JudithJeyafreedaAndrew@TamilNLP-ACL2022:CNN for Emotion Analysis in Tamil

Judith Jeyafreeda Andrew
The University of Manchester, Oxford road, Manchester, United Kingdom
judithjeyafreeda@gmail.com

Abstract

Using technology for analysis of human emotion is a relatively nascent research area. There are several types of data where emotion recognition can be employed, such as - text, images, audio and video. In this paper, the focus is on emotion recognition in text data. Emotion recognition in text can be performed from both written comments and from conversations. In this paper, the dataset used for emotion recognition is a list of comments. While extensive research is being performed in this area, the language of the text plays a very important role. In this work, the focus is on the Dravidian language of Tamil. The language and its script demands an extensive pre-processing. The paper contributes to this by adapting various pre-processing methods to the Dravidian Language of Tamil. A CNN method has been adopted for the task at hand. The proposed method has achieved a comparable result.

1 Introduction

Emotion Analysis is a task of classification of emotions in text. There are several application for this task such as reviews analysis in e-commerce, public opinion analysis, extensive search, personalized recommendation, healthcare, and online teaching (Sampath et al., 2022a; Ravikiran et al., 2022; Chakravarthi et al., 2022; Bharathi et al., 2022; Priyadharshini et al., 2022). A lot of research has been done on classifying comments, opinions, movie/product reviews, ratings, recommendations and other forms of online expression into positive or negative sentiments (Priyadharshini et al., 2021; Kumaresan et al., 2021; Chakravarthi, 2020; Chakravarthi and Muralidaran, 2021; Chakravarthi et al., 2020b).

Though there have been several research works around emotion recognition in English language, there are not many in Dravidian languages (Chakravarthi et al., 2021a,b, 2020a; Priyadharshini et al., 2020). The four major Dravidian languages are Tamil, Telugu, Malayalam and Kannada. This paper explores the idea of using deep neural networks specifically CNN for the purpose of Emotion Recognition in text from the Dravidian Language of Tamil (Ghanghor et al., 2021a,b; Yasaswini et al., 2021).

Tamil is one of the world's longest-surviving classical languages (Anita and Subalalitha, 2019b,a; Subalalitha and Poovammal, 2018; Subalalitha, 2019). According to A. K. Ramanujan, it is "the only language of modern India that is recognizably continuous with a classical history." Because of the range and quality of ancient Tamil literature, it has been referred to as "one of the world's major classical traditions and literatures." For about 2600 years, there has been a recorded Tamil literature (Sakuntharaj and Mahesan, 2021, 2017,?, 2016). The earliest period of Tamil literature, known as Sangam literature, is said to have lasted from from 600 BC to AD 300. Among Dravidian languages, it possesses the oldest existing literature. The earliest epigraphic documents discovered on rock edicts and "hero stones" date from the 6th century BC (Thavareesan and Mahesan, 2019, 2020a,b, 2021).

The task in (Sampath et al., 2022b) is categorized in two subtasks, both of which dealing with a corpus in the Dravidian language of Tamil. The first one aims at classifying social media comments in 8-10 classes where the classes are in English. The second subtask involves classifying text into one of the 30 classes, where the classes are also in tamil. The classification systems performance has been measured in terms of macro averaged Precision, macro averaged Recall and macro averaged F-Score across all the classes.

2 Related Work

With the increase in social media content in the recent past, a lot of focus has been given to Emotion Analysis. Several Machine Learning and Deep

Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, pages 58 - 63

Learning approaches have been developed for this cause. (Wiebe et al., 2005) proposed a manual corpus annotation for emotions and sentiments in news articles. (Strapparava and Mihalcea, 2008) describes an experiment for automatic identification of six different emotions in text including Anger, Disgust, Fear, Joy, Sadness and Surprise. The authors propose both knowledge based and corpus based methods for this purpose. (Liu, 2017) uses emotion detection to predict the future stock returns by applying a emotion classifier to tweets from the 2016 presidential election and financial tweets. (Gaind et al., 2019) uses a supervised model.The model developed is a hybrid one consisting of two completely different approaches. The first approach uses Emotion-Words Set and several textual features to classify and score text according to the emotions. The second approach uses standard classifiers like SMO and J48 to classify tweets. Finally, these approaches are combined to detect emotions in text more effectively. (Stojanovski et al., 2015) uses convolutional neural network architecture for emotion identification in Twitter messages. The model has been applied on Twitter messages for emotion identification related to public local services. This is an unsupervised method. (Savigny and Purwarianti, 2017) compared many methods for using word embedding in a classification task, namely average word vector, average word vector with TF-IDF, paragraph vector, and by using Convolutional Neural Network (CNN) algorithm. The authors showed that the accuracy of the classification increases while word embeddings are used in combination with CNN. (Zhang et al., 2018) addresses the problem where a sentence can evoke more than one emotion. For this purpose, the authors introduce an emotion distribution learning and propose a multi-task convolutional neural network for text emotion analysis.

(Andrew, 2020) proposes several machine learning techniques to classify sentiments from YouTube comments in the Dravidian languages of Tamil and Malayalam. The corpus in (Andrew, 2020) is YouTube comments in code mixed Dravidian languages of Tamil and Malayalam. It is noted that a Naïve Bayes method performs the best for sentiment analysis if YouTube comments on code mixed Dravidian language of Tamil. (Andrew, 2021) performs offensive language detection on YouTube comments in Dravidian languages of Tamil, Malayalam and Kannada. The authors perform a preprocessing step that allows the substitution of Dravidian language script to Latin script, replacement of emojis with words and the standard method of removing stop words. This is then followed by the use of several machine language techniques.

3 Data

The datset for the two subtasks are from (Sampath et al., 2022b).

3.1 Subtask A

The goal of subtask A is to classify emotions in Tamil text into 8-10 classes. The classes are in English. The classes are: Ambiguous, Anger, Anticipation, Disgust, Joy, Love, Neutral, Sadness and Trust. The train set consists of 14208 sentences, the development sets consists of 3552 sentences and the test set consists of 4440 sentences.

3.2 Subtask B

The goal of subtask B is to classify emotions in Tamil text into 30 classes. However, unlike subtask A, the classes are in Tamil as well. The train set consists of 30179 sentences, the development sets consists of 4269 sentences and the test set consists of 4268 sentences.

4 Pre-Processing

The Tamil text needs some pre-processing before training a deep learning algorithm. The preprocessing techniques are similar to ones in (Andrew, 2021).

- The words in the script of the Dravidian language of Tamil are replaced by latin text. For subtask B, both the text and the classes are replaced by latin text (IPA). This is performed using the anyascii package in Python.

- The emojis found in the text are replaced by the words that the emoji represents like happy, sad etc.

- Remove stop words and punctuations. For this purpose, python packages for language specific stop words. The advertools and stopwordsiso are used for language specific stopwords.

5 Deep Learning Methods for emotion classification

5.1 Pre-Processing

In this paper, a first preprocessing is done in order to change the script of Tamil to IPA, as described in the previous section. However, in order to be able to trained for a deep learning model, pre-processing methods like tokenization and stemming is performed on the transformed text. For this purpose, the inbuilt 'keras' python package is used.

5.2 Embedding

There have been several word embeddings proposed for the Dravidian language of Tamil. (Thavareesan and Mahesan, 2020c) proposes a word embedding-based Part of Speech (POS) tagger for Tamil, with experiments conducted on BoW, TF-IDF, Word2vec, fastText and GloVe. (Kumar et al., 2020) presents word-embedding for 14 different Indian languages including Tamil. A total of 422 embeddings have been released. In this paper, the embeddings from (Kumar et al., 2020) is used.

5.3 Deep Learning Models

In this paper, a Convolutional Neural Network (CNN) is used for emotion classification. The 'Keras' python CNN package is used for this purpose.

5.3.1 CNN

The central idea behind a CNN is the convolving or sliding pre-determined window of data. The data is first represented using word vectors. A weight matrix, called a filter consisting of an activation function, is then slid horizontally across the sentences by one step. Backpropagation will ensure that the weights of these filters are learned from the data. The next step is to calculate the convoluted feature. This layer is calculated by summing over the element-wise multiplication as each filter slides over the window of data one stride at a time and is multiplied by its corresponding weight in the filter. In cases where the filter doesnt exactly fit the matrix with a given number of slides, a **padding** is necessary. This can be done in two ways: (i) Pad the outer edges with zero vectors (zero-padding) (ii) ignore the part of the matrix that does not fit the filter (valid padding). In order to help the algorithm learn higher-order representations of the data while reducing the number of parameters, **pooling** can be performed. There are three types of pooling - Sum pooling, Max pooling and average pooling.

Finally, the fully connected layer receives the input from the previous pooling and convolutional layers. It then performs a classification task (cnn). This process is shown in Figure 1

Task	Precision	Recall	F1-score
A	0.150	0.122	0.094
B	0.094	0.068	0.057

Table 1: Results.

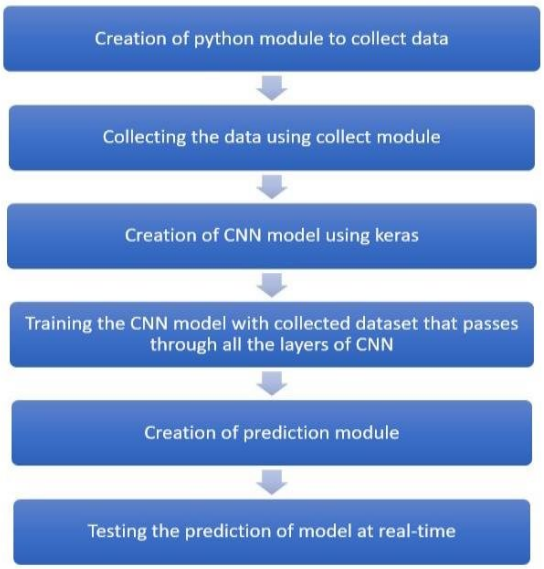

Figure 1: General Process Flow for a Convolution Neural Network (Pathak and Khan, 2021)

6 Results

The performance of the classification system has been evaluated in terms of macro averaged Precision, macro averaged Recall and macro averaged F-Score across all the classes. The evaluation has been performed with the sklearn package on python (Pedregosa et al., 2011).

The results for both tasks A and B are shown in Table 1.

A precision of 0.150, a recall of 0.122 and a F1-score of 0.094 is achieved for Task A. The highest scores of metrics achieved for Task A are precision is 0.220, recall is 0.250 and F1 score is 0.210.

A precision of 0.094, a recall of 0.068 and a F1-score of 0.054 is achieved for Task B. The highest scores of metrics achieved for Task B are precision is 0.15, recall is 0.171 and F1 score is 0.151.

In general this is quite low. It has to be kept in mind that task 2 had both the text and labels in the Dravidian language of Tamil.

It can be noted that when the language of the labels/category is in English, the results are better than when both the labels/category is in Tamil. (Andrew, 2021) shows that pre-processing Dravidian texts help improve the results when used with Machine Learning models, however, this does not seem to be the case with deep learning techniques. This is because deep learning techniques requires huge amount of training data. For a language like Tamil, such models are not easily available due to the lack if data. Using language models such as BERT trained for the Dravidian language of Tamil over a large corpus could help in more accurate classification of emotions.

There is clearly a huge amount of efforts that needs to go in encoding and decoding of Dravidian language scripts. Translating Dravidian Language scripts to Latin alphabets might not be the best approach for emotion classification. This is a critical point of pre-processing that needs to be considered in future works. Any new model built should be able to process the text with the script of the Dravidian language itself.

References

Nlp with cnns. https://towardsdatascience.com/nlp-with-cnns-a6aa743bdc1e.

Judith Jeyafreeda Andrew. 2020. Judithjeyafreeda@ dravidian-codemix-fire2020: Sentiment analysis of youtube comments for dravidian languages. In *FIRE (Working Notes)*, pages 522–527.

Judith Jeyafreeda Andrew. 2021. Judithjeyafreedaandrew@ dravidianlangtech-eacl2021: offensive language detection for dravidian code-mixed youtube comments. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 169–174.

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, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020a. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210, Marseille, France. European Language Resources association.

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, Vigneshwaran Muralidaran, Shardul Suryawanshi, Navya Jose, Elizabeth Sherly, and John P McCrae. 2020b. Overview of the track on sentiment analysis for Dravidian languages in code-mixed text. In *Forum for Information Retrieval Evaluation*, pages 21–24.

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

Bharathi Raja Chakravarthi, Priya Rani, Mihael Arcan, and John P McCrae. 2021b. A survey of orthographic information in machine translation. *SN Computer Science*, 2(4):1–19.

Bharat Gaind, Varun Syal, and Sneha Padgalwar. 2019. Emotion detection and analysis on social media. *arXiv preprint arXiv:1901.08458*.

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.

Saurav Kumar, Saunack Kumar, Diptesh Kanojia, and Pushpak Bhattacharyya. 2020. "a passage to India": Pre-trained word embeddings for Indian languages. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 352–357, Marseille, France. European Language Resources association.

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.

Clare H Liu. 2017. *Applications of twitter emotion detection for stock market prediction*. Ph.D. thesis, Massachusetts Institute of Technology.

Adarsh Pathak and Faraz Khan. 2021. Comparison of cnn and contour algorithm for number identification using hand gesture recognition.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Mani Vegupatti, and John P McCrae. 2020. Named entity recognition for code-mixed Indian corpus using meta embedding. In *2020 6th international conference on advanced computing and communication systems (ICACCS)*, pages 68–72. IEEE.

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. 2022a. 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.

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. 2022b. Findings of the shared task on Emotion Analysis in Tamil.

Julio Savigny and Ayu Purwarianti. 2017. Emotion classification on youtube comments using word embedding. In *2017 international conference on advanced informatics, concepts, theory, and applications (ICAICTA)*, pages 1–5. IEEE.

Dario Stojanovski, Gjorgji Strezoski, Gjorgji Madjarov, and Ivica Dimitrovski. 2015. Emotion identification in fifa world cup tweets using convolutional neural network. In *2015 11th International Conference on Innovations in Information Technology (IIT)*, pages 52–57. IEEE.

Carlo Strapparava and Rada Mihalcea. 2008. Learning to identify emotions in text. In *Proceedings of the 2008 ACM Symposium on Applied Computing*, SAC '08, page 15561560, New York, NY, USA. Association for Computing Machinery.

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. 2020c. 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.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2):165–210.

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.

Yuxiang Zhang, Jiamei Fu, Dongyu She, Ying Zhang, Senzhang Wang, and Jufeng Yang. 2018. Text emotion distribution learning via multi-task convolutional neural network. In *IJCAI*, pages 4595–4601.