Varsini_and_Kirthanna@DravidianLangTech-ACL2022-Emotional Analysis in Tamil

Varsini S, Kirthanna Rajan, Angel Deborah S, Rajalakshmi S, R.S.Milton, T.T.Mirnalinee
Department of Computer Science and Engineering
Sri Sivasubramaniya Nadar College of Engineering
Chennai , India
{varsini.sathianantha,kirthanna.rajan}@gmail.com,
{angeldeborahs,rajalakshmis,miltonrs,mirnalineett}@ssn.edu.in

Abstract

In this paper, we present our system for the task of Emotion analysis in Tamil. Over 3.96 million (Gaubys) people use these platforms to send messages formed using texts, images, videos, audio or combinations of these to express their thoughts and feelings. Text communication on social media platforms is quite overwhelming due to its enormous quantity and simplicity. The data must be processed to understand the general feeling felt by the author. We present a lexicon-based approach for the extraction emotion in Tamil texts. We use dictionaries of words labelled with their respective emotions. The process of assigning an emotional label to each text, and then capture the main emotion expressed in it. Finally, the F1-score in the official test set is 0.0300 and our method ranks 5th.

1 Introduction

Emotion Detection is the process of detecting the different human emotions such as anger, disgust, joy, sadness, surprise, love, anticipation, so on and so forth (Cherry). “Emotion Identification”, “Emotion Analysis” and “Emotion detection” all mean the same and can be used interchangeably (Munezero et al., 2014). People who are users of social media use these platforms as a way to express their feelings, thoughts and opinions on a wide range of topics. These feelings may be positive or neutral or negative. The book “Emotions In Social Psychology”, written by W. Gerrod Parrot, in 2001 (Parrott, 2001). In which he explained that the human emotion system can be formally classified into an emotion hierarchy with six classes at the primary level, namely Surprise, Love, Anger, Fear, Sadness and Joy, while certain other words fall in the secondary and tertiary levels.

Emotion detection is used in various fields such as the business world to analyze how people feel about their new product; in the medical world by identifying the way people respond to a pandemic (Ravikiran et al., 2022; Chakravarthi et al., 2022; Bharathi et al., 2022; Priyadharshini et al., 2022; Chakravarthi et al., 2021; Chakravarthi, 2020; Chakravarthi and Muralidaran, 2021).Emotion identification is also used in monitoring the feelings and emotions of the Users who use social platforms such as Instagram, Facebook, YouTube, Twitter and many more (Rajendram et al., 2017, 2022).

Analyzing the emotions felt by the author using the text is quite challenging while also interesting and essential, as most of the time these text messages not only express the emotion directly by using emotional words and emojis but also the interpretation of the meaning of concepts. Furthermore, new slang or terminologies or short-forms are being created as each day passes, which make emotion detection from text a more interesting as well as a challenging problem for us to tackle (Angel Deborah et al., 2021).

Tamil Language is a Dravidian Language that is natively spoken by the people of Tamil Nadu in South Asia. It is also the official Language of two sovereign nation, Sri Lanka and Singapore as well as the official language of the Union Territory of Puducherry (Sakuntharaj and Mahesan, 2021, 2017, 2016; Thavareesan and Mahesan, 2019, 2020a,b, 2021). Tamil was known as Tamilakam in the time period of the 6th century to the 3rd century CE. Tamil is the first Indian classical language to listed as classical language, and is one of the world’s oldest classical languages that is still spoken. There are 12 vowels, 18 consonants, and one special character, the aytam, in the present Tamil script. The vowels and consonants merge to form 216 compound characters, for a total of 247 characters (12 vowels + 18 consonants + 1 aytam + (12 x 18) combinations) (Chakravarthi et al., 2020; Anita and Subalalitha, 2019b,a; Subalalitha and Poovammal, 2018; Subalalitha, 2019; Srinivasan

May 26, 2022 ©2022 Association for Computational Linguistics

and Subalalitha, 2019; Narasimhan et al., 2018).

The paper is organized as follows. Section 2 discusses the related work on emotion analysis. The data set about the shared task is in Section 3.1. Section 4 outlines the features of the proposed system. Section 5 concludes the paper.

2 Related Work

Emotional Analysis from text is considered as an interesting as well as a challenging task in NLP. However due to lack of data set in Tamil language it is difficult to conduct high level research in this area.

The TamilEmo (Vasantharajan et al., 2022)introduced a labelled data set that is manually annotated of more than 42,000 Tamil YouTube comments and is labelled for 31 emotions including neutral. The main aim of the data set is to improve the detection of emotions from Tamil texts. They have also created three different groupings of emotions namely 3-class, 7-class and 31-class.

ACTSEA (Jenarthanan et al., 2019) presented a corpus for emotion analysis that is a scalable semi-automatic approach for creating annotated corpus for Tamil and Sinhala. They gathered data from an online social platform, Twitter, and then manually annotated them after cleaning it. They collected 6,00,280 Tamil Tweets and 3,18,308 Sinhala tweets which now make them have one of the largest data sets for the languages Tamil and Sinhala.

In the year 2007, two people - Strapparava and Mihalcea presented three detailed systems that took part in the SemEval 2007 Affective Text task. The three systems were rule-based, unsupervised and supervised systems. They noted that the rule based system performed the best for 4 emotion classes out of 6, while the supervised system did the best in the remaining two emotion classes. This was done for the language, English.

In the year 2007, Yang et al. proved that sequence labellers can outperform traditional classifier (Support Vector System) on a dataset of blogs, increasing the accuracy to 43.35 from 32.88.

3 Document Body

3.1 Data Description

Competition organizers provided data with text as features (Sampath et al., 2022). The text feature contains the 14,208 total data with emotions being classified as Ambiguous, Anger, Anticipation, Disgust, Fear, Joy, Love, Neutral, Sadness, Surprise, Trust. The detailed contents of the data set are shown in the Table 1.

Table 1. Class Distribution for Emotion Classification in Tamil

Label	Training set
Ambiguous	1689
Anger	834
Anticipation	828
Disgust	910
Fear	100
Joy	2,134
Love	675
Neutral	4,841
Sadness	695
Surprise	248
Trust	1,254

3.2 Emotion Identification using Keyword Spotting

The Emotion Identification is finding the frequency of the emotion word by checking the Emotion Word Knowledge Base (Jenarthanan14) and finding the frequency of the emoji by checking the Emotion Emoji Knowledge Base. This is done by tokenizing the given string (text message) into many substrings (words in the text message) and matching each substring to find a match in the Knowledge Bases. The Knowledge Base consists of emotions namely – anger, sadness, disgust, joy, surprise, fear, love, anticipation and so on and so forth.

This process for identifying emotion contains eight steps as given in Figure 1, here the text message is given as input and the returned value is the emotion felt in the text message. After getting input, we perform tokenization using space " " as the separating delimiter and create a list of substrings, that represent the words as well as the emojis in the text. These substrings are used to analyze the frequency of the emotion. Then the emotion is the output.

3.3 Lexical Affinity Method

The Lexicon-based method is a keyword-based search method that checks for emotion keywords assigned to some Emotional Classes(Abdaoui et al., 2017).

It is based on the idea of detecting emotions based on related keywords such as emotional words and

emojis. This is pretty easy to implement and a straightforward approach. This is more of an extension to the above “Emotion Identification using Keyword Spotting”, by assigning a number to the respective emotion. It increments the respective emotion variable’s value each time a word or emoji of that emotion is found. For example, if a smiley face is found 3 times in a text, the value of happiness is incremented by 3.
The flowchart is given in Figure 1.

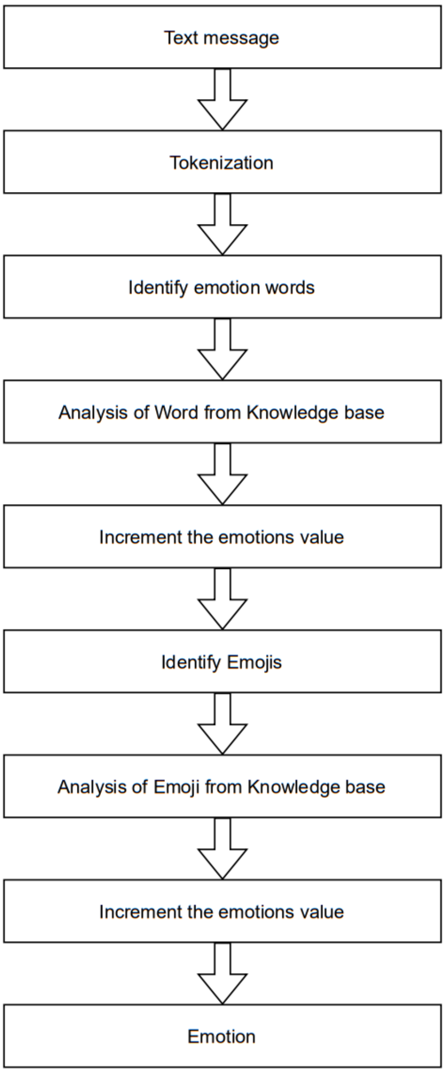
Text message
Tokenization
Identify emotion words
Analysis of Word from Knowledge base
Increment the emotions value
Identify Emojis
Analysis of Emoji from Knowledge base
Increment the emotions value
Emotion

Figure 1: Lexical Affinity Approach Flowchart

4 Our System

Methods described in the previous section, i.e., Section 3 are modified and integrated to extend their capabilities and to improve the performance for which a simple and easy to understand model is designed shown in Figure 2.

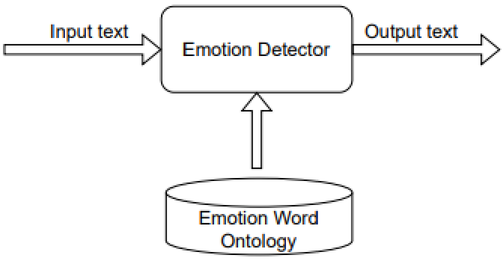

Figure 2: Our system model

4.1 Emotion Word Ontology

The Emotion Word Ontology is a combination of two Knowledge Bases of both words (in Tamil) (Jenarthanan14) and emojis.
The Word Knowledge Base consists of a list of emotional keywords that are matched to their respective emotion class. For example, words that express anger in Tamil are under the class "Anger" and are in all forms (past/present/future as well as singular/plural), while the words that express disgust in Tamil are under the class "Disgust" and are in all forms.
Similarly, there is an Emoji Knowledge Base that consists of emotion icons that are matched to their respective emotions. For example, icons that have a heart are under the class "Love", while the icons with tears are under the class "Sadness".

4.2 Emotion Detector

This is a function that helps in detecting the emotion of the Tamil message that is given as the input. The function assigns a value for each emotion, it also increments the values of the respective emotion variable when it encounters the same word or emoji in the Emotion Word Ontology. The emotion variable that has the greatest value is taken as the detected emotion.

5 Conclusion

In conclusion, the proposed system analyses emotions from text messages that are written in Tamil, using a very simple and straightforward method. Research in the domain of Emotion Analysis has flourished significantly over the past few years, making it a need to take a look back at the big picture that these individual works have led to. There are many methods and models to analyse emotions on text. (Tripathi et al., 2016) The dictionary-based approach is quite straightforward and adaptable to apply to any language.

Acknowledgement

We would like to thank Sri Sivasubramaniya Nadar College of Engineering and its Department of Computer Science and Engineering for providing us with the opportunity to work on this paper. We would also like to take this opportunity to thank The Association of Computational Linguistics (ACL) for inviting us to the DravidianLangTech Workshop and providing us with the opportunity to work on this problem and paper. We also thank Tamil-Sinhala-Emotion-Analysis (Jenarthanan14) (Jenarthanan et al., 2019)for providing us with their Data Set.

References

Amine Abdaoui, Jérôme Azé, Sandra Bringay, and Pascal Poncelet. 2017. Feel: a french expanded emotion lexicon. *Language Resources and Evaluation*, 51(3):833–855.

S Angel Deborah, TT Mirnalinee, and S Milton Rajendram. 2021. Emotion analysis on text using multiple kernel gaussian... *Neural Processing Letters*, 53(2):1187–1203.

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.

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*.

Bharathi Raja Chakravarthi, Navaneethan Rajasekaran, Mihael Arcan, Kevin McGuinness, Noel E. O'Connor, and John P. McCrae. 2020. Bilingual lexicon induction across orthographically-distinct under-resourced Dravidian languages. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 57–69, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Kendra Cherry. The 6 types of basic emotions and their effect on human behavior.

Justas Gaubys. How many people use social media in 2022? [updated jan 2022].

Rajenthiran Jenarthanan, Yasas Senarath, and Uthayasanker Thayasivam. 2019. Actsea: Annotated corpus for tamil amp; sinhala emotion analysis. In *2019 Moratuwa Engineering Research Conference (MERCon)*, pages 49–53.

Jenarthanan14. Jenarthanan14/tamil-sinhala-emotion-analysis.

Myriam Munezero, Calkin Suero Montero, Erkki Sutinen, and John Pajunen. 2014. Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE Transactions on Affective Computing*, 5(2):101–111.

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.

W Gerrod Parrott. 2001. *Emotions in social psychology: Essential readings*. psychology press.

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.

S Milton Rajendram, TT Mirnalinee, et al. 2017. Ssn_mlrg1 at semeval-2017 task 5: fine-grained sentiment analysis using multiple kernel gaussian process regression model. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 823–826.

S Milton Rajendram, Mirnalinee TT, et al. 2022. Contextual emotion detection on text using gaussian process and tree based classifiers. *Intelligent Data Analysis*, 26(1):119–132.

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, 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.

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.

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.

Vaibhav Tripathi, Aditya Joshi, and Pushpak Bhattacharyya. 2016. Emotion analysis from text: A survey. *Center for Indian Language Technology Surveys*.

Charangan Vasantharajan, Sean Benhur, Prasanna Kumar Kumarasen, Rahul Ponnusamy, Sathiyaraj Thangasamy, Ruba Priyadharshini, Thenmozhi Durairaj, Kanchana Sivanraju, Anbukkarasi Sampath, Bharathi Raja Chakravarthi, and John Phillip McCrae. 2022. Tamilemo: Finegrained emotion detection dataset for tamil.