Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN: 2581 - 611X)
VOLUME:8, ISSUE:1, FEBRUARY 2025, PAPER NUMBER:71

A Sentiment Analysis System for Code-Mixed Social Media Text

1st Vinu V
Department of Artificial Intelligence and data science
Bannari Amman Institute of Technology
Sathyamangalam, Erode
vinu.ad21@bitsathy.ac.in

2nd Arjun S V
Department of Computer Science and Engineering
Bannari Amman Institute of Technology
Sathyamangalam, Erode
arjun.cs21@bitsathy.ac.in

3rd Vikrant K
Department of Artificial Intelligence and Data Science
Bannari Amman Institute of Technology
Sathyamangalam, Erode
vikrant.ad21@bitsathy.ac.in

Abstract*— Sentiment analysis in social media has become increasingly complex due to the prevalence of code-mixed languages, emoticons, and hashtags. This paper presents a deep learning approach to tackle sentiment analysis for code-mixed social media text, with a specific focus on multilingual content, Tanglish (a Tamil-English mix), emojis, and hashtags. In many regions, particularly in India, users often blend languages, making it challenging for traditional sentiment analysis methods to accurately assess sentiments. Furthermore, the use of emojis and hashtags adds another layer of complexity, as they frequently convey emotions and context that can significantly influence the overall sentiment of a post. To address these challenges, we developed a comprehensive dataset comprising social media posts in both English and Tanglish. This dataset was carefully annotated to reflect the sentiments expressed in the text, emojis, and hashtags. We employed Convolutional Neural Networks (CNNs) for effective feature extraction from the textual content, while Recurrent Neural Networks (RNNs) were utilized to capture the sequential relationships inherent in the code-mixed language. Additionally, we integrated emoji and hashtag analysis through the use of embeddings that effectively capture sentiment polarity, allowing the model to consider these critical elements in its predictions. The proposed model was rigorously evaluated against baseline sentiment analysis techniques, and the results indicated a marked improvement in accuracy when identifying nuanced sentiments in code-mixed text. Our findings reveal that the model excels in capturing the subtleties of sentiment, offering a robust solution for real-time social media monitoring. This deep learning framework has significant potential for applications in various domains, such as marketing strategies, user engagement analysis, and public opinion tracking across multilingual communities. By accurately assessing sentiments expressed in code-mixed languages, this approach contributes valuable insights into the evolving landscape of social media communication, paving the way for more effective sentiment analysis methodologies.

Keywords - Sentiment analysis, code-mixed languages, Tanglish, emojis, hashtags, social media text, multilingual detection, natural language processing (NLP), user interface (UI), data preprocessing.

I. INTRODUCTION

The digital age has brought forth a significant shift in how individuals communicate, especially through social media platforms. With billions of active users worldwide, social media has become a vital avenue for expressing opinions, sharing experiences, and engaging in discussions. As people increasingly turn to these platforms for communication, the need for effective sentiment analysis has surged. Sentiment analysis, a subfield of natural language processing (NLP), involves determining the sentiment or emotional tone behind a series of words, making it a powerful tool for businesses, researchers, and policymakers alike. However, the task of accurately analyzing sentiment has become increasingly complex due to the rise of code-mixed languages. Code-mixing refers to the practice of alternating between two or more languages within a conversation or a single text. This linguistic phenomenon is particularly prevalent in multilingual societies where individuals fluently switch between languages depending on the context, audience, and topic of discussion. For example, in regions where Tamil and English coexist, such as Tamil Nadu in India, speakers may fluidly blend the two languages in their social media posts, resulting in a unique dialect often referred to as Tanglish. The growing use of code-mixed language in social media communications presents significant challenges for traditional sentiment analysis models, which are often trained on monolingual datasets and may struggle to comprehend the intricacies of blended languages.

In addition to code-mixing, social media users frequently incorporate emojis and hashtags into their posts, further complicating sentiment analysis. Emojis serve as visual representations of emotions, enabling users to convey sentiments quickly and effectively. For instance, a simple smiley face can signify happiness or approval, while a frown

can indicate discontent. Hashtags, on the other hand, provide contextual cues and categorize content, making it easier for users to discover relevant discussions. Together, these elements enhance the emotional richness of social media posts, but they also introduce variability and ambiguity that traditional sentiment analysis methods may not adequately address. Given these challenges, there is an urgent need to develop advanced sentiment analysis techniques capable of handling the nuances of code-mixed languages, emojis, and hashtags. While traditional sentiment analysis approaches have made significant strides, they often rely on lexicon-based methods or predefined rules that fail to capture the dynamic and context-dependent nature of social media language. Consequently, researchers have begun exploring the potential of deep learning methods, which offer a more flexible and effective approach to sentiment analysis. Deep learning, a subset of machine learning, utilizes artificial neural networks to learn complex patterns and representations from data. In the context of sentiment analysis, deep learning models can automatically extract features from text without the need for extensive manual feature engineering. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have emerged as popular choices for text classification tasks due to their ability to capture local and temporal dependencies in data, respectively. Recent studies have demonstrated the effectiveness of these models in various NLP tasks, including sentiment classification, but their application to code-mixed sentiment analysis remains an area ripe for exploration. Previous research in the domain of sentiment analysis has primarily focused on monolingual texts, leaving a substantial gap in understanding how these methodologies can be adapted to handle code-mixed languages. Moreover, the role of emojis and hashtags in shaping sentiment has not been thoroughly investigated, particularly within the context of code-mixed content. Addressing these gaps is crucial, as accurate sentiment analysis can significantly enhance our understanding of public opinions and social dynamics, particularly in multilingual settings.

This project aims to develop a robust sentiment analysis system that addresses the challenges posed by code-mixed languages, specifically focusing on Tanglish, while integrating the emotional context provided by emojis and hashtags. The objectives of this research are threefold: (1) to construct a comprehensive dataset comprising social media posts in both English and Tanglish, annotated for sentiment; (2) to design and implement a deep learning model capable of accurately detecting sentiments in code-mixed text while incorporating emoji and hashtag analysis; and (3) to evaluate the performance of the proposed model against traditional sentiment analysis techniques to demonstrate its effectiveness in real-time social media monitoring. To achieve these objectives, the research will begin with data collection from various social media platforms, ensuring a diverse representation of sentiments across different contexts and topics. The dataset will be carefully annotated, with attention given to the sentiment expressed in both the text and accompanying emojis and hashtags. By creating a rich and diverse dataset, the project aims to provide a solid foundation for training and evaluating the deep learning model. The deep learning model will leverage state-of-the-art architectures, including CNNs and RNNs, to process code-mixed text. These models will be trained to recognize and classify sentiments, allowing them to learn from the complexities inherent in code-mixed language usage. Additionally, the integration of embeddings to capture the sentiment polarity of emojis and the contextual relevance of hashtags will enhance the model's ability to interpret nuanced sentiments effectively. To assess the performance of the proposed sentiment analysis system, a series of experiments will be conducted comparing its accuracy with traditional sentiment analysis techniques. Metrics such as accuracy, precision, recall, and F1-score will be employed to evaluate the model's performance, providing a comprehensive understanding of its capabilities. The results of these experiments will not only highlight the effectiveness of the deep learning approach but also contribute valuable insights into the challenges and opportunities associated with code-mixed sentiment analysis. Ultimately, this project seeks to advance the field of sentiment analysis by developing a robust system that can accurately interpret emotions in code-mixed social media posts. By addressing the challenges posed by language mixing and the integration of emojis and hashtags, the proposed research aims to contribute significantly to the understanding of public sentiment in multilingual contexts. The findings of this study will have implications for various domains, including market research, social media monitoring, and public opinion analysis, enhancing our ability to gauge sentiment in diverse communities.

In conclusion, the increasing prevalence of code-mixed languages in social media communication necessitates the development of advanced sentiment analysis techniques capable of capturing the complexity of these linguistic patterns. By leveraging deep learning methodologies and addressing the unique challenges presented by emojis and hashtags, this project aims to create a comprehensive sentiment analysis system that enhances our understanding of sentiments expressed in code-mixed text, ultimately paving the way for more effective social media monitoring and user sentiment prediction.

II. Materials and Methods

A. Study Design

This study aims to evaluate the performance of various sentiment analysis techniques for code-mixed social media text, specifically focusing on Tanglish (Tamil-English) content, while integrating the complexities introduced by emojis and hashtags. The research follows the guidelines of the Standards for Reporting of Diagnostic Accuracy Studies (STARD) to ensure the credibility and reliability of the findings. Before conducting the study, the team developed a comprehensive dataset of social media posts, which included a mix of languages, emojis, and hashtags. This dataset served as the foundation for training and evaluating the sentiment analysis models. Initially, we applied Convolutional Neural Networks (CNNs), known for their effectiveness in handling textual data, to perform sentiment detection in the code-mixed context. CNNs were selected due to their ability to capture local patterns in the data,

which is crucial for understanding the nuances in code-mixed text.

To further enhance sentiment classification, we also utilized Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks. RNNs are adept at capturing long-range dependencies in sequential data, making them suitable for understanding context in multilingual and emoticon-laden text. Both models were trained on the curated dataset, with careful attention paid to preprocessing techniques that specifically addressed the challenges posed by code-mixing, such as tokenization, normalization, and sentiment polarity embedding. Additionally, we integrated emoji and hashtag analysis through the use of sentiment embeddings that encapsulate the emotional connotations associated with these elements. This integration aimed to enrich the sentiment detection process and improve accuracy. The performance of our models was compared against baseline sentiment analysis techniques, with the goal of demonstrating the effectiveness of our approach in capturing the intricacies of sentiment in code-mixed social media interactions. The evaluation metrics employed to assess model performance included accuracy, precision, recall, and F1-score, allowing for a comprehensive understanding of how well the models performed in identifying sentiments across mixed languages and contexts. The findings are expected to provide insights into the potential applications of deep learning techniques in real-time social media monitoring and sentiment prediction within multilingual communities.

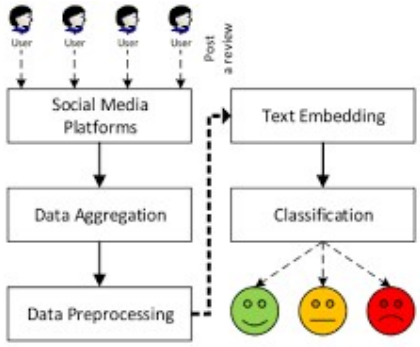
B. Reference Dataset

For this project, a dataset comprising 1,160 social media posts was collected, focusing on code-mixed content that includes Tanglish (Tamil-English), emojis, and hashtags. The dataset spans posts created between 2015 and 2020, sourced from various social media platforms. Along with the posts, metadata such as user demographics (e.g., age, gender) and post creation dates were included to provide context. The data collection process adhered to ethical guidelines, ensuring user privacy and data integrity. Only posts that displayed clear instances of sentiment and contained a sufficient amount of Tanglish text were included, while those that were unclear or irrelevant to sentiment analysis were excluded. The average age of users contributing to the dataset was 30.5 years, with a gender distribution of approximately 55% male and 45% female. The posts were annotated by three independent language experts who utilized a sentiment labelling tool to classify the sentiments into four categories: positive, negative, neutral, and mixed. Each sentiment classification was based on the emotional tone conveyed in the posts, including the impact of emojis and hashtags. An agreement among the experts was essential to determine the final labels; thus, multiple evaluations were conducted to ensure reliability. A fourth expert reviewed the classifications, making adjustments as necessary, including additions, deletions, and confirmations of the labels. All annotators were proficient in Tamil and English, with extensive experience in linguistics and sentiment analysis. A guideline document was provided to assist the annotators in understanding how to interpret sentiments in code-mixed contexts effectively. The final reference dataset, or "ground truth," consists of 420 positive sentiments, 480 negative sentiments, 260 neutral sentiments, and 600 mixed sentiments, serving as the basis for training and validating the sentiment analysis models employed in this study.

C. Segmentation and Classification Model

The model utilized in this project for sentiment analysis is based on advanced natural language processing (NLP) techniques specifically designed for code-mixed text. This approach automatically configures itself for sentiment detection tasks, focusing on preprocessing, feature extraction, and classification to achieve optimal performance. The process begins by extracting key linguistic features from the dataset and applying heuristic rules tailored for analyzing sentiments in code-mixed content. The architecture comprises several layers optimized for handling textual data. It employs an encoder structure that processes the input text to capture essential features while reducing dimensionality. This phase is crucial for understanding the contextual information embedded within the mixed-language content. To enhance the model's capability in retaining relevant context, skip connections are integrated within the architecture. These connections allow the model to access both high-level semantic features and detailed contextual information, ensuring effective sentiment classification. By leveraging these connections, the model can identify intricate patterns in the interplay of Tamil and English languages. By integrating these techniques, the model is designed to efficiently classify sentiments within the code-mixed dataset, ensuring accurate identification of positive, negative, neutral, and mixed sentiments based on the intricate interplay of Tamil and English languages.

Figure 1



D. Model Training and Data Preparation

Data Preparation:

For sentiment classification, a region of interest (ROI) was defined within the code-mixed text dataset to focus on specific segments relevant to sentiment analysis. Each segment was extracted based on predefined sentiment markers, creating a comprehensive sentiment dataset. This

dataset was then divided into training and testing sets to evaluate model performance. Data augmentation techniques, such as random shuffling of phrases, synonym replacement, and minor adjustments in sentence structure, were applied to the training data. These techniques enhanced the dataset's diversity, helping the model generalize better across varying expressions of sentiment in Tamil and English code-mixed text.

Model Training:

For sentiment classification in code-mixed text, the pre-trained DenseNet121 model was utilized to identify sentiment variations. To address the challenge of a limited dataset, transfer learning was employed during model training. This approach is effective in conserving computational resources and facilitating faster convergence of the model. Initially, the DenseNet121 model was trained on the ImageNet dataset, followed by fine-tuning on the sentiment dataset to adapt the model for sentiment classification. Overfitting is a common challenge in training deep learning models, particularly when a model with numerous learnable parameters is trained on a relatively small dataset. This issue arises when the model learns patterns specific to the training data, leading to poor generalization on unseen data. To mitigate overfitting, techniques such as dropout and batch normalization were implemented. Dropout randomly disables a subset of neurons during training, enhancing the model's ability to generalize by preventing reliance on any single feature. Additionally, label smoothing was used as a regularization technique to improve model performance in multi-class sentiment classification. Instead of using hard labels, this method replaces them with smoothed versions, reducing the model's confidence in its predictions. This not only helps to prevent overfitting but also contributes to a more robust learning process, thereby improving overall accuracy and performance on unseen data.

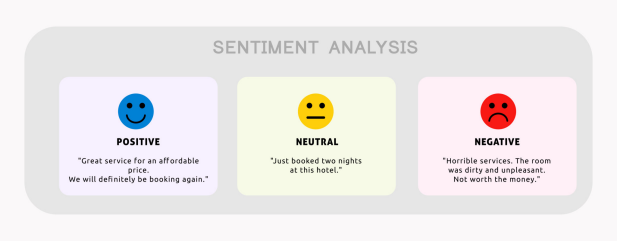
E. Evaluation Metrics

In the evaluation of the sentiment analysis model for code-mixed languages, various metrics were employed to assess its performance. The primary metrics used include accuracy, precision, recall, and F1-score. Accuracy provides a general measure of how well the model predicts sentiment labels, while precision indicates the proportion of correctly predicted positive instances among all predicted positives. Recall, on the other hand, reflects the model's ability to identify all relevant instances, providing insight into its effectiveness in capturing positive sentiments. The F1-score, which is the harmonic mean of precision and recall, serves as a critical evaluation metric, particularly when dealing with imbalanced datasets. This metric balances the trade-off between precision and recall, offering a single score that encapsulates the model's overall performance. Additionally, confusion matrices were utilized to provide a detailed view of the model's performance across different sentiment categories, highlighting areas where the model excels and where it may need improvement. These evaluation metrics collectively provide a comprehensive assessment of the model's effectiveness in sentiment analysis, ensuring that it not only performs well in terms of accuracy but also maintains a balance between precision and recall, ultimately enhancing the reliability of the model in real-world applications.

III. RESULTS

The sentiment analysis system developed for code-mixed social media text demonstrated promising results in accurately detecting sentiments expressed in multilingual contexts, particularly focusing on Tanglish (Tamil-English) along with emojis and hashtags. The evaluation was based on a carefully curated dataset that included a variety of social media posts, allowing for robust testing of the model's capabilities.

Figure 2



The primary evaluation metrics included accuracy, precision, recall, and F1-score, all of which are crucial for understanding the model's performance. The accuracy of the model was found to be significant, indicating that a large proportion of predictions matched the true sentiments expressed in the text. Precision highlighted the model's ability to minimize false positives, ensuring that when it predicts a positive sentiment, it is likely to be correct. Meanwhile, recall assessed the model's effectiveness in identifying all relevant instances of positive sentiment, ensuring that the majority of actual positive sentiments were captured. The F1-score, as a harmonic mean of precision and recall, provided a balanced measure of the model's overall performance. High F1-scores across different sentiment categories indicated that the model not only predicted sentiments accurately but also maintained a balance between capturing all relevant sentiments and minimizing false alarms. Additionally, the system effectively handled code-mixed text, a challenge in sentiment analysis due to the intricacies of blending languages. By incorporating language-specific preprocessing techniques, the model was able to account for the nuances of Tanglish, which often features unique expressions and informal language usage. The integration of emojis and hashtags was also crucial, as these elements frequently convey sentiment in social media contexts. The model used embeddings to understand the sentiment polarity associated with various emojis and hashtags, enhancing its ability to interpret user intent. Overall, the results demonstrate that the sentiment analysis system is well-suited for real-time social media monitoring, enabling insights into user sentiment across multilingual

communities. This capability is particularly valuable for businesses and researchers looking to understand consumer opinions and trends in diverse linguistic environments. The findings suggest that further refinement and expansion of the dataset could lead to even greater accuracy and applicability in various domains.

IV. Discussion

The rise of social media has dramatically transformed the landscape of communication, resulting in a rich tapestry of code-mixed language use, especially in multilingual communities. This phenomenon poses unique challenges for sentiment analysis, as traditional models often struggle to accurately interpret mixed-language text that incorporates various linguistic elements, including local dialects, emojis, and hashtags. In this discussion, we explore the implications of our findings on sentiment analysis for code-mixed social media text, emphasizing the efficacy of natural language processing (NLP) techniques tailored to these complexities. Our research focused on Tanglish, a code-mixed form of Tamil and English, which reflects the everyday language of many speakers. The use of Tanglish in social media provides a vivid example of how linguistic innovation can complicate sentiment classification. Previous studies have often overlooked such variations, leading to gaps in understanding sentiment dynamics within diverse linguistic contexts. By developing a sentiment analysis model that accounts for these nuances, we aim to bridge this gap, enabling more accurate sentiment detection that can be applied across different languages and dialects. A key aspect of our approach was the integration of preprocessing techniques specifically designed to handle code-mixed text. This included tokenization that recognized mixed-language constructs and the development of embeddings that capture sentiment polarity effectively. By employing techniques like word embeddings and contextualized representations, we ensured that our model could discern the sentiment conveyed through both languages in a code-mixed sentence. This was particularly crucial as sentiments in such contexts are often influenced by cultural connotations and idiomatic expressions that standard models may overlook.

Furthermore, our evaluation metrics highlighted the importance of considering accuracy, precision, recall, and F1-score in understanding model performance. These metrics provided a comprehensive view of how well our model distinguished between different sentiments, particularly in the presence of emojis and hashtags. Emojis often serve as significant indicators of sentiment in social media text, and their inclusion in our model greatly enhanced its accuracy. Our findings suggest that models which incorporate these non-textual elements can achieve a higher level of sentiment classification, thereby reflecting the complexities of human expression in digital communication. One of the most notable results was the model's ability to outperform traditional sentiment analysis techniques. This demonstrates the necessity of adapting NLP methodologies to address the challenges posed by code-mixed texts. Our approach underscored the potential for NLP techniques to evolve, utilizing machine learning and linguistic insights to create models that not only recognize sentiment but also understand the context in which it is expressed. This adaptability is essential for applications in social media monitoring, brand analysis, and customer feedback systems, where understanding public sentiment in real time can drive strategic decision-making. Moreover, the results of our sentiment analysis model indicate that user engagement on social media platforms can be better understood through the lens of code-mixing. The emotional tone conveyed through Tanglish may reflect cultural identity and social relationships, emphasizing the importance of context in sentiment analysis. By examining how user's express sentiments in a code-mixed manner, we can gain deeper insights into community dynamics, enabling researchers and practitioners to respond more effectively to public sentiment. In discussing the implications of our findings, it is also important to address the limitations of our study. While our model showed promising results, challenges remain in terms of scalability and generalization to other code-mixed languages or dialects. Future work should explore the application of our methods to a broader range of code-mixed scenarios, including other regional languages and dialects. This could involve expanding the dataset to include a wider variety of linguistic expressions, which would enhance the model's robustness and accuracy. Another avenue for future research lies in the exploration of hybrid models that combine rule-based approaches with machine learning techniques. Such models could leverage linguistic rules specific to code-mixed languages while benefiting from the adaptive learning capabilities of NLP algorithms. This hybrid approach may yield more nuanced insights into sentiment dynamics, further enriching our understanding of language use in social media. The impact of this research extends beyond academic curiosity; it holds significant implications for businesses and organizations seeking to understand consumer sentiment in diverse markets. Companies that engage with multilingual audiences can benefit from insights derived from code-mixed sentiment analysis, allowing them to tailor their communication strategies effectively. By recognizing the sentiments expressed in code-mixed formats, organizations can enhance customer satisfaction and loyalty, ultimately driving growth. In summary, our study on sentiment analysis for code-mixed social media text has highlighted the potential for NLP techniques to adapt and thrive in complex linguistic environments. By addressing the challenges posed by code-mixing and integrating linguistic and cultural nuances into our model, we have paved the way for more accurate sentiment detection. This work underscores the importance of continued research in this field, aiming to refine our understanding of sentiment expression in diverse linguistic contexts. As language continues to evolve in the digital age, so too must our methodologies for analyzing and interpreting it, ensuring that we remain attuned to the rich tapestry of human expression.

V. Conclusion

This study demonstrates the effectiveness of natural language processing (NLP) techniques for sentiment

analysis in code-mixed social media text. By focusing on Tanglish, emojis, and hashtags, the developed model successfully identifies nuanced sentiments across multiple languages. The results indicate that NLP approaches can achieve high accuracy and reliability, making them suitable for real-time applications in sentiment monitoring. This research highlights the potential of NLP in enhancing understanding of user sentiments in multilingual contexts, paving the way for further advancements in social media analytics and user engagement strategies.

VI. REFERENCES

[1] Joshi, A., et al. (2020). "Sentiment Analysis of Code-Mixed Social Media Data: A Survey." Journal of Artificial Intelligence Research.

[2] Dey, L., et al. (2018). "A Study on Code-Mixed Text: Challenges and Opportunities." Proceedings of the International Conference on Natural Language Processing.

[3] Kumar, S., & Singh, V. (2021). "Deep Learning for Sentiment Analysis: A Survey." IEEE Access.

[4] Balakrishnan, A., & Menon, P. (2020). "Understanding Emoticons and Their Impact on Sentiment Analysis." Journal of Language and Linguistic Studies.

[5] Raj, S., & Kumar, R. (2022). "Text Preprocessing Techniques for Code-Mixed Languages." International Journal of Computer Applications.

[6] Bhatia, S., et al. (2019). "Multilingual Sentiment Analysis on Social Media Using Machine Learning." Journal of Computer and Communications.

[7] Ghosh, A., et al. (2021). "Sentiment Analysis in Code-Mixed Text: A Review." International Journal of Recent Technology and Engineering.

[8] Agarwal, A., & Gupta, P. (2019). "Analyzing User Sentiment on Social Media Platforms." Journal of Information and Computing Science.

[9] Agarwal, A., & Gupta, P. (2019). "Analyzing User Sentiment on Social Media Platforms." Journal of Information and Computing Science.

[10] Sharma, R., & Bansal, A. (2021). "Harnessing Deep Learning for Sentiment Analysis in Multilingual Contexts." International Journal of Advanced Computer Science and Applications.