A Deep Neural Network-based Model for the Sentiment Analysis of Dravidian Code-mixed Social Media Posts

Jyoti Kumari[1], Abhinav Kumar[2]

[1]Department of Computer Science & Engineering, National Institute of Technology Patna, Patna, India
[2]Department of Computer Science & Engineering, Siksha 'O' Anusandhan Deemed to be University, Bhubaneswar, India

Abstract

Sentiment analysis is one of the most essential jobs in natural language processing. The research community has recently presented a slew of papers aimed at detecting sentiment from English social media posts. Despite this, research on recognising feelings in Dravidian Kannada-English, Malayalam-English, and Tamil-English postings has been limited. This study offers a dense neural network-based model for categorising postings in Kannada-English, Malayalam-English, and Tamil-English into five different sentiment classes. When character-level TF-IDF characteristics are combined with a dense neural network, encouraging results are obtained. The recommended model received weighted $F1$-scores of 0.61, 0.72, and 0.60 for Kannada-English, Malayalam-English, and Tamil-English social media postings, respectively. The code for the proposed models is available at https://github.com/Abhinavkmr/Deep-Neural-Network-based-Model-for-the-Sentiment-Analysis-of-Dravidian-Social-Media-Posts.git

Keywords

Sentiment analysis, Code-mixed, Tamil, Malayalam, Kannada, Machine learning, Deep learning

1. Introduction

Sentiment analysis is the process of finding polarity of a sentence. It can be negative or positive of neutral depending on the context of the text. It is most helpful in recognizing opinions/recommendations/queries/answers on a specific subject/product. It is gaining much attention these days due to its significant impact on businesses like e-commerce, spam detection [1, 2], recommendation system, social media monitoring, hate speech [3, 4], and disaster management [5, 6]. English is a common and acceptable language worldwide specially in the digital world. However, in a multilingual country like India, with more than 400 million internet users speaking more than one language for communications produces a new code-mixed language [7]. The presence of multiple script and language constructs in a text makes it more challenging . Most of the existing models are trained for single language's sentiment analysis and thus fails to capture a code-mixed language semantics. Extracting sentiments from code mixed user-generated texts becomes more difficult due to its multilingual nature [8].

Recently, the sentiment analysis of code-mixed language [9, 10] has drawn attention of the research community. Joshi et al. [11] have presented a model for Hinglish (Hindi-English)

FIRE 2021: Forum for Information Retrieval Evaluation, December 13-17, 2021, India
j2kumari@gmail.com (J. Kumari); abhinavanand05@gmail.com (A. Kumar)

dataset with subword representation of code-mix data and long short term memory (subword-LSTM). In [7], Patra et al. reported a model based on support vector machine using character n-grams features for Bengali-English code mixed data. Advani et al. [12] have used logistic regression with handcrafted lexical and semantic features to extract sentiments from Hinglish and Spanglish (Spanish + English) data. A morphological attention model has been proposed by Goswami et al. [13] for sentiment analysis on Hinglish data.

The Malayalam and Kannada languages are spoken in the Indian state of Kerala and Karnataka. There are around 38 million Malayalam speakers over the globe. Tamil is another Dravidian language spoken by Tamil people in India, Singapore, and Sri Lanka. The scripts of both the Dravidian languages are alpha-syllabic i.e., partially alphabetic and partially syllable-based. However, Roman script is frequently used for posting on social media because it is easy [6]. Thus, the majority of the available data on social media are code-mixed.

In this paper, we proposed a dense neural network-based model that utilizes character-level n-gram TF-IDF features to identify sentiments from the Kannada-English [14], Malayalam-English [15], and Tamil-English [16] social media posts. The proposed dense neural network-based model is validated by the dataset published in the *DravidianCodeMix FIRE 2021* [17, 18] track. The dataset provided by the organizer contains five different sentiment labels such as "positive," "negative," "mixed feelings," "unknown state," and "if the post is not in the mentioned Dravidian languages."

The rest of the paper is summarized as: related work is listed in Section 2, the dataset description, the proposed methodology is explained in Section 3. Various experiments and their finding is presented in Section 4. Finally, Section 5 concludes the discussion by highlighting the main findings of this study.

2. Related work

Recently, a number of works [19, 11, 7, 12, 9] have been made to extract sentiment from code-mixed social media posts. Mahata et al. [19] proposed a bi-directional LSTM-based model for detecting sentiment in social media posts containing both English and Tamil language. Joshi et al. [11] have presented a model for Hinglish (Hindi-English) dataset with subword representation of code-mix data and long short term memory (subword-LSTM). Mandalam et al. [20] proposed an LSTM-based model that uses sub-word level features and word embedding vectors to identify sentiment from code-mixed Tamil and Malayalam social media posts. In [7], Patra et al. reported a model based on support vector machine using character n-grams features for Bengali-English code mixed data.

Dowlagar et al. [21] proposed graph convolutional networks (GCN) for detecting sentiments in Dravidian social media posts. Balouchzahi et al. [22] proposed three different models to identify sentiments from Dravidian social media posts: SACo-Ensemble, SACo-Keras, and SACo-ULMFiT, using machine learning, deep learning, and transfer learning, respectively. Advani et al. [12] have used logistic regression with handcrafted lexical and semantic features to extract sentiments from Hinglish and Spanglish (Spanish + English) data. A morphological attention model has been proposed by Goswami et al. [13] for sentiment analysis on Hinglish data.

In line with the current literature, since social media code-mixed posts contain several

grammatical errors and non-standard abbreviations, the use of character-level features may perform well; thus, this paper investigates the usability of character-level features with a simple four-layered dense neural network for sentiment identification from Kannada, Malayalam, and Tamil social media posts.

3. Methodology

The systematic diagram for the proposed dense neural network-based model can be seen in Figure 1. The proposed model is validated against the Kannada-English, Malayalam-English, and Tamil-English datasets [18]. The overall data statistic for each language can be seen in Table 1.

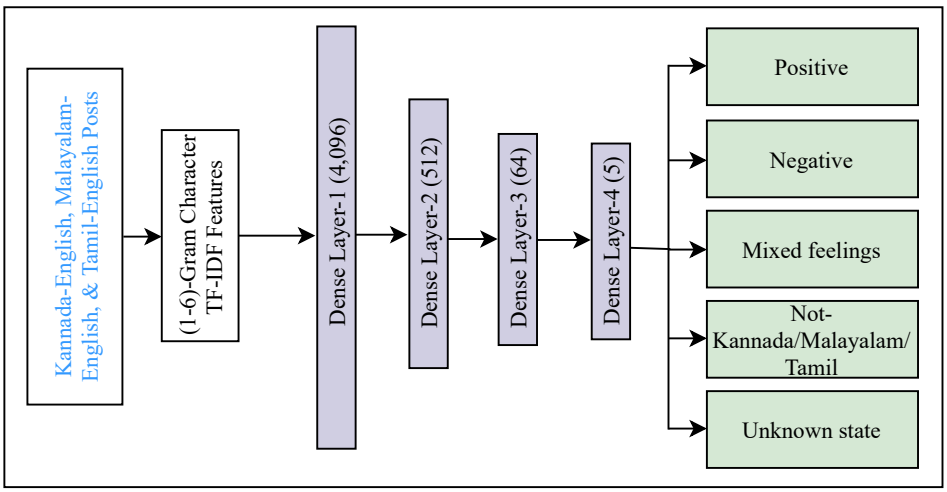

Figure 1: Proposed deep neural network-based model for sentiment classification from Kannada-English, Malayalam-English, and Tamil-English datasets

Table 1
Overall data statistic for Kannada, Malayalam, and Tamil dataset

Class	Kannada			Malayalam			Tamil		
	Train	Validation	Test	Train	Validation	Test	Train	Validation	Test
Mixed-feelings	574	52	65	926	102	134	4,020	438	470
Negative	1,188	139	157	2,105	237	258	4,271	480	477
Positive	2,823	321	374	6,421	706	780	20,070	2,257	2,546
Unknown state	711	69	62	5,279	580	643	5,628	611	665
Not-Kannada	916	110	110	-	-	-	-	-	-
Not-Malayalam	-	-	-	1,157	141	147	-	-	-
Not-Tamil	-	-	-	-	-	-	1,667	176	244
Total	6,212	691	768	15,888	1,766	1,962	35,156	3,962	4,402

In our experiments, we found that using character-level features with a dense neural network performed better than some complex models like convolutional neural network (CNN) and

Table 2
Best-suited hyper-parameters for the proposed model in case of Kannada-English, Malayalam-English, and Tamil-English

Hyper-parameters	Kannada-English	Malayalam-English	Tamil-English
Dense layes	4	4	4
Number of neurons at each layer	4,096, 512, 64, 5	4,096, 512, 64, 5	4,096, 512, 64, 5
Dropout	0.2	0.2	0.2
Activation function	ReLU, Softmax	ReLU, Softmax	ReLU, Softmax
Optimizer	Adam	Adam	Adam
Loss	Categorical cross-entropy	Categorical cross-entropy	Categorical cross-entropy
Learning rate	0.001	0.001	0.001
Batch size	20	20	20
Epochs	50	50	50

Table 3
Performance of the proposed model for Kannada, Malayalam, and Tamil social media posts

Class	Kannada-English			Malayalam-English			Tamil-English		
	Precision	Recall	F_1-score	Precision	Recall	F_1-score	Precision	Recall	F_1-score
Mixed-feelings	0.26	0.14	0.18	0.49	0.44	0.46	0.21	0.19	0.20
Negative	0.63	0.64	0.63	0.68	0.53	0.60	0.39	0.40	0.39
Positive	0.73	0.70	0.72	0.80	0.78	0.79	0.76	0.78	0.77
Unknown state	0.24	0.44	0.31	0.68	0.78	0.73	0.39	0.41	0.40
Not-Kannada	0.62	0.60	0.61	-	-	-	-	-	-
Not-Malayalam	-	-	-	0.78	0.74	0.76	-	-	-
Not-Tamil	-	-	-	-	-	-	0.64	0.49	0.55
Weighted Avg.	0.62	0.60	0.61	0.72	0.72	0.72	0.60	0.60	0.60

long-short-term memory (LSTM) when word-level features were used. As a result, we chose a four-layered dense neural network with character-level features to identify sentiments in Dravidian social media posts. To provide input to the dense neural network, experiment with the various combinations of character n-gram TF-IDF features is done. In this extensive experiment, we found that the first 50,000 one to six-gram character-level TF-IDF performs better for the Kannada-English dataset in comparison to the other combinations of n-gram. Similarly, for Malayalam-English and Tamil-English, the first 30,000 one to six-gram character-level TF-IDF features performed better. The extracted features are then passed through a four-layered dense neural network containing 4,096, 512, 64, and 5-neurons at consecutive layers, respectively. To train the proposed model, categorical cross-entropy and Adam is used as the loss function and optimizer, respectively. As the deep learning-based models are very sensitive to the chosen hyper-parameters, we performed a sensitivity analysis by varying the learning rate, batch size, dropout rate, and epochs. The best-suited hyper-parameters for the proposed dense neural network are listed in Table 2. The proposed system was implemented using Keras with Tensorflow as a backend.

4. Results

The performance of the proposed model is measured in terms of precision, recall, and F_1-score. Along with these metrics, we have plotted confusion matrix to show the prediction in each of

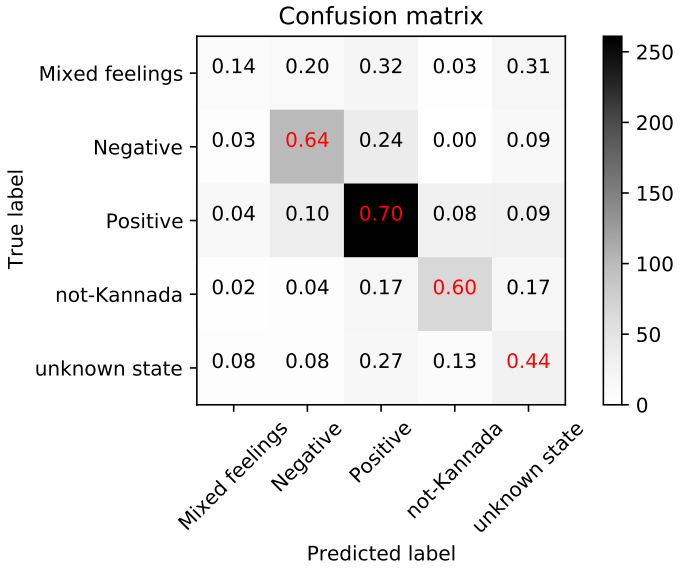

Figure 2: Confusion matrix (Kannada-English)

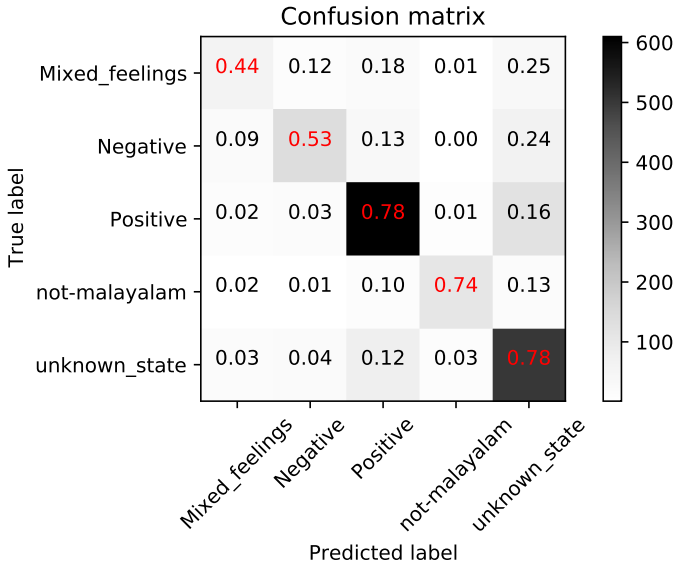

Figure 3: Confusion matrix (Malayalam-English)

the classes. The performance of the proposed model for Kannada-English, Malayalam-English, and Tamil-English datasets is listed in Table3.

In case of Kannada-English dataset, the proposed model has achieved weighted precision of 0.62, recall of 0.60, and F_1-score of 0.61. The confusion matrix for Kannada-English dataset can be seen in Figure 2. From the confusion matrix, it can be seen that the proposed dense neural

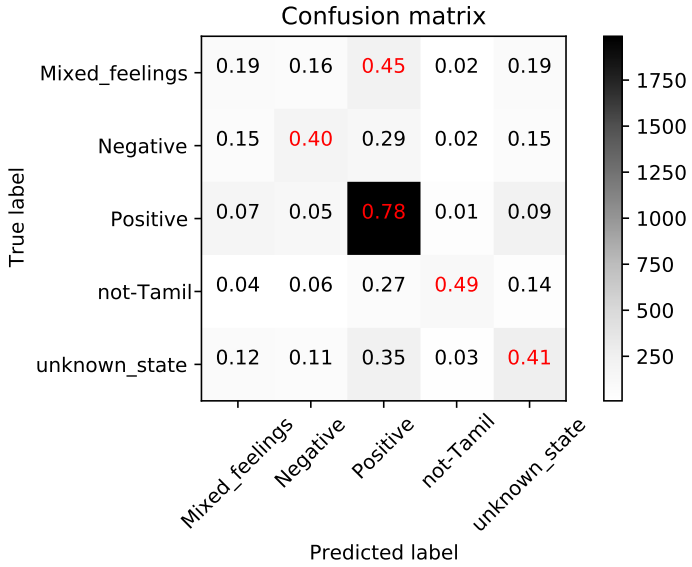

Figure 4: Confusion matrix (Tamil-English)

network perform significantly good for *Positive* class with the recall of 0.70, whereas it not good for *Mixed-feelings* class. The possible reason can be the mixed sentiment text, and the proposed model is not able to clearly distinguishe it. For Malayalam-English dataset, the proposed model is able to achieve a weighted precision, recall, and F_1-score of 0.72 (as can be seen in Table 3. The confusion matrix for Malayalam-English dataset can be seen in Figure 3. Similarly, for Tamil-English dataset, the proposed dense neural network has achieved a weighted precision, recall, and F_1-score of 0.60. The confusion matrix for Tamil-English dataset can be seen in Figure 4.

5. Conclusion

Sentiment analysis of textual content offers a wide range of applications in natural language processing. In this work, we have utilized character-level TF-IDF features with dense neural network to classify social media posts into five different classes. For Kannada-English, Malayalam-English, and Tamil-English social media postings, the suggested model has obtained weighted $F1$-scores of 0.61, 0.72, and 0.60, respectively. Since the usage of character-level TF-IDF features yields promising results, this feature may be further investigated in the future with various deep learning models to build a more robust system.

References

[1] S. Mani, S. Kumari, A. Jain, P. Kumar, Spam review detection using ensemble machine learning, in: International Conference on Machine Learning and Data Mining in Pattern

Recognition, Springer, 2018, pp. 198–209.
[2] K. Gaurav, P. Kumar, Consumer satisfaction rating system using sentiment analysis, in: Conference on e-Business, e-Services and e-Society, Springer, 2017, pp. 400–411.
[3] A. Kumar, S. Saumya, J. P. Singh, NITP-AI-NLP@ HASOC-Dravidian-CodeMix-FIRE2020: A machine learning approach to identify offensive languages from Dravidian code-mixed text., in: FIRE (Working Notes), 2020, pp. 384–390.
[4] A. Kumar, S. Saumya, J. P. Singh, NITP-AI-NLP@ HASOC-FIRE2020: Fine tuned BERT for the hate speech and offensive content identification from social media., in: FIRE (Working Notes), 2020, pp. 266–273.
[5] A. Kumar, J. P. Singh, Location reference identification from tweets during emergencies: A deep learning approach, International journal of disaster risk reduction 33 (2019) 365–375.
[6] A. Kumar, J. P. Singh, S. Saumya, A comparative analysis of machine learning techniques for disaster-related tweet classification, in: 2019 IEEE R10 Humanitarian Technology Conference (R10-HTC)(47129), IEEE, 2019, pp. 222–227.
[7] B. G. Patra, D. Das, A. Das, Sentiment analysis of code-mixed indian languages: An overview of SAIL_Code-Mixed Shared Task@ ICON-2017, arXiv preprint arXiv:1803.06745 (2018).
[8] B. R. Chakravarthi, R. Priyadharshini, V. Muralidaran, S. Suryawanshi, N. Jose, E. Sherly, J. P. McCrae, Overview of the track on sentiment analysis for Dravidian languages in code-mixed text, in: Forum for Information Retrieval Evaluation, 2020, pp. 21–24.
[9] B. R. Chakravarthi, R. Priyadharshini, V. Muralidaran, S. Suryawanshi, N. Jose, J. P. Sherly, Elizabeth McCrae, Overview of the track on Sentiment Analysis for Davidian Languages in Code-Mixed Text, in: Working Notes of the FIRE 2020. CEUR Workshop Proceedings., 2020.
[10] A. Kumar, S. Saumya, J. P. Singh, NITP-AI-NLP@ Dravidian-CodeMix-FIRE2020: A hybrid Cnn and Bi-LSTM network for sentiment analysis of Dravidian code-mixed social media posts., in: FIRE (Working Notes), 2020, pp. 582–590.
[11] A. Joshi, A. Prabhu, M. Shrivastava, V. Varma, Towards sub-word level compositions for sentiment analysis of hindi-english code mixed text, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016, pp. 2482–2491.
[12] L. Advani, C. Lu, S. Maharjan, C1 at SemEval-2020 Task 9: Sentimix: Sentiment analysis for code-mixed social media text using feature engineering, arXiv preprint arXiv:2008.13549 (2020).
[13] K. Goswami, P. Rani, B. R. Chakravarthi, T. Fransen, J. P. McCrae, ULD@ NUIG at SemEval-2020 Task 9: Generative morphemes with an attention model for sentiment analysis in code-mixed text, arXiv preprint arXiv:2008.01545 (2020).
[14] A. Hande, R. Priyadharshini, B. R. Chakravarthi, KanCMD: Kannada CodeMixed dataset for sentiment analysis and offensive language detection, in: Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media, Association for Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 54–63. URL: https://www.aclweb.org/anthology/2020.peoples-1.6.
[15] B. R. Chakravarthi, N. Jose, S. Suryawanshi, E. Sherly, J. P. McCrae, A sentiment analysis dataset for code-mixed Malayalam-English, 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), European Language Resources association, Marseille, France, 2020, pp. 177–184. URL: https://www.aclweb.org/anthology/2020.sltu-1.25.

[16] B. R. Chakravarthi, V. Muralidaran, R. Priyadharshini, J. P. McCrae, 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), European Language Resources association, Marseille, France, 2020, pp. 202–210. URL: https://www.aclweb.org/anthology/2020.sltu-1.28.

[17] B. R. Chakravarthi, R. Priyadharshini, S. Thavareesan, D. Chinnappa, D. Thenmozhi, E. Sherly, J. P. McCrae, A. Hande, R. Ponnusamy, S. Banerjee, C. Vasantharajan, Findings of the Sentiment Analysis of Dravidian Languages in Code-Mixed Text, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.

[18] R. Priyadharshini, B. R. Chakravarthi, S. Thavareesan, D. Chinnappa, D. Thenmozhi, E. Sherly, Overview of the DravidianCodeMix 2021 shared task on sentiment detection in tamil, malayalam, and kannada, in: Forum for Information Retrieval Evaluation, FIRE 2021, Association for Computing Machinery, 2021.

[19] S. Mahata, D. Das, S. Bandyopadhyay, Sentiment classification of code-mixed tweets using bi-directional RNN and language tags, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Kyiv, 2021, pp. 28–35. URL: https://aclanthology.org/2021.dravidianlangtech-1.4.

[20] A. V. Mandalam, Y. Sharma, Sentiment analysis of Dravidian code mixed data, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Kyiv, 2021, pp. 46–54. URL: https://aclanthology.org/2021.dravidianlangtech-1.6.

[21] S. Dowlagar, R. Mamidi, Graph convolutional networks with multi-headed attention for code-mixed sentiment analysis, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Kyiv, 2021, pp. 65–72. URL: https://aclanthology.org/2021.dravidianlangtech-1.8.

[22] F. Balouchzahi, H. L. Shashirekha, LA-SACo: A study of learning approaches for sentiments analysis inCode-mixing texts, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Kyiv, 2021, pp. 109–118. URL: https://aclanthology.org/2021.dravidianlangtech-1.14.