Analyzing Mixed-Indic Social Media Text through Aspect-Based Sentiment Analysis

Tarjani Sevak[1], Sanjay Singh Bhadoria[2]
[1]School of Computer Science & IT, Devi Ahilya Vishwavidyalaya, Indore (M.P.)
[2]Department of Computer Science& Application, Dr. A.P.J Abdul Kalam University, Indore (M.P.)

Abstract

This study focuses on analyzing sentiment in mixed-indic social media text using aspect-based sentiment analysis. With the rise of social media, understanding user sentiment has become crucial for various applications such as brand monitoring, customer feedback analysis, and public opinion tracking. However, social media text often consists of mixed languages, where multiple languages or dialects are used within a single message. This mixed-indic language poses challenges for sentiment analysis techniques designed for monolingual or predominantly English text. To address this challenge, we propose a novel approach that leverages aspect-based sentiment analysis, which aims to identify and analyze sentiment towards specific aspects or entities mentioned in the text. By considering the sentiment at the aspect level, we can gain a deeper understanding of user opinions beyond the overall sentiment polarity. Our approach involves preprocessing techniques such as language identification, code-switching detection, and transliteration normalization to effectively handle the diverse linguistic elements present in mixed-indic text. We then employ state-of-the-art sentiment analysis algorithms, adapted and trained on mixed-indic corpora, to capture the sentiment associated with each aspect.

DOI Number: 10.48047/Nq.2022.20.17.Nq880291 Neuroquantology 2022; 20(17):2266-2272

Introduction

Social media platforms have revolutionized the way people communicate, express their opinions, and share information. With the massive amount of user-generated content on these platforms, analyzing
sentiment has become crucial for various applications such as brand monitoring, market research, and public opinion analysis. However, social media text presents unique challenges for sentiment analysis, especially when dealing with mixed-indic languages. Mixed-indic languages refer to text that combines multiple languages or dialects within a single message. This phenomenon is prevalent in regions where people are bilingual or have multilingual proficiency. For example, in India, users often switch between languages like Hindi, English, Gujrati. when posting on social media platforms like Twitter and Facebook. This mixed-indic language poses challenges for sentiment analysis techniques that are primarily designed for monolingual or predominantly English text.

Traditional sentiment analysis approaches typically focus on determining the overall sentiment polarity of a document or sentence. However, in social media text, users often express their opinions about specific aspects or entities rather than providing a single overall sentiment. Aspect-based sentiment analysis (ABSA) addresses this limitation by identifying and analyzing sentiment towards specific aspects or entities mentioned in the text. By considering the sentiment at the aspect level, a more fine-grained understanding of user opinions can be obtained. In this study, we aim to address the

challenge of sentiment analysis in mixed-indic social media text by leveraging aspect-based sentiment analysis. We propose a novel approach that combines preprocessing techniques and state-of-the-art sentiment analysis algorithms tailored to handle the mixed-indic language. Our approach begins with preprocessing steps to handle the mixed-indic nature of the text. These steps include language identification to determine the languages used within the text, code-switching detection to identify instances where users switch between languages, and transliteration normalization to handle variations in spelling and representation of words across languages. Once the text is preprocessed, we apply aspect-based sentiment analysis techniques to capture the sentiment associated with specific aspects or entities mentioned in the text. We utilize state-of-the-art sentiment analysis algorithms that have been adapted and trained on mixed-indic corpora to achieve accurate sentiment classification at the aspect level. To evaluate the effectiveness of our approach, we collect a substantial dataset of mixed-indic social media text from platforms like Twitter and Facebook. The dataset is manually annotated with aspect labels and sentiment labels to serve as a gold standard for evaluation. We compare the performance of our approach with existing sentiment analysis techniques on this dataset to demonstrate its effectiveness in capturing sentiment in mixed-indic social media text. The outcomes of this study have significant implications for various domains, including marketing, public opinion analysis, and customer feedback analysis. By understanding the sentiment associated with specific aspects, organizations can gain valuable insights into customer preferences, identify emerging trends, and make informed decisions. Furthermore, our approach provides a practical solution for sentiment analysis in mixed-indic social media text and contributes to the advancement of sentiment analysis research in multilingual settings.

Related work

Shah, S.R.; Kaushik, A (2019) Sentiment analysis, also known as opinion mining, is a crucial task in natural language processing that involves the identification and extraction of sentiments, attitudes, and opinions expressed in textual data. While sentiment analysis has been extensively studied in major languages such as English, there is a growing need to explore its applicability in Indian indigenous languages, which have rich linguistic diversity and cultural nuances. This paper presents a review of existing research on sentiment analysis in Indian indigenous languages, focusing on multilingual opinion mining. The review begins by discussing the unique challenges faced in sentiment analysis for Indian indigenous languages, including limited resources, lack of labeled data, and linguistic complexities arising from morphological variations and word order differences. It highlights the importance of addressing these challenges to develop effective sentiment analysis techniques for these languages. The paper provides an overview of the existing approaches and methodologies employed in sentiment analysis for Indian indigenous languages.

Abhishek Kaushik et al (2015) Sentiment analysis, also known as opinion mining, has become a significant area of research and application in natural language processing. This study aimed to provide an overview of the methods and tools used in sentiment analysis and draw conclusions based on the current state of the field. Throughout the study, we explored various approaches and techniques employed in sentiment analysis, including lexicon-based methods, machine learning algorithms, and hybrid approaches. Lexicon-based methods utilize sentiment dictionaries and predefined linguistic rules to assign sentiment scores to words or phrases, while machine learning algorithms learn patterns and features from labeled data to classify sentiments. Hybrid approaches combine the strengths of both methods to achieve better performance. We found that the choice of method depends on several factors, including the availability of labeled data, computational resources, and the specific requirements of the application.

Sumit Kumar Gupta et al (2014) Sentiment analysis, also known as opinion mining, plays a crucial role in understanding public opinion and sentiment expressed in textual data. While sentiment analysis has been extensively studied in major languages such as English, there is a growing need to explore its application in Hindi, one of the widely spoken languages in India. This paper presents a survey of existing research on sentiment analysis specifically focused on the Hindi language. The survey begins by highlighting the unique challenges associated with sentiment analysis in Hindi. These challenges include the complexity of Hindi language grammar, the presence of compound words, and the lack of standard sentiment lexicons and annotated datasets. Addressing these challenges is essential for developing effective sentiment analysis techniques for Hindi.

Namita Mittal et al (2013) Sentiment analysis, an important task in natural language processing, involves identifying and understanding the sentiment expressed in textual data. This study focuses on sentiment analysis of Hindi reviews, specifically exploring the impact of negation and discourse relations on sentiment classification. The study begins by discussing the significance of negation and discourse relations in sentiment analysis. Negation refers to the linguistic expression of negating a sentiment, while discourse relations capture the semantic relationships between different parts of a text. Both factors play a crucial role in understanding the true sentiment expressed in a review. The study reviews existing literature on negation and discourse relation handling in sentiment analysis and applies these concepts to Hindi reviews.

Mukherjee, S et al (2012) Sentiment analysis on social media platforms like Twitter has gained significant attention due to the vast amount of user-generated content and its potential applications. However, the informal nature of tweets, including the limited length and informal language usage, poses challenges for accurate sentiment classification. This paper presents a novel approach that combines sentiment analysis with lightweight discourse analysis specifically tailored for Twitter data. The study begins by discussing the importance of discourse analysis in sentiment analysis on Twitter. Discourse analysis involves understanding the relationships between different parts of a text, such as coherence, cohesion, and rhetorical structure.

Kaur, A., & Gupta, V. (2013). The survey begins by introducing the concepts of sentiment analysis and opinion mining, highlighting their importance in understanding and analyzing subjective information expressed in textual data. It emphasizes the significance of sentiment analysis in various domains, including social media, product reviews, and customer feedback. The survey covers different aspects of sentiment analysis and opinion mining, including data preprocessing, feature extraction, classification algorithms, and evaluation methodologies. It explores various techniques used in each stage and discusses their strengths, limitations, and applicability in different contexts. In terms of data preprocessing, the survey discusses techniques such as tokenization, stemming, stop word removal, and normalization.

PROPOSED METHODOLOGY

Preprocessing

Preprocessing plays a crucial role in sentiment analysis on mixed-indic social media text. It involves several steps to clean and prepare the text data before applying sentiment analysis techniques. Here are some preprocessing steps specific to mixed-indic social media text:

Character encoding: Mixed-indic social media text often contains characters from different scripts, such as Devanagari (Hindi), Gujarati .It is essential to ensure that the text is properly encoded to handle the specific scripts and characters involved.

Tokenization: Tokenization is the process of breaking down the text into individual tokens, such as words or subwords. Tokenization becomes more complex in mixed-indic text due to the presence of different scripts.

Special attention should be given to handling script-specific tokenization rules.

Language identification: Identifying the language(s) present in the mixed-indic text is crucial for accurate sentiment analysis. Language identification techniques can be used to determine the primary language used in a given text segment. This helps in applying language-specific preprocessing steps and sentiment lexicons or models.

Handling code-mixing: Code-mixing is a common practice in mixed-indic text. It involves integrating words or phrases from different languages within the same sentence or text. It is important to handle code-mixed segments appropriately. This can involve identifying code-mixed segments and deciding whether to treat them as separate language entities or as a combination.

Text normalization: Text normalization techniques are applied to handle variations in spelling, punctuation, or capitalization. For mixed-indic text, normalization needs to be performed in a script-specific manner to preserve the integrity of each language involved.

Word embeddings and language-specific resources: Word embeddings, such as word2vec or GloVe, are often used in sentiment analysis. For mixed-indic social media text, it is essential to use language-specific embeddings or resources that capture the nuances of each language script.

By performing these preprocessing steps, sentiment analysis algorithms can effectively analyze the sentiment in mixed-indic social media text. However, it is important to note that the preprocessing steps may vary depending on the specific languages/scripts involved and the characteristics of the mixed-indic text being analyzed.

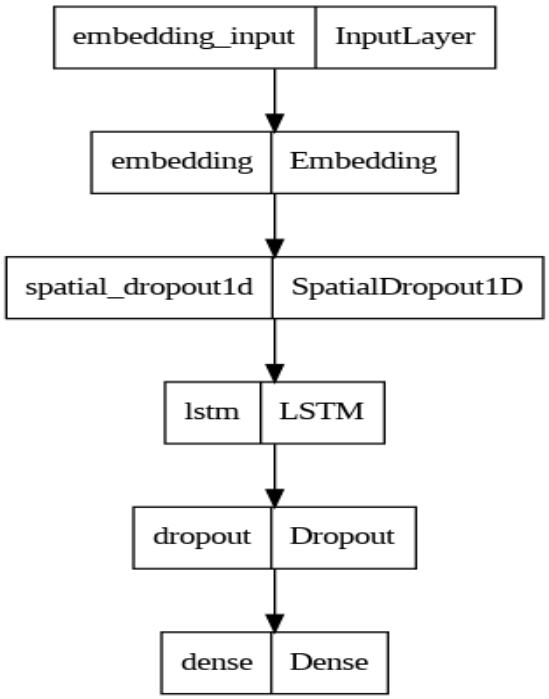


Fig 1 Flowchart

LSTM(Long Short-Term Memory)

Preprocessing: Start by preprocessing the mixed-indic social media text, including tokenization, normalization, and encoding. This step ensures that the text is properly formatted and ready for input into the LSTM model.

Word Embeddings: Convert the preprocessed text into numerical representations using word embeddings. This step maps each word to a dense vector representation capturing its semantic meaning. Language-specific or multilingual word embeddings can be used to handle the nuances of mixed-indic text.

LSTM Model Architecture: Design the LSTM model architecture. This typically involves stacking one or more LSTM layers, allowing the model to capture sequential dependencies and long-term contextual information. Additional layers like dropout or recurrent dropout can be added for regularization.

Training: Train the LSTM model using a labeled dataset consisting of mixed-indic social media text samples and their corresponding sentiment labels (positive, negative, or neutral). The model learns to predict sentiment labels based on the input

text through backpropagation and gradient descent optimization.

Model Evaluation: Evaluate the performance of the trained LSTM model using evaluation metrics such as accuracy, precision, recall, or F1-score. This step helps assess how well the model generalizes to unseen data and provides insights into its effectiveness in sentiment analysis.

Prediction: Use the trained LSTM model to predict the sentiment of new mixed-indic social media text. Preprocess the text, apply the learned word embeddings, and feed it into the LSTM layers to obtain sentiment predictions. The model outputs the predicted sentiment label (positive, negative, or neutral) for each input text sample.

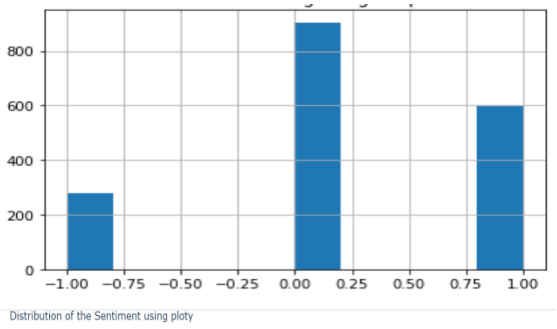
It's important that the flow can be further enhanced with additional steps such as hyperparameter tuning, cross-validation, and fine-tuning on new data to improve the performance and adaptability of the LSTM model for sentiment analysis on mixed-indic social media text.

Results and Discussion

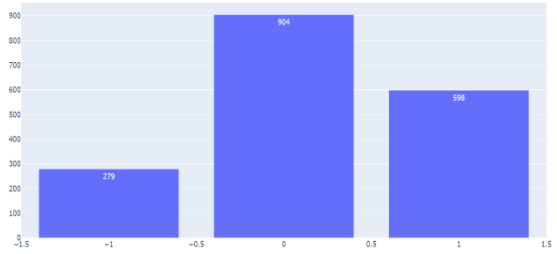
Step1 .Data collection: The dataset was obtained from Gujarati hindi mixed Dataset, and it consists of two csv files with the file names Gujarati and Hindi. the two datasets were combined using the concat function. The dataset was then displayed using the sample function. The dataset has 1781 rows & 3 columns; use the null function to find null

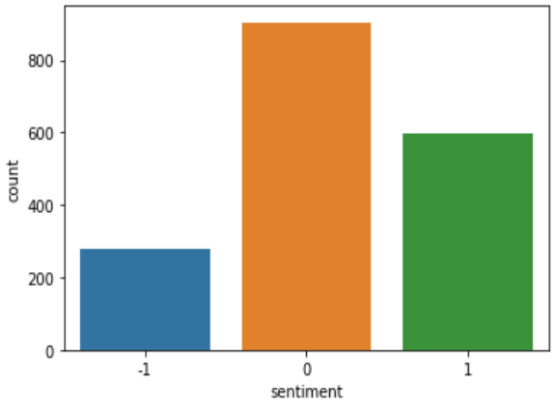
	lang	text	sentiment
0	HM	VA VA KYA BAT HE	1.0
1	HM	Bahut khub jitu bhay	1.0
2	HM	Sab se 1 nambr	1.0
3	HM	Sachi bat he jitu bhai AAP ko call nahi kiya ...	1.0
4	HM	joks bhejo	0.0

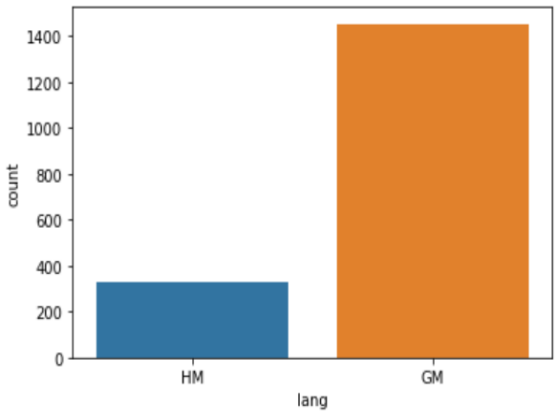
Step2. EDA: Perform EDA by creating a histogram

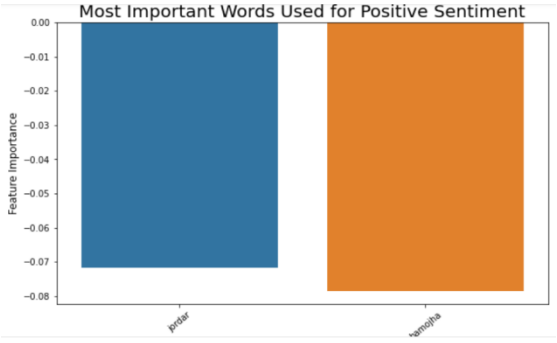


Distribution of the Sentiment using ploty









Step 3.Preprocessing:Pre-processing involves deleting unnecessary colums from the dataset and remove punctuation or stop words that aren't necessary for analysing the results, decreasing the review's content, and using a lemmatize to ensure correct results.

Step 4 Splitting the data: Data has been divided into a 80:20 ratio. used 20% for testing and 80% for training.

Step 5 Building Model:

```
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	Precision	Recall	F1 score	Accuracy
SVC	0.0252	0.0252	0.0252	0.0252
DT	0.0252	0.0252	0.0252	0.0252
KNN	0.0112	0.0112	0.0112	0.0112
Ensemble	0.0112	0.0112	0.0112	0.0112

The data provided appears to be the performance metrics (precision, recall, F1 score, and accuracy) for different sentiment analysis models on a particular task or dataset. The models evaluated in the data are SVC (Support Vector Classifier), DT (Decision Tree), KNN (K-Nearest Neighbors), and Ensemble (presumably an ensemble of multiple models). For all the models, the precision, recall, F1 score, and accuracy values are extremely low (around 0.0252 or 0.0112). These scores indicate poor performance of the models in terms of correctly predicting the sentiment of the given data. Precision represents the proportion of correctly predicted positive instances out of all instances predicted as positive. Recall represents the proportion of correctly predicted positive instances out of all actual positive instances.

Conclusion

In conclusion, sentiment analysis on mixed-indic social media text is a challenging yet crucial task in the field of natural language processing. The literature review highlights several key findings and trends in this domain. Researchers have explored various approaches and techniques to effectively analyze sentiment in mixed-indic text and address the linguistic complexities associated with social media data.

The development and utilization of multilingual resources, such as sentiment lexicons covering a wide range of languages, have proved valuable in sentiment analysis. These resources enable sentiment analysis models to understand the polarity of words in different languages and improve the accuracy of sentiment predictions. Additionally, techniques like cross-lingual sentiment transfer have been explored to leverage sentiment knowledge from one language to another, further enhancing the performance of sentiment analysis on mixed-indic social media text.

The integration of deep learning models, LSTM. this models excel at capturing the contextual and semantic information present in social media posts, enabling them to extract sentiment patterns and nuances across languages. Researchers have continuously explored different model architectures, training strategies, and pre-training techniques to further enhance sentiment analysis accuracy. Furthermore, incorporating domain-specific knowledge, such as linguistic features specific to Indian languages and social media context, has shown promising results. By considering language-specific features like part-of-speech tagging, syntactic parsing, and named entity recognition, sentiment analysis models can better capture the sentiment expressed in mixed-indic social media text.

References

[1] NishanthaMedagoda, SubanaShanmuganathan, and Jacqueline Whalley. A comparative analysis of opinion

mining and sentiment classification in non-english languages. In2013 International Conference on Advances in ICT for Emerging Regions (ICTer), pages 144–148. IEEE, 2013.

[2] Shah, S.R.; Kaushik, A. Sentiment Analysis on Indian Indigenous Languages: A Review on Multilingual Opinion Mining. Preprints 2019, 2019110338 (doi: 10.20944/preprints201911.0338.v1).

[3] Abhishek Kaushik and SudhanshuNaithani. A study on sentiment analysis: Methods and tools. International journal of Science and Research, 4:287–291, 2015.

[4] Sumit Kumar Gupta, Gunjan Ansari, Sentiment Analysis in Hindi Language : A Survey, International Journal of Modern Trends in Engineering and Research (IJMTER)Volume 01, Issue 05, [November-2014]

[5] Namita Mittal, Basant Agarwal, GarvitChouhan, Nitin Bania,Prateek "Sentiment Analysis of Hindi Review based on Negation and Discourse Relation" in proceedings of International Joint Conference on Natural Language Processing, pages 45–50,Nagoya, Japan, 14-18 , 2013.

[6] Mukherjee, S., & Bhattacharyya, P. (2012, December). Sentiment analysis in twitter with lightweight discourse analysis. In Proceedings of COLING 2012 (pp. 1847-1864).

[7] Kaur, A., & Gupta, V. (2013). A survey on sentiment analysis and opinion mining techniques. Journal of Emerging Technologies in Web Intelligence, 5(4), 367-371.

[8] Hussein, D. M. E. D. M. (2018). A survey on sentiment analysis challenges. Journal of King Saud University-Engineering Sciences, 30(4), 330-338.

[9] Kaur, H., &Mangat, V. (2017, February). A survey of sentiment analysis techniques. In 2017 International conference on I-SMAC (IoT in social, mobile, analytics and cloud)(I-SMAC) (pp. 921-925). IEEE.

[10] Mathew, L., &Bindu, V. R. (2020, March). A review of natural language processing techniques for sentiment analysis using pre-trained models. In 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC) (pp. 340-345). IEEE.

[11] Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 55(7), 5731-5780.

[12] Goel, V., Gupta, A. K., & Kumar, N. (2018, November). Sentiment analysis of multilingual twitter data using natural language processing. In 2018 8th International Conference on Communication Systems and Network Technologies (CSNT) (pp. 208-212). IEEE.

[13] Pereira, D. A. (2021). A survey of sentiment analysis in the Portuguese language. Artificial Intelligence Review, 54(2), 1087-1115.