International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Conference Proceedings: Volume: 05 ACME – 2026 DOI: 10.55041/ISJEM.ACME104
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Sentiment Analysis System for Hindi–English Text using Machine Learning

Ms . K Gangothri [1]
Assistant Professor, Department of AI&DS
Annamacharya Institute of Technology and Sciences, Tirupati – 517520, A.P.
gangakonkala@gmail.com

C Rajasekhar[2]
Department of AI&DS
Annamacharya Institute of Technology and Sciences, Tirupati – 517520, A.P.
ganar3215@gmail.com

A Ganesh [3]
Department of AI&DS
Annamacharya Institute of Technology and Sciences, Tirupati – 517520, A.P.
ganeshavula63@gmail.com

M Navya[4]
Department of AI&DS
Annamacharya Institute of Technology and Sciences, Tirupati – 517520, A.P.
navya20042@gmail.com

G Yashwanth [5]
Department of AI&DS
Annamacharya Institute of Technology and Sciences, Tirupati – 517520, A.P.
yashwanthganta507@gmail.com

ABSTRACT — With the rise of user-generated information in social media and other online platforms, there is an urgent requirement for the development of automatic systems that can analyze the opinions of people. Sentiment Analysis is one of the major applications of Natural Language Processing (NLP), which classifies the sentiments of given data into positive, negative, and neutral sentiments. This paper proposes a web-based sentiment analysis system for Hindi and English mixed text data using machine learning algorithms. The proposed system utilizes preprocessing steps such as tokenization, removal of stop words, and normalization of text data, and feature extraction using vectorization techniques. The system utilizes machine learning algorithms such as Naïve Bayes and Logistic Regression for classification purposes. The system is implemented using the Django platform, which provides an easy interface for real-time sentiment prediction. The results show that the proposed system using Logistic Regression achieves an accuracy of 79%, which is suitable for real-time applications. The proposed system is highly suitable and fills the gap between theoretical and practical applications.

Keywords — Sentiment Analysis, Natural Language Processing, Machine Learning, Hindi-English Text Data, Text Classification, Logistic Regression, Naïve Bayes Classifier, Tokenization, Stop Word Removal, Text Preprocessing, Feature Extraction, Bag of Words Model, TF-IDF Model, Django Framework, Web Application Development, User Interface Design, User Authentication System, Admin Module Development, Text Input System Design, File Upload Processing System, Real-Time Sentiment Prediction System, SQLite Database Management System, Backend Processing System, Frontend Interface Design, Python Programming Language, Data Processing System, Classification Model Development, Sentiment Prediction System Development, Web-Based

I.INTRODUCTION

The vast amounts of opinionated data in the modern digital world have been created by the continued exponentiating growth of user-generated material on different social media platforms, blogs, and online review sites. Manual analysis of such data is time consuming and is inefficient. Therefore, the necessity to have automated systems of sentiment analysis has been experienced. One of the significant NLP applications that highlights the identification and categorization of the emotional nuance of the information is Sentiment analysis. The information may be categorized into different sentiments which include positive, negative and neutral.

Although various studies have been carried out on the sentiment analysis of English texts, “the process of performing the same on the mixed texts containing both English and Hindi is quite challenging. The studies that have been carried out on the sentiment analysis of mixed texts are mostly focused on the evaluation and comparison of the algorithms. The application and usability of the system have been neglected in the studies.

The current paper discusses the development of a web-based sentiment analysis system for the analysis of Hindi-English mixed texts. The system is implemented with the integration of machine learning algorithms. The proposed system also includes the integration of various NLP preprocessing steps such as tokenization and feature extraction. The system is implemented with the integration of the Django framework. The implementation of the system provides the facility of an interactive interface for the user. The implementation of the machine learning algorithms in the proposed system provides the facility of performing the sentiment analysis in real-time.

II. RELATED WORK

Sentiment analysis is recognized as an important problem in the field of Natural Language Processing (NLP), with several techniques being proposed for sentiment analysis in different languages. Most of the earlier work on sentiment analysis was carried out using lexicon-based approaches, where SentiWordNet was used for sentiment analysis. However, lexicon-based approaches were not effective in capturing the meaning of sentences.

Approaches based on machine learning have also been trendy within recent times because of their capability to learn based on the data. The common machine learning-based methods that are being used in sentiment analysis are Naive Bayes, Support Vector Machine (SVM), Decision Tree, Logistic Regression, etc. An extensive comparative analysis of the application of these algorithms to Hindi and English text data is presented in . Based on the research, it is evident that Naive Bayes and SVM are very accurate.

Some of the recent developments in the field of sentiment analysis include the use of deep learning models such as Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU) models, along with word embedding techniques such as Word2Vec and fastText. These models have shown significant improvements in the accuracy of sentiment analysis by using semantic relationships between words.

For the specific case of code-mixed text such as Hindi-English, sentiment analysis poses several challenges in grammar, transliteration, and the availability of standard datasets. Several hybrid models using machine learning and lexicon-based approaches have also been proposed to improve the accuracy of sentiment analysis in code-mixed text. These models have shown moderate improvements in the accuracy of sentiment analysis.

Although several models have been proposed for sentiment analysis, most of them focus on the performance of the models, whereas few models have implemented the practical use of the models. Moreover, there is limited work on the development of user-friendly environments for the deployment of sentiment analysis models. With the limitations of the existing models, the proposed work focuses on developing a web-based sentiment analysis system using machine learning models.

III. METHODOLOGY

The system proposed is a web-based sentiment analysis system that is expected to analyze the given text in Hindi and English and classify it as positive, negative, or neutral sentiment. The system is based on machine learning and Natural Language Processing (NLP) techniques to analyze the input given by the user and provide real-time predictions. The methodology of the proposed system involves different stages, including text acquisition, preprocessing, feature extraction, classification, and result generation.

1.*System Overview*

The overall workflow of the sentiment analysis system is sequential in nature. The user is asked to provide the input text via the web interface. The input text is then passed through the NLP process and then the machine learning classifier. The output of the classifier is provided back to the user. The overall purpose of the system is to provide the user with an easy interface for the analysis of opinion in the mixed text of Hindi and English.

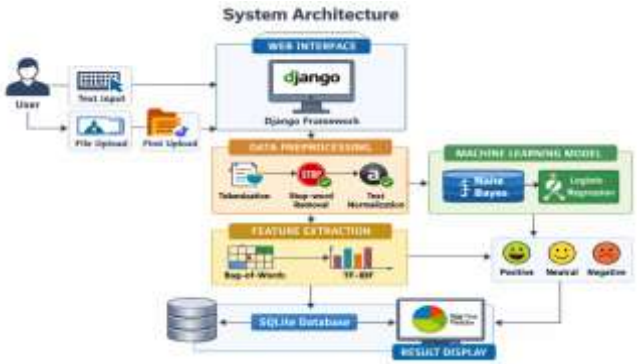

Fig 1 : System Architecture

2.*Text Acquisition and Input Handling*
The first step in the overall workflow of the system is acquiring the input. The system provides the user with the facility to either manually input the text or provide the text in the form of text files. This way, the system is able to process both long and short forms of text data. The input provided by the user is then passed on to the backend.

3.*Text Preprocessing*
In this step, the input text provided by the user is preprocessed. The preprocessing steps in this regard include removing the punctuations and special characters. The text is also converted to lowercase. The tokenization process is carried out to convert the text into individual words. The step of removing stop words is also carried out in this step.

4.*Feature Extraction*
After preprocessing, feature extraction techniques such as Bag-of-Words and TF-IDF convert the text into numerical form. Feature extraction techniques represent the importance of words in the text using feature vectors. The feature vectors obtained in this process are used in the next step of machine learning model implementation.

5.*Sentiment Classification Mechanism*
The next step in implementing the machine learning model is using machine learning algorithms such as Naïve Bayes and Logistic Regression. These algorithms classify the text into positive, negative, or neutral sentiments. It is observed that Logistic Regression provides good results in terms of accuracy. The process is efficient in terms of time complexity.

6.*Result Generation and Display*
Once the sentiment is predicted, it is displayed to the user through the interface. The system provides clear results on what category of sentiment is detected. The results are displayed instantly, enabling users to analyze text in real time.

IV. PERFORMANCE ANALYSIS

The performance of the proposed sentiment analysis system is analyzed using standard metrics for evaluating the performance of machine learning models. The proposed system can classify Hindi-English text into three classes: positive, negative, and neutral sentiment. The proposed system is evaluated on the basis of metrics such as accuracy, precision, recall, and F1-score.

The overall correctness of the proposed system is measured using accuracy which is the ratio of the number of samples which are handled correctly to the total number of samples which are passed to the system. Precision is a ratio of the amount of samples that are categorized as positive out of the total amount of samples that have been categorized as positive by the proposed system, whereas recall is a ratio of the amount of samples that are categorized as positive by the proposed system to the total number of samples that was given to the proposed system.

Among the algorithms implemented in the system, the performance of the Logistic Regression algorithm is better when compared to the Naïve Bayes classifier in terms of accuracy. The accuracy of the system is around 79%, which proves that the system performs well in real-time. The Naïve Bayes classifier also performs well in this regard but with low accuracy due to the assumption of the independent features used in the classifier.

The system is tested with various user inputs such as small sentences and long text data. The accuracy of the system in this regard is proved with the results obtained. The system performs well with all the inputs provided. However, the system faces difficulties in handling complex code-mixed sentences.

Another important aspect of the system is the real-time performance. The implementation of the system with the integration of various preprocessing techniques and the application of machine learning algorithms make the system perform well in this regard. The system generates the output in real-time after the input is provided.

The proposed system performs well in all aspects and can be implemented in various real-time scenarios.

V. RESULTS AND DISCUSSION

The sentiment analysis system proposed is tested on different types of text input on the English and Hindi language. This is to test the performance of the proposed system at large. The system proposed can be able to

classify the text into three categories; positive, negative, and neutral. Accuracy, precision, recall, and F1-score are taken into account to check the performance of the proposed system.

It is discovered in the proposed system that Naive Bayes is not performing well in comparison to Logistic Regression when used in an experiment. During experimentation, it is discovered that Logistic Regression is performing well against Naive Bayes in matters of consistency and accuracy. In this suggested system accuracy of 79% is obtained. This implies that the proposed system is performing better in real time applications. Naïve bayes is performing at a good rate but marginally low.
The system was tested with short sentences as well as longer text. It was found that the system performs well for opinions that are clearly expressed and for sentences that are simply structured. However, for handling complex sentences that are code mixed, there is a small degradation in performance. This is due to differences in grammar, colloquial usage, and vocabulary, which is often mixed in Hindi-English text.

When compared to other studies like , which focus on achieving higher accuracy with larger datasets and complex models, the proposed system focuses on usability and real-time performance. The incorporation of Django enables users to interact with the system and get results in real time.

Moreover, it is also observed that the system performs with quick response time. This is due to the optimization of the system in terms of processing and classification. This makes it suitable for real-world applications like analyzing social media, customer feedback, and reviews.

VII . CONCLUSION

This paper focused on the design and implementation of the web-based sentiment analysis system for the given text data in the Hindi and English languages. The overall objective of this research was to design and implement an efficient and user-friendly system that could analyze the given text data and classify the sentiments in the data in real time. The proposed system was successfully able to integrate the NLP techniques with the classification algorithms such as Logistic Regression and Naïve Bayes.

The experimental results clearly indicate that the proposed system was able to classify the sentiments in the text data with satisfactory accuracy. Among the various models implemented in the proposed system, the Logistic Regression approach was found to be performing better with an accuracy of around 79%", which clearly indicates the efficiency of the proposed system in handling the real-time text data. The proposed system was found to be performing well even with moderately long sentences. The overall contribution of this research was the implementation of the complete system rather than just comparing the models. The implementation of the proposed system with the Django framework provides the user with the facility to input the text data and get the output in real time. This helps in making the overall process of sentiment analysis more user-friendly and even non-technical users can make the most out of the proposed system.

Although the proposed system was found to be performing well in the overall process of sentiment analysis, some issues were observed with the accuracy of the output when the sentences were highly complex and code mixed.

REFERENCES

[1] Chandan Prasad Gupta, BalKrishan, ―Detecting Sentiment Analysis in Nepali Texts, IEEE, Page no (1-4),2015.
[2]ZhongkaiHu,JianqingHu,WeifengDing,XiaolinZheng ―Revie wSentiment Analysis based on Deep Learning‖,12th International Conference on e-business Engineering, IEEE, Page No(87-94), 2015. [3] Akhtar MS, Ekbal A, Bhattacharyya P (2016) Aspect-based sentiment analysis in Hindi: resource creation and evaluation. In: Proceedings of the tenth international conference on language resources and evaluation (LREC'16) [4] Xiadong Yan, Tao huang, ―Tibentian Sentence sentiment analysis based on the maximum entropy model‖.10th International Conference on Broadband and wireless Computing, Communication and Application, Page No(594- 597),2015. [5]ZhongkaiHu,JianqingHu,WeifengDing,XiaolinZheng ―Revie wSentiment Analysis based on Deep

Learning‖,12th International Conference on e-business Engineering, IEEE, Page No(87-94), 2015. [6] Santos I, Nedjah N, de Macedo Mourelle L (2017) Sentiment analysis using convolutional neural network with fastText embeddings. In: 2017 IEEE Latin American conference on computational intelligence (LA-CCI). IEEE

[7] Rehurek R, Sojka P (2010) Software framework for topic modelling with large corpora.In: Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks [8] Premjith B et al (2019) Embedding linguistic features in word embedding for preposition sense disambiguation in english—Malayalam machine translation context.Recent Adv Comput Intell 341–370

[9] Patra BG, Das D, Das A (2018)Sentiment analysis of codemixed Indian languages: an overview of SAIL_Code-Mixed Shared Task@ ICON-2017.arXiv preprint arXiv:1803.06745

[10] Nagamma P, Pruthvi H.R, Nisha K.K and Shweta NH, ―An Improved Sentiment Analysis of Online Movie Review Based on Clustering for Box Office Prediction‖, International Conference on Computing Communication and Automation (ICCCA 2015), Page No (933-937),2015.