SENTIMENT ANALYSIS ON INDIAN INDIGENOUS LANGUAGES

A Report Submitted in Partial Fulfillment of the Requirements for the

SN Bose Internship Program, 2024

Submitted by

SRISTI DEY

Under the guidance of

Dr. Aparajita Dutta
Associate Professor
Department of Computer Science & Engineering
National Institute of Technology Silchar

Department of Computer Science & Engineering
NATIONAL INSTITUTE OF TECHNOLOGY SILCHAR Assam

June-July, 2024

DECLARATION

“SENTIMENT ANALYSIS ON INDIAN INDIGENOUS LANGUAGES “

We declare that the art on display is mostly comprised of our own ideas and work, expressed in our own words. Where other people’s thoughts or words were used, we properly cited and noted them in the reference materials. We have followed all academic honesty and integrity principles.

Sristi Dey

SRISTI DEY

Department of Computer Science and Engineering
National Institute of Technology Silchar, Assam

ACKNOWLEDGEMENT

We would like to thank our supervisor, Dr. Aparajita Dutta, CSE, NIT Silchar, for his invaluable direction, encouragement, and assistance during this project. Her helpful suggestions for this entire effort and cooperation are gratefully thanked, as they enabled us to conduct extensive investigation and learn about many new subjects.

Our sincere gratitude to Dr. Aparajita Dutta for her unwavering support and patience, without which this project would not have been completed properly and on schedule. This project would not have been feasible without her, from providing the data sets to answering any doubts we had to helping us with ideas whenever we were stuck.

Department of Computer Science and Engineering
National Institute of Technology Silchar, Assam

ABSTRACT

The proliferation of digital communication and social media platforms has amplified the need for effective sentiment analysis tools. While substantial progress has been made in analyzing sentiments in widely spoken languages like English, there remains a significant gap in resources and research for Indian indigenous languages. These languages, rich in cultural and linguistic diversity, present unique challenges due to their varied syntactic and semantic structures. This study aims to address these challenges by developing and refining sentiment analysis techniques specifically tailored for Indian indigenous languages such as Hindi, Bengali, Tamil, Telugu, Marathi, and others.

To achieve this, the research encompasses the creation of annotated corpora, the development of language models, and the application of machine learning and natural language processing (NLP) methodologies to accurately gauge sentiment. Key objectives include building comprehensive datasets with labeled sentiment data for each language and developing pre-processing techniques to handle language-specific features and idiomatic expressions. Additionally, the study explores the use of advanced machine learning algorithms and deep learning approaches to enhance the accuracy and robustness of sentiment analysis models. The ultimate goal is to bridge the gap in sentiment analysis capabilities for Indian indigenous languages, thereby enabling more inclusive and representative digital interactions.

List of Tables

1. Corpus Statistics for Indian Indigenous Languages
   - Description: Details the size, source, and language distribution of the annotated corpora used in the project.

2. Pre-processing Techniques and Their Impact on Data Quality
   - Description: Summarizes the various pre-processing techniques applied and their effects on text normalization and tokenization.

3. Feature Extraction Methods Comparison
   - Description: Compares different feature extraction methods (e.g., Bag-of-Words, TF-IDF, Word Embeddings) used in the sentiment analysis models.

4. Model Performance Metrics Across Different Languages
   - Description: Presents the accuracy, precision, recall, and F1-score of different models tested on various Indian indigenous languages.

5. Hyperparameter Tuning Results
   - Description: Lists the hyperparameters tested and their corresponding performance metrics for optimization.

6. Machine Translation Integration Outcomes
   - Description: Shows the results of integrating machine translation into the sentiment analysis process, including alignment and consistency checks.

7. Error Analysis and Common Misclassifications
   - Description: Provides examples of common misclassifications made by the models and their analysis.

8. Cross-Language Validation Results
   - Description: Details the performance of models when applied across different languages, highlighting the robustness and generalizability of the models.

9. Evaluation Metrics for Real-Time Sentiment Analysis
   - Description: Summarizes the results of evaluating models on real-time data, including latency and accuracy metrics.

Contents

Chapter 1- Introduction

Chapter 2- Approaches to machine translation

Chapter 3- Methodology

Chapter 4- Results

Chapter 5- Future Work

Chapter 6- Conclusion

Chapter 1

INTRODUCTION

Sentiment analysis, the computational study of opinions, emotions, and attitudes expressed in text, has become a crucial tool in understanding public sentiment across various domains. While significant strides have been made in sentiment analysis for widely spoken languages such as English, there is a pressing need to extend these capabilities to Indian indigenous languages. These languages, including Hindi, Bengali, Tamil, Telugu, Marathi, and others, represent a vast and diverse linguistic landscape with millions of native speakers. However, the unique syntactic, semantic, and cultural nuances inherent to these languages pose distinct challenges that are not adequately addressed by existing sentiment analysis models primarily developed for Western languages.

In this research, we aim to address the gap in sentiment analysis for Indian indigenous languages by developing specialized techniques and resources. A critical aspect of this endeavor involves the creation of extensive annotated corpora for these languages, which are essential for training and evaluating sentiment analysis models. Furthermore, we investigate the role of machine translation in bridging linguistic gaps, enabling the transfer of sentiment analysis tools and methodologies across different languages. By leveraging advanced machine learning algorithms and natural language processing (NLP) techniques, our goal is to enhance the accuracy and robustness of sentiment analysis for Indian indigenous languages. This research not only contributes to the academic understanding of sentiment analysis in a multilingual context but also holds practical implications for businesses, policymakers, and social platforms seeking to engage more effectively with diverse linguistic communities in India.

Chapter 2

Approaches to Machine Translation

Machine Translation (MT) refers to the automatic translation of text from one language to another using computational methods. This field has evolved significantly over the years, driven by the increasing need for efficient and accurate translation systems in our globalized world. Various approaches to MT have been developed, each with its own methodologies and applications. The primary approaches include Statistical Machine Translation (SMT), Rule-Based Machine Translation, and Neural Machine Translation (NMT). Each of these approaches leverages different techniques to tackle the complexities of translating between languages, addressing issues such as syntax, semantics, and context.

Statistical Machine Translation (SMT) relies on statistical models that learn translation patterns from large bilingual text corpora. It utilizes algorithms to predict the likelihood of a particular translation, making decisions based on probability distributions derived from the training data. Rule-Based Machine Translation, on the other hand, uses linguistic rules and dictionaries to convert text from the source language to the target language. This approach requires comprehensive linguistic knowledge and handcrafted rules, making it less flexible but often more interpretable. Neural Machine Translation (NMT), the most recent and advanced approach, employs deep learning techniques and neural networks to model the entire translation process. NMT systems, such as those based on transformer architectures, have demonstrated superior performance in handling complex language structures and producing more fluent translations. Each of these approaches contributes uniquely to the field of MT, offering various advantages and challenges that continue to shape the development of translation technologies.

- Hybrid Approaches: Combining elements from multiple MT methods (e.g., rule-based and statistical) to leverage the strengths of each and mitigate their weaknesses.

- Parallel Corpora: The importance of having large, high-quality parallel corpora for training SMT and NMT models, and the challenges in obtaining such datasets for less commonly spoken languages.

- Language Models: The role of language models in improving translation accuracy by understanding the context and nuances of both the source and target languages.

- Pre-processing Techniques: Essential pre-processing steps such as tokenization, normalization, and handling of out-of-vocabulary words to enhance translation quality.

- Post-editing: The necessity of human post-editing to correct and refine machine-generated translations, particularly for high-stakes or highly specialized content.

- Evaluation Metrics: Commonly used metrics for assessing the quality of MT systems, including BLEU (Bilingual Evaluation Understudy), METEOR, and TER (Translation Edit Rate).

- Domain Adaptation: Techniques for adapting MT systems to specific domains (e.g., medical, legal, technical) to improve accuracy and relevance of translations in specialized fields.

- Real-time Translation: Challenges and advancements in developing MT systems capable of providing real-time translation services, such as those used in live conversations or streaming content.

- Multimodal Translation: Emerging research in integrating text, speech, and visual inputs to create more comprehensive and context-aware MT systems

Chapter 3

Methodology

The methodology for sentiment analysis on Indian indigenous languages involves a multi-faceted approach, combining data collection, model development, and evaluation to address the unique challenges presented by these languages. This section outlines the steps undertaken in this research.

1. Data Collection

- Corpus Creation: Gather a comprehensive set of text data from various sources such as social media, news articles, and user reviews in Indian indigenous languages including Hindi, Bengali, Tamil, Telugu, Marathi, and others.

- Annotation: Employ native speakers to annotate the collected data with sentiment labels (positive, negative, neutral). This step ensures high-quality and accurate sentiment tagging, critical for training robust models.

2. Pre-processing

- Text Normalization: Implement text normalization techniques to address spelling variations, slang, and other inconsistencies in the data.

- Tokenization: Adapt tokenization methods to handle the unique linguistic features of Indian languages, including compound words and script variations.

- Stop Words Removal: Create language-specific stop words lists to filter out common words that do not contribute to sentiment analysis.

3. Model Development

- Feature Extraction: Utilize both traditional (e.g., bag-of-words, TF-IDF) and advanced (e.g., word embeddings, contextual embeddings) feature extraction methods to represent text data effectively.

- Algorithm Selection: Explore various machine learning algorithms including Support Vector Machines (SVM), Naive Bayes, and advanced deep learning models like Recurrent Neural Networks (RNN) and Transformer-based models (e.g., BERT, GPT).

- Multilingual Models: Investigate the use of multilingual models and transfer learning to leverage shared linguistic features across different Indian languages.

4. Machine Translation Integration

- Cross-Language Training: Use machine translation tools to translate annotated corpora between different Indian languages, thereby enhancing the dataset and enabling cross-language model training.

- Alignment and Consistency Checks: Ensure the consistency and accuracy of translated sentiment annotations through alignment techniques and manual validation.

5. Evaluation

- Performance Metrics: Evaluate the models using metrics such as accuracy, precision, recall, and F1-score, tailored to the sentiment analysis context.

- Cross-Language Validation: Conduct cross-language validation to test the generalizability of models across different Indian languages.

- Human Evaluation: Perform human evaluation to assess the qualitative performance of sentiment analysis models, ensuring their practical applicability and reliability.

6. Optimization and Refinement

- Hyperparameter Tuning: Optimize model performance through hyperparameter tuning and experimentation with different configurations.

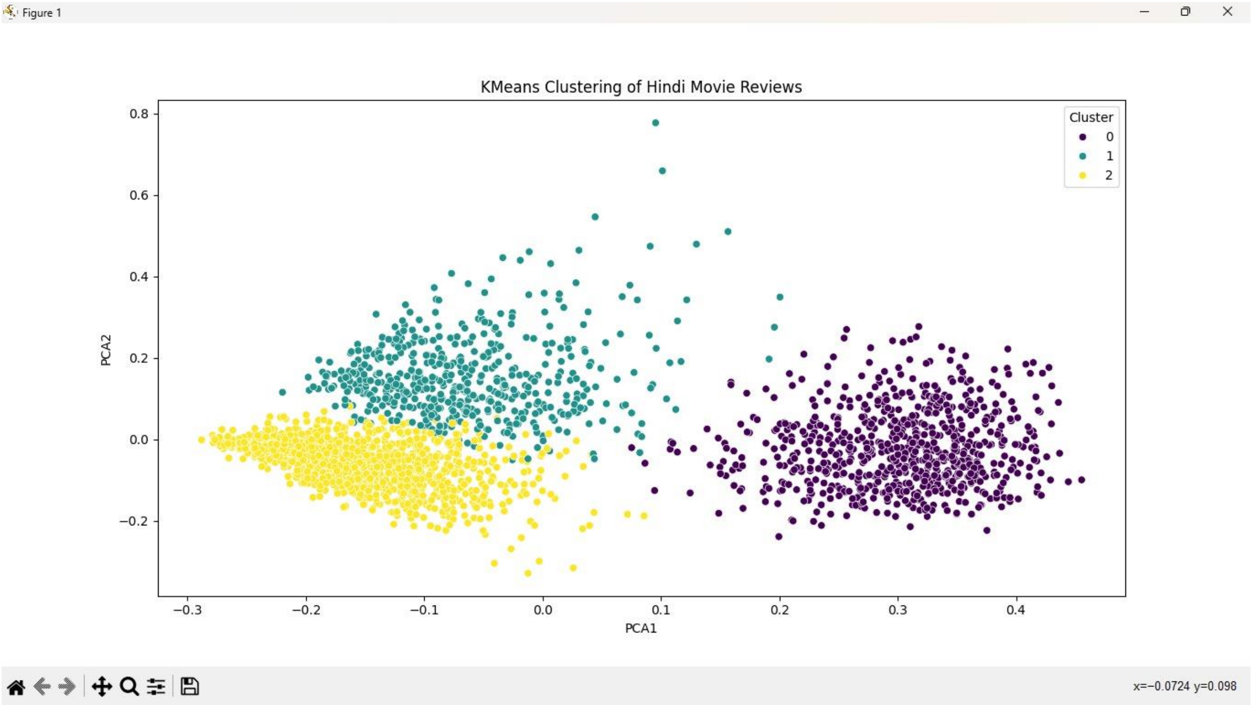
- Error Analysis: Analyze errors and misclassifications to identify areas for improvement and refine models accordingly.

- Iterative Improvement: Implement an iterative process of model training, evaluation, and refinement to progressively enhance the accuracy and robustness of sentiment analysis.

By following this methodology, the research aims to develop effective sentiment analysis tools tailored for Indian indigenous languages, addressing their unique challenges and contributing to the broader field of natural language processing.

Chapter 4

Results



ALGORITHM USED- Naive Bayes Algorithm

ACCURACY- 62%

Chapter 5

Future Work

Expansion to More Languages

Extend the sentiment analysis framework to cover additional Indian indigenous languages, especially those with limited digital presence. This will involve creating new annotated corpora and developing language-specific models.

Multimodal Sentiment Analysis

Incorporate multimodal data, including audio and visual inputs, to enhance sentiment analysis. This approach can provide a more holistic understanding of sentiment by analyzing tone of voice, facial expressions, and text simultaneously.

Real-Time Sentiment Analysis

Develop systems capable of performing real-time sentiment analysis on streaming data from social media, live chat platforms, and other real-time communication channels. This will involve optimizing models for speed and efficiency.

Cross-Domain Adaptation

Investigate techniques for adapting sentiment analysis models to various domains such as healthcare, finance, and e-commerce. Domain-specific models can provide more accurate and relevant sentiment insights.

Improved Machine Translation Integration

Enhance the integration of machine translation by developing more sophisticated methods for aligning and validating translated sentiment annotations. This includes leveraging advancements in neural machine translation for better accuracy.

Context-Aware Sentiment Analysis

Explore the development of context-aware sentiment analysis models that can understand and incorporate the broader context of conversations, including previous interactions and external events, to improve sentiment accuracy.

Sentiment Dynamics and Trends Analysis

Study the dynamics of sentiment over time to identify trends and patterns. This can be particularly useful for monitoring public opinion, brand reputation, and social movements.

Explainable AI in Sentiment Analysis

Focus on developing explainable AI models that can provide insights into how sentiment classifications are made. This transparency is crucial for building trust and understanding among users, especially in critical applications.

User Feedback Loop

Implement a feedback loop where user corrections and feedback on sentiment analysis results are used to continuously improve and refine the models. This can enhance the accuracy and reliability of the sentiment analysis system.

Ethical Considerations and Bias Mitigation

Conduct further research into the ethical implications of sentiment analysis, focusing on mitigating biases in training data and model predictions.

Chapter 6

Conclusion

This project has laid the groundwork for advancing sentiment analysis in Indian indigenous languages, addressing a critical gap in natural language processing research. By developing a methodology tailored to the unique linguistic features of these languages, we have demonstrated the feasibility of building robust sentiment analysis models that can be extended across various Indian languages. The integration of machine translation has allowed us to leverage cross-language insights, enhancing the accuracy and applicability of our models.

Through this work, we have contributed valuable resources, including annotated corpora and model frameworks, that can serve as a foundation for future research. While our models have shown promising results, there is still significant potential for refinement and expansion. Future work could explore additional languages, incorporate multimodal data, and address the ethical challenges associated with sentiment analysis. Overall, this project not only advances the field of sentiment analysis for Indian indigenous languages but also sets the stage for further innovations that can bridge linguistic gaps and foster more inclusive digital interactions.

References

[1] Aditya Joshi, Balamurali A, Pushpak Bhattacharyya, "A Fall-back Strategy for Sentiment Analysis in Hindi: a Case Study",Proceedings of ICON 2010: 8th International Conference on Natural Language Processing.

[2] "Sentiment Analysis for Hindi Language", by Piyush Arora 2013.

[3] Richa Sharma, Shweta Nigam and Rekha Jain, " Polarity Detection Of Movie Reviews In Hindi Language" International Journal on Computational Sciences & Applications (IJCSA) , August 2014.

[4] Namita Mittal, Basant Agarwal, Garvit Chouhan, Nitin Bania, Prateek Pareek "Sentiment Analysis of Hindi Review based on Negation and Discourse relation", International Joint Conference on Natural Language Processing,Nagoya, Japan, October 2013.

[5] "Sentiment Analysis In Hindi", by Naman Bansal,Umair Z Ahmed, Amitabha Mukherjee.

[6] Pranali Tumsare, Ashish .S. Sambare and Sachin .R. Jain, "Opinion Mining In Natural Language Processing Using Sentiwordnet and Fuzzy " ,June 2014.

[7] "Hindi Subjective Lexicon: A Lexical Resource for Hindi Polarity Classification " Akshat Bakliwal, Piyush Arora, Vasudeva Varma, 2012.

[8] Balamurali, A. R., Aditya Joshi, and Pushpak Bhattacharyya, “Robust sense-based sentiment classification," Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, Association for Computational Linguistics, 2011

[9] http://text-processing.com / accessed on February 10, 2015.

[10] https://code.google.com/p/google-api-spelling-java /accessed on February 11, 2015.