Sentiment Analysis for Hindi Movie Reviews

Prerna Katte
Dept. of Information Technology
Vidyalankar Institute of Technology
Mumbai, India
prerna.katte@vit.edu.in

Sandeep Dhenaki
Dept. of Information Technology
Vidyalankar Institute of Technology
Mumbai, India
sandeep.dhenaki@vit.edu.in

Sanika Gaikwad
Dept. of Information Technology
Vidyalankar Institute of Technology
Mumbai, India
sanika.gaikwad@vit.edu.in

Dr. Vipul Dalal
Dept. of Information Technology
Vidyalankar Institute of Technology
Mumbai, India
vipul.dalal@vit.edu.in

***Abstract* — This research paper presents the development and implementation of an advanced Sentiment Analysis system for Hindi movie reviews, employing diverse methodologies to enhance accuracy and applicability. The study encompasses the utilization of resource-based semantic analysis, in-language semantic analysis, and machine translation-based semantic analysis to effectively classify sentiments of Hindi movie reviews. The dataset, comprising 1000 movie reviews, is meticulously curated from diverse sources, including IIT-Bombay and Jagaran.com, providing a robust foundation for model training and evaluation. The project employs various classification techniques, such as decision trees and deep belief networks, offering a comprehensive understanding of the nuances in sentiment classification. By integrating multiple approaches, the research aims to provide an accurate and reliable sentiment analysis solution for Hindi movie reviews, contributing to the broader field of natural language processing and sentiment analysis. The paper emphasizes the significance of each approach, their respective strengths, and the potential for future improvements in sentiment analysis on multilingual datasets.**

Keywords—Sentiment Analysis, Deep Belief Network, Machine Learning, HindiSentiWordnet, Tensorflow.

I. INTRODUCTION

In the dynamic landscape of sentiment analysis for multilingual content, our research delves into the evolving field of natural language processing applied to Hindi movie reviews. In a world characterized by rapid digital advancements, the analysis of sentiments in non-English languages stands as a crucial area for exploration. This study addresses the intricacies of sentiment classification in the context of Hindi, a widely spoken language with a rich cultural and linguistic heritage. By employing resource-based semantic analysis, in-language semantic analysis, and a machine translation-based approach, our research seeks to unravel the diverse facets influencing sentiment in Hindi movie reviews.

Much like cutting-edge technologies revolutionizing various industries, our research endeavors to redefine the approach to sentiment analysis, aiming to create a nuanced and accurate system for understanding the emotional nuances in Hindi language movie reviews. The vast and varied dataset, carefully curated from multiple sources, provides a robust foundation for training and evaluating our sentiment analysis models. Through the application of advanced technologies, including deep belief networks and machine translation, our research aims to contribute to the broader discourse on sentiment analysis in a multilingual context.

As industries embrace AI for various applications, our research seeks to position sentiment analysis as an indispensable tool for understanding audience reactions in the context of Hindi cinema. The exploration of sentiment in Hindi movie reviews is not merely an examination of language but a deeper understanding of the cultural aspects and expressions unique to Hindi speakers. This research invites readers to join us on this journey, exploring the vast possibilities and challenges inherent in sentiment analysis in the Hindi language.

II. RELATED WORK

We conducted sentiment analysis on a dataset comprising 1000 Hindi movie reviews. Initially, we selected 250 reviews from the IIT-Bombay dataset, evenly distributed between positive and negative sentiments. Additionally, 750 reviews were manually gathered from Jagaran.com, a Hindi movie review website, and categorized into positive and negative sentiments, resulting in 375 reviews for each category. To enable a Machine Translation-based approach, where Hindi reviews are translated into English, we needed English reviews for training. We obtained a diverse set of English reviews from the NLTK dataset for effective translation-based sentiment analysis.

Our research methodology aligns with the importance of a diverse and well-curated dataset in sentiment analysis, as advocated by Joshi, Bhattacharyya, and Carman (2010) [1]. Following a similar approach, we meticulously curated a dataset of 1000 Hindi movie reviews, incorporating 250 labeled reviews from IIT-Bombay and 750 manually collected reviews from Jagran.com. This rigorous data collection is crucial for training and validating sentiment analysis models.

In [2] The authors devised a lexicon through a graph-based method, delving into the utilization of synonym and antonym relations via straightforward graph traversal to formulate a subjectivity lexicon. Their algorithm demonstrated commendable results, achieving an accuracy of around 79% in the classification of reviews, with a notable 70.4% agreement with human annotations. Building upon this foundation, our research extends the approach by leveraging HindiSentiWordnet. This enables the extraction of sentiment polarity and intensity from Hindi movie reviews, contributing to a comprehensive semantic understanding of the expressed sentiments.

In classifier and feature extraction, inspired by Pang and Lee (2008)[3], we explore various types of classifiers tailored to the unique challenges posed by sentiment analysis in Hindi movie reviews. It also dives deep into the process of feature extraction by means of unigrams, term-frequency, parts of speech, negation and other topic-oriented features.

Inspired by the work of He and Wu (2019)[4], we integrate machine translation into sentiment analysis using the Googletrans API to translate a subset of Hindi movie reviews into English, creating a parallel dataset. Evaluating classifier performance on both translated English reviews and original Hindi reviews enables us to assess the feasibility and accuracy of a machine translation-based approach in Hindi sentiment analysis.

In reference to paper [5], the authors demonstrated that incorporating discourse markers into a bag-of-words model for the English language results in an enhancement of sentiment classification accuracy by 2 - 4%.

Published in 2009, the study by Apoorv Agarwal, Fadi Biadsy, and Kathleen R. McKeown [6] focuses on analyzing contextual phrase-level polarity. This involves utilizing a combination of lexical affect scoring and syntactic n-grams, proposing a technique to evaluate the sentiment polarity of phrases within context by considering both lexical information and syntactic structures.

The paper "A Fall-back Strategy for Sentiment Analysis in Hindi: a Case Study"[7] presents three distinct approaches for sentiment analysis in the context of the Hindi language. The first method entails the creation of a sentiment-annotated corpus specifically for Hindi movie reviews, followed by training a classifier on this corpus. This trained classifier is then utilized for the analysis of new Hindi documents. In the second approach, the given Hindi document is translated into English, and a classifier trained on English movie reviews is employed to conduct sentiment analysis. The third method introduces a novel lexical resource named Hindi-SentiWordNet (H-SWN) and applies a majority score-based strategy for classifying sentiment in the given Hindi document. These approaches contribute to a comprehensive understanding of sentiment analysis challenges and strategies tailored to the unique linguistic characteristics of Hindi.

The research on Sentiment Analysis of Hindi Review [8] introduces a method to enrich the coverage of Hindi SentiWordNet, thereby elevating the accuracy of sentiment classification. This approach also delves into examining the impact of negation and discourse rules on Hindi sentiment analysis, achieving an overall accuracy of 80.21%.

By amalgamating these approaches, our research endeavors to contribute to the progressing field of sentiment analysis for Hindi movie reviews, addressing the distinct challenges arising from the multilingual and culturally diverse nature of the content. The insights drawn from existing literature form a robust basis for our inventive approach to enhance sentiment analysis accuracy in the context of Hindi cinema.

III.PROPOSED METHODOLOGY

Our proposed methodology for sentiment analysis on Hindi movie reviews aims to bring about a transformative shift in the field by integrating advanced technologies such as resource-based semantic analysis, in-language semantic analysis, and machine translation-based semantic analysis. The initial phase involves the curation of a diverse dataset, comprising 1000 Hindi movie reviews, with contributions from IIT-Bombay's labeled dataset and manually collected reviews from Jagran.com. Through meticulous preprocessing, we standardize the dataset to ensure uniformity and eliminate noise, setting the foundation for a robust sentiment analysis framework.

We also explore the potential of machine translation-based analysis through the Googletrans API. We assess the system's performance using accuracy and F1 score metrics and integrate user feedback for ongoing improvement. Privacy measures are in place, and scalability is a key consideration for handling larger datasets and future growth. This approach ensures an effective and adaptable system for sentiment analysis in Hindi movie reviews.

Steps followed for Sentiment Analysis:

1. Data Collection and Preprocessing:

Curate a diverse dataset comprising 1000 Hindi movie reviews, including 250 labeled reviews from IIT-Bombay and 750 manually collected reviews from Jagran.com. Employ meticulous preprocessing techniques to clean and standardize the dataset, ensuring uniformity and eliminating noise.

2. Resource-Based Semantic Analysis:

Implement HindiSentiWordnet as a resource for sentiment analysis, leveraging its semantic information and sentiment scores. Develop algorithms to extract sentiment polarity and intensity from the reviews using the HindiSentiWordnet resource.

3. Language Semantic Analysis:

Utilize a variety of classifiers and feature extraction methods, including Bag of Words models with Term Frequency (TF) or Term Frequency-Inverse Document Frequency (TFIDF) values. Experiment with different classifiers such as Decision Trees, Support Vector Machines, and Naive Bayes to evaluate their effectiveness in capturing sentiment patterns in Hindi reviews

4. Machine Translation-Based Semantic Analysis:

Translate a subset of Hindi reviews into English using the Googletrans API to create a parallel dataset. Train classifiers on the translated English reviews and evaluate their performance on the original Hindi reviews to assess the feasibility and accuracy of the machine translation-based approach.

5. Evaluation and Feedback Mechanisms:

Implement rigorous evaluation metrics to assess the performance of each approach, considering factors such as accuracy and F1 score. Incorporate user feedback mechanisms for continuous improvement, allowing the system to adapt and enhance its accuracy over time.

6. Privacy and Security Considerations:

Implement robust privacy measures to ensure the confidentiality of user data and compliance with data protection regulations.
Integrate secure data transmission protocols to safeguard user information throughout the sentiment analysis process.

7. Scalability:

Design the system with scalability in mind, enabling it to handle larger datasets and accommodate future expansions seamlessly.
By combining these methodologies, our research aims to create a comprehensive and accurate sentiment analysis system for Hindi movie reviews, addressing the unique challenges posed by the multilingual and culturally rich nature of the content.

IV.IMPLEMENTATION AND RESULTS

The methodology of Resource-Based Semantic Analysis employing HindiSentiWordnet for sentiment classification in Hindi reviews involves a nuanced exploration of sentiment resources tailored to the intricacies of the Hindi language. HindiSentiWordnet, an extension of the well-established English SentiWordnet, serves as a foundational lexicon enriched with sentiment scores for Hindi words. This approach addresses the scarcity of domain-specific sentiment resources for Hindi, making it an invaluable asset in the sentiment analysis landscape.

The process begins with the compilation of a sentiment lexicon specific to Hindi, where each word is associated with its sentiment polarity and intensity. This lexicon acts as a reference guide, attributing sentiment scores to individual words and enabling the classification of reviews based on the cumulative sentiment orientation of their constituent words.

In a typical workflow, the Hindi text undergoes pre-processing, including tokenization and removal of stopwords, to enhance the accuracy of sentiment assignment. Each word in the pre-processed text is then matched with its corresponding entry in the HindiSentiWordnet lexicon. The sentiment scores of individual words are aggregated to determine the overall sentiment polarity of the text.

While the approach showcases effectiveness in leveraging existing sentiment resources for Hindi, it is not without challenges. The scarcity of domain-specific sentiment annotations for Hindi and potential discrepancies in sentiment nuances pose inherent limitations. Despite these challenges, Resource-Based Semantic Analysis shows an accuracy of 53.51% and an F-measure of 0.527

The In-Language Semantic Analysis approach is tailored to the specific nuances of Hindi, employing machine learning classifiers for sentiment analysis. The process begins with data collection and preprocessing, compiling positive and negative Hindi movie reviews. Feature generation relies on a foundational Bag of Words model, with options for Term Frequency (TF) or TFIDF representations. Decision Trees and Naive Bayes classifiers are trained on these features, and models undergo validation using techniques like k-fold cross-validation. Challenges addressed include the effective handling of negations and idiomatic expressions in Hindi. The possibility of hybrid approaches, combining predictions from multiple classifiers, is explored for enhanced accuracy.

Term Frequency (TF): TF serves as a fundamental measure, quantifying the frequency of each term within a given document. For sentiment analysis in Hindi reviews, TF becomes a valuable metric, reflecting the prevalence of specific words within individual critiques.

Term-Frequency-Inverse-Document-Frequency(TFIDF): Recognizing the limitations of TF, the In-Language Semantic Analysis approach incorporates TFIDF to provide a more comprehensive view. TFIDF extends beyond mere term frequency by factoring in the inverse document frequency, effectively weighing the importance of a term not just within a specific document but across the entire collection of reviews.

Machine Translation-Based Semantic Analysis is a pioneering approach in sentiment classification for Hindi reviews, harnessing the power of machine translation to bridge language gaps and enable sentiment analysis on a cross-linguistic level. This methodology extends the boundaries of sentiment analysis by training classifiers on English reviews and subsequently translating Hindi reviews into English for evaluation.

The workflow commences with the acquisition of parallel datasets comprising both Hindi and English reviews. The English dataset serves as the training ground for sentiment classifiers, allowing them to learn the intricacies of sentiment patterns in English text. Popular machine translation tools, such as the Googletrans API, facilitate the translation of Hindi reviews into English, ensuring linguistic nuances are preserved during the process.

The training phase involves the application of various classifiers, ranging from traditional machine learning models to deep learning architectures, on the English dataset. This step aims to equip the classifiers with the ability to discern sentiment orientations effectively. Once trained, these classifiers are employed to predict sentiments in the translated Hindi reviews.

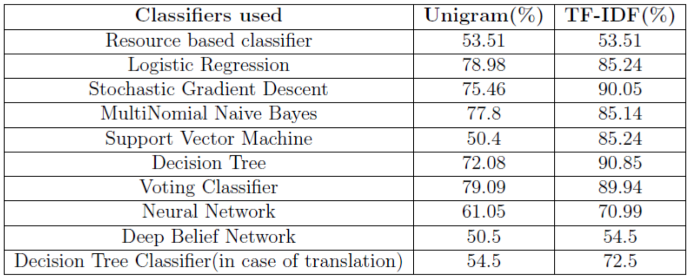
While the Machine Translation-Based approach presents a novel solution to overcome the scarcity of labeled sentiment data in Hindi, it introduces challenges related to translation accuracy and potential loss of context during language conversion. The choice of an optimal translation model and the impact of linguistic variations between Hindi and English further influence the method's efficacy.

DBN Neural Network for In-Language Classification: Introducing deep learning through Deep Belief Networks (DBN), this approach utilizes Unigram and TFIDF features for sentiment classification. DBN, a neural network architecture, captures intricate patterns in the reviews, showcasing the potential of deep learning in sentiment analysis. However, the complexity of DBN requires careful tuning and a substantial amount of training data. This approach offers a glimpse into the capabilities of deep learning but demands computational resources and expertise in neural network configurations. The DBN utilized in this study comprises multiple layers of latent variables, forming a hierarchical structure that allows the network to capture complex patterns in the data. TensorFlow serves as a cornerstone in the implementation of the DBN, providing a versatile and powerful framework for constructing, training, and optimizing deep learning models.

TensorFlow's extensibility allows researchers and practitioners to customize the DBN architecture, experiment with different configurations, and incorporate additional features as needed. This flexibility ensures adaptability to the unique requirements of sentiment analysis on Hindi movie reviews. TensorFlow is instrumental in preprocessing the input data, including tokenization, feature extraction, and numerical representation.

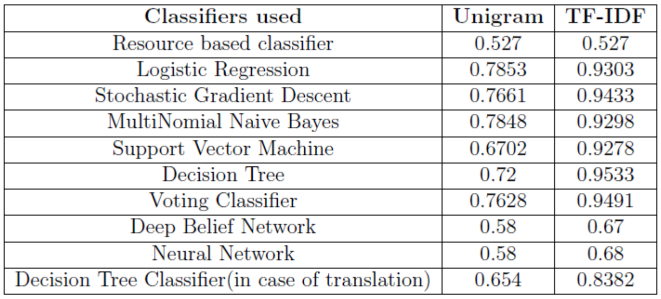
Results:

Accuracy:



Classifiers used	Unigram(%)	TF-IDF(%)
Resource based classifier	53.51	53.51
Logistic Regression	78.98	85.24
Stochastic Gradient Descent	75.46	90.05
MultiNomial Naive Bayes	77.8	85.14
Support Vector Machine	50.4	85.24
Decision Tree	72.08	90.85
Voting Classifier	79.09	89.94
Neural Network	61.05	70.99
Deep Belief Network	50.5	54.5
Decision Tree Classifier(in case of translation)	54.5	72.5

F-measure:



Classifiers used	Unigram	TF-IDF
Resource based classifier	0.527	0.527
Logistic Regression	0.7853	0.9303
Stochastic Gradient Descent	0.7661	0.9433
MultiNomial Naive Bayes	0.7848	0.9298
Support Vector Machine	0.6702	0.9278
Decision Tree	0.72	0.9533
Voting Classifier	0.7628	0.9491
Deep Belief Network	0.58	0.67
Neural Network	0.58	0.68
Decision Tree Classifier(in case of translation)	0.654	0.8382

V. CONCLUSION

This research project explores three distinct approaches for sentiment analysis on Hindi movie reviews, namely Resource-Based Semantic Analysis, In-Language Semantic Analysis, and Machine Translation-Based Semantic Analysis.

Resource-Based Semantic Analysis utilizes HindiSentiWordnet as a valuable lexical resource, mapping words to sentiment scores. This approach leverages the richness of the Hindi language, but it inherently relies on the availability and accuracy of lexical resources. In-Language Semantic Analysis employs various machine learning classifiers and deep learning techniques to directly analyze Hindi text. The use of classifiers and deep belief networks offers a data-driven approach, capturing intricate patterns and relationships within the reviews. Machine Translation-Based Semantic Analysis introduces an innovative solution by translating Hindi reviews into English and applying sentiment analysis in the translated domain.

Out of all the three approaches we can conclude that the second approach of In-Language classification performed using annotated corpus in the same language gave the best results. Also among the unigram and TF-IDF approach, TF-IDF proved to generate better results as compared to the other approach. It also shows us that the decision tree gives the best accuracy of 90.85% in case of TF-IDF representation and in case of unigram we have been able to achieve an accuracy of 79% through voting classifier.

Future Aspects:

The future trajectory of this project involves refining and enhancing each approach to achieve more accurate sentiment predictions. For Resource-Based Semantic Analysis, expanding and fine-tuning sentiment lexicons specific to the domain of Hindi movie reviews will be imperative. In-Language Semantic Analysis can benefit from a larger and more diverse labeled dataset for training, enabling the models to generalize better across different genres and styles of reviews. Exploring state-of-the-art natural language processing techniques and pre-trained language models may further elevate the accuracy of sentiment predictions.

Machine Translation-Based Semantic Analysis could see improvements by fine-tuning translation models specifically for movie reviews, preserving sentiment nuances during the translation process. Additionally, investigating the integration of contextual embedding and cross-lingual sentiment analysis models could enhance the overall performance.

Collaborations with linguists, domain experts, and the continual incorporation of user feedback will be crucial for refining the models and ensuring their applicability to a broader range of Hindi movie reviews. As the field of sentiment analysis evolves, embracing emerging technologies and methodologies will be essential for staying at the forefront of understanding sentiment.

REFERENCES

[1]Your Sentiment Precedes You: Using an author's historical tweets to predict sarcasm Anupam Khattri1 Aditya Joshi2,3,4 Pushpak Bhattacharyya2 Mark James Carman3 1 IIT Kharagpur, India, 2 IIT Bombay, India, 3Monash University, Australia

[2] Bakliwal, A., Arora, P., Varma, V.: Hindi Subjective Lexicon: A Lexical Resource for Hindi Polarity Classification (2012)

[3] Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2), 1–135 (2008).

[4] Fine Grained Human Evaluation for English-to-Chinese Machine Translation: A Case Study on Scientific Text
Ming Liu, He Zhang, Guanhao Wu

[5]Mukherjee, S., Bhattacharyya, P.: Sentiment Analysis in Twitter with Lightweight Discourse Analysis. In: Proceedings of the 24th International Conference on Computational Linguistics, COLING 2012 (2012)

[6] Apoorv Agarwal, Fadi Biadsy, and Kathleen R. Mckeown. 2009. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams.

[7] Joshi, A.R., Balamurali, P.: A Fall-Back Strategy For Sentiment Analysis In Hindi: A Case
Study. In: International Conference on Natural Language Processing, ICON (2010)

[8] Discourse Based Sentiment Analysis for Hindi Reviews
Namita Mittal, Basant Agarwal, Garvit Chouhan, Prateek Pareek, and Nitin Bania.