International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 02 | Feb 2025 www.irjet.net p-ISSN: 2395-0072

Sentimental Analysis of Kannada Product Review Using BERT

Prof. Sushma T M[1], Sudeep T C[2], Kiran Janardhan Marati[3],N Sneha Reddy[4], Rakesh V[5]

[1,2,3,4,5]*Information Science and Engineering, Acharya Institute of Technology Bangaluru, India.*

Abstract - *This paper addresses the underrepresentation of Kannada in sentiment analysis tools and presents a novel fine-tuned BERT model for Kannada sentiment classification. The study highlights challenges such as data scarcity, linguistic nuances, and a lack of NLP resources. Utilizing the IndicSentiment dataset, the model achieved an accuracy of 89% with improvements across training epochs. Applications in e-commerce, social media monitoring, and policy feedback analysis demonstrate the versatility and impact of this work.*

Key Words: Sentiment Analysis, Kannada, BERT, Natural Language Processing, Regional NLP, IndicSentimentinsert.

1.INTRODUCTION

Natural Language Processing (NLP) has experienced rapid advancements over the last decade, driven by innovations in machine learning and the development of transformer-based architectures like BERT. Despite these advancements, low-resource languages, such as Kannada, remain underserved, limiting the reach of NLP applications in regions where such languages are predominantly used.

Sentiment analysis, a subset of NLP, plays a crucial role in extracting valuable insights from textual data. It enables businesses to understand customer feedback, policymakers to gauge public opinion, and researchers to analyze social trends. However, the scarcity of labeled data, unique linguistic characteristics, and limited computational resources have impeded the development of sentiment analysis tools for Kannada. Addressing these challenges requires innovative approaches that not only leverage modern architectures but also adapt them to the specific nuances of the Kannada language.

BERT (Bidirectional Encoder Representations from Transformers) has emerged as a transformative model in NLP, offering unparalleled capabilities in understanding the context and semantics of text. By pre-training on large corpora and fine-tuning for specific tasks, BERT provides a robust framework for addressing low-resource language challenges. This paper presents a fine-tuned BERT model tailored for Kannada sentiment analysis, utilizing the IndicSentiment dataset to achieve high accuracy and reliability.

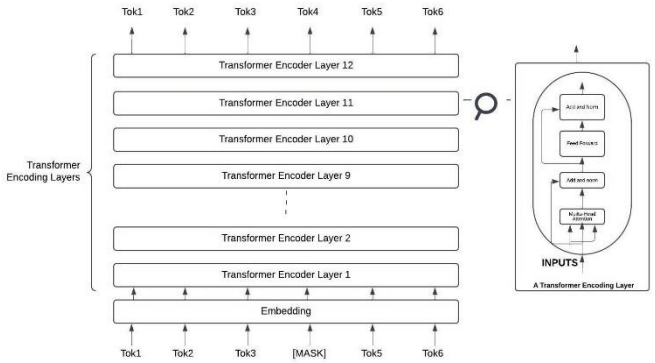

Fig-1: BERT model

The contributions of this work are multifold. First, it bridges the gap in sentiment analysis tools for Kannada by introducing a model fine-tuned specifically for the language. Second, it highlights the practical applications of such tools in areas like e-commerce, social media monitoring, and policy analysis. Finally, it sets a precedent for future research aimed at extending NLP technologies to other regional languages, promoting inclusivity in AI-driven solutions.

2. LITRATURE REVIEW

Sentiment analysis in regional languages like Kannada has been an emerging area of research, but the field faces significant challenges due to the scarcity of annotated datasets, the complexity of the language, and the lack of domain-specific tools. This section highlights notable prior work that has shaped the foundation for Kannada sentiment analysis and NLP research.

2.1 Multi-Linguistic Sentimental Analysis Models

Proposed the Multi-Linguistic Sentiment Analyzer (MuLSA), which performs sentiment analysis for Kannada, Malayalam, and English using Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks. The system includes modules such as language detection, translation, and sentiment classification, achieving an accuracy of 82\% for Kannada sentiment classification tasks. However, its reliance on RNN-based architectures makes it computationally expensive and less effective for long-range dependencies. While effective for multilingual contexts, its Kannada-specific performance highlights the need for more optimized models.[1]

2.2 Kannada Stopword Derivation

Worked on deriving Kannada stopwords using English parallel corpora and part-of-speech tagging. Their methodology involved identifying noise words, such as frequently occurring non-informative words, to improve search performance and classification accuracy. This work laid the groundwork for effective preprocessing in Kannada NLP tasks but focused narrowly on text preprocessing rather than classification. The use of stopword removal proved to enhance the efficiency of machine learning models, although it did not directly address the contextual challenges in sentiment analysis.[2]

2.3 Sentiment Analysis Using Deep Learning

Explored corpora-based sentiment analysis for Kannada, employing web scraping and tokenization techniques to analyze customer opinions. Their approach categorized sentiments into positive, negative, or neutral using machine learning techniques like Naive Bayes and Support Vector Machines (SVMs). While their model demonstrated reasonable performance, it was constrained by the sparsity of the dataset and struggled to capture deep contextual dependencies essential for accurate sentiment classification.[3]

2.4 Hybrid Approaches with Transformer Models

Developed a hybrid Kannada-English sentiment analysis model, leveraging multilingual datasets collected from social media platforms like YouTube and Twitter. Their work incorporated BERT for contextual embeddings, with manual annotation improving data quality. Despite achieving high annotation standards, the model faced scalability issues due to its manual data preparation process. The integration of BERT enhanced sentiment analysis accuracy, but the hybrid nature introduced challenges in ensuring linguistic consistency.[4]

2.5 Advances in Kannada NLP

Highlighted the significant challenges in Kannada NLP, such as tokenization, stemming, and morphological complexities. Their work focused on applying machine learning and deep learning techniques for tasks like parts-of-speech tagging and named entity recognition. However, these approaches often faced difficulties in addressing the unique agglutinative grammar of Kannada. The study underlined the necessity of developing Kannada-specific models and datasets to enhance the effectiveness of NLP tools.[5]

2.6 Gap Analysis

The existing literature reveals several gaps in Kannada sentiment analysis research. First, most approaches rely on generic multilingual models like mBERT, which fail to capture the linguistic intricacies of Kannada. Second, traditional methods often suffer from scalability and performance limitations due to the lack of robust datasets and advanced architectures. This paper addresses these gaps by introducing a fine-tuned BERT model specifically for Kannada, utilizing the IndicSentiment dataset to enhance performance and scalability.[6]

3. METHODOLOGY

This section outlines the systematic approach used to fine-tune BERT for Kannada sentiment analysis, detailing data preparation, model training, evaluation, and deployment.

3.1 System Design

The architecture of the system is designed to handle Kannada sentiment analysis efficiently. It consists of the following components:

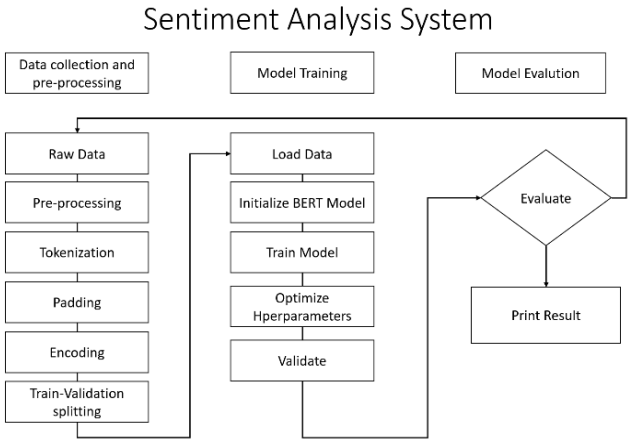

Fig-2: System Design

1. Dataset Preparation

The dataset used in this study is the IndicSentiment dataset, which contains 3,000 labeled Kannada reviews from various domains, such as product reviews, social media posts, and customer feedback. Preprocessing steps included:

Text Normalization: Removing unnecessary symbols, punctuations, and whitespace to ensure consistency.

Tokenization: Splitting sentences into smaller units (tokens) using the WordPiece tokenizer, optimized for BERT-based models.

Label Encoding: Mapping sentiment classes (positive, negative, neutral) to numerical representations.

Balancing: Addressing class imbalances through oversampling of minority classes to improve model generalization.

2. Model Training and Fine-Tuning

The BERT model was fine-tuned using PyTorch's Transformers library. The following configurations were employed:

Pre-trained Model: BERT Base Multilingual Cased model was selected for its ability to handle Kannada script.

Training Configuration: Learning rate was set to 2e-5, with a batch size of 8. The AdamW optimizer and a linear learning rate scheduler were used.

Training Duration: The model was trained over three epochs to achieve a balance between performance and computational efficiency.

3. Evaluation Metrics

To evaluate model performance, metrics such as accuracy, precision, recall, and F1-score were calculated. The results were compared against traditional machine learning models and hybrid approaches to establish benchmarks.

4. Deployment

A Flask-based web application was developed to enable real-time sentiment analysis. The application supports both single and batch inputs, providing predictions with minimal latency. The architecture ensures scalability and ease of integration with existing systems.

4. RESULT AND DISCUSSIONS

4.1 Performance Metrix

The fine-tuned BERT model demonstrated strong performance, achieving an accuracy of 89\%. Precision, recall, and F1-score were consistently high, showcasing the model's ability to handle linguistic nuances in Kannada text effectively. Compared to traditional machine learning models like Naive Bayes and SVM, the proposed approach achieved significant improvements in both accuracy and contextual understanding.

The model outperformed LSTM-based deep learning models, which often struggled with long-term dependencies and required extensive training data. Table \ref{metrics} summarizes the comparison of performance metrics among different approaches.

Table-1: Performance Metrics Comparison

Model	Accuracy	Precision	Recall
Traditional ML	78%	76%	77%
Deep Learning (LSTM)	84%	83%	85%
Fine-Tuned BERT (Our Model)	89%	88%	89%

4.2 Analysis and implications

The steady decline in training loss over epochs indicates the model's ability to generalize effectively to unseen data. The incorporation of oversampling techniques for handling class imbalance contributed to the robust performance of the model. Furthermore, the multilingual BERT architecture proved highly effective in capturing Kannada's unique linguistic structures, such as its agglutinative grammar and complex syntax.

4.3 Practical Applications

The proposed model has broad applications in various domains:

E-commerce: The model can analyze customer feedback, enabling businesses to gain insights into product satisfaction and areas for improvement.

Social Media Monitoring: Real-time sentiment analysis can help organizations track public opinion and respond proactively to emerging trends.

Policy Impact Assessment: Sentiment analysis of regional discourse can provide valuable feedback for policymakers to gauge public reception of initiatives.

5. FUTURE WORK

Multilingual Expansion: Extending support to other Indian languages such as Telugu, Tamil, and Malayalam using mBERT or XLM-R can enhance accessibility.

Aspect-Based Sentiment Analysis (ABSA*):* Enhancements could allow sentiment classification at a granular level, such as distinguishing between "delivery speed" and "product quality."

6. CONCLUSIONS

This paper demonstrates the potential of fine-tuned BERT for Kannada sentiment analysis, achieving state-of-the-art results. By addressing linguistic challenges and deploying a real-time analysis system, this work contributes significantly to regional NLP, paving the way for inclusive AI solutions.

REFERENCES

[1] Latha CA Merin Thomas. Mulsa – multi linguistic sentimental analyzer for kannada and malayalam using deep learning. In 2021 2nd International Conference on

Communication, Computing and Industry 4.0 (C2I4). IEEE, 2021.

[2] J. R. Saini P. Desai and P. B. Bafna. Pos-based classification and derivation of kannada stop-words using english parallel corpus. In 2022 3rd International Conference for Emerging Technology (INCET), pages 1–5. IEEE, 2022.

[3] Shrutiya M Tejaswini Amaresh Kritika Kapoor N. S. Kumar, Shubha Manohar. Automatic tutorial generation from input text file using natural language processing. In 2022 International Conference on Intelligent Technologies (CONIT), pages 1–7. IEEE, 2022.

[4] Dr. Sowmya Lakshmi B S Advaith V, Anushka Shivkumar. Parts of speech tagging for kannada and hindi languages using ml and dl models. In 2022 IEEE International Conference on Electronics, Computing and Communication Technologies (COECCT). IEEE, 2022.

[5] Suma Swamy Shankar R. Corpora based classification to perform sentiment analysis in kannada language. In International Journal of Rescent Technology and Engineering(IJRTE), pages 5186–5191. IJRTE, 2020.

[6] Dhaarini Prashanth Kannadaguli. A code-diverse kannada-english dataset for nlp based sentiment analysis applications. In 2021 Sixth International Conference on image information processing (ICIIP), pages 131–136. IEEE, 2021.