# CUET-NLP@TamilNLP-ACL2022: Multi-Class Textual Emotion Detection from Social Media using Transformers

**Nasehatul Mustakim**$^{\Psi}$, **Rabeya Akter Rabu**$^{\Psi}$, **Golam Sarwar Md. Mursalin**$^{\Psi}$,
**Eftekhar Hossain**$^{\$}$, **Omar Sharif**$^{\Psi}$ and **Mohammed Moshiul Hoque**$^{\Psi}$
$^{\Psi}$Department of Computer Science and Engineering
$^{\$}$Department of Electronics and Telecommunication Engineering
$^{\$\Psi}$Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
{u1604109, u1604127, u1604014}@student.cuet.ac.bd
{eftekhar.hossain, omar.sharif, moshiul_240}@cuet.ac.bd

## Abstract

Recently, emotion analysis has gained increased attention by NLP researchers due to its various applications in opinion mining, e-commerce, comprehensive search, healthcare, personalized recommendations and online education. Developing an intelligent emotion analysis model is challenging in resource-constrained languages like Tamil. Therefore a shared task is organized to identify the underlying emotion of a given comment expressed in the Tamil language. The paper presents our approach to classifying the textual emotion in Tamil into 11 classes: ambiguous, anger, anticipation, disgust, fear, joy, love, neutral, sadness, surprise and trust. We investigated various machine learning (LR, DT, MNB, SVM), deep learning (CNN, LSTM, BiLSTM) and transformer-based models (Multilingual-BERT, XLM-R). Results reveal that the XLM-R model outdoes all other models by acquiring the highest macro $f_1$-score (0.33).

## 1 Introduction

Textual emotion analysis is the automatic process of specifying a text into an emotion class from predefined connotations (Parvin et al., 2022). With the unprecedented growth of the internet, online and social media platforms significantly influence people's lives and interactions. People share opinions, expressions, information, feelings, emotions, ideas and concerns online (Ghanghor et al., 2021a,b; Yasaswini et al., 2021). People seek emotional support from their relatives, friends, or even virtual platforms when they go through challenging or adverse times (Priyadharshini et al., 2021; Kumaresan et al., 2021; Chakravarthi, 2020; Chakravarthi and Muralidaran, 2021). Textual emotion analysis (TEA) has been proven helpful in various applications, for example, consumer feedback on services and products (Hossain et al., 2021b; Mamun et al., 2022). The positive and negative customer experiences help to assess the demand for products and services (Hossain et al., 2021a). However, one cannot fully express his/her attitude only through positive and negative sentiments. For example – *I threw my iPhone in the water, and now it is not working, so I feel awful* (Sadness) vs *What a pain my new iPhone is not working* (Anger). Both texts express negative sentiment, but the first is sadness, and the latter is considered anger. Thus, emotion analysis is very crucial to understand the actual state of mind (Staiano and Guerini, 2014). In recent years, plenty of research has been conducted to analyze textual emotion. However, low-resource languages (i.e. Tamil and Bengali) remained out of focus, and very few research activities have been conducted to date. This deficiency occurs due to the scarcity of resources, limited corpora and unavailability of text processing tools (Sampath et al., 2022; Ravikiran et al., 2022; Chakravarthi et al., 2021, 2022; Bharathi et al., 2022; Priyadharshini et al., 2022). This shared task paper aims to mitigate this gap by presenting computational models for emotion analysis in Tamil.

Tamil is the predominant language of the majority of people living in Tamil Nadu, Puducherry (in India), and the Northern and Eastern regions of Sri Lanka (Sakuntharaj and Mahesan, 2021, 2017, 2016; Thavareesan and Mahesan, 2019, 2020a,b, 2021). The language is spoken by tiny minority communities in various Indian states such as Karnataka, Andhra Pradesh, Kerala, Maharashtra, and in specific places of Sri Lanka such as Colombo and the hill country. Tamil or varieties of it were widely employed as the main language of governance, literature, and general usage in the state of Kerala until the 15th century AD (Subalalitha, 2019; Srinivasan and Subalalitha, 2019; Narasimhan et al., 2018). Tamil was also commonly employed in inscriptions unearthed in the southern Andhra Pradesh regions of Chittoor and Nellore until the 12th century AD. Tamil was employed for inscriptions in southern Karnataka regions such as Kolar, Mysore,

*Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 199 - 206
May 26, 2022 ©2022 Association for Computational Linguistics

Mandya, and Bangalore from the 10th to 14th century (Anita and Subalalitha, 2019b,a; Subalalitha and Poovammal, 2018).

The significant contribution of this work illustrates in the following:

- Developed transformer-based computation models for classifying emotion in Tamil considering 11 predefined emotion categories.
- Investigated the performance of various machine learning (ML), deep learning (DL) and transformer-based techniques to address the task followed by detailed error analysis.

## 2 Related Work

In the past few years, emotion analysis research has attracted researchers from diverse domains such as computer science, psychology and healthcare. Chaffar and Inkpen (2011) developed a model to recognize six basic emotions from the affective text on ALM's Dataset (1250 texts). They employed several ML techniques where support vector machine (SVM) achieved the highest performance with bag of words (BoW) features. Huang et al. (2019) proposed a contextual model to detect emotion. They combined two LSTM layers hierarchically and formed an ensemble with the BERT model, which achieved 77% accuracy. Vijay et al. (2018) developed a model with SVM and RBF kernel to identify the fear, disgust and surprise emotions from 2866 Hindi-English code-mixed tweets. Wadhawan and Aggarwal (2021) experimented with several DL (CNN, LSTM, BiLSTM) and transformer-based models for recognizing emotions from 149088 Hindi-English code mixed tweets. The transformer-based BERT model outperformed all other techniques and obtained an accuracy of 71.43%. Iqbal et al. (2022) presented a Bengali emotion corpus (BEmoC) containing 7000 texts with six basic emotion categories: *joy, anger, sad, fear, surprise, disgust*. Das et al. (2021) performed an investigation of various ML, DNN, and transformer-based techniques on BEmoD dataset containing 6523 texts. Their results showed that XLM-R outdoes others providing an $f_1$-score of 69.61%. In a similar work, Parvin et al. (2022) implemented various DL techniques (CNN, GRU, BiLSTM) with different ensemble combinations to recognize six emotions from a corpus containing 9000 Bengali texts. The ensemble of CNN and BiLSTM outperformed other models by achieving $f_1$-score of 62.46%.

## 3 Task and Dataset Descriptions

The emotion analysis shared task in Tamil comprises two tasks. We have participated in Task-a, where multi-class categorization of textual emotion is performed. The organizers[1](Sampath et al., 2022) provided the annotated dataset having 11 types of emotions: Ambiguous, Anger, Anticipation, Disgust, Fear, Joy, Love, Neutral, Sadness, Surprise and Trust. The dataset consists of training, validation and test sets containing 14208, 3552 and 4440 texts. Table 1 shows the number of samples for each set in each class that reveals the dataset's imbalanced nature. Very few samples belong to the fear and surprise classes compared to the neutral class.

| **Classes** | **Train** | **Valid** | **Test** |
|---|---|---|---|
| Neutral | 4,841 | 1,222 | 1,538 |
| Joy | 2,134 | 558 | 702 |
| Ambiguous | 1,689 | 437 | 500 |
| Trust | 1,254 | 272 | 377 |
| Disgust | 910 | 210 | 277 |
| Anger | 834 | 184 | 244 |
| Anticipation | 828 | 213 | 271 |
| Sadness | 695 | 191 | 241 |
| Love | 675 | 189 | 196 |
| Surprise | 248 | 53 | 61 |
| Fear | 100 | 23 | 33 |
| Total | 14,208 | 3,552 | 4,440 |

Table 1: Class-wise distribution of Tamil emotion dataset

To get better insights, we further analyzed the training set. Table 2 shows the detailed statistics of the training set after removing inconsistencies from the texts. The neutral class retained the highest number of words and unique words, whereas the fear class had the least. On average, all the classes have ≈8-10 words; however, the texts from joy, love and surprise classes tend to be shorter than other classes.

## 4 Methodology

This work employed four ML, three DL and two transformer-based approaches to identify the underlying emotions of social media comments in

[1] https://competitions.codalab.org/competitions/36396

| Classes | Total words | Unique words | Max. length (words) | Avg. words (per text) |
|---|---|---|---|---|
| Neutral | 37,344 | 17,033 | 169 | 7.7 |
| Joy | 14,624 | 6,746 | 84 | 6.9 |
| Ambiguous | 14,579 | 8,309 | 114 | 8.6 |
| Trust | 11,757 | 6,318 | 110 | 9.3 |
| Disgust | 8,996 | 5,651 | 128 | 9.9 |
| Anger | 7,879 | 5,149 | 116 | 9.4 |
| Anticipation | 8,489 | 5,131 | 86 | 10.3 |
| Sadness | 6,911 | 4,485 | 76 | 9.9 |
| Love | 4,598 | 2,705 | 65 | 6.8 |
| Surprise | 1,633 | 1,362 | 55 | 6.9 |
| Fear | 1,040 | 864 | 108 | 10.4 |

Table 2: Detailed statistics of each class in the training set

Tamil. Initially, the unwanted characters (i.e., numbers, extra space, punctuation and URLs) and stop words are removed from the texts. Afterwards, different feature extraction techniques (i.e., TF-IDF, Word2Vec (Mikolov et al., 2013) extract the textual features. Figure 1 depicts the schematic diagram of the emotion classification system.

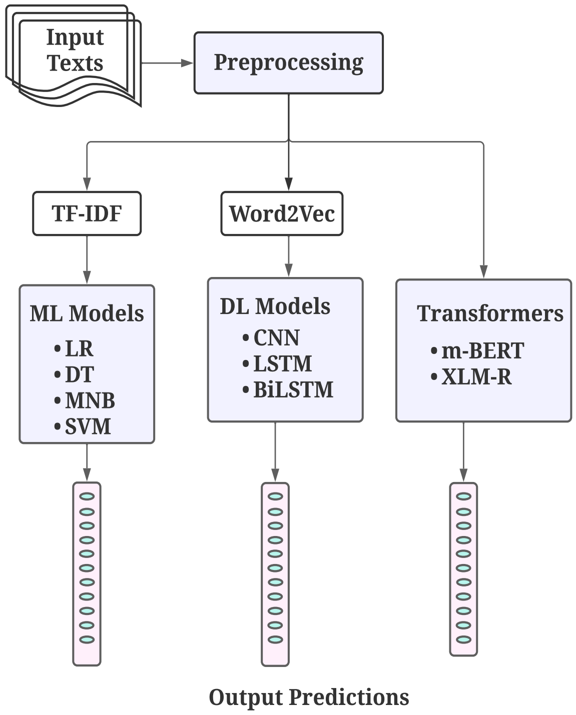


Figure 1: Abstract process of textual emotion classification

### 4.1 Feature Extraction

To train the ML models, we use the TF-IDF values of the unigram and bigram features, where maximum features are settled to 40000. On the other hand, Word2Vec embedding features are used to develop the DL based methods. The Keras embedding layer is applied to generate the embedding vectors of dimension 100.

### 4.2 ML-based Methods

Four traditional ML methods such as logistic regression (LR), decision tree (DT), support vector machine (SVM) and multinomial naive Bayes (MNB) have been employed to accomplish the emotion classification task. The models are implemented by using the 'Scikit-learn'[2] library. The LR model is constructed by setting the regularization parameter C at 10, solver to 'lbfgs' along with a balanced class weight. For the DT model, the 'gini' criterion is used for splitting the nodes. Similarly, in the case of MNB, the smoothing parameter $\alpha$ is fine-tuned at 1.50. For SVM, the 'rbf' kernel is used with a regularization value of 7.

### 4.3 DL-based Methods

This work also employed several DL methods such as CNN, LSTM and BiLSTM to address the task. All the models take word embedding vectors (Word2Vec) as features. We construct a CNN (Kim, 2014) architecture consisting of one convolution layer of 128 filters and a max-pooling layer with a pool size of 2. The flattened output of the pooled layer is then passed to the softmax layer for the classification. Likewise, a layer of LSTM and BiLSTM network of 128 units is developed with a drop-out rate of 0.2 to dissuade the overfitting problem. Finally, the output sentence representation is transferred to the softmax layer for predicting the emotion class. The DL models are implemented by using the Keras library[3] with the TensorFlow (Abadi et al., 2015) backend. 'Adam' (Kingma and Ba, 2014) optimizer with a learning rate of 0.001 is used to compile the models, whereas the 'sparse_categorical_ crossentropy' loss function is used to calculate the errors during the training. We also use the Keras callbacks methods to choose the best intermediate model.

#### 4.3.1 Transformers

Recent advancements in NLP have demonstrated that the transformer-based architecture is superior in solving several classification problems (Puranik et al., 2021; Li et al., 2021; Sharif et al., 2021) irrespective of the language variation. In this work, two

[2] https://scikit-learn.org/stable/

[3] https://keras.io/

| **Hyperparameters** | **CNN** | **LSTM** | **BiLSTM** | **m-BERT** | **XLM-R** |
|---|---|---|---|---|---|
| Input length | 300 | 300 | 300 | 150 | 150 |
| Embedding dimension | 100 | 100 | 100 | - | - |
| Filters (layer 1) | 128 | - | - | - | - |
| Pooling type | max | - | - | - | - |
| Kernel size | 5 | - | - | - | - |
| LSTM units | - | 128 | 128 | - | - |
| Dropout | - | 0.2 | 0.2 | | |
| Optimizer | ‘adam’ | ‘adam’ | ‘adam’ | ‘adam’ | ‘adam’ |
| Learning rate | $1e^{-3}$ | $1e^{-3}$ | $1e^{-3}$ | $2e^{-5}$ | $2e^{-5}$ |
| Epochs | 20 | 3 | 20 | 3 | 5 |
| Batch size | 32 | 32 | 32 | 12 | 12 |

Table 3: Summary of tuned hyperparameters for DL and Transformer-based models

widely used transformer models such as – m-BERT (Devlin et al., 2018) and XLM-R (Conneau et al., 2019) are employed to address the task. Specifically, we culled the ‘bert-base-multilingual-cased’ and ‘xlm-roberta-base’ versions of the models from Huggingface [4] transformers library and fine-tuned them on the dataset. We have trained the models up to five epochs with the help of the Ktrain (Maiya, 2020) package and used the ‘adam’ optimizer with a learning rate of $2e^{-5}$. Table 3 illustrates the various hyperparameters of the developed models.

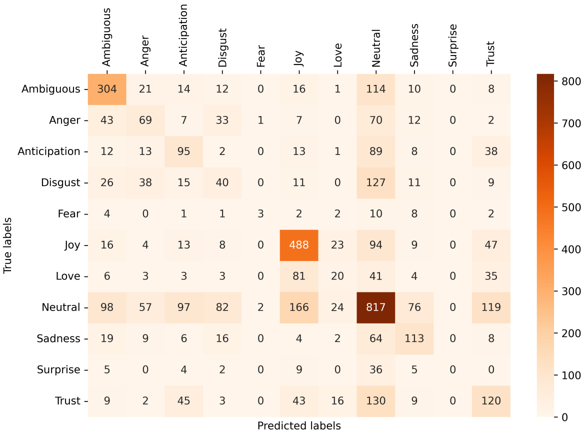
## 5 Results and Analysis

Table 4 reports the performance comparison of the different approaches. The efficacy of the models is determined based on the macro $f_1$-score. It is observed that amid the ML models, LR achieved the highest $f_1$-score of 0.23 while MNB performed poorly on the test set. On the other hand, DL based methods did not surpass the performance of the best ML model ($f_1$-score = 0.23) as both CNN and BiLSTM achieved an identical score of 0.21. However, the transformer model, XLM-R, outperformed all the models by achieving the highest accuracy (0.47), precision (0.36), recall (0.33) and macro $f_1$-score (0.33).

Table 5 shows the class-wise performance of each model in terms of $f_1$-score. The XLM-R model achieved the highest $f_1$-score in seven classes out of eleven as these classes have the most instances in training set. The LR and m-BERT models obtained the highest score in love (0.16) and neutral (0.54) classes, while BiLSTM acquired maximal scores in the remaining classes: fear (0.21) and surprise (0.04).

### 5.1 Error Analysis

Table 4 illustrates that XLM-R acquired the highest score and outperformed all the other approaches. A quantitative error analysis of the best model has been carried out by using the confusion matrix (Figure 2). It is observed that the model identified 817 instances of the ‘neutral’ class correctly and incorrectly reckoned 166 and 119 instances as from ‘joy’ and ‘trust’ emotion class, respectively. Alternatively, it predicted the ‘surprise’ class as ‘neutral’ and ‘joy’ mostly. Furthermore, we noticed that the model becomes confused among the emotions of ‘neutral’, ‘joy’, ‘trust’ and ‘surprise’. The main reason behind this might be the class imbalance problem. There might be plenty of words that are similar for some classes. Apart from this, the number of training texts in the surprise class is only 248, which is inadequate for the model to learn the context effectively. Moreover, the considerable diversity of the Tamil language can also be a potential cause. We have also observed that the most true predictions were made for the neutral and joy class, and an apparent reason for it is that the model saw plenty of texts of that class during the training.



| | Ambiguous | Anger | Anticipation | Disgust | Fear | Joy | Love | Neutral | Sadness | Surprise | Trust |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Ambiguous | 304 | 21 | 14 | 12 | 0 | 16 | 1 | 114 | 10 | 0 | 8 |
| Anger | 43 | 69 | 7 | 33 | 1 | 7 | 0 | 70 | 12 | 0 | 2 |
| Anticipation | 12 | 13 | 95 | 2 | 0 | 13 | 1 | 89 | 8 | 0 | 38 |
| Disgust | 26 | 38 | 15 | 40 | 0 | 11 | 0 | 127 | 11 | 0 | 9 |
| Fear | 4 | 0 | 1 | 1 | 3 | 2 | 2 | 10 | 8 | 0 | 2 |
| Joy | 16 | 4 | 13 | 8 | 0 | 488 | 23 | 94 | 9 | 0 | 47 |
| Love | 6 | 3 | 3 | 3 | 0 | 81 | 20 | 41 | 4 | 0 | 35 |
| Neutral | 98 | 57 | 97 | 82 | 2 | 166 | 24 | 817 | 76 | 0 | 119 |
| Sadness | 19 | 9 | 6 | 16 | 0 | 4 | 2 | 64 | 113 | 0 | 8 |
| Surprise | 5 | 0 | 4 | 2 | 0 | 9 | 0 | 36 | 5 | 0 | 0 |
| Trust | 9 | 2 | 45 | 3 | 0 | 43 | 16 | 130 | 9 | 0 | 120 |

Figure 2: Confusion matrix of the best model (XLM-R)

[4] https://huggingface.co/

| Approach | Classifier | Accuracy | Precision | Recall | $f1$-score |
|---|---|---|---|---|---|
| ML | LR | 0.31 | 0.23 | 0.23 | 0.23 |
| | DT | 0.26 | 0.19 | 0.19 | 0.19 |
| | MNB | 0.38 | 0.11 | 0.16 | 0.08 |
| | SVM | 0.40 | 0.18 | 0.35 | 0.20 |
| DL | CNN | 0.29 | 0.21 | 0.21 | 0.21 |
| | LSTM | 0.35 | 0.09 | 0.03 | 0.05 |
| | BiLSTM | 0.31 | 0.20 | 0.23 | 0.21 |
| Transformers | m-BERT | 0.44 | 0.27 | 0.23 | 0.23 |
| | XLM-R | **0.47** | **0.36** | **0.33** | **0.33** |

Table 4: Performance comparison of various models on the test set

| Classes | LR | DT | SVM | MNB | CNN | LSTM | BiLSTM | m-BERT | XLM-R |
|---|---|---|---|---|---|---|---|---|---|
| Ambiguous | 0.31 | 0.26 | 0.27 | 0.00 | 0.25 | 0.00 | 0.21 | 0.54 | **0.58** |
| Anger | 0.18 | 0.15 | 0.09 | 0.00 | 0.15 | 0.00 | 0.16 | 0.17 | **0.30** |
| Anticipation | 0.17 | 0.14 | 0.09 | 0.00 | 0.17 | 0.00 | 0.16 | 0.30 | **0.33** |
| Disgust | 0.12 | 0.09 | 0.03 | 0.00 | 0.13 | 0.00 | 0.14 | 0.06 | **0.17** |
| Fear | 0.20 | 0.18 | 0.11 | 0.00 | 0.17 | 0.00 | **0.21** | 0.00 | 0.15 |
| Joy | 0.52 | 0.46 | 0.53 | 0.35 | 0.45 | 0.00 | 0.47 | 0.58 | **0.63** |
| Love | **0.16** | 0.11 | 0.11 | 0.00 | 0.15 | 0.00 | 0.14 | 0.08 | 0.14 |
| Neutral | 0.38 | 0.33 | 0.52 | 0.52 | 0.39 | 0.51 | 0.44 | **0.54** | 0.52 |
| Sadness | 0.26 | 0.19 | 0.19 | 0.00 | 0.20 | 0.00 | 0.16 | 0.02 | **0.45** |
| Surprise | 0.00 | 0.00 | 0.03 | 0.00 | 0.02 | 0.00 | **0.04** | 0.00 | 0.00 |
| Trust | 0.24 | 0.18 | 0.19 | 0.03 | 0.21 | 0.00 | 0.19 | 0.21 | **0.31** |

Table 5: Class-wise performance of models in terms of $f_1$-score

## 6 Conclusion

This paper investigated four ML, three DL and two transformer-based models to classify emotion from Tamil texts. Among all models, the XLM-R obtained the highest macro $f_1$-score of 0.33. Since this work did not use any pre-trained embedding, it might adversely affect the performance of the DL model. Thus, we opt to experiment with pre-trained word embedding in the future. Moreover, we plan to explore other advanced transformer-based models (i.e., Indic-BERT, MuRIL) and ensemble approaches to address the emotion analysis task. Since the dataset is imbalanced, it will be interesting to investigate the impact of resampling on the models in the future.

## Acknowledgements

This work was supported by the CUET NLP Lab and Chittagong University of Engineering & Technology, Bangladesh.

## References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

R Anita and CN Subalalitha. 2019a. An approach to cluster Tamil literatures using discourse connectives. In *2019 IEEE 1st International Conference on Energy, Systems and Information Processing (ICESIP)*, pages 1–4. IEEE.

R Anita and CN Subalalitha. 2019b. Building discourse parser for Thirukkural. In *Proceedings of the 16th International Conference on Natural Language Processing*, pages 18–25.

B Bharathi, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, N Sripriya, Arunaggiri Pandian, and Swetha Valli. 2022. Findings of the shared task on Speech Recognition for Vulnerable Individuals in Tamil. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Soumaya Chaffar and Diana Inkpen. 2011. Using a heterogeneous dataset for emotion analysis in text. In *Canadian conference on artificial intelligence*, pages 62–67. Springer.

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 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.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John Phillip McCrae, Paul Buitaleer, Prasanna Kumar Kumaresan, and Rahul Ponnusamy. 2022. Findings of the shared task on Homophobia Transphobia Detection in Social Media Comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Kayalvizhi Sampath, Durairaj Thenmozhi, Sathiyaraj Thangasamy, Rajendran Nallathambi, and John Phillip McCrae. 2021. Dataset for identification of homophobia and transophobia in multilingual YouTube comments. *arXiv preprint arXiv:2109.00227*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Avishek Das, Omar Sharif, Mohammed Moshiul Hoque, and Iqbal H. Sarker. 2021. Emotion classification in a resource constrained language using transformer-based approach. pages 150–158.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Nikhil Ghanghor, Parameswari Krishnamurthy, Sajeetha Thavareesan, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2021a. IIITK@DravidianLangTech-EACL2021: Offensive language identification and meme classification in Tamil, Malayalam and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 222–229, Kyiv. Association for Computational Linguistics.

Nikhil Ghanghor, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021b. IIITK@LT-EDI-EACL2021: Hope speech detection for equality, diversity, and inclusion in Tamil , Malayalam and English. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 197–203, Kyiv. Association for Computational Linguistics.

Eftekhar Hossain, Omar Sharif, Mohammed Moshiul Hoque, and Iqbal H. Sarker. 2021a. Sentilstm: A deep learning approach for sentiment analysis of restaurant reviews. In *Hybrid Intelligent Systems*, pages 193–203, Cham. Springer International Publishing.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2021b. Sentiment polarity detection on bengali book reviews using multinomial naïve bayes. In *Progress in Advanced Computing and Intelligent Engineering*, pages 281–292, Singapore. Springer Singapore.

Chenyang Huang, Amine Trabelsi, and Osmar R Zaïane. 2019. Ana at semeval-2019 task 3: Contextual emotion detection in conversations through hierarchical lstms and bert. *arXiv preprint arXiv:1904.00132*.

MD Iqbal, Avishek Das, Omar Sharif, Mohammed Moshiul Hoque, and Iqbal H Sarker. 2022. Bemoc: A corpus for identifying emotion in bengali texts. *SN Computer Science*, 3(2):1–17.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Prasanna Kumar Kumaresan, Ratnasingam Sakuntharaj, Sajeetha Thavareesan, Subalalitha Navaneethakrishnan, Anand Kumar Madasamy, Bharathi Raja Chakravarthi, and John P McCrae. 2021. Findings of shared task on offensive language identification in Tamil and Malayalam. In *Forum for Information Retrieval Evaluation*, pages 16–18.

Xiangyang Li, Yu Xia, Xiang Long, Zheng Li, and Sujian Li. 2021. Exploring text-transformers in aaai 2021 shared task: Covid-19 fake news detection in english. In *International Workshop on Combating On line Ho st ile Posts in Regional Languages dur ing Emerge ncy Si tuation*, pages 106–115. Springer.

Arun S Maiya. 2020. ktrain: A low-code library for augmented machine learning. *arXiv preprint arXiv:2004.10703*.

Md Mashiur Rahaman Mamun, Omar Sharif, and Mohammed Moshiul Hoque. 2022. Classification of textual sentiment using ensemble technique. *SN Computer Science*, 3(1):1–13.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality.

Anitha Narasimhan, Aarthy Anandan, Madhan Karky, and CN Subalalitha. 2018. Porul: Option generation and selection and scoring algorithms for a tamil flash card game. *International Journal of Cognitive and Language Sciences*, 12(2):225–228.

Tanzia Parvin, Omar Sharif, and Mohammed Moshiul Hoque. 2022. Multi-class textual emotion categorization using ensemble of convolutional and recurrent neural network. *SN Computer Science*, 3(1):1–10.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumar Kumaresan. 2022. Findings of the shared task on Abusive Comment Detection in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Sajeetha Thavareesan, Dhivya Chinnappa, Durairaj Thenmozhi, and Rahul Ponnusamy. 2021. Overview of the DravidianCodeMix 2021 shared task on sentiment detection in Tamil, Malayalam, and Kannada. In *Forum for Information Retrieval Evaluation*, pages 4–6.

Karthik Puranik, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. 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*, pages 98–106.

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 in code-mixed Tamil-English comments. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ratnasingam Sakuntharaj and Sinnathamby Mahesan. 2016. A novel hybrid approach to detect and correct spelling in Tamil text. In *2016 IEEE International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 1–6.

Ratnasingam Sakuntharaj and Sinnathamby Mahesan. 2017. Use of a novel hash-table for speeding-up suggestions for misspelt Tamil words. In *2017 IEEE International Conference on Industrial and Information Systems (ICIIS)*, pages 1–5.

Ratnasingam Sakuntharaj and Sinnathamby Mahesan. 2021. Missing word detection and correction based on context of Tamil sentences using n-grams. In *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 42–47.

Anbukkarasi Sampath, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, Ruba Priyadharshini, Subalalitha Chinnaudayar Navaneethakrishnan, Kogilavani Shanmugavadivel, Sajeetha Thavareesan, Sathiyaraj Thangasamy, Parameswari Krishnamurthy, Adeep Hande, Sean Benhur, Kishor Kumar Ponnusamy, and Santhiya Pandiyan. 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*. Association for Computational Linguistics.

Omar Sharif, Eftekhar Hossain, and Mohammed Moshiul Hoque. 2021. NLP-CUET@DravidianLangTech-EACL2021: Offensive language detection from multilingual code-mixed text using transformers. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 255–261, Kyiv. Association for Computational Linguistics.

R Srinivasan and CN Subalalitha. 2019. Automated named entity recognition from tamil documents. In *2019 IEEE 1st International Conference on Energy, Systems and Information Processing (ICESIP)*, pages 1–5. IEEE.

Jacopo Staiano and Marco Guerini. 2014. Depechemood: a lexicon for emotion analysis from crowd-annotated news. *arXiv preprint arXiv:1405.1605*.

C. N. Subalalitha. 2019. Information extraction framework for Kurunthogai. *Sādhanā*, 44(7):156.

CN Subalalitha and E Poovammal. 2018. Automatic bilingual dictionary construction for Tirukural. *Applied Artificial Intelligence*, 32(6):558–567.

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.

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.

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.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2021. Sentiment analysis in Tamil texts using k-means and k-nearest neighbour. In *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 48–53.

Deepanshu Vijay, Aditya Bohra, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. Corpus creation and emotion prediction for hindi-english code-mixed social media text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 128–135.

Anshul Wadhawan and Akshita Aggarwal. 2021. Towards emotion recognition in hindi-english code-mixed data: A transformer based approach. *arXiv preprint arXiv:2102.09943*.

Konthala Yasaswini, Karthik Puranik, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. 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*, pages 187–194, Kyiv. Association for Computational Linguistics.