LIDOMA@DravidianLangTech: Convolutional Neural Networks for Studying Correlation Between Lexical Features and Sentiment Polarity in Tamil and Tulu Languages

Moein Shahiki-Tash and **Jesús Armenta-Segura** and **Zahra Ahani** and **Olga Kolesnikova** and **Grigori Sidorov** and **Alexander Gelbukh**
Instituto Politécnico Nacional,
Centro de Investigación en Computación
Mexico City, Mexico.
{mshahikit2022, jarmentas2022, zahani2023, kolesnikova, sidorov, gelbukh}@cic.ipn.mx

Abstract

With the prevalence of code-mixing among speakers of Dravidian languages, DravidianLangTech proposed the shared task on Sentiment Analysis in Tamil and Tulu at RANLP 2023. This paper presents the submission of LIDOMA, which proposes a methodology that combines lexical features and Convolutional Neural Networks (CNNs) to address the challenge. A fine-tuned 6-layered CNN model is employed, achieving macro F1 scores of 0.542 and 0.199 for Tulu and Tamil, respectively.

1 Introduction

In recent years, there has been a significant surge of interest in sentiment analysis on social media platforms for Dravidian languages. The linguistically diverse and multicultural environments in which these languages are spoken have contributed to the prevalence of a linguistic phenomenon known as code-mixing. Code-mixing refers to the occurrence of multiple languages within a single document or utterance (E. Ojo et al., 2022). This phenomenon is particularly prominent in written texts, where non-native scripts and hybrid words combine elements from more than one language.

The Shared Task on Sentiment Analysis in Tamil and Tulu, proposed by DravidianLangTech at RANLP 2023 (B et al., 2023; Hegde et al., 2023), aims to address the challenges associated with sentiment analysis in code-mixed text. This shared task seeks to introduce a new gold standard corpus specifically designed for sentiment analysis in the context of Tamil-English and Tulu-English code-mixing language. Moreover, their dataset (Chakravarthi et al., 2020; Hegde et al., 2022) also has class imbalance problems depicting real-world scenarios.

The main focus of the proposed approach is to identify sentiment polarity in code-mixed comments and posts extracted from social media platforms. These comments and posts often contains more than one sentence, making the sentiment analysis task more complex.

In order to tackle the proposed shared tasks, this paper presents an approach that utilizes lexical features and convolutional neural networks (CNNs) (Fukushima, 1980; LeCun et al., 1989). Lexical features have demonstrated a strong correlation with various pragmatic phenomena, including sentiment analysis tasks such as hope and hate speech detection (Dowlagar and Mamidi, 2021; Balouchzahi et al., 2023). They have also been effective in other pragmatic tasks such as user preferences predictions in entertainment domains (Armenta-Segura and Sidorov, 2023). Additionally, lexical features have shown significant relevance in sentiment analysis when code-mixing is involved, as demonstrated in (E. Ojo et al., 2022) with Kannada and English languages.

On the other hand, CNNs have proven to be effective in detecting relevant features associated with sentiments across different classes (Shahiki-Tash et al., 2023), which is the reason why they were employed on this work. The presented model consists of a 6-layered CNN with the following structure: The first layer generates an embedding from a bag of words vectorization. The second and third layers are convolutional layers designed to learn the lexical features that have the strongest relationship with the labeling of each sample, which in this particular case are *positive*, *negative*, *neutral*, and *mixed feelings*. The fourth and fifth layers help prevent overfitting and reduce the dimensionality of the output by fine-tuning the lexical feature extraction using max pooling (Yamaguchi et al., 1990). Finally, the sixth layer utilizes a sigmoid ac-

Third Workshop on Speech and Language Technologies for Dravidian Languages, pages 180–185, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-085-4_025

tivation function to relate the learned features with the binary golden label. The achieved F1 scores were 0.542 for the Tulu-English dataset and 0.199 for the Tamil-English dataset.

The structure of this paper is as follows: in Section 2 it is described some state-of-the-art works on sentiment polarity detection. In Section 3, the methodology is detailed. In Section 4, it is provided a brief description of both datasets, and the experimental workflow is outlined. In Section 5, it is discussed the results of the experiments. Finally, in Section 6, the paper is concluded.

2 Related Work

Sentiment polarity analysis is considered one of the pioneering tasks in computational sentiment analysis. One of the earliest approaches in this field are the General Inquirer (Stone and Hunt, 1963), which is a 1961 IBM system capable to perform content analysis for behavioral sciences, most particular pattern detection in text for categorizing words according to their semantics, related to positive or negative sentiments. In 1997, a most focused approach was proposed with the system Smokey (Spertus, 1997), designed to detect abusive messages by using a rule-based approach to identify offensive language and contexts.

Following on the line of negative sentiment detection, in (Warner and Hirschberg, 2012), the authors proposed a lexicon-based approach for hate speech detection. Their approach focused on analyzing the sense in which selected words were used in sentences to identify hateful or offensive content, making the task close similar to word sense disambiguation. However, they discovered that this hypothesis is vulnerable when faced with incomplete datasets, especially in cases where a word only appears in one type of speech.

On the other hand, in the domain of positive speech, a notable line of research is the peace speech line initiated in (Palakodety et al., 2019b,a), where the authors primarily focused on analyzing peace-oriented discourse, particularly in the context of a conflict between Pakistan and India.

Furthermore, in (Chakravarthi, 2020), the authors focuses more towards the themes of equality, diversity, and inclusion. Notably, Chakravarthi also organized a series of shared tasks (Chakravarthi et al., 2022; Chakravarthi and Muralidaran, 2021), where team LIDOMA utilized a Convolutional Neural Network (CNN) to address the specified task (Shahiki-Tash et al., 2023). This model is a variation to the model presented in this paper.

About code-mixing detection, several computational approaches have been done to address the task in languages from India. For instance, in (Shekhar et al., 2020), the authors worked on code-mixing between Hindi and English, presenting a methodology for language identification in a dataset comprising Facebook, Twitter, and WhatsApp messages. In (Patwa et al., 2020), the authors proposed a shared task at SemEval-2020, in which team LIMSI_UPV (Banerjee et al., 2020) proposed a recurrent convolutional neural network architecture to address the task. In (Ansari et al., 2021), the authors expanded this line by incorporating Urdu into the analysis and utilizing transformer models with attention mechanisms, specifically employing BERT models.

In (Yasir et al., 2021), the authors considered code-mixing involving Saraiki and Bengali. They employed recurrent neural networks and word vectorizations to address the task of language identification in code-mixed texts.

In (Dutta, 2022), the author proposed a setting that aligns closely with the shared tasks mentioned earlier, but with a focus on English-Hindi and English-Bengali code-mixing. Additionally, she introduced an index to measure the level of mixing within the corpora, providing insights into the degree of code-mixing present in the data.

Furthermore, in (E. Ojo et al., 2022), the authors proposed an n-gram-based approach to tackle the task of language identification in Kannada-English code-mixed texts.

3 Methodology

Diving further in the structure outlined in the introduction, the overall followed procedure is explained now, along with the used hiperparameters.

3.1 Preprocessing

All samples written in the latin alphabet were preprocessed by lowercasing and removing special characters. All samples containing kannadian, Tamil and Tulu alphabet characters were letting intact. All URL patterns were removed in all samples. This process helped to enhance the results due to the noise reduction, as in (Shahiki-Tash et al., 2023). After that, word-based tokenization was performed creating a Bag-of-Words representation, ready to be feeded into the first layer of the 6-

layered CNN (see Figure 1 for a summary and an example).

3.2 Layers of the network

The first layer of the CNN embeds the input tokens into a dense vector representation, capturing semantic relationships between them, in a straightforward standard way to convert text into vectors. Concretely, it maps the bag-of-words tokens into 32-dimentional dense vectors. The layer allows a maximum of 2000 features and processes sequences with a maximum length of 40 tokens. Additionally, it applies L_2 regularization with a strength of 0.0005 to the embedding weights. All these hyperparameters were determined through a trial and error fine-tuning process, picking the ones who brought better results. In general, all hyperparameters for every layer in this model were determined in this same fashion.

The second layer is convolutional with small kernels of size 3, allowing it to capture better local parameters. Also, it consists of 128 filters. The kernel regularizer was L_2, with a strength of 0.0005 to the output weights. To prevent overfitting, a bias regularizer is also applied, which is the same as the one applied to the kernels. The chosen activation function for this layer is ReLu (Fukushima, 1969), which maps a value x to $Max\{0, x\}$.

The third layer is similar to the second, but it employs half the number of filters. We included it aiming to refine the output of the second layer.

The fourth layer is a Flatten layer. Its purpose is to reshape the input data to a flat one-dimensional representation, required for the employment of a dense layer.

The fifth layer is a 32-dimensioned dense layer with ReLu as activation function. It also includes a L_2 regularizer for the kernels and a bias reguarizer, both with strenght of 0.001. Its function is to convert the vector into a suitable string able to become a prediction in the last layer.

Finally, the output layer is 4-dimentional and has a sigmoid activation function (Cramer, 2002; Verhulst, 1845). It also includes the same regularizers as the previous dense layer.

4 Experimental Setup

4.1 Data

The Tulu training set contains $6,457$ samples with labels Positive, Neutral, Negative and Mixed Feelings. The Tamil training set contains $33,989$ sam-

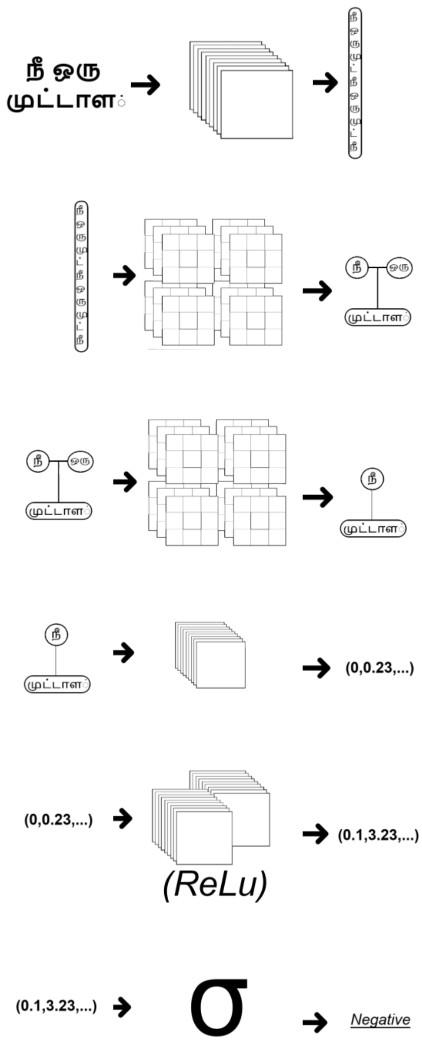

Figure 1: From top to down, ilustrations of the six layers of our CNN model. The example text can be written in latin alphabet as *Nī oru muṭṭāḷ*, which means *you are an idiot* in Tamil. In the first layer, the tokenized text is converted into a dense vector. In the second and third layer, the 3×3 kernels extracts patterns relevant to the golden labels (in this example, represented as a link betweet the tokens *Nī* and *muṭṭāḷ* -*you* and *idiot*-). The fourth layer convert these patterns into a vector. The fifth layer uses ReLu and, finally, the sixth layer makes a prediction using the sigmoid function. The final output can be *positive*, *negative*, *neutral* and *mixed feelings*.

Figure 2: Label distribution among the training sets. Recall that Unknown State corresponds to Neutral in the Tamil training set.

Sample	Polarity
Vani bhojam fans hit like solli 500 like Vangida Vendiyathu than	Neutral
Ithu yethu maathiri illama puthu maathiyaala irukku	Positive
Wow! Back to Baasha mode. thalaivaaaa. petta paraakkkkk	Negative
Kaagam karaindhu koodi unnum, Manidham ennum moodar koodam koodi serdhu pagaimai kollum... Idil yaar uyarthinai yaar agrinai	Mixed Feelings

Table 1: Latin alphabet examples from the Tamil training sets.

ples with labels Positive, Unknown State (Neutral), Negative and Mixed Feelings. In Figure 2 it is shown the distribution of every sample, along with the precise number of samples for each class. In Table 2 it is shown examples per label in the Tulu training set. In Table 1 it is shown for the Tamil training set.

4.2 Experimental Workflow

Every dataset was splitted into a 75 : 25 ratio for training the model. The CNN was trained during 30 epochs.

5 Results

After the 30-epoch training, the model achieved a macro F1 score of 0.516 in the Tulu evaluation set, and 0.199 in the Tamil evaluation set. The most important factors for these results were the notable differences between kannada, Tamil, Tulu and latin alphabets, in which this network was designed, and the nature of the labelling: regardless previous experiences where variations of this CNN was employed, the datasets employed for this task includes the categories of Neutral and Mixed Feelings, while in the other sentiment analysis tasks the labelling was binary in terms of a single polarity,

Sample	Polarity
Bega 2 nd part padle	Neutral
Devdas kapikad no1	Positive
Enchi pankda comedy	Negative
Yan 4 class d uppunaga kallamundkur du thutina cha parka thandada suruta drama	Mixed Feelings

Table 2: Latin alphabet examples from the Tulu training sets

and not mixing it.

Another important factor was the balance of the dataset. As shown in Figure 2, there is a high imbalance in the dataset which led to a general low performance in the proposed methods, being macro F1-score of 0.32 the best for Tamil and 0.542 the best for Tulu, not so far of our results.

6 Conclusions

In this paper, it was presented the LIDOMA submission for the shared task on Sentiment Analysis in Tamil and Tulu, proposed by DravidianLangTech at RANLP2023. They employed CNN's, who have proven being effective in sentiment polarity tasks.

The proposed methodology involved the conversion of labels into categorical values, then basic preprocessing of the samples and finally the training of a 6-layered CNN. The findings highlight the complexities involved in handling non-balanced datasets along with the merge of polarities within the *Mixed Feelings* cathegory.

Future work will focus on adapt the CNN architecture to deal better with mixed cathegories, along with adding more steps of preprocessing adapted to kannada, Tamil and Tulu alphabets. Also, it is possible to add the use of atention mechanisms to enhance results in this and other similar datasets.

Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20232138, 20232080, 20231567 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

Mohd Zeeshan Ansari, M M Sufyan Beg, Tanvir Ahmad, Mohd Jazib Khan, and Ghazali Wasim. 2021. Language identification of hindi-english tweets using code-mixed bert. In *2021 IEEE 20th International Conference on Cognitive Informatics Cognitive Computing (ICCI*CC)*, pages 248–252.

Jesús Armenta-Segura and Grigori Sidorov. 2023. Anime Success Prediction Based on Synopsis Using Traditional Classifiers. In *Proceedings of Congreso Mexicano de Inteligencia Artificial, COMIA*.

Premjith B, Jyothish Lal G, Sowmya V, Bharathi Raja Chakravarthi, Rajeswari Natarajan, Nandhini K, Abirami Murugappan, Bharathi B, Kaushik M, Prasanth SN, Aswin Raj R, and Vijai Simmon S. 2023. Findings of the shared task on multimodal abusive language detection and sentiment analysis in tamil and malayalam. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Fazlourrahman Balouchzahi, Grigori Sidorov, and Alexander Gelbukh. 2023. Polyhope: Two-level hope speech detection from tweets. *Expert Systems with Applications*, 225:120078.

Somnath Banerjee, Sahar Ghannay, Sophie Rosset, Anne Vilnat, and Paolo Rosso. 2020. LIMSI_UPV at SemEval-2020 task 9: Recurrent convolutional neural network for code-mixed sentiment analysis. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1281–1287, Barcelona (online). International Committee for Computational Linguistics.

Bharathi Raja Chakravarthi. 2020. Hopeedi: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*.

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, Subalalitha Cn, John McCrae, Miguel Ángel García, Salud María Jiménez-Zafra, Rafael Valencia-García, Prasanna Kumaresan, Rahul Ponnusamy, Daniel García-Baena, and José García-Díaz. 2022. Overview of the shared task on hope speech detection for equality, diversity, and inclusion. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 378–388, Dublin, Ireland. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020. 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.

J. S. Cramer. 2002. The origins of logistic regression. *Econometrics eJournal*.

Suman Dowlagar and Radhika Mamidi. 2021. Edione@lt-edi-eacl2021: Pre-trained transformers with convolutional neural networks for hope speech detection. *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, page 86 – 91. Cited by: 6.

Aparna Dutta. 2022. Word-level language identification using subword embeddings for code-mixed Bangla-English social media data. In *Proceedings of the Workshop on Dataset Creation for Lower-Resourced Languages within the 13th Language Resources and Evaluation Conference*, pages 76–82, Marseille, France. European Language Resources Association.

O. E. Ojo, A. Gelbukh, H. Calvo, A. Feldman, O. O. Adebanji, and J. Armenta-Segura. 2022. Language identification at the word level in code-mixed texts using character sequence and word embedding. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON): Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts*, pages 1–6, IIIT Delhi, New Delhi, India. Association for Computational Linguistics.

Kunihiko Fukushima. 1969. Visual feature extraction by a multilayered network of analog threshold elements. *IEEE Transactions on Systems Science and Cybernetics*, 5(4):322–333.

Kunihiko Fukushima. 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202.

Asha Hegde, Mudoor Devadas Anusha, Sharal Coelho, Hosahalli Lakshmaiah Shashirekha, and Bharathi Raja Chakravarthi. 2022. Corpus creation for sentiment analysis in code-mixed Tulu text. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 33–40, Marseille, France. European Language Resources Association.

Chakravarthi Bharathi Raja Hegde, Asha, Hosahalli Lakshmaiah Shashirekha, Rahul Ponnusamy, SUBALALITHA CN, Lavanya S K, Thenmozhi D, Martha Karunakar, Shreya Shreeram, and Sarah Aymen. 2023. Findings of the shared task on sentiment analysis in tamil and tulu code-mixed text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.

Shriphani Palakodety, Ashiqur R. KhudaBukhsh, and Jaime G. Carbonell. 2019a. Hope speech detection: A computational analysis of the voice of peace. In *European Conference on Artificial Intelligence*.

Shriphani Palakodety, Ashiqur R. KhudaBukhsh, and Jaime G. Carbonell. 2019b. Voice for the voiceless: Active sampling to detect comments supporting the rohingyas. In *AAAI Conference on Artificial Intelligence*.

Parth Patwa, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, Srinivas PYKL, Björn Gambäck, Tanmoy Chakraborty, Thamar Solorio, and Amitava Das. 2020. SemEval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 774–790, Barcelona (online). International Committee for Computational Linguistics.

Moein Shahiki-Tash, Jesús Armenta-Segura, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023. LIDOMA at HOPE2023@IberLEF: Hope Speech Detection Using Lexical Features and Convolutional Neural Networks. In *Proceedings of IberLEF*.

Shashi Shekhar, Dilip Kumar Sharma, and MM Sufyan Beg. 2020. Language identification framework in code-mixed social media text based on quantum lstm—the word belongs to which language? *Modern Physics Letters B*, 34(06):2050086.

Ellen Spertus. 1997. Smokey: Automatic recognition of hostile messages. In *AAAI/IAAI*.

Philip J. Stone and Earl B. Hunt. 1963. A computer approach to content analysis: studies using the general inquirer system. *Proceedings of the May 21-23, 1963, spring joint computer conference*.

P.F. Verhulst. 1845. *Recherches mathématiques sur la loi d'accroissement de la population, par P.F. Verhulst ...* M. Hayez.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, Canada. Association for Computational Linguistics.

Kouichi Yamaguchi, Kenji Sakamoto, Toshio Akabane, and Yoshiji Fujimoto. 1990. A neural network for speaker-independent isolated word recognition. *First International Conference on Spoken Language Processing (ICSLP 1990)*.

Muhammad Yasir, Li Chen, Amna Khatoon, Muhammad Amir Malik, and Fazeel Abid. 2021. Mixed script identification using automated dnn hyperparameter optimization. *Computational Intelligence and Neuroscience*.