Transformer based Sentiment Analysis in Dravidian Languages

Pawan Kalyan Jada[a], D Sashidhar Reddy[a], Konthala Yasaswini[a], Arunaggiri Pandian K[b], Prabakaran Chandran[b], Anbukkarasi Sampath[c] and Sathiyaraj Thangasamy[d]

[a]Indian Institute of Information Technology Tiruchirappalli
[b]Thiagarajar College of Engineering, Madurai, India
[b]Mu Sigma Inc., Bengaluru, Karnataka, India
[c]Kongu Engineering College, Erode, Tamil Nadu, India
[d]Sri Krishna Adithya College of Arts and Science, Coimbatore, Tamil Nadu, India

Abstract

The development of social media platforms have enabled users to express their thoughts and opinions about entities freely, without any inadvertent implications it may have on a person/group. Due to the volume of active social media users, it is becoming increasingly apparent for the need of automated sentiment analysis systems for social media. This paper describes our work on the task of Sentiment Analysis in Dravidian language-DravidianCodeMix 2021. We propose a soft voting classifier with the help of other fine-tuned multilingual language models, achieving the best weighted F1-Score of 0.752, 0.619, and 0.648 in Malayalam, Tamil, and Kannada respectively. Our approach achieved the best results in Tamil, securing 3^{rd} rank in the language. The source codes of our systems are published[1].

Keywords

Sentiment Analysis, Dravidian languages, Transfer learning, Transformers

1. Introduction

Social media is a powerful and robust tool that has led an inherent impact on the users [1]. Over the years, the internet has gained more and more users jumping from 738M in the year 2000 all the way up to 3.6B in 2020. It provided a medium for people of different age, background, ethnicity to interact accounting for more cultural exchanges as well as exposure to new ideologies from different users. It also enabled us to access information across the planet, to socialize and stay up-to-date with the latest technologies and to share our ideas and thoughts to the world. With millions of active users, content generated on social media is difficult to be moderated by human beings. Analysing the opinions expressed by the users is important to identify the areas of disagreements and differences among the users [2].

[1]https://github.com/PawanKalyanJada/dravidian-code-mix

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

pawankj@iiitt.ac.in (P. K. Jada); duddukuntasr19e@iiitt.ac.in (D. S. Reddy); konthalay18c@iiitt.ac.in (K. Yasaswini); arunabimanyu123@gmail.com (A. P. K); prabakaran.chandran98@gmail.com (P. Chandran); anbu.1318@gmail.com (A. Sampath); sathiyarajt@skacas.ac.in (S. Thangasamy)

Users can openly share their ideas on social media sites such as YouTube, Facebook, Instagram, and Twitter. Certain people's perspectives can be harmful to a specific community, gender, religion, or race. These unpleasant posts/comments might be detrimental to one's mental health. Sentiment analysis is the technique of categorising a statement based on its polarity. Sentiment analysis aids in evaluating consumer satisfaction with the products and services that many businesses give [3], as well as understanding public opinion, which may aid in making better decisions in the future. Indefinitely, it became a prominent subject of study in the Natural Language Processing research field. Because of the huge quantity of data created on a daily basis, the study into evaluating the sentiment on social media postings has grown exponentially.

The majority of data found on social media is frequently code-mixed. The combination of two or more languages in a phrase is known as code-mixing [4, 5, 6]. Because of variations in syntax, vocabulary, and meaning, code-mixed writing is far more difficult to read than standard language. As a result, achieving good results in activities such as Sentiment Analysis, Named Entity Recognition, POS Tagging, and so on becomes extremely difficult.

Tamil evolved from the Proto-Dravidian language, which is estimated to have existed prior to 500 BC. Tamil is the official language of the Indian state of Tamil Nadu, as well as Singapore and Sri Lanka [7]. There are around 77 million Tamil speakers worldwide. The Tamil-Brahmi script was the parent script from which the subsequent Vatteluttu and Tamil scripts evolved. It consists of 12 vowels, 18 consonants, and 1 aytam (voiceless velar fricative).

Kannada and Malayalam are two more Dravidian languages that are widely spoken in Karnataka and Kerala, respectively. Kannada can alternatively be spelled Kanarese or Kannana. Kannada is spoken by about 40 million people and is recognised as a classical language. The earliest Kannada inscription comes from around 450 CE. Kannada literature was influenced by the Lingayat and Haridasa movements and began with Kavirajamarga and Pampa Bharata. Ramacharitan is the oldest surviving literacy text in Malayalam. Malayalam contains 15 vowels, 36 consonants, and a variety of additional symbols. The Valleluttu script is incorporated in the contemporary Malayalam. These Dravidian languages generate a massive volume of code-mixed data [8].

The rest of the paper is organised as follows, Section 2 comprises of the related work in sentiment analysis. Section 3 entails the dataset used and task descriptions, while Section 4 provides a detailed description of the architecture used for this task. Section 5 discusses about the results of our models in the shared task, and finally, Section 6 concludes our work and talks about potential directions for future works.

2. Related Work

Sentiment analysis is essential in introspection [9]. The availability of code-mixed data from social media was critical to extracting data for sentiment analysis [10]. The topic of code-mixing in Dravidian languages is explored in [11, 12]. Sentiment analysis tasks were completed in the late 1990s by classifying text or phrases [13]. Finn Arup Nielsen, Opinion Finder, and General Inquirer produced a new word list in order to provide a score to each term [14]. The sentiment of a sentence is determined by the individual score of each word in the sentence. Two typical

ways to solving a sentiment analysis problem are machine learning approaches and lexicon-based approaches [15]. Opinion lexicon is employed in the Lexicon-based technique to identify sentence polarity [16, 17]. Naive Bayes is one approach for dealing with sentiment analysis. In previous years, N-grams were proposed to extract sentiments [18]. These approaches were ineffective due to the dynamic nature of data. Many studies have been conducted in recent years to integrate deep learning and machine learning approaches for effective sentiment categorisation.

In [19], the authors have developed a model using Conditional Random Field for part-of-speech tagging on mixed script social media text which contained two or three languages, among which English is one language and the others are Hindi, Bengali and Tamil.

Several types of multilingual and cross-lingual embeddings were employed in order to efficiently transfer knowledge from monolingual text to code-mixed language for sentiment analysis of code-mixed text in. These embeddings have shown to improve the performance of sentiment analysis on code-mixed text. [20] presented the *Sentiment Analysis of Code-Mixed Text* (SACMT) model, which consists of twin Bidirectional LSTM networks for sentiment analysis of code-mixed text and address the problem by projecting the sentences onto a single sentiment space using shared parameters. This method outperforms the state-of-the-art sentiment analysis methods on code-mixed data. For the sentiment analysis of Dravidian code-mixed dataset, [21] presented meta embeddings with the transformer and GRU model. When sarcasm is employed in negative polarity remarks, the system is unable to determine the sentiment. [22, 23] employed a hybrid model of bidirectional LSTM and CNN architectures which extracts character features from each word. Crowdsourcing approaches were utilised in [24, 25] to manually rate polarity in twitter posts. To classify a sentence into one of the sentiment classes, a parallel ensemble of two models - a traditional machine learning model and an end-to-end deep learning model was employed in [26].

The authors of [27] provides a unique technique to detecting sentiment polarity in Twitter messages by extracting a vector of weighted nodes from the WordNet graph which presents a domain-independent non-supervised solution. An end-to-end Cross-lingual sentiment analysis (CSLA) model that eliminates the requirement for unsupervised cross-lingual word embeddings (CLWE) by utilising unlabelled data in different languages and domains was introduced by the authors of [28]. A significant element of [29] is the development of a new multimodal opinion database labelled at the utterance level. In [30], a benchmark dataset, a comprehensive corpus of around 12000 Bengali reviews, was introduced and the performance of supervised machine learning (ML) classifiers was evaluated in a machine-translated English dataset and compared to the source Bengali dataset. Several researchers bench marked multi-task learning on auxiliary tasks on Dravidian languages [31].

A surge of information is created everyday as a result of world-wide internet usage, which poses huge risks, because online texts with high toxicity can cause personal assaults, mental health problems, online harassment, and bullying behaviours [32]. In [33] the authors integrated the outcomes of three feature-based classifiers for identifying cyber hate speech on Twitter and investigated the benefits of ensembles of different classifiers. The authors of [34] investigated the challenge of hate speech identification in code-mixed texts and provided a dataset of code-mixed Hindi-English tweets from Twitter. The authors of [35] suggested a typology that encapsulates the key similarities and distinctions across subtasks, and addressed

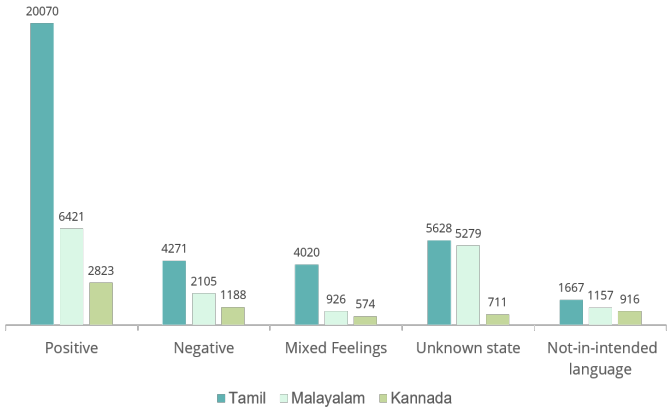

Figure 1: Class-wise distribution of the Training Set

the consequences for feature construction and data annotation, based on the previous work on hate speech, cyberbullying, and online abuse.

3. Dataset and Task description

In this section, we describe the dataset provided by the organisers to the participants and the task [36, 37, 10].

3.1. Dataset

The organisers of FIRE-2021 provided training and validation code-mixed data sets in Tamil-English, Kannada-English and Malayalam-English [38, 39, 40]. The datasets consist of comments collected from Youtube that are annotated with sentiment polarity. In the data sets, there are three types of code-mixed sentences : Inter-Sentential Code-Mixing, Intra-Sentential Code-Mixing and Tag switching. The training and validation data sets comprises of sentences in five classes :

1. ***Positive state*** - The comment provides an explicit or implicit indication that the speaker is in a positive state.
2. ***Negative state*** - The comment provides an explicit or implicit indication that the speaker is in a negative state.
3. ***Mixed feelings*** - The comment provides an explicit or implicit indication that the speaker is experiencing both positive and negative feeling.
4. ***Neutral state*** - The comment provides no explicit or implicit indication of the speaker's emotional state.
5. ***Not in intended language*** - Comment not in Tamil/Malayalam/Kannada.

The Tamil code-mix data set consists of 35,656 comments for the train set, 3,962 for the validation set and 4,403 comments for testing the model. In the Kannada code-mix data set,

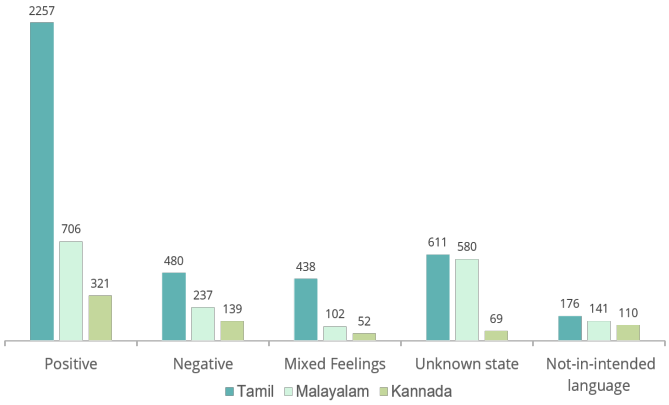

Figure 2: Class-wise distribution of the Validation Set

there are 6,212 comments for training, 6,91 for validating and 7,68 for testing the model. The Malayalam code-mix data set comprises of 15,888 comments in training set, 1,766 comments in validation set and 1,963 comments in test set.

3.2. Task description

The participants are required to produce labels indicating the sentiment polarity of a given code-mixed comment. Each sentence should be classified into one of these labels : Positive, Negative, Neutral, Mixed feelings, Not-in-intended language. At the beginning of the task, the training and development data sets were already made available to the participants. Only the comments from the test split were eventually made accessible to participants via Codalab. The weighted-average F1 scores were considered for official ranking since the labels in the task were not balanced.

Table 1
Examples of code-mixing sentences from the dataset

Text	Language	Class
Ithu yethu maathiri illama puthu maathiyaala irukku	Tamil	Positive
Pulikku pakaram patti odande vere mattam onnum ella.	Malayalam	Mixed feelings
ആദ്യ നൂറു കോടി വേണ്ടവർ ... adei mwonoose like	Malayalam	Neutral
ರಂಗಿತರಂಗದ ಇತಿಹಾಸ ಮರುಕಳಿಸುವಂತಿದೆ!	Kannnada	Positive
Are bhai Yek dum phel diya	Kannada	Not-kannada

4. System Description

To determine the sentiment of a particular text, we employed pre-trained transformer models. The models employed for the cause are MuRIL [41], mBERT [42], DistilmBERT [43] and XLM-Roberta [44]. These models are then fine-tuned for this particular task. For all three languages,

the same models were utilised. After obtaining the probability scores from different models, we soft vote [45] these scores to get our final result. Soft Voting computes the weighted sum of all the probabilities for each class label and then forecasts the class label with the highest likelihood. Each individual classifier in soft voting offers a probability value that a certain data point belongs to a specified target class. The predictions are weighted by the significance of the classifier and totaled. The target label with the highest sum of weighted probability then receives the vote.

4.1. MuRIL

MuRIL [41] is an Indic language model that has been extensively trained and improved to perform better in Indian languages. It supports around 17 languages, including English and 16 other Indian languages. MuRIL surpassed multilingual BERT on all benchmark data sets of Indic languages. Masked Language Modeling (MLM) [46] and Translation Language Modeling (TLM) are two approaches used in MuRIL's pre-training phase. TLM makes use of parallel translation data where it takes a sequence of parallel sentences from the translation data and randomly mask tokens from the source as well as from the target sentence, hence establishing a cross-lingual mapping among the tokens. MuRIL is the outcome of pre-training a BERT-based Encoder model with MLM and TLM objectives. MuRIL was also pre-trained on PMINDIA and Dakshina datasets. It comprises of 236M parameters.

4.2. XLM-Roberta

XLM-Roberta [44] is a variant of Roberta that is multilingual. The hybrid model XLM-Roberta was trained on 2.5 TB of commoncrawl data and combines XLM and Roberta. It was trained using the multilingual MLM loss on 100 different languages. XLM-R achieved state-of-the-art results on multiple cross lingual benchmarks. *xlm-roberta-base* is fine-tuned for our sentiment analysis task, which contains 12-layers, 768-hidden-state, 8-heads and a parameter size of 270M.

4.3. BERT

The Encoder of a Transformer is utilised in the design of Bidirectional Encoder Representation from Transformers (BERT). During its pre-training phase, BERT is trained on the whole English Wikipedia and the Brown Corpus. It is trained with two language modelling objectives: Masked Language Modeling (MLM), in which 15% of the tokens are randomly masked, and Next Sentence Prediction (NSP), in which the model must predict whether the first sentence precedes the second sentence or not. Here, we adopt a *bert-base-multilingual-cased* [47] model trained on top of the largest Wikipedia corpus, which contains 104 languages. This model consists of 12 layers, 12 Attention heads, and approximately 179M parameters.

4.4. DistilBERT

DistilBERT [48] is a modified version of BERT model. It uses triple loss language modelling that combines cosine distance loss with knowledge distillation. In comparison to the MLM loss, the

two distillation losses in the triple loss have a significant impact on model performance. The authors found it useful to include a cosine embedding loss, which tends to align the directions in big model distillation. Knowledge distillation is a compression approach that involves training a small model to mimic the behaviour of a bigger model. DistilBERT is not only 60% faster than BERT, but it also includes 40% less parameters. In this case, we use a cased multilingual distilbert model with 6 layers, 768 dimensions, and 12 Attention heads.

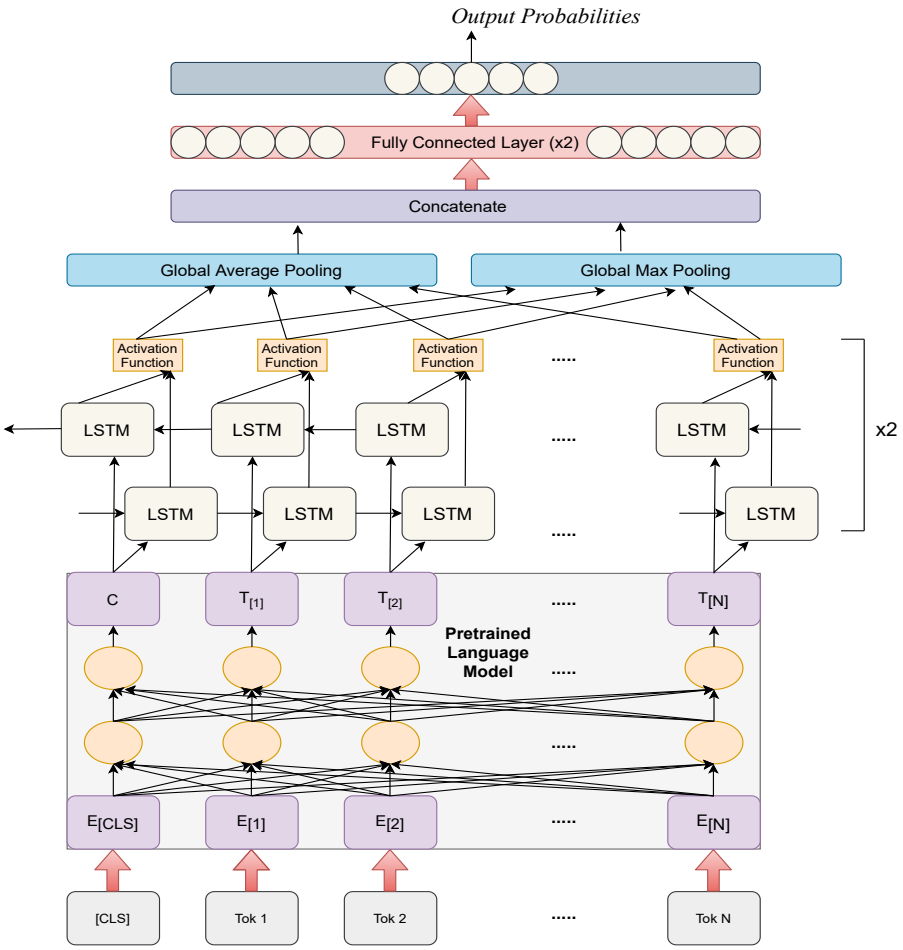

Figure 3: System Architecture based on Transformers

4.5. Methodology

First, we preprocess the data by removing emojis and punctuation. The text is then tokenized using the tokenizer of the corresponding language model, and all sequences are padded to the same length. The sequence output is then retrieved and then sent up to two BiLSTM [49] layers with units 200 and 100, respectively. The generated output values are concatenated after being fed into a global average pooling layer and a global max pooling layer. To acquire the probability scores, this is then fed into several Fully Connected layers, followed by a softmax activation function as shown in Figure 3. Refer Table 2 for the parameters used in the models.

Parameters	Values
Optimizer	Adam
Dropout Rate	0.5
Batch Size	64
Max Length	200
Learning Rate	1e-3
Activation Function	Softmax
Loss Function	cross-entropy

Table 2
Parameters used for training the Models

5. Results and Analysis

We have fine-tuned various transformer models, like MuRIL, BERT, XLM-RoBERTa, DistilBERT. We used the Tensorflow implementation of the models, provided by the Hugging-face library. Based on the results released by the organisers, we have secured third position with an F1-score of 0.626 on the Tamil test set. We received F1-scores of 0.609 and 0.708 on the Kannada and Malayalam test sets, respectively. We anticipated the models would provide similar results based on the soft-voting scores we obtained on the validation sets. In comparison to the Tamil models, the Kannada and Malayalam models performed relatively poor on the Kannada and Malayalam test sets, contrary to our expectations. The results of soft-voting technique on test data sets is shown in table 3.

Language	F1-score
Tamil	0.626
Malayalam	0.708
Kannada	0.609

Table 3
Weighted F1-scores of soft-voting method on test data sets

We submitted the results obtained through the soft-voting technique because it produced the best results for the three Dravidian languages. We also submitted the scores of distilBERT for Tamil and XLM-Roberta for Kannada and Malayalam. Soft-voting yielded F1-scores of 0.619, 0.752, and 0.648 for Tamil, Malayalam, and Kannada, respectively. Table 4 shows the weighted

average Precision, Recall, and F1-scores of the transformer models and soft-voting approach evaluated on development data sets of the three Dravidian languages.

Among the transformer models, distilBERT performed better on the Tamil validation set with an F1-score of 0.607, while XLM-Roberta performed better on the Malayalam and Kannada validation sets with F1-scores of 0.721 and 0.621, respectively. MuRIL is the model which gave rather poor performance on Tamil and Kannada data sets despite being specifically built for Indian languages.

Table 4
Weighted F1-scores of the models on the data sets

Model	**Code-mixed data set**								
	Malayalam			Tamil			Kannada		
	W(P)	**W(R)**	**W(F1)**	**W(P)**	**W(R)**	**W(F1)**	**W(P)**	**W(R)**	**W(F1)**
BERT	0.679	0.663	0.668	0.7461	0.585	0.608	0.619	0.608	0.608
XLM-R	0.720	0.725	0.721	0.590	0.613	0.596	0.634	0.618	0.621
DistilBERT	0.672	0.678	0.672	0.597	0.625	0.607	0.628	0.627	0.617
MuRIL	0.678	0.674	0.675	0.586	0.618	0.582	0.601	0.631	0.606
Soft-voting	0.751	0.757	**0.752**	0.613	0.649	**0.619**	0.656	0.656	**0.648**

When compared to the results of soft-voting technique on the Tamil test set and development set, the F1-score improved from 0.619 to 0.626. We have also observed a decrease in the performance of soft-voting in the Kannada and Malayalam languages. One of the causes is the data sets discrepancy in class distribution. The majority of the texts fall into the positive category, followed by the unknown state and negative categories. Our models performed well in the majority class and poorly in the minority class. We also notice that the F1-score of the Mixed Feeling class is quite low when compared to the non-Tamil class, despite the fact that the number of sentences belonging to the former label is significantly higher than the latter in the Tamil validation-set. Another anomaly we noticed is that the F1-Score of not-malayalam label is the highest even though the data set has more samples belonging to the positive class in Malayalam data set.

6. Conclusion

In this paper, we describe our work on sentiment analysis of Dravidian languages for Kannada, Malayalam, and Tamil. We fine-tuned various pre-trained multilingual natural language models such as BERT, XLM-R, DistilBERT, and MuRIL to classify the sequence into one of these 5 classes: Positive, Negative, Neutral, Mixed-feelings, and Not-in-indented language. Overall, XLM-Roberta performed competently compared to other models. Soft Voting technique is applied to the individual model's probabilities, enhancing the overall performance. The problem of class imbalance had a serious impact on performance of the model in the low support classes. The soft-voting technique achieved weighted-average F1 Scores of 0.626, 0.708 and

0.609 for Tamil, Malayalam, and Kannada respectively. In the future, we intend to apply class weighting techniques and semi-supervised approaches to further improve our performance.

References

[1] T. Aichner, M. Grünfelder, O. Maurer, D. Jegeni, Twenty-five years of social media: a review of social media applications and definitions from 1994 to 2019, Cyberpsychology, Behavior, and Social Networking 24 (2021) 215–222.

[2] P. Sobkowicz, M. Kaschesky, G. Bouchard, Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web, Government information quarterly 29 (2012) 470–479.

[3] M. P. Anto, M. Antony, K. Muhsina, N. Johny, V. James, A. Wilson, Product rating using sentiment analysis, in: 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), IEEE, 2016, pp. 3458–3462.

[4] U. Barman, A. Das, J. Wagner, J. Foster, 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, 2014, pp. 13–23.

[5] P. Muysken, P. C. Muysken, et al., Bilingual speech: A typology of code-mixing, Cambridge University Press, 2000.

[6] A. Hande, K. Puranik, R. Priyadharshini, B. R. Chakravarthi, Domain identification of scientific articles using transfer learning and ensembles, in: Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2021 Workshops, WSPA, MLMEIN, SDPRA, DARAI, and AI4EPT, Delhi, India, May 11, 2021 Proceedings 25, Springer International Publishing, 2021, pp. 88–97.

[7] J. W. Christie, The medieval tamil-language inscriptions in southeast asia and china, Journal of Southeast Asian Studies (1998) 239–268.

[8] A. Pratapa, G. Bhat, M. Choudhury, S. Sitaram, S. Dandapat, K. Bali, Language modeling for code-mixing: The role of linguistic theory based synthetic data, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 1543–1553.

[9] E. De Saa, L. Ranathunga, Self-reflective and introspective feature model for hate content detection in sinhala youtube videos, in: 2020 From Innovation to Impact (FITI), volume 1, IEEE, 2020, pp. 1–6.

[10] 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, 2020, pp. 54–63.

[11] K. Krishnasamy, Code mixing among tamil-english bilingual children, International Journal of Social Science and Humanity 5 (2015) 788.

[12] S. N. Sridhar, On the functions of code-mixing in kannada (1978).

[13] V. Hatzivassiloglou, K. McKeown, Predicting the semantic orientation of adjectives, in: 35th annual meeting of the association for computational linguistics and 8th conference of the european chapter of the association for computational linguistics, 1997, pp. 174–181.

[14] F. Å. Nielsen, A new anew: Evaluation of a word list for sentiment analysis in microblogs, arXiv preprint arXiv:1103.2903 (2011).
[15] T. K. Patel, L. Habimana-Griffin, X. Gao, B. Xu, S. Achilefu, K. Alitalo, C. A. McKee, P. W. Sheehan, E. S. Musiek, C. Xiong, et al., Dural lymphatics regulate clearance of extracellular tau from the cns, Molecular neurodegeneration 14 (2019) 1–9.
[16] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede, Lexicon-based methods for sentiment analysis, Computational linguistics 37 (2011) 267–307.
[17] K. Puranik, A. Hande, R. Priyadharshini, T. Durairaj, A. Sampath, K. P. Thamburaj, B. R. Chakravarthi, Attentive fine-tuning of Transformers for Translation of low-resourced languages @LoResMT 2021, in: Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages, European Association for Machine Translation, Online, 2021.
[18] F. Aisopos, D. Tzannetos, J. Violos, T. Varvarigou, Using n-gram graphs for sentiment analysis: an extended study on twitter, in: 2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService), IEEE, 2016, pp. 44–51.
[19] S. Ghosh, S. Ghosh, D. Das, Part-of-speech tagging of code-mixed social media text, in: Proceedings of the second workshop on computational approaches to code switching, 2016, pp. 90–97.
[20] N. Choudhary, R. Singh, I. Bindlish, M. Shrivastava, Sentiment analysis of code-mixed languages leveraging resource rich languages, 2018. arXiv:1804.00806.
[21] S. Dowlagar, R. Mamidi, Cmsaone@dravidian-codemix-fire2020: A meta embedding and transformer model for code-mixed sentiment analysis on social media text, 2021. arXiv:2101.09004.
[22] K. Yasaswini, K. Puranik, A. Hande, R. Priyadharshini, S. Thavareesan, B. R. Chakravarthi, 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, Kyiv, 2021, pp. 187–194. URL: https://aclanthology.org/2021.dravidianlangtech-1.25.
[23] K. Puranik, A. Hande, R. Priyadharshini, S. Thavareesan, B. R. Chakravarthi, 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, Kyiv, 2021, pp. 98–106. URL: https://aclanthology.org/2021.ltedi-1.13.
[24] N. A. Diakopoulos, D. A. Shamma, Characterizing Debate Performance via Aggregated Twitter Sentiment, Association for Computing Machinery, New York, NY, USA, 2010, p. 1195–1198. URL: https://doi.org/10.1145/1753326.1753504.
[25] S. Thabasum Aara, P. K. Arunaggiri, T. S. Sai Kumar, A. Prabalakshmi, A novel convolutional neural network architecture to diagnose covid-19, in: 2021 3rd International Conference on Signal Processing and Communication (ICPSC), 2021, pp. 595–599. doi:10.1109/ICSPC51351.2021.9451701.
[26] M. G. Jhanwar, A. Das, An ensemble model for sentiment analysis of hindi-english code-mixed data, 2018. arXiv:1806.04450.
[27] A. Montejo-Ráez, E. Martínez-Cámara, M. T. Martin-Valdivia, L. A. U. López, Random walk weighting over sentiwordnet for sentiment polarity detection on twitter, in: Pro-

ceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis, 2012, pp. 3–10.

[28] Y. Feng, X. Wan, Towards a unified end-to-end approach for fully unsupervised cross-lingual sentiment analysis, in: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 1035–1044. URL: https://aclanthology.org/K19-1097. doi:10.18653/v1/K19-1097.

[29] V. Pérez-Rosas, R. Mihalcea, L.-P. Morency, Utterance-level multimodal sentiment analysis, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2013, pp. 973–982.

[30] S. Sazzed, Cross-lingual sentiment classification in low-resource Bengali language, in: Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020), Association for Computational Linguistics, Online, 2020, pp. 50–60. URL: https://aclanthology.org/2020.wnut-1.8. doi:10.18653/v1/2020.wnut-1.8.

[31] A. Hande, R. Priyadharshini, A. Sampath, K. P. Thamburaj, P. Chandran, B. R. Chakravarthi, Hope speech detection in under-resourced kannada language, 2021. arXiv:2108.04616.

[32] S. V. Georgakopoulos, S. K. Tasoulis, A. G. Vrahatis, V. P. Plagianakos, Convolutional neural networks for toxic comment classification, in: Proceedings of the 10th Hellenic Conference on Artificial Intelligence, SETN '18, Association for Computing Machinery, New York, NY, USA, 2018. URL: https://doi.org/10.1145/3200947.3208069. doi:10.1145/3200947.3208069.

[33] P. Burnap, M. Williams, Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making, Policy & Internet 7 (2015) 223–242.

[34] A. Bohra, D. Vijay, V. Singh, S. S. Akhtar, M. Shrivastava, A dataset of Hindi-English code-mixed social media text for hate speech detection, in: Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, Association for Computational Linguistics, New Orleans, Louisiana, USA, 2018, pp. 36–41. URL: https://aclanthology.org/W18-1105. doi:10.18653/v1/W18-1105.

[35] Z. Waseem, T. Davidson, D. Warmsley, I. Weber, Understanding abuse: A typology of abusive language detection subtasks, 2017. arXiv:1705.09899.

[36] B. R. Chakravarthi, P. K. Kumaresan, R. Sakuntharaj, A. K. Madasamy, S. Thavareesan, P. B, S. Chinnaudayar Navaneethakrishnan, J. P. McCrae, T. Mandl, Overview of the HASOC-DravidianCodeMix Shared Task on Offensive Language Detection in Tamil and Malayalam, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.

[37] B. R. Chakravarthi, R. Priyadharshini, S. Thavareesan, D. Chinnappa, T. Durairaj, E. Sherly, J. P. McCrae, A. Hande, R. Ponnusamy, S. Banerjee, C. Vasantharajan, Findings of the Sentiment Analysis of Dravidian Languages in Code-Mixed Text 2021, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.

[38] B. R. Chakravarthi, R. Priyadharshini, V. Muralidaran, N. Jose, S. Suryawanshi, E. Sherly, J. P. McCrae, Dravidiancodemix: Sentiment analysis and offensive language identification

dataset for dravidian languages in code-mixed text, CoRR abs/2106.09460 (2021). URL: https://arxiv.org/abs/2106.09460. arXiv:2106.09460.

[39] 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://aclanthology.org/2020.sltu-1.25.

[40] 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://aclanthology.org/2020.sltu-1.28.

[41] S. Khanuja, D. Bansal, S. Mehtani, S. Khosla, A. Dey, B. Gopalan, D. K. Margam, P. Aggarwal, R. T. Nagipogu, S. Dave, S. Gupta, S. C. B. Gali, V. Subramanian, P. Talukdar, Muril: Multilingual representations for indian languages, CoRR abs/2103.10730 (2021). URL: https://arxiv.org/abs/2103.10730. arXiv:2103.10730.

[42] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[43] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, ArXiv abs/1910.01108 (2019).

[44] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, CoRR abs/1911.02116 (2019).

[45] P. Kalyan, D. Reddy, A. Hande, R. Priyadharshini, R. Sakuntharaj, B. R. Chakravarthi, IIITT at CASE 2021 task 1: Leveraging pretrained language models for multilingual protest detection, in: Challenges and Applications of Automated Extraction of Socio-political Events from Text, Association for Computational Linguistics, Online, 2021, pp. 98–104. URL: https://aclanthology.org/2021.case-1.13.

[46] W. L. Taylor, "cloze procedure": A new tool for measuring readability, Journalism quarterly 30 (1953) 415–433.

[47] T. Pires, E. Schlinger, D. Garrette, How multilingual is multilingual bert?, arXiv preprint arXiv:1906.01502 (2019).

[48] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019).

[49] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation 9 (1997) 1735–1780.