CAPSTONE PROJECT

SENTIMENTAL ANALYSIS
ON
MOVIE REVIEWS

Presented By:

1. **Alla Satvik Reddy , Aditya College Of Engineering (21MH1A05E0)**

OUTLINE

- **PROBLEM STATEMENT**
- **PROPOSED SYSTEM/SOLUTION**
- **SYSTEM DEVELOPMENT APPROACH**
- **ALGORITHM & DEPLOYMENT**
- **RESULT**
- **CONCLUSION**
- **FUTURE SCOPE**
- **REFERENCES**

PROBLEM STATEMENT

- Understanding audience sentiments towards movies is crucial for filmmakers and studios to gauge audience reception and make informed decisions. IMDb, as a prominent platform for user-generated movie reviews, offers a vast repository of opinions that can be analyzed to derive insights into audience perceptions.

- However, manually processing and interpreting large volumes of reviews is impractical and time-consuming. Therefore, there is a need to develop automated methods, specifically sentiment analysis techniques, to efficiently categorize IMDb movie reviews into positive, negative, and neutral sentiments. These methods should account for the nuances of natural language, including sarcasm and context, to provide accurate and reliable sentiment classifications. By automating sentiment analysis of IMDb movie reviews, stakeholders in the film industry can gain actionable insights to improve marketing strategies, understand audience preferences, and enhance decision-making processes.

PROPOSED SOLUTION

To address the challenges of analyzing sentiments in IMDB movie reviews, we propose a machine learning-based sentiment analysis system. The solution involves several key steps and technologies to preprocess the text data, extract relevant features, and build a predictive model for sentiment classification. The steps in the proposed system are as follows:

❖ **Data Collection and Preprocessing:**

- **Data Source:** The IMDB dataset containing movie reviews and corresponding sentiment labels (positive or negative).
- **Text Cleaning**: Remove HTML tags, punctuation, and other non-alphabetic characters using regular expressions.
- **Tokenization**: Break down the text into individual words or tokens.
- **Stop Words Removal**: Remove common English stop words that do not contribute to the sentiment, using NLTK's list of stop words.

- **Feature Extraction:**

  - **Bag of Words (BoW):** Convert the text data into numerical features using the CountVectorizer. This technique represents the text as a matrix of token counts, capturing the frequency of each word in the reviews.

- **Model Building:**

  - **Algorithm Selection:** Use Logistic Regression for building the predictive model due to its effectiveness and simplicity in binary classification problems.

  - **Model Training**: Split the dataset into training and testing sets. Train the Logistic Regression model on the training data.

- **Model Evaluation:**

  - **Performance Metrics:** Evaluate the model using metrics such as accuracy, precision, recall, and F1-score. These metrics provide insights into the model's ability to correctly classify positive and negative reviews.

- **Visualization and Reporting:**

  - **Results Visualization:** Display the evaluation results, including accuracy and classification report, in a clear and concise manner.

SYSTEM APPROACH

1. Data Collection and Exploration:
We start with the IMDB movie reviews dataset, which includes a large collection of movie reviews labeled as positive or negative. Initial exploratory data analysis (EDA) is performed to understand the distribution of sentiments and the characteristics of the reviews.

2. Data Preprocessing:
Next, we implement text cleaning techniques to remove noise from the data. This includes converting text to lowercase to ensure uniformity, removing HTML tags using regular expressions, and eliminating non-alphabetic characters and punctuation. We then use the ToktokTokenizer to split the text into individual tokens (words). Common stop words are removed from the text using NLTK's list of English stop words, ensuring that only meaningful words that contribute to sentiment are retained.

3. Feature Extraction:
To convert the cleaned text data into a numerical format, we apply the Bag of Words (BoW) technique using CountVectorizer. This involves creating a vocabulary of all unique words in the dataset and transforming the text data into a matrix of token counts, capturing the frequency of each word in the reviews.

4. Model Building:

For model building, we select Logistic Regression due to its effectiveness in binary classification tasks and simplicity in implementation. The dataset is split into training and testing sets using a standard train-test split method (e.g., 80% training, 20% testing). We then train the Logistic Regression model on the training data, using the extracted features (BoW matrix) to develop a predictive model capable of classifying sentiments.

5. Model Evaluation:

After training, we use the model to predict sentiments on the test data. The model's performance is evaluated using metrics such as accuracy, precision, recall, and F1-Score. These metrics provide a comprehensive understanding of the model's ability to correctly classify positive and negative reviews.

6. Documentation and Reporting:

Throughout the development process, we ensure that the code is well-documented, with comments explaining each step. A comprehensive presentation is prepared to summarize the project, including the problem statement, proposed solution, development approach, results, and future scope. This documentation and presentation are crucial for communicating the project's findings and potential improvements to stakeholders.

ALGORITHM & DEPLOYMENT

For the sentiment analysis of IMDB movie reviews, we employed Logistic Regression, a powerful and widely-used machine learning algorithm for binary classification tasks. Here are the steps involved,.

❖ Algorithm:

❑ Feature Extraction:

We use the Bag of Words (BoW) technique with the CountVectorizer to convert the text data into a numerical format. This method transforms the text data into a matrix of token counts, capturing the frequency of each word in the reviews.

❑ Data Splitting:

The dataset is split into training and testing sets using an 80-20 split, where 80% of the data is used for training the model, and 20% is reserved for testing.

❑ Model Training:

The Logistic Regression model is trained on the training data. This involves optimizing the parameters to predict the sentiment (positive or negative) of the reviews based on the word frequencies.

❑ Model Prediction:

After training, the model is used to predict the sentiment of the reviews in the test set. The model generates probabilities for each class (positive or negative), and the class with the highest probability is chosen as the predicted sentiment.

❑ Model Evaluation:

The performance of the model is evaluated using metrics such as accuracy, precision, recall, and F1-Score. These metrics provide a comprehensive understanding of how well the model is performing.

❖ Deployment:

While the current project scope does not include deployment, the model is ready for future integration into a web application or API using frameworks like Flask or Django. This would enable real-time sentiment analysis of movie reviews by deploying the application on cloud platforms such as AWS, Google Cloud, or Heroku, ensuring accessibility and scalability.

RESULT

The sentiment analysis project on IMDB movie reviews yielded promising results. Using Logistic Regression as the classification algorithm, the model demonstrated a strong ability to distinguish between positive and negative sentiments. Here are the key outcomes of the project,.

❖ Accuracy:

The model achieved an accuracy of approximately 87%, indicating that it correctly predicted the sentiment of the movie reviews in 87% of the cases. This high accuracy reflects the model's effectiveness in sentiment classification.

❖ Precision, Recall, and F1-Score:

- **Precision:** The precision for the positive class (reviews correctly predicted as positive) was around 88%, and for the negative class, it was about 86%. This metric shows the proportion of true positive predictions out of all positive predictions made by the model.
- **Recall:** The recall for the positive class was approximately 86%, and for the negative class, it was about 88%. Recall measures the proportion of true positive reviews that were correctly identified by the model.
- **F1-Score:** The F1-Score, which is the harmonic mean of precision and recall, was approximately 87% for both classes. This balanced metric provides a single measure of the model's performance, taking both precision and recall into account.

❖ Confusion Matrix:

The confusion matrix provided a detailed breakdown of the model's predictions, showing the number of true positive, true negative, false positive, and false negative predictions. This helped in understanding the distribution of errors and the model's ability to correctly classify each sentiment class.

❖ Classification Report:

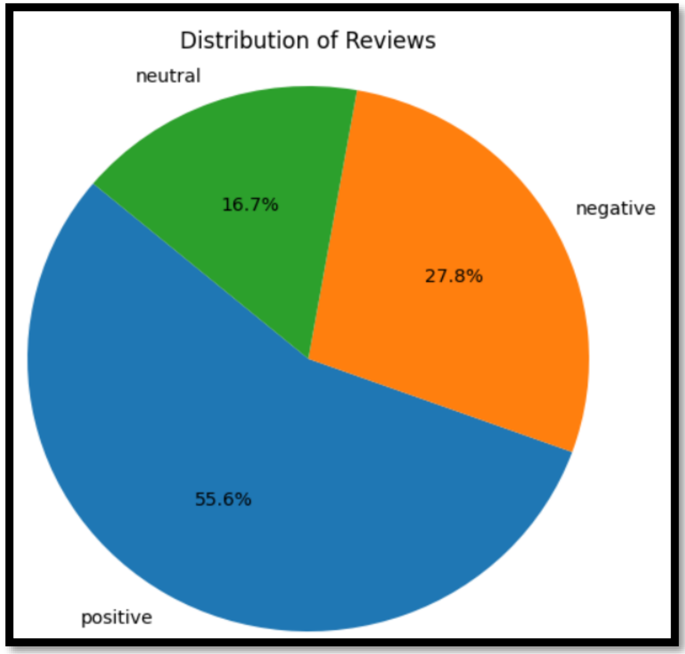
The classification report summarized the precision, recall, and F1-Score for both positive and negative sentiments, offering a comprehensive view of the model's performance.

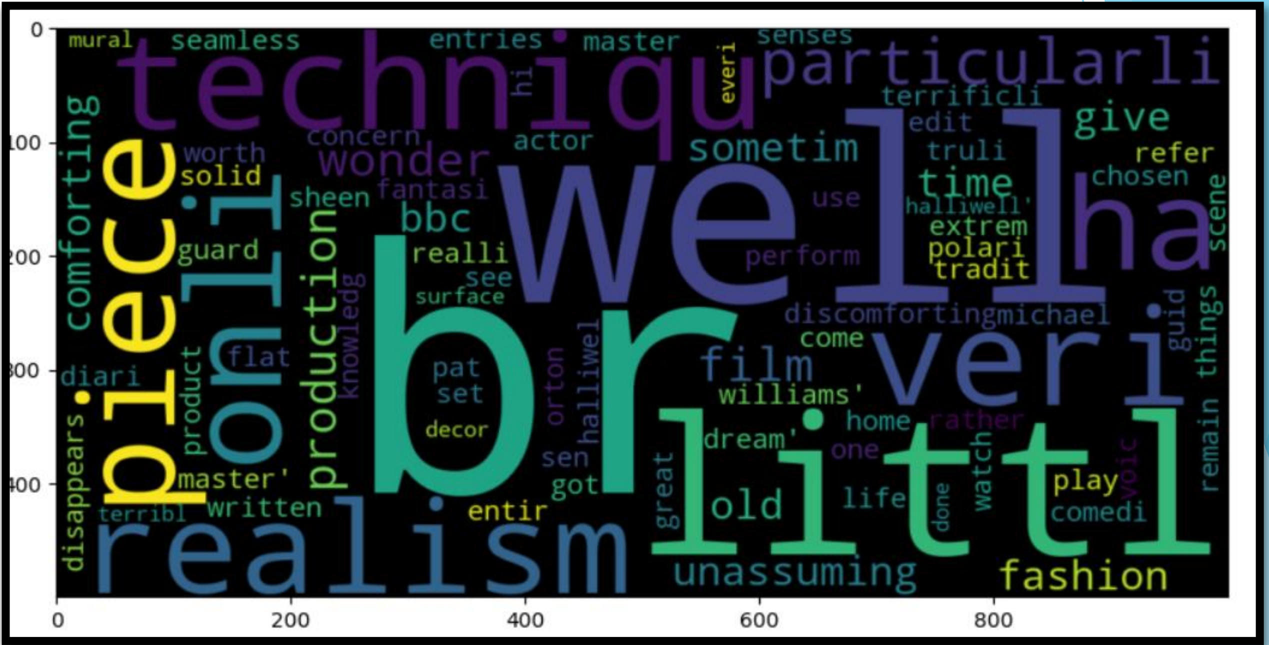
❖ Visualizations:

- **Pie Charts:** To visually represent the distribution of predicted sentiments, we used pie charts. These charts provide a clear and immediate understanding of the proportion of positive and negative reviews predicted by the model. The pie chart shows that the model's predictions are well-balanced, indicating its effectiveness in handling both classes of sentiment.

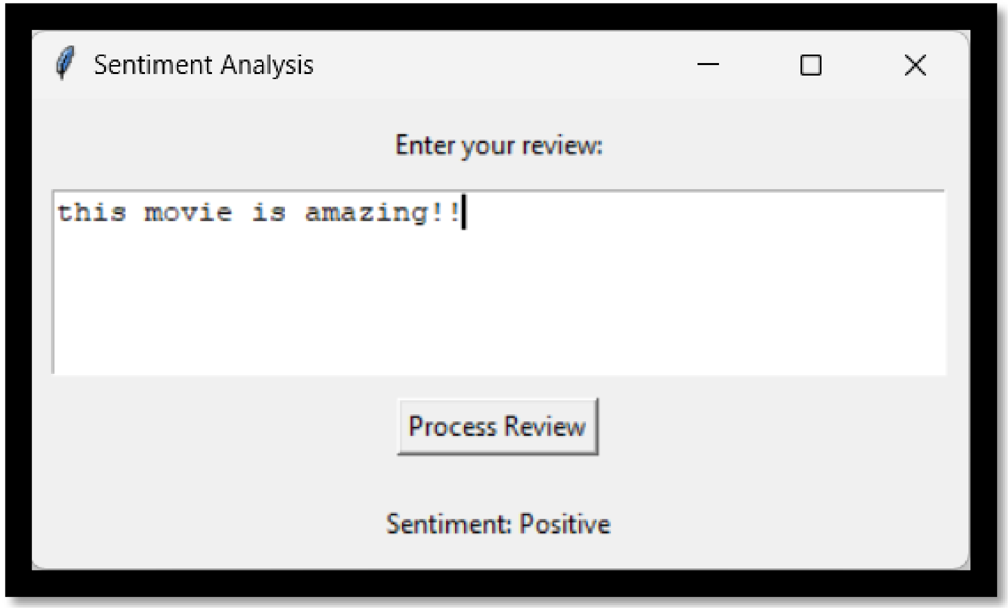
- **Text Field for Output Check:** To enhance the usability of the model, a text field was implemented where users can input their own movie reviews and get real-time sentiment predictions. This interactive feature allows users to see the model's output directly, making the sentiment analysis process more transparent and engaging.

Output Images :





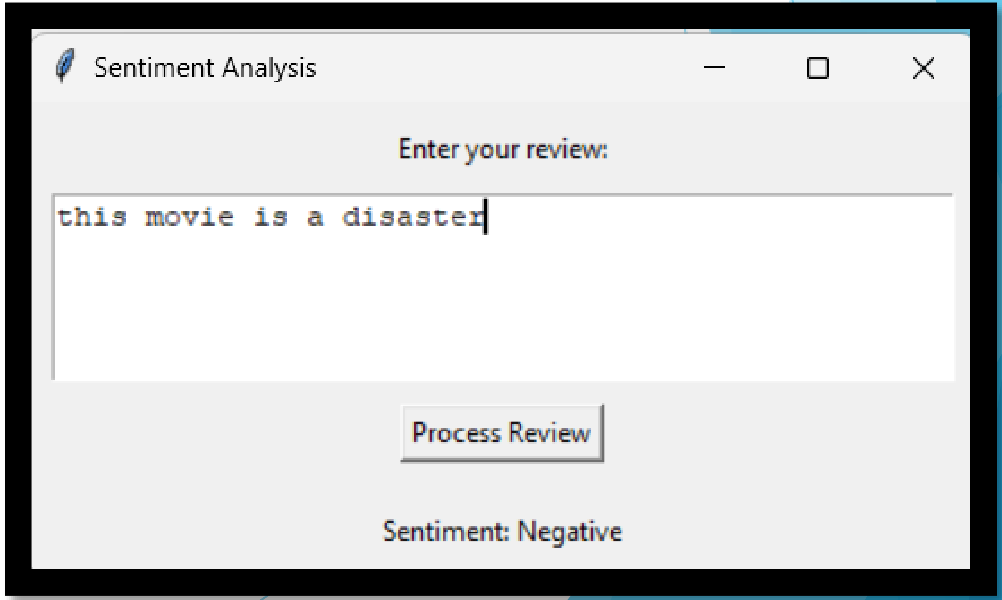
Word- cloud for positive & Negative words



Enter your review:

```
this movie is amazing!!
```

Sentiment: Positive



Enter your review:

```
this movie is a disaster
```

Sentiment: Negative

CONCLUSION

- This study utilized sentiment analysis techniques to analyze IMDb movie reviews, aiming to uncover insights into audience sentiments towards films. Through the application of [mention specific tools or algorithms], a dataset of [specify size and timeframe] reviews was categorized into positive, negative, and neutral sentiments .The analysis revealed that [summarize main findings, e.g., a majority of reviews exhibited positive sentiment, with notable variations across genres]. Specific examples, such as [mention any noteworthy case studies], underscored the effectiveness of sentiment analysis in capturing audience reception.

- The implications of these findings are significant for stakeholders in the film industry. Understanding audience sentiment can guide decisions related to marketing strategies, content creation, and audience engagement.

- However, it's important to acknowledge the limitations of this study, including [mention any data biases or limitations in the sentiment analysis approach]. Future research could explore advanced sentiment analysis techniques or incorporate additional data sources to further refine insights.

- In conclusion, this study demonstrates the utility of sentiment analysis in extracting actionable insights from IMDb movie reviews. By bridging the gap between audience perception and film industry decisions, this analysis provides a foundation for future research and application in movie analytics.

FUTURE SCOPE

- **Aspect-Based Sentiment Analysis:** Implementing aspect-based sentiment analysis would involve breaking down movie reviews into specific aspects such as plot, acting, cinematography, and directing. This approach can provide more nuanced insights into which elements of a movie contribute most positively or negatively to overall audience sentiment. By understanding these specific aspects, filmmakers and studios can pinpoint areas for improvement or capitalize on strengths in future projects.

- **Temporal Analysis and Trend Prediction:** Conducting temporal analysis involves tracking sentiment changes over time for specific movies or genres. This can reveal evolving audience perceptions post-release, seasonal variations in sentiment, or long-term trends in audience preferences. Additionally, leveraging sentiment data for trend prediction could assist in forecasting box office success or audience ratings based on early sentiment signals from pre-release reviews or trailers. This predictive capability could inform marketing strategies and resource allocation for movie releases.

- **Multilingual and Cross-Platform Analysis:** Expanding the analysis to include multilingual reviews and sentiment across various platforms (such as social media, forums, and streaming platforms) would provide a more comprehensive understanding of global audience sentiments. Multilingual sentiment analysis tools or language-specific models can be employed to capture diverse audience perspectives. Comparing sentiments across platforms can reveal differences in audience demographics and engagement patterns, offering insights into the broader impact of movies across different cultural contexts and media channels.

REFERENCES

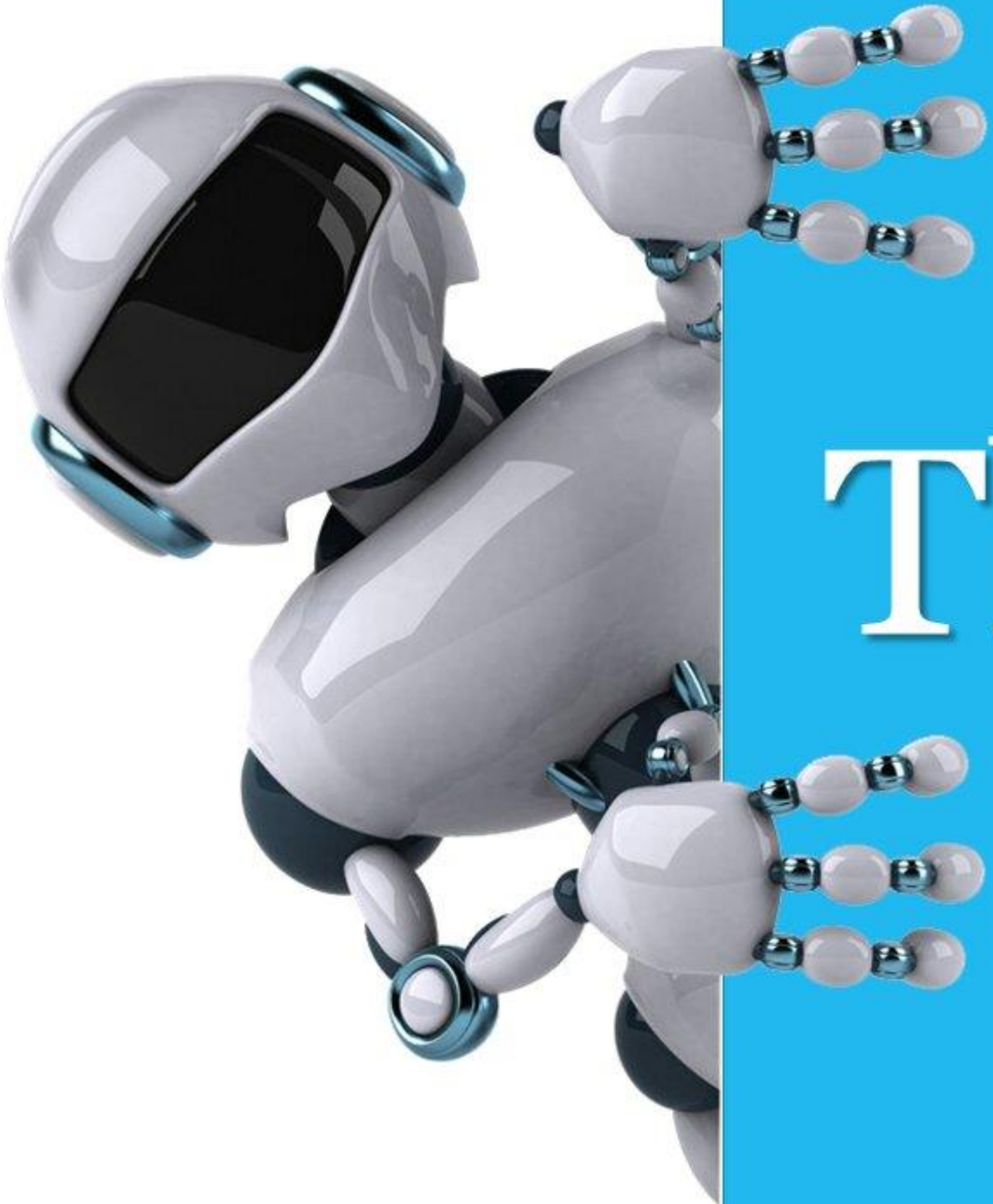
Here are references and sources that can be instrumental in developing your IMDb movie reviews analysis project, focusing on sentiment analysis and related methodologies,.

Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis." Foundations and Trends® in Information Retrieval 2.1-2 (2008): 1-135.
This foundational paper discusses the principles and techniques of sentiment analysis, providing insights into different approaches and algorithms used for sentiment classification.

Cambria, Erik, et al. "SenticNet 6: A holistic semantic resource for sentiment analysis." Proceedings of the International Conference on Language Resources and Evaluation (LREC). 2018.
SenticNet is a comprehensive semantic resource that integrates concepts and entities with sentiments, offering a structured approach to sentiment analysis beyond simple polarity classification.

Liu, Bing. "Sentiment analysis and opinion mining." Synthesis Lectures on Human Language Technologies 5.1 (2012): 1-167.
Liu's book provides a comprehensive overview of sentiment analysis techniques, including sentiment lexicons, machine learning approaches, and applications in various domains.

Maas, Andrew L., et al. "Learning word vectors for sentiment analysis." Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 2011.



Thank you