Sentiment Analysis on IMDb Movie Reviews Using Machine Learning

Mrs.Vaidehi Kulkarni

Department of Information Technology, PES MCOE Pune, Maharashtra

Ravichandra Udupa, Shaikh Mohammed Zaid, Divya Thorat , Om Khatri,

UG Student, Department of Information Technology, PES MCOE Pune, Maharashtra

Abstract: Sentiment analysis is an artificial intelligence branch that focuses on expressing human emotions and opinions in the sense of data. Social Networking sites have become popular and common places for sharing wide range of emotions through short texts. These emotions include happiness, sadness, anxiety, fear, etc. Short texts help in identifying the sentiment expressed by the crowd. Sentiment Analysis identifies the overall opinion expressed by a reviewer towards a movie. The proposed model clearly differentiates between a positive review and negative review. In the proposed work, we show that the use of Since understanding the context of the reviews plays an important role in classification and helps in capturing the context of the movie reviews and hence increases the accuracy of classification. We look at the expression of emotions to determine whether a movie review is positive or negative, then standardize and use these features to train multiple label classifiers to identify movie reviews on the right label.

Keywords: Natural language processing, Sentimental analysis, lemmatization,
Stochastic Gradient Descent (SGD), Image Processing), Support Vector Machine (SVM)

I. INTRODUCTION

Social media has become an integral part of human living in recent days. People want to share each and every happening of their life on social media. The text plays a vital aspect in information shared, where users share their opinions on trending topics, politics, movie reviews, etc. These opinions which people share on social networking sites are generally known as Short Texts because of its length Short text have gained its importance over traditional blogging because of their simplicity and effectiveness in influencing the crowd. They are even used by search engines in the form of queries. Apart from their popularity, it has certain challenges like identification of sarcasm, sentiment, use of slang words, etc. Therefore, it becomes important to understand short texts and derive meaningful insights from them.

On the other side, a growing number of people are turning to the internet as a means of communicating their opinions to those they do not know. Sentiment is a feature that decides whether a text's expressed opinion is positive or negative, or whether it is about a particular subject. "Logan is a really nice movie, highly recommended 10/10," for example, sharing positive feelings about the film Logan, which is also the title of this post. For example, in the text "I am amazed that so many people put Logan in their favorite movies, I felt a good watch but not that good," the author's feelings about the movie might be good, but not the same message. Above everything, Emotional analysis is a collection of processes, strategies, and tools for gathering and converting knowledge below the level of language, such as viewpoint and attitude.

II. LITERATURE SURVEY

Avi Ajmera[1], in his paper devised and compared Bag of Words, TF-IDF, Word2Vec calculation methods were combined with a variety of classifier, including the Decision tree, Random Forest, Naive Bayes classifier, Support vector machine, , Decreased stochastic gradient, and checks, to determine emotions in the IMDB's database of movie reviews.Author found that Word2Vec with Stochastic Gradient Descent is the right mixture of emotions for the IMDB database segmentation issue based on test results.

To demonstrate the effectiveness of the IMDB review, H M Keerthi Kumar[2] et al focused on the Maximum Entropy with correlation shows the best results in terms of both accuracy and F-measure when compared to other classifiers. Use of Hybrid Feature Extraction Method (HFEM) makes the model more efficient in terms of accurate classification by adding the advantages of individual feature extraction method. HFEM improves the space complexity by reducing the input space to minimal number of features that are sufficient to represent the review content. space complexity and classification accuracy.

Palak Baid[3] et al in their research, devised various techniques to identify the polarity of the tweets. The algorithms performed were Naïve Bayes, K-Nearest Neighbour, Random Forest. The best results were given by Naïve Bayes classifier.

III. METHODOLOGY

1.Data Collection and Preprocessing:

IMDb movie reviews need to be collected from reliable sources or through APIs provided by IMDb. The raw data often requires cleaning to remove irrelevant information, such as HTML tags, special characters, and duplicate entries.

2.Natural Language Processing (NLP) Tools:

- Breaking down reviews into tokens (words or phrases) for analysis.
- Reducing words to their base or root form for consistency.
- Identifying parts of speech to understand the grammatical structure.
- Identifying entities such as movie names, actors, and locations mentioned in the reviews.

Natural language grammar, also known as NLP, refers to a computer program's capacity to comprehend human speech in its natural form. Natural language processing is a part of artificial intelligence. The development of natural language processing (NLP) applications is notoriously challenging since computers typically require people to "speak" to them in the language of an accurate, unambiguous, and well-structured system, or with a limited number of voice commands that are clearly written. Human speech, on the other hand, is not always accurate; it is frequently vague, and the structures of languages can be influenced by a variety of factors, such as slang, regional languages, and the social context in which they are spoken. Syntax analysis and semantic naming are the two primary methods that are utilized in natural language processing.

3.Feature Extraction:

Extracting relevant features from the reviews, such as specific keywords, adjectives, or phrases that indicate sentiment.

4.Model Training and Evaluation:

Training sentiment analysis models on labeled datasets and evaluating their performance using metrics like accuracy, precision, recall, and F1-score.

5.Text Pre-Processing and Normalization

Cleaning, pre-processing, and standardizing the text to bring text artefacts such as phrases and words to some degree format is one of the most critical steps before joining the stateof-the-art engineering and modelling process. This allows for ranking over the entire text, which aids in the creation of meaningful functionality and reduces the noise that can be imported because of a variety of factors such as inactive characters, special features characters, XML and HTML tags, and so on.

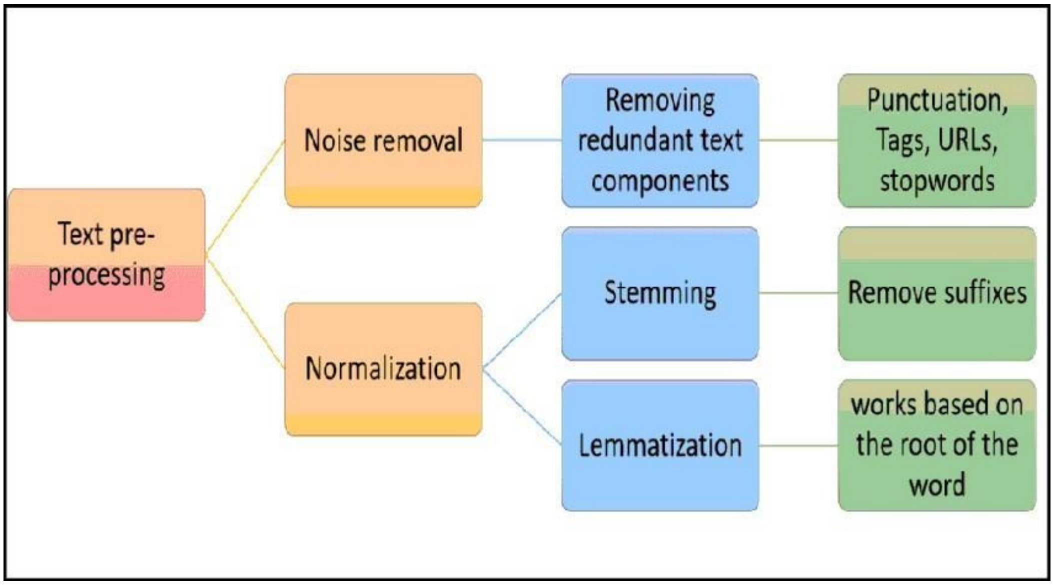

Text pre-processing
Noise removal
Normalization
Removing redundant text components
Stemming
Lemmatization
Punctuation, Tags, URLs, stopwords
Remove suffixes
works based on the root of the word

Fig.1.1 Text preprocessing

IV.SYSTEM FEATURES

4.1 Functional requirements

Functional requirements for a sentiment analysis system on IMDb movie reviews using machine learning outline the specific features and capabilities the system must possess to meet user needs and expectations. Here are key functional requirements for such a system:

- User Authentication and Authorization:
- Users should be able to create accounts, log in, and manage their profiles:
- Different levels of access should be provided for various user classes

- Data Collection and Retrieval:

- The system should collect IMDb movie reviews, either through APIs or web scraping.
- It should allow users to search for specific movies and retrieve their reviews.

- Text Preprocessing:
  - Text data should be cleaned, tokenized, and processed for further analysis
  - Stop words, special characters, and noise should be removed.

- Sentiment Analysis Models:
  - The system should employ machine learning models to classify reviews as positive, negative, or neutral
  - Multiple models should be available to cater to different languages and genres

V. DEVELOPMENT RESOURCES

1.Hardware Resources Required :

RAM : Min 8 GB, Storage : Min 256 GB,

Notebook,PyCharm, Computer : Standard Laptop or Desktop sufficient for development

GPU : NVIDIA, Web Server : AWS,Azure,Heroku

2.Software Resources Required

Operating System : Windows or Linux

IDE : Jupyter, Visual Studio

Frontend : HTML,CSS,JavaScript

Backend : Database : SQLite , PostgreSQL, MySQL

Version Control : GitHub or GitLab

VI. SYSTEM ARCHITECTURE

1.Activity Diagram:

Activity Diagram is a behavioural diagram presenting the actors their functions performed.

• They also include the swim lanes and the forks and joins.

• The represent the individual lane as their entire activities and the functionality carried out by that particular actor in the respective lanes.

• Input is being provided by the user and the output is being given back to the user.

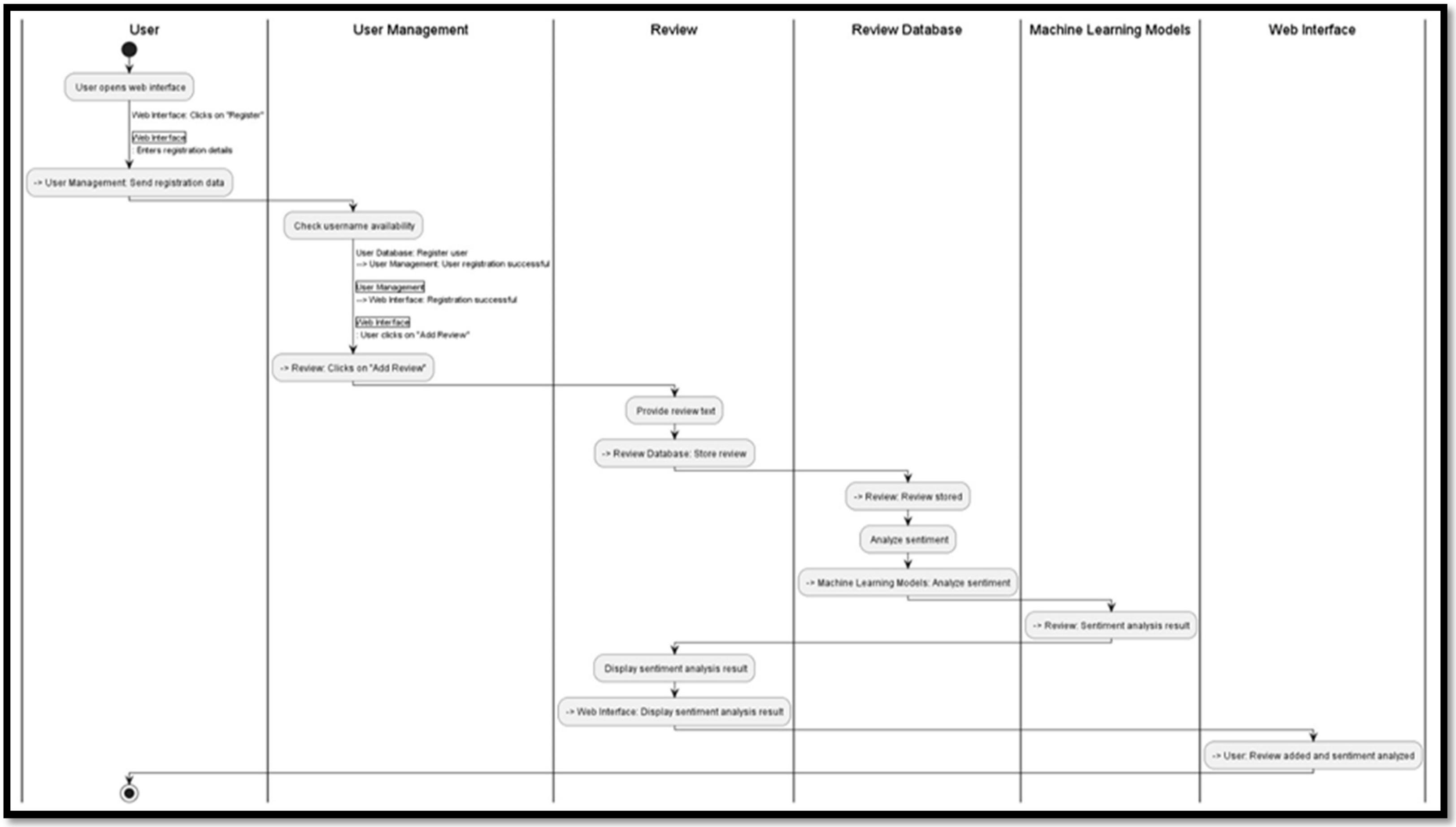

User
User Management
Review
Review Database
Machine Learning Models
Web Interface
User opens web interface
Web Interface: Clicks on "Register"
Web Interface
: Enters registration details
-> User Management: Send registration data
Check username availability
User Database: Register user
--> User Management: User registration successful
User Management
--> Web Interface: Registration successful
Web Interface
: User clicks on "Add Review"
-> Review: Clicks on "Add Review"
Provide review text
-> Review Database: Store review
-> Review: Review stored
Analyze sentiment
-> Machine Learning Models: Analyze sentiment
-> Review: Sentiment analysis result
Display sentiment analysis result
-> Web Interface: Display sentiment analysis result
-> User: Review added and sentiment analyzed

Fig. 6.1.Activity diagram of text summarizing system

2.UML Diagram:

Use Case diagram is used for representing the problem statement that is the actors in it, their functionality in an behavioural manner. They are useful when the system is to be in the programmatic execution.

In the above Figure 4.6 there are different actors:

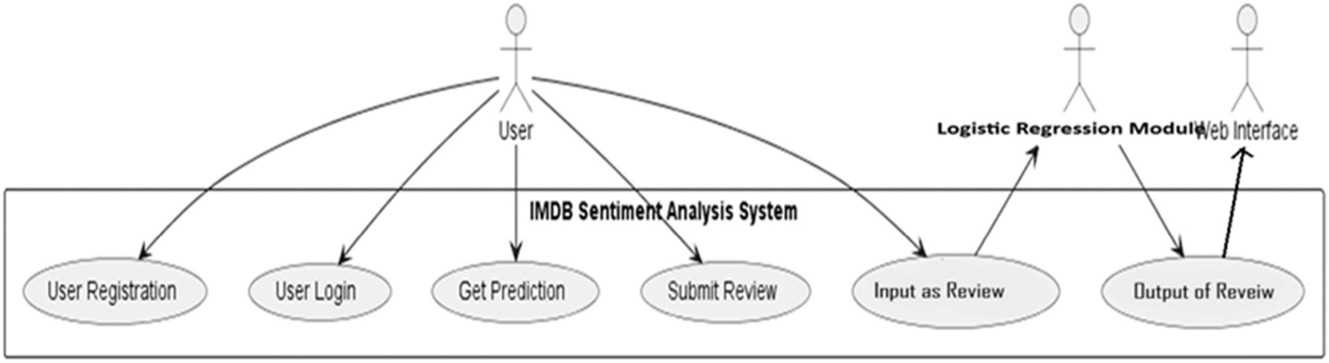
• Actors

– User

– Summarization System

• Functional Blocks include the stepwise execution of the entire system that is the form of different procedures (functionality).

• Input to the every function in this diagram is the output from the previous state.



VII. FUTURE WORK

The IMDb Review Sentiment Analysis project holds significant potential for future enhancements and expansions, including:

1. **Multilingual Support:**
   - Expanding language support to reach a wider global audience, enabling users to interact with IMDb content in their preferred languages.
2. F**ine-Grained Sentiment Analysis**:
   - Enhancing the sentiment analysis capabilities to provide more detailed and nuanced insights into user sentiments, including specific aspects and emotions expressed in reviews.
3. **Personalized Recommendations:**
   - Further customizing recommendations for registered users, utilizing their review sentiments and viewing history to provide tailored content suggestions.
4. **Real-Time Sentiment Analysis:**
   - Implementing real-time sentiment analysis for trending movies and TV shows, enabling users and the industry to stay updated on audience reactions.

VIII. CONCLUSION

This paper examines the issue of sentiment analysis. Basically we have used Logistic Regression with Natural Language Processing to determine emotions in the IMDB's database of movie reviews. It is versatile and effective classification algorithm for sentiment analysis tasks. It works by modelling the review belonging to a particular sentiment class, allowing us to classify reviews as positive, negative or neutral. Also we have used NLP preprocessing techniques which include tolenization,removing of stop word, stemming or lemmatization and feature engineering to transform text data into a suitable format for machine learning.

We have successfully developed a system that leverages the power of machine learning, specifically the Logistic Regression algorithm, to provide valuable insights into the sentiments expressed in IMDb reviews. This project has been an exciting exploration between technology and entertainment.

IX. REFERENCE

[1] Avi Ajmera," Sentiment Analysis of IMDb Movie Reviews", International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; Volume 10 Issue XII Dec 2022.

[2] H. M. Keerthi Kumar B. S. Harish, H. K. Darshan, "Sentiment Analysis on IMDb Movie Reviews Using Hybrid Feature Extraction Method", International Journal of Interactive Multimedia and Artificial Intelligence, Vol. No.5, 7 December 2018

[3] Palak Baid, Apoorva Gupta, Neelam Chaplot , "Sentiment Analysis of Movie Reviews using Machine Learning Techniques", International Journal of Computer Applications (0975 – 8887) Volume 179 – No.7, December 2017

[4]Refer H. M. Kumar and B. S. Harish. "Classification of Short Text Using Various Preprocessing Techniques: An Empirical Evaluation." In Recent Findings in Intelligent Computing Techniques pp. 19-30. Springer, Singapore, 2018.

[5] Refer K. S. Srujan, S. S. Nikhil, H. Raghav Rao, K. Karthik, B. S. Harish, and H. M. Kumar. "Classification of Amazon Book Reviews Based on Sentiment Analysis." In Information Systems Design and Intelligent Applications, pp. 401-411. Springer, Singapore, 2018.

[6] Refer H. M. Kumar, B. S. Harish, S. V. Kumar, and V. N. Aradhya. "Classification of sentiments in short-text: an approach using mSMTP measure". In Proceedings of the 2nd International Conference on Machine Learning and Soft Computing. pp. 145-150. ACM. 2018