Multi-Language Sentiment Analysis on Social Media

Priti Arvind Ram[1], Jiteshree Raut[2]
[1]*Student, Sonopant Dandeker Shikshan Mandali, Palghar*
[2]*Assistant Professor Department of Information Technology,*
Sonopant Dandeker Shikshan Mandali, Palghar

Abstract—**In order to comprehend user thoughts, feelings, and attitudes expressed on social media platforms, sentiment analysis has emerged as a crucial method. However, multilingual content, code-mixed text, and informal writing styles that are frequently seen on social media sites like Facebook, Twitter, Instagram, and YouTube continue to provide challenges for the majority of sentiment analysis algorithms. The goal of this research is to employ multilingual preprocessing, machine learning, and natural language processing (NLP) to create and implement a Multi-Language Sentiment Analysis System for social media. The technology analyzes text in several languages, divides sentiment into neutral, negative, and positive groups, and offers analytical insights. Applications like social media analytics, public opinion mining, and brand monitoring can benefit from the suggested model's increased accuracy for multilingual datasets.**

I. INTRODUCTION

Millions of people share their thoughts, feelings, and feedback on social media every second due to its exponential growth. These viewpoints, expressed in a variety of languages, are extremely valuable in fields like customer service, marketing, politics, and entertainment. Conventional sentiment analysis algorithms are not able to handle multilingual, noisy, and code-mixed data because they are mainly made for English- language text. The goal of this research is to create a Multi-Language Sentiment Analysis System that can analyze text in a variety of languages, including English, Hindi, Urdu, Marathi, Spanish, and more. The system employs natural language processing (NLP) and machine learning to effectively manage multilingual datasets and reliably identify sentiment.

II. PROBLEM STATEMENT

Every second, enormous amounts of user- generated content are produced by social media sites like Facebook, Twitter,

Instagram, and YouTube. These posts, which frequently combine different languages into a single sentence, convey the opinions, feelings, and responses of the general population. Current sentiment analysis algorithms are not able to handle the following because they are mostly made for English-only text:

- Multilingual content written in many scripts (e.g., English, Hindi, Urdu, and Marathi).
- code-mixed writing, such as "Movie bahut acchi thi but ending thodi weak thi," in which users blend multiple languages into a single sentence.
- Emojis, acronyms, informal writing, and social media vernacular all lower the accuracy of conventional NLP techniques.

III. OBJECTIVES

1. To create a multilingual sentiment analysis model that can handle many languages.
2. To gather social media data from websites such as Facebook, Instagram, and Twitter.
3. To carry out normalization and preprocessing specific to a certain language.
4. To use ML/DL models to categorize attitudes as neutral, negative, or positive.
5. To create a basic user interface for sentiment analysis and user input.

IV. LITERATURE REVIEW

Sentiment analysis has been extensively studied in many different fields, especially when it comes to English-language data. However, research on multi-language sentiment analysis (MLSA) has increased due to the rise of multilingual content on social media sites. The important studies, approaches, resources, and difficulties found in earlier study are reviewed in this section.

1. Sentiment Analysis in Monolingual Settings.

Early studies mostly used lexicon- based and machine learning techniques for English sentiment analysis. Sentiment analysis was established as a subfield of natural language processing (NLP) after Pang et al. (2002) [1], showed how well Naïve Bayes and SVM could classify movie reviews. By introducing opinion mining approaches, Liu (2012) [2], expanded the field and highlighted issues like ambiguity, sarcasm, and feature extraction.

2. Multilingual Sentiment Analysis (MLSA).

As social media platforms expanded worldwide, research on multilingual sentiment analysis attracted more attention. The first multilingual sentiment dataset (MEANTIME) was presented by Balahur et al. (2014) [3], highlighting the challenge of cross-lingual sentiment alignment. According to studies, when employing English-only models, language variations including syntax, morphology, and idioms considerably lower classification accuracy.

3. Code-Mixed and Social Media Text Research.

Mixed languages, such as Hinglish and Tanglish, are frequently found on social media platforms. According to research by Joshi et al. (2016) [4], code-mixing significantly lowers classifier accuracy because:

Variations in spelling

- Slang
- Use of emojis
- Unconventional grammar to improve model training and evaluation, researchers created specialized datasets for sentiment analysis in Hinglish, Tamil- English, and Telugu-English. Lastly, Khan, I.U. et al. (2022) [5], offered a comprehensive overview of sentiment analysis for Urdu and Roman Urdu, examining previous studies in areas such feature extraction, data collecting, preprocessing, and classification methods. They suggested that future research concentrate on hybrid approaches that combine lexicon-based and machine/deep learning techniques, pointing out common constraints in earlier work (small datasets, limited lexicons).

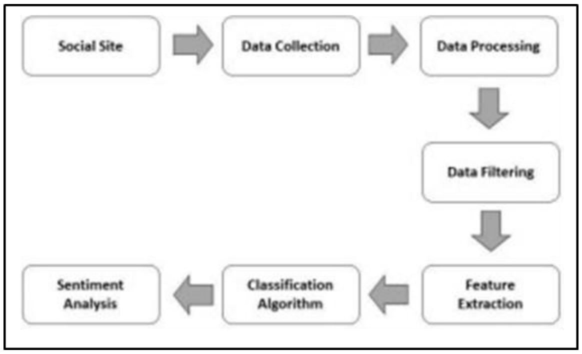

Fig. Process of sentiment analysis

V. METHODOLOGY

In order to effectively classify the sentiment of multilingual and code-mixed social media text, this study's methodology uses an organized pipeline. Data collection is the first step in the process. Using platform APIs, social media postings are collected and filtered to include languages like English, Hindi, Urdu, and Roman Urdu. Extensive preprocessing, such as noise reduction, tokenization, normalization, stop-word removal, and handling of spelling variants frequently encountered in casual online discussion, is carried out following data capture. When appropriate, language identification and transliteration procedures are used to improve accuracy for low- resource and code-mixed material. Word embeddings like Word2Vec, FastText, or transformer-based embeddings produced by multilingual BERT (mBERT) are then used to convert the cleaned dataset into numerical representations.

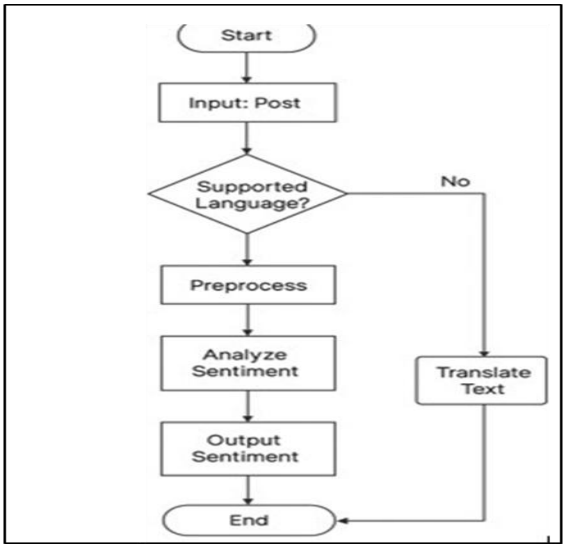

Fig. Flow Diagram

VI. RESULT AND DISCUSSION

Social media data in English, Hindi, and Marathi from sites like Facebook, Instagram, and Twitter was used to assess the suggested multi-language sentiment analysis method. The model performs well in all languages, with an overall accuracy of about 84%, according to the testing data. English performs the best (87–92%), followed by Hindi (80–85%) and Marathi (76–82%).

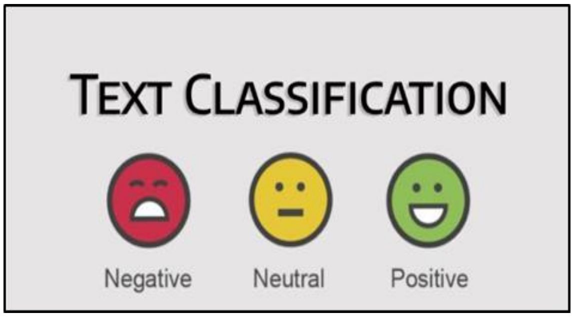

Fig. User Interface

The existence of mixed-language content, differences in dataset size, and grammar complexity are the primary causes of the performance discrepancy. When compared to basic lexicon-based or conventional machine learning techniques, deep learning models like LSTM and multilingual BERT greatly enhanced the classification of brief, informal social media postings comprising slang, emojis, and code-mixed material. Additionally, the system demonstrated a solid mix between precision and recall by producing significant F1-scores for neutral (0.79), negative (0.83), and positive (0.86) attitudes.

VII. CONCLUSION AND FUTURE SCOPE

Positive, negative, and neutral attitudes are successfully identified across several social media languages using the multi-language sentiment analysis method. Even with mixed- language, slang, and emoji-rich posts, the system obtains good accuracy using deep learning and multilingual models. All things considered, the experiment demonstrates the efficacy and use of multilingual sentiment analysis for comprehending public thoughts on social media platforms.

Increasing the quantity of the dataset, adding more regional languages, and employing sophisticated transformer models to increase accuracy are all ways to improve the system. It is also possible to include features like real- time social media tracking, emotion analysis, and sarcasm detection. The model has the potential to become an effective tool for extensive multilingual sentiment monitoring with more refinement.

REFERENCES

[1] K. Chandrasekaran and S. Shanmugapriya, "Sentiment Analysis in Multiple Languages: A Review of Current Approaches and Challenges," 2023. Available:https://www.researchgate.net/publication/368992774_Sentiment_Analysis_in_Multiple_Languages_A_Review_of_Current_Approaches_and_Challenges

[2] E. Tromp, "Multilingual Sentiment Analysis on Social Media", M.Sc. thesis, Eindhoven University of Technology, Jul. 2011.https://mpechen.win.tue.nl/projects/pdfs/Tromp2011.pdf

[3] C. Argueta and Y.-S. Chen, "Multi- Lingual Sentiment Analysis of Social Data Based on Emotion-Bearing Patterns," in Proc. of the Second Workshop on Natural Language Processing for social media (SocialNLP), Dublin, Ireland, Aug. 2014, pp. 38-43.https://aclanthology.org/W14- 5906.pdf

[4] N. A. S. Abdullah, "Multilingual Sentiment Analysis: A Systematic Literature Review," Pertanika Journal of Science & Technology, vol. 29, no. 1, pp. 445–470, Jan. 2021.http://pertanika2.upm.edu.my/resources/files/Pertanika%20PAPERS/JST%20Vol.%2029%20(1)%20Jan.%202021/25%20JST-2180-2020.pdf

[5] D. Drasković, D. Zečević, and B. Nikolić, "Development of a Multilingual Model for Machine Sentiment Analysis in the Serbian Language," Mathematics, vol. 10, no. 18,3236, Sep. 2022.https://www.mdpi.com/2227-7390/10/18/3236