Sentiment Analysis in Hindi

Pratham R Shetty
Dept. Computer Science and Technology
PES University
Bengaluru, India
prathamshetty0826@gmail.com

Pradeep V Naik
Dept. Computer Science and Technology
PES University
Bengaluru, India
knaik1935@gmail.com

Pragnya Vempati
Dept. Computer Science and Technology
PES University
Bengaluru, India
vempati.pragnya@gmail.com

Abstract-Sentiment analysis, a text analytics technique leveraging natural language processing (NLP) and machine learning, serves as a powerful tool in determining the emotional tone of messages. Also referred to as "opinion mining" or "emotion artificial intelligence," it discerns emotions like happiness, frustration, anger, and sadness. Sentiment analysis has been predominantly focused on English, neglecting widely spoken languages like Hindi. In this project, we explore a pertinent use case focusing on the client-provider relationship in a multilingual environment, particularly when the client expresses opinions in Hindi, and the provider lacks proficiency in the language.

Keywords-

I INTRODUCTION

In today's digital age, social media and online interactions play a crucial role in how businesses connect with customers and understand their needs. Sentiment analysis, also known as opinion mining, is a powerful tool that uses natural language processing (NLP) and machine learning to detect emotions in text. It helps businesses determine whether customer feedback is positive, negative, or neutral. It can also be used to decipher slightly more complex human emotions such as happiness, frustration, anger, and sadness from written expressions.

Sentiment analysis plays a critical role in a variety of contexts, from gauging public sentiment on social issues to understanding consumer reactions to products and services. Its applications are important in areas where the textual data is multilingual, which presents unique challenges and opportunities for global businesses.

One significant issue is the language barrier between service providers and clients. This is especially prevalent in cases where clients express opinions in a language not fully mastered by the service providers, as often occurs with Hindi—a language spoken by over 340 million people as a first language.

Our project focuses on filling this gap by developing a system that can analyze sentiments in Hindi text.It addresses the challenge businesses face in understanding multilingual customer feedback. Accurately gauging customer satisfaction in languages like Hindi is crucial for businesses in a globalized market. By leveraging sentiment analysis, this project empowers businesses to understand and respond to customer feedback, ultimately improving products and services. Our system will allow businesses to better understand and respond to customer feedback in Hindi. This understanding can lead to more informed decisions and improvements in products and services, catering specifically to the needs of Hindi-speaking customers. Through this project, we aim to empower businesses to overcome language barriers and enhance communication with their customers, ensuring their needs are met and their satisfaction is achieved.

II RELATED WORK

In exploring the domain of sentiment analysis, various researchers have provided deep insights into the application of machine learning and natural language processing techniques across different languages and contexts. Pipalia et al.[1] conducted a comprehensive comparative analysis of various transformer-based architectures such as BERT, RoBERTa, DistilBERT, XLNet, and T5, highlighting the superior performance of XLNet with an accuracy of 96.2% on the IMDb reviews dataset. Takawane et al. [13]focused on enhancing BERT models' effectiveness through language augmentation techniques for code-mixed Hindi–English text, demonstrating significant improvements in model performance. Despite their success, they acknowledged the model's limited applicability to other language pairs and the overarching reliance on pre-trained models, which might hinder broader applicability.

Pathak et al. [11] introduced innovative ensemble methods using multilingual BERT (mBERT) models to handle ABSA in Hindi, achieving significant advancements in accuracy and F1-scores. However, they highlighted issues such as the dependency on the quality of pre-training and the need for computational resources. Yadav et al. [9] have also explored Aspect-Based Sentiment Analysis (ABSA) in Hindi for e-commerce product reviews, utilizing Support Vector Machine (SVM) techniques to achieve nuanced sentiment classification, though they noted the challenge of limited resources for Hindi sentiment analysis.

Shrestha et al. [10]addressed sentiment analysis in Hindi through Natural Language Processing (NLP), employing classifiers like Naïve Bayes and Decision Trees to analyze text, but they too recognized the limitations related to the complexity of the language and the need for large, robust datasets for machine learning applications. Goel [3]focused on sentiment analysis of multilingual Twitter data, employing algorithms like Recurrent Neural Networks (RNN) and Naive Bayes to analyze sentiments expressed in tweets. The research by Rakshitha et al. [4]applies sentiment analysis to customer reviews in Indian regional languages on Twitter using techniques like Naive Bayes and lexicon-based approaches. These studies collectively advance our understanding of multilingual sentiment analysis and highlight the importance of targeted methodologies and resources to address the unique challenges presented by each language.

Lo et al. [2] addressed the challenges of multilingual sentiment analysis, particularly the analysis of informal or mixed-language content on social media. by advocating for a hybrid approach combining knowledge-based and statistical methods to improve accuracy in sentiment classification across diverse linguistic landscapes. Nanavati [5] focuses on the sentiment analysis in various Indian languages, underscoring the need for advanced preprocessing and feature extraction techniques to handle the linguistic diversity found in languages such as Hindi, Bengali, and Tamil. Nanavati's research emphasizes the integration of machine learning and lexicon-based approaches to improve the accuracy of sentiment analysis. In a similar vein, Nasukawa [6]delves into the semantic intricacies of sentiment analysis by employing syntactic parsers and sentiment lexicons. His work is pivotal in highlighting the importance of understanding the relationships between sentiment expressions and their subjects, which is crucial for extracting sentiments with high precision from complex text sources like web pages and news articles.

Singh and Lefever [7] investigate the challenges associated with code-mixed social media text, particularly in Hinglish, which blends Hindi and English. They explore the use of unsupervised cross-lingual embeddings to enhance sentiment analysis capabilities, demonstrating the effectiveness of these embeddings in handling the linguistic complexities of code-mixed text. Pandey [8] addresses the specific challenges of sentiment analysis in Hindi, focusing on movie reviews. Her work with HindiSentiWordNet to parse and understand sentiment in Hindi texts illustrates the potential and necessity for specialized resources and techniques tailored to specific languages and contexts.

Joshi et al.a [12] developed a Subword-LSTM model to analyze sentiment in Hindi-English code-mixed text, demonstrating the model's superior performance over traditional approaches, yet they cited challenges like dataset size and the availability of NLP tools for code-mixed languages.

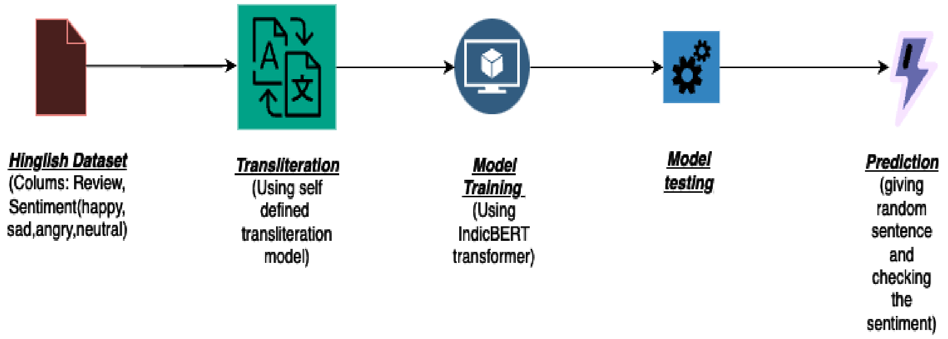
These studies collectively showcase the dynamic approaches and methodologies in sentiment analysis, from sophisticated transformer-based models addressing broad multilingual datasets to focused analyses of specific linguistic features in less commonly addressed languages like Hindi. Each research effort contributes to a deeper understanding of the nuanced requirements and potential advancements in the field of sentiment analysis.

III DATASET

The dataset we picked is

IV PROPOSED METHODOLOGY

A. Architecture Diagram


Hinglish Dataset
(Colums: Review, Sentiment(happy, sad,angry,neutral)
Transliteration
(Using self defined transliteration model)
Model Training
(Using IndicBERT transformer)
Model testing
Prediction
(giving random sentence and checking the sentiment)

B. MODEL DEFINITIONS

We have tested various models against our dataset.

They include LSTM, CNN+LSTM, and transformers like BERT, mBERT, IndicBERT, DistillBERT all fine-tuned to suit our dataset and get the highest possible accuracy.

1) LSTM

Long Short-Term Memory (LSTM) networks, a sophisticated subclass of recurrent neural networks (RNNs), are particularly effective in addressing sequence prediction challenges. These networks are distinguished by their unique internal architecture comprising three distinct types of gates: input, forget, and output. Each gate plays a critical role in managing information flow within the network by selectively retaining, discarding, or allowing information to pass through at each timestep, thereby addressing the key challenges of vanishing gradients that plague traditional RNNs. This gating mechanism equips LSTMs to excel in applications requiring nuanced comprehension of lengthy data sequences, such as in natural language processing tasks like speech recognition and text generation, as well as complex predictive modeling in finance and healthcare sectors.

2) CNN-LSTM Architectures

The CNN-LSTM architecture integrates the spatial feature extraction prowess of Convolutional Neural Networks (CNNs) with the sequential data processing capabilities of Long Short-Term Memory (LSTM) networks. This hybrid model is particularly adept at handling applications involving sequences of spatial data, such as video classification and frame sequence analysis in domains like sports analytics and security surveillance. Within this architecture, CNN layers are tasked with extracting spatial features from individual frames, whereas LSTM layers process these frames in sequence to capture temporal dynamics, thereby offering a comprehensive analysis of both spatial and temporal elements. The dual-layer integration, however, increases the model's complexity and computational demands, necessitating significant resources for training and operational execution.

3) BERT

Bidirectional Encoder Representations from Transformers (BERT) utilizes a transformer-based architecture to revolutionize natural language processing. Unlike traditional models that process words sequentially, BERT evaluates the context of each word in conjunction with all other words in the sentence simultaneously. This bidirectional approach allows for a more nuanced understanding of language, as it captures context from both left and right sides of each word. BERT has significantly enhanced the performance of numerous NLP tasks, including text translation, sentiment analysis, and entity recognition. Despite its advantages, the complexity and computational intensity of BERT's architecture limit its deployment primarily to environments with substantial processing capabilities.

4) mBERT

Multilingual BERT (mBERT) extends the groundbreaking features of BERT to multiple languages by training on a diverse multilingual corpus. This expansion enables mBERT to effectively handle NLP tasks across different languages, making it an invaluable tool for global applications such as multilingual customer support and international market analysis. However, the performance of mBERT may vary with languages that have fewer training data, underlining the challenge of maintaining high performance uniformly across a wide range of languages.

5) IndicBERT

IndicBERT leverages the ALBERT architecture, a lighter and more efficient variant of BERT, to specifically address the linguistic challenges presented by the multitude of Indian languages. By focusing on the nuances and diversity of Indian linguistic features, IndicBERT is optimized for tasks like sentiment analysis, regional language translation, and local content generation. This model is particularly beneficial for applications designed to cater to India's varied linguistic landscape. Nonetheless, the variation in dialects and scripts among Indian languages may affect the consistency of its performance across different linguistic contexts.

6) DistilBERT

DistilBERT streamlines the original BERT architecture to maintain a large proportion of its effectiveness while significantly reducing its size and complexity. This distilled version is particularly suited for deployment in environments constrained by computational resources, such as mobile and edge devices. DistilBERT successfully balances performance and efficiency, enabling the deployment of advanced NLP capabilities in resource-limited settings without sacrificing the ability to handle complex or nuanced tasks. This makes it an ideal choice for applications requiring both high performance and real-time processing efficiency.

V EXPERIMENTAL RESULTS

Model	**Accuracy**

LSTM	0.7888
BERT	0.8234
mBERT	0.9157
DistillBERT	0.8
IndicBERT	0.8703

VI CONCLUSION

References

1. Pipalia, K., Bhadja, R., & Shukla, M., "Comparative Analysis of Different Transformer Based Architectures Used in Sentiment Analysis."
2. Lo, S. L., Cambria, E., Chiong, R., & Cornforth, D. (2016). "Multilingual sentiment analysis: from formal to informal and scarce resource languages." Artif Intell Rev. DOI: 10.1007/s10462-016-9508-4.
3. Goel, V. "Sentiment Analysis of Multilingual Twitter Data using NLP and Deep Learning Methods."
4. Rakshitha, K., Ramalingam, H. M., Pavithra, M., Advi, H. D., & Hegde, M. (2021). "Sentimental Analysis of Indian Regional Languages on Social Media." Global Transitions Proceedings. DOI: https://doi.org/10.1016/j.gltp.2021.08.039.
5. Nanavati, N., "Exploration of Sentiment Analysis in Indian Languages," Sarvajanik College of Engineering & Technology, Surat, India.
6. Nasukawa, T., "Sentiment Analysis: Capturing Favorability Using Natural Language Processing," IBM Research, Tokyo, Japan.
7. Singh, P., Lefever, E., "Unsupervised Cross-Lingual Embeddings for Sentiment Analysis in Hinglish," Ghent University, Belgium.
8. Pandey, P., "Sentiment Analysis Framework for Hindi Movie Reviews Using HindiSentiWordNet," Department of Computer Engineering, PIIT, New Panvel, India.
9. Yadav, V., Verma, P., & Katiyar, V. (2021). "E-commerce Product Reviews Using Aspect Based Hindi Sentiment Analysis," Proceedings of the 2021 International Conference on Computer Communication and Informatics.
10. Shrestha, H., Dhasarathan, C., Munisamy, S., & Jayavel, A. (2020). "Natural Language Processing Based Sentimental Analysis of Hindi (SAH) Script an Optimization Approach," International Journal of Speech Technology.
11. Pathak, A., Kumar, S., Roy, P.P., & Kim, B.-G. (2021). "Aspect-Based Sentiment Analysis in Hindi Language by Ensembling Pre-Trained mBERT Models," Electronics.
12. Joshi, A., Prabhu, A., Shrivastava, M., & Varma, V. (2016). "Towards Sub-Word Level Compositions for Sentiment Analysis of Hindi-English Code Mixed Text," Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics.
13. Takawane, G., Phaltankar, A., Patwardhan, V., Patil, A., Joshi, R., & Takalikar, M.S. (2023). "Language Augmentation Approach for Code-Mixed Text Classification," Natural Language Processing Journal.