IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

A Comparative Analysis of Bilingual Sentiment Classification: Integrating English and Marathi Social Media Streams

Vedika Toke[1] and Anvisha Sawant[2]
MCA Department
Navinchandra Mehta Institute of Technology and Development, Dadar, India
anvishaanandsawant@gmail.com, vedikatoke1211@gmail.com

Abstract: *There is an urgent need for sentiment analysis since there is a need for sentiment from the information available on the Web as text data. It is now well understood by the social networking community that two or more languages can be used in a single text file. This research proposes to solve this issue with the help of modeling of sentiment analysis using English and Marathi text. There are also previous methods used to do this where the text file undergoes pre-processing, then feature extraction using TF-IDF, and finally classification using Naïve Bayes, Logistic Regression, and Support vector machine with 89% accuracy.*

Keywords: Bilingual Sentiment Analysis, NLP, TF-IDF, Social Media Analytics, Marathi Linguistics, Machine Learning.

I. INTRODUCTION

Natural Language Processing refers to a subject that combines intelligence and the study of languages. Natural Language Processing makes it possible for the computer to understand the text written in languages that people use every day. Natural Language Processing also looks at things like sentiment analysis. Sentiment analysis is a way to automatically figure out if a piece of text is positive, negative or neutral.

Natural Language Processing and sentiment analysis are very useful. However the methods we have now for sentiment analysis are really good when we are looking at texts that are written in English. When we look at texts in languages the results are not very accurate.

In a country like India people speak different languages, including English and Marathi. When people in India use the internet they often write in English or Marathi to express their feelings. Natural Language Processing is important, for understanding what people mean when they write things on the internet.

II. PROBLEM STATEMENT

The monolingual nature of the algorithm employed in most sentiment analysis machine learning models presents one of the greatest challenges to present-day sentiment analysis. It may be difficult for even the highly efficient models to cope with the nuances of language and its vocabulary in case of analysis of texts in different languages. The absence of a common processing software for languages like Marathi makes the situation even more complicated.

III. OBJECTIVES

- To address these challenges, the following aspects will be central to our research:
- Building a Sentiment Analysis Pipeline that can process both English and Marathi language text with same time.
- Improving quality of multilingual data by implementing efficient noise removal approaches.

DOI: 10.48175/IJARSCT-37261

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

- Utilizing the TF-IDF approach to transform unstructured data into machine-readable form.
- Evaluating the performance of different classifiers such as Naïve Bayes, Logistic Regression, and SVMs.
- Finding out the best classifier to implement in real-life applications within social media platforms.

IV. LITERATURE REVIEW

In many instances, the academic literature review concerning sentiment analysis highlights the use of Naïve Bayes and Logistic Regression algorithms for their efficiency and effectiveness with regard to textual data sets. However, it appears that the current body of literature on the topic focuses on English-language content predominantly. The deep learning-based method, BERT, which would prove beneficial for multi-language sentiment analysis, consumes vast computational resources.

V. DATASET DESCRIPTION

The current work involves an exhaustive analysis based on a selected set of over 700 social media tweets. Due to the paucity of quality Marathi data for sentiment analysis, we have manually annotated and augmented the Marathi corpus to ensure the adequate representation is made in the multilingual dataset. This results in the balancing of the dataset and reduces any potential bias that might exist in the model.

VI. METHODOLOGY

We followed a defined set of steps to convert unstructured social media conversations into structured information. The process was systematic and orderly. This ensures that each word gets its proper weight before machine learning algorithms get involved with classification.

Fig. 1: Proposed System Architecture for Multilingual Sentiment Analysis

This ensures that the each of this word gets its proper weight before machine learning algorithms get involved with classification.

VII. IMPLEMENTATION

A. Data Preprocessing: Setting the Stage

The need for preprocessing is more pronounced in a multilingual environment handled simultaneously. An effective data pipeline was created that handles diverse scripts like Devanagari and Latin, strips away redundant content, and normalizes the data. With respect to the highly inflected nature of Marathi language, preprocessing allows the focus to be narrowed to relevant sentimental words.

B. Feature Extraction with TF-IDF

TF-IDF, standing for Term Frequency–Inverse Document Frequency, is used as a link between natural language and computational approach. This statistic method measures the significance of a particular word in a document compared to other documents in a whole corpus. This method helps get rid of unnecessary terms and focus on specific emotional keywords in both English and Marathi.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

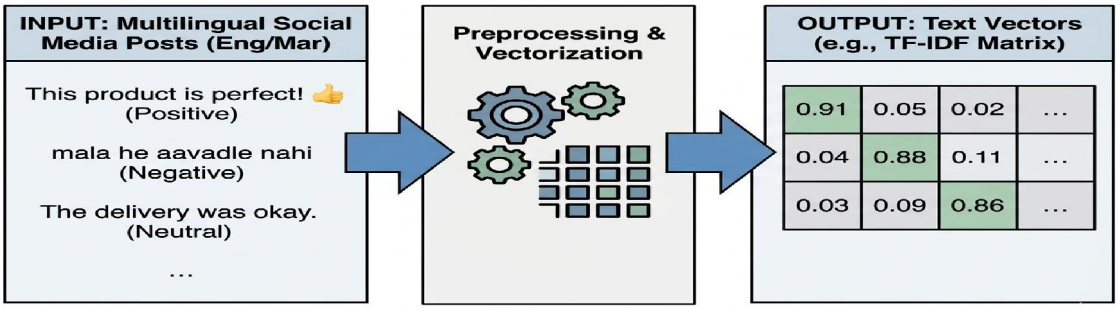

Fig. 2: Conceptual breakdown of input text transforming into high-dimensional TF-IDF vectors

C. The Classification Models

The study applied the following algorithms to establish the appropriate algorithm for the data:

Naive Bayes algorithm: This is a quick and accurate probabilistic approach that works well in simple text classifications.

Logistic Regression algorithm: Logistic Regression is a supervised machine learning classification algorithm used to predict sentiment categories such as positive, negative, and neutral.

Support Vector Machine (SVM) algorithm: This is a classifier that is efficient in separating data using hyperplanes.

VIII. RESULTS AND ANALYSIS

The models were evaluated based on how accurately they predicted the sentiment labels. The results are as follows:

Model	Accuracy (%)
Naive Bayes	82%
Logistic Regression	87%
Support Vector Machine (SVM)	**89%**

Table I: Classification Model Accuracy Comparison

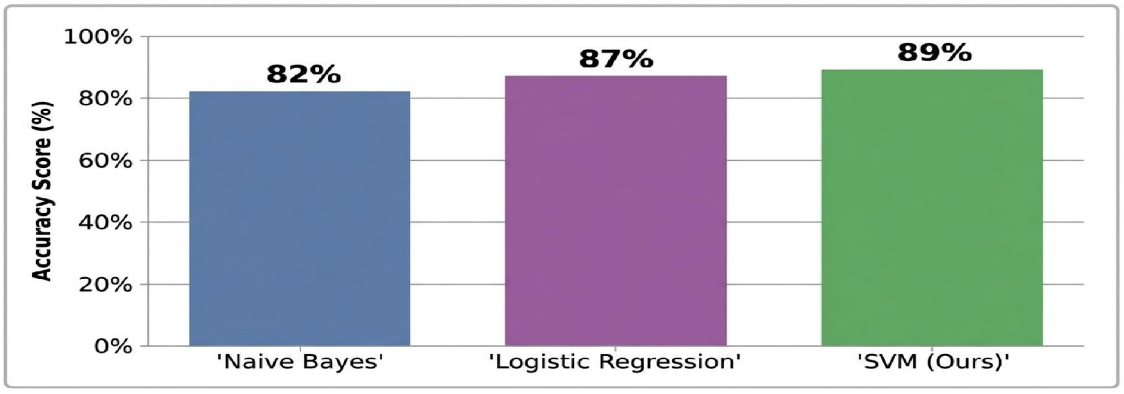

Fig. 3: Comparative Accuracy Analysis of Three Sentiment Classifiers

This visual comparison replaces Table 1, showing SVM (89%) as the top-performing model

Discussion: The reason why SVM was the best performer is because of its ability to handle the difficult nature of the high dimensional space formed by TF-IDF. Naive Bayes performed well as a benchmark, although it had the

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

disadvantage of having relatively low accuracy because of its simple model to deal with sentences using multiple languages. Compared to Naive Bayes, the logistic regression did slightly better; however, SVM had the upper hand.

IX. CONCLUSION

This study demonstrates that sentiment analysis using machine learning algorithms is an effective approach for identifying and classifying opinions in textual data on English and Marathi texts is indeed possible and very accurate. With the help of proper pre-processing and SVM classifiers, an accurate result was obtained. In the end of the research paper, it can be said that having a system that supports regional languages makes more sense.

X. FUTURE SCOPE

In the future, some possible directions for future work could be taken with this research are:
Extended language capabilities: Consider other local languages like Hindi and Tamil.
Enhanced technology: Try to leverage neural networks such as BERT for capturing deeper context.
Live monitoring: Create an application to analyze trending topics in real-time.
Larger datasets: Use more data to enhance the performance of the model.

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, 2019.
[2] A. Joshi, V. Prabhu, P. Goyal, and P. Bhattacharyya, "SentimixHindi: A Sentiment Analysis Dataset for Hindi," Proceedings of the 8th Workshop on South and Southeast Asian Natural Language Processing, pp. 45–52, 2021.
[3] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," Proceedings of the European Conference on Machine Learning (ECML), pp. 137–142, 1998.
[4] A. McCallum and K. Nigam, "A Comparison of Event Models for Naive Bayes Text Classification," AAAI Workshop on Learning for Text Categorization, pp. 41–48, 1998.
[5] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008.
[6] F. Sebastiani, "Machine Learning in Automated Text Categorization," ACM Computing Surveys, vol. 34, no. 1, pp. 1–47, 2002.
[7] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs Up? Sentiment Classification Using Machine Learning Techniques," Proceedings of EMNLP, pp. 79–86, 2002.
[8] J. Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries," Proceedings of the First Instructional Conference on Machine Learning, 2003.
[9] A. K. Uysal and S. Gunal, "The Impact of Preprocessing on Text Classification," Information Processing & Management, vol. 50, no. 1, pp. 104–112, 2014.
[10] P. Bhattacharyya, Machine Translation and I