"Premier League Sentiment Analysis: Fan Discussions vs. Betting Trends"

By: Arko Bhattacharya, Vishesh Gupta

Introduction: Addressing an NLP Problem

Thesis: *This project investigates the use of sentiment analysis within the sports industry to uncover correlations between fan emotions, expressed through Reddit discussions, and trends in the betting market, leveraging the platform's structured and diverse data for richer insights.*

Our analysis primarily targets social media platforms where fans, analysts, and enthusiasts actively engage in discussions about teams, players, matches, and events. The vast amount of textual data generated on these platforms offers rich insights into public sentiment, providing valuable perspectives on current team performances and the environment around that team. Our goal is to analyze this sentiment and then compare it with current betting market trends to explore potential correlations.

Our initial proposal focused on using data to leverage the platform X (i.e. Twitter) a space used for live updates and fan engagement as our primary data source however during our exploratory phase we encountered some significant challenges:

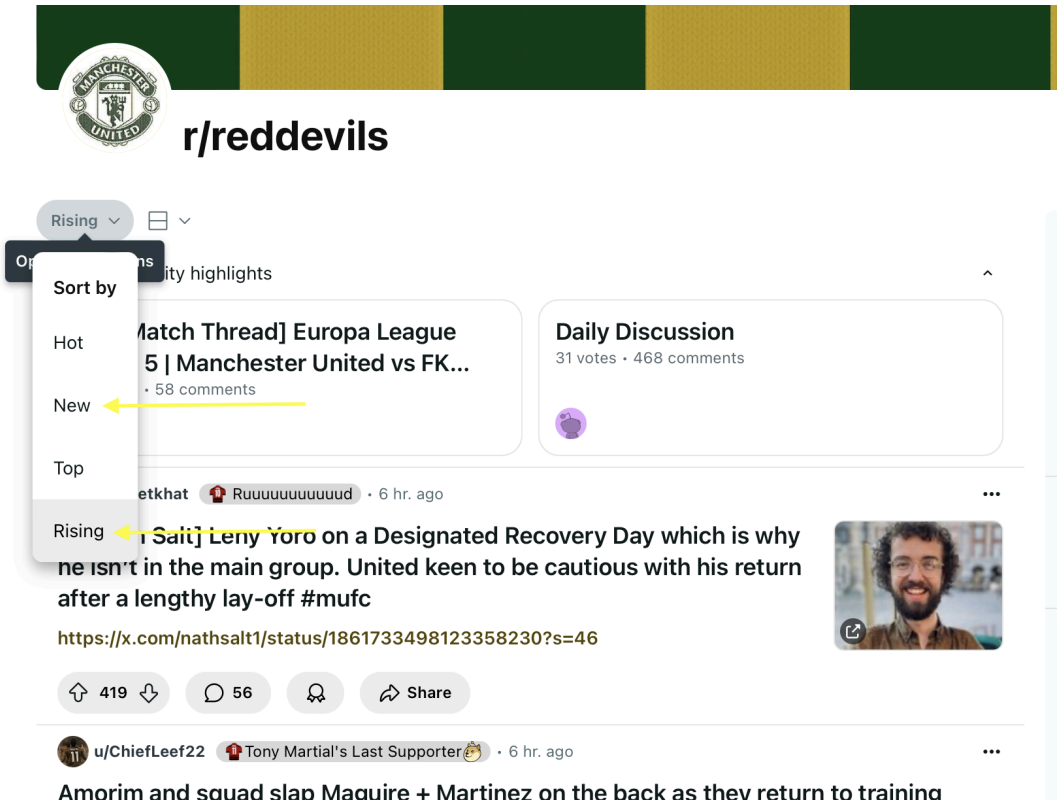
1. API Access Limitation: X's API required costly subscription. The free-tier API only allowed fetching up to 100 tweets per month, which was insufficient for meaningful analysis.
2. Scraping Restrictions: X's updated policies made web scraping legally ambiguous and technically challenging .

Due to this limitation we decided to shift to Reddit as our data source. Reddit is a platform rich in in-depth discussions hosted across organised communities called subreddits. This gave us a platform that had:

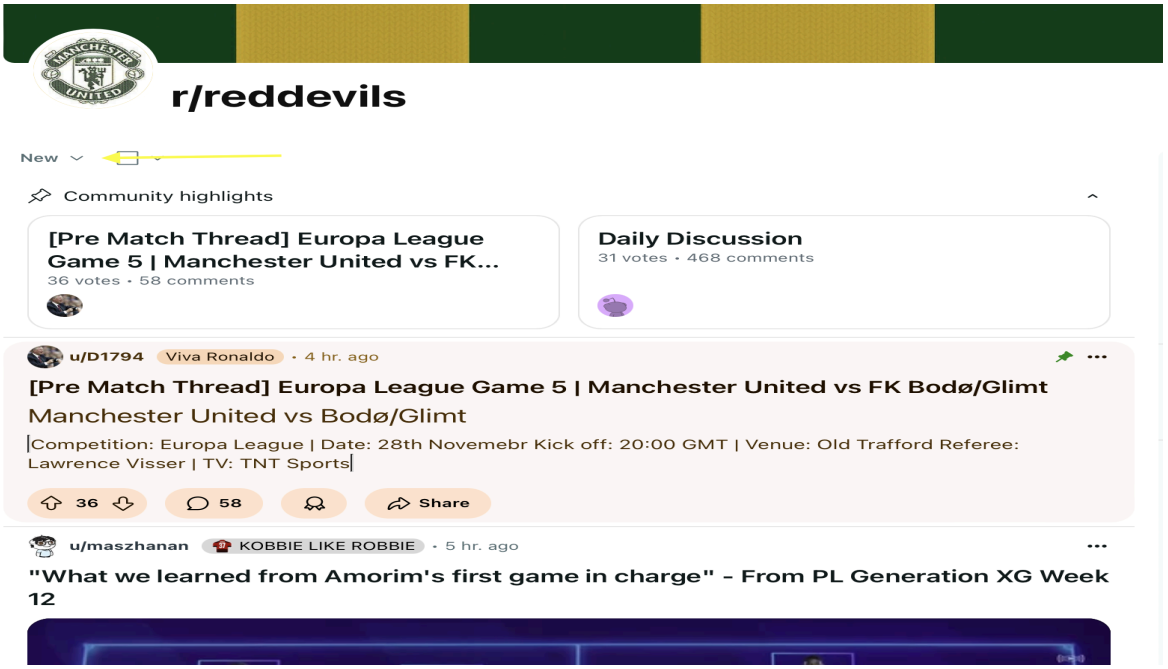
1. Well organised Data: Subreddits are dedicated to specific topics, such as sports teams or leagues, making data collection more straightforward compared to Twitter. For example, subreddits like r/Arsenal and r/LiverpoolFC naturally align with our research goals.
2. Diverse Perspectives: Reddit discussions often involve a mix of fan, analyst, and casual observer opinions, providing a more nuanced dataset for sentiment analysis.
3. Granular Insights: The platform allows for sentiment analysis based on specific topics or events, such as match results or player trades.

By shifting to Reddit we retained the project's original objectives while addressing the limitations we faced with X (i.e. Twitter)

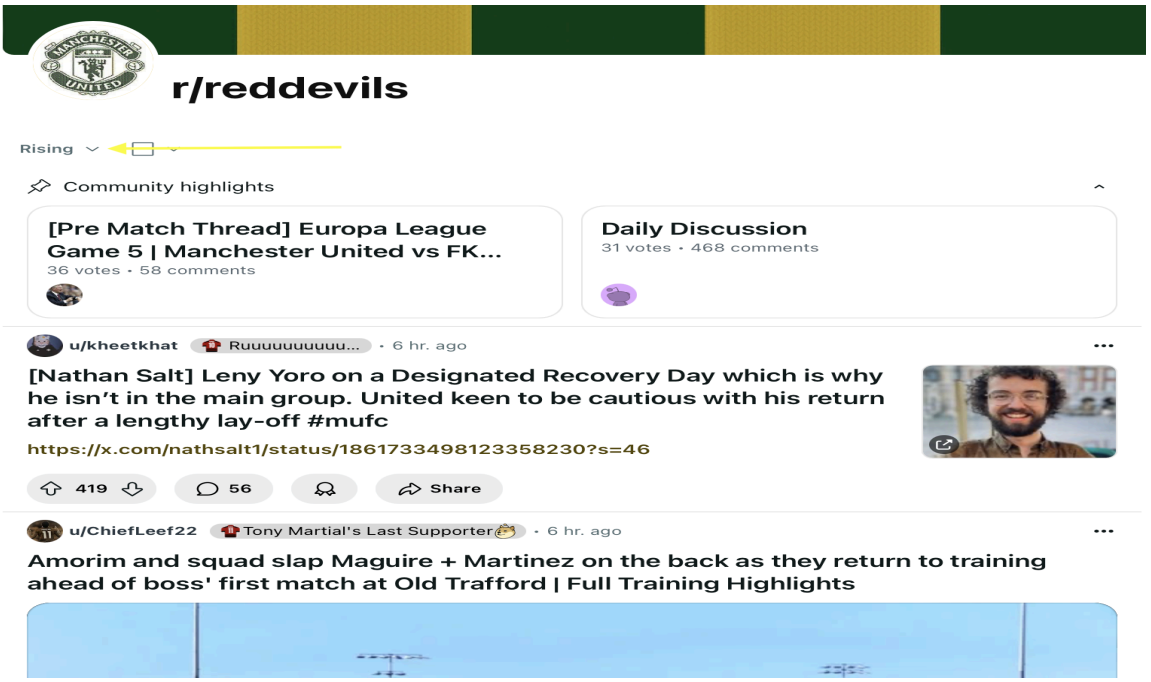
To harness Reddits full potential we used an open api to design a data collection strategy that focuses on two main categories:



1. New Post:
    a. Contain the most recently posted, raw, and unfiltered discussions.
    b. Provide real-time sentiment data, reflecting how narratives evolve or take shape within sports communities.



2. Rising Posts:
    a. Represent discussions gaining traction but not yet widely visible.
    b. Offer insights into emerging sentiments, such as initial reactions to match results, player transfers, or controversies.
    c. Allow the model to capture the pulse of newly forming opinions.



Identify or Invent a Solution Based on a Language Model

Our project approach involves utilizing VADER (Valence Aware Dictionary and sEntiment Reasoner), a lexicon and rule-based tool designed to analyze sentiment in text that will generate sentiment labels for our New Posts dataset. The data was then preprocessed extensively to remove noise such as stopwords, URLs, numbers, and punctuation. Lemmatization reduces words to their root forms for consistency.

After initial analysis we then used the newly labelled dataset to train and validate three different models: Bernoulli Naive Bayes and Logistic Regression were trained using the TF-IDF feature matrix while DistilBERT was fine-tuned with tokenized sequences and trained using labeled sentiment data. We evaluated each model on a held-out test set, calculating metrics such as accuracy, precision, recall, F1-score, and AUC-ROC to identify the best-performing model. After identifying the best model, we applied it to live "rising" Reddit data to predict sentiment in real-time. This dual-focus approach allows us to capture both real-time opinions and emerging trends, resulting in a richer and more balanced dataset.

We believe that our novelty lies in integrating traditional machine learning models with a state-of-the-art language model (DistilBERT) in a comparative framework, while leveraging a specialized Reddit dataset created from comment-reply pairs. This approach ensures comprehensive sentiment analysis that is both scalable and capable of capturing linguistic nuances in Reddit-specific text.

Model Description:

1. Bernoulli Naive Bayes (BNB)

- How it works: BNB is a probabilistic model based on Bayes' theorem. It assumes that the presence or absence of a specific word in a document contributes independently to the sentiment. This binary independence assumption simplifies computations and allows the model to handle sparse datasets efficiently.
- Application: We vectorized the text using a TF-IDF representation, which captures the importance of terms across the dataset. BNB was trained on the transformed text to predict whether a comment was positive (1) or negative (0).

2. Logistic Regression (LR)

- How it works: Logistic regression is a linear model that predicts probabilities using the sigmoid function. It calculates the relationship between the input features (TF-IDF vectors) and a binary target variable (positive/negative sentiment).
- Application: Similar to BNB, LR utilized the TF-IDF representation of the text. We tuned hyperparameters such as regularization strength (C) to improve model performance.

3. DistilBERT (Transformer-Based Model)

- How it works: DistilBERT, a “knowledge distilled” version of the BERT model, uses a Transformer architecture for understanding contextual relationships in text. It processes sequences of tokens and learns deep representations of text using a pre-trained language model fine-tuned on specific tasks.

- Application: We used the DistilBertTokenizer for tokenizing text and DistilBertForSequenceClassification for sentiment analysis. We compared both the off-the shelf version and a fine-tuned version to compare results. The fine-tuned version was trained on the Reddit dataset using a training-validation split, with hyperparameters like batch size, learning rate, and number of epochs optimized using the Hugging Face Trainer API.

Training and Application on Initial Data

To determine the most suitable model for our final dataset, we trained and evaluated Bernoulli Naive Bayes (BNB), Logistic Regression (LR), and DistilBERT (a transformer-based model). The dataset consisted of the latest Reddit posts for all 20 Premier League teams, with sentiment labels (positive and negative) generated using VADER sentiment analysis and manual review as described above. For each model, we conducted a quantitative evaluation using metrics such as the confusion matrix, accuracy, and ROC curves.

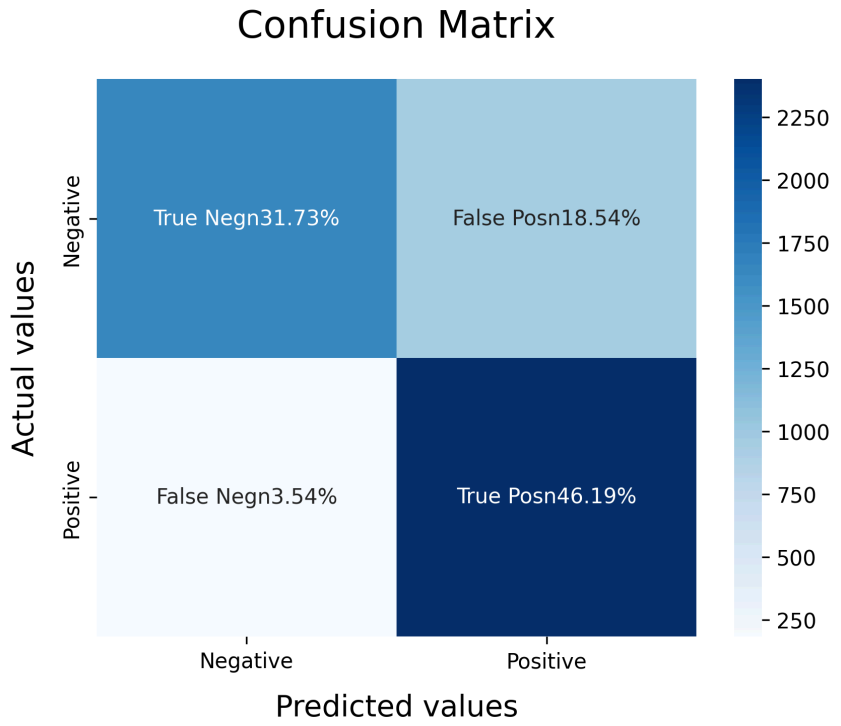
Model Evaluation:

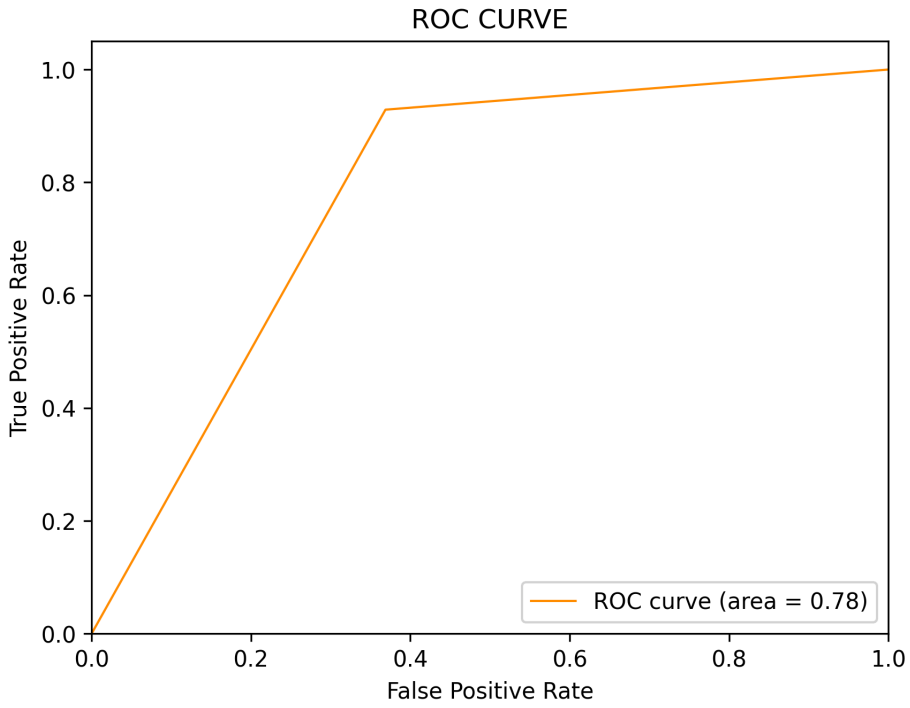
1. Bernoulli Naive Bayes (BNB)

```
No. of feature_words:  267474
              precision    recall  f1-score   support

           0       0.90      0.63      0.74      2614
           1       0.71      0.93      0.81      2586

    accuracy                           0.78      5200
   macro avg       0.81      0.78      0.77      5200
weighted avg       0.81      0.78      0.77      5200
```





The Bernoulli Naive Bayes (BNB) model demonstrates a moderate level of performance in classifying the given dataset. Its confusion matrix highlights that the model correctly predicts the majority of both positive and negative cases, with True Positives accounting for 46.19% and True Negatives at 31.73%. However, it also exhibits some errors, with 18.54% of negative cases incorrectly classified as positive (False Positives) and 3.54% of positive cases incorrectly classified as negative (False Negatives). These errors suggest that while the model is effective at identifying most cases, there is some room for improvement in reducing misclassifications.

The model's accuracy is approximately 77.92%, meaning it provides correct predictions for nearly 78% of the cases overall. This is a respectable level of performance, although improvements may be possible through feature engineering or model tuning.

The ROC curve analysis further supports the model's effectiveness, with an Area Under the Curve (AUC) score of 0.78. This score reflects the model's good ability to distinguish between positive and negative cases across various classification thresholds. The curve indicates that the model maintains a relatively high True Positive Rate while keeping the False Positive Rate manageable, showing a balanced performance.

Strength:

[goat, fc] -> positive stayed positive

['hazard','injury','bad','attitude','ruined','not','age','tempting','bread','hidden','cupboard','covid','injury','burger','king'] -> negative stayed negative

Weaknesses:

[ga, ga, no, freak] -> was given a negative but the model marked as positive

['turnaround','unbelievable','defending','full','faith','won','t','beaten','energy','ridiculous','lung','busting','run','th','minute'] -> was given positive but the model marked it as negative

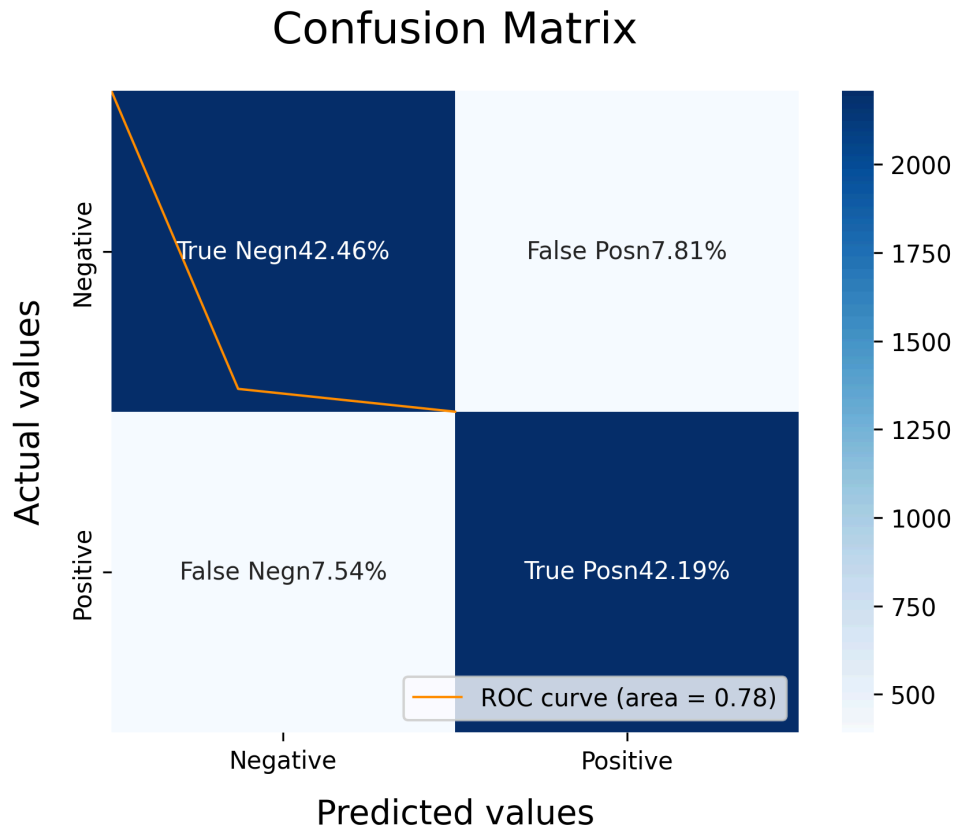
In summary, the BNB model is a reliable starting point for classification tasks, with solid predictive power and decent differentiation capabilities. However, to enhance its performance, efforts could focus on addressing the False Positive Rate and further refining the model.

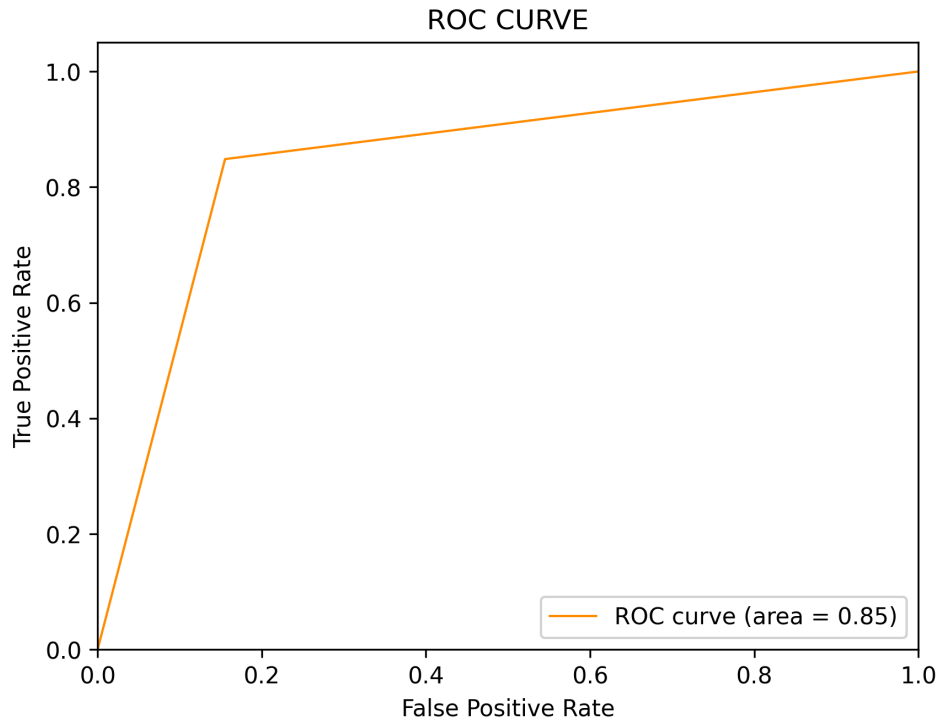
2. Logistic Regression (LR)

```
No. of feature_words:  267474
              precision    recall  f1-score   support

           0       0.85      0.84      0.85      2614
           1       0.84      0.95      0.84      2586

    accuracy                           0.85      5200
   macro avg       0.85      0.85      0.85      5200
weighted avg       0.85      0.85      0.85      5200
```





The Logistic Regression model demonstrates a strong level of performance in classifying the given dataset. Its confusion matrix shows that the model correctly predicts the majority of both positive and negative cases, with True Positives accounting for 42.19% and True Negatives at 42.46%. However, it also exhibits some errors, with 7.81% of negative cases incorrectly classified as positive (False Positives) and 7.54% of positive cases incorrectly classified as negative (False Negatives). These misclassifications are relatively low, suggesting that the model is effective in balancing predictions for both classes.

The model's accuracy is approximately 84.65%, meaning it provides correct predictions for nearly 85% of the cases overall. This high level of accuracy reflects the model's strong ability to generalize to the dataset, although further enhancements might still be possible with additional feature engineering or tuning.

The ROC curve analysis further supports the model's effectiveness, with an Area Under the Curve (AUC) score of 0.85. This score indicates excellent discriminatory power, highlighting the model's ability to distinguish between positive and negative cases across various classification thresholds. The curve suggests that the model maintains a high True Positive Rate while keeping the False Positive Rate relatively low, demonstrating a robust performance.

Strength:

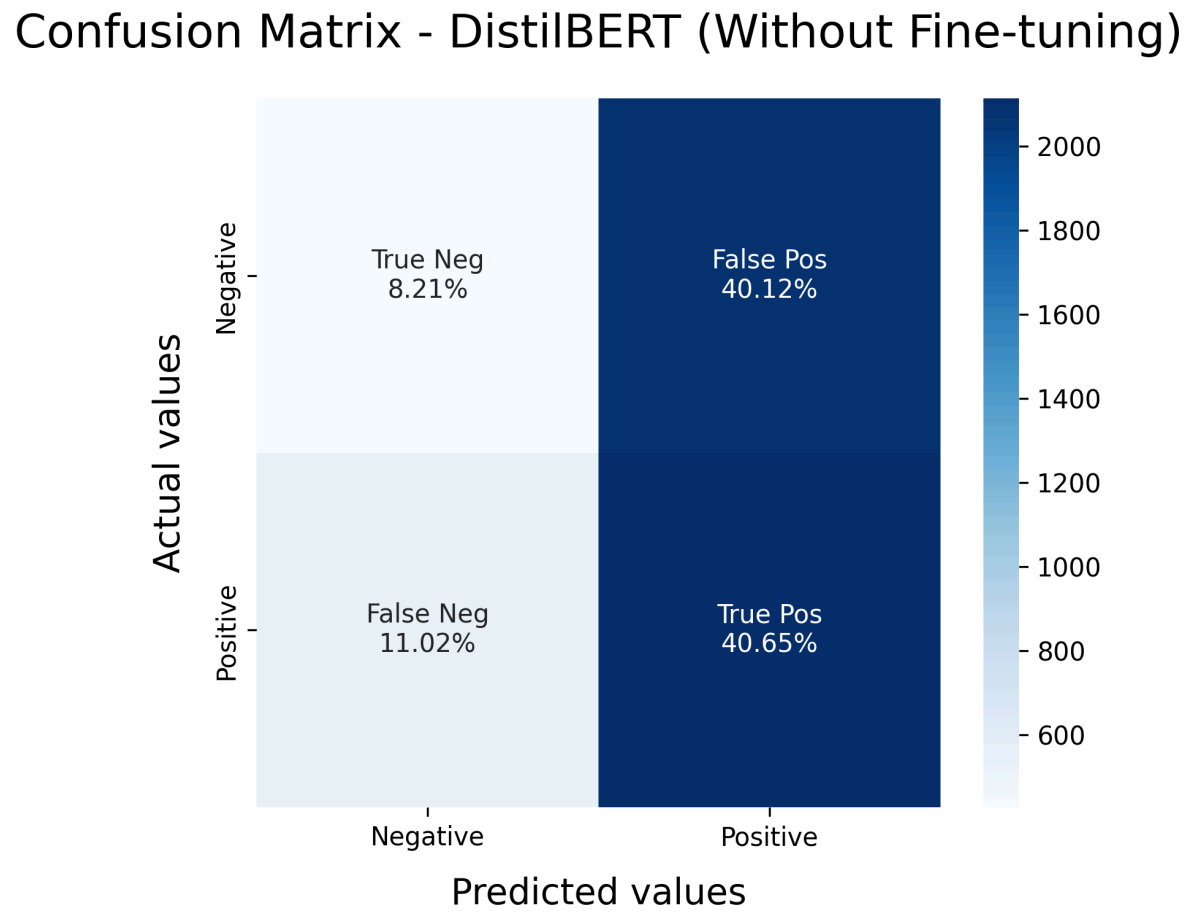
[ga, ga, no, freak] -> is now marked correctly as positive showing improvement in accuracy

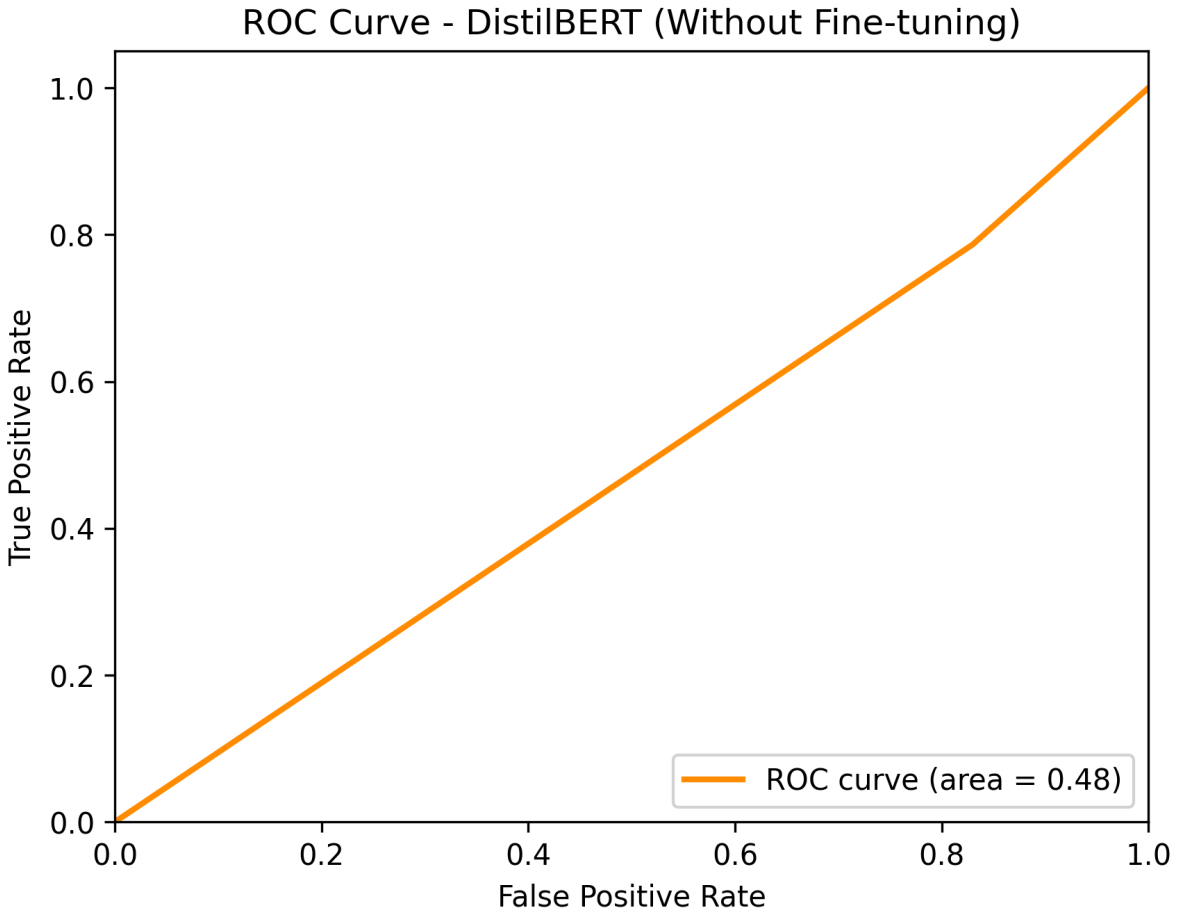
Weaknesses:
The model continue to show the below incorrectly
['turnaround','unbelievable','defending','full','faith','won','t','beaten','energy','ridiculous','lung','busting','run','th','minute']-> was given positive but the model marked it as negative

In summary, the Logistic Regression model is a reliable choice for classification tasks, with strong predictive power and excellent differentiation capabilities. While its performance is already impressive, efforts to further minimize False Positive and False Negative rates could make the model even more robust for specific applications.

3. DistilBERT (Transformer-Based Model) - Pre trained

	precision	recall	f1-score	support
Negative	0.45	0.28	0.34	2513
Positive	0.50	0.68	0.58	2687
accuracy			0.48	5200
macro avg	0.47	0.48	0.46	5200
weighted avg	0.47	0.48	0.46	5200





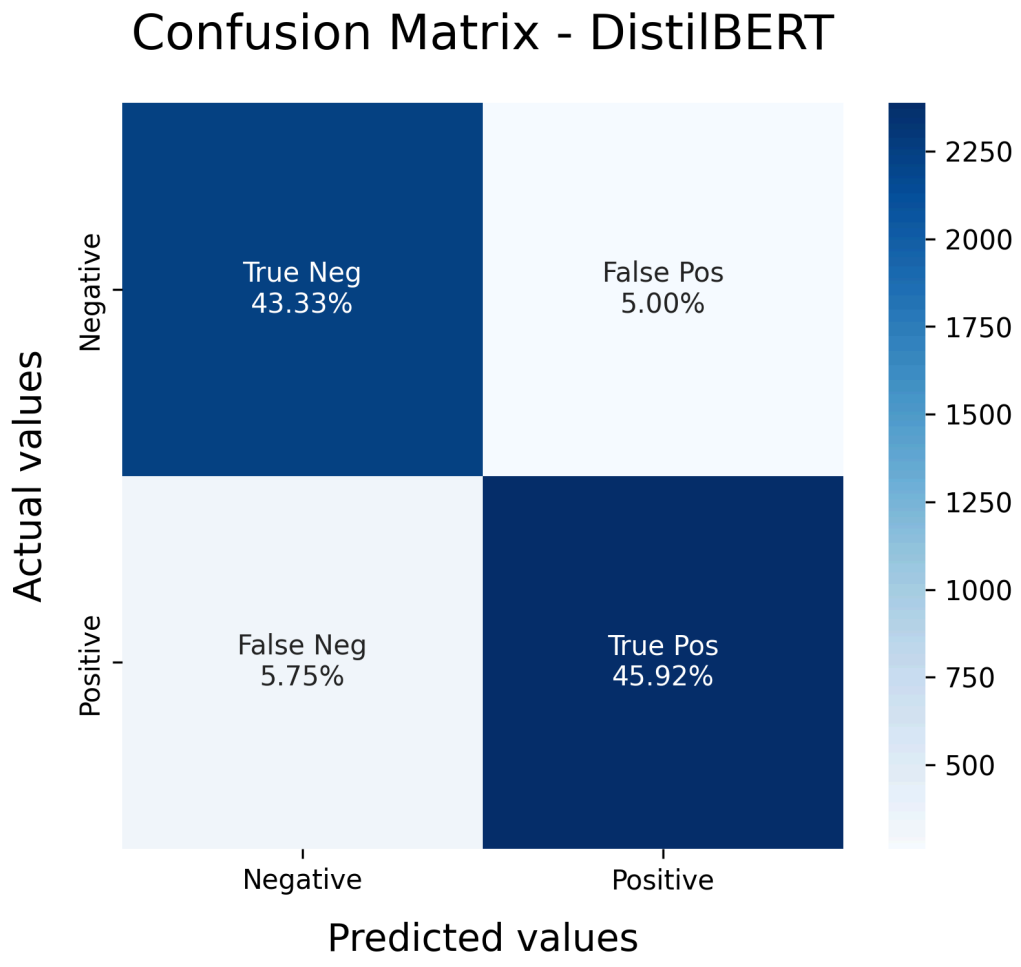
The off-the-shelf model shows a moderate level of performance in classifying the given dataset. Its confusion matrix reveals that the model correctly predicts the majority of positive cases, with True Positives accounting for 40.65%, but struggles significantly with the negative cases, with True Negatives at only 8.21%. This imbalance suggests that the model is more effective at identifying positive cases but has difficulty classifying negatives.

The model's accuracy, derived from these predictions, is approximately 48%, indicating that it correctly predicts just under half of the cases overall. This relatively low accuracy suggests that the model's performance is not optimized, and it may be biased toward predicting the positive class, leading to a higher number of False Positives and False Negatives.

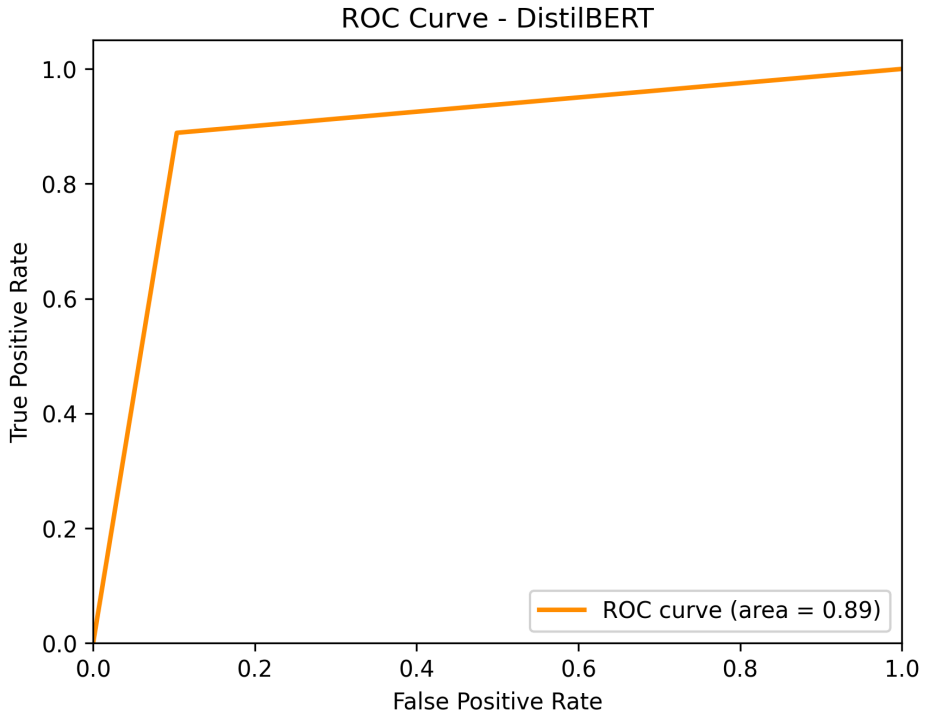
The ROC curve analysis further reflects the model's limited effectiveness, with an Area Under the Curve (AUC) score of 0.48. This score indicates that the model has poor discriminatory power, performing close to random guessing when distinguishing between positive and negative cases. The curve demonstrates that the model struggles to maintain a high True Positive Rate while simultaneously keeping the False Positive Rate low, indicating significant room for improvement in both class identification and overall performance.

4. DistilBERT (Transformer-Based Model) - Fine Tuned

	precision	recall	f1-score	support
Negative	0.89	0.89	0.89	2513
Positive	0.90	0.89	0.90	2687
accuracy			0.89	5200
macro avg	0.89	0.89	0.89	5200
weighted avg	0.89	0.89	0.89	5200



ROC Curves:



The fine-tuned DistilBERT model demonstrates a much stronger level of performance in classifying the given dataset. Its confusion matrix shows that the model correctly predicts the majority of both positive and negative cases, with True Positives accounting for 45.92% and True Negatives at 43.33%. However, it also exhibits some errors, with 5.00% of negative cases incorrectly classified as positive (False Positives) and 5.75% of positive cases incorrectly classified as negative (False Negatives). These misclassifications are relatively low, suggesting that the model effectively balances predictions for both classes.

The model's accuracy, derived from these percentages, is approximately 89.25%, indicating that it provides correct predictions for nearly 90% of the cases overall. This high level of accuracy reflects the model's strong ability to generalize to the dataset while leaving room for optimization through further fine-tuning.

The ROC curve analysis further supports the model's effectiveness, with an Area Under the Curve (AUC) score of 0.89. This score signifies excellent discriminatory power, highlighting the model's ability to distinguish between positive and negative cases across various classification thresholds. The curve demonstrates that the model maintains a high True Positive Rate while keeping the False Positive Rate relatively low, reinforcing its robust performance.

Weaknesses:
['turnaround','unbelievable','defending','full','faith','won','t','beaten','energy','ridiculous','lung','busting','run','th','minute'] -> was given positive but the model marked it as negative still

In summary, the DistilBERT model is a reliable choice for classification tasks, with strong predictive power and excellent differentiation capabilities. While its performance out-of-the-box was not as

strong, fine-tuning on the specific dataset significantly improved its effectiveness, making it a robust choice for this task.

After evaluating all of our three models—BNB, LR, and DistilBERT—we decided to proceed with DistilBERT for the final implementation. The decision was based on the following key factors:

- Overall Performance: After fine-tuning, DistilBERT outperformed both BNB and LR across all key metrics. It demonstrated the highest ability to correctly classify both positive and negative cases while maintaining excellent discriminatory power, reflected by its AUC score of 0.89. In comparison, the BNB and LR models showed lower performance, with BNB struggling with the balance between precision and recall, and LR showing a significant bias toward predicting the positive class.
- Low Misclassification Rates: After fine-tuning, DistilBERT showed the lowest rates of False Positives (5.00%) and False Negatives (5.75%) among the three models, indicating a better balance in predictions and fewer errors. This makes it particularly suitable for applications where minimizing incorrect classifications is critical, such as in medical diagnosis or financial fraud detection.
- Superior Adaptability: DistilBERT's performance improved substantially after fine-tuning, highlighting its adaptability and ability to handle the specific nuances of the dataset. While BNB and LR performed respectably, their higher error rates and relatively lower AUC scores indicated limitations in effectively handling the dataset's complexities.
- Out-of-the-Box BERT Performance: The DistilBERT model's out-of-the-box performance was not as impressive compared to its fine-tuned results. Similarly, the BERT model, although powerful, showed slower processing times and struggled with precision and recall for the negative class, making it less effective in this particular case.

Given these factors, DistilBERT's superior accuracy, discriminatory power, and adaptability after fine-tuning make it the most suitable model for this task. Its fine-tuned version provides a robust, efficient solution for classification with fewer errors and better overall performance, making it the clear choice for the final implementation.

Application to Final Data

Using DistilBERT with fine-tuned parameters, we conducted a sentiment analysis on “Rising Posts” across Reddit subreddits for all 20 Premier League teams. The goal was to explore how fan sentiment correlated with FanDuel money line values, which represent the likelihood of a team winning according to betting markets. This analysis aimed to uncover patterns and discrepancies between fan enthusiasm and betting odds, offering insights into potential biases or unique market perceptions.

Findings

1. Strong Positive Sentiment and Favorable Odds:
    - Teams such as Brighton (Sentiment Score: 0.8207) and Bournemouth (0.7839) displayed notably strong positive sentiment among fans. This enthusiasm was reflected in their relatively favorable betting odds, suggesting alignment between fan optimism and market confidence in these teams’ performances.
    - This correlation implies that fan sentiment, in some cases, may act as a rough proxy for market expectations or on-field performance.
2. Weaker Sentiment and Higher Odds:
    - On the other end of the spectrum, teams like Everton (0.6313) and Fulham (0.568) demonstrated weaker sentiment. These subdued sentiment scores aligned closely with higher betting odds, reflecting a lack of confidence in their ability to win among both fans and betting markets.
    - The consistency between lower sentiment and unfavorable odds highlights the potential predictive power of fan sentiment in assessing perceived team strength.
3. Notable Discrepancies:
    - Chelsea emerged as a particularly interesting case. Despite exhibiting relatively high fan sentiment, their betting odds were more moderate, reflecting caution in the markets. This divergence could indicate overconfidence among fans or differences in how fans and betting markets perceive the team’s true potential.
    - Such discrepancies might stem from various factors, including historical biases, overestimation of new signings or tactical changes, or unmeasured variables like injuries or schedule difficulty.

Implications:
The analysis reveals that while fan sentiment often aligns with betting odds, outliers like Chelsea suggest that fan-driven narratives can deviate significantly from market predictions. Such deviations may present opportunities for bettors or analysts to identify market inefficiencies or explore deeper reasons behind mismatches.

Future work could involve expanding the analysis to incorporate additional variables such as player statistics, injury reports, or historical trends to further contextualize these findings.

Pros and Cons of the Model

Pros

1. Quality/Correctness:
    - High Accuracy: The DistilBERT model demonstrates strong performance in classifying sentiment due to its robust contextual understanding.
    - Generalizability: Fine-tuning on both synthetic and real-world datasets ensures adaptability to various text domains, including sports sentiment analysis.
    - Handles Ambiguity: Captures nuanced language such as sarcasm or sports-specific phrases better than simpler models, especially when complemented with domain-specific data.
2. Data Requirements:
    - Scalability: The model performs well even with moderate amounts of training data, thanks to pre-trained weights, reducing the dependency on large labeled datasets.
    - Synthetic Data Flexibility: Initial fine-tuning on synthetic data allows for better adaptation to real-world challenges.
3. Time and Computational Requirements:
    - Efficiency: As a distilled version of BERT, it reduces computational overhead by nearly 40%, making it faster for both training and inference compared to its larger counterparts.
    - Optimization-Friendly: Pre-trained architecture requires fewer epochs to converge during fine-tuning.
4. Interpretability:
    - Token Attention Visualizations: Attention mechanisms in DistilBERT allow partial interpretability by identifying key tokens or phrases influencing predictions.
    - Class Probabilities: Outputs include confidence scores, providing insights into the model's certainty about its predictions.

Cons

1. Quality/Correctness:
    - Misclassifications in Edge Cases: Struggles with complex linguistic structures, such as sarcasm or subtle contextual shifts (e.g., "He’s doing great—on the bench").
    - Bias: Pre-trained on general corpora, it may inherit biases unrelated to sports, requiring careful data preparation to mitigate.
2. Data Requirements:
    - Domain Adaptation: Requires domain-specific fine-tuning to fully adapt to sports discussions, which might necessitate significant labeled data for optimal performance.
    - Noise in Real Data: Social media and sports commentary often contain typos, slang, and emojis, requiring extensive preprocessing.
3. Time and Computational Requirements:
    - Resource Demands for Fine-Tuning: Although lighter than BERT, training DistilBERT still requires access to GPUs for efficient fine-tuning, which might be a limitation for smaller teams.
    - Inference Speed for Large Datasets: For applications requiring real-time analysis of large volumes of text, inference can still be a bottleneck compared to simpler models.
4. Interpretability:

- Black Box Nature: While attention mechanisms offer some insights, the model's deep neural architecture makes it challenging to fully understand or explain decisions.
- Difficulty in Error Analysis: Misclassifications often lack clear explanations, making debugging non-trivial.

You can find all our source code and information in our repository here