**ESPN**

*Introduction*

I undertook a project to analyze sentiment in IPL 2020 news articles using Natural Language Processing (NLP) techniques. This involved gathering data, cleaning it, performing sentiment analysis, and training a machine learning model to classify article sentiments. This document outlines the process I followed and my findings.

*Step 1: Data Collection*

I started by scraping IPL 2020 news articles from a chosen website. Using Selenium, I extracted article headlines, content, and links. Once all the relevant articles were scraped, I saved them into a CSV file.

A screenshot of a computer

AI-generated content may be incorrect.

*Step 2: Data Cleaning*

Upon inspecting the dataset, I noticed irrelevant data such as match schedules and tables. I filtered out non-news articles and focused only on articles containing IPL 2020-related content. I also handled missing values by removing empty articles.

A screenshot of a screenshot of a news article

AI-generated content may be incorrect.

*Step 3: Sentiment Analysis*

For sentiment analysis, I used two different approaches:

1. TextBlob Analysis: This method provided sentiment polarity scores, categorizing articles as positive, negative, or neutral.

2. VADER Sentiment Analysis: I used the SentimentIntensityAnalyzer from the NLTK library, which gave more refined sentiment scores.

After comparing the results, I noticed a significant skew towards positive sentiment, which suggested a bias in the dataset. To ensure more balanced results, I refined the classification approach.

A close-up of a text

AI-generated content may be incorrect.

A red and blue graph

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A white background with black text

AI-generated content may be incorrect.

A close-up of a computer screen

AI-generated content may be incorrect.

*Step 4: Machine Learning Model Training*

With the sentiment-labeled dataset, I trained multiple machine learning models:

- Logistic Regression

- Random Forest Classifier

- Support Vector Machine (SVM)

- Naive Bayes Classifier

Initially, I encountered extremely high accuracy scores (100%), which indicated overfitting due to class imbalance. To address this, I balanced the dataset using the SMOTE (Synthetic Minority Over-sampling Technique) method.

After balancing the dataset, I retrained the models and evaluated their performance using accuracy scores and classification reports.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer screen

AI-generated content may be incorrect.

*Step 5: Model Evaluation*

After balancing the dataset, the models achieved realistic accuracy scores of around 89%. This indicated that the sentiment classification was more reliable.

A screenshot of a computer

AI-generated content may be incorrect.

*Findings & Observations*

- The majority of articles had a positive sentiment, possibly due to the nature of sports reporting.

- After balancing the dataset, the models performed more effectively and classified sentiments more accurately.

- Some misclassifications occurred, which suggested that further fine-tuning of the dataset and models could improve performance.

- The logistic regression model performed best, and I saved it as the final model for future sentiment predictions.

*Conclusion*

Through this project, I successfully scraped IPL 2020 news articles, cleaned the dataset, performed sentiment analysis, and built a predictive model for sentiment classification. The insights gained indicate the overall tone of IPL-related news coverage and demonstrate the importance of balancing datasets when training models.