MUCS@DravidianLangTech@ACL2022: Ensemble of Logistic Regression Penalties to Identify Emotions in Tamil Text

Asha Hegde[1 a], **Sharal Coelho**[1 b], **Hosahalli Lakshmaiah Shashirekha**[1 c]
[1]Department of Computer Science, Mangalore University, Mangalore, India
{[a]hegdekasha,[b]sharalmucs,[c]hlsrekha}@gmail.com

Abstract

Emotion Analysis (EA) is the process of automatically analyzing and categorizing the input text into one of the predefined sets of emotions. In recent years, people have turned to social media to express their emotions, opinions or feelings about news, movies, products, services, and so on. These users' emotions may help the public, governments, business organizations, film producers, and others in devising strategies, making decisions, and so on. The increasing number of social media users and the increasing amount of user generated text containing emotions on social media demands automated tools for the analysis of such data as handling this data manually is labor intensive and error prone. Further, the characteristics of social media data makes the EA challenging. Most of the EA research works have focused on English language leaving several Indian languages including Tamil unexplored for this task. To address the challenges of EA in Tamil texts, in this paper, we - team MUCS, describe the model submitted to the shared task on Emotion Analysis in Tamil at DravidianLangTech@ACL 2022. Out of the two subtasks in this shared task, our team submitted the model only for Task a. The proposed model comprises of an Ensemble of Logistic Regression (LR) classifiers with three penalties, namely: L1, L2, and Elasticnet. This Ensemble model trained with Term Frequency - Inverse Document Frequency (TF-IDF) of character bigrams and trigrams secured 4th rank in Task a with a macro averaged F1-score of 0.04. The code to reproduce the proposed models is available in github[1].

1 Introduction

Emotions are a form of psychological state of human mind and in texts the emotions are commonly represented through content bearing words such as happiness, anger, joy, disgust, boredom, depression, etc. The process of automatically analyzing and categorizing the input text into one of the predefined sets of emotions like happy, sad, angry and so on is called Emotion Analysis (Priyadharshini et al., 2021; Kumaresan et al., 2021). Analyzing the text for emotions helps to improve an existing process, grab new opportunities, capture the real response of audiences for their movies and reality shows, recognize and predict the market trends, and so on (Sampath et al., 2022; Ravikiran et al., 2022; Chakravarthi et al., 2022; Bharathi et al., 2022; Priyadharshini et al., 2022). Today internet and social media have become a popular platform for users to express the emotions, views, sentiments and opinions. The freedom to users to express their emotions about anything and everything on social media is increasing the social media text containing emotions (Chakravarthi, 2020; Chakravarthi and Muralidaran, 2021). Further, the freedom of the use of language on social media makes the analysis of social media text very challenging. The large volume and complexity of social media data makes the analysis of such data very challenging and interesting.

Most the EA works focus on English language leaving the task in several Indian languages including Tamil unexplored for the task (Vasantharajan et al., 2022). Due to the availability of a large volume of user-generated social media data in Tamil containing different emotions, EA in Tamil is gaining popularity (Jenarthanan et al., 2019). In recent years, there has been an increase in the EA of text in classical languages like Tamil The growing number of Tamil users on social media platforms and the increasing number of posts and comments shared by these users are making it nearly impossible to track and control the content manually (Sakuntharaj and Mahesan, 2021, 2017, 2016; Thavareesan and Mahesan, 2019, 2020a,b, 2021). Hence, there is a need for tools or models to analyse the emotions in the social media comments automatically. EA is an open-ended issue because of the creative users' cre-

[1]https://github.com/hegdekasha/Emotion-analysis-in-Tamil

Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, pages 145 - 150

ative posts on social media (B and A, 2021b,a). To address the challenges of EA in Tamil, in this paper, we - team MUCS, describe the model submitted to "Emotion Analysis in Tamil"[2] shared task organized by DravidianLangTech@ACL 2022. This task aims to classify the input comment in Tamil into one of eleven emotion categories. The proposed methodology consists of an Ensemble of LR classifiers with different regularizations or penalties, namely: LASSO (L1) regularization, Ridge (L2) regularization and Elasticnet regularization. TF-IDF of character bigrams and trigrams is used to train the LR classifiers and soft voting is used to classify the input comment into one of eleven categories.

The following is a breakdown of the paper's structure. Section 2 contains the literature review and Section 3 explains the proposed methodology. Section 4 describes the experiments conducted to identify and determine type of emotions, as well as the outcomes and the paper concludes in Section 5 with future work.

2 Literature Review

Researchers are trying to develop tools for processing the Tamil language for various applications such as EA, Text Summarization, Sentiment Analysis (SA) and so on (Nandwani and Verma, 2021).

Chiorrini et al. (2021) analyzed the performance of SA and emotion recognition using Bidirectional Encoder Representations from Transformers (BERT) models on real-world Twitter dataset. The experimental results showed that the models scored 0.92 and 0.90 accuracies for SA and emotion recognition, respectively. Vasantharajan et al. (2022) developed the largest manually annotated dataset of over 42k Tamil YouTube comments and categorized them into 31 emotions in order to recognize emotional statements. They established three distinct groups of emotions that are of 3-class, 7-class, and 31-classes. For the 3-class group dataset, they used Multilingual Representations for Indian Languages (MuRIL) pre-trained model trained on English and 16 Indian languages and obtained a macro average F1-score of 0.60. For 7-class and 31-class groups, the Random Forest (RF) model performed well with macro average F1-scores of 0.42 and 0.29, respectively.

To determine the category of emotions, Alotaibi (2019) trained Support Vector Machine (SVM), k-Nearest Neighbors (kNN), and LR classifiers on the International Survey on Emotion Antecedents and Reactions (ISEAR) dataset using TF-IDF features. LR classifier obtained 0.86 and 0.85 as precision and F1-score respectively. Using the benefits of Convolutional Neural Network (CNN) and Bidirectional Long Short Term Memory (BiLSTM), Ahmad et al. (2020) proposed an attention-based C-BiLSTM model to classify emotional states of poetry texts into different emotional states like love, joy, hope, sadness, anger, etc. Experimental results showed an accuracy of 88% for their model.

Even though several techniques have been developed to detect emotions in the text, very few attempts have been made for the Tamil language. This opens up lots of possibilities to conduct experiments on EA of Tamil texts including social media data.

3 Methodology

Inspired by Anusha and Shashirekha (2020) and Balouchzahi and Shashirekha (2020) an Ensemble of LR classifiers is proposed to identify the emotions in Tamil text and classify them into one of the given eleven categories and the framework of the proposed model is shown in Figure 1. The proposed model consists of three modules, namely: Pre-processing, Feature Extraction and Classifier Construction which are described briefly below:

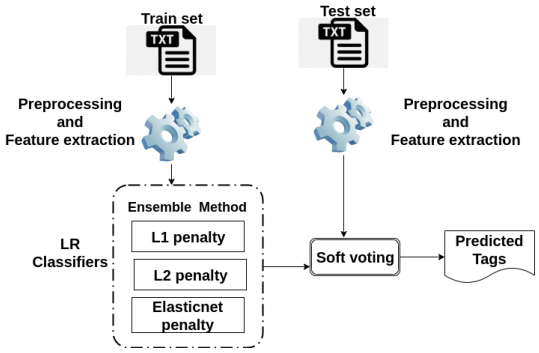

Figure 1: Framework of the proposed model

3.1 Pre-processing

Pre-processing step is essential to clean the text to improve the quality of data. The text is preprocessed by removing punctuation marks, digits, unrelated characters, and stopwords, as these features do not contribute to the task of classification. Tamil stopwords[3] list available in github repository

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

[3]https://gist.github.com/arulrajnet/

are used to remove Tamil stopwords from the given corpus as stop words do not contribute to the classification. Further, emojis are also removed as the dataset has enough textual content.

3.2 Feature Extraction

Feature extraction is one of the key steps in classification. TF-IDF expresses the relative importance between a word in the document and the entire corpus and TF-IDF of character n-grams has shown good performance (Kanaris et al., 2007). Hence, all the character bigrams and trigrams are extracted from the dataset and are vectorized using TFidfVectorizer[4]. The number of character bigrams and trigrams extracted from the datasets amounts to 13,808.

3.3 Classifier Construction

Model performance is heavily dependent on the features of the dataset and the classifier employed. No classifier produces good results for every dataset. Due to this, in general, no classifier can be considered as the best. An ensemble of classifiers, where the weakness of one classifier is compensated by the strength of another, produces better results than a single classifier. The proposed Ensemble of LR models with L1, L2 and Elasticnet penalties are trained on character bigrams and trigrams and soft voting is used to classify the input text into one of the emotion categories.

LR algorithm is a Machine Learning (ML) classifier used to predict categorical variables with the use of dependent variables and regularization to reduce overfitting (Indra et al., 2016). The penalties used in the LR models are described below:

- **L1 regularization -** The term LASSO stands for Least Absolute Shrinkage and Selection Operator and is also known as L1 regularization. In L1 regularization, L1 penalty which is equal to the absolute magnitude of coefficients is added to the loss function. L1 penalty uses shrinkage to determine regression coefficients and shrinkage occurs when a data value is shrunk towards zero.

- **L2 regularization -** The Ridge regularization also known as L2 regularization adds a squared magnitude of the coefficient to the loss function as a penalty. If the loss is zero then the regularization leads to an ordinary least square.

- **Elasticnet regularization -** L1 regularization eliminates many features, whereas L2 regularization manifests the loss by adding large weights. Elasticnet regularization is a popular type of regularized LR that combines L1 and L2 penalties. More precisely, elasticnet combines feature elimination from L1 regularization and feature coefficient reduction from L2 regularization to improve the model's predictions.

[4]https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

Classes	Training set	Dev set
Neutral	4,841	1,222
Joy	2,134	558
Ambiguous	1,689	437
Trust	1,254	272
Disguist	910	210
Anger	834	184
Anticipation	828	213
Sadness	695	191
Love	675	189
Surprise	248	53
Fear	100	23

Table 1: Statistics of Tamil dataset used for Task a

4 Experiments and Results

Statistics of the dataset for Task a in the EA shared task is summarized in Table 1 and the sample Tamil comments with their corresponding labels are shown in Table 2. The observation of the dataset shows the imbalance in the distribution of samples.

Several experiments were conducted with different values of the hyperparameters for the classifiers. The values of the hyperparameters which gave good results on the Development (Dev) set were used to conduct experiments on the Test set and such values of the hyperparameters are given in Table 3. For final evaluation and ranking, the predicted outputs on the Test set were submitted to the organizers of the shared task. A macro-averaged Precision, macro-averaged Recall, and macro-averaged F1-score were used by the organizers to measure the performance of the classifier for EA task and the results of the proposed model are shown in Table 4. The comparison of the performances of the best models of the shared task

Sl. No	Tamil sentences	Label
1.	அண்ணன் கிட்டுக்கு வாழ்த்துக்கள்	Joy
2.	வேல்ராஜ் வேலையா தான் இருக்கும்	Anticipation
3.	ஆமா நானும் இதான் யோஎத்தேன்	Trust
4.	இவர் சொன்னது உன் மை	Neutral
5.	எந்த ஊர் சொல்லுங்க அக்கா	Ambiguous
6.	ஏஒறுவனிடம் தோர்க போகும் பழசு	Sadness
7.	இந்த நிமிடமும் தமிழகத்தின் முதல்வர் எடப்பாடமு	Surprise
8.	பொணம் இன்னி பசங்க அரசாங்கம்	Anger
9.	உங்களை பார்களும் சகோதரா	Love
10.	இதே வேலை யா போச்சு இவனுகளுக்கு	Disguist
11.	நாம் டம்ளர் டெபாசிட் போச்சா	Fear

Table 2: Sample Tamil comments with their labels

Type of regularizations	Hyperparameters
L1	C=1, penalty="l1", tol=0.01, solver="saga"
L2	C=1, penalty="l2", tol=0.01, solver="saga"
Elasticnet	penalty="elasticnet", l1_ratio=0.5

Table 3: Details of hyperparameters used in the proposed model

with that of the proposed model in terms of macro-averaged F1-score is shown in Figure 2.

The proposed model obtained macro-averaged F1-scores of 0.20 and 0.04 for Dev set and Test set respectively. It is clear that the scores for Dev set and Test set were low because of imbalanced nature of the dataset. Class distribution has a significant impact on the predictions and the same is reflected in the results. 'Neutral' class has the maximum distribution of 34.07% of the overall distribution, whereas 'Fear' class has a least distribution of 0.70%. The proposed model exhibited a low F1-score because of the large difference between the number of samples in the classes.

Datasets	Precision	Recall	Macro averaged F1-score
Dev set	0.38	0.19	0.20
Test set	0.11	0.13	0.04

Table 4: Results of the proposed model

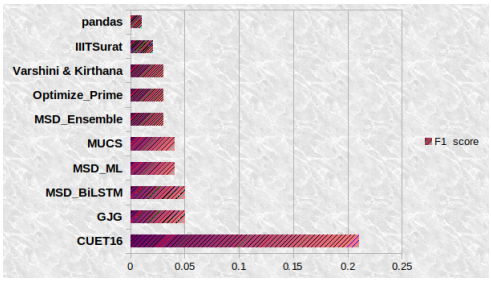

Figure 2: Comparison of the macro-averaged F1-scores of the proposed model with that of the other best models in the shared task

5 Conclusion and Future Work

In this paper, we, team MUCS, have presented the description of the proposed model submitted to a shared task on EA in Tamil at Dravidian-LangTech@ACL 2022 to identify the different categories of emotions from social media comments in Tamil. The proposed Ensemble of LR classifiers with L1, L2 and Elasticnet penalties obtained macro-averaged F1-score of 0.04 and secured 4th place in the shared task. In future, we intend to investigate sets of features and different re-sampling methods for identifying emotions in Tamil text.

References

Shakeel Ahmad, Muhammad Zubair Asghar, Fahad Mazaed Alotaibi, and Sherafzal Khan. 2020. Classification of Poetry Text into the Emotional States using Deep Learning Technique. volume 8, pages 73865–73878. IEEE.

Fahad Mazaed Alotaibi. 2019. Classifying Text-based Emotions using Logistic Regression.

M. D. Anusha and H. L. Shashirekha. 2020. An Ensemble Model for Hate Speech and Offensive Content Identification in Indo-European Languages. In *FIRE (Working Notes)*, pages 253–259.

Bharathi B and Agnusimmaculate Silvia A. 2021a. SSNCSE_NLP@DravidianLangTech-EACL2021: Meme classification for Tamil using machine learning approach. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 336–339, Kyiv. Association for Computational Linguistics.

Bharathi B and Agnusimmaculate Silvia A. 2021b. SSNCSE_NLP@DravidianLangTech-EACL2021: Offensive language identification on multilingual code mixing text. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 313–318, Kyiv. Association for Computational Linguistics.

Fazlourrahman Balouchzahi and H. L. Shashirekha. 2020. LAs for HASOC-Learning Approaches for Hate Speech and Offensive Content Identification. In *FIRE (Working Notes)*, pages 145–151.

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.

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.

Andrea Chiorrini, Claudia Diamantini, Alex Mircoli, and Domenico Potena. 2021. Emotion and Sentiment Analysis of Tweets using BERT. In *EDBT/ICDT Workshops*.

ST Indra, Liza Wikarsa, and Rinaldo Turang. 2016. Using Logistic Regression Method to Classify Tweets into the Selected Topics. In *2016 international conference on advanced computer science and information systems (icacsis)*, pages 385–390. IEEE.

Rajenthiran Jenarthanan, Yasas Senarath, and Uthayasanker Thayasivam. 2019. ACTSEA: Annotated Corpus for Tamil & Sinhala Emotion Analysis. In *2019 Moratuwa Engineering Research Conference (MERCon)*, pages 49–53. IEEE.

Ioannis Kanaris, Konstantinos Kanaris, Ioannis Houvardas, and Efstathios Stamatatos. 2007. Words versus Character N-grams for Anti-spam Filtering. volume 16, pages 1047–1067. World Scientific.

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.

Pansy Nandwani and Rupali Verma. 2021. A Review on Sentiment Analysis and Emotion Detection from Text. volume 11, pages 1–19. Springer.

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.

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.

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.

Charangan Vasantharajan, Sean Benhur, Prasanna Kumar Kumarasen, Rahul Ponnusamy, Sathiyaraj Thangasamy, Ruba Priyadharshini, Thenmozhi Durairaj, Kanchana Sivanraju, Anbukkarasi Sampath, Bharathi Raja Chakravarthi, et al. 2022. TamilEmo: Finegrained Emotion Detection Dataset for Tamil.