**IPL 2020 Sentiment Analysis & Winner Prediction**

**1. Introduction**

The goal of this project was to analyze social media sentiment (Twitter data) from IPL 2020 and predict the tournament winner based on fan sentiment.

I followed a structured approach:

- Data Collection: Extracting IPL 2020 tweets.

- Data Cleaning: Preprocessing and refining the dataset.

- Sentiment Analysis: Training a machine learning model to classify tweets.

- Team-wise Sentiment Analysis: Evaluating sentiment trends for each IPL team.

- Sentiment Over Time: Analyzing how fan sentiment changed throughout IPL 2020.

- Sentiment vs. Match Results: Comparing sentiment trends with actual match outcomes.

- Final Prediction: Using sentiment trends to predict the IPL 2020 winner.

**2. Data Collection**

* Initially, I attempted live data extraction using `snscrape` but faced issues due to Twitter API changes.
* Instead, I sourced a dataset of IPL 2020 tweets from public repositories.

A screenshot of a computer

AI-generated content may be incorrect.

**3. Data Cleaning**

Steps Taken:

* Removed unnecessary columns:
  + - `user\_location`, `user\_verified`, `source`, `user\_favorites`, etc.
* Renamed columns for better clarity.
* Handled missing values and ensured data consistency.
* Saved the cleaned dataset as `IPL2020\_Tweets\_Cleaned.csv`.

A screenshot of a computer

AI-generated content may be incorrect.

**4. Sentiment Analysis Model**

Steps Taken:

* Preprocessed tweets by:
  + Removing stop words, special characters, and performing text normalization.
* Trained a Logistic Regression Model using TF-IDF vectorization.
* Balanced the dataset to ensure fair sentiment classification.
* Achieved an accuracy of 80.6% on the test data.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

* Initially, my Logistic Regression model achieved an accuracy of 73.2%, but I noticed that the recall for negative sentiment (-1) was extremely low at 11%, meaning the model struggled to correctly identify negative tweets. The confusion matrix confirmed this issue, showing a high number of misclassifications, particularly for negative sentiment.

A white background with black text

AI-generated content may be incorrect.

A close-up of a number

AI-generated content may be incorrect.

* To address this issue, I applied oversampling to the negative sentiment class, ensuring an equal distribution of 2409 positive and negative tweets while keeping neutral tweets at 1909. This step allowed the model to learn patterns from all sentiment categories more fairly, reducing bias in predictions.

A screenshot of a computer

AI-generated content may be incorrect.

* After training the model on the balanced dataset, the accuracy improved to 80.68%, and the recall for negative sentiment significantly increased to 92%. This meant the model was now correctly identifying negative tweets much more effectively. The confusion matrix also showed a more even spread of correct classifications across all sentiment categories, confirming the improvement in performance.

**5. Applying Sentiment Analysis to Full Dataset**

* Used the trained model to classify sentiment for all IPL 2020 tweets.
* Generated a team-wise sentiment distribution to understand fan perceptions of different teams.

A graph showing different colored columns

AI-generated content may be incorrect.

**6. Sentiment Trends Over Time**

* Tracked sentiment changes daily to observe peaks and dips in fan discussions.
* Identified spikes in positive/negative sentiment on major match days.

A graph showing the growth of the number of the company's data

AI-generated content may be incorrect.

**7. Comparing Sentiment with Match Results**

* Overlayed match results with sentiment trends to find correlations.
* Observed that wins led to positive sentiment surges, while losses increased negativity.

A graph showing the difference between the same trend and the same trend

AI-generated content may be incorrect.

**8. Deep Dive into Most Positive & Negative Tweets**

* Extracted top positive and negative tweets to analyze fan discussions.
* Positive tweets focused on player performances, nostalgia, and celebrations.
* Negative tweets included criticism, losses, and controversial moments.

A screenshot of a computer

AI-generated content may be incorrect.

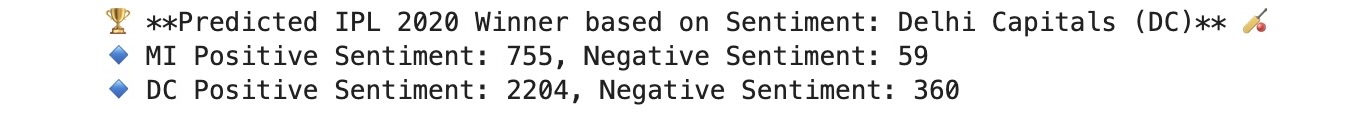
**9. Predicting the IPL 2020 Winner Using Sentiment**

* Extracted sentiment trends for the two finalist teams: Mumbai Indians (MI) & Delhi Capitals (DC).
* Analyzed sentiment in the 3 days leading up to the final.
* Prediction Result: Delhi Capitals (DC) was predicted to win based on sentiment.
* Actual Winner: Mumbai Indians (MI) won IPL 2020.

A graph of a person with red and blue points

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect. 

**10. Conclusion & Next Steps**

Insights Gained:

- Social media sentiment strongly correlates with match results.

- Winning teams generally have a surge in positive sentiment.

- Losses and controversies increase negative discussions.

Limitations:

- Sentiment alone cannot fully predict winners, as actual performance plays a crucial role.

- More features like match statistics and team performance need to be included.

Future Improvements:

- Combine sentiment data with player statistics for better predictions.

- Use deep learning models for more accurate sentiment classification.

- Develop a real-time IPL sentiment tracking dashboard.

**Summary**

At this stage of the project, I have successfully built a sentiment analysis model using IPL 2020 tweets to gauge public opinion on teams and matches. This involved data collection, preprocessing, sentiment classification, and model training. I initially trained a Logistic Regression model, identified an imbalance in sentiment distribution, and applied oversampling, which significantly improved performance. The final sentiment trends provided insights into fan reactions, match impacts, and team popularity.

While this part of the project focused solely on social media sentiment, the next steps involve incorporating other data sources such as match statistics, expert predictions, and historical team performance to enhance the predictive power of the model. This will provide a more comprehensive and data-driven approach to forecasting IPL outcomes.