Understanding the role of Emojis for emotion detection in Tamil

Ratnavel Rajalakshmi [1,*], **Faerie Mattins R**[1], **Srivarshan Selvaraj**[1],
Antonette Shibani[2], **Anand Kumar M**[3],
Bharathi Raja Chakravarthi[‡]
[1] School of Computer Science and Engineering, Vellore Institute of Technology, Chennai
[2] University of Technology Sydney, Sydney
[3] National Institute of Technology Karnataka Surathkal.
[‡]School of Computer Science, University of Galway, Ireland
rajalakshmi.r@vit.ac.in,faeriemattins.r2019@vitstudent.ac.in,
srivarshan.2019@vitstudent.ac.in, Antonette.Shibani@uts.edu.au,
manandkumar@nitk.edu.in
bharathiraja.asokachakravarthi@universityofgalway.ie

Abstract

Emotions are commonly discerned by a persons facial expression and body movements. Detecting emotion only through text using Natural language processing (NLP) is a challenging research area for low-resource languages like Tamil. One way to identify emotion is with the help of emojis that are indicative of the emotion expressed by the writer. This paper presents a study on how emojis represent emotion in text and their usage in building machine-learning techniques to detect emotion. Feature extraction techniques like TF-IDF and MuRIL are used with classifiers like Logistic Regression, Random Forest, and XGBoost to detect emotions in Tamil YouTube comments. The most commonly used emojis and the number of times an emoji is repeated in a specific text are analyzed, as well as how they relate to emotion recognition. A combination of TF-IDF and XGBoost achieves the best performance of 0.32 weighted-average F1 score, with the emojis in the text substituted with phrases that depict them.

1 Introduction

The technique of recognizing a person's emotional state of mind by facial expression and demeanor is known as emotion detection (ED) (Chakravarthi and Muralidaran, 2021; Chakravarthi et al., 2022). Detecting a person's emotion in the text is difficult since it seldom provides phrases that explicitly stress the individual's feelings, and emotion is only discovered by interception of concepts through text data. ED is critical in many rapidly evolving fields including e-commerce, social media, comprehensive search, and advertising. Despite past work on ED including speech and facial expressions, text-based ED is limited (Acheampong et al., 2020). Furthermore, ED in Tamil texts is harder than in English due to the scarcity of corpora and NLP tools for low resource languages like Tamil (Thavareesan and Mahesan, 2019, 2020a,b).

People comment on posts/videos on social media sites such as Twitter, YouTube, and Instagram and express their emotions (Chakravarthi, 2020, 2022a,b). Because facial expressions cannot be observed in writing, emojis can be used to infer how the person is feeling (Hande et al., 2022; Shanmugavadivel et al., 2022; Subramanian et al., 2022). Emoji usage is also quite widespread on social media since it allows individuals to express themselves. However, little progress has been made in comprehending the significance of emojis in ED in texts, particularly in low-resource languages like Tamil. Along with the fundamental emotions of fear, anger, joy, sorrow, disgust, and surprise (Cherbonnier and Michinov, 2021),

Proceedings of the Workshop on Multimodal Machine Learning in Low-resource Languages, pages 9 - 17
December 15, 2022 ©2022 Association for Computational Linguistics

five additional categories of neutral, ambiguous, anticipation, love, and trust are used in this study.

This research investigates Tamil YouTube comments to determine the emotions they represent. The primary purpose of this article is to investigate how effectively emojis assist in text emotion detection.

2 Prior Works

Prior works that used transformer-based models like Multilingual-BERT and XLM-R to to categorize Tamil YouTube comments into eleven emotions demonstrate how XML-R outperformed all other models with a macro F1-score of 0.33 (Mustakim et al., 2022). These multilingual transformers are also used to detect offensive hate and offensive content in Tamil YouTube comments (Rajalakshmi et al., 2023). The work details how the process of stemming and affix stripping makes a difference by giving better results in BERT inputs, especially in MuRIL. Prior works also focus on aspect-based and (Ganganwar and Rajalakshmi, 2019) and context aware sentiment with attention-enhanced features from bidirectional transformers (Sivakumar and Rajalakshmi, 2022). Various text embedding techniques/traditional algorithms are proposed, particularly for short text classification.(Rajalakshmi, 2014, 2015; Rajalakshmi and Aravindan, 2018; Rajalakshmi et al., 2018; Rajalakshmi and Xaviar, 2017; Rajalakshmi et al., 2020)

Multilingual BERT models like Indic Bert and XLMRoberta are used to detect offensive content in code-mixed Hindi-English tweets. Using them as embedding models with ensemble models as downstream classifiers seem to provide better performance than other classifiers (Rajalakshmi et al., 2021c). BERT based approaches see their usage not only in Tamil but also Arabic tweets. Including emoji in these approaches show an improvement in the performance of the models in identifying hate speech (Althobaiti, 2022). The work also states that the incorporation of textual emoji descriptions as features may enhance or degrade the performance of the models, depending on the number of examples per class and whether emojis are a distinguishing characteristic between classes. In previous works, sentiment analysis and span detection is performed using transformers models in code-mixed languages like Tamil-English and Hindi-English (Ravikiran et al., 2022)(Rajalakshmi et al., 2022c)(Rajalakshmi et al., 2021b)(Kannan et al., 2021).

Emoji embedding is one way to develop features for sentiment analysis tasks and can be seen in works that implement in Bi-LSTM based models for enhanced performace (Liu et al., 2021). Other works in Indian languages like Hindi and Marathi demonstrate the advantage of using XGBoost for multiclass classification (Rajalakshmi et al., 2021a). Sentiment analysis on the English Twitter dataset shows that the inclusion of emojis using TF-IDF as a feature extraction technique shows marginal improvement over excluding the emojis (Yoo and Rayz, 2021). Investigations show that a CNN can be used for emotion detection in Tamil when used with embedding approaches like BoW, TFIDF, Word2vec, fastText, and GloVe (Andrew, 2022). Moreover, there are works which focus on detecting signs of depression using XGBoost and detecting abusive comments in Tamil using transformer models from social media (Rajalakshmi et al., 2022a)(Sharen and Rajalakshmi, 2022).

The amount of emoji usage and the presence of text and emoji in expressing sentiments have been examined using the web documents of well-known male and female celebrities and compares the overall emoji usage among the most popular Twitter users (Gupta et al., 2020). The work demonstrates how sentiment analysis for both text and emoji is more thorough and accurate. Prior works also used other deep learning models like self-attentive LSTM, BiLSTM-CRF and hybrid convolutional bidirectional recurrent neural network for sentiment analysis (Sivakumar and Rajalakshmi, 2021)(Rajalakshmi et al., 2022b)(Soubraylu and Rajalakshmi, 2021).

3 Proposed Methodology

Understanding the function of emojis in Tamil text during sentiment analysis is the primary motivation behind this work. Tamil comments from YouTube are used as the input texts, which are preprocessed and vectorized using TF-IDF and MuRIL. Logistic Regression and

Notation	Emotion	Count	Notation	Emotion	Count
Joy	Joy	585	Ant	Anticipation	73
Neu	Neutral	401	Dis	Disgust	69
Tru	Trust	183	Ang	Anger	59
Lov	Love	143	Sur	Surprise	34
Amb	Ambiguous	139	Fea	Fear	12
Sad	Sadness	120			

Table 1: Dataset description

ensemble models like Random Forest and XGBoost are then trained on the features. Cross-validation is performed on the results and further analysis on the impact of emoji in text is discussed.

3.1 Dataset

The dataset consists of text from 22200 Tamil YouTube comments (Sampath et al., 2022). The text are classified into 11 different emotions: Neutral, Anger, Joy, Disgust, Trust, Anticipation, Ambiguous, Love, Surprise, Sadness and Fear. Some texts contain emojis while most of them don't. Since the primary focus of this research is to understand the role of emojis, the texts without emoji are removed and the dataset is constricted to 1818 texts with atleast one emoji. The description of the dataset and the notations used can be seen in Table 1. It can also be noticed that some texts in the dataset have one emoji, some have multiple emojis while others have the same emoji repeated multiple times.

3.2 Preprocessing

In an attempt to understand the contributions of emoji in a text, two additional input variations are considered for comparison. The first variation has no emoji and the other variation has the emojis name rather than the emoji itself. Figure 1 shows the example of the the input text variations. Another important thing to note down is that, emoji names for emojis with various skin tones are also mentioned down in the text with emoji name column. Further preprocessing is performed on all the columns with text. Stopword removal is done by utilizing the 125 stopwords suggested by Ashok R. (Ashok, 2016). An affix stripping iterative stemming algorithm (Porter, 2001) is used to reduce derivative words to their root form. After this, feature extraction of text takes place.

3.3 Feature Extraction

To vectorize the text data, this research employs two types of feature extraction techniques, TF-IDF and MuRIL. Further, cross-validation is performed on this vectorized data with a K-fold of 5.

3.3.1 TF-IDF

Term FrequencyInverse Document Frequency (TF-IDF) is employed for vectorizing the dataset. TF-IDF determines how pertinent a word is to a corpus or series of words in a text. The frequency of a term in the corpus offsets the way that meaning changes as a word appears more frequently in the text.

3.3.2 MuRIL

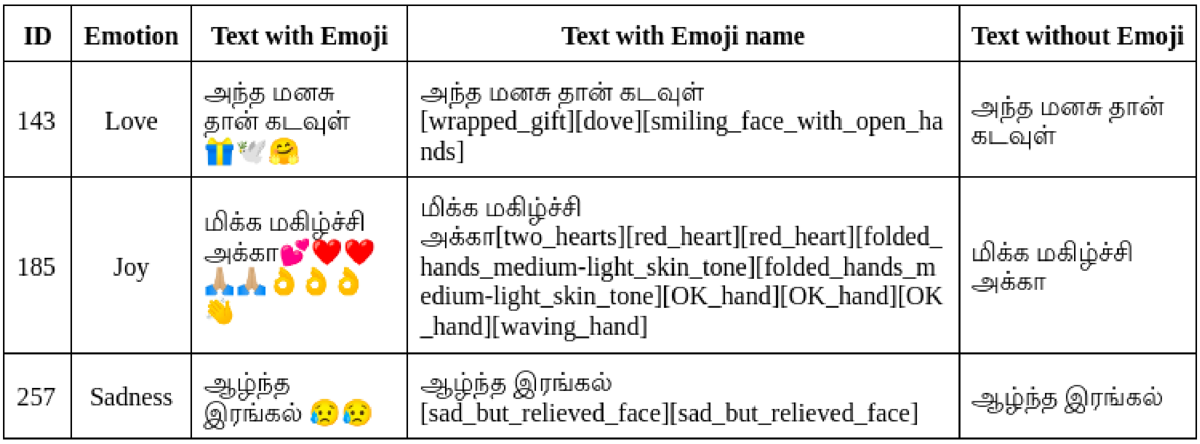
MuRIL is a pre-trained BERT model from Google's Indian research division (Khanuja et al., 2021). It is a multilingual language paradigm that has only been trained on corpora containing English and 16 additional Indian languages, including Tamil. Masked language modeling and translation language modeling are the two stages of training. Here, the MuRIL model is used as an embedding layer.

3.4 Classifiers

In this research, Logistic Regression, Random Forest and XGBoost are utilized to train the vectorized data. Hyperparameter tuning was performed for all classifiers and the parameters are presented in Table 2.

3.4.1 Logistic Regression

It is a machine learning model used for classification. The linear regression model is the source of its development. A logistic function is fitted with the output of the linear regression model to forecast the target variable. In this paradigm, a decision boundary is used. This



ID	Emotion	Text with Emoji	Text with Emoji name	Text without Emoji
143	Love	அந்த மனசு தான் கடவுள் 🎁🕊️🤗	அந்த மனசு தான் கடவுள் [wrapped_gift][dove][smiling_face_with_open_hands]	அந்த மனசு தான் கடவுள்
185	Joy	மிக்க மகிழ்ச்சி அக்கா💕❤️❤️🙏🏼🙏🏼👌👌👌👋	மிக்க மகிழ்ச்சி அக்கா[two_hearts][red_heart][red_heart][folded_hands_medium-light_skin_tone][folded_hands_medium-light_skin_tone][OK_hand][OK_hand][OK_hand][waving_hand]	மிக்க மகிழ்ச்சி அக்கா
257	Sadness	ஆழ்ந்த இரங்கல் 😥😥	ஆழ்ந்த இரங்கல் [sad_but_relieved_face][sad_but_relieved_face]	ஆழ்ந்த இரங்கல்

Figure 1: Examples of input text variations

Classifier	Hyperparameter used
Logistics Regression	’C’: 1, ’dual’: False, ’fit intercept’: False, , ’penalty’: ’l2’, ’solver’:’newton-cg’
Random Forest	’bootstrap’: True, ’class weight’: None, , ’criterion’: ’entropy’, ’max features’: ’log2’, ’n estimators’: 100, ’oob score’: False, ’warm start’: False
XGBoost	’booster’: ’gbtree’, ’grow policy’: ’depthwise’, , ’learning rate’: 0.1, ’max depth’: 6, ’sampling method’: ’uniform’, ’tree method’: ’hist’

Table 2: Hyperparameters used for classifiers

establishes a cutoff point separating one class of variables from another.

3.4.2 Random Forest

It is an ensemble method, which entails combining numerous little decision trees, or estimators, each of which produces its own predictions. The random forest model incorporates the estimators’ predictions to deliver a more precise prediction. Additionally, massive datasets with a variety of dimensions and feature types can be handled by random forests.

3.4.3 XGBoost

The gradient boosting framework is used by the decision tree-based ensemble machine learning method known as XGBoost. The XGBoost classifier is reliable and produces effective results in a variety of distributed situations. It also offers a wrapper class that enables models to be used in the scikit-learn framework as classifiers or regressors.

4 Results and Discussion

In this research, the evaluation metrics taken into account are weighted precision, weighted recall and weighted F1-Score. Since the dataset is imbalanced, weighted metrics are taken into account. While recall measures how effectively the positives are recognized, precision measures how accurately the predictions are made. F1-score is a culmination of the values of precision and recall.

It can be inferred from the Table 3 that both TF-IDF and MuRIL feature extraction methods achieve greater results when used with the XGBoost ensemble model. The XGBoost algorithm builds upon the Random Forest algorithm by introducing gradient boosting. By attempting to minimize error before adding further decision trees, the XGBoost algorithm (Chen and Guestrin, 2016) outperforms the Random Forest algorithm and thus in turn the Logistic Regression algorithm. It is also made abundantly clear that the presence of emojis in the text increases the performance. However, the way in which the emojis are represented also seems to play a role in the performance of the model. It can be noted that text with emoji receives a slightly better F1-score than plain text in both feature extraction scenarios. This may be explained by the fact that the addition of the emojis increases the feature space of the input vector, thus providing more information for the classifiers to train on.

It can be inferred that in both TF-IDF and MuRIL, text with emoji name has the best results. This could account to the fact that, when emojis are converted to its name state, it has more repetitive terms. For instance, the key word "Heart" appears in the phrases "Red Heart❤️", "Growing Heart💗", "Sparkling Heart💖", "Purple Heart💜", "Blue Heart💙", and so forth. The meaning of a heart is the same regardless of how it is shown in an emoji. Every emoji has a distinct unicode. This attributes to the fact that during vectorization, all these emojis are taken as unique features even when they have something in common. This issue is avoided when the emoji is converted to textual format, where more emphasis is given to each word while increasing the models performance.

Pre-trained transformer models generally perform well in NLP tasks. Their ability to do NSP (next sentence prediction) is used to learn the context between words, which can be used in a variety of tasks. Surprisingly Googles MuRIL transformer model trained on multilingual data including Tamil, fails to perform better than its TF-IDF counterpart when used as an embedding layer. The model overemphasizes one particular emotion, leading to all the predictions being that particular emotion and losing generalizability across the other emotions. This might be due to the imbalanced nature of the dataset. This case can also be viewed in Vaishali Ganganwar et al work where they proved that MuRIL showed underperformance due to dataset imbalance for Tamil text (Ganganwar and Rajalakshmi, 2022).

Figure 4 details the six most occurring emojis in the corpus taken and their occurrence across all emotions. Taking a look at the distribution of the emojis one can state that these popularly used emojis though being extensively used in two or three emotions are quite ambiguous and are used in unexpected emotions. The 🤣 emoji sees its main usage in the Joy emotion, which is to be expected as it represents rolling on the floor in laughter. However it also sees use in categories such as Ambiguous and Disgust which can confidently said is not represented by the emoji. From this it can be said that even though emojis can denote the emotion of the author of the text, they cannot be solely relied on and have to be used in combination with the words in the text. This is especially true in the case of sarcasm, where the emoji might denote an emotion which is not interpreted when one reads the entire text.

Another interesting area to draw insights from is the number of times an emoji is repeated. One might assume that the repetition of a single emoji multiple times in a text would be a strong indicator to a particular emotion. However, from table 5 it is evident that it is not the case. It is evident that the frequency of use of 2- and 3-repeating emojis against a single emoji differs. However, the coefficient of variance reveals that their distribution across all emotions is almost the same. To test the validity of this hypothesis that the distribution is same for all occurrences, a two sample t-test was performed on each pair of occurrences. The results of this test in Table 6 show that the p-value is greater than 0.05 for all occurrences. Thus we fail to reject the null hypothesis which signifies that the mean is not affected by the number of times emojis are repeated. This can also be seen in Figure 2 as the curves for the different occurrences have similar patterns even if they differ in magnitude. Thus it is noted that, the frequency of occurrence of an emoji in the text is not indicative of the conveyed emotion.

5 Conclusion and Future works

This study investigates the influence of emojis on the detection of emotion portrayed through Tamil YouTube comments. Text embedding was performed using TF-IDF and the MuRIL pre-trained model, while downstream classifiers included Logistic Regression, Random Forest, and XGBoost. The combination of TF-IDF and XGBoost yielded the best results, with a weighted-average F1 score of 0.32. Replacing the emoji with a word that represents it outperformed expressing it using UTF encoding

Feature Extraction	**Category**	**LR**			**RF**			**XGBoost**		
		P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
TF-IDF	Plain Text	0.24	0.29	0.25	0.23	0.29	0.24	0.25	0.31	**0.25**
	Text + Emoji	0.25	0.30	0.26	0.24	0.30	0.25	0.26	0.32	**0.26**
	Text + Emoji name	0.30	0.36	0.31	0.31	0.37	0.31	0.31	0.36	**0.32**
MuRIL	Plain Text	0.10	0.32	0.16	0.26	0.30	0.21	0.23	0.30	**0.24**
	Text + Emoji	0.10	0.32	0.16	0.26	0.32	0.23	0.25	0.31	**0.25**
	Text + Emoji name	0.10	0.32	0.16	0.23	0.33	0.22	0.29	0.35	**0.28**

Table 3: Performance of different classifiers on TF-IDF and MuRIL. Here, P represents Weighted Precision, R represents Weighted Recall and Weighted F1-Score.

Emoji	**Occ**	**Joy**	**Neu**	**Tru**	**Lov**	**Amb**	**Sad**	**Ant**	**Dis**	**Ang**	**Sur**	**Fea**
🙏	420	176	73	73	52	7	18	14	3	0	2	2
😂	211	86	47	6	3	16	4	8	22	8	9	2
👍	176	74	42	33	7	5	2	6	1	3	2	1
❤️	96	40	13	12	22	2	2	3	0	2	0	1
🤣	96	41	1	3	1	10	1	1	8	6	4	2
😭	88	6	14	3	4	3	49	2	3	1	3	0

Table 4: Occurrence of the 6 most frequently used emojis and their distribution across all predicted emotions

Emoji	**Occ**	**Coeff**	**Joy**	**Neu**	**Tru**	**Lov**	**Amb**	**Sad**	**Ant**	**Dis**	**Ang**	**Sur**	**Fea**
🙏🙏🙏	162	0.1360	94	12	24	24	2	3	3	0	0	0	0
🙏🙏	60	0.1205	31	7	8	5	0	6	1	1	0	1	0
🙏	223	0.1371	131	31	34	13	3	8	1	1	0	0	1
👍👍👍	45	0.1445	28	7	6	1	1	1	0	0	0	1	0
👍👍	27	0.1644	19	6	2	0	0	0	0	0	0	0	0
👍	109	0.1343	61	19	20	2	0	1	3	1	1	0	1
❤️❤️ ❤️	30	0.1346	14	5	0	11	0	0	0	0	0	0	0
❤️❤️	14	0.1229	7	1	3	2	0	0	1	0	0	0	0
❤️	52	0.1334	30	3	7	7	1	1	2	0	1	0	0

Table 5: Analysis on the occurrence of emoji repetition

Test Condition	**t**	**p**	**df**	**Diff**	**95% C.I.**
🙏 vs 🙏🙏	1.2359	0.2308	20	14.82	-10.19 to 39.83
🙏 vs 🙏🙏🙏	0.3854	0.7040	20	5.55	-24.47 to 35.56
🙏🙏 vs 🙏🙏🙏	1.0500	0.3062	20	9.2	-9.15 to 27.69
👍 vs 👍👍	1.2761	0.2165	20	7.45	-4.73 to 19.64
👍 vs 👍👍👍	0.9521	0.3524	20	5.82	-6.93 to 18.57
👍👍 vs 👍👍👍	0.5364	0.5976	20	-1.64	-8.00 to 4.73
❤️ vs ❤️❤️	1.2696	0.2188	20	3.45	-2.22 to 9.13
❤️ vs ❤️❤️❤️	0.6541	0.5205	20	2.00	-4.38 to 8.38
❤️❤️ vs ❤️❤️❤️	0.8716	0.3938	20	-1.45	-4.94 to 2.03

Table 6: Statistics associated with number of times emoji used for expressing emotion intensity.

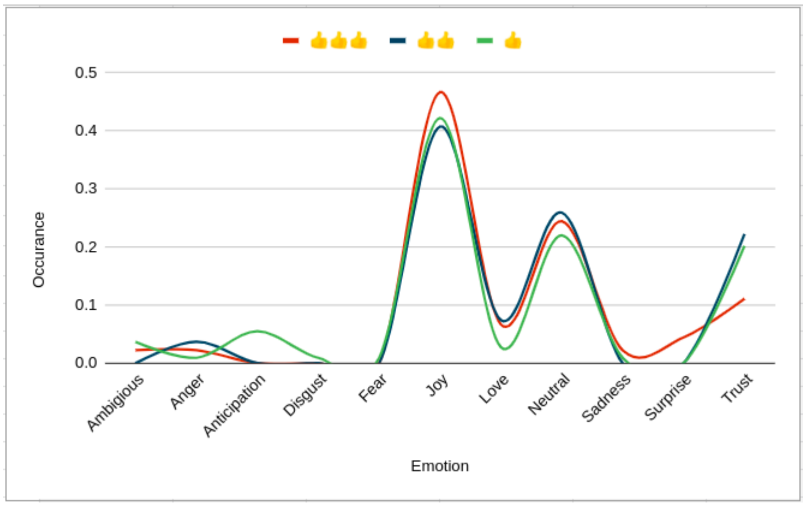


Figure 2: Repetition of 👍 in text and its normalized occurrence across all emotions.

or deleting it entirely from the text. The most often used emojis appear on text that convey an emotion very different to the one indicated by the emojis, demonstrating that one cannot rely just on these emojis to predict the emotion, but rather utilize them in conjunction with the text as has previously been shown useful. Repeated use of an emoji in the same text does not produce a greater link with any particular emotion than a single use of the same emoji as has been proved by a test of significance.

Because the introduction of social media and messaging applications has limited humans to utilizing text and emoticons as the primary mode of communication, this field of research has enormous promise. Emojis can give insight into the emotion that the author wishes to convey, but they can also be deceptive, thus other clues are necessary. More research may be done on the distinct combination of emojis and the emotion that they convey, as well as how they vary if the emojis were present independently. The dataset's imbalance was a big impediment, and working on a balanced dataset might provide better results. A bigger dataset with similar categories of emotions can be employed in future work to generalize findings from the study.

References

Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. 2020. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2(7):e12189.

Maha Jarallah Althobaiti. 2022. Bert-based approach to arabic hate speech and offensive language detection in twitter: Exploiting emojis and sentiment analysis. *International Journal of Advanced Computer Science and Applications*, 13(5).

Judith Jeyafreeda Andrew. 2022. Judith-JeyafreedaAndrew@TamilNLP-ACL2022:CNN for emotion analysis in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 58–63, Dublin, Ireland. Association for Computational Linguistics.

R Ashok. 2016. Tamilnlp. `https://github.com/AshokR/TamilNLP/`.

Bharathi Raja Chakravarthi. 2020. 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*, pages 41–53, Barcelona, Spain (Online). Association for Computational Linguistics.

Bharathi Raja Chakravarthi. 2022a. Hope speech detection in youtube comments. *Social Network Analysis and Mining*, 12(1):1–19.

Bharathi Raja Chakravarthi. 2022b. Multilingual hope speech detection in English and Dravidian languages. *International Journal of Data Science and Analytics*, 14(4):389–406.

Bharathi Raja Chakravarthi, Adeep Hande, Rahul Ponnusamy, Prasanna Kumar Kumaresan, and Ruba Priyadharshini. 2022. How can we detect homophobia and transphobia? experiments in a multilingual code-mixed setting for social media governance. *International Journal of Information Management Data Insights*, 2(2):100119.

Bharathi Raja Chakravarthi and Vigneshwaran Muralidaran. 2021. Findings of the shared task on hope speech detection for equality, diversity, and inclusion. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 61–72, Kyiv. Association for Computational Linguistics.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Anthony Cherbonnier and Nicolas Michinov. 2021. The recognition of emotions beyond facial expressions: Comparing emoticons specifically designed to convey basic emotions with other modes of expression. *Computers in Human Behavior*, 118:106689.

Vaishali Ganganwar and R Rajalakshmi. 2019. Implicit aspect extraction for sentiment analysis: a survey of recent approaches. *Procedia Computer Science*, 165:485–491.

Vaishali Ganganwar and Ratnavel Rajalakshmi. 2022. MTDOT: A Multilingual Translation-based Data augmentation technique for Offensive content identification in Tamil text data. *Electronics*, 11(21).

Shelley Gupta, Archana Singh, and Jayanthi Ranjan. 2020. Sentiment analysis: usage of text and emoji for expressing sentiments. In *Advances in Data and Information Sciences*, pages 477–486. Springer.

Adeep Hande, Siddhanth U Hegde, and Bharathi Raja Chakravarthi. 2022. Multi-task learning in under-resourced dravidian languages. *Journal of Data, Information and Management*, 4(2):137–165.

R Ramesh Kannan, Ratnavel Rajalakshmi, and Lokesh Kumar. 2021. IndicBERT based approach for sentiment analysis on code-mixed Tamil tweets.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Chuchu Liu, Fan Fang, Xu Lin, Tie Cai, Xu Tan, Jianguo Liu, and Xin Lu. 2021. Improving sentiment analysis accuracy with emoji embedding. *Journal of Safety Science and Resilience*, 2(4):246–252.

Nasehatul Mustakim, Rabeya Rabu, Golam Md. Mursalin, Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. CUET-NLP@TamilNLP-ACL2022: Multi-class textual emotion detection from social media using transformer. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 199–206, Dublin, Ireland. Association for Computational Linguistics.

Martin F Porter. 2001. Snowball: A language for stemming algorithms.

R Rajalakshmi, Hans Tiwari, Jay Patel, Ankit Kumar, and R Karthik. 2020. Design of kids-specific URL classifier using Recurrent Convolutional Neural Network. *Procedia Computer Science*, 167:2124–2131.

R Rajalakshmi and Sanju Xaviar. 2017. Experimental study of feature weighting techniques for URL based webpage classification. *Procedia computer science*, 115:218–225.

Ratnavel Rajalakshmi. 2014. Supervised term weighting methods for URL classification. *J. Comput. Sci.*, 10(10):1969–1976.

Ratnavel Rajalakshmi. 2015. Identifying health domain URLs using SVM. In *Proceedings of the Third International Symposium on Women in Computing and Informatics*, pages 203–208.

Ratnavel Rajalakshmi and Chandrabose Aravindan. 2018. An effective and discriminative feature learning for URL based web page classification. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1374–1379. IEEE.

Ratnavel Rajalakshmi, Ankita Duraphe, and Antonette Shibani. 2022a. DLRG@ DravidianLangTech-ACL2022: Abusive comment detection in tamil using multilingual transformer models. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 207–213.

Ratnavel Rajalakshmi, Faerie Mattins, S Srivarshan, and L Preethi Reddy. 2021a. Hate speech and offensive content identification in Hindi and Marathi language tweets using Ensemble techniques. pages 1 – 11. CEUR Workshop Proceedings.

Ratnavel Rajalakshmi, Mohit More, Bhamatipati Shrikriti, Gitansh Saharan, Hanchate Samyuktha, and Sayantan Nandy. 2022b. DLRG@ tamilnlp-acl2022: Offensive span identification in Tamil using BiLSTM-CRF approach. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 248–253.

Ratnavel Rajalakshmi, S Ramraj, and R Ramesh Kannan. 2018. Transfer learning approach for identification of malicious domain names. In *International Symposium on Security in Computing and Communication*, pages 656–666. Springer.

Ratnavel Rajalakshmi, Preethi Reddy, Shreya Khare, and Vaishali Ganganwar. 2022c. Sentimental analysis of code-mixed Hindi language. In *Congress on Intelligent Systems*, pages 739–751. Springer.

Ratnavel Rajalakshmi, Yashwant Reddy, and Lokesh Kumar. 2021b. DLRG@ dravidianlangtech-eacl2021: Transformer based approach for offensive language identification on code-mixed Tamil. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 357–362.

Ratnavel Rajalakshmi, Srivarshan Selvaraj, Faerie Mattins R., Pavitra Vasudevan, and Anand Kumar M. 2023. HOTTEST: Hate and offensive content identification in Tamil using transformers and enhanced STemming. *Computer Speech Language*, 78:101464.

Ratnavel Rajalakshmi, S Srivarshan, Faerie Mattins, E Kaarthik, and Prithvi Seshadri. 2021c. Conversational Hate-offensive detection in code-mixed Hindi-English tweets. pages 1 –11. CEUR Workshop Proceedings.

Manikandan Ravikiran, Bharathi Raja Chakravarthi, Anand Kumar Madasamy, Sangeetha Sivanesan, Ratnavel Rajalakshmi, Sajeetha Thavareesan, Rahul Ponnusamy, and Shankar Mahadevan. 2022. Findings of the shared task on offensive span identification from code-mixed Tamil-English comments. *arXiv preprint arXiv:2205.06118*.

Anbukkarasi Sampath, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, Ruba Priyadharshini, Subalalitha Cn, Kogilavani Shanmugavadivel, Sajeetha Thavareesan, Sathiyaraj Thangasamy, Parameswari Krishnamurthy, Adeep Hande, et al. 2022. Findings of the shared task on emotion analysis in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 279–285.

Kogilavani Shanmugavadivel, Sai Haritha Sampath, Pramod Nandhakumar, Prasath Mahalingam, Malliga Subramanian, Prasanna Kumar Kumaresan, and Ruba Priyadharshini. 2022. An analysis of machine learning models for sentiment analysis of tamil code-mixed data. *Computer Speech & Language*, page 101407.

Herbert Sharen and Ratnavel Rajalakshmi. 2022. DLRG@ LT-EDI-ACL2022: Detecting signs of depression from social media using XGBoost method. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 346–349.

Soubraylu Sivakumar and Ratnavel Rajalakshmi. 2021. Analysis of sentiment on movie reviews using word embedding self-attentive LSTM. *International Journal of Ambient Computing and Intelligence (IJACI)*, 12(2):33–52.

Soubraylu Sivakumar and Ratnavel Rajalakshmi. 2022. Context-aware sentiment analysis with attention-enhanced features from bidirectional transformers. *Social Network Analysis and Mining*, 12(1):1–23.

Sivakumar Soubraylu and Ratnavel Rajalakshmi. 2021. Hybrid convolutional bidirectional recurrent neural network based sentiment analysis on movie reviews. *Computational Intelligence*, 37(2):735–757.

Malliga Subramanian, Rahul Ponnusamy, Sean Benhur, Kogilavani Shanmugavadivel, Adhithiya Ganesan, Deepti Ravi, Gowtham Krishnan Shanmugasundaram, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2022. Offensive language detection in tamil youtube comments by adapters and cross-domain knowledge transfer. *Computer Speech & Language*, 76:101404.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2019. Sentiment analysis in tamil texts: A study on machine learning techniques and feature representation. In *2019 14th Conference on Industrial and Information Systems (ICIIS)*, pages 320–325. IEEE.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020a. Sentiment lexicon expansion using word2vec and fasttext for sentiment prediction in tamil texts. In *2020 Moratuwa Engineering Research Conference (MERCon)*, pages 272–276. IEEE.

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

Byungkyu Yoo and Julia Taylor Rayz. 2021. Understanding emojis for sentiment analysis. In *The International FLAIRS Conference Proceedings*, volume 34.