An Evaluation of Twitter Sentiment Data and Machine Learning Classifiers to Predict NFL Match Outcomes

University of London DSM500 Final Project

By Patrick O'Neill

Abstract

In recent years, there has been an explosion in the sports betting market and this research examines the effectiveness of machine learning classifiers and Twitter sentiment in predicting the winning team against the spread in National Football League (NFL) games. This study evaluates AdaBoost, Naive Bayes, Neural Network, Random Forest and XGBoost classifiers, along with three Twitter sentiment features that include a stock market charting strategy to measure relative changes. Tweets are collected based on team hashtags and sentiment determined with a pre-trained Twitter-tuned RoBERTa model. The study integrates Twitter sentiment features with basic football statistics from the 2022 NFL season to train several models.

This research finds AdaBoost to be the most accurate classifier, with a top accuracy of 65.1%, significantly higher than other leading research predicting point spread outcomes. Additionally, Naive Bayes (56.6%), Neural Network (59.0%) and XGBoost (57.9%) also exhibit significantly higher model accuracies than the 52.4% needed to achieve profitability when selecting spread bet winners. The study concludes that while the effectiveness of Twitter data was inconclusive, four out of five classifiers utilize sentiment features in their most accurate models, including three that use the swing sentiment features. However the accuracy increases are not found to be significant. Future studies with more data may strengthen the evidence to support the use of Twitter sentiment to predict NFL outcomes.

Table of Contents

Abstract **1**
Table of Contents **2**
1. Introduction **4**
2. Literature Review **7**
2.1 Background 7
2.2 Previous Literature 8
2.2.1 Twitter Sentiment Analysis 8
2.2.2 European Football and Twitter Sentiment Analysis 9
2.2.3 NFL and Twitter Sentiment Analysis 10
2.2.4 Machine Learning and Sports Predictions 12
3. Methodology **13**
3.1 Data Collection 13
3.2 Sentiment Analysis 14
3.3 Feature Selection 17
3.3.1 Football Statistic Features 17
3.3.2 Twitter Sentiment Features 19
3.4 Data Preprocessing 20
3.5 Datasets 21
3.5.1 Football Features 21
3.5.2 Sentiment Features 22
3.6 Model Creation and Tuning 23
3.6.1 AdaBoost 23
3.6.2 Neural Network 23
3.6.3 Random Forest 24
3.6.4 Naive Bayes 24
3.6.5 XGBoost 24
3.7 Training and Evaluation 25
4. Experiments and Discussion **26**
4.1 Baseline Models With No Sentiment 26
4.2 Experiment 1: Profitable Accuracy and Leading Work Comparisons 27
4.2.1 Profitable Accuracy 27
4.2.2 Leading Research Benchmark Accuracy 28
4.3 Experiment 2: Effectiveness of Twitter Sentiment Data 29
4.4 Experiment 3: Comparison of ML Classifiers 32
4.5 Experiment 4: Swing Sentiment 33
5. Conclusions and Future Work **34**
6. References **36**
Appendices **39**
Appendix A: Feature Descriptions 39

Appendix B: Baseline Dataset Experiment Results 40
Appendix C: Experiment 2 Results 42
Appendix D: Experiment 2 Significance Tests 44

1. Introduction

Betting on the outcome of events has a long rich history that dates back thousands of years, from the Ancient Romans' chariot races where citizens would bet on chariot races (Sports Encyclopedia, 2023) to the modern-day sports betting market worth an estimated $250 billion USD. With the expansion of online betting tools and the reduction of gambling restrictions, this industry is expected to have continued growth into the foreseeable future (SportsBettingDime, 2022). There are an abundance of factors of varying significance that come together to determine a winner. Coaching, game tactics, weather, injuries, refereeing, location, weather and more, are all intertwined making the outcome of matches notoriously difficult to predict. These factors vary widely and evolve over the course of years, seasons or even within matches. The challenge that predicting sports outcomes presents is a key driver of its popularity, the element of uncertainty appealing to fans and gamblers alike.

The NFL is the most valuable sports league in the world and the championship game, the Super Bowl, drew an estimated $16 billion in wagers in 2023 (PBS, 2022). The league was formed in 1920 and consists of 32 football teams located in the United States. Wagering on the NFL is incredibly popular, drawing an estimated $100 billion USD in wagers during the 2022 NFL season. Gamblers can make a variety of bets including picking the outright winner, picking the winner based on the point spread, or betting on different events throughout the game.

There are many stakeholders when it comes to the prediction of sports match outcomes. Teams are looking for advantages against opponents, owners look to increase profits, sports bookmakers need to set the best odds to maximize profits, and gamblers are hoping to win big. The difficult nature of predicting sports outcomes has also attracted the interest of academics. A large variety of publicly accessible datasets makes it an excellent field to test the effectiveness of predictive modeling techniques. Gambling markets are highly efficient, arbitrage opportunities are quickly accounted for with an adjustment of odds. Gamblers are incentivized to devise unique betting strategies to beat the market and turn a profit, while bookkeepers set the odds to limit any arbitrage opportunities that could lead to a loss in revenue from an unbalanced betting

market. Research in this field dates back to the 1970's with the use of mixed linear models to predict American football outcomes (Harville, 1977) . Modern techniques such as boosted decision trees (Landers, 2019), neural networks (Hsu, 2021) and the comparison of several machine learning (ML) classifiers (Beal, 2020) have recently been applied to this problem. The experts in this field are the sportsbooks themselves, they set the betting lines and point spreads of matches using a combination of sophisticated models, expert domain knowledge, and crowdsourced information from bets. Bookkeepers set the gold standard baseline models to beat, making it an excellent research space to test predictive models and techniques.

Twitter is a real-time microblogging app founded in 2006 and has grown to one of the biggest social media platforms. With over 500 million tweets published daily and over 400 million active users (Business of Apps, 2022), Twitter provides a rich source of crowdsourced data, making it an invaluable research tool for academics and businesses. Users post short messages called tweets, about any topic of interest and these can be analyzed using natural language processing techniques. Insights and predictions can be derived for a variety of subjects including crime, stock markets, political elections, public opinion polls, public health, and even the detection of natural disasters. Users may use hashtags to associate their tweet with certain trends or topics. In the case of sports, a hashtag may refer to a certain team they are tweeting about such as "#bengals" or "#ravens". By analyzing user-annotated hashtags and filtering tweets based on certain topics, researchers can discover trends or learn about public opinion. One such type of analysis is sentiment analysis where texts are classified as containing either positive or negative sentiment. Twitter sentiment analysis has been used with some success to predict sporting outcomes as well, being harnessed as a 'wisdom of the crowds' feature.

While previous research has had varying success predicting NFL outcomes through the use of either ML classifiers or Twitter sentiment analysis, the combination of these approaches has yet to be explored in depth. Beal (2020) tested the effectiveness of several ML classifiers to predict NFL outcomes but noted that further research could attempt to include the human element of sports. Twitter sentiment analysis has the potential to harness the wisdom of the crowds, adding that missing human element into ML classification models similar to those tested by Beal.

The aim of this paper is to evaluate the use of machine learning classifiers with Twitter sentiment features to predict NFL outcomes against the spread. The research questions that will be explored in this project are as follows:

1. Utilizing the combination of football performance and Twitter sentiment data, are machine learning (ML) classifiers able to exceed the 52.4% benchmark accuracy to be profitable in predicting NFL outcomes against the spread?
2. Does the addition of Twitter sentiment features improve a ML classifier's accuracy predicting the outcome of NFL games?
3. Which ML classifier, between AdaBoost (AB), Random Forest (RF), Naive Bayes (NB), Neural Network (NN) and XGBoost (XGB), produces the highest accuracy predicting NFL outcomes against the spread?
4. Does the technical charting Twitter sentiment feature derived by Schumaker (2017) improve a ML classifier's accuracy predicting the outcome of NFL games?

Game statistics are collected for the entire 2022 season while the collection of Tweets ranges from weeks 12 to 19, including the first week of playoffs. This data will be used to create different feature sets for which to use to train the different ML classifiers. Models will be trained on data that contains different sentiment features, or none at all. Publicly available football statistics will be included in the model alongside Twitter data. The performance of these models will be evaluated on their ability to correctly identify the winner with the point spread. In order to be profitable when making point spread bets, a model must pick the winner 52.4% of the time.

The rest of the paper is structured as follows, a literature review is conducted of previous studies concerning Twitter sentiment analysis, sports prediction with ML classifiers, and sports prediction using Twitter sentiment data. Section 3 presents the methodology of the paper including data collection, feature selection, and construction of the final datasets. The different models, training and evaluation are described in detail. Section 4 presents and discusses the results of the experiments conducted while section 5 concludes the study.

2. Literature Review

2.1 Background

The NFL has an 18 week long regular season where teams play weekly with the exception of a bye week for a total of 17 games per season. 14 teams make a single elimination playoff format that ends with the championship game called the Super Bowl. Most games are played on Sunday in addition to a single game on Monday and Thursday night with some exceptions to this schedule during holidays and playoffs. The regularity of the NFL's weekly schedule makes it particularly convenient for study, in this case, the collection of tweets can be done in regular intervals. This would be much more challenging if games were only a day apart such as in other sports such as baseball or hockey. There is generally a minimum of 4 days between games, with the exception of teams playing in the Thursday night game if they played on the previous Sunday afternoon or morning. NFL teams have around 10 coaches and 55 players on their roster with each team having 11 players on the field at one time. Teams have separate offense and defensive players and coaches. Teams can either run or pass the ball with the goal of scoring a touchdown by getting the football into the opponent's end zone, or kicking a field goal worth 3 points.

There are many ways to bet on NFL games. Wagers can be made on the outright winner, where higher returns are earned with successful bets placed on the underdog teams rather than the favored team. Point spread betting is another popular wager and the focus of this paper. Bookmakers set a "spread" based on which team is the favorite to win, and gamblers bet on either team to cover the spread. An example point spread is the Detroit Lions at +4.5 underdogs to the favored Chicago Bears at -4.5. Chicago is favored to win by 4.5 points, so in order to win a spread bet placed on them, Chicago would have to win by 5 points or more to cover the spread. The Lions would cover the spread by either losing by less than 4.5 points or winning outright. A push results when the score of the game equals the spread and the wager is returned to the bettor, there are no winners or losers. The spread is generally set at ½ point intervals to avoid this because sportsbooks earn revenue from a commission on losing bets, and a push results in no losing bets. Generally, a bettor will have to bet $110 to win $100 by betting on the spread. In order to be profitable you must be more than 52.4% accurate to cover the sportsbook commission. The spread is set with the goal of attracting an even number of bets

on either team, creating a balanced book. If the initially set odds attract too many bets to one team over the other, bookmakers may adjust the odds to encourage bets on the other team, rebalancing the book.

2.2 Previous Literature

2.2.1 Twitter Sentiment Analysis

Kharde (2016) and Zimbra (2018) conducted research surveys of Twitter sentiment analysis approaches and techniques. Zimbra identified three main challenges faced in this domain: novel language as a result of the short length of text from a character limit, sentiment class imbalance (Hagen 2015) leading to biased models, and stream-based tweet generation. Kharde (2016) noted further challenges in identifying the subjective part of the text, domain dependance and sarcasm detection. Zimbra confirmed the issue of domain dependence, finding that general classification systems performed poorly in comparison to trained domain-specific systems. While the study reviewed five different domains, none were sports-related. Therefore, sentiment analysis of sports related tweets could present specific challenges due to the specific language and terms used. Top-performing models used an ensemble of machine learning classifiers, producing better results than deep learning methods and achieving an average accuracy of 71% across all tested domains. Zimbra concluded that ensemble methods could overcome the challenges presented by class imbalance and poor sentiment recall. Kharde concluded that machine learning methods had the highest accuracy in their survey of the literature and testing while lexicon-based models performed well in certain situations.

Barbieri (2020) devised a novel evaluation framework, TweetEval, to identify top classification models and create benchmarks for seven heterogenous Twitter-specific classification tasks including sentiment analysis. Experiments using this framework found that it is effective to take a generic pre-existing model and continue training it with Twitter data. A baseline Twitter sentiment classification model is established by building off RoBERTa (Liu 2019), a NN text classification model. The pre-trained model is fine-tuned with the addition of a dense layer to reduce the output dimensions to the appropriate number of labels for each classification task. This sentiment classification model was 72.9% accurate in the Twitter sentiment task, a result competitive with the top classification models tested by Zimbra.

2.2.2 European Football and Twitter Sentiment Analysis

Twitter sentiment analysis has been applied to various domains such as finance, healthcare, and politics but few studies have explored its potential to predict sports match outcomes. Wunderlich (2020) found that lexicon-based sentiment tools could effectively classify batches of 1000 tweets related to European football. While some research has removed hashtag information from the tweets before classification, Wunderlich includes them believing they may contain information beneficial to determining the appropriate sentiment of the tweet. This is a reasonable decision because of Twitter's character limit. Removing hashtags would effectively reduce an already limited dataset, making sentiment classification more challenging.

Half of the tweet samples in the study were found to contain less than 10 words, making some classification difficult. The example short tweets provided appear to be written in reaction to in-match events. Our current research is limited to intergame tweets, which should reduce the proportion of short, reactionary tweets and increase overall classification accuracy. The number of tweets associated with each team and game in this study ranges from several hundred to five thousand, similar to current research. Wunderlichs findings validate that lexicon-based tools can accurately classify batches of European football tweets that are of a similar size to that of this paper.

Schumaker's (2016) study utilized OpinionFinder, a Twitter sentiment analysis tool which classifies tweets by both tone and polarity, to predict European football match outcomes. Interestingly, Schumacher found that a teams' ranking in the standings did not always correlate with the baseline sentiment of their fanbase. Some top ranked teams did not have the highest level of positive sentiment and vice versa. A comparison of an odds-only baseline model to a sentiment model found that the odds-only model had the highest accuracy predicting the outright winner but the sentiment models generally had higher payouts. Sentiment models demonstrated higher accuracies predicting longshot winners resulting in higher returns on their wagers. The study also found that the inclusion of the relative surge or drop in positive sentiment compared to the club average yielded the highest payout. These findings highlight the significance of relative differences in sentiment between teams rather than gross sentiment comparisons.

Wunderlich (2022) examines the ability of Twitter sentiment analysis to forecast goals in real time, during European football matches. Tweets are collected throughout the match and sentiment was classified using three different lexicon-based methods. However, due to multicollinearity and redundancy issues, no sentiment features were included in the training of the final model. Instead, tweet intensity was included, a measure of the number of tweets posted per minute. In-game goal forecasting regression and RF models were unable to improve upon the betting odds set ahead of the match. The study found that neither Twitter sentiment or tweet rate improved the accuracy of models forecasting in-game goals. Of interest was the finding that tweet length became shorter once the match started, consistent with the findings of Wunderlich's previous study in 2020. This result supports the assumption that reactionary in-game tweets are shorter in length.

2.2.3 NFL and Twitter Sentiment Analysis

Previous research examining the use of Twitter data to predict the outcome of NFL games has been limited. Only a few studies have explored this area, including the work of Hong (2010) and Sinha (2013), who combined football statistics with Twitter data to create outcome models. In contrast, Schumacher (2017) focused exclusively on using Twitter sentiment features.

Hong utilized a lexicon-based sentiment classification tool to analyze multiple types of media, including newspapers, blogs and Twitter. Twitter data was only used for a portion of the study however as Twitter was a relatively new platform at the time of research. Hong devised an equation which considers daily positive and negative sentiment counts for opposing teams to determine a favored team, and combined this with football statistics in a linear regression model. A main result was that model accuracy improved in the second half of seasons, which was believed to be due to more evidence of a team's ability from previous games being available. Hong found that cumulative sentiment averages over the duration of the season were found to be better predictors than week-to-week sentiment scores, and that the sentiment-based model performed best during the second half of the season compared to football data models. This model produced a point spread betting strategy that was profitable for the second half of each season studied, achieving a 60% winning rate. The winning rate is determined by a betting strategy, where 30 games in each season are selected where the model output is most

significantly different from the point spread. Overall accuracy is not reported in the study but from the charts included it can be estimated to be around 50%. So while the study achieved a betting strategy winning rate of 60%, the model may not have been profitable when betting on all games.

In their study, Sinha (2013) adopts a similar approach of combining both Twitter features and performance statistics to predict NFL game results. However, instead of sentiment features, they employed a logistic regression classifier with Twitter rate and unigram features. Tweet volumes for individual teams are then calculated for each week and rate changes are then measured and added to the model. The authors found that the best model combined a set of sport performance features with Twitter rate features, achieving an accuracy of 57.2% predicting the winner with the spread. Different models performed better at different points in the season and the authors noted that the selection of the best feature set was a challenge. They suggested the use of different ML classifiers to address this challenge which is what this research aims to complete. There is a vast amount of performance statistics available in sports, and ML models have the potential to determine the most predictive ones.

Schumaker (2017) extended previous research by incorporating the use of technical stock market charting techniques to Twitter sentiment analysis. Building on their earlier study of European football, which found that different fan bases have distinct baseline measures of sentiment, this study uses relative differences and trends in Twitter sentiment for opposing teams. Sentiment is tracked using moving averages, looking for “golden cross” or “death cross” patterns that are more generally used in stock market analysis. These patterns arise when short-term moving averages cross the long-term moving averages, indicating potential gains or losses. 96-hour and 24-hour moving average windows are calculated and then combined to form a “swing” statistic to find any cross patterns in a teams sentiment.

Moving averages normalize the sentiment data of each team, taking into account that fan bases may have different baseline sentiments about their respective teams. The model containing the technical charting pattern “swing” statistic had the highest payout, due to its enhanced ability to predict long shot winners, a similar result to Schumaker (2016). The addition of football performance features to this model has the potential to increase the accuracy achieved by this study.

To conclude, the use of Twitter data has had varying degrees of success predicting European and American football outcomes. While Hong's sentiment-based model achieved a profitable spread betting strategy, the overall accuracy may not have been above the 52.4% needed to be profitable. Sinha's combination of Twitter rate and unigram features with football performance statistics did manage to achieve a profitable accuracy of 57.2%, but did so with a logistic regression model, finding feature selection a large challenge. A ML classifier could help deal with this challenge. Schumaker's study on NFL games incorporated technical stock market charting techniques to effectively normalize the Twitter sentiment data of each team, after finding in previous research that different fan bases had different baseline levels of sentiment not correlated with the team's success. Schumacher's study once again found Twitter sentiment a valuable tool to predict longshot winners better. Overall, these studies demonstrate the potential of using Twitter data in predicting sports outcomes, but further research could build off these results.

2.2.4 Machine Learning and Sports Predictions

Beal (2020) conducted a study to test and compare several ML classification models for predicting the outright winner of NFL games using datasets consisting of 42 football performance features. An NB model in this study achieved the top accuracy of 67.53% predicting outright winners over five NFL seasons. This is higher than the bookmakers baseline for this type of bet which is around 63%.

The study found that the NB classifier performed the best followed by AB (66.3%), RF (64.31%), Decision Tree (63.5%) and a NN (60.7%). There were some seasons in which the NB model was outperformed by AB and RF models, suggesting that an ensemble classifier using all three methods could improve overall prediction consistency. The top model was able to slightly outperform the bookmakers odds but this was not demonstrated to be a statistically significant result. It is also worth noting that the results include a misleading comparison to Landers' (2019) research, which predicted winners with the spread with an accuracy of 58.1%, one of the best results in the literature for point spread betting. In contrast, the baseline accuracy of this betting is 52.4%, while for outright winners, it is around 63%.

Landers (2019) uses a number of engineered football features based on the favored and underdog trees to predict winners WTS spread over three NFL seasons. A boosted decision

tree (BDT) classifier was able to achieve an accuracy of 58.1%, which was shown to be significantly higher than and the profitable baseline of 52.4%. This model performed better than 50 humans in a league picking winners WTS on a weekly basis over an NFL season.

Overall, Beal's study showed that different ML classifiers do have varying results predicting NFL games. The most promising were NB and AB, achieving an accuracy higher than the baseline odds, while DT and NN were close. The authors noted that the inclusion of wisdom of the crowds features and the point spread could potentially improve the accuracy of their findings, both of which are attempted in this paper. Landers demonstrated that a BDT classifier with football statistics can produce a profitable spread betting model.

3. Methodology

3.1 Data Collection

There are three main pieces of data required for this project: tweets containing the hashtags of NFL teams, football performance statistics and scores, and the point spread of each game. While game statistics and the point spread are readily available online, the collection of Twitter data was challenging and limited the amount of data included in the study. Basic access to Twitter data only allows tweets posted within the last week, and does not allow for the retrieval of historical data. Special academic access is needed for this but the application was unsuccessful. Tweet collection began as early as possible into the study in order to gather as much data as possible, starting in week 12 of the 2022 NFL season.

Tweets were collected using Tweepy (Tweepy, 2023), a python library allowing access to the Twitter API. Searches were conducted using the official team hashtags (NFL Picks Pro, 2022) and one or two unofficial hashtags that were found to be frequently used when searching Twitter. Following Sinha's (2013) approach, tweets that contained more than one team's hashtag were removed. This allows the tweet's sentiment to be assigned to a single team. Schumaker (2017) applies another filter, limiting tweets to one per user during each moving window analyzed (96 and 24 hour) to avoid a single user skewing the results. For this study however, multiple tweets from single users were retained because high-rate tweeters are likely to be news or spam and then classified as neutral sentiment and not included in the training

data. Moreover, fans with multiple negative or positive sentiment tweets could help to capture particularly negative or positive events leading into the games.

Football statistics and the point spread for all games during the 2022 season were collected from pro-football-reference.com (Pro-Football-Reference, 2022). This website contains a comprehensive list of statistics for each team and game which were collected using a python web scraping library, Selenium (Selenium, 2022). Performance statistics were collected for the whole 2022 season to allow for the creation of lagged performance statistics that can be averaged over any number of games for that season.

3.2 Sentiment Analysis

Sentiment analysis was conducted on all tweets collected using the Twitter-tuned RoBERTa model, a state-of-the-art sentiment classification model. This was chosen because of its accessibility, ease of use, social media training data, and a Twitter sentiment classification test accuracy comparable with top results from previous literature. The RoBERTa base is trained on five different English language corpora, including 38 GBs of Reddit data, another social media platform containing many shorthand acronyms and slang. The Barbieri (2020) baseline model RoBERTa (Liu et al. 2019) achieved a general sentiment accuracy of 72.9% on the TweetEval sentiment classification baseline test set. This result is competitive when compared to the top models found by Zimbra's Twitter sentiment classification review which were around 70%.

The RoBERTa model is easily accessible through huggingface.com, an open-source community for pre-trained machine learning models. The latest version of this model is trained on approximately 124 million tweets written from 2018 to 2021 (Hugging Face, 2023), which is important to counteract the challenge presented by stream-based data generation (Zimbra, 2018). Twitter language evolves over time, and models must be trained on newer data to maintain high levels of accuracy when exposed to new data. Tweets are processed according to direction by the model authors: usernames were replaced with "@user", links replaced with "http" and hashtags left intact, and then tokenized using the model's auto-tokenizer. Finally, a TensorFlow implementation was used to classify all tweets with the Twitter tuned RoBERTa model, outputting positive, negative and neutral sentiment classifications. Example classifications are shown in Table 1.

Table 1

Example Tweets and their corresponding Sentiment Classification

Tweet	Sentiment Classification
“So we’ve gone from: “#Bears roster stinks. Just need to see growth from Fields,” to… “Fields needs to win a game in the 4th quarter,” to… “Fields needs to play with a separated shoulder b/c all great QBs rise to the occasion.” Like what are we even doing here?”	Negative
“#49ers Deebo Samuel (hamstring) who is listed as Questionable, will likely play against the Saints, per @RapSheet”	Neutral
“My only day off until December 10th, what better way to spend it than drinking and watching @49ers. As always it's #fttb and let's get that win against @Saints.”	Positive

Table 2 presents the results of the sentiment classification performed on over 600,000 tweets. Only 10% of tweets contained negative sentiment, while 48% were classified as neutral and 42% positive sentiment. The high percentage of neutral tweets can be attributed to the fact that many tweets are likely news stories updates or team updates not expressing any particular sentiment. The gross counts of positive and negative sentiment tweets are presented in Chart 1, while Chart 2 shows the ratio of negative to positive sentiment for each team. These show the level of variance between teams, supporting the need for a way to capture the relative difference rather than fixed sentiment measures.

Table 2

Breakdown of sentiment classification showing the proportion of negative sentiment is very low in comparison to positive and neutral sentiment.

	Count	Percentage
Negative	60,227	10.0%
Neutral	291,036	48.4%
Positive	250,433	41.6%
Total	601,696	

Chart 1

Total counts of positive and negative sentiment tweets by team

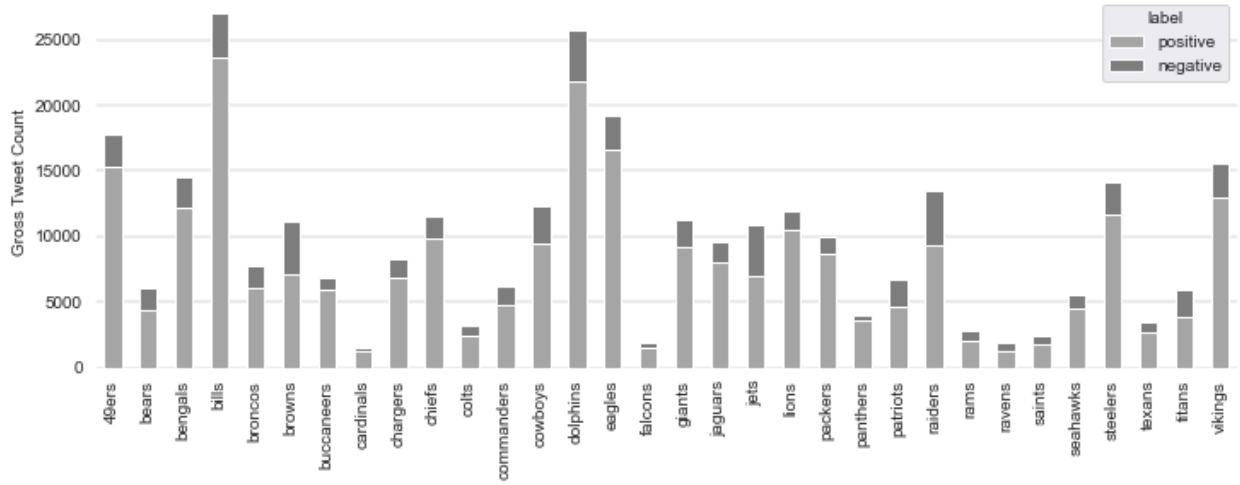
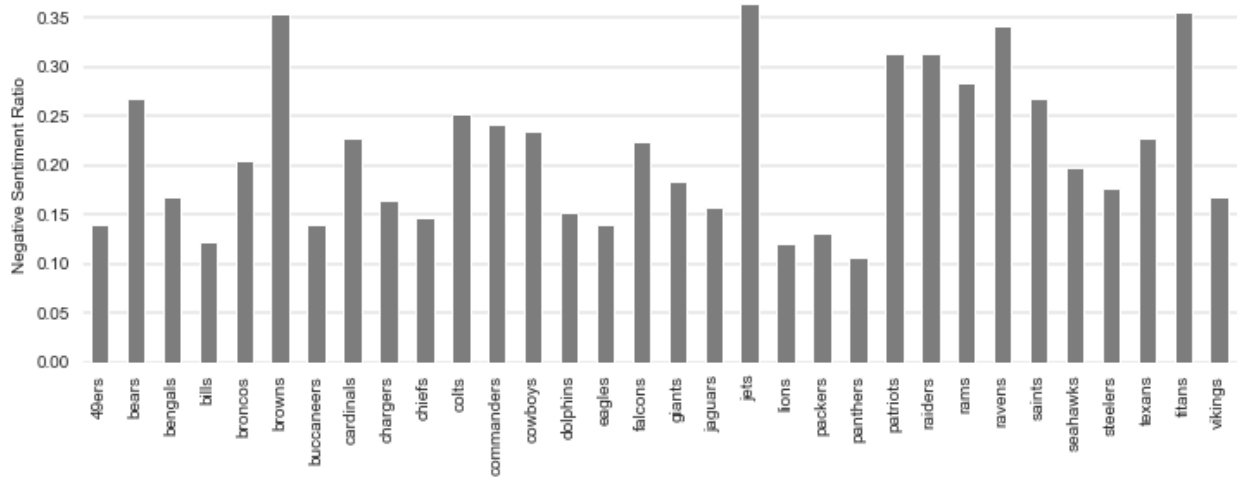


Chart 2

Ratio of negative to positive sentiment for each team



3.3 Feature Selection

This section provides an overview of the football statistical features and Twitter sentiment features chosen for the datasets. Unlike previous studies (Sinha, 2013 and Landers, 2019), this research will use football statistics without any custom feature engineering. The focus of this study is to evaluate the effectiveness of several ML classifiers and different Twitter sentiment features rather than engineered football features.

3.3.1 Football Statistic Features

Given the vast amount of football statistics available, some statistics that are believed to have minimal effects on the game or have little variance are removed to reduce the total number of features. Additionally, any statistics that are mathematical combinations of other features, such as averages or percentages, are also removed as they exhibit perfect multicollinearity with their base statistics. For example, pass completion percentage, the amount of passes caught as a percentage of total passes attempted is removed as both features are already included. Similarly, rushing yards per attempt, pass net yards per attempt, passing yards per attempt are also removed. This approach makes the model easier to interpret, reduces noise and decreases training times.

Other statistics have been removed based on domain knowledge and previous studies. For instance, 4th down conversion data have been removed due to their sparsity, with many teams not recording any stats in this category each week. In line with previous studies which did not include any kicking data, field goal statistics are removed. The scoring potential of field goals is already reflected in the total points scored.

Turnovers are kept as a single metric rather than being divided into fumbles and interceptions because the type of turnover is insignificant compared to the impact it has on the game. Turnovers can result in momentum swings that have large effects on the outcome of the game. Teams that force more turnovers than they give up generally have a better record. Penalties can also result in significant swings in the momentum of the game for one team or the other. When a team takes a penalty they lose a certain number of yards in field position

dependent on the type of penalty. Teams that are able to take less penalties than their opponents have a better chance to win.

Time of possession, how long each team has control of the football, is also included. A team that is able to control possession of the ball and remain on offense longer will have greater success. Following Beal (2020), expected points for defense, a measure constructed by pro-football-reference.com statisticians is included. This statistic represents how much the defense contributed to the point differential at the end of the game. While this is an engineered statistic, it is included in this study because it is both available online and also included in previous research.

Table 3

List of football performance features used for both the home and away teams for each game

Football Game Features (Both Home and Away Teams)	
Team	Offense & Defense
Time of Possession	Pass Completions
# Penalties	Pass Attempts
Penalty Yards	Pass Yards
# Opponent Penalties	Pass TDs
Opponent Penalty Yards	Rush Attempts
Team	Rush Yards
Day of Week	Rush TDs
Spread	Sacks
3rd Down Attempts	Sack Yards
3rd Downs Made	1st Downs
Pass QB Rating	Turnovers
Defense Expected Points	Points Scored / Allowed

The point spread is included because this contains a lot of information about the expected result and is information gamblers would base their decision as well. Both Sinha (2013) and Landers (2019) include the spread as a feature in their models, while Beal (2020) did not, but did however suggest future studies evaluate its inclusion. The point spread has been found to be a valuable feature in predicting NFL outcomes in previous research (Glickman and Hal, 1998). The use of a home field coefficient as used in Beal (2020) is not included because the point spread already contains this information. Bookkeepers take the home field advantage into consideration when setting the spread. Table 3 lists all football features included in the models and appendix A contains further descriptions of included football statistics.

3.3.2 Twitter Sentiment Features

This study uses the Twitter sentiment features proposed by Schumaker (2017) to further test their applicability in ML classification models combined with football statistical data. Tweets are gathered up to 96 hours prior to the start of the game, and divided into two windows: a 96-hour window and a 24-hour window. However, if a team plays on a Thursday night, there is a chance their previous game could fall within the 96 hour window and in that case tweets are collected four hours after the start of the previous game to ensure no in-game tweets are included. The sentiment for the 96-hour windows is calculated for each team and game using Eq. (1). The sentiment for the 24-hour window is calculated using the same equation but with the appropriate time frame. To measure the relative change in sentiment and identify any surges or drops, the “swing” statistic is calculated using Eq. (2).

$$Sentiment_{Team,96hr} = \frac{Sentiment_{Team,\,96hr,\,Pos}}{Sentiment_{Team,\,96hr,Total}} - \frac{Sentiment_{Team,\,96hr,\,Neg}}{Sentiment_{Team,\,96hr,Total}} \quad (1)$$

$$Swing_{Team} = Sentiment_{Team,\,24hr} - Sentiment_{Team,96hr} \quad (2)$$

3.4 Data Preprocessing

For each game, features are split into home and away such as *home team offensive passing yards* and *away team offensive passing yards*. A binary target representing whether the away team wins or loses with the spread (WTS) is created according to Table 4. It should be noted that one game in the dataset resulted in a push, where the spread was equal to the difference in score. Pushes, which result in no winners or losers, are extremely rare and so this datapoint is removed as the prediction of a push is out of the scope of this study.

Prior to feeding the data into any NNs, normalization is necessary to avoid issues with gradient updates and ensure network convergence. It is not recommended to input data with large values or heterogeneous data with varying ranges, as this can cause problems. Therefore, the features in the dataset are normalized to have a mean of 0 and a standard deviation of 1. This will also facilitate future feature comparison. The only missing values found in the dataset were for features where a 0 was expected, so the empty values were replaced with 0s. To prevent multicollinearity, the team names and day of games are one hot encoded, and one value is dropped.

Table 4

Calculations for determining the target variable of winner or loser with the spread. Points scored by the away team are added to the away spread and compared to the points scored by the home team.

Result	Value	Calculation
Away Winner (WTS)	1	$Pts_{Away} + Spread_{Away} > Pts_{Home}$
Away Loser (WTS)	0	$Pts_{Away} + Spread_{Away} < Pts_{Home}$

3.5 Datasets

3.5.1 Football Features

Multiple football statistics datasets are constructed using averages of a certain number of previous games a team has played. Landers (2019) uses feature statistics calculated as an average of the season up until that current week in their models. Beal (2020) includes this average as well in addition to last season averages. In this research, models only include current season data due to the limited availability of Twitter data. However, it is believed that due to the high rate of player and coach turnover in the offseason, previous seasons data may not be very relevant to current season predictions. This is further supported by Hong's finding that models were better able to predict the second half of seasons. Bad teams may fire coaches and cut players while successful teams may have to let go of star players that may not fit in the salary cap structure.

It is difficult to know intuitively how many previous games are the most indicative of a teams' next performance. Whether it be a whole season worth of data, the last few games or simply just the last game played. Star player injuries can significantly affect a team's chances to win and this could be captured in shorter term average windows while the effects of outlier performances could be minimized with longer windows. For this reason seven different averages, varying the number of games and the weights of games, are created according to Table 5**.** The use of a weighted average gives higher weighting to a team's most recent performance. The effectiveness of the seven different average window datasets are tested for each classifier without sentiment data to identify the best football features.

Table 5

No-sentiment datasets, included averages both standard and weighted of a team's performance over the last 1, 3, 7 games and whole season up to the current game.

Dataset	Description
1 Game	Last game
3 Games	Last 3 games (unweighted mean)
3 Games (wtd)	Last 3 games (weighted mean)
7 Games	Last 7 games, (unweighted mean)
7 Games (wtd)	Last 7 games (weighted mean)
Season	Current seasons games up until current week (unweighted mean)
Season (wtd)	Current seasons games up until current week (weighted mean)

3.5.2 Sentiment Features

Four different models will be constructed to test the effectiveness of Twitter sentiment data. The baseline dataset does not include any Twitter sentiment features. Datasets two and three will contain the Twitter sentiment calculated within the 96-hour and 24-hour windows, respectively. A fourth dataset includes the Twitter sentiment swing feature, a measure of the relative change in Twitter sentiment prior to each game. These datasets are described in Table 6.

Table 6

Sentiment test datasets, three datasets with each sentiment feature: 96-hour, 24-hour, and swing, plus a fourth baseline dataset with no-sentiment features.

Dataset	Description
No Sentiment	Baseline data (football performance features only)
96-Hour Sentiment	Baseline data + 96-hour sentiment ratio from Eq. (1)
24-Hour Sentiment	Baseline data + 24-hour sentiment ratio from Eq. (1)
Swing Sentiment	Baseline data + Swing sentiment from Eq. (2)

Sinha (2013) excluded the final week in their model because some games played will have no bearing on determining which teams will make the playoffs. Some teams have already solidified their playoff berth and choose to rest their star players while others have been eliminated. However, this study includes the final week to increase the amount of data available. Even if games do not have playoff implications, players and coaches have a strong incentive to win in order to maintain their jobs for the following year. Bookkeepers must set a spread with all available information, and bettors still have the potential to win money.

3.6 Model Creation and Tuning

3.6.1 AdaBoost

AdaBoost, or Adaptive Boosting, was introduced by Freund (1999). The algorithm combines multiple weak learners in order to create a single strong classifier. Weights are assigned to each training example, and after each iteration, the weights are adjusted. Misclassified data is given higher weights and iterations continue until a strong enough classifier is achieved. The AB package from scikit-learn will be used to construct and test this model (scikit-learn, 2023)).

3.6.2 Neural Network

Neural networks are a ML technique that are modeled after the human brain. They contain many interconnected neurons that can recognize patterns. They contain an input layer, numerous hidden layers, and an output layer. NNs may be well suited to this task due to the large amount of predictive features however the small amount of data could be limiting. There are many different types of NN layer patterns that can be used depending on the classification problem. Beal (2020) uses a NN structure that contains a multi-layer perceptron (MLP) with ‘relu’ activation function and 100 layers. Khan (2003) also uses a back propagation MLP which allows for the use of supervised learning. The NN model will be constructed using Keras Tensorflow and the Keras random search tuner package was used to narrow the field of possible network architectures and parameters (TensorFlow, 2023). The results of this search are used to conduct a fore focused grid-search to find the optimal network architecture and hyperparameters. Top results included 3 to 5 blocks of paired dense and dropout layers.

3.6.3 Random Forest

Random Forests make use of multiple decision trees to produce more accurate and stable predictions in comparison to a single decision tree (Brieman L., 2001). They use bagging where multiple trees are trained on random samples of different subsets of the data in order to reduce correlation. Bagging offers an advantage over boosting as it can be done in parallel. Similar to AB, multiple weak learners are combined to form one strong learner. The RF Classifier from scikit-learn will be used to construct the model (scikit-learn, 2023).

3.6.4 Naive Bayes

Naive Bayes classifiers are based on the Bayes theorem with the underlying assumption of conditional independence between every pair of features. Even with this simplified assumption this classifier still can perform quite well for many classification tasks. The simplified assumptions allow for NB classifiers to be trained quickly and require only small amounts of data (scikit-learn, 2023). The scikit-learn NB Classifier will be used to train this model.

3.6.5 XGBoost

XGBoost is an open source scalable decision tree boosting system (Chen, 2016) that was selected for this study due to the success of Landers (2019). Their study achieved an impressive accuracy of 58.1% in predicting winners WTS using a boosted decision tree classifier. To optimize the performance of the model, a grid search method is utilized to find the best hyperparameters for each dataset. The xgboost python library is used to implement this classifier (XGBoost, 2023) .

3.7 Training and Evaluation

The choice of the appropriate training and evaluation methodology is an essential aspect of any machine learning research, and is generally selected based on the characteristics of the dataset. This research contains a relatively small number of datapoints, specifically just 105. To maximize the use of the available data, repeated *k*-fold cross-validation is used. *K*-fold cross-validation divides the data into *k* subsamples. *K* models are then trained using *k*-1 folds, and evaluated using the final test fold which changes for each test. The final accuracy is calculated as the average accuracy of all test sets over the *k* folds. Results can still be noisy and repeating this process helps to further reduce variance. Repeated *k*-fold repeats the same process for a specified amount of times, with the folds being reshuffled after each repeat to reduce the error of the estimate. Stratified *k*-fold cross-validation is not needed because the dataset is balanced, with 49.5% of away teams winning WTS and 50.5% of home teams winning.

The choice of the number of folds to use has tradeoffs, mainly in training time and performance of the model. If *k* is too small, this could result in high variance and insufficient predictive power, while if *k* is too large, the model could overfit and training times be very long. In this study, we follow Beal's (2020) approach and use 10 folds, a common choice for *k*-fold cross-validation. Landers (2019) uses three folds, one for each season of data, while Sinha (2013) employs time-series cross-validation, where previous weeks are used as training folds and the latest week as the holdout test fold. Sinha's strategy makes the most sense so as to avoid including any future game information to predict the result of past games. For example, the results of games played in 2017 would not be available to a model or bettor making bets on games played in 2016. However, due to the limited data size in this study, time-series cross-validation is not feasible. Therefore, the use of ten fold cross-validation with ten repeats is used to evaluate the classifiers.

A hold out test set is not used due to the small amount of training data. Doing so would greatly reduce the amount of training data and likely result in reduced model performance. The lack of a hold out test set could result in the overestimation of model performance. Models are evaluated based on the average accuracy over ten repeats of a ten fold cross-validation procedure, for a total of 100 recorded accuracies per model. As this is a binary classification

problem with a balanced dataset, accuracy is an appropriate evaluation metric. Moreover, there is no preference between the two outcomes, whether the away team wins or loses WTS. If the away team is predicted to not cover the spread, then the home team is predicted to win WTS and should be bet on.

4. Experiments and Discussion

In this section, the experiments and results of the study are presented and discussed. The best average accuracy of each tuned classifier and dataset is reported after a repeated *k*-fold cross-validation process. The 100 model accuracies produced from this process are used to determine whether there is a statistically significant difference between model accuracies.The results are used to draw conclusions about this study's research questions. 105 games from the 2022 season were used in training and testing the models. All significance testing is conducted using the python SciPy stats package (SciPy, 2023).

4.1 Baseline Models With No Sentiment

Initial tests were conducted on the football features to determine the best baseline data to use for each classifier, prior to the addition of Twitter sentiment features. Classifiers were trained on the seven football performance datasets (as described in section 3.5.1) and tuned using a grid search procedure.

Table 7

Accuracy results and top dataset for the best model each classifier after training and testing on the no-sentiment datasets

	Accuracy	
	Mean	Std. Dev
AdaBoost (1 game)	**61.7%**	**15.7%**
Neural Network (1 game)	57.9%	13.4%
Naive Bayes (7 games)	55.4%	13.0%
Random Forest (ssn wtd.)	54.6%	14.7%
XGBoost (1 game)	53.9%	13.7%

The best dataset specific to each classifier is then combined with Twitter sentiment features to create three new datasets. The results presented in Table 7 show that while there was no single dataset that performed the best across all classifiers, the 1-game dataset was the best choice for AB, NN, and XGB. The best baseline classifier was an AB model with an average accuracy of 61.74%. Appendix B shows the results of all baseline no-sentiment testing.

4.2 Experiment 1: Profitable Accuracy and Leading Work Comparisons

4.2.1 Profitable Accuracy

In this experiment, we test whether ML classifiers trained on a combination of performance of Twitter sentiment data can exceed the accuracy needed to be profitable when making point spread bets. The model must pick the correct winner over 52.4% of the time. T-tests are conducted to evaluate whether the average accuracy of each classifier is significantly greater than 52.4% and results are presented in Table 8. The AB, NB, NN and XGB accuracies were all significantly greater (p-values < 0.01), while the RF classifier was not. We can conclude that AB, NB, NN, and XGB models using both football and Twitter sentiment features can be profitable making point spread bets with 99% certainty. Looking more closely at the RF p-value of 0.052, we can conclude that with 90% certainty a RF model would also be profitable.

Table 8

T-test results of whether the best average model accuracy is significantly higher (α = 0.05) than the accuracy needed to be profitable (52.4%). All models are significantly higher at this level except for RF.

Model	Dataset	tstat	p-value	Accuracy	Reject Ho
AdaBoost	1 Game Cross Sent.	9.367	p < 0.001	0.652	TRUE
Naive Bayes	7 Games 24h Sent.	3.139	0.0011	0.566	TRUE
Neural Network	1 Game 24h Sent.	3.846	p < 0.001	0.578	TRUE
Random Forest	Ssn Wtd Cross Sent.	1.640	0.052	0.548	FALSE
XGBoost	1 Game Cross Sent.	3.641	p < 0.001	0.579	TRUE

4.2.2 Leading Research Benchmark Accuracy

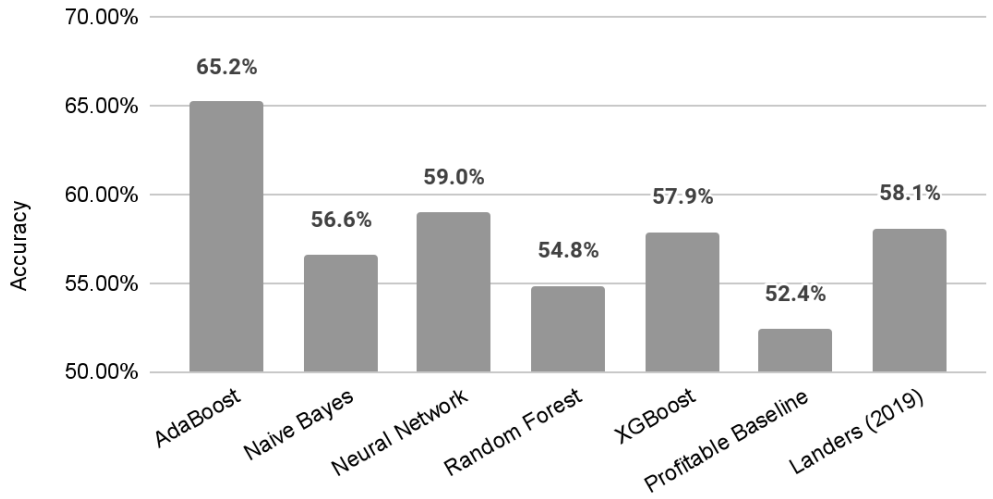
The best accuracy previously achieved predicting NFL spread outcomes was achieved by Landers (2019). An average accuracy of 58.1% was reported using a boosted decision tree model predicting games over three seasons. The top performing model of this study, an AB classifier using data from 1-game and Twitter sentiment swing features, achieved an accuracy of 65.21%. This result is significantly higher than the accuracy achieved by Landers. The AB classifier with no sentiment features was also significantly greater with a recorded accuracy of 62.0%. Results are presented in Chart 3.

Landers recorded a top accuracy of 59.4% for a single season and it is worth comparing this result as this study is only looking at one season, which may be a more predictable season. The results of a significance test again conclude that both the swing feature and no-sentiment feature AB models were significantly more accurate than Landers' top season with p-values of 0.24E-5 and 0.047 respectively. While no other model produced accuracies significantly higher than this benchmark, only RF models produced model accuracies that were all significantly lower. This shows that NB, NN, and XGB classifiers using both football and Twitter sentiment data are able to produce results competitive to that of the leading work.

These results are important as they extend previous research using ML classifiers to predict NFL outcomes. The evidence presented shows that these classifiers are capable of producing significant results when using both Twitter sentiment and football data. There is strong evidence that the AB model is significantly higher than leading work. While this experiment was limited to only eight weeks of the 2022 NFL season, it did include the final week of the season in the data which previous research did not, citing the difficult nature of prediction when teams have nothing to play for. This study also includes a week of playoff games, data which has not been included in previous studies. The lack of a holdout test set could mean reported accuracies may be overestimated, however significance testing shows that even when compared to the best reported single season accuracy of Landers, the AB classifier accuracy is significantly greater with over 99% confidence.

Chart 3

Model prediction accuracy comparisons to profitable baseline and leading work. AB, NB, NN, XGB were significantly higher than the profitable baseline and AB significantly higher than Landers' leading work.



4.3 Experiment 2: Effectiveness of Twitter Sentiment Data

This experiment is to determine whether the addition of Twitter sentiment features will improve the accuracy of ML models in predicting point spread winners. Models without sentiment features are compared to three models incorporating Twitter sentiment features: 96-hour pregame sentiment, 24-hour pregame sentiment, and the swing sentiment. To create new models, the three sentiment features are added to the top baseline dataset reported for each classifier. A fourth dataset with no Twitter sentiment features is included as the baseline. Each of these four datasets is uniquely trained and tuned for each classifier. The accuracy from the repeated *k*-fold cross-validation processes are used as the final model accuracy with results shown in Chart 4. Detailed results are shown in Appendix C.

The NN was the only classifier where the no-sentiment dataset had the highest accuracy while the other 4 classifiers included Twitter sentiment features. The average for each dataset across all classifiers is included in the final column to show that the no-sentiment dataset accuracy was on average lower than swing but higher than 96-hour and 24-hour sentiments. To determine the statistical significance, an ANOVA test for each classifier is performed and results presented in Table 9.

Chart 4

Model accuracies for each sentiment dataset and classifier. Average of each dataset across all classifiers included in the last column. Swing sentiment has the highest overall average accuracy, topping 3 of 5 classifiers.

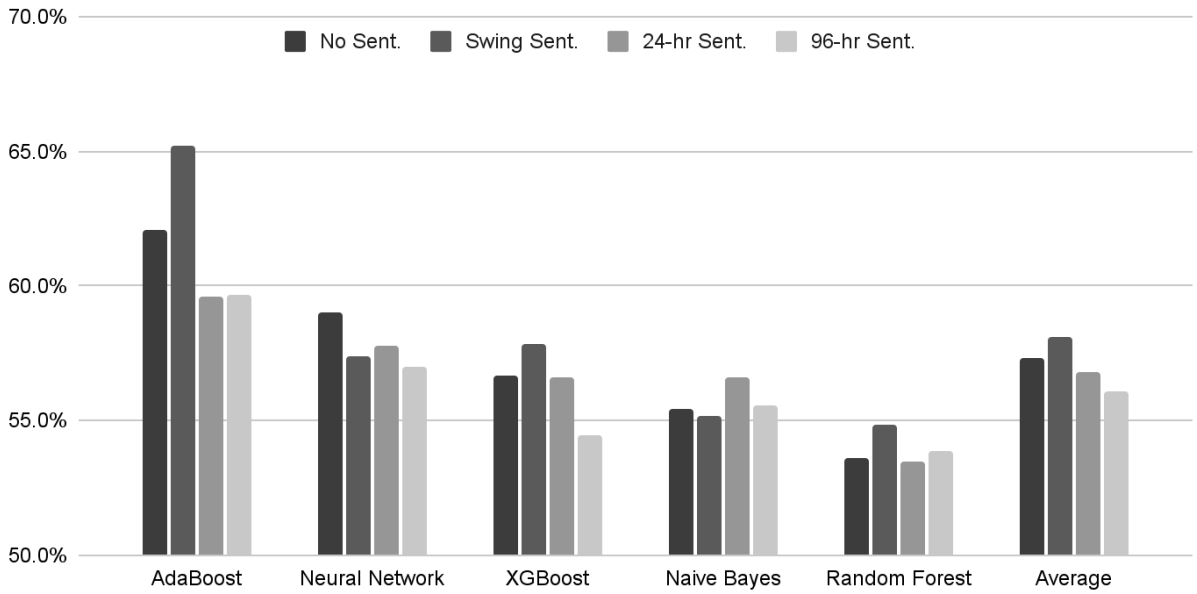


Table 9

Results of ANOVA testing of the difference between dataset average accuracies for each classifier. The null hypothesis that the average of all means are equal is rejected at an alpha level of 0.01

ANOVA Test of Difference Between Group Means		
	Statistic	p-value
AdaBoost	38144510.92	$p < 0.001$
Naive Bayes	47.48	$p < 0.001$
Neural Network	33.68	$p < 0.001$
Random Forest	29.89	$p < 0.001$
XGBoost	560.39	$p < 0.001$

The null hypothesis of this test is that there is no statistical difference between all datasets, meaning that none of the average accuracies are significantly different in the test group. For each classifier, this is rejected in favor of the alternative that at least two dataset accuracies are significantly different. When comparing test statistics, AB had a statistic three orders of magnitude higher than all other models. We can infer from this that the AB classifier is able to extract significantly more information from the sentiment data than other models, resulting in bigger differences between dataset accuracies.

A post-hoc Tukey test compares individual pairs of the group of datasets to determine any statistically significant differences in accuracies. While the ANOVA test showed significant difference between all the mean accuracies, the Tukey tests failed to find significant differences between individual models, with the exception of the AB models. Significance testing results found the swing model to be significantly different from the 96-hour and 24-hour models. Accuracies of the AB models are shown in Chart 5 and the results of the significance testing in Table 10. The remainder of these findings are presented in Appendix D.

Chart 5

AdaBoost classifier model accuracies. Swing sentiment achieved the highest accuracy overall (65%) and no-sentiment second highest (62%).

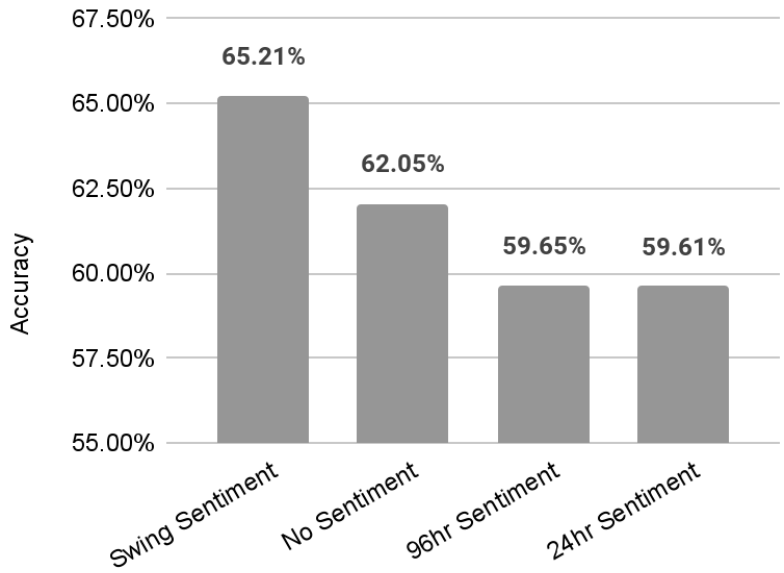


Table 10

Significance testing of differences in individual Adaboost models. Swing sentiment was significantly higher than both 24 and 96 hour sentiment.

Datasets		Mean Diff.	p-value	Reject Ho (α=0.05)
24h Sent.	96h Sent.	0.0004	1.0000	FALSE
24h Sent.	Swing Sent.	0.0560	0.0355	TRUE
24h Sent.	No Sent.	0.0245	0.6384	FALSE
96h Sent.	Swing Sent.	0.0556	0.0373	TRUE
96h Sent.	No Sent.	0.0241	0.6495	FALSE
Swing Sent.	No Sent.	-0.0315	0.4234	FALSE

The result of the ANOVA test finding group differences but Tukey tests failing to find significant differences between pairs is likely due large variances in model accuracies combined with the reduced statistical power of the test from the comparison of four different groups. Future testing might have greater success determining significant differences between the datasets with more data and reduced noise in the results, lowering the variance of model accuracies. There were however significant differences in the AB models in spite of the high variances. The accuracies in Chart 5 show the swing model was 3.1% more accurate than no-sentiment but this difference did not fall outside random chance. Overall, it cannot be concluded that the increase in model accuracy from the addition of Twitter sentiment features is not from random chance. However, four of the five classifiers' top datasets were aided with the use of sentiment features. The addition of more data in future studies could potentially find a more significant result for this problem.

4.4 Experiment 3: Comparison of ML Classifiers

This experiment is performed to determine whether different ML classifiers' are better able to predict winner WTS outcomes of NFL games. Five different classifiers are examined with each having been trained and tested with multiple datasets, with the best result being used for this experiment. The final accuracies are compared using an ANOVA test of the null hypothesis that there is no significant difference between model accuracies. This is rejected at the 95% confidence level in favor of the alternative that at least two accuracies are significantly different.

Table 11

Results of Tukey test of significant differences between classifiers. AdaBoost was significantly more accurate than all other classifiers.

		Mean diff	p-adj	Reject Ho (α = 0.05)
AdaBoost	Naive Bayes	-0.0862	0.0002	**TRUE**
AdaBoost	Neural Network	-0.0621	0.0173	**TRUE**
AdaBoost	Random Forest	-0.1036	0	**TRUE**
AdaBoost	XGBoost	-0.0733	0.0026	**TRUE**
Naive Bayes	Neural Network	0.0241	0.7492	FALSE
Naive Bayes	Random Forest	-0.0175	0.907	FALSE
Naive Bayes	XGBoost	0.0129	0.9675	FALSE
Neural Network	Random Forest	-0.0415	0.2321	FALSE
Neural Network	XGBoost	-0.0112	0.9808	FALSE
Random Forest	XGBoost	0.0304	0.5517	FALSE

A post-hoc Tukey test is again used to determine which individual models are significantly different from one another and the results reported in Table 11. AB was the most accurate classifier (65.2%), followed by NN (59.0%), XGB (57.9%), NB (56.6%), and lastly RF (54.8%), presented in Chart 3. The results of this testing find that the AB's accuracy is significantly higher than every other classifier. It may be worth exploring this comparison with each classifier trained on a single dataset of football statistics rather than one chosen specifically to maximize the accuracy for each classifier. This would make any differences in results more clearly attributed to differences in classifier rather than dataset.

4.5 Experiment 4: Swing Sentiment

The final experiment was conducted to determine whether the technical charting method of measuring sentiment would significantly increase model accuracy. The XGB, AB and RF classifiers used the swing sentiment dataset to achieve their most accurate models. To test the significance of these results, a two-sample t-test was conducted for each classifier. The null

Table 12

Results of t-tests comparing model accuracies of swing sentiment models to no sentiment models

	No Sent. (%)	Swing Sent. (%)	t-statistic	p-value	reject Ho
Naive Bayes	55.4%	55.2%	0.1583	0.8744	FALSE
Random Forest	53.6%	54.8%	0.5810	0.5619	FALSE
Neural Network	59.0%	57.4%	0.7747	0.4394	FALSE
AdaBoost	62.1%	65.2%	1.5138	0.1317	FALSE
XGBoost	56.7%	57.9%	0.5731	0.5672	FALSE

hypothesis that there is no difference in average accuracy from the swing dataset and the no-sentiment baseline is tested at the 95% confidence level and results shown in Table 12**.** AB had the largest test statistic and a p-value of 0.1317, meaning that there is over an 85% chance that the result is not due to chance, but this is still short of the 95% needed to be a definitive result. The AB classifier swing dataset was shown to be significantly more accurate than the 96-hour and 24-hour AB sentiment models in Experiment 1. This evidence supports Schumaker's 2017 findings that sentiment features that capture relative changes are better predictors than those that don't take into account the baseline differences between teams and games.

While the results show some evidence that a swing sentiment statistic does improve model accuracies predicting winners WTS, the results fall short of a statistically significant conclusion. Once again, a larger dataset has the potential to reduce the variance of the resultant average accuracies and produce more statistically significant results. We are unable to conclude that the swing feature significantly increases model accuracy over no-sentiment, however it was a significant improvement over the 96 and 24-hour sentiment datasets for AB, evidence that this feature is a stronger predictor.

5. Conclusions and Future Work

In this paper, experiments were done to test the effectiveness of machine learning classifiers and Twitter sentiment features to predict winners with the spread of NFL games. Seven football datasets, five classifiers and three unique Twitter sentiment features were evaluated. The study found conclusive evidence that ML classifiers trained on a combination of

football and Twitter sentiment data can be profitable. AB, NB, NN, and XGB models had results significantly higher than this benchmark accuracy (52.4%) at the 99% confidence level and RF at the 90% level. The top-performing AB model accuracy (65.2%) was significantly higher than the leading work (58.1%) while NN (59.0%) and XGB (57.9%) were statistically equivalent to this benchmark accuracy.

This study evaluated the inclusion of Twitter sentiment features but failed to find enough evidence to conclude the efficacy of these features. Even though a significant result was not reached, there is evidence to support the future study of sentiment swing features. The best models for four of the five classifiers tested included Twitter sentiment features, three of which were the swing measure. The swing feature was also used in the most accurate model of the study. The AB swing sentiment model was significantly more accurate than the 96-hour and 24-hour sentiment datasets, supporting the use of relative measures of sentiment for prediction over fixed. The study concluded that AdaBoost was the most effective classifier of those studied, producing two models: swing-sentiment (65.2%) and no-sentiment (62.1%), which were both significantly more accurate than any other classifiers.

In summary, our findings indicate that ML classifiers are effective in producing profitable models utilizing readily available football statistics and Twitter sentiment data. Twitter sentiment features resulted in higher accuracies than models relying strictly on football data, but the difference was not great enough to confidently conclude the improvements were outside random chance. This study was novel in its approach of using Twitter sentiment features, specifically a swing feature to measure relative changes in sentiment, within ML models to predict spread outcomes and the models produced achieved accuracies significantly higher than previous work. This study also included the final week of the season and playoff game data. The limitations of collecting Twitter data led to a smaller dataset and higher variance in the results. Further examination with more data would increase the power of significance tests, reducing the chance of type II errors while also allowing for the use of a holdout test set, lowering the chance of the results being overestimated. The use of time-series cross-validation to predict games on a weekly basis using past data would be a valuable extension of this research. Additionally, testing these findings on other sports, such as European football, could further enhance the generalizability of the findings.

Word Count: 9301

6. References

Barbieri, F., Ballesteros, M., & Saggion, H. (2020). TweetEval: Unified Benchmark for Tweet Classification. arXiv preprint arXiv:2010.12421.

Beal, A., & Lombriser, R. (2020). A Comparison of Machine Learning Classifiers for Predicting NFL Outcomes. In Proceedings of the 2nd International Conference on Machine Learning and Computing (pp. 239-243).

Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

Business of Apps. (n.d.). Twitter Statistics (2022) & Usage Facts. Retrieved February 22, 2023, from https://www.businessofapps.com/data/twitter-statistics/.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794. doi: 10.1145/2939672.2939785

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1), 119-139.

Glickman, M. E., & Stern, H. S. (1998). A state-space model for National Football League scores. Journal of the American Statistical Association, 93(441), 25-35.

Hagen, M., Potthast, M., & Stein, B. (2015). A Webis Ensemble for Twitter Sentiment Detection. In Proceedings of the 8th International Workshop on Semantic Evaluation (pp. 747-751).

Harville, D. (2012). Predictions for National Football League Games Via Linear-Model Methodology. Journal of Quantitative Analysis in Sports, 8(2), Article 5. doi:10.1515/1559-0410.1399

Hong, L., Davison, B. D., & Golbeck, J. (2010). Predicting the NFL Using Twitter. In Proceedings of the 3rd International Conference on Weblogs and Social Media (pp. 192-199).

Hugging Face. (n.d.). Sentiment analysis model using RoBERTa on Twitter data. Retrieved January 20, 2023, from https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest.

Khan, K. A., Ahmad, F., & Imran, M. (2003). Neural network prediction of NFL football games. Journal of computing sciences in colleges, 19(6), 273-280.

Kharde, V. (2016). Sentiment Analysis of Twitter Data. International Journal of Computer Applications, 146(11), 1-5.

Landers, R. N., & Liao, C. (2019). Machine Learning Approaches to Fantasy Football Performance. Journal of Applied Sport Management, 11(1), 24-33.

Liu, B., Li, X., Li, W., Li, T., & Liu, B. (2019). Fine-tuned Language Models for Text Classification. arXiv preprint arXiv:1907.11692.

NFL Picks Pro. (2022, September 7). 2022 NFL Team Hashtags for Twitter. Retrieved January 23, 2023, from https://nflpickspro.com/news/2022-nfl-team-hashtags-for-twitter/.

PBS NewsHour. (2022, February 4). Super Bowl betting estimated to reach $16 billion. Retrieved February 10, 2023, from https://www.pbs.org/newshour/nation/super-bowl-betting-estimated-to-reach-16-billion

Pro-Football-Reference.com. (n.d.). Pro-Football-Reference.com. Retrieved December 12, 2022, from https://www.pro-football-reference.com/.

Schumaker, R. P., & Chen, H. (2016). Predicting Wins and Spread in the Premier League Using Sentiment Analysis of Twitter Data. Journal of Prediction Markets, 10(1), 63-78.

Schumaker, R. P., & Zhang, H. (2017). A Study of NFL Twitter Sentiment Using Technical Stock Market Charting Techniques. Journal of Computer Science and Applications, 5(4), 123-132.

scikit-learn. (n.d.). Scikit-learn. Retrieved January 12, 2023, from https://scikit-learn.org.

SciPy. (n.d.). SciPy v1.7.1 Reference Guide. Retrieved January 29, 2023, from https://docs.scipy.org/doc/scipy/index.html.

Selenium with Python. (n.d.). Selenium with Python documentation. Retrieved December 15, 2022, from https://selenium-python.readthedocs.io/.

Sinha, A., Khattar, A., & Varshney, L. R. (2013). Predicting NFL Games Using Twitter. In Proceedings of the 2nd International Conference on Advances in Computing, Communication and Control (pp. 183-188).

SportsBettingDime. (2022, January 12). Global Sports Betting Market Projected to Reach Nearly $155 Billion by 2024. Retrieved February 10, 2023, from https://www.sportsbettingdime.com/guides/finance/global-sports-betting-market/

Sportsecyclopedia. (n.d.). Betting History: The Tank Scandal. Retrieved February 10, 2023, from https://sportsecyclopedia.com/tank/tank/bettinghistory.html

TensorFlow. (n.d.). TensorFlow. Retrieved January 5, 2023, from https://www.tensorflow.org/.

Tweepy. (n.d.). Tweepy: Twitter for Python. Retrieved January 23, 2023, from https://www.tweepy.org/.

Wunderlich, N., & Wieneke, A. (2022). Twitter during Soccer Matches: Do Tweets Contain Information for In-Play Forecasting?. International Journal of Sports Science & Coaching, 17(1), 139-147.

Wunderlich, N., Schanz, M., & Wieneke, A. (2020). Lexicon-based Sentiment Analysis as a Tool to Analyze Sports Twitter. International Journal of Sports Science & Coaching, 15

XGBoost Documentation. (n.d.). XGBoost 1.6.1 documentation. Retrieved January 19, 2023, from https://xgboost.readthedocs.io/en/stable/.

Zimbra, D. K. (2018). State of the Art in Twitter Sentiment Analysis. Journal of Computer Science and Applications, 6(1), 1-6.

Appendices

Appendix A: Feature Descriptions

Team Features		Offense and Defense Features	
Feature	Description	Feature	Description
Time of Possession	Amount of game time a team is on offense	**Pass Completions**	# of completed passing plays
# Penalties	Team's number of penalties	**Pass Attempts**	# of attempted passing plays
Penalty Yards	Team's total penalty yards	**Pass Yards**	# of passing yards gained
# Opp Penalties	Opponent's number of penalties	**Pass TDs**	# of passing touchdowns scored
Opp Penalty Yards	Opponent's number of penalty yards	**Sacks**	# of times the quarterback tackled for a loss
Team	One-hot encoded team name	**Sack Yards**	# of yards lost from sacks
Day of Week	One-hot encoded day of week	**Rush Attempts**	# of rushing plays
Spread	The point spread	**Rush Yards**	# of rushing yards
3rd Down Attempts	# of attempted 3rd down conversions	**Rush TDs**	# of rushing touchdowns scored
3rd Downs Made	# of successful 3rd down conversions	**1st Downs**	# of first downs gained
Pass QB Rating	Quarterback passer rating statistic used to measure QB performance	**Turnovers**	# of times ball was turned over to opponent by fumble or interception
		Points Scored	Total points scored and given up
		Expected Points	Estimated point value at the start of a given play, based on down, distance, and field position (sports-reference.com)

Appendix B: Baseline Dataset Experiment Results

Final accuracies recorded from 10 repeated 10 fold cross validation on each classifier with different football datasets. Hyperparameters are included.

Naive Bayes	Accuracy	Std. Dev	Smoothing
7 Games	0.5545	0.1299	5.34E-01
7 Games Wtd.	0.5545	0.1299	5.34E-01
1 Game	0.5495	0.1454	1.00E-04
Season	0.5420	0.1341	1.87E-02
Season Wtd.	0.5420	0.1341	1.87E-02
3 Games	0.5390	0.1478	1.23E-09
3 Games Wtd	0.5390	0.1478	1.23E-09

AdaBoost	Accuracy	Std. Dev.	N Est.	Base Est. Depth	Learning Rate
1 Game	0.6174	0.1567	100	1	0.1
7 Games	0.5631	0.1458	10	1	10
7 Games Wtd.	0.5631	0.1458	10	1	10
3 Games	0.5535	0.1539	100	1	10
3 Games Wtd.	0.5535	0.1539	100	1	10
Season	0.5309	0.1270	5000	1	1
Season Wtd.	0.5309	0.1270	5000	1	1

Neural Network	Accuracy	Std. Dev.	# Layers	# Units	Dropout Rate	Learning Rate
1 Game	0.5790	0.1336	4	256	0.3	1.85E-03
Season	0.5470	0.1500	4	256	0.3	3.00E-02
7 Games	0.5420	0.1305	2	2048	0.3	1.30E-02
7 Games Wtd	0.5410	0.1530	10	128	0.1	1.00E-04
Season Wtd	0.5260	0.1604	2	2048	0.3	1.30E-02
3 Games Wtd	0.5110	0.1568	4	256	0.3	1.85E-03
3 Games	0.5030	0.1565	4	256	0.3	1.85E-03

Random Forest	Accuracy	Std. Dev.	N Est.	Max Depth	Min Sample Split	Min Sample Leaf
Season Wtd.	0.5455	0.1470	5	75	8	4
7 Games	0.5374	0.1274	8	150	2	8
Season	0.5349	0.1461	5	50	2	2
1 Game	0.5335	0.1567	8	100	4	2
7 Games Wtd.	0.5299	0.1332	10	10	4	4
3 Games	0.4926	0.1553	8	100	12	4
3 Games Wtd.	0.4908	0.1574	5	15	8	4

XGBoost	Accuracy	Std. Dev.	Eta	Depth	Gamma
1 Game	0.5395	0.1375	0.75	1	3
7 Games	0.5216	0.1336	1	1	0.5
7 Games Wtd.	0.5216	0.1336	1	1	0.5
Season	0.5149	0.1329	1	2	0
Season Wtd.	0.5149	0.1329	1	2	0
3 Games	0.5105	0.1345	0.5	2	1
3 Games Wtd.	0.5105	0.1345	0.5	2	1

Appendix C: Experiment 2 Results

Final accuracies recorded from 10 repeated 10 fold cross validation on each classifier with different Twitter sentiment features and a baseline no-sentiment dataset. Hyperparameters are included.

Naive Bayes	Accuracy	Std. Dev	Smoothing
24 hr Sentiment	0.5659	0.1328	0.3511
96 hr Sentiment	0.5556	0.1299	0.4329
No Sentiment	0.5545	0.1299	0.5337
Swing Sentiment	0.5515	0.1287	0.4329

AdaBoost	Accuracy	Std. Dev.	N_estimators	Base Est Depth	Learning Rate
Swing Sentiment	0.6521	0.1361	75	1	0.1
No Sentiment	0.6205	0.1564	75	1	0.1
96hr Sentiment	0.5965	0.1478	75	1	0.1
24hr Sentiment	0.5961	0.1410	75	1	0.05

Neural Network	Accuracy	Std. Dev.	# Layers	# Units	Dropout Rate	Learning Rate
No Sentiment	0.5900	0.1360	4	256	0.3	1.00E-03
24 hr Sentiment	0.5780	0.1397	5	512	0.3	1.85E-03
Swing Sentiment	0.5740	0.1540	3	256	0.2	1.85E-03
96 hr Sentiment	0.5700	0.1459	3	256	0.2	1.85E-03

Random Forest	Accuracy	Std. Dev.	N Est.	Max Depth	Min Sample Split	Min Sample Leaf
Swing Sentiment	0.5485	0.1484	10		8	8
96 hr Sentiment	0.5385	0.1407	5	15	4	1
No Sentiment	0.5361	0.1510	10	75	2	1
24 hr Sentiment	0.5349	0.1530	8	200	4	8

XGBoost	Accuracy	Std. Dev.	Eta	Depth	Gamma	Alpha	Lambda	Subsample
Swing Sentiment	0.5788	0.1498	0.1	1	0	0	0	1
No Sentiment	0.5669	0.1425	0.1	1	0	0	0	1
24 hr Sentiment	0.5660	0.1539	0.05	2	0.3	0.25	0	1
96 hr Sentiment	0.5449	0.1356	0.1	1	0	0	0.25	1

Appendix D: Experiment 2 Significance Tests

Pairwise Tukey test comparing to determine which models were significantly different. The only models showing a significant difference were the AB Swing Sentiment was significantly greater than both the 24 hour sentiment and the 96 hour sentiment models.

AdaBoost (1 Game)		**Mean Diff.**	**p-value**	**Reject Ho (alpha = 0.05)**
24h Sent.	96h Sent.	0.0004	1.0000	FALSE
24h Sent.	**Swing Sent.**	**0.0560**	**0.0355**	**TRUE**
24h Sent.	No Sent.	0.0245	0.6384	FALSE
96h Sent.	**Swing Sent.**	**0.0556**	**0.0373**	**TRUE**
96h Sent.	No Sent.	0.0241	0.6495	FALSE
Swing Sent.	No Sent.	-0.0315	0.4234	FALSE
Naive Bayes (7 Games)		**Mean Diff.**	**p-value**	**Reject Ho (alpha = 0.05)**
24h Sent.	96h Sent.	-0.0103	0.9453	FALSE
24h Sent.	Swing Sent.	-0.0144	0.8656	FALSE
24h Sent.	No Sent.	-0.0115	0.9262	FALSE
96h Sent.	Swing Sent.	-0.0041	0.9962	FALSE
96h Sent.	No Sent.	-0.0012	0.9999	FALSE
Swing Sent.	No Sent.	0.0029	0.9986	FALSE
Random Forest (Ssn. wtd.)		**Mean Diff.**	**p-value**	**Reject Ho (alpha = 0.05)**
24h Sent.	96h Sent.	0.0035	0.9983	FALSE
24h Sent.	Swing Sent.	0.0135	0.9182	FALSE
24h Sent.	No Sent.	0.0012	0.9999	FALSE
96h Sent.	Swing Sent.	0.0100	0.9647	FALSE
96h Sent.	No Sent.	-0.0024	0.9995	FALSE
Swing Sent.	No Sent.	-0.0124	0.9362	FALSE
Neural Network (1 Game)		**Mean Diff.**	**p-value**	**Reject Ho (alpha = 0.05)**
24h Sent.	96h Sent.	-0.0080	0.9797	FALSE
24h Sent.	Swing Sent.	-0.0040	0.9974	FALSE
24h Sent.	No Sent.	0.0120	0.9363	FALSE
96h Sent.	Swing Sent.	0.0040	0.9974	FALSE
96h Sent.	No Sent.	0.0200	0.7629	FALSE
Swing Sent.	No Sent.	0.0160	0.8629	FALSE

XGBoost (1 Game)		Mean Diff.	p-value	Reject Ho (alpha = 0.05)
24h Sent.	96h Sent.	-0.0211	0.7384	FALSE
24h Sent.	Swing Sent.	0.0128	0.9259	FALSE
24h Sent.	No Sent.	0.0009	1.0000	FALSE
96h Sent.	Swing Sent.	0.0339	0.3583	FALSE
96h Sent.	No Sent.	0.0220	0.7123	FALSE
Swing Sent.	No Sent.	-0.0119	0.9394	FALSE