International Journal of Ethics in Engineering & Management Education
Website: www.ijeee.in (ISSN: 2348-4748, Volume 11, Issue 1, January 2024)

IMDB-SENTI: A Nuance Handling Sentiment Analysis on Movie Review

SURESH KUMAR MANDALA
School of Computer Science & Artificial Intelligence
SR university, Telangana.
mandala.suresh83@gmail.com

NEELIMA GURRAPU
School of Computer Science & Artificial Intelligence
SR university, Telangana.
gneelima83@gmail.com

KARMILLA VINIL
School of Computer Science & Artificial Intelligence
SR university,
Telangana.karmillavinil@gmail.com

GURRAM KEERTHAN
School of Computer Science & Artificial Intelligence
SR university,
Telangana.keerthan4152@gmail.com

BALYA SAI POOJITHA
School of Computer Science & Artificial Intelligence
SR university,
Telangana.saipoojithabalya@gmail.com

SYED AKRAMUDDIN
School of Computer Science & Artificial Intelligence
SR university, Telangana.
syedsjunnu@gmail.com

KUMBA BHARGAVI
School of Computer Science & Artificial Intelligence
SR university,
Telangana.kumbabhargavi1234@gmail.com

Abstract: **In the field of natural language processing, sentiment analysis has grown in importance. It helps to derive insights from textual data and understand the expressed sentiment. This project is a valuable resource for understanding and implementing sentiment analysis techniques, especially for gauging public opinion from unstructured text data. This research explores sentiment analysis in detail by comparing the traditional methodology of Logistic Regression, convolutional neural networks (CNN), and the probabilistic approach of Naive Bayes. This project aims to provide an automated and accurate sentiment analysis solution for the film industry and other stakeholders. The solution will help them make informed decisions, improve user engagement, and gain valuable insights from user-generated reviews on IMDb.**

Keywords: **Natural Language Processing, Sentiment Analysis, Logistic Regression, Convolution Neural Networks, Naïve Bayes, IMDb.**

1. INTRODUCTION

Sentiment analysis, also known as opinion mining, is a branch of natural language processing that seeks to identify the underlying emotional tone in a given piece of text. With the abundance of user-generated content on the internet, grasping the public's sentiment has become crucial for businesses, governments, and individuals alike.

Sentiment analysis is a valuable approach for analyzing text and classifying it based on the emotional connotations of the words and phrases used. This technique entails determining whether the sentiments expressed in the text are positive, negative, or neutral. While some methods employ a rating scale that ranges from positive to slightly positive, neutral, slightly negative, and negative, sentiment analysis can be more complex than that. Nonetheless, it is a powerful tool that can be used to monitor customer feedback on products and services, analyze public opinion on a particular topic, and gain insights into the emotional responses of individuals or groups to specific events or situations.

Sentiment analysis is a popular application of natural language processing (NLP) and machine learning techniques that involves analyzing movie reviews to determine the sentiment expressed by reviewers, which is usually classified as positive, negative, or neutral. This kind of analysis is useful for understanding how audiences perceive and respond to movies, as well as for evaluating critical reception. Sentiment analysis on movie reviews is a relatively recent development, closely linked to advancements in natural language processing (NLP) and machine learning.

2. RELATED WORK

Sentiment analysis in the context of movie reviews has been extensively researched. Various methodologies and models have been proposed to extract insights from user-generated content. Notable studies include:

- Bag-of-words models were often used for sentiment analysis of reviews, but they struggled with contextual nuances while capturing explicit sentiments.
- Researchers have explored the application of machine learning models, such as Support Vector Machines (SVM)

International Journal of Ethics in Engineering & Management Education
Website: www.ijeee.in (ISSN: 2348-4748, Volume 11, Issue 1, January 2024)

and Decision Trees. These models consider feature interactions but face challenges in handling semantic complexities.

- The emergence of deep learning has brought a significant shift in sentiment analysis. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks have shown improved performance in capturing sequential dependencies in reviews.
- Ensemble models aim to improve sentiment prediction accuracy by combining multiple algorithms, mitigating the limitations of individual models.
- Some works have developed region-specific sentiment lexicons and models tailored to cultural nuances, contributing to more accurate sentiment interpretation.

It is important to note that although previous studies have provided valuable insights, there is still a gap in addressing the nuanced aspects of sentiment in movie reviews. Most models cannot handle subtleties and cultural variations effectively. To bridge this gap, IMDB-SENTI adopts a hybrid approach by integrating diverse models and incorporating cultural sensitivity to provide a more comprehensive sentiment analysis.

3. PROBLEM STATEMENT

Sentiment analysis in movie reviews faces challenges in capturing nuanced sentiments and cultural variations. Existing models often struggle with subtle expressions and fail to adapt to diverse user bases.

The specific problems identified include:

- Many sentiment analysis models fail to capture nuanced expressions in movie reviews, where sentiments are often subtle and context-dependent. This lack of nuance handling may cause current approaches to misinterpret the sentiment polarity.
- Sentiment analysis models may be inaccurate due to cultural variations. Differences in cultural context may lead to differing interpretations.
- Studies often focus on individual models, like bag-of-words or deep learning, without exploring model integration. Combining various models can enhance sentiment predictions.
- Language is constantly evolving, and sentiments are expressed in new ways. Traditional models may struggle to keep up.

a. REQUIREMENT ANALYSIS

3.1.1. Functional Requirements:

- User Input: Users can easily submit movie reviews for sentiment analysis.
- Multi-Model Integration: The system would integrate logistic regression, CNN, and Naive Bayes models for sentiment prediction.
- Nuance Handling: IMDB-SENTI would accurately analyze movie reviews for better predictions.
- Cultural Adaptability: To accurately interpret sentiments across diverse user bases, the system would take into account cultural variations. It is important to ensure that the system's analysis of sentiment is not biased or inaccurate.
- Dynamic Language Model: IMDB-SENTI would use a dynamic language model to adapt to evolving expressions in movie reviews.
- Output: Users would receive sentiment predictions indicating confidence.

3.1.2. Non-Functional Requirements:

- Usability: The system would have a simple and user-friendly interface, accessible to users of all technical backgrounds.
- Performance: IMDB-SENTI would promptly provide sentiment predictions for smooth UX.
- Scalability: The system is designed to handle a growing user base and expanding movie review dataset.
- Security: User data would be handled securely to protect privacy and preserve sentiment analysis integrity.
- Accuracy: The system would continuously undergo performance evaluation and improvement for accurate sentiment predictions.

3.1.3. Technical Requirements

The following requirements will be helpful for the project.

- Programming knowledge of Python is required for coding with Machine Learning frameworks such as sci-kit-learn and Tensor Flow, as well as Python libraries for Natural Language Processing like NLTK and spaCy.
- We need an open-source dataset of diverse IMDB movie reviews with labeled sentiments.
- General hardware requirements to run a machine learning model and train it with deep learning architecture include a computer with a sufficient amount of RAM and an efficient GPU.
- This project may require funds for cloud computing resources and datasets.

b. RISK ANALYSIS

Here are some potential risks and corresponding mitigation strategies for IMDB-SENTI, a sentiment analysis system for movie reviews:

3.2.1. Data Bias:

Risk: The training data might contain biases, leading to skewed sentiment predictions.

Mitigation: To address this, we will regularly update the training dataset, implement data augmentation techniques, and conduct bias audits.

3.2.2. Model Complexity:

Risk: Integrating multiple models could add complexity and impact performance.

International Journal of Ethics in Engineering & Management Education
Website: www.ijeee.in (ISSN: 2348-4748, Volume 11, Issue 1, January 2024)

Mitigation: To mitigate this risk, we will conduct extensive testing, optimize code, and choose model combinations that provide a good trade-off between accuracy and complexity.

3.2.3. Cultural Sensitivity:

Risk: Misinterpretation of sentiments due to cultural nuances in reviews.

Mitigation: To address this, we will implement a cultural analysis module, continually update the system's cultural understanding, and provide options for users to specify cultural context.

3.2.4. User Privacy:

Risk: Storing and processing user reviews could raise privacy concerns.

Mitigation: To mitigate this risk, we will implement robust encryption for user data, anonymize reviews during model training, and comply with relevant privacy regulations.

3.2.5. Dynamic Language Changes:

Risk: Rapid changes in language usage could affect the models' accuracy.

Mitigation: To address this, we will implement a mechanism for dynamic language model updates, keeping the system aligned with evolving linguistic expressions.

3.2.6. Scalability Issues:

Risk: Increased user base and review submissions could strain system resources.

Mitigation: To mitigate this risk, we will design the system with scalability in mind, employ cloud services if necessary, and periodically assess resource usage.

3.2.7. Model Evaluation:

Risk: Inaccurate evaluation metrics may lead to a misrepresentation of model performance.

Mitigation: To address this, we will use multiple evaluation metrics, conduct thorough cross-validation, and involve domain experts in assessing results.

3.2.8. Security Threats:

Risk: Potential vulnerabilities may be exploited, leading to unauthorized access or data breaches.

Mitigation: To mitigate this risk, we will regularly update security protocols, conduct penetration testing, and employ robust access controls.

3.2.9. User Adoption:

Risk: Users may not trust or adopt the sentiment predictions.

Mitigation: To address this, we will provide clear explanations of predictions, share model accuracy metrics, and offer user-friendly features for feedback and improvement.

By taking these steps, we aim to build a resilient and trustworthy sentiment analysis system for movie reviews.

c. FEASIBILITY ANALYSIS

The task at hand involves the implementation and integration of three distinct sentiment analysis models, namely Logistic Regression, CNN, and Naive Bayes. This requires technical proficiency in machine learning, natural language processing, and software development.

Evaluation: Open-source frameworks and libraries facilitate model integration, and the team possesses the requisite technical know-how. Realistic.

i. Financial feasibility

The creation and upkeep of IMDB-SENTI entail expenses linked to the procurement of data, training models, and server infrastructure. It is necessary to take into account monetization possibilities like subscription plans and ad integration.

Evaluation: Although initial expenditures for building the model and data are necessary, future income streams may be able to cover these expenses.

ii. Ethical and Legal feasibility:

The project needs to follow sentiment analysis ethics, respect user privacy, and abide by data protection rules.

Evaluation: Complying with legal and moral requirements is crucial. The group pledges to use ethical AI techniques and strong data protection protocols.

iii. Commercial Viability:

Requirements for IMDB-SENTI's success include user happiness and adoption. Market acceptance depends on having a thorough understanding of user preferences and wants.

Assessment: To make sure the system meets user expectations; market research and user input tools will be used.

4. PROPOSED ALGORITHM

IMDB-SENTI uses multiple approaches for sentiment analysis, taking advantage of the unique capabilities of Logistic Regression, Convolutional Neural Network (CNN), and Naive Bayes. Rather than integrating these models, the project focuses on comparing their performance in analyzing the sentiments of movie reviews. Each model is applied independently to provide a complete understanding of their strengths and weaknesses.

Logistic Regression is a powerful statistical model that is commonly used for binary classification tasks. It is often applied to analyze linear relationships between features extracted from sources such as movie reviews and their corresponding sentiments, whether positive or negative. The model boasts several strengths, including its simplicity, interpretability, and efficiency in processing large datasets. Using the interpretable machine learning technique known as logistic regression, we attempt to extract both the TF-IDF and the bag of words from the provided text. Thus, techniques like GloVe and Word2Vec are applied. Following this, the features are given weights in order to determine if the review is favorable, unfavorable, or neutral. Metrics including accuracy, precision, recall, F1-Score, and ROU-AUC are used in the model performance testing. For the model to perform better, one of the most crucial Fine Tune parameter considerations is also made.

Convolutional Neural Networks (CNN) is designed to capture spatial hierarchies, which makes them particularly suitable for analyzing image and sequence data. In the context of IMDB-SENTI, CNN can be adapted for natural language processing. Specifically, convolutional layers can be used to identify local patterns in word sequences. The strengths of CNN are its

International Journal of Ethics in Engineering & Management Education
Website: www.ijeee.in (ISSN: 2348-4748, Volume 11, Issue 1, January 2024)

ability to understand spatial relationships in textual data and to capture nuanced features that might otherwise be missed. For text analysis, the CNN model will work well with sequential text. In order to capture patterns, this model will be created in a way that allows it to automatically learn the hierarchical features from the text. Thus, the initial feature extraction from the text input is done using the CNN model. The model is able to pick up on crucial details from the reviews. The model features are extracted using the pooling and embedding layers. Naive Bayes is a probabilistic model that is based on Bayes' theorem, which assumes that the features are independent of each other. It is commonly used to determine the likelihood of a review belonging to a specific sentiment class based on word occurrences. One of its main strengths is its ability to handle large feature spaces with efficiency, simplicity, and effectiveness. With probabilistic concepts as its basis, the system is easy to understand and put into practice. Naive Bayes works well for problems involving text classification, such as sentiment analysis. Text characteristics and other high-dimensional data are managed well by it. Since the algorithm relies on the premise of feature independence, text data can be processed with particular efficiency. The "naive" assumption appears reasonable because words in a document are usually viewed as independently occurring in sentiment analysis.

4.1 DESIGN

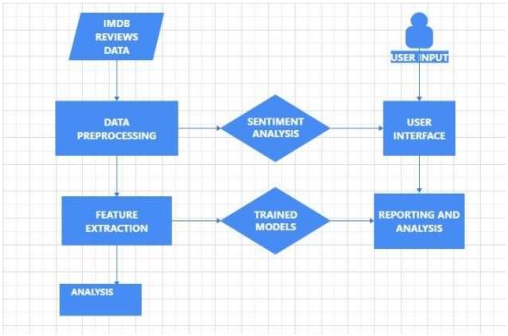


Fig. 7.1 Architecture Diagram

The Architecture diagram depicts the IMDB-SENTI sentiment analysis project's sequential data flow and activities are depicted in this diagram: Data Processing Pipeline: IMDb Reviews serve as the source of data. Utilizes data pre-processing to organize and sanitize the text. After that, data is sent to feature extraction and model training. In a separate step, the data is used for Model Training with various methods such as bag-of-words, word embeddings, and TF-IDF. Assessment and Comparative Analysis: The performance of the trained models is assessed. The advantages and disadvantages of each model are determined through comparative analysis. User Communication- Through a User Interface, users can submit movie reviews (GUI).The trained models are used by the GUI to initiate sentiment analysis. User Input and Reporting: Input from users is recorded to help improve the system. Reports and analysis shed light on the overall performance of the project.

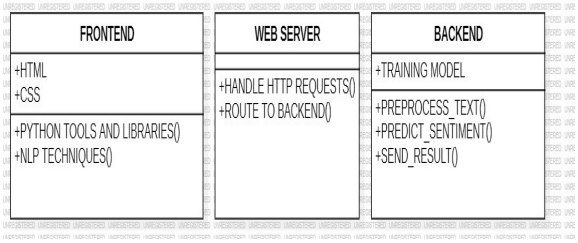


Fig 7.2 CLASS Diagram

The frontend uses HTML form components and CSS functions for user interaction. The backend uses a model and pre-processing operations to predict sentiments, with weights and parameters determined through training.

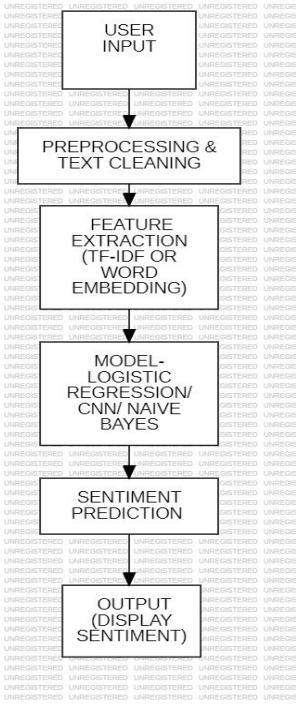
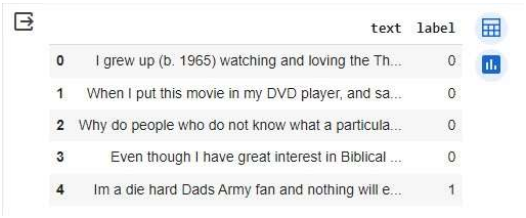


Fig. 7.3 Data Flow Diagram

“User input” indicates the input text (user reviews). "Pre-processing" Includes cleanup, punctuation, and other word processing features. “Feature extraction”, this is where features are extracted from the processed documents. “Model" Perform a sentiment analysis using the given model. “Sentiment prediction" here the system predicts the emotion (positive or negative). "Output" indicates the final (unexpected) result.

International Journal of Ethics in Engineering & Management Education
Website: www.ijeee.in (ISSN: 2348-4748, Volume 11, Issue 1, January 2024)

4.2 IMPLEMENTATION



	text	label
0	I grew up (b. 1965) watching and loving the Th...	0
1	When I put this movie in my DVD player, and sa...	0
2	Why do people who do not know what a particula...	0
3	Even though I have great interest in Biblical ...	0
4	Im a die hard Dads Army fan and nothing will e...	1

Fig 8.1 Description of the dataset which has two columns text and label

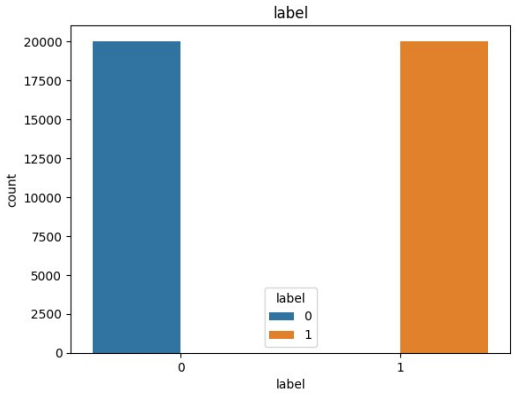


Fig 8.2 the figure shows that the data is balanced

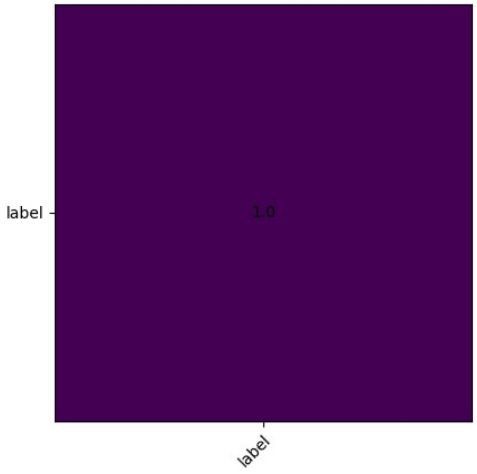


Fig 8.3 Heatmap showing the correlation value as 1.0

	text	label	tokens
0	i grew up b watching and loving the thunderbir...	0	[i, grew, up, b, watching, and, loving, the, t...
1	when i put this movie in my dvd player and sat...	0	[when, i, put, this, movie, in, my, dvd, playe...
2	why do people who do not know what a particula...	0	[why, do, people, who, do, not, know, what, a,...
3	even though i have great interest in biblical ...	0	[even, though, i, have, great, interest, in, b...
4	im a die hard dads army fan and nothing will e...	1	[im, a, die, hard, dads, army, fan, and, nothi...

Fig 8.4 Tokenization

Text Cleaning: The clean text function uses regular expressions to get rid of unnecessary whitespace, special characters, and digits from the text. For uniformity, it changes the text to lowercase. Tokenization: The cleaned text is tokenized using the word tokenize function from the NLTK package. In the 'text' column, tokenization is applied to every movie review. Outcome: A new column called "tokens" has the cleaned and tokenized data. 'Tokens' is a column that now has word lists taken directly from the related movie reviews. In order to guarantee that the text data is in a consistent and useable format, text cleaning is an essential pre-processing step. By dissecting the text into individual words, tokenization creates a foundation for additional analysis.

5 RESULTS AND COMPARATIVE ANALYSIS

Logistic regression

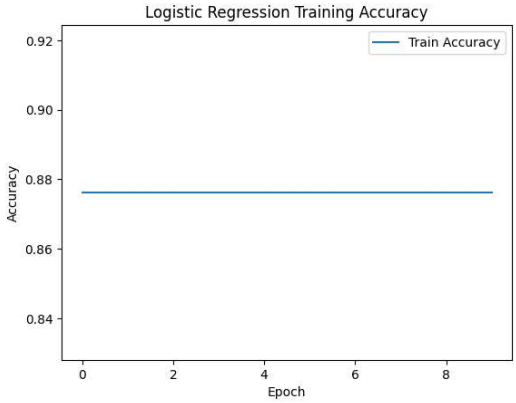
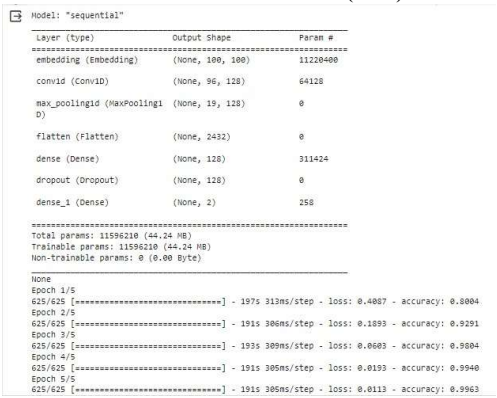


Fig 9.1 Accuracy curve of Logistic Regression

Convolution neural networks (CNN)



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 100, 100)          11220400

 conv1d (Conv1D)             (None, 96, 128)           64128

 max_pooling1d (MaxPooling1  (None, 19, 128)           0
 D)

 flatten (Flatten)           (None, 2432)              0

 dense (Dense)               (None, 128)               311424

 dropout (Dropout)           (None, 128)               0

 dense_1 (Dense)             (None, 2)                 258

=================================================================
Total params: 11596210 (44.24 MB)
Trainable params: 11596210 (44.24 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
None
Epoch 1/5
625/625 [==============================] - 197s 313ms/step - loss: 0.4087 - accuracy: 0.8004
Epoch 2/5
625/625 [==============================] - 191s 306ms/step - loss: 0.1893 - accuracy: 0.9291
Epoch 3/5
625/625 [==============================] - 193s 309ms/step - loss: 0.0603 - accuracy: 0.9804
Epoch 4/5
625/625 [==============================] - 191s 305ms/step - loss: 0.0193 - accuracy: 0.9940
Epoch 5/5
625/625 [==============================] - 191s 305ms/step - loss: 0.0113 - accuracy: 0.9963
```

Fig 9.2 CNN training data with 5 epochs

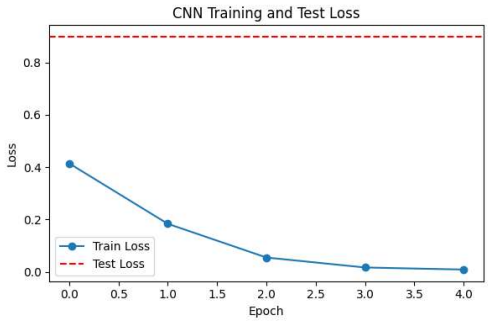


Fig 9.3 CNN training and test loss curves

International Journal of Ethics in Engineering & Management Education
Website: www.ijeee.in (ISSN: 2348-4748, Volume 11, Issue 1, January 2024)

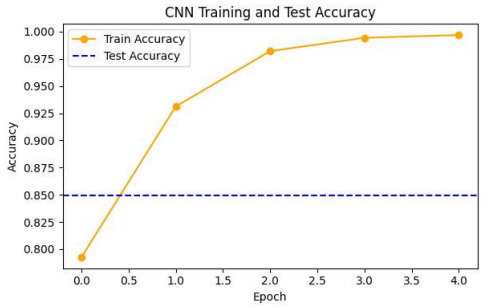


Fig 9.4 CNN training and test accuracy curve

Naïve bayes

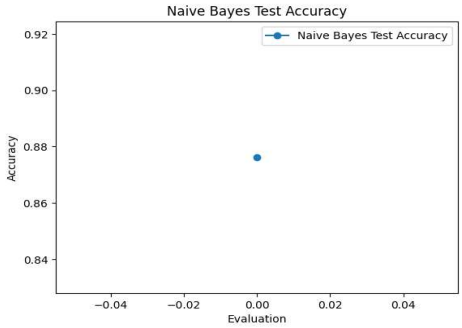


Fig 9.5 Test accuracy curve of Naïve bayes

Graphical user interface

Fig 9.6 GUI of the IMDB-SENTI

The User-Interface is user friendly. It has a hollow space for entering the user input. The interface also has two buttons namely Analyse sentiment and Clear. Analyse sentiment button performs the expected sentiment analysis and gives the output below as positive or negative. Clear button clears the input area so that the user can give another input.

Comparative Analysis

Logistic Regression Classification Report:

	precision	recall	f1-score	support
0	0.51	0.47	0.49	2582
1	0.48	0.51	0.49	2418
accuracy			0.49	5000
macro avg	0.49	0.49	0.49	5000
weighted avg	0.49	0.49	0.49	5000

Fig 9.7 Classification report of Logistic regression model

157/157 [==============================] - 3s 18ms/step

CNN Classification Report:

	precision	recall	f1-score	support
0	0.85	0.85	0.85	2495
1	0.85	0.85	0.85	2505
accuracy			0.85	5000
macro avg	0.85	0.85	0.85	5000
weighted avg	0.85	0.85	0.85	5000

Fig 9.8 Classification report of CNN model

Naive Bayes Classification Report:

	precision	recall	f1-score	support
0	0.88	0.86	0.87	2495
1	0.86	0.89	0.87	2505
accuracy			0.87	5000
macro avg	0.87	0.87	0.87	5000
weighted avg	0.87	0.87	0.87	5000

Fig 9.9 Classification report of Naïve bayes model

Accuracy Comparison Table:

	Model	Accuracy
0	Logistic Regression	0.8724
1	CNN	0.8490
2	Naive Bayes	0.8515

Fig 9.10 Accuracy Comparison table

After training and evaluating the three models, Logistic regression model has achieved an accuracy of 87%, Convolution Neural Networks model has achieved an accuracy of 84% and the Naïve bayes model has achieved 85% accuracy.

6 CONCLUSION WITH CHALLENGES

After comparing all the parameters in the classification reports of three models and their accuracies, We Conclude that, Logistic Regression and Naïve Bayes are better-suited models than Convolution Neural Networks (CNN) model while performing Sentiment Analysis on Movie reviews.

To sum up, the sentiment analysis project has been a worthwhile educational experience that has shed light on several different facets of machine learning and natural language processing (NLP). The application of several models, such as CNN, Naive Bayes, and Logistic Regression, made it possible to thoroughly investigate several methods for text classification.

The project's application was further improved with the creation of a graphical user interface (GUI), which made it simple for users to interact with and acquire sentiment forecasts. The project's collaborative style, which involves team members and maybe version control systems like Git, is indicative of how data science initiatives work in the real world.

The voyage was not without difficulties, though. Here are a few noteworthy difficulties: Managing unstructured textual data frequently presents issues with data quality and necessitates extensive preparation to address noise and irregularities. Selecting suitable models and fine-tuning hyper parameters can be difficult and necessitate a thorough comprehension of the advantages and disadvantages of each model. There are several obstacles to overcome while moving from model development to deployment, such as deciding on the best deployment platform and guaranteeing the model's scalability. It's a constant struggle to address moral issues with bias in models and data privacy.

In addition to providing a sentiment analysis solution, this project gave the team transferable abilities that they could use in a variety of data science and machine learning applications. As the project comes to a close, the lessons learned and obstacles surmounted provide a solid basis for future undertakings and a dedication to continuous education in this ever-evolving sector.

REFERENCES

[1]. Pang, B., Lee, L., & Vaithyanathan, S. (2002). Sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing (Vol. 10, pp. 79-86).

[2]. Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1631-1642)

[3]. Hutto, C. J., & Gilbert, E. E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In Eighth International Conference on Weblogs and Social Media (ICWSM-14).

[4]. Bishop, C. M. (2006). Pattern recognition and machine learning. springer.

[5]. Selvi, P. S., Arivoli, D. (2017). A Survey of Sentiment Analysis Techniques. International Journal of Computer Applications, (0975 – 8887).

[6]. LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

[7]. Rish, I. (2001). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46).