Eureka-CIOL@DravidianLangTech 2025: Using Customized BERTs for Sentiment Analysis of Tamil Political Comments

Enjamamul Haque Eram, Anisha Ahmed, Sabrina Afroz Mitu, Azmine Toushik Wasi[†]
Shahjalal University of Science and Technology, Sylhet, Bangladesh
[†]**Correspondence:** azmine32@student.sust.edu

Abstract

Sentiment analysis on social media platforms plays a crucial role in understanding public opinion and the decision-making process on political matters. As a significant number of individuals express their views on social media, analyzing these opinions is essential for monitoring political trends and assessing voter sentiment. However, sentiment analysis for low-resource languages, such as Tamil, presents considerable challenges due to the limited availability of annotated datasets and linguistic complexities. To address this gap, we utilize a novel dataset encompassing seven sentiment classes, offering a unique opportunity to explore sentiment variations in Tamil political discourse. In this study, we evaluate multiple pre-trained models from the Hugging Face library and experiment with various hyperparameter configurations to optimize model performance. Our findings aim to contribute to the development of more effective sentiment analysis tools tailored for low-resource languages, ultimately empowering Tamil-speaking communities by providing deeper insights into their political sentiments. Our full experimental codebase is publicly available at: ciol-researchlab/NAACL25-Eureka-Sentiment-Analysis-Tamil

1 Introduction

Sentiment analysis is a crucial aspect of Natural Language Processing (NLP) that facilitates the categorization of textual opinions into various sentiment classes, such as positive, negative, neutral, and more. It has significant applications in understanding political discourse, enabling tasks such as forecasting election outcomes, analyzing public sentiment, and formulating targeted policies. Social media platforms like X (formerly Twitter) have gained immense popularity and have become a major hub for political discussions. In India, a substantial number of individuals actively use X to express their thoughts and opinions on political

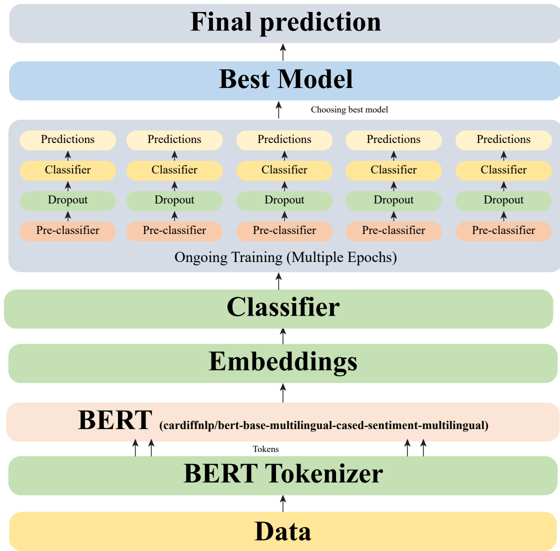

Figure 1: Model architecture, containing tokenizer, pre-trained model, classifier and other components

issues. As of 2018, the platform reported over 321 million monthly active users (Wang et al., 2012), with approximately 34.4 million users from India. Although this constitutes a small fraction of India's total population, the platform is extensively used by politicians, influencers, celebrities, and well-educated individuals whose tweets can significantly influence public sentiment. Their opinions can shape political narratives and sway the perspectives of their followers, highlighting the importance of structured sentiment analysis in this domain. Categorizing political content through sentiment analysis enables individuals to identify and engage with similar or opposing viewpoints, fostering a more organized exchange of information and enhancing public participation and awareness of political issues (Ansari et al., 2020).

Twitter data have been leveraged for various research applications, including sentiment analysis (Kouloumpis et al., 2011), stock market prediction (Bollen et al., 2011), trend detection (Mathioudakis and Koudas, 2010), information credibility assessment (Castillo et al., 2011), and event detection

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 126–131
May 3, 2025 ©2025 Association for Computational Linguistics

(Becker et al., 2021). However, sentiment analysis in Tamil presents unique challenges due to the language's complex linguistic structure and diverse dialectical variations. While previous studies have explored sentiment analysis in Tamil, its application within the political discourse domain remains relatively underexplored. Addressing this gap, our study aims to develop effective sentiment analysis models tailored to Tamil political content, contributing to a deeper understanding of public opinion within this low-resource language setting.

Tamil-speaking individuals play a crucial role in shaping Indian politics, influencing public opinion and driving political discourse. Tamil, a prominent member of the Dravidian language family, is one of the oldest and most widely spoken languages in India, with a rich cultural and linguistic heritage. Despite its significance, there is a notable lack of dedicated resources and tools for analyzing political sentiment in Tamil. The Dravidian languages, including Telugu, Kannada, and Malayalam, share common linguistic features but also exhibit distinct characteristics, making them a unique challenge for NLP tasks (Chakravarthi et al., 2021, 2022, 2023, 2024). With the increasing prominence of social media as a platform for political discussions, this gap presents a significant challenge. Political opinions expressed on social media are inherently complex, often involving sarcasm, ambiguous viewpoints, and contextually nuanced arguments. Traditional NLP techniques, such as rule-based parsing, frequently struggle to handle these complexities, necessitating the development of more sophisticated approaches to effectively capture sentiment (Maynard and Funk, 2012).

Our study addresses the existing gap by focusing on the classification of political sentiment in Tamil tweets using a novel dataset and state-of-the-art machine learning models from the Hugging Face library. We systematically evaluate multiple transformer-based models and experiment with diverse hyperparameter configurations to achieve optimal performance. Our experiments demonstrate that the *"cardiffnlp/bert-base-multilingual-cased-sentiment-multilingual"* model achieves a training accuracy of 79%, with a precision of 80% and an F1 score of 80%. However, during validation, the model attains an accuracy of 33%, precision of 35%, and an F1 score of 34%. Despite the inherent challenges of sentiment analysis in Tamil, our results surpass those of other teams working on similar tasks, highlighting the potential of our approach. The findings of this study contribute to the development of more effective and reliable tools for political sentiment analysis in Tamil, addressing the needs of policymakers, researchers, and stakeholders interested in understanding Tamil public opinion. Our work paves the way for future advancements in sentiment analysis for low-resource languages, fostering deeper insights into political discourse in Tamil-speaking communities.

2 Related Works

Recent research on sentiment analysis of Tamil political comments using customized BERT models has focused on improving language comprehension and classification accuracy in Dravidian languages. TamilCogniBERT enhances Tamil text understanding through a pre-trained BERT framework with self-learning techniques (G et al., 2024). Task-specific pre-training and cross-lingual transfer learning have been shown to improve sentiment classification in Tamil-English code-mixed data (Gupta et al., 2021). Multi-task learning frameworks help tackle the issue of limited annotated data, enhancing sentiment and offensive language detection across Tamil, Malayalam, and Kannada (Hande et al., 2021). The DravidianCodeMix dataset, containing 60,000+ annotated comments, provides a strong foundation for model evaluation and training (Chakravarthi, 2022). Despite these advancements, challenges persist in handling diverse dialects and code-mixing, requiring further research to enhance model robustness.

3 Problem Description

Problem Statement. Sentiment analysis plays a crucial role in understanding public opinion, particularly in the political domain, where sentiments influence strategic decisions, policy-making, and public engagement. With the rise of social media platforms such as X (formerly Twitter), individuals now have a direct channel to express their political views (Chakravarthi et al., 2025). However, analyzing and classifying sentiments in Tamil tweets pose significant challenges due to the language's complex socio-cultural and linguistic characteristics.

The Political Multiclass Sentiment Analysis of Tamil X (Twitter) Comments Shared Task was conducted as an integral part of DravidianLangTech@NAACL 2025 (Chakravarthi et al., 2025). This task focused on the Political Multiclass

Sentiment Type	Train	Test
Opinionated	1361	153
Sarcastic	790	115
Neutral	637	84
Positive	575	69
Substantiated	412	52
Negative	406	51
None of the above	171	20

Table 1: Sentiment Distribution in Train and Test Sets

Sentiment Analysis of Tamil tweets, categorizing them into seven distinct sentiment classes: *Substantiated, Sarcastic, Opinionated, Positive, Negative, Neutral, and None of the Above*. Accurate classification requires not only linguistic proficiency but also an understanding of context, cultural nuances, and intent, adding substantial complexity to the task. Tamil, a Dravidian language with unique vocabulary and syntactic structures, poses challenges for conventional NLP techniques, especially in political discourse involving sarcasm and subjective expressions. Additionally, class imbalance in the dataset and the lack of robust pre-trained models for Tamil necessitate customized approaches for accurate sentiment classification.

Dataset. The dataset used for Tamil political sentiment analysis is divided into three subsets: training, validation, and testing. The **training set** consists of 4,352 Tamil political tweets, each labeled into one of seven sentiment classes: *Substantiated, Sarcastic, Opinionated, Positive, Negative, Neutral,* and *None of the Above*, serving as the primary source for model learning. The class distribution is added in Table 1. The **validation set** contains 544 labeled tweets and is used to fine-tune the model, ensuring generalization and preventing overfitting. Lastly, the **test set** comprises 544 unlabeled tweets, which are used for final evaluation by assessing the model's predictive performance. The dataset presents challenges such as class imbalance, with certain sentiment categories being underrepresented, making it difficult to achieve consistent accuracy across all classes. Despite these challenges, the dataset provides a valuable resource for developing and benchmarking sentiment analysis models tailored for Tamil political discourse.

4 System Description

Model. For sentiment analysis, we employ the *cardiffnlp/**b**ert-**b**ase-**m**ultilingual-**c**ased-**s**entiment-**m**ultilingual* model (BBMCSM, in short) (Antypas et al., 2022), which is a fine-tuned variant of the BERT base multilingual cased architecture. This model is specifically enhanced to classify sentiments in multilingual tweets, leveraging BERT's bidirectional processing capability to capture contextual meanings across various languages effectively. The model is designed for text classification tasks, particularly in analyzing sentiments expressed in social media content. The fine-tuning process utilizes the `TweetNLP` toolkit (Camacho-Collados et al., 2022), which is specialized for processing and analyzing tweet data. During evaluation on the test set, the model achieved an F1 score of 0.616 for both micro and macro measurements and an accuracy of 0.617, demonstrating a moderate level of reliability in sentiment detection for multilingual tweets. This model holds potential for automating sentiment analysis in social media, aiding in the understanding of public opinions and emotions across diverse linguistic contexts.

Implementation Details. The dataset used in this study consists of training, validation, and test sets, each serving a distinct purpose. The training set is utilized to optimize the model's weights through backpropagation, while the validation set is employed for hyperparameter tuning, ensuring the best configuration for generalization. The test set, which remains unseen during training, is used for the final evaluation of the model's performance on real-world data. For model training, we experimented with different hyperparameter configurations to achieve optimal performance. The final selected hyperparameters include an *input dimension* equal to the feature size of the training set, *number of classes* set to 7, and *hidden dimensions* of 1,536 and 786. We utilized a *batch size* of 32 and trained the model for *50 epochs* with a *learning rate* of 0.001 and a *dropout probability* of 0.3 to prevent overfitting. A fixed *random seed* of 42 was used to ensure reproducibility. Performance metrics such as Accuracy, Precision, Recall, and F1-Score were recorded to assess both training and validation outcomes. This systematic approach provides valuable insights for building efficient multilingual sentiment classification models with practical applications in political discourse analysis.

5 Experimental Findings

Training and Validation Results. The training and validation results demonstrated in Table 2

Table 2: Model Performance in Different Setups (Training and Validation Data)

Hidden dims	LR	dropout	T Acc	T Prec	T Rec	T F1	V Acc	V Prec	V Rec	V F1
1536, 786	0.001	0.3	0.7980	0.8040	0.7968	0.8002	0.3290	0.3429	0.3433	0.3391
1536, 786	0.001	0.4	0.7824	0.7924	0.7824	0.7850	0.3150	0.3285	0.3357	0.3262
1028, 786	0.001	0.3	0.6523	0.6711	0.6409	0.6527	0.3438	0.3496	0.339	0.3383
1028, 786	0.001	0.4	0.6250	0.6456	0.6123	0.6253	0.3334	0.3350	0.3243	0.3210
786, 256	0.001	0.3	0.3619	0.4052	0.2960	0.2582	0.3290	0.2654	0.2961	0.2382
786, 256	0.001	0.4	0.3557	0.3645	0.2834	0.2313	0.3107	0.2353	0.2675	0.2036
512, 256	0.001	0.3	0.3660	0.4086	0.3022	0.2692	0.2831	0.2284	0.2673	0.2254
512, 256	0.001	0.4	0.3500	0.3523	0.2934	0.2823	0.2934	0.2243	0.2723	0.2323

Table 3: Macro F1 Scores on Test Data

Submission	F1 Score (Macro)
Mean	0.2769
Median	0.2769
Our Result (Best)	**0.3187**
Our Rank	**4th**

shows the importance of hyperparameter optimization in achieving high performance for sentiment analysis tasks. Our experiments focused on fine-tuning a multilingual BERT model, with varying configurations of hidden dimensions, learning rates (LR), and dropout rates. The results highlight the influence of these hyperparameters on the model's ability to generalize across both training and validation sets. The results indicate that dropout plays a crucial role in preventing overfitting but can lead to underfitting when set too high. When the dropout rate is 0.3, the larger hidden dimensions (1536, 786) perform the best in both training and validation, achieving high accuracy, precision, recall, and F1-Score, with only a moderate drop in validation performance. However, increasing the dropout rate to 0.4 leads to a slight decrease in training performance and a more significant drop in validation results, especially for smaller hidden dimension configurations (1028, 786 and 512, 256). This suggests that while dropout helps regularize the model, a higher dropout rate can overly restrict the model's ability to learn from the data, particularly for smaller architectures. In the case of the 512, 256 configuration, the combination of smaller hidden dimensions and higher dropout results in poor training and validation performance, confirming the importance of selecting an appropriate model capacity for the task. Interestingly, the larger hidden dimensions maintain better generalization, as they are more robust to dropout, particularly in validation. This highlights the importance of balancing dropout and model size for optimal performance. Overall, a dropout rate of 0.3 is the most effective for achieving good generalization, particularly when using larger hidden dimensions, while higher dropout rates tend to hinder performance, especially for smaller models.

Test Results. Our customized BERT model also performed well on the test set, achieving an MF1 score of 0.3187, surpassing all other models. The average and median MF1 scores across all teams were 0.2769. This suggests that our approach, through hyperparameter optimization and improvement of the multilingual model, effectively captures sentiment patterns in Tamil political contexts. These results validate our approach and provide a foundation for further applications.

Overall, this study sets a benchmark for Tamil political sentiment analysis and opens avenues for future work, such as dataset expansion and exploring alternative architectures for multilingual sentiment analysis.

6 Conclusion

Sentiment analysis in low-resource languages presents unique challenges, and this work significantly contributes to Tamil political sentiment analysis. By developing a new annotated dataset and benchmarking transformer-based models, we demonstrated the feasibility of capturing subtle political sentiments in Tamil. Our fine-tuned multilingual BERT model achieved strong results, showcasing the effectiveness of NLP techniques and hyperparameter optimization in complex linguistic tasks. Despite challenges like class imbalance and nuanced sentiment expressions, this study provides a solid foundation for future research. The insights from this work can extend to other low-resource languages, advancing the goal of making NLP more inclusive across linguistic contexts.

Limitations

Despite our approach performing well, there are limitations to address. We were constrained by computational resources, preventing the use of larger, more complex models, which could improve accuracy by capturing deeper structures. Challenges such as class imbalance and ambiguous sentiments in the training and evaluation sets also impacted model performance. Additionally, while our fine-tuned multilingual model showed decent results, further domain-specific pretraining on Tamil data could enhance its understanding of political sentiment. Addressing these constraints in future work could lead to a more robust sentiment analysis framework.

Broader Impact

This work tackles sentiment analysis in low-resource languages, focusing on Tamil political discourse. By creating a new annotated dataset and experimenting with transformer-based models, we showcase the potential of multilingual BERT for capturing subtle political sentiments. Despite challenges like class imbalance and complex expressions, this study lays the foundation for future research in low-resource languages. Our findings contribute to making NLP more inclusive and adaptable across diverse linguistic and cultural contexts.

Acknowledgement

We express our sincere gratitude to Computational Intelligence and Operations Laboratory (CIOL) for their invaluable guidance, unwavering support, and continuous assistance throughout this journey. We are deeply appreciative of their efforts in organizing the CIOL Winter ML Bootcamp (Wasi et al., 2024), which provided an enriching learning environment and a strong foundation for collaborative research. The research mentoring and structured support offered by CIOL played a pivotal role in shaping this work, fostering innovation, and empowering participants to contribute meaningfully to the field of computational linguistics.

References

Mohd Zeeshan Ansari, M.B. Aziz, M.O. Siddiqui, H. Mehra, and K.P. Singh. 2020. Analysis of political sentiment orientations on twitter. *Procedia Computer Science*, 167:1821–1828.

Dimosthenis Antypas, Asahi Ushio, Jose Camacho-Collados, Leonardo Neves, Vitor Silva, and Francesco Barbieri. 2022. Twitter Topic Classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Hila Becker, Mor Naaman, and Luis Gravano. 2021. Beyond trending topics: Real-world event identification on twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1):438–441.

Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

Jose Camacho-Collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa-Anke, Fangyu Liu, Eugenio Martínez-Cámara, et al. 2022. TweetNLP: Cutting-Edge Natural Language Processing for Social Media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Abu Dhabi, U.A.E. Association for Computational Linguistics.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, WWW '11. ACM.

Bharathi R. Chakravarthi, Ruba Priyadharshini, Anand Kumar M, Sajeetha Thavareesan, and Elizabeth Sherly, editors. 2023. *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*. INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria.

Bharathi Raja Chakravarthi. 2022. Dravidiancodemix: sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text. *Language Resources and Evaluation*, 56(3):765–806.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Anand Kumar M, Parameswari Krishnamurthy, and Elizabeth Sherly, editors. 2021. *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics, Kyiv.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Anand Kumar Madasamy, Parameswari Krishnamurthy, Elizabeth Sherly, and Sinnathamby Mahesan, editors. 2022. *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics, Dublin, Ireland.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Anand Kumar Madasamy, Sajeetha Thavareesan, Elizabeth Sherly, Rajeswari Nadarajan, and Manikandan Ravikiran, editors. 2024. *Proceedings of the Fourth Workshop on Speech, Vision, and Language*

Technologies for Dravidian Languages. Association for Computational Linguistics, St. Julian's, Malta.

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Elizabeth Sherly, Thenmozhi Durairaj, Sathiyaraj Thangasamy, Ratnasingam Sakuntharaj, Prasanna Kumar Kumaresan, Kishore Kumar Ponnusamy, Arunaggiri Pandian Karunanidhi, and Rohan R. 2025. Overview of the Shared Task on Political Multiclass Sentiment Analysis of Tamil X(Twitter) Comments: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ashwinraj G, Sarfraz Hussain M, and Madhu Perkin T. 2024. Tamilcognibert: Enhancing tamil language comprehension using self learning. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pages 1–7.

Akshat Gupta, Sai Krishna Rallabandi, and Alan W. Black. 2021. Task-specific pre-training and cross lingual transfer for sentiment analysis in dravidian code-switched languages. page 73–79.

Adeep Hande, Siddhanth U Hegde, Ruba Priyadharshini, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. Benchmarking multi-task learning for sentiment analysis and offensive language identification in under-resourced dravidian languages. *arXiv: Computational Complexity*.

Efthymios Kouloumpis, Theresa Wilson, and Johanna D. Moore. 2011. Twitter sentiment analysis: The good the bad and the omg! In *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*, pages 538–541. AAAI Press.

Michael Mathioudakis and Nick Koudas. 2010. Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, SIGMOD/PODS '10. ACM.

Diana Maynard and Adam Funk. 2012. *Automatic Detection of Political Opinions in Tweets*, page 88–99. Springer Berlin Heidelberg.

Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. 2012. A system for real-time Twitter sentiment analysis of 2012 U.S. presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, pages 115–120, Jeju Island, Korea. Association for Computational Linguistics.

Azmine Toushik Wasi, MD Shakiqul Islam, Sheikh Ayatur Rahman, and Md Manjurul Ahsan. 2024. Ciol presnts winter ml bootcamp. 6 December, 2024 to 6 February, 2025.