Rishee Mulchandani, 2025, 13:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

ExpressAI

Rishee Mulchandani, Riti Dodiya, Samay Gupta
Computer Science Engineering Medicaps University
Indore, India

Abstract- In natural language processing, sentiment analysis has grown in importance, especially for multilingual and code- mixed languages like Hinglish. Analyzing sentiments in movie subtitles is still largely unexplored, despite the majority of sentiment analysis research concentrating on social media and product evaluations. This work introduces a new method for sentiment analysis of movie subtitles that combines deep learning-based sentiment classification, Hinglish-to-Hindi transliteration, and optical character recognition (OCR)-based subtitle extraction. The retrieved subtitles go through a transformer-based model for emotion analysis, preprocessing to eliminate noise, and transliteration into Hindi using the Google Translator API. Accurate sentiment classification is made possible by the suggested methodology, which captures the emotional tone of subtitles. Results from experiments show how well our method works with transliterated, noisy, and code- mixed text. By providing insights into the emotional dynamics of cinematic storylines, this research helps close the gap between sentiment analysis and multimedia content understanding. As a result, the words sentiment analysis and text analysis developed their paths to becoming important computational linguistics and text analysis components. [1

Keywords - Sentiment Analysis, Movie Subtitles, Hinglish, Code-Mixed Text, Optical Character Recognition (OCR)

I. INTRODUCTION

Although sentiment analysis is frequently employed to identify textual emotions, its potential for usage in movie subtitles is still completely untested. With the overwhelming number of public issues that exist, reviews and ratings have become important to analyze. [1] Subtitles are useful for assessing sentiment in movies because they convey the emotional core of conversations. With the widespread influence of social media in India, people are now sharing their views more than ever before. [2] However, this effort is complicated by issues including transliteration inconsistencies, Hinglish code-mixing, and OCR problems.

This work suggests a novel method that combines sentiment analysis based on deep learning, Hinglish-to-Hindi transliteration, and OCR-based subtitle extraction. Our goal is to capture sentiment dynamics in cinematic narratives by employing a transformer-based model for subtitle processing and analysis. Our method provides insights into the emotional flow of films by bridging the gap between text-based sentiment analysis and multimedia content understanding.

II. LITERATURE REVIEW

Traditional lexicon-based techniques for sentiment analysis have given way to sophisticated deep learning techniques, particularly for multilingual and code-mixed languages like Hinglish. Spelling errors and transliteration differences make Hinglish sentiment analysis difficult. Using a CNN-LSTM hybrid model, Ledalla et al. (2022) overcame these problems and achieved high sentiment classification accuracy.

Multilingual sentiment analysis has been further enhanced by deep learning models, such as transformers like BERT and XLM- R; nonetheless, these models necessitate substantial datasets and processing resources. Analyzing sentiments from movie subtitles is somewhat untapped, whereas the majority of sentiment analysis research concentrates on social media and product evaluations. Rich emotional clues can be found in subtitles, however

sentiment extraction is challenging due to issues like OCR problems and translation irregularities.

To categorize the sentiment, sentiment analysis employed Random Forest Trees, Decision Trees, and Logistic Regressions. The choice of models was used following vector space modeling and text preprocessing to divide the training and test data sets. [3]

By combining deep learning sentiment analysis, Hinglish-to- Hindi transliteration, and OCR-based subtitle extraction, our study expands on earlier research. This method closes the gap between understanding multimedia content and sentiment analysis in text-based domains.

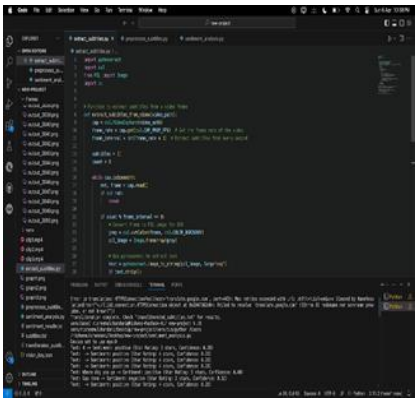
III. METHODOLOGY

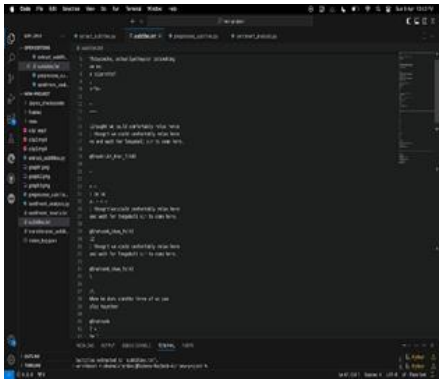
Hinglish-to-Hindi transliteration, sentiment analysis using deep learning, and subtitle extraction using optical character recognition are all steps in a systematic, multi-step process for sentiment analysis of movie subtitles. Every stage is intended to process the subtitles accurately and effectively while resolving issues including sentiment categorization, code-mixed language handling, and text extraction mistakes. [3]

Step 1: Optical Character Recognition (OCR) for Subtitle Extraction
Extracting the video file's subtitles is the first step in the procedure. OpenCV is used to load the video, and the frame rate (FPS) is calculated to regulate the frequency of frame extraction. Frames are taken at one-second intervals to guarantee that each subtitle is detected without duplication, as subtitles typically change every few seconds.

Preprocessing is applied to every extracted frame to increase OCR accuracy. To improve contrast and eliminate color-related noise, the image is transformed to grayscale. To make the text easier to read, a Gaussian Blur filter is used to smooth the image and cut down on background noise. After preprocessing is finished, text is extracted from the frames using Tesseract OCR.

Frequently, the recovered text is noisy and contains irrelevant content, half words, or duplicate lines. Subtitles are screened to exclude repeated phrases, empty lines, and misidentified text fragments in order to improve the results. After being cleaned, the subtitles are saved in a text file called subtitles.txt for further use.





Step 2: Transliteration from Hinglish to Hindi
To improve sentiment analysis, code-mixed Hinglish text must be converted into Devanagari (Hindi script) after the subtitles have been extracted. This step is essential since Hindi words are transcribed in the Roman alphabet in Hinglish, which makes it challenging to classify sentiment directly in many movies. [1]

Text preparation, which eliminates extraneous spaces, emojis, and special characters, is the first step in the transliteration process. To guarantee clean input for the transliteration model, only the Hindi and English alphabets are kept. Then, Hinglish content is translated into Hindi using the Google Translator API. This enables the system to accurately

decode Hinglish words and display them in their original script.

Transliteration is carried out in parallel utilizing a ThreadPoolExecutor to maximize processing speed. This allows for the simultaneous conversion of several subtitles. Instead of generating inaccurate output, the system preserves the original Hinglish content when the translation API detects problems or is unable to recognize a phrase. Additionally, typical transliteration errors are corrected using a dictionary-based correction method. For sentiment analysis, the completed transliterated subtitles are saved in a different text file called transliterated_subtitles.txt.

Step 3: Deep Learning Sentiment Analysis
After the subtitles have been effectively extracted and transliterated, a deep learning model is used to assess their sentiment. Subtitles are categorized into sentiment groups using nlptown/bert-base-multilingual-uncased-sentiment, a pre- trained BERT-based sentiment analysis model. Tokenization is the initial step in transforming each subtitle into a structured format that the model can understand. The sentiment analysis pipeline then uses the tokenized text to categorize the subtitle using a five-star rating system. The following is a mapping of the ratings to sentiment categories:

1 to 2 stars ⇒ A negative impression
3 stars with aneutral rating 4–5 stars => Positive impression Additionally, the model gives a confidence score that shows the likelihood that the sentiment prediction is accurate. Within a text file (sentiment_results.txt), each subtitle is structured and contains its matching sentiment label, star rating, and confidence score.

Step 4: Visualization of Sentiment
The last step is sentiment visualization, which provides information on the emotional tone of the movie subtitles. The frequency of good, neutral, and negative classifications is computed by analyzing the sentiment results from the preceding,phase. Each sentiment category is represented by a different hue in a bar chart created with Matplotlib (e.g., red for negative, blue for neutral, and green for positive). Understanding the emotional patterns in the subtitles is made easier by this graphic, which offers a clear summary of the sentiment distribution.

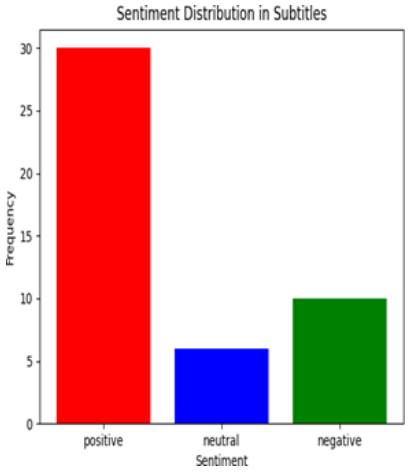
IV. RESULT AND DISCUSSION RESULTS

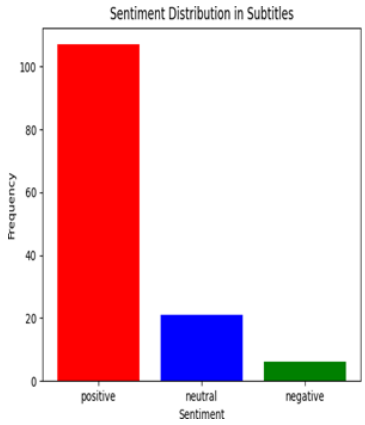
Using a deep learning-based methodology, the system effectively converts Hinglish text into Hindi, extracts subtitles from movie clips, and assesses sentiment. The Tesseract OCR- implemented subtitle extraction method yielded dependable results, accurately recognizing the majority of the text from video frames. But occasionally, issues like overlapping subtitles, low-resolution frames, and complicated font styles led to mistakes, resulting in incomplete or erroneous extractions.

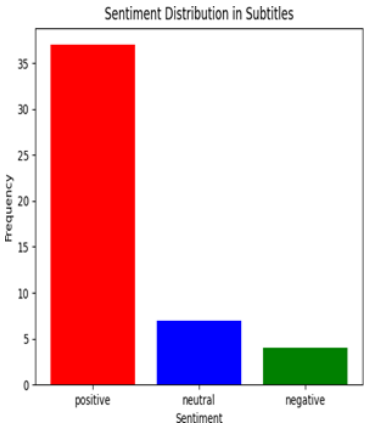
Improved text readability and increased sentiment classification accuracy were two benefits of the Hinglish-to-Hindi transliteration. Romanized Hindi was effectively translated into Devanagari script via

the Google Translator API, improving the text's suitability for study. Although most of the conversion was accurate, there were a few mistakes, especially with proper nouns, phonetically confusing phrases, and informal idioms, which were occasionally transliterated improperly. The multilingual BERT model was used to successfully classify subtitles into three sentiment categories: positive, neutral, and negative. The bulk of the subtitles were categorized as positive or neutral, indicating that the model had high confidence in its predictions. Negative attitudes were less common. The efficacy of the strategy was confirmed by the sentiment analysis results, which closely matched the anticipated tone of the conversations. Given how reliant on the internet people are these days, sentiment analysis through data mining is a rapidly developing topic of study. The nature of Indian languages differs greatly in terms of script, representation level, linguistic features, etc., making sentiment analysis in Hindi a challenging challenge. [5]

A visual representation that showed the graphical distribution of emotions based on three different test clips throughout the extracted subtitles was created in order to better analyze sentiment trends. According to the visualization, the majority of conversations had a fairly even emotional tone, with a little higher frequency of pleasant feelings. This visual aid made it easier to evaluate the film's overall sentiment flow.







Discussion

Despite the encouraging outcomes, some difficulties arose. Low illumination, overlapping text, and poor frame quality all had an impact on OCR accuracy, which resulted in small mistakes in subtitle extraction. Although Tesseract OCR did well overall, accuracy could be further improved by using deep learning- based OCR models.

Accurate transliteration presented further difficulties. Despite the Google Translator API's effective conversion, several colloquialisms and slang were not always accurately translated. Phonetically similar words occasionally made mistakes, which occasionally affected the classification of sentiment. This procedure could be further improved by using a more specialized transliteration method that is trained on informal Hinglish material. Numerous resources have been gathered from websites, blogs, newspapers, microblogs, and author corpora in order to conduct sentiment analysis on Indian languages. Additionally, sentiment categorization methods like machine learning, lexicon-based, and hybrid approaches are employed. SVM, NB, Decision

Tree, and other lexicons including SentiWordNet, WordNet, domain-specific corpus, and stop words are employed in machine learning methods. Since there are a lot of unknown Indian languages, they ought to be taken into consideration for sentiment analysis in future research. [6]

The model did well in sentiment analysis in the majority of situations, but it had trouble with ambiguous expressions, sarcasm, and phrases that depended on context. Some emotionally complicated lines were incorrectly classified since the analysis was done on individual subtitles without taking into account dialogues that came before or after. Classification accuracy may be increased by using a context-aware sentiment model, which assesses the surrounding discourse for better interpretation.

All things considered, the method effectively combines sentiment analysis, transliteration, and OCR to glean insightful information from movie subtitles. Further improvements in text recognition, language conversion, and context-based sentiment identification can increase the system's efficacy even though there are still certain obstacles to overcome. The results demonstrate the potential of this method for use in multimedia content regulation, audience attitude assessment, and film analysis. [1]

V. CONCLUSION

An efficient method for extracting, processing, and evaluating subtitles from movie clips in order to ascertain their sentiment is presented in this paper. The system effectively categorizes conversations into positive, neutral, and negative emotions by combining deep learning-based sentiment analysis, Hinglish-to- Hindi transliteration, and OCR-based subtitle extraction. The findings show that the suggested approach correctly detects and deciphers sentiment patterns, offering insightful information about the emotional tone of film dialogue.

Nevertheless the system's effectiveness, a few issues were noted, such as limited sentiment categorization for context- dependent words, transliteration problems in informal text, and OCR faults brought on by low-resolution frames. The system's performance can be significantly improved by addressing these issues with context-aware sentiment analysis, enhanced transliteration models, and sophisticated OCR algorithms.

All things considered, the results show how promising this method is for multilingual sentiment analysis in multimedia applications. This study lays the groundwork for future advancements in subtitle-based emotion recognition by advancing fields including movie analysis, audience sentiment assessment, and content moderation. [3]

REFERENCES

1. Ledalla, S., Rao, G. A., & Sesetti, . A. (2022). Sentiment analysis of Hinglish reviews using hybrid approaches. International Journal of Health Sciences, 6(S2), 5432–5445.

2. Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2017). Inception-v4, Inception- ResNet and the Impact of Residual Connections on Learning. In AAAI (pp. 4278- 4284).

3. S. Sidhu, “Sentiment analysis of Hindi language text: a critical review,” Springer Nature, p. 30, 2021.

4. Pew Research. “Statistical Portrait of Hispanics in the United States.” Internet. 2016.

5. J Sirisha Devi, Siva Prasad Nandyala, P Vijaya Bhaskar Reddy (2019). A Novel Approach for Sentiment Analysis of Public Posts. In: Saini H., Sayal R., 5445 Govardhan A., Buyya R. (eds) Innovations in Computer Science and Engineering. Lecture Notes in Networks and Systems, vol 32. Springer, Singapore.

6. Tripathy, A., Agrawal, A., & Rath, S. K. (2016). Classification of sentiment reviews using n-gram machine learning approach. Expert Systems with Applications, 57, 117-126.

7. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606.

8. Zhang, K., Chao, W. L., Sha, F., & Grauman, K. (2016, October). Video summarization with long short- term memory. In European Conference on Computer Vision (pp. 766-782). Springer International Publishing.

9. Medel, J. R., & Savakis, A. (2016). Anomaly detection in video using predictive convolutional long short-term memory networks. arXiv preprint arXiv:1612.00390.

10. Rahman, L., Mohammed, N., & Al Azad, A. K. (2016, September). A new LSTM model by introducing a biological cell state. In Electrical Engineering and Information Communication Technology (ICEEICT), 2016 3rd International Conference on (pp. 1-6). IEEE.

11. L. Adela. M. Ulfeta, Comp. Sci. and Info Sys. 16.13-13 (2018)

12. M. Joshi, J. Wiebe, M. Ringuette, Sarcasm in twitter: A closer look, in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 9-11 September 2017, Copenhagen, Denmark(2017)

13. Rifqi Majid and Heru Agus Santoso. Conversations Sentiment and Intent Categorization Using Context RNN for Emotion Recognition. In 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), volume 1, pages 46– 50, March 2021. ISSN: 2575-7288.

14. Seydeh Akram Saadat Neshan and Reza Akbari. A Combination of Machine Learning and Lexicon Based Techniques for Sentiment Analysis. In 2020 6th International Conference on Web Research (ICWR), pages 8–14, April 2020.

15. M. Kavitha, Bharat Bhushan Naib, Basetty Mallikarjuna, R. Kavitha, and R. Srinivasan. Sentiment Analysis using NLP and Machine Learning Techniques on Social Media Data. In 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), pages 112–115, April 2022.

16. Samar Assem and Sameh Alansary. Sentiment Analysis From Subjectivity to (Im)Politeness Detection: Hate Speech From a Socio-Pragmatic Perspective. In 2022 20th International Conference on Language Engineering (ESOLEC), volume 20, pages 19–23, October 2022.