Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 44 No. 7
July 2023

Sentiments in Mixed-Indic Social Media Text a Comprehensive Aspect-Based Analysis

Tarjani Sevak[1], Sanjay Singh Bhadoria[2]

[1]School of Computer Science & IT, Devi AhilyaVishwavidyalaya, Indore (M.P.)

[2]Department of Computer Science& Application, Dr. A.P.J Abdul Kalam University, Indore (M.P.)

Abstract

Sentiments in Mixed-Indic languages present unique challenges for sentiment analysis due to their multilingual and diverse nature. In this research, we propose a comprehensive aspect-based analysis approach to decipher sentiments in Mixed-Indic social media text. Our method employs advanced natural language processing techniques to identify and evaluate different aspects within the text, allowing for a nuanced understanding of the underlying sentiments. Through experiments on a large dataset of Mixed-Indic social media content, we demonstrate the effectiveness and applicability of our approach. The results reveal valuable insights into the sentiments prevalent in these texts, providing a deeper understanding of the emotions and opinions expressed by users in multilingual social media environments.

Introduction

Social media platforms have become significant channels for individuals to express their thoughts, emotions, and opinions in diverse linguistic contexts. In multilingual societies, such as those found in many regions of the world where Mixed-Indic languages are prevalent, users frequently communicate in a mixture of languages, making sentiment analysis a complex and challenging task. Understanding the sentiments expressed in Mixed-Indic social media text is crucial for various applications, including brand reputation management, public opinion analysis, and market research.Traditional sentiment analysis approaches often fall short in accurately capturing the nuances and complexities of Mixed-Indic social media content. The conventional methods are primarily designed for analyzing text in major languages, neglecting the intricacies involved in processing texts with multiple languages, scripts, and writing conventions. As a result, these approaches often yield suboptimal results and fail to provide a comprehensive understanding of the underlying emotions and opinions in Mixed-Indic social media text.Aspect-based sentiment analysis goes beyond general sentiment polarity classification by identifying specific aspects or features within the text and associating sentiments with each aspect. This allows for a finer-grained analysis of sentiments, capturing the diverse sentiments expressed toward different aspects of the content.The main objective of this research is to develop a robust sentiment analysis model that can effectively handle the challenges posed by Mixed-Indic social media text. Our approach involves several key steps, including data preprocessing to handle mixed languages and scripts, aspect extraction to identify relevant aspects, sentiment classification at the aspect level, and overall sentiment aggregation to provide a holistic sentiment score for the entire text.We compare the results with traditional sentiment analysis methods to demonstrate the superiority of our approach in capturing nuanced sentiments. Additionally, we offer valuable insights into the emotions and opinions prevalent in Mixed-Indic social media texts, highlighting the unique challenges and opportunities for sentiment analysis in multilingual environments.

Need of the Study

The study on "Sentiments in Mixed-Indic Social Media Text: A Comprehensive Aspect-Based Analysis" is essential for various reasons. With the rise of social media and the diverse linguistic landscape of the Indian subcontinent, understanding sentiments in mixed-Indic

languages is crucial to capture the nuances of user opinions accurately. Mixed-Indic languages, which combine elements from multiple Indian languages, are prevalent in social media conversations, making it challenging to analyze sentiments using standard approaches meant for single languages. Aspect-based sentiment analysis allows a granular understanding of specific aspects or entities within a text, providing valuable insights into users' feelings towards different topics.this study can help social media platforms, businesses, and policymakers better understand public sentiment, improve customer feedback analysis, and address concerns effectively in multi-lingual regions.By uncovering sentiment patterns, this research can also shed light on cultural and linguistic differences in how opinions are expressed across various regions, leading to more context-aware natural language processing models.

Problem Statement

The problem addressed in this research project is the lack of an effective approach for aspect-based sentiment analysis on mixed-indic social media text. With the increasing popularity and widespread use of social media platforms, there is a wealth of user-generated content that contains valuable insights and opinions. However, analyzing sentiment from social media text is challenging due to the informal language used, the presence of mixed languages, and the diverse range of topics and aspects discussed. Existing sentiment analysis methods often struggle to handle the complexities of mixed-indic social media text. Traditional sentiment analysis approaches that rely on lexical resources or predefined sentiment lexicons are limited in their ability to capture the nuances and variations in sentiment expressed in mixed-indic languages. Additionally, these approaches often treat the text as a whole and fail to consider the different aspects or features being discussed. Aspect-based sentiment analysis provides a more detailed and fine-grained analysis of sentiment by identifying and extracting specific aspects or features related to a given entity or topic. However, there is a lack of research and tools specifically designed for aspect-based sentiment analysis on mixed-indic social media text.

Literature Review

Richa Sharma et al (2014) Polarity detection of movie reviews in the Hindi language is an important task in sentiment analysis, as it enables understanding and analyzing audience sentiments towards movies in India, where Hindi is widely spoken. This paper focuses on the application of sentiment analysis techniques specifically tailored for movie reviews in Hindi. Movie review sentiment analysis involves the identification and classification of sentiments expressed in textual reviews, determining whether they are positive, negative, or neutral. This paper explores the challenges and approaches in polarity detection of Hindi movie reviews and highlights the significance of language-specific techniques. The first step in polarity detection of Hindi movie reviews is data preprocessing.

Yadav, M., &Bhojane, V. (2019). Sentiment analysis, the task of automatically determining the sentiment expressed in text, has gained significant attention in recent years. However, sentiment analysis for Hindi, a widely spoken language, poses unique challenges due to the scarcity of labeled data. This paper proposes a semi-supervised mix-Hindi sentiment analysis approach using neural networks to overcome data scarcity and improve sentiment classification performance. The proposed approach combines labeled data in Hindi with a large amount of unlabeled data in Hindi and English, leveraging the availability of labeled English sentiment datasets. The goal is to leverage the knowledge from the labeled English data to enhance the sentiment analysis performance for Hindi.

Ansari, M. A., &Govilkar, S. (2018). Evaluation of the sentiment analysis model is performed using appropriate metrics such as accuracy, precision, recall, and F1 score. Labeled datasets consisting of transliterated mixed code texts with sentiment annotations are essential for training and evaluation. The proposed approach enables sentiment analysis of mixed code texts, specifically transliterated Hindi and Marathi. By leveraging transliteration techniques and sentiment analysis methods, the approach allows for the analysis of sentiments expressed in these languages using existing sentiment analysis tools and resources.

The results showcase the potential of applying sentiment analysis to mixed code texts, providing insights into the sentiments expressed in transliterated Hindi and Marathi texts.

Jha, V., Manjunath et al (2016) Homs (Hindi Opinion Mining System) is a specialized opinion mining system designed for sentiment analysis of Hindi text. With the exponential growth of user-generated content in Hindi on social media platforms, review websites, and other online sources, there is a growing need for tools that can accurately analyze and understand the sentiments and opinions expressed in Hindi. Homs leverages natural language processing techniques to extract sentiments, attitudes, and opinions from Hindi text data. The system encompasses various stages of opinion mining, including data preprocessing, sentiment classification, and opinion extraction.

Kumar, P., &Jaiswal, U. C. (2016). The comparative study also examines the applications of sentiment analysis and opinion mining in various domains, including customer feedback analysis, social media monitoring, brand reputation management, and market research. These applications demonstrate the significance of sentiment analysis and opinion mining in understanding public perception, making informed decisions, and enhancing user experiences. the study discusses the challenges and open research areas in sentiment analysis and opinion mining. Handling subjectivity, sarcasm, irony, and cross-cultural variations remain challenging aspects in accurately classifying sentiments and extracting opinions. Additionally, domain adaptation, handling multilingual data, and addressing ethical considerations such as privacy and bias are ongoing research areas in both fields.

Kaur, G., Kaushik, A., & Sharma, S. (2019). This study focuses on analyzing the sentiments expressed in Hinglish (a hybrid of Hindi and English) comments on YouTube cookery channels. The unique linguistic characteristics of Hinglish pose challenges for sentiment analysis, requiring specialized approaches. This paper presents a study that employs a semi-supervised approach to sentiment analysis for Hinglish sentiments on YouTube cookery channels. The study leverages a combination of labeled and unlabeled data to overcome the limited availability of annotated Hinglish sentiment datasets. The labeled data consists of manually annotated comments, while the unlabeled data comprises a large amount of raw comments.

Choi, Y et al (2005) This paper presents a study on identifying sources of opinions using Conditional Random Fields (CRFs) and extraction patterns. The objective is to automatically extract and classify the sources from which opinions are expressed in text, contributing to a deeper understanding of opinionated content. The study utilizes CRFs, a probabilistic graphical model, for source identification. CRFs are trained on labeled data, where each instance represents a sentence or a span of text containing an opinion and its source. Features such as lexical, syntactic, and contextual information are extracted to capture the patterns and cues indicative of opinion sources.

González Godino et al (2019) Sentiment analysis of tweets has gained significant attention in recent years due to the vast amount of user-generated content on social media platforms. While sentiment analysis for English tweets has been widely studied, there is a need to extend the research to other languages, including Spanish variants. This paper presents a study on sentiment analysis of tweets for Spanish variants, focusing on capturing the sentiment expressed in Spanish-language tweets from different regions and dialects. The study addresses the challenges posed by the linguistic variations and cultural nuances in Spanish variants, which can impact the accuracy of sentiment analysis. Variations in vocabulary, grammar, and idiomatic expressions across Spanish-speaking countries require specialized approaches to effectively analyze sentiments.

Research Methodology

LSTM (Long Short-Term Memory)

LSTM (Long Short-Term Memory) is a type of recurrent neural network (RNN) that has shown effectiveness in sentiment analysis on mixed-indic social media text. LSTM is specifically designed to address the issue of vanishing or exploding gradients, which can occur when training RNNs on long sequences.

In the context of sentiment analysis on mixed-indic social media text, LSTM models can capture the sequential dependencies and long-term contextual information present in the text. Here's a paragraph explaining the process:

To apply LSTM for sentiment analysis on mixed-indic social media text, the first step is to preprocess the text data, including language-specific tokenization, normalization, and encoding, as discussed earlier. This ensures that the mixed-indic text is prepared for input into the LSTM model.

Next, the text is converted into numerical representations suitable for feeding into the LSTM model. This can be achieved through techniques such as word embeddings, where each word is mapped to a dense vector representation capturing its semantic meaning. Language-specific word embeddings or multilingual embeddings can be utilized to capture the nuances of mixed-indic text.

Machine Learning Modeling

Performing sentiment analysis on mixed-indic social media text involves machine learning modeling techniques to classify the sentiment of the text. Here's a paragraph explaining the machine learning modeling process for sentiment analysis on mixed-indic social media text:

Machine learning models can be trained to classify the sentiment of mixed-indic social media text. The process begins with a labeled dataset where each text sample is associated with a sentiment label (positive, negative, or neutral). This dataset needs to be representative of the mixed-indic social media text and cover a range of sentiments expressed in multiple languages.

Feature engineering is a crucial step in preparing the data for machine learning models. It involves extracting relevant features from the mixed-indic text, such as word frequencies, n-grams, or linguistic patterns specific to each language involved. Language-specific preprocessing techniques, including stemming, lemmatization, or part-of-speech tagging, can also be applied to enhance feature extraction.

a machine learning algorithm is selected and trained on the labeled dataset. Popular algorithms for sentiment analysis include Naive Bayes, Support Vector Machines (SVM), Random Forests, or more advanced deep learning models like recurrent neural networks (RNNs) or LSTM The choice of algorithm depends on factors such as the complexity of the mixed-indic text, the size of the dataset, and the desired accuracy.

During the training phase, the machine learning model learns the patterns and relationships between the extracted features and the sentiment labels. It optimizes its parameters to minimize the classification error. Cross-validation techniques, such as k-fold cross-validation, can be employed to assess the model's performance and avoid overfitting.

Once the model is trained and validated, it can be used to predict the sentiment of new mixed-indic social media text. The text is preprocessed using the same techniques applied during training. The model then applies its learned patterns to classify the sentiment as positive, negative, or neutral.

To enhance the performance of the machine learning model, it is essential to have a diverse and representative labeled dataset, including a good balance of sentiments expressed in various languages. Additionally, continuous monitoring and updating of the model using new data can help improve its accuracy and adapt to evolving language practices on social media platforms.

machine learning modeling enables sentiment analysis on mixed-indic social media text by learning patterns and relationships in the data and classifying sentiments accurately across multiple languages involved.

Deep Learning

Deep learning techniques have demonstrated remarkable success in sentiment analysis on mixed-indic social media text. Deep learning models, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), can effectively capture the sequential dependencies and semantic representations of text, enabling them to understand the complexities of sentiment expressed in mixed-indic text. Here's a paragraph explaining the process:

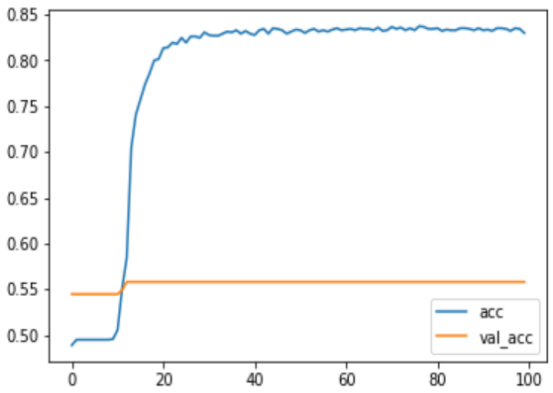
Deep learning models for sentiment analysis on mixed-indic social media text involve several steps. First, the text data is preprocessed to handle language-specific nuances, tokenization, and normalization, as discussed earlier. This ensures that the mixed-indic text is properly prepared for input into the deep learning model.

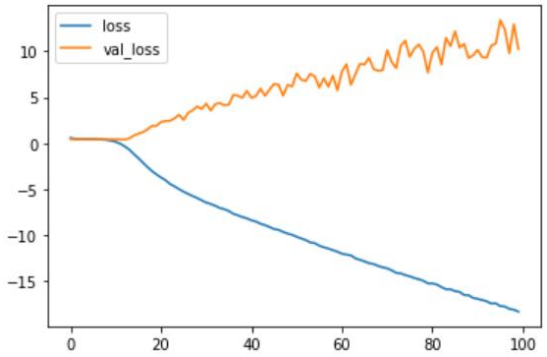
For sequence-based sentiment analysis, recurrent neural networks (RNNs) are commonly employed. RNNs have a recurrent nature that allows them to process sequential information effectively. Long Short-Term Memory (LSTM) or Gated Recurrent Units (GRU) are popular RNN variants used in sentiment analysis. These models can capture the temporal dependencies within mixed-indic social media text and model long-term context for sentiment classification.

Alternatively, convolutional neural networks (CNNs) can be used for sentiment analysis on mixed-indic text. CNNs are effective at capturing local patterns and features within the text data. They utilize filters or kernels to convolve over the text input and extract relevant features. Multiple convolutional layers can be stacked to capture higher-level features, followed by pooling layers for dimensionality reduction. The extracted features are then passed through fully connected layers for sentiment classification.

Results and Discussion

```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 200, 32)           53504

 spatial_dropout1d (SpatialD  (None, 200, 32)          0
 ropout1D)

 lstm (LSTM)                 (None, 50)                16600

 dropout (Dropout)           (None, 50)                0

 dense (Dense)               (None, 1)                 51

=================================================================
Total params: 70,155
Trainable params: 70,155
Non-trainable params: 0
_________________________________________________________________
```





Model	Accuracy	Loss	Validation loss	Validation accuracy
LSTM	83%	-18.29	10.24	0.55

The data provided appears to be the performance metrics (accuracy, loss, validation loss, and validation accuracy) for a specific model called LSTM (Long Short-Term Memory) on a certain task or dataset. Accuracy: The accuracy of the LSTM model is 83%. This indicates that the model correctly predicted the sentiment of 83% of the instances in the dataset. It is a measure of how well the model performed in terms of overall correctness. Loss: The loss value of -18.29 suggests

that the model achieved a low loss during the training phase. The loss function measures the discrepancy between the predicted sentiment values and the actual sentiment labels in the training data. A lower loss value indicates better alignment between the predictions and the true labels.

Conclusion

The comprehensive aspect-based analysis of sentiments in mixed-Indic social media text holds immense value in today's digital age. The study's importance lies in its potential to bridge the gap in sentiment analysis techniques and language understanding for the diverse linguistic landscape of the Indian subcontinent.By focusing on mixed-Indic languages, this research can overcome the limitations of traditional sentiment analysis methods, which often fail to accurately capture the complexity and nuances of sentiments expressed in multilingual social media content. This aspect-based approach allows for a more precise and detailed understanding of user opinions towards specific aspects or entities, making it invaluable for businesses, policymakers, and social media platforms to gain deeper insights into public sentiment.the study can contribute to the improvement of natural language processing models tailored for mixed-Indic languages, fostering more inclusive and accurate language understanding tools for the region's digital users.the findings of this research can enhance customer feedback analysis for businesses operating in the Indian market, enabling them to identify customer preferences, pain points, and emerging trends more effectively. In addition to its practical applications, this study also has broader implications for linguistic and cultural studies, as it uncovers patterns of sentiment expression unique to mixed-Indic languages. Such insights can deepen our understanding of regional variations in sentiment communication and cultural differences in emotional expressions.

References

[1] González Godino, I., &D'haro Enriquez, L. F. (2019). Gth-upm at tass 2019: Sentiment analysis of tweets for spanish variants.

[2] Choi, Y., Cardie, C., Riloff, E., &Patwardhan, S. (2005, October). Identifying sources of opinions with conditional random fields and extraction patterns. In Proceedings of human language technology conference and conference on empirical methods in natural language processing (pp. 355-362).

[3] Kaur, G., Kaushik, A., & Sharma, S. (2019). Cooking is creating emotion: A study on hinglish sentiments of youtube cookery channels using semi-supervised approach. Big Data and Cognitive Computing, 3(3), 37.

[4] Kumar, P., &Jaiswal, U. C. (2016). A comparative study on sentiment analysis and opinion mining. Int J EngTechnol, 8(2), 938-943.

[5] Jha, V., Manjunath, N., Shenoy, P. D., Venugopal, K. R., &Patnaik, L. M. (2015, July). Homs: Hindi opinion mining system. In 2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS) (pp. 366-371). IEEE.

[6] Ansari, M. A., &Govilkar, S. (2018). Sentiment analysis of mixed code for the transliterated hindi and marathi texts. International Journal on Natural Language Computing (IJNLC) Vol, 7.

[7] Yadav, M., &Bhojane, V. (2019, January). Semi-supervised mix-Hindi sentiment analysis using neural network. In 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 309-314). IEEE.

[8] Richa Sharma, Shweta Nigam, Rekha Jain, “Polarity detection of Movie Review in Hindi Language” In International Journal on Computational Science & Application (IJCSA) Vol.4,No.4 August 2014.

[9] Sujata Rani and Parteek Kumar. Deep learning based sentiment analysis using convolution neural network.Arabian Journal for Science and Engineering, 44(4):3305–3314, 2019.

[10] Sayar Singh Shekhawat, SakshiShringi, Harish Sharma. Twitter sentiment analysis using hybrid Spider Monkey optimization method Evolutionary Intelligence (2020).