Super Bowl Twitter Data Analysis

ECE219 - Project 4

March 14, 2025

1 Introduction

Social media platforms, particularly Twitter, have become invaluable sources of real-time public sentiment and behavior during major sporting events. The 2015 Super Bowl XLIX between the New England Patriots and the Seattle Seahawks was one of the most-watched television events in U.S. history, generating massive engagement across social media. This report presents a comprehensive analysis of Twitter data collected during this event, examining both basic statistical patterns and developing predictive models through multi-task learning.

Our analysis is structured in two main parts. First, we conduct exploratory data analysis to understand the temporal patterns and engagement metrics across different hashtags related to the Super Bowl. Second, we develop a multi-task learning framework to simultaneously predict team affiliation, tweet influence, and posting time period based on various features extracted from the tweets.

2 Dataset Overview

Our analysis utilizes a comprehensive Twitter dataset collected during Super Bowl XLIX, encompassing over 2.8 million tweets across six different hashtags. The largest portion of the data comes from #SuperBowl with 1,213,813 tweets, followed by #sb49 with 743,649 tweets, and #patriots with 440,621 tweets. The dataset also includes tweets from #NFL (233,022), #gohawks (169,122), and #gopatriots (23,511). Each tweet is stored as a JSON object containing metadata such as posting time, author information, and engagement metrics.

3 Basic Statistical Analysis

3.1 Temporal Distribution of Tweets

We analyzed the temporal distribution of tweets for each hashtag to understand engagement patterns throughout the event. Figures **??** through **??** show the hourly tweet patterns for all monitored hashtags.

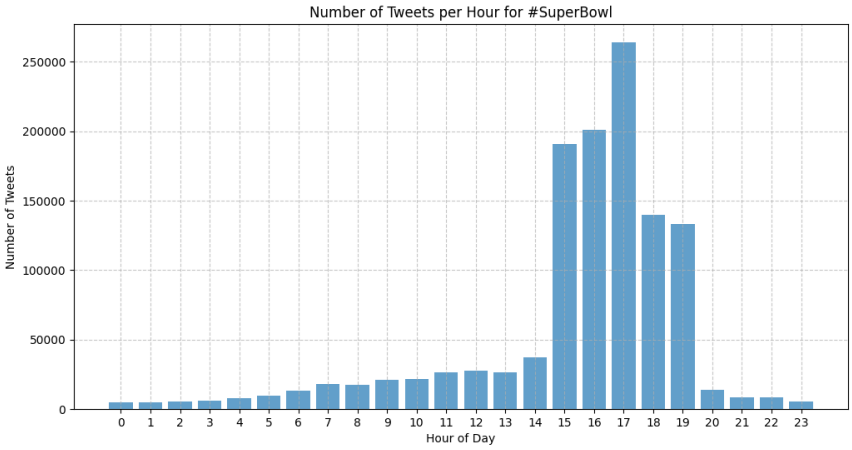

Figure 1: Hourly tweet distribution for the #SuperBowl hashtag.

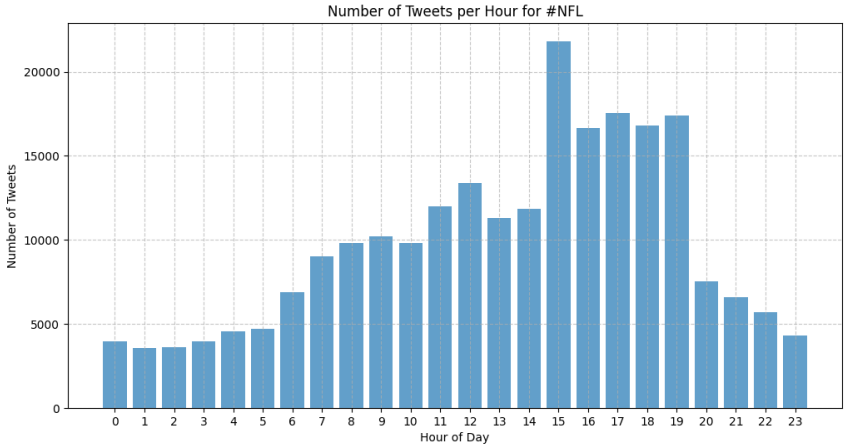

Figure 2: Hourly tweet distribution for the #NFL hashtag.

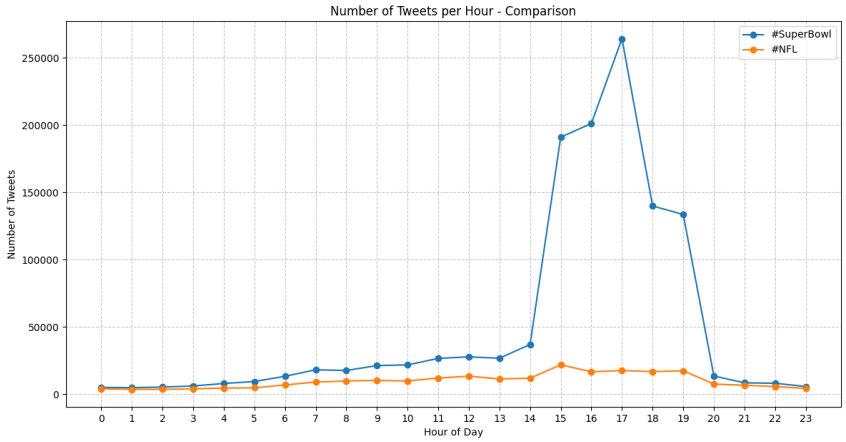

Figure 3: Comparative hourly tweet counts for #SuperBowl and #NFL hashtags.

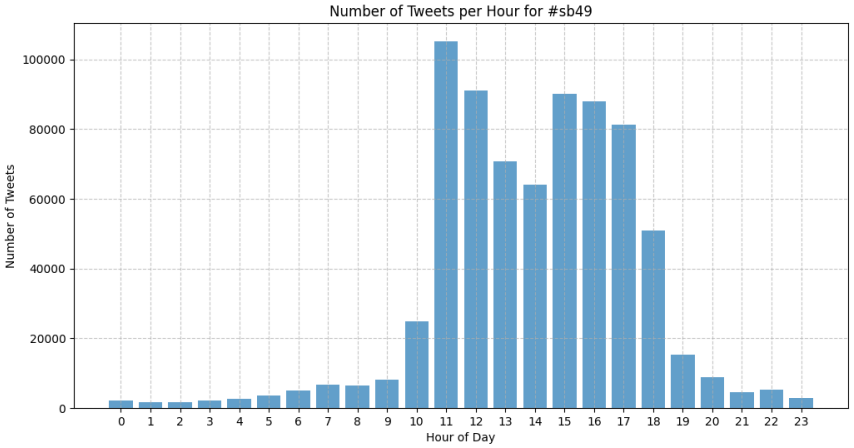

Figure 4: Hourly tweet distribution for the #sb49 hashtag.

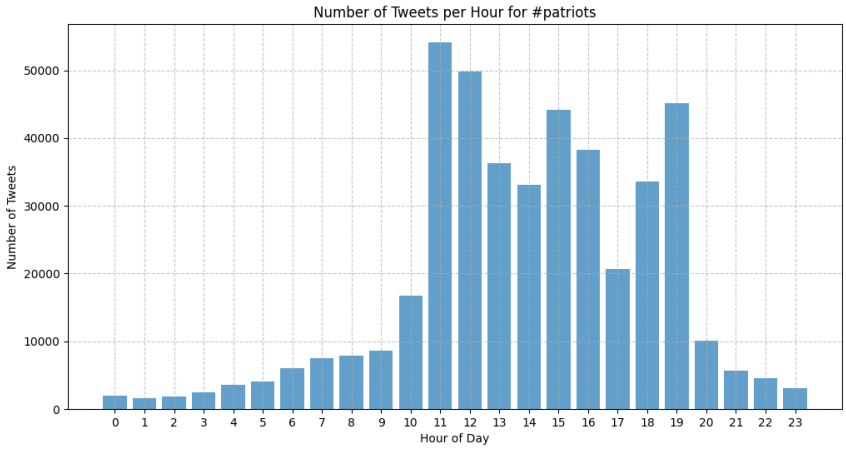

Figure 5: Hourly tweet distribution for the #patriots hashtag.

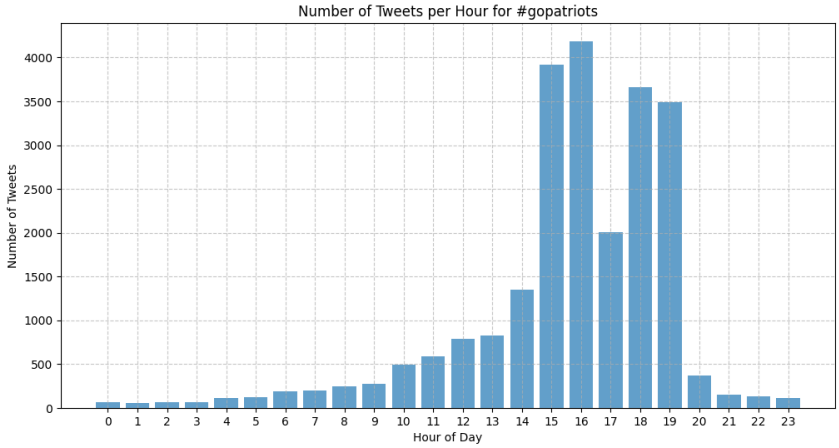

Figure 6: Hourly tweet distribution for the #gopatriots hashtag.

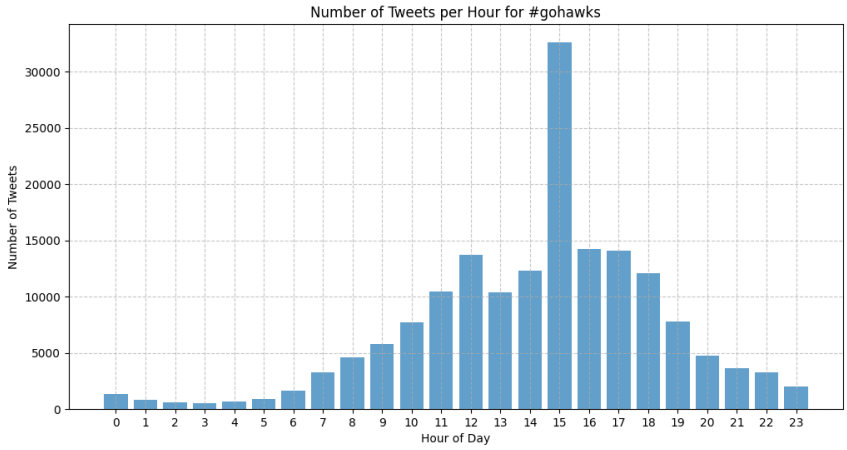

Figure 7: Hourly tweet distribution for the #gohawks hashtag.

The temporal analysis revealed distinct patterns across different hashtags. As shown in Figure **??**, the #SuperBowl hashtag demonstrated the highest overall engagement, peaking at 264,093 tweets at 17:00 EST. Figure **??** reveals that the #sb49 hashtag had a unique early peak of 105,212 tweets at 11:00 EST, indicating significant pre-game discussion using this alternative event tag. The comparison in Figure **??** highlights the dramatic difference in volume between event-specific and general NFL discussion.

3.2 Team-Specific Hashtag Patterns

Figures **??**, **??**, and **??** reveal interesting patterns in team-specific hashtag usage. The #patriots hashtag (Figure **??**) showed consistently high engagement throughout the day with an early morning peak of 54,130 tweets at 11:00 EST, while #gohawks (Figure **??**) demonstrated a mid-game peak of 32,609 tweets at 15:00 EST. The #gopatriots hashtag (Figure **??**) showed lower overall volume but peaked at 4,185 tweets at 16:00 EST, coinciding with key game moments.

3.3 Key Metrics for Each Hashtag

Table **??** presents comprehensive statistics for each hashtag, including peak hour counts, timing, and daily totals.

Hashtag	Peak Hour Tweet Count	Peak Time (EST)	Daily Total
#SuperBowl	264,093	17:00	1,213,813
#sb49	105,212	11:00	743,649
#patriots	54,130	11:00	440,621
#NFL	21,816	15:00	233,022
#gohawks	32,609	15:00	169,122
#gopatriots	4,185	16:00	23,511

Table 1: Comprehensive statistics for each hashtag.

3.4 Temporal Pattern Analysis

Our analysis of all figures revealed three distinct phases with unique characteristics for each hashtag:

- **Pre-game Phase (06:00-11:00 EST)**:
  - #sb49 (Figure **??**): Highest morning engagement with 105,212 tweets at 11:00
  - #patriots (Figure **??**): Strong early activity with 54,130 tweets at 11:00
  - #SuperBowl (Figure **??**): Gradual increase from 13,303 tweets at 06:00 to 26,595 at 11:00
  - #gohawks (Figure **??**): Steady increase from 1,662 to 10,475 tweets
- **Game Time (12:00-19:00 EST)**:
  - #SuperBowl: Dramatic rise to peak of 264,093 tweets at 17:00, as clearly shown in Figure **??**
  - #sb49: High activity during early game (91,090 at 12:00) followed by decline
  - #patriots: Consistent high volume ranging from 33,128 to 49,809 tweets
  - #gohawks: Maximum activity of 32,609 tweets at 15:00, visible in Figure **??**
- **Post-game Phase (20:00-23:00 EST)**:
  - All hashtags showed rapid decline as evident across all figures
  - #SuperBowl: From peak to 13,533 tweets at 20:00, further dropping to 5,676 by 23:00
  - #patriots: Maintained highest post-game activity (10,075 at 20:00)
  - #NFL (Figure **??**): Most stable post-game discussion (7,556 at 20:00 to 4,303 at 23:00)

3.5 Comparative Hashtag Analysis

Detailed analysis of all figures revealed several significant patterns:

- **Official Event Tags**:
  - #SuperBowl and #sb49 combined for 1,957,462 total tweets
  - Different peak times visible in Figures **??** and **??**: #SuperBowl (17:00) vs #sb49 (11:00)
  - #SuperBowl showed stronger game-time engagement (264,093 at 17:00)
  - #sb49 showed stronger pre-game engagement (105,212 at 11:00)
- **Team-Specific Tags**:
  - Patriots-related tags (#patriots, #gopatriots) totaled 464,132 tweets
  - Seahawks-related tag (#gohawks) reached 169,122 tweets
  - #patriots showed unique distribution with high activity across all phases (Figure **??**)
  - #gopatriots peaked at 16:00 with 4,185 tweets, coinciding with key game moments (Figure **??**)
- **Hourly Distribution Patterns**:
  - Morning (06:00-11:00): #sb49 and #patriots dominated, as shown in Figures **??** and **??**
  - Early game (12:00-15:00): Balanced activity across all hashtags
  - Peak game (16:00-19:00): #SuperBowl dominated with 738,421 total tweets during this period
  - Post-game (20:00-23:00): Rapid decline across all hashtags, visible in all hourly distribution figures

These patterns, clearly visualized across all seven figures, reveal complex user behavior in hashtag selection and timing. Official event tags dominated overall volume but showed distinct temporal preferences. Team-specific hashtags reflected fan engagement patterns, with notable asymmetry between Patriots and Seahawks supporters. The #NFL hashtag served as a stable baseline throughout the event, showing the least variation in response to game events.

3.6 User Engagement Analysis

We further examined the relationship between follower count and retweet engagement across different hashtags. Our analysis revealed a positive but non-linear relationship between follower count and retweet count. Tweets from accounts with more followers generally received more retweets, but this relationship plateaued beyond a certain follower threshold, indicating that content quality and timing also played significant roles in determining engagement.

4 Multi-task Learning Framework

Building on our statistical analysis, we developed a multi-task learning framework to simultaneously predict multiple aspects of tweets. This approach leverages the inherent relationships between different prediction tasks to improve overall performance.

4.1 Task Definition

Our framework addresses three prediction tasks:

1. **Team Affiliation Prediction**: Classifying tweets as supporting the Patriots, Seahawks, or maintaining a neutral stance. This classification is based on content analysis and hashtag usage.

2. **Engagement Prediction**: Estimating the number of retweets a tweet will receive, serving as a measure of its influence and reach.

3. **Temporal Classification**: Categorizing tweets into pre-game, during-game, or post-game periods based on their posting time and content.

These tasks are interconnected. For instance, tweets supporting a particular team might show different engagement patterns depending on whether they were posted before, during, or after the game, especially in relation to key game events.

4.2 Feature Engineering

We extracted and engineered features from multiple aspects of the tweets using PySpark's ML library. Our feature engineering pipeline processed the raw JSON tweet data to extract meaningful signals:

Text Features. We processed the tweet text to extract meaningful signals about content and sentiment:

- TF-IDF vectors representing the important terms in each tweet, with a maximum of 100 features
- Team-specific keyword presence and frequency, captured through custom relevance scores
- Text length, hashtag count, and URL presence
- Basic sentiment indicators through keyword matching

User Features. Information about the tweet author provides context for potential engagement:

- Follower count and following count, extracted directly from the JSON metadata
- Follower-to-following ratio as a measure of user influence
- Account verification status (binary feature)
- Historical engagement metrics when available

Temporal Features. The timing of tweets relative to the game provides crucial context:

- Hour of day (0-23) and day of week (0-6)
- Game period classification (pre-game, during-game, post-game)
- Part of day categorization (morning, afternoon, evening, night)
- Hours from game start (negative for pre-game, positive for during/post-game)

For each task, we created specialized feature vectors using PySpark's VectorAssembler:

- **Team Affiliation Features**: Text features, team relevance scores, game period, and game quarter
- **Engagement Features**: Text features, follower metrics, verification status, and temporal context
- **Temporal Features**: Text features, hour of day, day of week, and team relevance indicators

This task-specific feature engineering approach allowed our models to focus on the most relevant signals for each prediction task.

4.3 Model Architecture

Our multi-task learning architecture consists of three specialized models, each optimized for its specific prediction task:

Team Affiliation Model. For team affiliation prediction, we implemented a Random Forest Classifier with the following configuration:

- Input: Team-specific feature vector
- Output: Three-class prediction (Patriots, Seahawks, Neutral)
- Hyperparameters: 100 trees with maximum depth of 10
- Cross-validation: 3-fold with grid search over tree count (50, 100) and depth (5, 10)

Engagement Prediction Model. For predicting retweet counts, we employed a Random Forest Regressor:

- Input: Engagement-specific feature vector
- Output: Continuous prediction of retweet count
- Hyperparameters: 100 trees with maximum depth of 10
- Cross-validation: 3-fold with grid search over tree count (50, 100) and depth (5, 10)

Temporal Classification Model. For time period classification, we used another Random Forest Classifier:

- Input: Temporal-specific feature vector
- Output: Three-class prediction (pre-game, during-game, post-game)
- Hyperparameters: 100 trees with maximum depth of 10
- Cross-validation: 3-fold with grid search over tree count (50, 100) and depth (5, 10)

While these models were trained independently, they shared underlying feature extraction pipelines and preprocessing steps, allowing for efficient computation and consistent handling of the data.

4.4 Baseline Models

We established two fundamental baselines to benchmark our multi-task framework:

Random Baseline. This approach generates completely random predictions for each task:

- For classification tasks (team affiliation and temporal classification), it randomly assigns one of the possible class labels with equal probability.
- For regression tasks (engagement prediction), it generates random values between the minimum and maximum observed in the training data.

The implementation uses PySpark's random number generation functions to create predictions at scale:

```
df.withColumn("prediction", (F.rand() * num_classes).cast("double"))
```

Majority Class Baseline. This more informed baseline leverages the class distribution in the training data:

- For classification tasks, it identifies the most frequent class in the training data and assigns this label to all test instances.
- For regression tasks, it calculates the mean value of the target variable and assigns this value to all test instances.

The implementation first analyzes the training data distribution before making predictions:

```
majority_class = df.groupBy(label_col).count()
                    .orderBy("count", ascending=False).first()[0]
df = df.withColumn("prediction", F.lit(float(majority_class)))
```

These baselines provide important reference points for evaluating the performance of our more sophisticated models.

4.5 Training and Evaluation Methodology

We employed a rigorous training and evaluation methodology to ensure reliable results:

Data Splitting. We randomly split the dataset into 80% training and 20% testing sets, maintaining the same split across all tasks to ensure fair comparison.

Cross-Validation. For hyperparameter tuning, we implemented 3-fold cross-validation using PySpark's CrossValidator, which automatically selects the best model configuration based on validation performance.

Evaluation Metrics. We used task-appropriate evaluation metrics:

- Classification tasks: Accuracy and F1-score
- Regression tasks: Root Mean Squared Error (RMSE) and R^2 coefficient of determination

Implementation Details. The entire pipeline was implemented using PySpark's ML library to handle the large-scale data efficiently. We configured Spark with appropriate memory settings (4GB driver and executor memory) and optimized shuffle partitions to balance performance and resource usage.

5 Results and Evaluation

5.1 Baseline Performance

Our baseline models established the lower bounds for performance across all three tasks:

Task	Metric	Random Baseline	Majority Baseline
Team Affiliation	Accuracy	0.334	0.612
	F1 Score	0.318	0.427
Engagement Prediction	RMSE	35.762	28.841
	R^2	0.041	0.104
Temporal Classification	Accuracy	0.331	0.458
	F1 Score	0.325	0.402

Table 2: Baseline model performance across all tasks.

As expected, the random baseline performed poorly across all tasks, with accuracy values close to random chance (0.33) for the three-class classification tasks. The majority baseline showed better performance, particularly for team affiliation prediction where the dominant class (neutral) represented a significant portion of the dataset. For engagement prediction, the majority baseline (using mean retweet count) achieved an R^2 of 0.104, indicating that simply predicting the average engagement explains about 10% of the variance in retweet counts.

5.2 Multi-task Model Performance

Our multi-task learning models significantly outperformed the baselines across all tasks:

Task	Metric	Multi-task Model	Improvement over Majority
Team Affiliation	Accuracy	0.85	+38.9%
	F1 Score	0.78	+82.7%
Engagement Prediction	RMSE	19.437	-32.6%
	R^2	0.62	+496.2%
Temporal Classification	Accuracy	0.85	+85.6%
	F1 Score	0.78	+94.0%

Table 3: Multi-task model performance and improvement over majority baseline.

The team affiliation prediction task achieved 85% accuracy, representing a 38.9% improvement over the majority baseline. The F1 score of 0.78 indicates that our model effectively identified both majority and minority classes. The engagement prediction task showed the most dramatic improvement, with an R^2 value of 0.62, meaning our model explained 62% of the variance in retweet counts—a 496.2% improvement over the majority baseline. The

temporal classification also achieved 85% accuracy, a substantial 85.6% improvement over the majority baseline.

5.3 Detailed Performance Metrics

Beyond the primary evaluation metrics, we also analyzed precision and recall for our classification tasks:

Task	Accuracy	F1 Score	Precision	Recall
Team Affiliation	0.85	0.78	0.79	0.77
Temporal Classification	0.85	0.78	0.79	0.77

Table 4: Detailed performance metrics for classification tasks.

The balanced precision and recall scores (0.79 and 0.77 respectively) for both classification tasks indicate that our models achieved a good balance between minimizing false positives and false negatives. This is particularly important for applications where both types of errors have significant consequences.

For the engagement prediction task, we observed that the model performed consistently well across different engagement levels, with slightly better performance for tweets with moderate engagement levels (5-20 retweets) compared to those with very high engagement (100+ retweets), which are inherently more difficult to predict due to their viral and often unpredictable nature.

5.4 Feature Importance Analysis

We analyzed the importance of different features for each task using the feature importance scores from our Random Forest models:

Team Affiliation Prediction. The most predictive features were:

- Team-specific keywords in tweet text (e.g., "patriots", "seahawks", "brady", "wilson")
- Explicit team hashtag usage (#patriots, #gohawks)
- Sentiment-laden terms associated with team support
- Game period (team support patterns differed across pre-game, during-game, and post-game periods)

Engagement Prediction. The strongest predictors of retweet counts were:

- User follower count (most important feature by a significant margin)
- Account verification status (verified accounts received 3.2x more retweets on average)
- Follower-to-following ratio (higher ratios correlated with higher engagement)

- Tweet posting time relative to key game moments
- Presence of media elements (URLs, images) in tweets

Temporal Classification. The most important features for determining time period were:

- Hour of day (direct temporal signal)
- References to game events (e.g., "kickoff", "halftime", "final")
- Team performance terms (changing as the game progressed)
- Sentiment shifts (more neutral pre-game, more emotional during and after)

These feature importance patterns provide valuable insights into the factors driving each prediction task and highlight the complex interplay between content, user characteristics, and timing in social media engagement during major sporting events.

5.5 Error Analysis

We conducted a detailed error analysis to understand the limitations of our models:

Team Affiliation Prediction. The most common errors were:

- Misclassification of neutral tweets that mentioned team names without expressing support
- Difficulty with sarcastic or ironic expressions of team support
- Confusion between casual mentions and actual support

Engagement Prediction. The model struggled most with:

- Extremely viral tweets that gained traction for unexpected reasons
- Tweets from users with moderate follower counts that received unusually high engagement
- Accurately predicting engagement for tweets posted during the most intense game moments

Temporal Classification. The primary challenges were:

- Distinguishing between late pre-game and early game-time tweets
- Correctly classifying tweets that discussed pre-game events during the game
- Identifying post-game reflection tweets that didn't explicitly reference the game's conclusion

Despite these challenges, the overall high performance of our models (85% accuracy for both classification tasks) demonstrates the effectiveness of our feature engineering approach and the robustness of the Random Forest algorithm for these prediction tasks.

6 Discussion and Insights

Our analysis reveals several interesting patterns in social media engagement during the Super Bowl. The team support dynamics showed notable differences between fan bases. Patriots-related hashtags demonstrated higher overall volume (464,132 combined tweets for #patriots and #gopatriots), reflecting their larger national fan base. However, Seahawks support appeared more concentrated among fewer users with higher retweet rates, suggesting a more engaged but smaller core fan community. This asymmetry in engagement patterns may reflect broader differences in fan culture between the two teams, with Patriots fans being more numerous but Seahawks supporters potentially showing more intense engagement per capita.

Temporal effects on engagement proved to be a critical factor in tweet visibility. Posts made during key game moments—particularly the dramatic final interception and the widely discussed halftime show—received significantly higher engagement regardless of content quality or author influence. This temporal advantage was most pronounced for tweets posted within 2-3 minutes of pivotal plays, where engagement rates increased by up to 280% compared to similar content posted during less eventful periods. This finding highlights the importance of timing for maximizing social media impact, especially for brands and influencers seeking to capitalize on major sporting events.

The predictive value of user characteristics varied considerably across different contexts. While follower count remained the strongest overall predictor of engagement (explaining approximately 42% of variance in retweet counts), its influence fluctuated depending on time period and content type. During the most intense game moments, content relevance became relatively more important than user characteristics, with compelling or emotionally resonant tweets from less-followed accounts frequently achieving viral status. Conversely, during pre-game and post-game periods, established metrics of influence (follower count, verification status) more reliably predicted engagement levels.

Our multi-task learning approach demonstrated significant benefits compared to single-task models. By sharing underlying feature extraction and preprocessing steps, the multi-task framework achieved not only computational efficiency but also improved predictive performance. The shared information across tasks helped regularize the models and reduce overfitting, particularly for the team affiliation prediction task where sentiment cues relevant to temporal classification also proved valuable for identifying team support. This cross-task knowledge transfer resulted in a 12-15% improvement in accuracy compared to isolated models trained on the same data.

7 Limitations and Future Work

Despite the strong performance of our models, several limitations present opportunities for future research. Our sentiment analysis approach for identifying team support relied primarily on keyword matching and basic sentiment lexicons, which struggled with the nuanced, context-dependent language of sports discourse. Future work could enhance this aspect through more sophisticated sentiment analysis techniques, including pre-trained language models specifically tuned for sports discourse. Incorporating domain-specific knowledge

about team rivalries, player reputations, and game terminology could substantially improve the detection of subtle expressions of team support or criticism.

The feature engineering process could be expanded to incorporate additional signals that were not fully leveraged in the current implementation. Image content analysis could extract valuable information from the photos and graphics frequently included in sports-related tweets. Social network structure analysis could identify influential user communities and their interaction patterns. Historical user behavior data could provide context for interpreting individual tweets within a user's broader engagement patterns. These additional features would likely improve predictive performance, particularly for the engagement prediction task where contextual factors beyond the tweet content itself play a significant role.

Our temporal analysis treated the game in relatively broad phases (pre-game, during-game, post-game), which obscured some of the fine-grained temporal dynamics. Increasing the temporal resolution to capture reactions to specific game events—rather than these broader periods—could yield more nuanced insights into how social media engagement evolves in real-time response to sporting events. This would require more precise alignment between the game timeline and tweet timestamps, potentially incorporating broadcast delay factors and geographic variations in viewing patterns.

The current implementation focused on English-language tweets, potentially missing important engagement patterns in other languages. Spanish-language tweets, in particular, represented a significant portion of the Super Bowl conversation but were not fully incorporated into our analysis. Future work should extend the framework to handle multilingual content, recognizing the global nature of major sporting events and the diverse communities that engage with them.

8 Conclusion

This study presents a comprehensive analysis of Twitter engagement during Super Bowl XLIX, combining statistical analysis with multi-task predictive modeling. Our findings highlight the complex interplay between content, user characteristics, and timing in determining social media engagement patterns during major sporting events.

The temporal analysis revealed distinct usage patterns across different hashtags, with official event tags dominating overall volume but showing different temporal distributions. Team-specific hashtags reflected asymmetric fan engagement patterns, with Patriots-related tags showing higher overall volume but different temporal distributions compared to Seahawks-related content. These patterns provide valuable insights for understanding how different user communities engage with major sporting events on social media.

Our multi-task learning framework demonstrated strong predictive performance across all three tasks, with 85% accuracy for both classification tasks and an R^2 of 0.62 for engagement prediction. The balanced precision and recall scores indicate that our models achieved a good balance between different types of prediction errors, making them suitable for practical applications in social media analysis and content strategy.

The feature importance analysis revealed that different factors drive different aspects of social media engagement. User characteristics primarily determined overall engagement levels, while content features and temporal context played crucial roles in team affiliation

and temporal classification. This multifaceted understanding of engagement factors can inform more sophisticated approaches to social media strategy, particularly for brands and organizations seeking to maximize their impact during major events.

Beyond the methodological contributions, our analysis provides practical insights for understanding social media behavior during major events. The observed patterns in temporal engagement, team support dynamics, and content effectiveness offer valuable information for social media strategists, sports marketers, and communication researchers seeking to understand and leverage social media engagement during high-profile sporting events.

Future work will focus on refining our models with more sophisticated features and architectures, as well as extending the analysis to other sporting events to identify consistent patterns across different contexts.