ML&AI_IIITRanchi@DravidianLangTech: Fine-Tuning of Indic-BERT for Exploring Language-Specific Features for Sentiment Classification in Code-Mixed Dravidian Language

Kirti Kumari
IIIT Ranchi, India
kirti@iiitranchi.ac.in

Shirish Shekhar Jha
IISER Bhopal, India
shirish20@iiserb.ac.in

Zarikunte Kunal Dayanand
IISER Bhopal, India
zarikunte20@iiserb.ac.in

Praneesh Sharma
KIIT, Bhubaneswar, India
praneeshsharma30@gmail.com

Abstract

Code-mixing presents challenges to sentiment analysis due to limited availability of annotated data found on low-resource languages such as Tulu. To address this issue, comprehensive work was done in creating a gold-standard labeled corpus that incorporates both languages while facilitating accurate analyses of sentiments involved. Encapsulated within this research was the employed use of varied techniques including data collection, cleaning processes as well as preprocessing leading up to effective annotation along with finding results using fine tuning Indic-BERT and performing experiments over TF-IDF and Bag-of-Words. The outcome is an invaluable resource for developing custom-tailored models meant solely for analyzing sentiments involved with code mixed texts across Tamil and Tulu domain limits; allowing a focused insight into what makes up such expressions. Remarkably, the adoption of hybrid models yielded promising outcomes, culminating in a 10^{th} rank achievement for Tulu, and a 14^{th} rank achievement for Tamil, supported by an macro F1-Score of 0.471 and 0.124, respectively.

1 Introduction

Sentiment analysis is instrumental when it comes to understanding subjective information contained within written language data. Due to extensive language mixing within the realm of electronic communications lately, it's critical that we employ sentiment analysis in practice. This study focuses primarily on conducting a thorough examination of how best sentiments are analyzed when it comes to Tulu and Tamil code-mixed texts since these two languages typically fuse regularly across various social media sites. Code-mixed (Hedge et al., 2023) text becomes uniquely challenging when it blends multiple languages within one sentence or utterance. The addition of Tulu and Tamil to this mix together with English or other regional tongues significantly complicates any attempts at sentiment analysis. One barrier that stands out is the absence of resources as well as annotated datasets for code-mixed Tulu and Tamil scripts' emotion assessment. Consequently, there is an urgent need to compile a corpus that can enable effective interpretation of sentiments revealed through these dialects as well as guidelines on how best to proceed with analyses. In this research paper, we propose an inclusive approach towards developing such a corpus through diverse data collection channels ranging from social media platforms, online discussion forums down other digital sources that exhibit diverse language mixing designs encompassed by these texts; subsequently utilizing human experts' annotation skills to deliver precision sentiment labels required for designing valid models capable of evaluating such analyses. To increase the precision of sentimental analyses applied to mixed-language Tulu-Tamil texts, an interdisciplinary team plans on exploring several critical linguistic and contextual features. They intend on investigating methods like customized lexicons geared towards capturing each language's subtle nuances in expressing emotions - such as part-of-speech-tagging approaches or syntactic parsing methods - amongst others. Leveraging machine-learning-algorithms along with cutting-edge natural-language-processing technologies will enable them to construct reliable models capable of acutely categorizing various emotional states present within mixed-language texts. This research paper has practical implications that could benefit future studies on emotional analysis of code-mixed Tulu and Tamil language texts.

Firstly, the data set developed during this research provides useful information that can be harnessed by researchers exploring this field in greater depth. Secondly, our proposed methodology improves ac-

Third Workshop on Speech and Language Technologies for Dravidian Languages,
pages 192–197, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-085-4_027

curacy to capture sentiments expressed within hybrid language forms so that results may provide more accurate readings when compared with previous studies or models. Lastly, our insights into developing sentiment-aware applications could have wide-ranging potential benefits for social media monitoring tools, mental health assessment platforms and customer feedback systems. In general, through the utilization of machine-aided analyses to expand upon essential linguistic features, we successfully tackled gaps in current comprehension related to mixed language forms.

Our work is summarised in the following sections. Related Work, it outlines the work that has been done in the related field, Datasets summarises the efforts and intricacies about the dataset on which the task has been performed, Experimental Results and Discussion, discusses about the various methodologies employed and their results obtained, lastly in Section 5, we wrap up our discussion and findings.

2 Related Work

In recent times due to the growth of social media, Sentiment Analysis referred as SA in the text has become significantly important and there is extensive research being carried out on SA of monolingual texts belonging to high-resource languages such as English, German, Russian. However, only few work have been reported on SA in Tulu and Tamil languages and very less number of SA works are found for other Dravidian languages too.

The authors of (Rani et al., 2020) have become imperative to identify hate speech in social media communication in order to avoid confrontations and control bad behaviour. Using techniques created for monolingual datasets to detect hate speech offers a problem because of the presence of multilingual speakers who regularly transition between languages. In this work, the aim was to analyse, identify, and compare hate speech in text from code-mixed social media platforms. Additionally, they collected a dataset of posts and comments from Facebook and Twitter that are code-mixed with text in Hindi and English. Their test findings show that deep learning models that have been trained on this code-mixed corpus perform better.

The research (Agrawal et al., 2018) focuses on finding a solution to the difficulty of handling idioms in Natural Language Processing (NLP) jobs specifically focused at Indian languages. The goal of this study is to close the gap between Indian languages and NLP applications in idiom handling.The authors offer a thorough analysis of idiomatic expressions used in Indian languages and suggest a cutting-edge method for handling and interpreting idioms in the context of NLP tasks. They use linguistic tools like dictionaries and corpora to build a database of idioms in the Indian language. The information in this repository is quite helpful for deciphering and analysing idiomatic idioms.The research also presents a framework for idiom disambiguation that takes into account the syntactic and semantic characteristics of idioms in Indian languages. The suggested framework uses statistical and rule-based methods to separate the meanings of idioms in various circumstances and the authors test their methodology through tests on several NLP tasks, including sentiment analysis and machine translation, utilising datasets in Indian languages. The outcomes show how their method to handling idioms was successful in enhancing these tasks' performance.The accuracy and effectiveness of NLP applications for Indian languages are improved since it offers insights and methods for handling idioms.

The research (Cieliebak et al., 2017) outlines the investigation into the development of a thorough corpus and benchmark resources designed exclusively for sentiment analysis in the German language utilizing Twitter data. Given the distinct linguistic traits and contextual complexities of the German language, the authors recognized the necessity for high-quality datasets and evaluation measures in the field of German sentiment analysis. They compiled a sizable German Twitter corpus by gathering tweets that contain sentiment-related hashtags in order to fill this gap. The corpus includes a wide range of subjects and emotions covered by German tweets.Apart from creating the corpus, the authors of the paper also constructed benchmark resources specifically designed for German sentiment analysis. They performed manual annotation on a portion of the collected tweets, assigning sentiment labels to ensure the availability of labeled data for training and evaluating sentiment analysis models. The labeling process involved categorizing the tweets as positive, negative, or neutral based on their underlying sentiment. To evaluate how well sentiment analysis models perform on the German Twitter corpus, the authors introduced evaluation

metrics that are specifically tailored for German sentiment analysis. These metrics take into consideration the intricacies and subtle nuances of sentiment expression in the German language, offering a comprehensive and accurate evaluation framework.

The research (Jiang et al., 2019) centers around tackling aspect-based sentiment analysis (ABSA) through the introduction of a challenge dataset and the proposal of effective models. The authors acknowledged the necessity for high-quality datasets and robust models in ABSA, which involves discerning sentiment polarity towards specific attributes or aspects of a given target entity. To overcome the limitations of current datasets, they meticulously curated a challenge dataset tailored to test the capabilities of ABSA models. This challenge dataset encompasses a diverse array of reviews from various domains such as restaurants, laptops, and hotels. Each review is meticulously annotated with sentiment labels at the aspect-level, indicating the sentiment polarity associated with different aspects mentioned within the review. The dataset provides a comprehensive evaluation framework for assessing the performance of ABSA models. Additionally, the authors put forth effective models for ABSA. They introduce an attention-based neural network model that harnesses the contextual information of words to capture sentiment towards different aspects. This model adeptly addresses the challenges posed by the intricate relationships between aspects and sentiments in ABSA tasks. To assess the proposed model and compare its performance against existing approaches, the authors conducted extensive experiments utilizing the challenge dataset. The results demonstrated that their model surpasses several state-of-the-art models in aspect-based sentiment analysis, underscoring its effectiveness and superiority.

The authors of (Rogers et al., 2018) present research conducted to create a rich sentiment analysis dataset specifically designed for Russian-language social media content. The authors recognized the need for high-quality datasets that meet the unique challenges of sentiment analysis in Russian social media. To overcome the limitations of existing datasets, they developed RuSentiment. This is a large dataset carefully designed to capture the complex nuances and characteristics of emotional expression in Russian social media texts. The RuSentiment dataset contains a diverse collection of Russian social media posts from platforms such as Twitter, VKontakte, and LiveJournal. The dataset covers a wide range of topics and includes various sentiment categories such as positive, negative, neutral, and ambiguous. Additionally, the dataset is enriched with additional annotations such as sentiment strength, sentiment goal, and sentiment frame. Sentiment Strength indicates the strength of the sentiment conveyed in each post, while Sentiment Target identifies the specific entity or aspect that the sentiment targets. Emotion frames provide contextual information about the situations in which emotions are expressed. To ensure the quality and reliability of the dataset, the authors used a careful annotation process involving multiple annotators and checked agreement between the annotators. We have also developed customized guidelines and annotation schemes specifically for sentiment analysis in Russian social media. This paper also describes a baseline experiment using the RuSentiment dataset to evaluate the performance of our sentiment analysis model. The results demonstrate the effectiveness of the dataset in capturing the complex nuances of emotional expression in Russian social media and highlight its potential in developing advanced sentiment analysis models.

The research (Mandalam and Sharma, 2021) presented a method for implementation to classify Dravidian code-shuffled comments according to their polarity. Due to the availability of Tamil and Malayalam datasets with mixed codes, his three methods are proposed: subword-level models, word-embedding-based models, and machine-learning-based architectures. Subword and word embedding-based models use Long Short Term Memory (LSTM) networks with language-specific preprocessing, while machine learning models use Term Frequency-Inverse Document Frequency (TF-IDF) vectors with logistic regression models conversion was used.

In the article (Gupta et al., 2021), the authors explore two popular approaches, namely task-specific pre-training and cross-lingual transfer, for handling code-switched data in the context of sentiment analysis. Specifically, they focus on two Dravidian Code-Switched languages, Tamil-English and Malayalam-English, and evaluate the performance of four different BERT-based models.The goal was to compare the impact of task-specific pre-training and cross-lingual transfer on sentiment analysis tasks. The authors find that task-specific

pre-training yields superior results in zero-shot and supervised settings compared to leveraging cross-lingual transfer from multilingual BERT models.

By conducting experiments on our newly created sentiment analysis corpus, we aim to evaluate the performance of fine-tuned Indic-BERT models in comparison to TF-IDF and Bag-of-Words approaches. These experiments will provide insights into the strengths and limitations of each method and their suitability for sentiment analysis in code-mixed Tulu and Tamil text.

3 Datasets

The dataset (Hegde et al., 2022; Chakravarthi et al., 2020, 2022) used for this task aims to address the challenge of identifying sentiment polarity in code-mixed comments and posts in two language pairs: Tamil-English and Tulu-English. These comments and posts were collected from social media platforms, providing a real-world context for sentiment analysis research. The dataset contains annotations of sentiment polarity at the comment or post level, allowing for a comprehensive analysis of the overall sentiment expressed in each instance.

One notable characteristic of the dataset is the sentence structure. While a comment or post may consist of multiple sentences, the average sentence length across the corpora is approximately one sentence. This aspect of the dataset simplifies the task by focusing on individual comments or posts as self-contained units for sentiment analysis.

The dataset captures the intricacies of code-mixing, which refers to the phenomenon of mixing two or more languages within a single communication instance. In this case, the dataset specifically focuses on Tamil-English and Tulu-English code-mixing. Code-mixing is prevalent in multilingual societies, particularly in social media conversations, and understanding sentiment in such mixed-language contexts is crucial for gaining insights into users' opinions and attitudes.

One important consideration when working with this dataset is the presence of class imbalance. Class imbalance refers to an uneven distribution of sentiment polarities in the dataset, where certain sentiments may be over–represented while others are underrepresented. This class imbalance accurately reflects real–world scenarios, as sentiment distribution in social media is often skewed, with certain sentiments being more prevalent than others. Addressing class imbalance is a significant challenge in sentiment analysis, and this dataset has provided an opportunity to explore and develop effective techniques to handle this issue.

Researchers and practitioners can leverage this dataset for various tasks related to sentiment analysis, such as sentiment classification, sentiment intensity estimation, or sentiment trend analysis. The availability of sentiment annotations at the comment or post level enables a fine-grained analysis of sentiment in code-mixed social media content. By utilizing this dataset, one can gain insights into the sentiment patterns and dynamics within the Tamil-English and Tulu-English code-mixed social media landscape.

The dataset for us served as a valuable resource for training and evaluating sentiment analysis models. We were able to develop and fine-tune machine learning or deep learning models using this dataset to improve the accuracy and performance of sentiment analysis algorithms in code-mixed contexts. Additionally, the presence of class imbalance in the dataset provided an opportunity to explore techniques for handling imbalanced data and improving model robustness in real-world scenarios.

In terms of potential challenges, code-mixing introduces linguistic complexities that may pose difficulties for sentiment analysis. Sentiment expression can vary depending on the language used, and the presence of code-mixed phrases, idioms, or cultural references adds an additional layer of complexity to sentiment interpretation. While working with this dataset we considered these linguistic intricacies and explore techniques that effectively capture sentiment in code-mixed contexts.

Furthermore, the dataset's domain, which comprises social media comments and posts, presents its own set of challenges. Social media platforms are characterized by informal language, abbreviations, emoticons, and non-standard grammar, which can impact sentiment analysis accuracy. It was crucial for us to account for these unique characteristics and develop models that can effectively handle the noisy nature of social media data.

4 Experimental Results and Discussion

To evaluate the performance of sentiment analysis models, fine-tuning of Indic-BERT was employed.

Fine-tuning of In-Domain contextualized BERT (Indic-BERT): the fine-tuning of indic BERT

Table 1: Indic-BERT results on the test dataset

Method	Macro F1-Score
Indic-BERT Tamil	0.124
Indic-BERT Tulu	0.471

demonstrated notable improvements in sentiment classification accuracy. By leveraging pre-trained language representations and fine-tuning them on the code-mixed Tulu and Tamil sentiment analysis task, the model achieved enhanced performance in capturing the nuances of sentiment expressed in the code-mixed texts.

Indic-BERT (Kakwani et al., 2020), an ALBERT model, has been exclusively pretrained on 12 major Indian languages. It underwent pre-training using unique monolingual corpus comprising approximately 9 billion tokens, followed by evaluation on diverse tasks. Despite having fewer parameters compared to other multilingual models like mBERT and XLM-R, Indic-BERT achieves comparable or superior performance. It utilizes a vast corpus of text data from Indic languages to acquire contextual representations of words and sentences. Through training on a diverse range of Indic language data, the model captures the specific linguistic patterns, syntactic structures, and semantic relationships inherent in these languages. This empowers the model to comprehend and generate meaningful representations for Indic text.

Notably, Indic-BERT excels in handling code-mixed text, where multiple languages are combined within a single instance of communication. Indic-BERT has been purposefully trained to effectively handle code-mixed text, rendering it suitable for a wide array of applications involving mixed-language data.

The findings in Table 1 highlight the significance of contextualized language models like Indic-BERT in capturing the complex sentiments expressed in code-mixed texts. The fine-tuning process enables the model to adapt specifically to the characteristics of Tulu and Tamil code-mixed data, leading to improved classification performance. Figure 1 shows the utilized architecture of our model.

The utilization of the current corpus and the findings of this research have significant implications for sentiment analysis in code-mixed Tulu and Tamil texts.

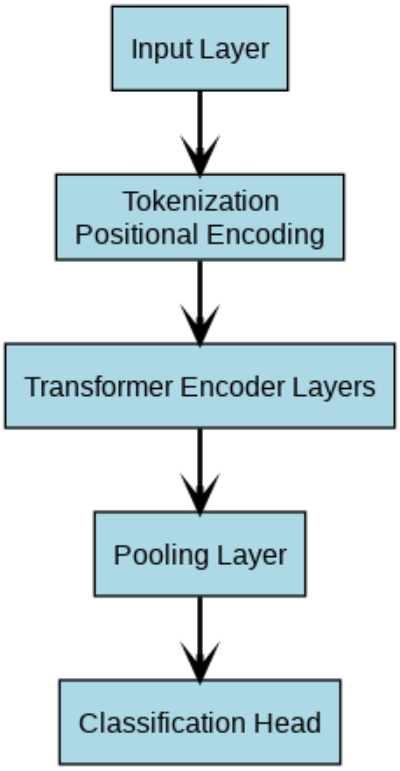

Figure 1: Utilized Indic-BERT Architecture

5 Conclusion

The corpus serves as a valuable resource for future studies and advancements in sentiment analysis techniques for code-mixed languages. Researchers can utilize this corpus to develop and evaluate more accurate sentiment analysis models specific to Tulu and Tamil code-mixed text. Moreover, the incorporation of fine-tuned Indic-BERT and the exploration of traditional approaches contribute to the existing body of knowledge in sentiment analysis. These findings can guide future research in developing effective sentiment analysis models for code-mixed languages.And also the outcomes of this research have practical implications for various applications. The accurate sentiment analysis of code-mixed Tulu and Tamil texts can be leveraged in brand reputation management, political analysis, customer feedback analysis, and mental health monitoring systems. By understanding and interpreting the sentiments expressed in these languages, decision-makers can make informed decisions and provide necessary support.

Now concluding, the research highlights the importance of corpus used for sentiment analysis in code-mixed Tulu and Tamil texts. The incorporation of fine-tuned Indic-BERT and experimentation

with traditional approaches contributes to the advancement of sentiment analysis techniques. The developed corpus and the obtained results provide valuable resources for future studies and applications in sentiment analysis, benefiting various domains that rely on accurate sentiment interpretation in code-mixed Tulu and Tamil texts.

Acknowledgements

We are thankful to Indian Institute of Information Technology Ranchi for all the support during our research.

References

Ruchit Agrawal, Vighnesh Chenthil Kumar, Vigneshwaran Muralidharan, and Dipti Misra Sharma. 2018. No more beating about the bush: A step towards idiom handling for indian language nlp. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

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.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P McCrae. 2022. Dravidiancodemix: Sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text. *Language Resources and Evaluation*, 56(3):765–806.

Mark Cieliebak, Jan Milan Deriu, Dominic Egger, and Fatih Uzdilli. 2017. A twitter corpus and benchmark resources for german sentiment analysis. In *5th International Workshop on Natural Language Processing for Social Media, Boston MA, USA, 11 December 2017*, pages 45–51. Association for Computational Linguistics.

Akshat Gupta, Sai Krishna Rallabandi, and Alan Black. 2021. Task-specific pre-training and cross lingual transfer for code-switched data. *arXiv preprint arXiv:2102.12407*.

Asha Hedge, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rahul Ponnusamy, Subalalitha Chinnaudayar Navaneethakrishnan, Lavanya S.K, Durairaj Thenmozhi, Martha Karunakar, Shreya Sriram, 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.

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.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. A challenge dataset and effective models for aspect-based sentiment analysis. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 6280–6285.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*.

Asrita Venkata Mandalam and Yashvardhan Sharma. 2021. Sentiment analysis of Dravidian code mixed data. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 46–54, Kyiv. Association for Computational Linguistics.

Priya Rani, Shardul Suryawanshi, Koustava Goswami, Bharathi Raja Chakravarthi, Theodorus Fransen, and John Philip McCrae. 2020. A comparative study of different state-of-the-art hate speech detection methods in hindi-english code-mixed data. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, pages 42–48.

Anna Rogers, Alexey Romanov, Anna Rumshisky, Svitlana Volkova, Mikhail Gronas, and Alex Gribov. 2018. Rusentiment: An enriched sentiment analysis dataset for social media in russian. In *Proceedings of the 27th international conference on computational linguistics*, pages 755–763.