# PANDAS@TamilNLP-ACL2022: Emotion Analysis in Tamil Text using Language Agnostic Embeddings

**Krithika S**
SSN College of Engineering
krithika2010039@ssn.edu.in

**Divyasri K**
SSN College of Engineering
divyasri2011037@ssn.edu.in

**Gayathri G L**
SSN College of Engineering
gayathri2010090@ssn.edu.in

**Durairaj Thenmozhi**
SSN College of Engineering
theni_d@ssn.edu.in

**B. Bharathi**
SSN College of Engineering
bharathib@ssn.edu.in

**B. Senthilkumar**
SSN College of Engineering
senthil@ssn.edu.in

## Abstract

As the world around us continues to become increasingly digital, it has been acknowledged that there is a growing need for emotion analysis of social media content. The task of identifying the emotion in a given text has many practical applications ranging from screening public health to business and management. In this paper, we propose a language agnostic model that focuses on emotion analysis in Tamil text. Our experiments yielded an F1-score of 0.010.

## 1 Introduction

Over the past decade, advances in technology have progressed at an extraordinary rate, which in turn has rapidly transformed our methods of communication, expedited by the need for all institutions and establishments to 'go digital'. People have adapted to online modes of communication such as social networking sites and online discussion forums. These platforms have many advantages such as the ability to bring people with similar passions together and enable them to exchange their views, or the capacity to allow people to rally together for a common cause. It would be useful for these platforms to identify people with common interests by filtering through their comments. On the other hand, there are also some disadvantages that arise from the abuse of these tools, which are not limited to, but include, the posting of inappropriate or hurtful comments (O'Keeffe et al., 2011), (Gao et al., 2020). To ensure moderation in content, it is necessary to monitor posts and comments published on various social media (Naslund JA, 2020). Research suggests that monitoring social media can be useful in surveying, understanding and predicting public health (Chancellor and Choudhury, 2020), (Aiello et al., 2020), (Brenda K. Wiederhold, 2020). Social media analytics can also aid in enterprise management (Lee, 2018) and national security (Sykora et al., 2013). A commonly used method to monitor social media is emotion analysis.

Emotions are subjective mental states that are brought about as reactions to our thoughts or memories, or as reactions to external events in our surroundings. Speech and text are both generally associated with an emotion of some kind - joy, sorrow, fear, anger, etc. Classifying a given text into one of the many categories of emotions is a constructive way to analyze and obtain some understanding of the text (Kim and Klinger, 2018). Such textual emotion analysis has many practical applications. For example, Unilever analyzes the emotional expressions of prospective candidates for its entry-level jobs, which helps in saving a significant amount of time during the candidate screening process.

The task of Emotion Analysis in Tamil - DravidianLangTech@ACL 2022[1] (Sampath et al., 2022) aims to classify a set of given comments by predicting the probable emotions associated with each one. A particular trial faced here is that of the dataset being in the Tamil language, for which comparatively less resources are available. In this paper, we have used a Language-Agnostic Sentence Embedder called LaBSE, a popular multilingual BERT embedding model to identify the emotions in Tamil text.

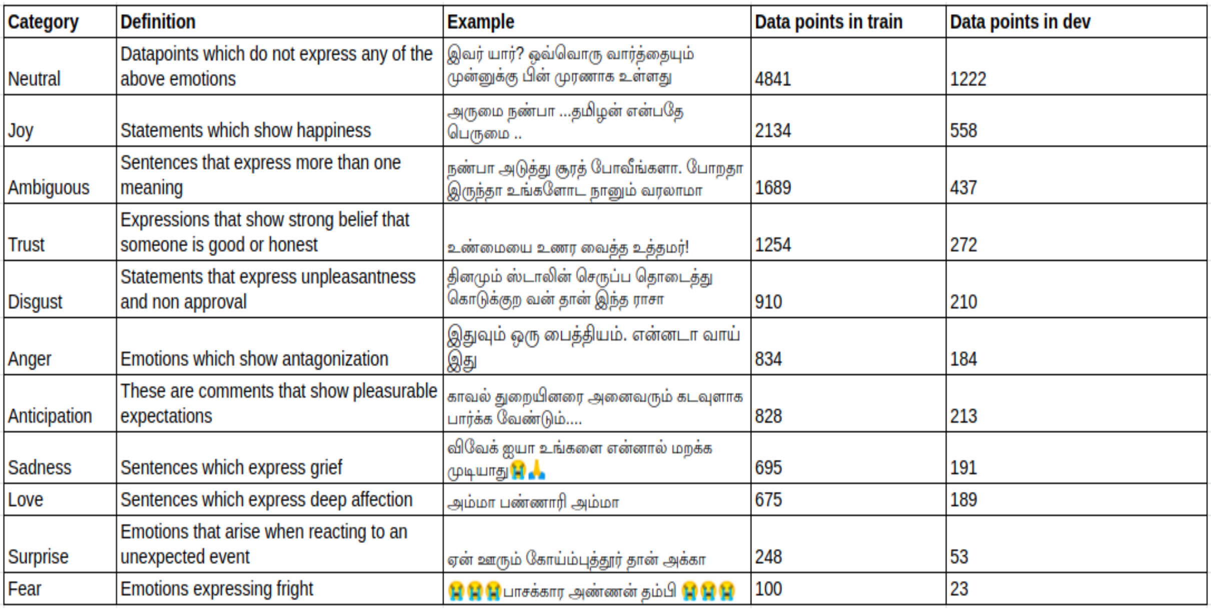
The rest of the paper is organized as follows: section 2 outlines any related works ascertained by a literature survey; section 3 provides a description of the dataset; section 4 details the methodology used for this task; section 5 discusses the results and in section 6, a conclusion is put forward.

## 2 Related Work

The field of Emotion Analysis allows for a multitude of approaches to be used, some of which have been documented in prominent literature. With the recent research and development in speech to text concepts, various researchers have gone about building and testing fast and accurate models for

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

*Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 105 - 111
May 26, 2022 ©2022 Association for Computational Linguistics



| Category | Definition | Example | Data points in train | Data points in dev |
|---|---|---|---|---|
| Neutral | Datapoints which do not express any of the above emotions | இவர் யார்? ஒவ்வொரு வார்த்தையும் முன்னுக்கு பின் முரணாக உள்ளது | 4841 | 1222 |
| Joy | Statements which show happiness | அருமை நண்பா ...தமிழன் என்பதே பெருமை .. | 2134 | 558 |
| Ambiguous | Sentences that express more than one meaning | நண்பா அடுத்து சூரத் போவீங்களா. போறதா இருந்தா உங்களோட நானும் வரலாமா | 1689 | 437 |
| Trust | Expressions that show strong belief that someone is good or honest | உண்மையை உணர வைத்த உத்தமர்! | 1254 | 272 |
| Disgust | Statements that express unpleasantness and non approval | தினமும் ஸ்டாலின் செருப்ப தொடைத்து கொடுக்குற வன் தான் இந்த ராசா | 910 | 210 |
| Anger | Emotions which show antagonization | இதுவும் ஒரு பைத்தியம். என்னடா வாய் இது | 834 | 184 |
| Anticipation | These are comments that show pleasurable expectations | காவல் துறையினரை அனைவரும் கடவுளாக பார்க்க வேண்டும்.... | 828 | 213 |
| Sadness | Sentences which express grief | விவேக் ஐயா உங்களை என்னால் மறக்க முடியாது😭🙏 | 695 | 191 |
| Love | Sentences which express deep affection | அம்மா பண்ணாரி அம்மா | 675 | 189 |
| Surprise | Emotions that arise when reacting to an unexpected event | ஏன் ஊரும் கோய்ம்புத்தூர் தான் அக்கா | 248 | 53 |
| Fear | Emotions expressing fright | 😭😭😭பாசக்கார அண்ணன் தம்பி 😭😭😭 | 100 | 23 |

Table 1: Data distribution

sentiment analysis, especially using inputs from Indian languages. One such work was done by (Uma and Kausika) (V.Uma et al., 2016) by applying the SVM model on an independent corpus of Tamil and English tweets, which were segregated into positive, neutral and negative labels.

Anand Kumar Madasamy, Soman Kotti Padannayil (Seshadri et al., 2016) formulated a tri-layer RNN for the SAIL task of classifying tweets in Indian languages with 1500 lines of Tamil, Hindi and Bengali, which was applied on the Indian textual tweets. This was compared with a Naive Bayes model given by the SAIL task, which gave a significantly lesser accuracy.

Sajeetha Thavareesan and Sinnathamby Mahesan (Thavareesan and Mahesan, 2021) modeled 5 different approaches for sentiment analysis of Tamil texts. They experimented on the UJ_Corpus_Opinions and SAIL-2015 corpus .They extracted the features using TF, BoW , TF-IDF , Word2vec and fastText. They subsequently used Lexicon based and ML based approaches (using SVM , Extreme Gradient Boost EGB, Random Forest RF, Neural Network NN, Linear Regression LR, k nearest Neighbours kNN). They observed that feature extraction using fastText and EGB outperformed the rest. Xiaotian Lin et al (Lin et al., 2021), performed a multilingual text classification - classification into positive, negative ,neutral and mixed emotions using a plain laBSE model, which was comparatively better than the models XLM, XLM RoBERTa and Multilingual BERT. They used the MLM strategy to achieve the desired result. A research work done by Niveditha et al (Nivedhitha et al., 2016) modeled an unsupervised approach on the 2015 SAIL dataset to feature extraction using SentiWordNet and Word2vec embeddings. Kamal et al. (Sarkar, 2015) participated in the SAIL shared task, which included classifying tweets given in the Hindi and Bengali languages. They processed the tweets with the emoticons and used Multinomial Naive Bayes model. For opinion mining in Hindi (into positive, negative, neutral), they used POS tagging in which adjectives were analyzed to perform the task of mining.

Other state of the art models include CNN, RNN, BiLSTM models and ML techniques implemented on these embeddings.

The conclusion from the aforementioned literature is that LaBSE shows encouraging results in feature extraction, and such datasets comprising Tamil and other subcontinental languages are best served by this transformer. Emotional Analysis and classification of emotion is found to be done most effectively by SVM classifier. In summation, a model which incorporates elements of SVM and LaBSE can be expected to be a good approach for this ACL task, and it is also a novel outlook, which has not been examined by the authors listed above.

## 3 Dataset

The dataset under consideration for the task consists of comments made by YouTube users in the

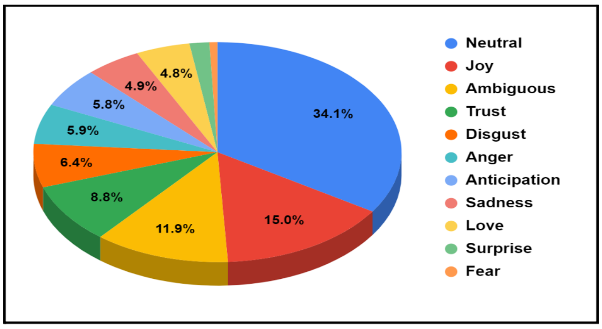


Figure 1: Data distribution of the Training dataset

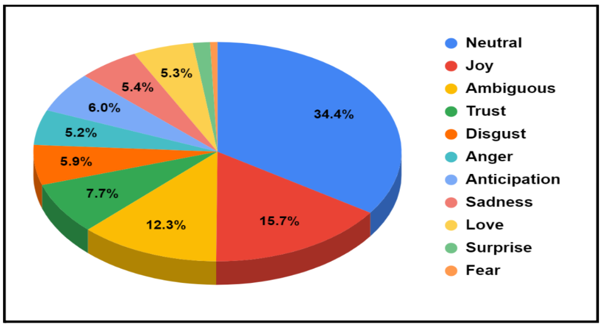


Figure 2: Data distribution of the Development dataset

Tamil language. One of the Dravidian languages and predominantly spoken in Tamil Nadu (India), Tamil has a unique script and an alphabet that is made up of 12 vowels and 18 consonants that when combined in various ways, can give rise to about 216 compound characters.

The comments in this dataset are grouped under 11 different categories based on the emotion that each conveys, as is shown in Table 1. Figure 1 and Figure 2 depict the variance in the training and development datasets. The comments in the training dataset have an average length of 1.18 sentences, with the longest comment having 13 sentences. The average word count per comment is 9.7.

## 4 Methodology

The proposed methodology for this task includes extracting structural features from the processed data and applying classifier models to them. A schematic diagram illustrating the procedure is given in Figure 3.

### 4.1 Preprocessing

Any given raw dataset may contain inconsistencies in its data or may contain some unnecessary data known as noise. Before feeding the data to the required algorithm, it is therefore important to clean the dataset. This process of cleaning the data is known as preprocessing and involves a series of steps. The procedure adopted in this task is as follows:

1. Checking for inconsistencies in the dataset: Many models cannot be trained if any inconsistencies, such as empty rows or mismatched values, are present in the dataset. These anomalies were first removed from the dataset.

2. Removal of punctuation and special characters: The model used focuses on identifying words in the text and creating a corpus of the most frequent words in every category of text in the dataset. Punctuation and special characters interfere with this process and hence, they were removed from the text using a list of punctuation marks from the string library and a custom-made list of special characters. In this case, emoticons were also considered to be special characters and have been removed from the text.

3. Transformation of the data: In this step, the text is converted into a form suitable for the mining process. To establish uniformity in the data and thereby reduce misinterpretation, the text was normalized by the conversion of all text to lowercase.

4. Reduction of the data: In any text, there is a considerable amount of fillers or stop words, i.e., words that do not convey any information necessary for the task of analyzing the text. These words may be important for the grammar of the language, but are redundant in the mining process. Such words have been removed from the text using a custom-made list of stop words in Tamil.

5. Balancing of the dataset: As is observed from the dataset, there is an imbalance in the distribution of data in the training dataset. This can lead to huge inaccuracies in the predicted results. To balance out the data, Synthetic Minority Oversampling Technique (or SMOTE)[2] was used. It is a statistical oversampling technique that helps to overcome an imbalance in data by generating synthetic data for the

[2] https://towardsdatascience.com/machine-learning-multiclass-classification-with-imbalanced-data-set-29f6a177c1a

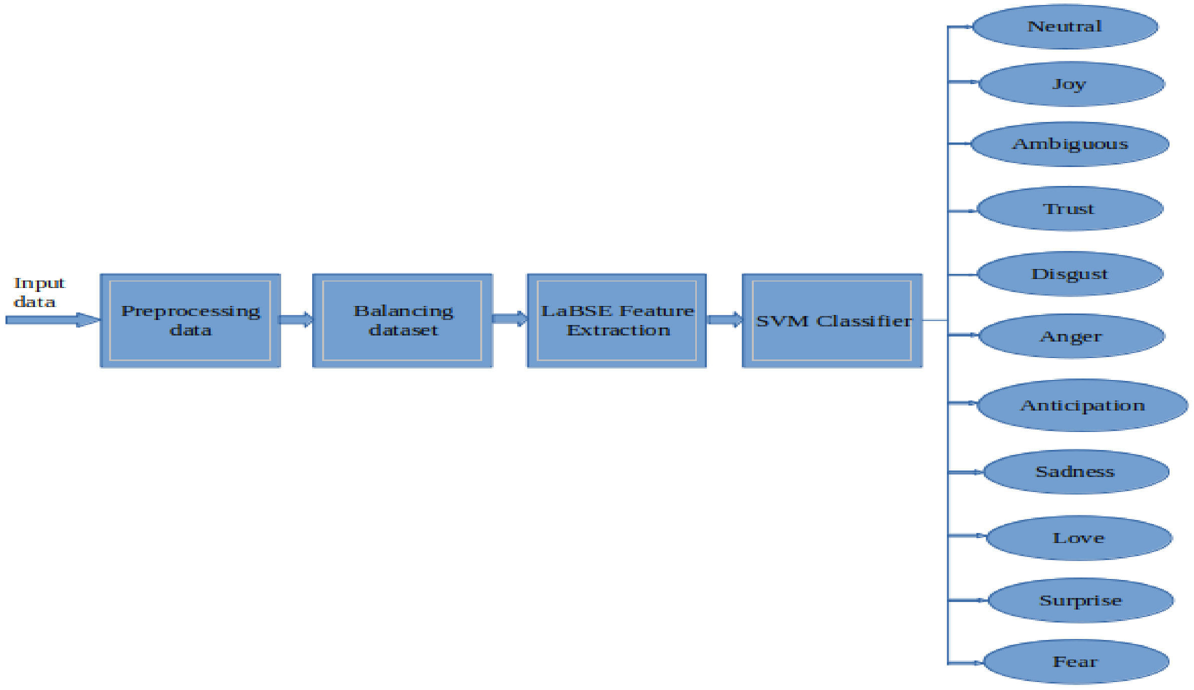


Figure 3: Schematic diagram of the methodology

minority categories in the dataset. It utilizes a k-nearest neighbors algorithm by choosing a minority input vector and adding a new datapoint anywhere between it and one of its nearest neighbors. The process is repeated until the dataset is balanced.

### 4.2 Embeddings and Feature Extraction

For analyzing text, embedding is used to represent words in the form of real-valued vectors that encode the meaning of the words with the intention that words which are expected to have similar meanings are grouped together.

A feature is a characteristic or property by which a given text can be measured or quantified. Raw data is complex and contains a vast number of features, which makes the process of training a model on the dataset cumbersome. Feature extraction reduces the number of dimensions required to define a large dataset by creating a smaller set of new features and rejecting the larger number of existing ones. In this stage, raw data is transformed into numerical features that can be further processed.

#### 4.2.1 LaBSE feature extraction

Language-Agnostic BERT Sentence Embedding, or LaBSE, is a multilingual language model developed by Google, based on the BERT model. It performs tokenization using Wordpiece, the subword-based tokenization algorithm.

LaBSE is a dual encoder model, with each of its two encoders encoding source and target sentences independently, which are then fed to a scoring function to rank them based on their similarity. This latest technique for sentence embedding encodes sentences into a shared embedding space wherein similar sentences are stored next to each other.

LaBSE is currently a popular model for feature extraction. (Rodríguez et al., 2021) used LaBSE both for feature extraction and as an end to end model for classification and reported that its usage improves the performance for both mono-lingual and cross-lingual sets of data.

For this task, we used LaBSE for embedding the preprocessed data, which was then passed to a Classifier model for classification of the given text based on the emotions associated with them. We used the default parameters for the laBSE model with the learning rate set to 0.001. The model includes 630 dense layers and 1 sigmoid layer.

### 4.3 Models applied

The models we experimented on for this task include the SVM Classifer and some simple transformers like LaBSE and IndicBERT[3]. After some consideration, we decided on implementing a model that combines LaBSE feature extraction with the SVM Classifier.

[3] https://github.com/AI4Bharat/indic-bert/blob/master

#### 4.3.1 SVM Classifier

Support Vector Machine, or SVM, is a supervised machine learning algorithm that is widely used for classification-based problems (Yang et al., 2015). It works by mapping data to a high-dimensional feature space in which the data points can be easily categorized. Scaling up the dimensionality greatly contributes to the probability of the data being classified accurately, even when linear separation of the data is not possible.

### 4.4 Experimentation with SMOTE

Since the dataset is highly biased towards 'Neutral' text, we considered the effects of balancing out the data before training the model. The SMOTE technique is used for making a dataset balanced, as discussed above in subsection 4.1. We have tabulated our results for the classification model in Table 2 to compare the case in which SMOTE was used with the case in which it is not used.

| Model | Precision | Recall | F1 | Accuracy |
|---|---|---|---|---|
| LaBSE for feature extraction with SVM Classifier without SMOTE preprocessing | 0.40 | 0.23 | 0.25 | 0.43 |
| LaBSE for feature extraction with SVM Classifier after SMOTE preprocessing | 0.40 | 0.25 | 0.27 | 0.44 |

Table 2: Training results for the models under consideration

#### 4.4.1 LaBSE embedding with SVM Classifier

The data, stripped of special characters, stop words and inconsistencies, was passed through a LaBSE embedding model for feature extraction. The output was then given to an SVM Classifier and the results were recorded.

#### 4.4.2 LaBSE embedding with SVM Classifier and SMOTE preprocessing

The data, preprocessed with an additional step of balancing out the dataset using an oversampling technique (SMOTE), was allowed to undergo feature extraction using LaBSE, followed by classification of the text using an SVM Classifier.

## 5 Results and Analysis

It is apparent that the results are slightly better when SMOTE is used along with the classification model, although the dataset seems to be so highly unbalanced that there is very little difference between the two results.

### 5.1 Performance metrics

This task is evaluated on the macro averages of three performance metrics - Precision, Recall and F1-score[4]. The metrics are computed separately for each class and then the scores are averaged to ensure equal priority for each performance class.

In classification, precision refers to the probability that a classification has been performed accurately. It is the ratio of correctly classified data points to the total number of data points that have been predicted to be of that class.

$$Precision = \frac{TP}{TP + FP}$$

Recall gives some measure of the number of classifications belonging to a particular category that are performed without error. It is the ratio of the correctly classified points of a particular class to the sum of the correctly and incorrectly classified points of the same class.

$$Recall = \frac{TP}{TP + FN}$$

F1-score is the weighted average of precision and recall, and is often used when a balance of both these metrics is needed or when a large class imbalance is encountered.

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

| Team | Precision | Recall | F1 | Rank |
|---|---|---|---|---|
| CUET16 | 0.220 | 0.250 | 0.210 | 1 |
| GJG_Emotion Analysis_taskA | 0.110 | 0.160 | 0.050 | 2 |
| MSDBLSTM _TamilData | 0.090 | 0.080 | 0.050 | 2 |
| MSD | 0.090 | 0.100 | 0.040 | 2 |
| **pandas_tamil** | **0.080** | **0.070** | **0.010** | **8** |

Table 3: Performance results for the Emotion Analysis task

[4] https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

### 5.2 Results

The development dataset was used for the evaluating the performance of the models after training them. The final performance results on the test dataset for the task are recorded in Table 3.

For the given dataset, LaBSE feature extraction along with the SVM Classifier yielded better results than other models that were experimented with. The accuracy was slightly increased when the data was further preprocessed using the SMOTE technique.

Our submission secured the 8th rank in Task B, i.e., Emotion Analysis on a Tamil dataset. Our model procured an F1-Score of 0.010, a Precision score of 0.080 and a Recall score of 0.070.

## 6 Conclusion

In this research paper, we have presented a multilingual transformer model for the emotion analysis of Tamil text as required by the DravidianTechLang ACL 2022 shared task. LaBSE, a pre-trained language agnostic BERT model, was found to perform comparatively well on the Tamil dataset. This model yielded an F1-score of 0.010 on the given dataset. We believe that these results can be improved upon highly by using custom embeddings, based on statistical analysis of the language, to process the data before training the model.

## References

Allison E. Aiello, Audrey Renson, and Paul N. Zivich. 2020. Social media– and internet-based disease surveillance for public health. *Annual Review of Public Health*, 41(1):101–118. PMID: 31905322.

MBA BCB BCN Brenda K. Wiederhold, PhD. 2020. Cyberpsychology, behavior, and social networking. *International Association of CyberPsychology, Training, and Rehabilitation (iACToR)*, 25.

Stevie Chancellor and Munmun Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *npj Digital Medicine*, 3.

Junling Gao, Pinpin Zheng, Yingnan Jia, Hao Chen, Yimeng Mao, Suhong Chen, Yi Wang, Hua Fu, and Junming Dai. 2020. Mental health problems and social media exposure during covid-19 outbreak. *PLOS ONE*, 15(4):1–10.

Evgeny Kim and Roman Klinger. 2018. A survey on sentiment and emotion analysis for computational literary studies. *CoRR*, abs/1808.03137.

In Lee. 2018. Social media analytics for enterprises: Typology, methods, and processes. *Business Horizons*, 61.

Xiaotian Lin, Nankai Lin, Kanoksak Wattanachote, Shengyi Jiang, and Lianxi Wang. 2021. Multilingual text classification for dravidian languages. *CoRR*, abs/2112.01705.

Torous J Aschbrenner KA Naslund JA, Bondre A. 2020. Social Media and Mental Health: Benefits, Risks, and Opportunities for Research and Practice. *J Technol Behav Sci.*, (12):245–257.

E. Nivedhitha, Shinde Pooja Sanjay, M. Anand Kumar, and K. P. Soman. 2016. Unsupervised word embedding based polarity detection for tamil tweets. *Control theory & applications*, 9.

Gwenn Schurgin O'Keeffe, Kathleen Clarke-Pearson, Council on Communications, and Media. 2011. The Impact of Social Media on Children, Adolescents, and Families. *Pediatrics*, 127(4):800–804.

Sebastián E. Rodríguez, Héctor Allende-Cid, and Héctor Allende. 2021. Detecting hate speech in cross-lingual and multi-lingual settings using language agnostic representations. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications - 25th Iberoamerican Congress, CIARP 2021, Revised Selected Papers*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 77–87. Springer Science and Business Media Deutschland GmbH. Funding Information: Acknowledgment. This work was supported in part by Basal Project AFB 1800082, in part by Project DGIIP-UTFSM PI-LIR-2020-17. Héctor Allende-Cid work is supported by PUCV VRIEA. Publisher Copyright: © 2021, Springer Nature Switzerland AG.; null ; Conference date: 10-05-2021 Through 13-05-2021.

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

Kamal Sarkar. 2015. A sentiment analysis system for indian language tweets. volume 9468.

S. Seshadri, M. Kumar, and Soman Kp. 2016. Analyzing sentiment in indian languages micro text using recurrent neural network. 7:313–318.

Martin Sykora, Tom Jackson, Ann O'Brien, and Suzanne Elayan. 2013. National security and social media monitoring: a presentation of the emotive and related systems.

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.

Samir E. Abdelrahman V.Uma, N. Kaushikaa, Umar Farooq, and Tej Prasad Dhamala. 2016. Sentiment analysis of english and tamil tweets using path length similarity based word sense disambiguation.

Yujun Yang, Jianping Li, and Yimei Yang. 2015. The research of the fast svm classifier method. *2015 12th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages 121–124.