Introduction

The NCAA Division I Men's Basketball Tournament, known informally as the March Madness Tournament, is a single elimination tournament featuring 68 college basketball teams and 67 games that are played mostly during March. The tournament bracket (see next page) is a grid of all the teams in the tournament and the round by round schedule of games with the path teams follow (if they win) to reach the national championship. One of the reasons for the tournament's popularity is trying the fill out the bracket accurately beforehand. Because there are 67 games, there are 2^{67} or 147.57 quintillion possible tournament outcomes.

Given the tournament's popularity, March Madness is widely discussed on social media, particularly Twitter. One of the goals of this project is to analyze the sentiment of March Madness related tweets, or a computational measure of whether a tweet is positive, negative, or neutral. In this project, the sentiment of a tweet was calculated using a recurrent neural network (RNN).

The purpose of this project is to analyze the relationship (if any) between whether a team wins, the average sentiment of tweets about the team, and the team's Basketball Power Index (BPI) score. BPI is a game prediction measure developed by ESPN, which accounts for opponent strength, pace of play, site, travel distance, altitude, and rest, and is calculated by simulating the game 10,000 times.

Sentiment Analysis

A recurrent neural network (RNN) was used to compute the sentiment of March Madness related tweets. The sentiment of a tweet is a computational measure of whether a tweet is positive, negative, or neutral. RNNs are a type of neural network that make use of sequential information, and have a "memory" which captures information about what has been calculated so far. For example, in calculating the sentiment of a particular word in a tweet, it makes sense to look at the context in which the word was used -- to take into account the sentiment of words that came before it. The RNN was trained on the sentiment140 dataset, which contains 1,600,000 annotated tweets extracted from the twitter api.

One important consideration was to use "Word Embeddings", or a natural language processing technique that aims to map semantic meaning into a geometric space. As explained in Keras documentation, "this is done by associating a numeric vector to every word in a dictionary, such that the distance between any two vectors would capture part of the semantic relationship between the two associated words. The geometric space formed by these vectors is called an embedding space." In a good embedding space, the path or vector to go between 2 words would capture the semantic relationship between them.

Dataset

The entire dataset for this project is on the next page. A description of each column in the dataset is below.

game_id: An integer key from 1 to 63, used as an identifier for each game played in the tournament. The full tournament includes 67 games, but the "first four" play in round of 4 game is omitted from this analysis. Additionally, games (10 in total) for which at least 15 tweets per team could not be obtained are excluded.

round: An integer from 1 to 6 corresponding to the game's round number.
Round 1: 64 teams -- Round of 64
Round 2: 32 teams -- Round of 32
Round 3: 16 teams -- Sweet 16
Round 4: 8 teams -- Elite 8
Round 5: 4 teams -- Final 4
Round 6: 2 teams -- National Championship

team: College basketball team name.

seed: The tournament is comprised of 4 regions. Each region has 16 teams, which are ranked from 1 to 16. The corresponding rank is the team's seed.

win: Boolean Value -- 0 (False, Lost) or 1 (True, Won) -- indicating whether the team won the corresponding game

win_margin: Team's margin of victory or defeat. Positive for wins, negative for losses.

bpi: ESPN's game prediction measure, which accounts for opponent strength, pace of play, site, travel distance, altitude, and rest, and is calculated by simulating the game 10,000 times. A value closer to 100 indicates a higher likelihood of winning the corresponding game.

win_bpi: Boolean Value -- 0 (False, Lose) or 1 (True, Win) -- indicating the a prediction of the team's outcome in the corresponding game based on higher BPI score

num_tweets: The number of tweets obtained for sentiment analysis. Tweets were obtained using a publicly available Twitter web scraper (https://github.com/Jefferson-Henrique/GetOldTweets-python). Search criteria involved the team name and game date.

avg_pos_conf: Measure of positive sentiment (scale of 0 to 1), averaged over all tweets obtained for a team on the game date.

win_pos_conf: Boolean Value -- 0 (False, Lose) or 1 (True, Win) -- indicating a prediction of the team's outcome in the corresponding game based on higher positive sentiment

χ^2 Test for Independence

- Multinomial Experiment: n independent identical trials (games) => 1 of 2 categories (win/loss)
- Investigate dependency between 2 classification criteria using contingency table

1) What is the relationship between winning and average positive sentiment?

	Win	**Lose**	Totals
Win -- Pos Sentiment	29	24	53
Lose -- Pos Sentiment	24	29	53
Totals	53	53	106

H_0: Winning and Average Positive Sentiment are independent
H_a: Winning and Average Positive Sentiment not are independent (dependent)

Test Statistic: $\varphi = \sum [(\text{observed} - \text{expected})^2 / \text{expected}] \sim \chi^2$ w/1 d.f. (for 2x2 table)
Expected = 53*53 / 106 = 26.5 => Observed Test Statistic = $\varphi_0 = 0.9434$

p-value = $P(\chi^2_1 \geq \varphi) = P(\chi^2_1 \geq 0.9434) = 0.3314$
For a significance level $\alpha = 0.05$, p-value > α. Therefore, do not reject H_0

=> Winning and Average Positive Sentiment are likely independent according to this dataset. In other words, knowing a team has a higher or lower average positive sentiment value does not affect the team's probability of winning.

2) What is the relationship between winning and BPI score?

	Win	**Lose**	Totals
Win -- BPI Score	40	13	53
Lose -- BPI Score	13	40	53
Totals	53	53	106

H_0: Winning and BPI score are independent
H_a: Winning and BPI score not are independent (dependent)

Test Statistic: $\varphi = \sum [(\text{observed} - \text{expected})^2 / \text{expected}] \sim \chi^2$ w/1 d.f. (for 2x2 table)
Expected = 53*53 / 106 = 26.5 => Observed Test Statistic = $\varphi_0 = 27.5094$

p-value = $P(\chi^2_1 \geq \varphi) = P(\chi^2_1 \geq 27.5094) = 1.5633 \times 10^{-7}$
For a significance level $\alpha = 0.05$, p-value < α. Therefore, reject H_0

=> Winning and BPI score are likely dependent according to this dataset. In other words, knowing a team has a higher or lower BPI score does affect the team's probability of winning.

χ^2 Test for Independence *(continued)*

3) In games where BPI failed to accurately predict the winner, what is the relationship between winning and average positive sentiment?

	Win	**Lose**	Totals
Win -- Pos Sentiment	11	2	13
Lose -- Pos Sentiment	2	11	13
Totals	13	13	26

H_0: In games where BPI was wrong, winning and average positive sentiment are independent
H_a: In games where BPI was wrong, winning and average positive sentiment are not independent

Test Statistic: $\varphi = \sum$ [(observed - expected)^2 / expected] ~ χ^2 w/1 d.f. (for 2x2 table)
Expected = 13*13 / 26 = 6.5 => Observed Test Statistic = φ_0 = 12.4615

p-value = $P(\chi^2_1 \geq \varphi) = P(\chi^2_1 \geq 12.4615) = 4.1542 \times 10^{-4}$
For a significance level $\alpha = 0.05$, p-value $< \alpha$. Therefore, reject H_0

=> In games where BPI failed to accurately predict the winner, winning and average positive sentiment are likely dependent according to this dataset. In other words, if BPI was wrong, knowing a team has a higher or lower average positive sentiment value does affect the team's probability of winning.

Assumptions of the Chi-Square Test

An important assumption of the chi-square test that is violated is that each subject or trial contributes data to only one cell. Here, if you look at each game as a trial, then each game contributes to two cells, once for the winning team and once for the losing team. For example, in questions 1 and 2, the total of cell frequencies in the contingency table is 106 but the actual number of "independent" trials is really 53. For these reasons, the results of the chi-square test may be invalid, and analysis for matched-pairs makes more sense (see next section). Regardless, the chi-square test is still included in this analysis because even though the assumptions are violated, reasonable conclusions are still achieved.

Nonparametric Analysis: Sign Test for Matched Pairs Experiment

- Nonparametric Statistics: work well under fairly general assumptions about probability distributions or parameters
- Matched Pairs Experiment: n pairs of random and independent observations in form (X_i, Y_i)

1) Is the distribution of average positive sentiment for winning teams the same as the distribution of average positive sentiment for losing teams?

- No assumptions made on probability distribution for average positive sentiment.
- Assume sentiment is independent with respect to each tournament game (reasonable)

Let X denote average positive sentiment for winning teams. $X \sim F_x = F(x)$
Let Y denote average positive sentiment for losing teams. $Y \sim F_Y = F(X - \theta)$
n = # of games = 53; X_i and Y_i denote sentiment values for teams in ith game.
=> actual form of F_x unknown (non parametric model)

H_0: $\theta = 0$; $X \overset{d}{=} Y$; $P(X > Y) = 0.5$ -- distribution of sentiment is same for winning/losing teams
H_a: $\theta < 0$; $X \overset{s}{>} Y$; $P(X > Y) > 0.5$ -- positive sentiment is stochastically greater for winning teams

Test Statistic: $M = \sum_{i=0}^{n} D_i \sim$ Binomial(n, p) where $D_i = 1(X_i > Y_i)$, n= 53, and $p = p_0 = 0.5$
In other words, M = # of positive differences where $Di = X_i - Y_i$
Observed Test Statistic = $M_0 = 29$

p-value = $P(\text{Binomial}(n, p) \geq M_0) = P(B(53, 0.5) \geq 29) = 0.2916$
p-value: If H_0 true, p-value is the probability of observing a more extreme test statistic
For a significance level $\alpha = 0.05$, p-value > α. Therefore, do not reject H_0

=> The distribution of average positive sentiment is likely the same for winning and losing teams. In other words, according to this dataset, no accurate determination can be made on the basis of sentiment alone on whether a team is likely to win or lose. This is likely due to noise in the tweets obtained for a particular game/team.

2) Is the distribution of BPI for winning teams the same as the distribution of BPI for losing teams?

- No assumptions made on probability distribution for BPI.
- Assume BPI is independent with respect to each tournament game (true assumption)

Let X denote BPI for winning teams. $X \sim F_x = F(x)$
Let Y denote BPI for losing teams. $Y \sim F_Y = F(X - \theta)$
n = # of games = 53; X_i and Y_i denote BPI values for teams in ith game.

H_0: $\theta = 0$; $X \overset{d}{=} Y$; $P(X > Y) = 0.5$ -- distribution of BPI is same for winning/losing teams
H_a: $\theta < 0$; $X \overset{s}{>} Y$; $P(X > Y) > 0.5$ -- BPI is stochastically greater for winning teams

Nonparametric Analysis: Sign Test for Matched Pairs Experiment *(continued)*

2) Is the distribution of BPI for winning teams the same as the distribution of BPI for losing teams? (continued)

Test Statistic: $M = \sum_{i=0}^{n} D_i \sim \text{Binomial}(n, p)$ where $D_i = 1(X_i > Y_i)$, n= 53, and $p = p_0 = 0.5$
Observed Test Statistic = $M_0 = 40$

p-value = $P(\text{Binomial}(n, p) \geq M_0) = P(B(53, 0.5) \geq 40) = 0.00013$
For a significance level $\alpha = 0.05$, p-value < α. Therefore, reject H_0

=> The distribution of BPI is likely different for winning and losing teams. Winning teams have a stochastically greater BPI compared to losing teams. In other words, according to this dataset, BPI is a reliable predictor of whether a team is likely to win or lose.

3) In games where BPI failed to accurately predict the winner, is the distribution of average positive sentiment for winning teams the same as the distribution of average positive sentiment for losing teams?

- No assumptions made on probability distribution for sentiment in games where BPI was wrong
- Assume sentiment is independent with respect to each tournament game (reasonable)

Let X denote average positive sentiment for winning teams (where BPI wrong). $X \sim F_x = F(x)$
Let Y denote average positive sentiment for losing teams (where BPI wrong). $Y \sim F_Y = F(X - \theta)$
n = # of games where BPI wrong = 13; X_i and Y_i denote sentiment values for teams in ith game.

H_0: $\theta = 0$; $X \overset{d}{=} Y$; $P(X > Y) = 0.5$ -- distribution of sentiment when BPI wrong same for win/lose
H_a: $\theta < 0$; $X \overset{s}{>} Y$; $P(X > Y) > 0.5$ -- positive sentiment is stochastically greater for winning teams

Test Statistic: $M = \sum_{i=0}^{n} D_i \sim \text{Binomial}(n, p)$ where $D_i = 1(X_i > Y_i)$, n= 53, and $p = p_0 = 0.5$
Observed Test Statistic = $M_0 = 11$

p-value = $P(\text{Binomial}(n, p) \geq M_0) = P(B(13, 0.5) \geq 11) = 0.01123$
For a significance level $\alpha = 0.05$, p-value < α. Therefore, reject H_0

=> The distribution of average positive sentiment in games where BPI failed to accurately predict the winner is likely different for winning and losing teams. Winning teams have stochastically greater average positive sentiment compared to losing teams. In other words, according to this dataset, when BPI is wrong, average positive sentiment is a better predictor whether a team is likely to win or lose.

Analysis and Conclusions

The purpose of this project was to analyze the relationship (if any) between whether a team wins, the average sentiment of tweets about the team, and the team's Basketball Power Index (BPI) score. Using nonparametric statistics, the sign test for matched pairs experiments indicates the following (according to this dataset).

1. No accurate determination can be made on the basis of sentiment alone on whether a team is likely to win or lose. This is likely due to noise in the tweets obtained for a particular game/ team and the reliability of the web scraper for pulling tweets. As an aside, the official twitter API was not used to obtain tweets because it did not support getting tweets in a specific date range.
2. Winning teams have a stochastically greater BPI compared to losing teams. This conclusion verifies that BPI is a reliable predictor of whether a team is likely to win or lose.
3. In games where BPI failed to accurately predict the winner, winning teams have a stochastically greater average positive sentiment compared to losing teams. This is likely due to the phenomenon of an "upset." If BPI is wrong, then the winning team was not the favorite, or was not expected to win. Thus, more reaction can be expected on social media, possibly explaining the higher positive sentiment.