery, which they believed made Tamil and other Dravidian states vulnerable to external political dominance.

Despite the vast amounts of primary and secondary speakers, Kannada is a low resource language concerning language technology. It primarily speaks by people in Karnataka, India, and is also the state's official language. Catanese, the Kannada script, is an alpha-syllabary of the scripts of the Brahmic family evolving into the Kadamba script and used to write other under-resourced languages like Tulu, Konkani, and Sankethi. The Kannada script has 13 vowels (14 if the obsolete vowel includes), 34 consonants, and 2 yogavahakas (semiconsonants: part-vowel, part-consonant). Malayalam used Vatteluttu script and Pallava-Grantha script. However, by 2020 language mixing of foreign languages in the Dravidian language has become very frequent. English is seen as a predominant language economically and culturally by Dravidian languages speakers, so social media users often adopted Roman script and mixed native script.

The Dravidian-CodeMix task was introduced in 2020 and aimed to explore the sentiment analysis of code mixed comments in Dravidian languages. In 2020, we released the data for Tamil and Malayalam in Roman script. The dataset included 15,000 instances for Tamil and 6,000 instances for Malayalam. In 2021, apart from Tamil and Malayalam, we introduce a Sen-

[1] https://en.wikipedia.org/wiki/Maraimalai_Adigal

[2] https://en.wikipedia.org/wiki/Bharathidasan

[3] https://en.wikipedia.org/wiki/Devaneya_Pavanar

[4] https://en.wikipedia.org/wiki/Perunchithiranar_(Tamil_nationalist)

timent Analysis dataset for Kannada Thus, in 2021 we will include three languages Tamil, Malayalam, and Kannada. Our dataset contains all kinds of code-mixing, ranging from simple script mixing to mixing at the morphological level. The challenge is to determine the polarity of sentiment in a code-mixed dataset of comments or posts in Tamil-English, Malayalam-English, and Kannada-English [4]. More details about the annotation of the dataset can be found in [3, 5, 6]

Our shared task seeks to promote a study on how sentiment communicates on Dravidian social media language in a code-mixed setting and aim for better social media analysis [7, 8]. We presented the training, development and test set to the participants. This paper presents an overview of the task description, dataset, description of the participating systems, analysis, and provide insights from the shared task.

2. Task Description

This task aims at the classification of sentence-level polarities. The main objective of the proposed systems is to classify the polarity of a given YouTube comment into mixed feelings, negative and positive or identify if the given comment does not belong to one of the following languages of this shared task: Tamil-English, Malayalam-English, and Kannada-English. The comments provided to the participants were written in a mixture of Latin script, native script, and both Latin script and native script. Some of the comments followed the grammar of one of the Dravidian languages: Tamil or Malayalam, or Kannada, but are written using the English lexicon. Other comments followed the lexicon of the Dravidian languages and were written using English grammar. The participants were provided with the development, training and test dataset. This is a message-level polarity classification task. Participants' systems have to classify a Youtube comment into positive, negative, neutral, mixed emotions, or not in the intended languages.

The following examples are from the Tamil dataset illustrate dataset code-mixing.

- **Epo pa varudhu indhe padam** - *When will this movie come out?* Tamil words written in Roman script with no English switch.
- **Yaru viswasam teaser ku marana waiting like pannunga** - *Who is waiting for Viswarm teaser, please like* Tag switching with English words.
- **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.

The following examples are from the Malayalam dataset.

- **Orupaadu nalukalku shesham aanu ithupoloru padam eranghunnathu.** - *A movie like this is coming out after a long time.* Malayalam words written in Roman script with no English switch.

- **Malayalam industry ku thriller kshamam illannu kaanichu kodukku anghotu.** - *Show that there is no shortage for thriller movies in Malayalam film industry.* Tag switching with English words.

- **Manju chechiyude athyugran performancenayi kaathirikunnu. The Lady superstar of Malayalam industry.** - *Waiting for the awesome performance of Manju sister. The Lady superstar of Malayalam film industry.* Inter-sentential switch

- **Next movie ready for nammude swantham dhanush.** - *Next movie ready for our dear Dhanush.*

- **Orupaadu nalukalku shesham aanu ithupoloru padam eranghunnathu.** - *A movie like this is coming out after a long time.* Malayalam words are written in Roman script with no English switch.

The following examples are from the Kannada dataset.

- **Yaru tension agbede yakandre dislike madiravru mindrika kadeyavru** – *No one needs to worry as the people who disliked this are fans of Mandrika.* Intra-sentential switch between clauses

- **Gottilla Rakshit Shettru natana nanu fida. Boss waiting for movie Charitre bareyo ella lakshana ide. All the best for you bright future** –*Don't know why, I am obsessed with Rakshit Shetty's acting. waiting for your movie, expecting it to be a blockbuster. All the best for your bright future.* Inter-sentential and intra-sentential mix. (Kannada written in both Latin and Kannada script)

- **Nanage ansutte ee video vanna rashmika mandanna fans dislike madirbahudu** –*I feel that this video has been disliked by the fans of Rashmika Mandana.*Intra-sentential switch between clauses.Code-switching at morphological level: (written in both Kannada and Latin script)

- **My favorite song in 2019 is Taaja Samachara. Sahitya priyare omme ee haadu kelidre kelthane irabeku ansutte. Everybody watch this.** –*My favourite song in 2019 is Taaja Sanachara. Literature admirers, please listen to the song once; you would want to listen to it over and over again. Everybody watch this.* Inter-sentential code-mixing: Mix of English and Kannada (Kannada written in Kannada script itself)

The data was annotated for sentiments according to the following schema.

- **Positive state:** The text contains an explicit or implicit indication that the speaker is in an optimistic mood, i.e., joyful, admiring, relaxed, and forgiving.

- **Negative state:** The text contains an explicit or implicit indication that the speaker is in an unfavourable condition, i.e., depressed, angry, nervous, or aggressive.

- **Mixed feelings:** The text contains an explicit or implicit indication indicating that the speaker is experiencing both good and negative emotions. Comparing two films

Language	Tamil	Malayalam	Kannada
Number of words	513,311	224,207	65,002
Vocabulary size	94,928	57,566	20,665
Number of comments	44,020	19,616	7,671
Number of sentences	52,750	24,014	8,472
Average number of words per sentence	11	11	8
Average number of sentences per comment	1	1	1

Table 1
Corpus statistics of the dataset

Class	Tamil	Malayalam	Kannada
Negative	5,228 (11.87 %)	2,600 (13.25 %)	1,484 (19.34 %)
Not in intended language	2,087 (4.74 %)	1,445 (736 %)	1,136 (14.80 %)
Neutral state	6,904 (15.68 %)	6,502 (33.14 %)	842 (10.97 %)
Mixed feelings	4,928 (1119 %)	1,162 (5.92 %)	691 (9.00 %)
Positive	24,873 (56.50 %)	7,907 (40.30 %)	3,518 (45.86 %)
Total	44,020	19,616	7,671

Table 2
Class-wise Dataset Distribution

	Tamil	Malayalam	Kannada
Training	35,220	15,694	6,136
Development	4,398	1,960	767
Test	4,402	1,962	768
Total	44,020	19,616	7,671

Table 3
Train-Development-Test Data Distribution with 90%-5%-5% train-dev-test split

- **Neutral state:** There is no explicit or implicit indication of the speaker's emotional state: examples include requests for likes or subscriptions, as well as inquiries about the release date or movie dialogue. This is a state that can be termed neutral.
- **Not in intended language:** For Kannada, if the sentence does not contain Kannada, then it is not Kannada.

The annotators were provided with Tamil, Kannada, and Malayalam translations of the above to facilitate better understanding. A minimum of three annotators annotated each sentence. Dataset corpus statistics are given in Table 1, Table 2, and Table 3.

No.	TeamName	Precision	Recall	F1-Score	Rank
1	SSNCSE_NLP [9]	0.64	0.66	0.63	1
2	MUCIC [10]	0.62	0.66	0.63	2
3	CIA_NITT [11]	0.63	0.64	0.63	3
4	SOA-NLP [12]	0.64	0.65	0.62	4
5	IIITT-Karthik Puranik [13]	0.62	0.63	0.62	5
6	Dynamic Duo [14]	0.67	0.65	0.62	6
7	KBCNMUJAL [15]	0.62	0.64	0.62	7
8	IIITT-Pawan [16]	0.61	0.61	0.61	8
9	AI ML	0.62	0.60	0.61	9
10	SSN_NLP_MLRG [17]	0.60	0.59	0.60	10
11	Amrita_CEN [18]	0.60	0.58	0.57	11
12	IIIT_DWD	0.57	0.54	0.55	12
13	LogicDojo	0.43	0.56	0.48	13
14	MUM [19]	0.41	0.49	0.37	14
15	IRLab@IITBHU [20]	0.29	0.35	0.32	15

Table 4
Rank list based on weighted average F1-score along with other evaluation metrics (Precision and Recall) for the Kannada track

3. Methodology

We received 54 submissions for the task, out of which 17 were for the Malayalam track, 22 were for the Tamil track, and 15 were for the Kannada track. The rank lists for the Kannada track, Tamil track, and the Malayalam track are shown in Tables 4, 5 and 6 respectively. Below we briefly describe the systems proposed by the top 3 teams in both tracks.

- MUCIC [28]: The authors extracted the character level and syllable level features from the text, which were then used to create the TF-IDF feature vectors. The authors have documented three models, namely: a logistic regression model, an LSTM classifier, and a multilayer perceptron classifier, to classify the messages. The TF-IDF feature vectors are fed to these models, which in turn are trained on the classification task.

- CIA_NITT [29]: The authors proposed a system that uses a pretrained XLM-RoBERTa for sequence classification. They tokenize the input text using the SentencePiece tokenizer, which is then fed as embeddings to be fine-tuned for the XLM-RoBERTa model. .

- ZYBank-AI [30]: The authors based their experiments on the XLM-RoBERTa as well. To improve the results, they have added self-attention to the 12 hidden layers of the XLMRoBERTA. Furthermore, they propose a two-stage pipeline for the task at hand. In the first stage, the model is trained on data from Dravidian-CodeMix-FIRE 2020. In the second stage, the pre-trained model is fine-tuned on the Dravidian-CodeMix-FIRE 2021 and evaluated on test data.

- IIITT-Pawan [31]: The authors proposed an ensemble of several fine-tuned language models for sequence classification: BERT, MuRIL, XLM-RoBERTa, DistilBERT. Each of

No.	TeamName	Precision	Recall	F1-Score	Rank
1	CIA_NITT [11]	0.71	0.71	0.71	1
2	ZYBank-AI [21]	0.68	0.68	0.68	2
3	IIITT-Pawan [16]	0.62	0.65	0.63	3
4	IIITT-Karthik Puranik [13]	0.62	0.64	0.62	4
5	MUCIC [10]	0.61	0.64	0.62	5
6	SOA_NLP [12]	0.61	0.65	0.62	6
7	Ryzer [22]	0.60	0.61	0.60	7
8	SSN_NLP_MLRG [17]	0.60	0.61	0.60	8
9	AIML [23]	0.60	0.60	0.60	9
10	KBCNMUJAL [15]	0.58	0.60	0.59	10
11	SSNCSE_NLP [9]	0.60	0.64	0.59	11
12	KonguCSE	0.57	0.62	0.57	12
13	MUM [19]	0.58	0.62	0.56	13
14	LogicDojo	0.54	0.59	0.56	14
15	IIIT DWD [24]	0.55	0.56	0.56	15
16	IIIT Surat [25]	0.54	0.57	0.55	16
17	Amrita_CEN [18]	0.64	0.50	0.53	17
18	SSN-NLP	0.62	0.49	0.51	18
19	DLRF	0.34	0.58	0.42	19
20	IRLab@IITBHU [20]	0.38	0.46	0.41	20
21	SSNHacML [26]	0.38	0.45	0.41	21
22	SSN_IT_NLP [27]	0.38	0.39	0.38	22

Table 5
Rank list based on weighted average F1-score along with other evaluation metrics (Precision and Recall) for the Tamil track

the classifiers is separately trained on training data. During testing, soft voting is employed among all of these classifiers to predict the most likely class.

- SOA_NLP[32]: The authors proposed the following two ensemble models for tackling the problem at hand: an ensemble of support vector machine, logistic regression and random forest for Kannada-English texts and an ensemble of support vector machine and logistic regression for Malayalam-English and Tamil-English texts.

- SSNCSE_NLP [33]: The authors have carried out experiments with different features such as TF-IDF vectors, count vectorizer and contextual transformer embeddings on primitive machine learning models.

- IIIT DWD [34]: The authors used pre-trained Word2Vec word embeddings and a parallel RNN model to feed the embeddings into, and have reported their findings on all three datasets.

- IIIT Surat [35]: The authors used several feature extraction and preprocessing techniques and then used GLoVe word embeddings and then fed those embeddings to Bi-directional Long-Short Term Memory (Bi-LSTM) model for further processing. For Char embedding,

No.	TeamName	Precision	Recall	F1-Score	Rank
1	ZYBank-AI Team [21]	0.80	0.81	0.80	1
2	CIA_NITT [11]	0.75	0.76	0.75	2
3	SOA_NLP [12]	0.73	0.73	0.73	3
4	MUCIC [10]	0.73	0.73	0.73	4
5	AIML [23]	0.72	0.72	0.72	5
6	IIITT-Pawan[16]	0.71	0.71	0.71	6
7	KBCNMUJAL [15]	0.73	0.71	0.71	6
8	SSN_NLP_MLRG [17]	0.70	0.71	0.70	7
9	SSNCSE_NLP [9]	0.69	0.69	0.69	8
10	Dynamic Duo [14]	0.69	0.70	0.69	9
11	IIITT-Karthik Puranik [13]	0.65	0.67	0.65	10
12	IRLab@IITBHU [20]	0.65	0.67	0.65	10
13	Amrita_CEN [18]	0.64	0.61	0.62	11
14	IIIT DWD [24]	0.62	0.62	0.62	12
15	IIIT Surat [25]	0.63	0.63	0.61	13
16	MUM [19]	0.62	0.63	0.60	14
17	LogicDojo	0.52	0.58	0.55	15

Table 6
Rank list based on weighted average F1-score along with other evaluation metrics (Precision and Recall) for the Malayalam track

64 units of Bi-LSTM were used, whereas for processing the words, 32 units of Bi-LSTM was used.

- SSN_NLP_MLRG [36]: The authors experimented with several machine learning algorithms during the validation process and then fine-tuned the MBERT model to build the system and predict the sentiment polarity for the Tamil-English, Malayalam-English, and Kannada-English languages.

- IRLab@IITBHU [37]: The authors examined if the use of meta embeddings such as FastText will give an edge over pre-trained embeddings such as mBERT. The authors feed meta embeddings into a multiheaded attention based transformer encoder and then over a BiLSTM layer and concatenating it with TF-IDF embeddings to obtain the final outputs.

- Amrita_CEN [38]: The authors implemented three architectures: Deep Neural Network (DNN), Bi-LSTM, and finally, Convolution Neural network (CNN) combined to a hybrid model for all the three test sets. Additionally, the authors use a class-weight optimization technique to handle class imbalance.

- SSNHacML [39]: The authors proposed an ensemble framework called Ensemble of Convolutional Neural Network and Multi-Head Attention with Bidirectional GRU (ECMAG) to map the code-mixed user comments to their corresponding sentiments. The model has been tested on the Tamil-English Code mixed dataset. The model takes XLMRoberta multilingual sub-word embeddings of the processed text data as input.

- MUM [40]: The authors converted the text data into feature vectors and then fed it into a BiLSTM network. The authors submit their predictions to the code-mixed test sets of Kannada, Malayalam, and Tamil.

- AIML [41]: The authors extracted character-level features from the text. The dense neural network then uses the extracted features to classify them into different sentiment classes.

- KBCNMUJAL [42]: The authors presented their systems for all three Dravidian Languages (Kannada-English, Tamil-English and Malayalam-English). They use models such as Multinomial Bayes (MNB), CNN and neural networks.

- Dynamic Duo [43]: The authors used a pre-trained language-based Model (BERT), wrapped with ktrain (a python library for model training and testing) to train and validate the data. The authors present their findings on the code-mixed Kannada-English dataset.

- Ryzer[44]: The authors used conventional translation and transliteration algorithms to convert the corpus into a native Tamil script and then fed the data into pre-trained language models like mBert, ULMFit, DistilBert. Additionally, They tested the approach on CNN-BiLSTM and ULMFiT.

- SSN_IT_NLP [45]: The authors used a conventional machine learning algorithm. The TF-IDF features are extracted and used for sentiment classification using a Random Forest classifier.

- SSNCSE_NLP [9]: The authors employed a variety of feature extraction techniques and concluded that the count, TF-IDF based vectorization, and multilingual transformer encoding technique performs well on the code-mix polarity labelling task. With these features, and acheived a weighted F1 score of 0.588 for the Tamil-English task, 0.69 for the Malayalam-English task and 0.63 for the Kannada-English tasks respectively.

4. Evaluation

The distribution of the sentiment classes is imbalanced in both datasets. This takes into account the varying degrees of importance of each class in the dataset. We used a classification report tool from Scikit learn[5].

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

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

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

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

$$P_{\text{weighted}} = \sum_{i=1}^{L} (P \text{ of } i \times \text{ Weight of } i) \tag{4}$$

$$R_{\text{weighted}} = \sum_{i=1}^{L} (R \text{ of } i \times \text{ Weight of } i) \tag{5}$$

$$F - Score_{\text{weighted}} = \sum_{i=1}^{L} (F - Score \text{ of } i \times \text{ Weight of } i) \tag{6}$$

5. Results and Discussion

The sentiment analysis shared task was organized for three languages Tamil, Kannada, and Malayalam. Overall, there are 120 participants registered for this shared task, yet 22 teams have submitted their working notes for Tamil, 17 for Malayalam, and 15 for Kannada.Table 4, Table 5, and Table 6 show the rank lists of Tamil, Malayalam and Kannada in that order. Most of the submissions submit their systems for the three languages, as specified earlier. Here in this section, we highlight the results of all three languages, which have ranked top positions on the dataset. The results are sorted based on the weighted F1 scores. Most of the teams have used transformer-based models such as BERT, DistilBERT, XLM-RoBERTa or other language models that follow its architecture, in spite of not being pretrained on code-mixed text. Due to the presence of a non-native script in our corpus, the teams got the pre-trained model from the libraries and adopted it for our corpus by fine-tuning. Some teams have used Long Short Term Memory (LSTM) and ULMFiT in their experiments. Also, a few other submissions adopted traditional machine learning algorithms such as Naive Bayes (NB), K-Nearest Neighbors, etc., to solve the problem.

However, LSTM and traditional machine learning algorithms did not yield good results compared to the transformer-based models. Out of all the proposed models, XLM-RoBERTa and the transformer-based model produced the best outcomes. Even though many systems with different approaches with F1-score less than the baseline, we accepted those papers to encourage diverse research methods to solve the problem in Dravidian Languages. Most working notes reported class-wise precision, recall, and F1-score. We used weighted F1 scores as our primary evaluation metric.

Among the Tamil teams, CIA_NITT [29] got the first position with an F1-score of 0.71. This system achieved 0.71 as the precision and recall score is the same as the F-score. The team from ZYBank-AI [21] achieved the second position with an F-score of 0.68, lagging the top by 0.03. The top five teams attained an F1 score higher than 0.62. Teams placed in the top positions utilized the transformer-based models for their experiments, particularly XLM-RoBERTa. Contextual embeddings are also found to be effective in this method to reach the top positions. In Malayalam, ZYBank-AI [21] and CIA_NITT [29] teams switched positions with an F1-score of 080 and 0.75, respectively. Team IIITT-Pawan [31] reached the third position with an F1-score of 0.63 and lagged the top team by only 0.08. According to the Kannada benchmark, CIA_NITT [29] secured the third position while SSNCSE_NLP [33] and MUCIC [28] teams reached first

Language	Team Name	Rank
Tamil	CIA_NITT [29]	1
	ZYBank-AI [21]	2
	IIITT-Pawan [31]	3
Kannada	SSNCSE_NLP [33]	1
	MUCIC [28]	2
	CIA_NITT [29]	3
Malayalam	ZYBank-AI Team [21]	1
	CIA_NITT [29]	2
	SOA_NLP [32]	3

Table 7
Overall Results with Top Three Ranks

and second places in the benchmark, respectively. Also, both teams have used traditional machine learning algorithms such as Logistic Regression, SVM with TF-IDF feature vectors. As we can see, these models have overcome the transformer-based models based on the performance and became the best in the Kannada benchmark.

Table 7 shows the overall results and teams that are placed in the top three positions. As we can see, only one team(CIA_NITT [29]) managed to be in the top 3 systems for the languages, along with achieving the best performance on the code-mixed Tamil dataset. Among the systems submitted during the evaluation period, we observe that the best performing models scored a weighted F1-score of 0.63 in Kannada, 0.80 in Malayalam, and 0.71 in Tamil.

6. Conclusion

We present the results of the sentiment analysis shared task on Tamil, Malayalam, and Kannada. The dataset used in the shared tasks included code-mixed instances obtained from social media. Specifically, the dataset was created from Youtube comments following human annotation. Most of the participants fine-tuned pretrained multilingual language models. At the same time, the top-performing systems involved the application of attention layers on the contextualized word embeddings and fine-tuning the models pretrained on the previous edition, DravidianCodeMix-2020's training data. Results indicate that there is room for improvement in all three languages Tamil, Malayalam, and Kannada. The increase in the number of participants and the better performance of the systems shows an increase in interest in Dravidian NLP.

Acknowledgments

This publication is the outcome of the research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2 (Insight_2), and Irish Research Council grant IRCLA/2017/129 (CARDAMOM-Comparative Deep Models of Language for Minority and Historical Languages).

References

[1] A. Hande, S. U. Hegde, R. Priyadharshini, R. Ponnusamy, P. K. Kumaresan, S. Thavareesan, B. R. Chakravarthi, Benchmarking Multi-Task Learning for Sentiment Analysis and Offensive Language Identification in Under-Resourced Dravidian Languages, ArXiv abs/2108.03867 (2021).

[2] B. R. Chakravarthi, R. Priyadharshini, R. Ponnusamy, P. K. Kumaresan, K. Sampath, D. Thenmozhi, S. Thangasamy, R. Nallathambi, J. P. McCrae, Dataset for Identification of Homophobia and Transophobia in Multilingual YouTube Comments, ArXiv abs/2109.00227 (2021).

[3] A. Hande, R. Priyadharshini, B. R. Chakravarthi, 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, Association for Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 54–63. URL: https://aclanthology.org/2020.peoples-1.6.

[4] A. Hande, K. Puranik, K. Yasaswini, R. Priyadharshini, S. Thavareesan, A. Sampath, K. Shanmugavadivel, D. Thenmozhi, B. R. Chakravarthi, Offensive Language Identification in Low-resourced Code-mixed Dravidian languages using Pseudo-labeling, ArXiv abs/2108.12177 (2021).

[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. 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.

[7] B. R. Chakravarthi, 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, Association for Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 41–53. URL: https://aclanthology.org/2020.peoples-1.5.

[8] A. Hande, R. Priyadharshini, A. Sampath, K. Thamburaj, P. Chandran, B. R. Chakravarthi, Hope Speech detection in under-resourced Kannada language, ArXiv abs/2108.04616 (2021).

[9] B. B, S. G. U, Machine learning based approach for sentiment analysis on Multilingual Code Mixing Text, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.

[10] F. Balouchzahi, H. L. Shashirekha, G. Sidorov, CoSaD- Code-Mixed Sentiments Analysis for Dravidian Languages, in: Working Notes of FIRE 2021 - Forum for Information

Retrieval Evaluation, CEUR, 2021.

[11] Y. Prakash Babu, R. Eswari, K. Nimmi, CIA_NITT@Dravidian-CodeMix-FIRE2020: Malayalam-English Code Mixed Sentiment Analysis Using Sentence BERT And Sentiment Features, in: FIRE (Working Notes), 2020.

[12] A. Kumar, S. Saumya, J. P. Singh, An ensemble-based model for Sentiment Analysis of Dravidian Code-mixed Social Media Posts, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.

[13] K. Puranik, B. B, B. S. Kumar, Transliterate or translate? Sentiment analysis of code-mixed text in Dravidian languages, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.

[14] S. Dutta, H. Agrawal, P. K. Roy, Sentiment Analysis on Multilingual Code Mixing Text using BERT, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.

[15] V. Pathak, M. Joshi, P. Joshi, M. Mundada, T. Joshi, KBCNMUJAL@HASOC-Dravidian-CodeMix-FIRE2020: Using Machine Learning for Detection of Hate Speech and Offensive Codemix Social Media text, in: FIRE (Working Notes), 2020.

[16] P. K. Jada, D. S. Reddy, K. Yasaswini, A. P. K, P. Chandran, A. Sampath, S. Thangasamy, Transformer based Sentiment Analysis in Dravidian Languages, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.

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

[18] P. K. P.H.V, P. B, S. Jp, S. Kp, ADeep Learning based Sentiment analysis forMalayalam,Tamil and Kannada languages , in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.

[19] A. M D, S. H L, BiLSTM-Sentiments Analysis in Code MixedDravidian Languages, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.

[20] A. Saroj, S. Pal, IRLab@IIT-BHU@Dravidian-CodeMix-FIRE2020: Sentiment Analysis on Multilingual Code Mixing Text Using BERT-BASE, in: FIRE (Working Notes), 2020.

[21] Y. Bai, B. Zhang, Y. Gu, T. Guan, Q. Shi, Automatic Detecting the Sentiment of Code-Mixed Text by Pre-training Model , in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.

[22] S. Sivapiran, C. Vasantharajan, U. Thayasivam, Sentiment Analysis in Dravidian Code-Mixed YouTube Comments and Posts , in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.

[23] J. Kumari, A. Kumar, A Deep Neural Network-based Model for the Sentiment Analysis of Dravidian Code-mixed Social Media Posts , in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.

[24] A. K. Mishra, S. Saumya, A. Kumar, Sentiment analysis of Dravidian-CodeMix language , in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.

[25] P. K. Roy, A. Kumar, Sentiment Analysis on Tamil Code-Mixed Text using Bi-LSTM , in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.

[26] P. D, S. J B, T. Durairaj, ECMAG - Ensemble of CNN and Multi-Head Attention with

Bi-GRU for Sentiment Analysis in Code-Mixed Data, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.

[27] S. N, D. S, Opinion And Attitude Investigation , in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.

[28] F. Balouchzahi, H. L. Shashirekha, G. Sidorov, CoSaD- Code-Mixed Sentiments Analysis for Dravidian Languages, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.

[29] Y. Prakash Babu, R. Eswari, Sentiment Analysis on Dravidian Code-Mixed YouTube Comments using Paraphrase XLMRoBERTa Model, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.

[30] Y. B. Bai, B. Zhang, Y. Gu, T. Guan, Q. Shi, Automatic Detecting the Sentiment of Code-Mixed Text by Pre-training Model, in: FIRE (Working Notes), 2020.

[31] P. K. Jadaa, S. Reddy, K. Yasawini, A. Pandian K, P. Chandran, A. Sampath, S. Thangasamy, Transformer based Sentiment Analysis in Dravidian Languages, in: FIRE (Working Notes), 2020.

[32] A. Kumar, S. Saumya, J. P. Singh, An ensemble-based model for Sentiment Analysis of Dravidian Code-mixed Social Media Posts, in: FIRE (Working Notes), 2020.

[33] B. Bharathi, G. Samyuktha, Machine learning based approach for sentiment Analysis on Multilingual Code Mixing Text, in: FIRE (Working Notes), 2020.

[34] A. K. Mishra, S. Saumya, A. Kumar, Sentiment analysis of Dravidian-CodeMix language, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.

[35] P. K. Roy, A. Kumar, Sentiment Analysis on Tamil Code-Mixed Text using Bi-LSTM, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.

[36] K. Adaikkan, T. Durairaj, Multilingual Sentiment Analysis in Tamil, Malayalam, and Kannada code-mixed social media posts, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.

[37] S. Chanda, R. P. Singh, S. Pal, Is Meta Embedding better than pre-trained word embedding to perform Sentiment Analysis for Dravidian Languages in Code-Mixed Text?, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.

[38] P. Kumar, P. B, S. J.P, S. KP, Deep Learning based Sentiment analysis for Malayalam,Tamil and Kannada languages, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.

[39] P. D, S. J B, T. Durairaj, ECMAG - Ensemble of CNN and Multi-Head Attention with Bi-GRU for Sentiment Analysis in Code-Mixed Data, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.

[40] A. MD, H. L. Shashirekha, BiLSTM-Sentiments Analysis in Code MixedDravidian Languages, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.

[41] J. Kumari, A. Kumar, A Deep Neural Network-based Model for the Sentiment Analysis of Dravidian Code-mixed Social Media Posts, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.

[42] P. Joshi, V. Pathak, Offensive Language Identification on Code-mixed Dravidian Languages, A Non-linguistic Approach, in: Working Notes of FIRE 2021 - Forum for Infor-

mation Retrieval Evaluation, CEUR, 2021.

[43] S. Dutta, H. Agrawal, P. K. Roy, Sentiment Analysis on Multilingual Code Mixing Text using BERT, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.

[44] S. Sivapiran, C. Vasantharajan, U. Thayasivam, Sentiment Analysis in Dravidian Code-Mixed YouTube Comments and Posts, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.

[45] S. N, D. S, Opinion And Attitude Investigation, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.