NBA Analytics: Predicting 2023 Season Conference Leaders and Analyzing Twitter and Reddit Text to Observe for Correlations between Team Performances and Online Social Activity.

Corbin Cahalan, Levina (Minxin) Dong, Thomas James
Master of Applied Data Science, University of Michigan School of Information SIADS 696: Milestone II
October 24, 2022

Content warning: This report includes discussion about a dataset that documents topics such as sentiment analysis, sarcasm levels, and potential moods NBA players possibly could be having before and/or after games. In no way is this data guaranteed to provide definite proof of the behaviors of NBA players.The data from this project should not be used to judge NBA player activity.

Table of Contents

Introduction

The National Basketball Association (NBA) is one of the most famous national sport leagues in the world. Millions of Americans are big fans of the NBA, and many news sites, articles, and discussion boards are devoted to NBA-related topics. In this project, we use supervised learning to make predictions for the 2023 NBA regular season games. Unsupervised learning was used to study the types of influences social media could potentially have on NBA players, and explore the moods and word-correlations of the most common NBA topics discussed amongst Twitter and Reddit users. Depending on the effectiveness of our models, a sports analyst, sports bettor, or general NBA fan would appreciate our project.

Part A. Supervised Learning

Motivation

Predicting which teams will make it to the NBA playoffs is valuable to sports bettors and general NBA fans. Our main question was, “Which teams will make it to the Finals in the 2022-2023 NBA season?” Using the last seven years of regular season data, we attempted to predict the 8 teams from both Western and Eastern conferences based on the number of games each team will win.

Data Sources

Our two main data sources are the balldontlie API (Park, 2019) and Basketball Reference (2022). We were able to extract all features we used for our models using the balldontlie API, and we collected the games schedule for the upcoming NBA season using Basketball Reference. In addition to the details below, you can view how we conducted data extraction and manipulation in the nba_supervised_learning.ipynb file.

Balldontlie API

We used balldontlie to extract data from three of their APIs: Games, Stats, and Teams. All three of the APIs return data in JSON format, which we used pandas.json_normalize() to convert into a flattened datatable. For all balldontlie APIs, we collected data for the last seven regular seasons (i.e. no playoff games). The Games API mainly collects the home and away team scores per historical game, which enabled us to create a boolean column whether the home team won. Using games from the last 7 seasons, we collected a total of 8,384 games. The important features were game ID, season, home team ID, home team score, visitor team ID, and visitor team score. The Stats API collects game statistics at the player level, which enabled us to aggregate statistics at the team level. We collected a total of 214,343 results over the last seven seasons. The important features collected from the Stats API were game ID, season, team id, 3-point field goal attempts and makes, field goal attempts and makes, free throw attempts and makes, steals, blocks, defensive rebounds, and assists. The Teams API allowed us to collect basic information on each team, such as the team name along with an arbitrary team ID, which was useful as a primary key to connect with the 2022-2023 games schedule on Basketball Reference. We collected 30 results from the Teams API, which is the total number of teams in the NBA.

Basketball Reference

We used Basketball Reference to collect the game information for the 2022-2023 season. This was a simple task of downloading the game data in CSV format from the 2022-23 NBA Schedule and Results page (Basketball Reference, 2022). In total, we collected all 1,230 games that are planned to be played this season. The important columns were the visitor team name and the home team name per game. Basketball Reference uses the same naming convention as the balldontlie API, except for the LA Clippers, so it was pretty easy to use both of these sources together.

Methods & Evaluation

To predict the winning teams for the 2022-2023 regular season, we had to make predictions for every game this upcoming season whether the winning team or losing team would win each game. Because we are looking at a dependent variable that can only have two values (win, loss), we tested out three different supervised learning classification algorithms: logistic regression, random forest classifier, and XGBoost classifier. Logistic regressions are powerful and common when predicting binomial output. Additionally, using

decision tree-based models such as a random forest classifier can reduce overfitting. XGBoost is known to be a strong model for both classification and regression, so we wanted to make sure to test it out as well.

Feature Selection

In any game of basketball, there are a lot of important variables that can impact a team's chance of winning. Our team ultimately decided that it would be important to use both offensive and defensive statistics as inputs for our models. To train the models, we used the raw game-level statistics per historical game. As inputs to make 2023 game predictions, we used the 7-year average of those same statistics when one team is up against another team. For example, we used the 7-year average number of steals for the LA Lakers as the home team when they played against the Golden State Warriors as the away team to predict the 2023 game when the LA Lakers are the home team and the Golden State Warriors are the away team.

In total, we used 17 features to train our models and make game predictions. For defense features, we used defensive rebounds, steals, and blocks (one per away and home team). For offensive features, we used 3-point percentage, field goal percentage, and free throw percentage (one per away and home team). Lastly, we also used the home team ID, away team ID, and season year. Below are the input features to make predictions for each 2023 game:

Variable	Description
home_team_id / away_team_id	Arbitrary ID to distinguish each team using integers
season	The integer value for the season year
home_fg3pct / away_fg3pct	The 3-point percentage per team
home_fgpct / away_fgpct	The field goal percentage per team
home_ftpct / away_ftpct	The free throw percentage per team
home_ast_avg / away_ast_avg	The average assists per game per team
home_blk_avg / away_blk_avg	The average blocks per game per team
home_dreb_avg, away_dreb_avg	The average defensive rebounds per game per team
home_stl_avg, away_stl_avg	The average steals per game per team

Table A-1: Features used for 2023 game prediction

Before inputting our data into our models, we used RobustScaler() to normalize our data. RobustScaler() allows us to remove potential outliers in our data using the interquartile range and balance the scale per feature. For example, one home team made 24 blocks in one game when the IQR represents 3-7 blocks for the typical home team in the last 7 years. Normalizing our data across metrics reduces the importance of those outlier game statistics and potentially leads to a more effective prediction model.

Model Training & Tuning

After manipulating player-level data per game to be team-level, we had all of our features prepared for our models. Only ~2.7% of our full dataset had to be dropped due to missing values. In terms of training data size vs. accuracy, we had originally tested a few different training and testing dataset splits. With more than 8,000 games, we had ample data to create a sufficient model. We ultimately found that it was best to split up our dataset into 75% training data and 25% test data for model evaluation instead of lower percentages such as 15% or 20%. This also helps reduce overfitting, in which some of the models we used are prone to overfit data. For each of the three algorithms (LogisticRegression, RandomForestClassifier, XGBClassifier), we used RandomizedSearchCV to determine the optimal parameters to use. Because we had many parameters to use across our three models, RandomizedSearchCV allowed us to iterate quickly and determine reasonable parameter levels. Please view the nba_supervised_learning.ipynb file for more details on the parameters we selected.

Evaluation

As a baseline, we started out by saying that our model should be at least better than random chance (i.e. 50% accuracy). However, an even better baseline is whether the home team wins a game. In the last 7 years, if someone chose the home team to win every game, they'd be correct ~57% of the time, so we ultimately decided that our model should be better than selecting the home team to win every game. After normalizing our data and tuning our parameters, we found that our LogisticRegression model performed the best. All of our models achieved accuracy scores around 80-85%. Because the LogisticRegression model performed the best, we moved forward with using it for evaluation, analysis, and prediction.

Our winning model: LogisticRegression(solver='liblinear', penalty='l1', C=1, random_state=0)

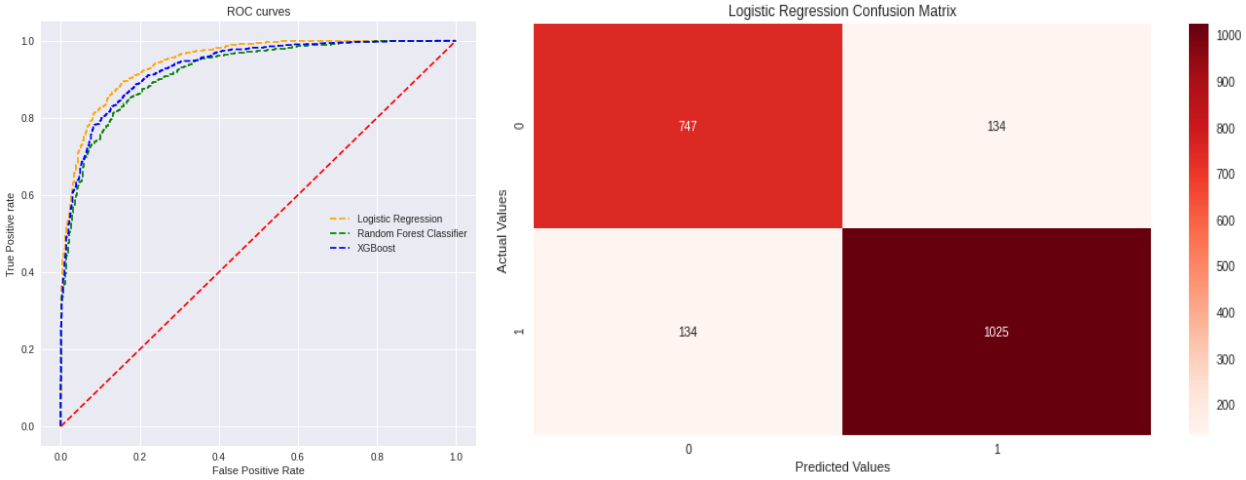


Figure A-1: ROC curves per model and confusion matrix for our logistic regression model

To evaluate our models, we used accuracy scores, ROC curves and their respective AUC scores, and a confusion matrix. Our winning logistic regression model achieved an AUC score of 94.5% and an accuracy score of 86.9%. These are very high scores, which put us on alert for potential overfitting. Very few games were inaccurately predicted with only 134 false positives and 134 false negatives in the test data. Putting that aside, we moved forward with making predictions for the 2022-2023 NBA season.

Results

Using the 7-year average data for each team when playing against another team, we were able to make predictions per game for the 2022-2023 NBA season. Here are the top 8 teams per conference that we predict will head to the NBA finals, along with their total number of wins:

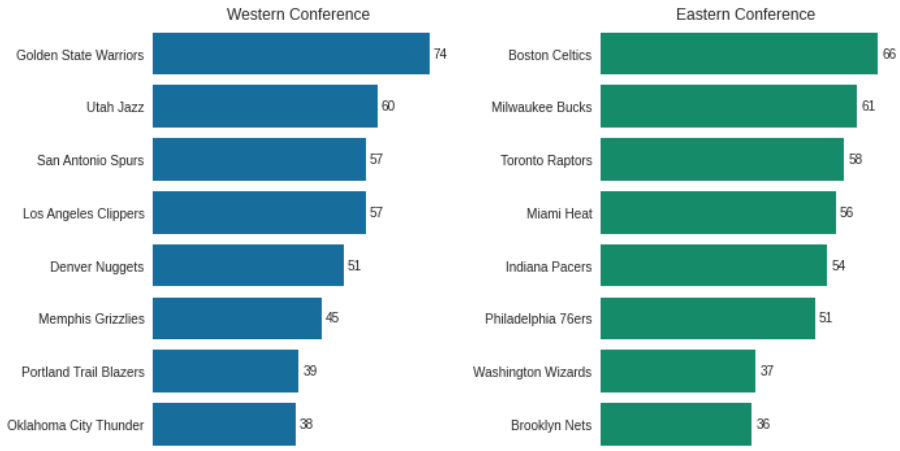


Figure A-2: Our 2022-2023 NBA Regular Season Top 8 Team Per Conference

Failure Analysis

Looking at the first few games that have been played in the 2022-2023 season, our model has predicted a handful of game winners correctly, but definitely not at the accuracy that we had with the last seven seasons of regular season data! Where could we have gone wrong, and what is missing?

First, the 7-year average stats we used to predict 2022-2023 regular season game wins is not an apples-to-apples metric with raw metrics we used to train our models. While we were able to train our models using the real statistics that occurred in each game, we had to use the 7-year averages for one team against another as input to predict the results, since we don't know the stats before a game occurs. This assumes that teams will only perform the average of how they've performed in the last seven years. Teams change so much each season that using the 7-year average might be too wide of a window. For example, the Utah Jazz and San Antonio Spurs are very unlikely to make the top 8 teams in the Western Conference due to trades made in the offseason, although their teams have performed relatively well in the last few years. In fact, San Antonio's head coach recently shared that nobody should expect the Spurs making the playoffs this year (Karlis, 2022). Some teams will definitely outperform their average performance in the last seven years, and some will underperform. Ideally, we would want the training metric to be the same as the prediction model input, so we should use a rolling 7-year average per year of game data in the training data to align with the 7-year average input we used to predict 2022-2023 season games. For example, games from 2018 would use the average from the previous seven years similar to how our 2022-2023 data used an average of the last seven years. Alternatively, we could also use metrics with different windows, such as performance during the last 10 games to predict how the next game will go down.

Overall, our model is probably overfitting because it is too simple. We would need to consider a different combination and number of features that accurately predicts team wins. For example, our model predicted that the Golden State Warriors will win 74 out of 82 regular season games. The all-time best team season score was the Warriors in 2016 with 73 game wins (Champs or Chumps, 2022). While it isn't unimaginable that the Warriors can achieve 74 wins, it has never happened before. To improve our model, we might want to look at more advanced statistics such as players' defensive and offensive scores and true shooting percentages. A time component might also be useful to consider, such as "first half of season" vs. "second half of season" because teams shift their strategies throughout a season.

Unfortunately, our data also doesn't consider unpredictable issues such as injuries and team drama, both of which play a huge factor in a team's success. In order to improve our model, we would likely want to make predictions based on player level data that predicts whether a team wins a game or not. For example, if Zion Williamson gets injured, our model should remove his likely contribution to the Pelicans winning any game. Starting our model with player-level predictions per game could potentially lead to better team-level statistics when putting two teams up against each other. To factor in team drama, we'd have to create an entirely separate model that possibly uses NLP for sentiment analysis of online articles and social media.

Part B. Unsupervised Learning: NBA Text Analysis through Reddit

Introduction

For unsupervised learning, we worked on text analysis using natural language processing with a focus on social media data. In 2021, the most popular Google search in the United States was "NBA" (National Basketball Association), outranking other top searches such as "Squid Game," "Mega Millions," and "stimulus check" (Kennedy, 2022). Social community platforms such as YouTube, Twitter, and Reddit showcase various levels of NBA awareness among the public. We wanted to explore how NBA audiences and fans react to NBA games and to identify popular topics they discuss during the season. Our team decided to use data from Reddit and Twitter to perform the unsupervised learning. We used Reddit data to discover general topics that interest NBA fans during a season, and we used Twitter data to analyze people's reactions on each NBA team's Twitter account. Importantly, Reddit data are verified news, which implies they reflect the most influential new topics during the season. Twitter data comes from users' microblogging, meaning tweets are segregated based on trends and may contain both true and false information. This is especially true when people use tweets to attract attention, so Twitter data contains more fluctuations than Reddit data. In other words, the polarity gap in Twitter data is larger than that of Reddit data. Furthermore, controversial information on Reddit and Twitter, much of which may be false, travels faster than noncontroversial information (Dizikes, 2018; Jasser, 2021).

Motivation

Athlete branding is an important tool in measuring a team's performance on social media (Palka, 2022). Strong social media interaction can provide value to the team by increasing the team's opportunities and financial income. To understand NBA fans' interests during a game period, we used Reddit data to examine NBA followers' behaviors of social interaction. Our team used Reddit data to perform sentiment analysis to measure polarity within the Reddit community and used topic modeling to predict the top news that NBA audiences and fans are most interested in during game periods (such as playoff seasons). This information provides us with social media trends and audience preferences during NBA game periods. Such details could help NBA teams capture the audience's attention and increase awareness of NBA games.

Data Source

In contrast to Twitter data, Reddit captures aggregated information. Unlike Twitter's 280-character limit, Reddit data come in the form of long discussion posts, articles related to the subject posts, and user questions. Reddit also creates smaller communities called "subreddits." A subreddit brings together people with similar interests to discuss popular content from the web. Importantly, Reddit is used more for discussion than connection and attention like Twitter. This makes Reddit a good source for topic modeling data.

To extract data from Reddit, we chose to use Pushshift Reddit API. This API can build robust data aggregations and provide extensive capabilities for searching Reddit data. There are two main endpoints in which this API can search through public information: comment and submission. Within these two options, submissions are public posts on discussion topics of interest to users. Because our analysis focused on NBA audiences' points of view, we decided to utilize submissions for our data collection. Next, we decided to reduce the data to the NBA subreddit. This subreddit contains 5.4 million members and generates most of the NBA data. This smaller community can provide us with more precise information on the NBA and ensure all submissions are related to NBA news and discussion.

We used six important parameters to limit the quantity of Reddit data prior to data extraction:

Parameter	Description	Accepted Values
q	The *q* field identifies the query being examined in the submissions. For example, when we searched for submissions related to playoff seasons only, our q was "playoff."	String/quoted string for phrases
size	The *size* field determines the number of submissions that can attract each link.	Integer </= 500
sort_type	The *sort_type* field arranges the submissions; we used "score" as our *sort_type* because "score" is measured by users' votes on the posts. Depending on how significant the news is, users cast votes that are then separated and viewed (Yadav, 2022). The higher the score, the more important the submission.	"score," "num_comments," or "created_utc"
subreddit	The *subreddit* field is used to select the target community; we used the subreddit "NBA."	String or comma-delimited string (multiple values allowed)
after	The *after* and *before* fields are used to select the timeframe of the chosen posts. We limited our timeframe to the 2022 NBA playoff season (April 16–May 29,2022).	Epoch value or integer + "s," "m," "h," or "d" (i.e., "30d" for 30 days)
before		

After assigning a parameter to Pushshift, the API returns a JSON format. Because the volume of online social media data is large, our team decided to export the data into CSV format for easy modification.

Data	Type	Size	Date Range	Observations
NBA Playoff Reddit Text Data	CSV	11.4MB	March 31, 2022 - June 29, 2022	40,185

Unsupervised Learning Methods

Sentiment Analysis

We used Reddit data to run a sentiment analysis on each post and assess their word distributions according to the following types: positive, negative, and neutral. We then utilized topic modeling to depict the most influential topics during the 2022 seasons.

For the sentiment analysis, we used the NLTK package SentimentIntensityAnalyzer (SIA) polarity score to measure the score for each submission. This model is easy to use because it does not require existing training data. The SIA assigned four scores to each post: neg, neu, pos, and compound. After creating the scoring record, we generated a label for each post by using the compound score, which was the normalized sum of the post's positive, negative, and neutral scores. When the score is greater than zero, the post is determined to be positive. When the score is equal to zero, the post is neutral, and when the score is below zero, the post is negative.

```
Positive headlines:

['highlight embiid erostep poster', 'highlight crazy seqence leads  holiday  giannis', 'best player
never  reglar season', 'highlight saddiq  knocks triple pistons  forcing sixers take timeot', 'schmi
tz imagine pistons team chet holmgren paolo banchero jabari smith ftre bright cade cnningham clearly
sperstar making detroit brass also  several singles loved isaiah livers pick wold gotten  rond looks
healthy']

Negative headlines:

['possible  fire  coach right before  playoffs', 'post game thread  detroit pistons defeat philadelp
hia', 'cade cnningham steals shooting otplays embiid harden victory', 'qestion  pistons  granted pos
session overtrned goaltend rled block', 'philadelphia  bench combined points tonight loss detroit pi
stons']
```

Figure B-1: The Top Five Posts for Positive and Negative Submission

In Figure B1, most of the top posts, those with strong scores (high positive and negative scores), described players' abilities and a game. Our findings suggest NBA fans are positive in their discussions of players' ability to score and their performances during a season. For example, Reddit users used "best player" to highlight players who perform well. On the opposite sides, on the negative posts, users seem to be attracted to news related to the team, such as team overall performance. The Reddit submission did not contain much negative information on the losing team players. Reddit users appeared unemotional.

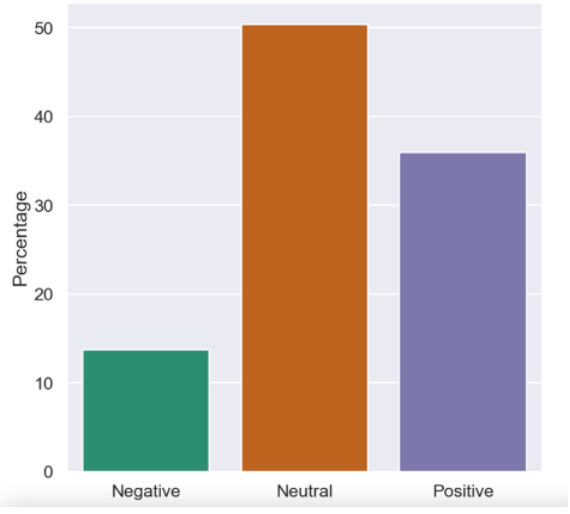


Figure B-2: % of Polarity Scores

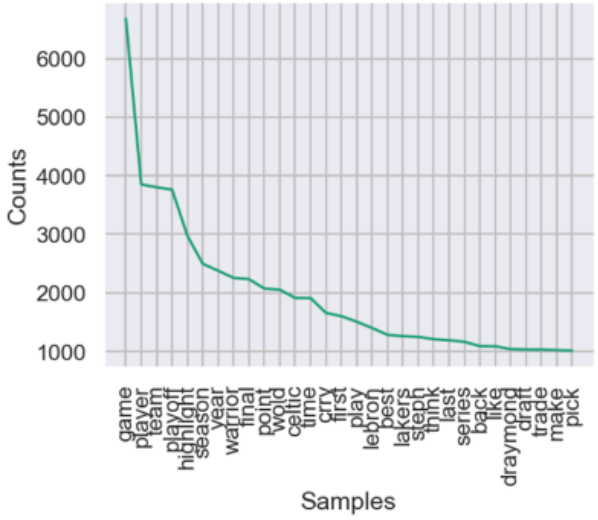


Figure B-3: Top 30 Words From the Frequency Distribution

Furthermore, the distribution of the submission polarity score was expected in Figure B2. Almost 50.39% of the Reddit submissions belonged to the neutral category, whereas approximately 35.93% were positive, and 13.67% were negative. Reddit submissions contained primarily discussions and did not clearly show users' emotions, so these results are not surprising.

Word Distribution

Before topic modeling, our team explored word distribution to discover the most frequent word appearing in the overall discussion. We used tokenization and stemming to remove duplicate words and words with similar meanings to make the frequency distribution more accurate.

As shown in Figure B3, the top 10 words are “game,” “player,” “team,” “playoff,” “highlight,” “season,” “year,” “warrior,” “final,” and “point.” Most of the words are common words that appeared in the NBA news and discussions. Both Golden State Warriors and Boston Celtics appeared in the top 30 words, which is expected because these teams played in the 2022 season finals.

Figure B-4: Word Clouds for Positive and Negative Posts

In the word distribution for each group (positive and negative) in Figure B4, most words were from posts submitted during the last game of the season. For example, in the playoff data, the majority of words in the cloud came from discussions about the final game. The words “warrior,” “championship,” “good,” and “free throw” were mentioned, presumably because the Warriors won the playoffs. In addition, “LeBron James” also showed up in the word cloud, likely because of discussion of the news that James missed half of the playoff season (Polacek, 2022). Moreover, because the words “LeBron” and “Laker” appeared in the negative post, we assumed there were many people discussing the news that, without James, the Lakers were eliminated from the 2022 playoffs. This phenomenon also implies that famous players are a target for news when something bad happens.

Topic Modeling

Based on the chaos in the Reddit data, we used Gensim library simple_preprocess() to clean the data again for topic modeling. This tool can help us eliminate words that are too long and too short. In addition, we added a lemmatization with part-of-speech tag to only include words that are nouns and verbs. By excluding other parts of speech such as pronouns, adverbs, and adjectives, we can focus on important events, names, and actions.

For the topic modeling process, we first chose to use CountVectorizer to build the matrix based on term frequency because we were not analyzing the document, so we did not choose TfidfVectorizer, which builds a matrix based on term frequency on a collection of raw documents. For the parameter setting, we added min_df = 10 to set the minimum requirement for the occurrence of a word and token_pattern='[a-zA-Z0-9]{2,}' to keep the character above two characters.

For the basic Latent Dirichlet Allocation (LDA) model from the Sklearn package, we set the following parameters:n_components = 20, we first tested with 20 topics; 1. max_iter = 10, the maximum learning iteration; learning_method = ‘online,’ we chose to use “online” rather than “batch” because an online update will be much faster than a batch and can be controlled by learning_decay in later hyperparameter tuning; random_state = 100, a status can replicate the results;batch_size = 128, default setting for learning_method; evaluate_every = -1, compute perplexity every n iteration; n_jobs = -1, use all available CPU to increase speed.

This basic model returned a log likelihood of -1105105.55 and a perplexity score of 891.55. So, for the hyperparameter tuning, we decided to use the GridSearchCV to determine the best LDA model.

We chose to use the search_params = {'n_components': [10, 15, 20, 25, 30], 'learning_decay': [.5, .7, .9]} to be the test components.

```
Best Model's Params:  {'learning_decay': 0.7, 'n_components': 10}
Best Log Likelihood Score:  -230030.71409809758
Model Perplexity:  758.3829468788875
```

Figure B-5: Best Model Parameter for LDA

As shown in Figure B5, after cross-validation in GridSearchCV, the best model contains 10 topics with a learning decay rate of 0.7. Moreover, the model perplexity is approximately 758.38, which is below the basic model. A low perplexity value suggests the sample may be accurately predicted by the probability distribution. Thus, on the final LDA model, we decided to use n_components = 10 and learning_decay = 0.7.

Unsupervised Evaluation

```
Topic 0: playoff series win net state look miss knick simmon pass
Topic 1: year think player history career know play star sixer thing
Topic 2: make playoff come lebron lead feel klay hit performance qestion
Topic 3: team player trade leage year rond wold today throw offer
Topic 4: say sorce tell start wiggin sign watch contract shot ask
Topic 5: warrior laker steph want championship fan time need talk conference
Topic 6: season point game coach head assist rebond piston average stat
Topic 7: game final play time basketball thread bck grizzly beat clipper
Topic 8: celtic draft pick highlight kyrie maverick jazz ball net try
Topic 9: heat shoot leave point people deal tonight score westbrook title
```

Figure B-6: Top 10 Topics for the LDA Model

One possible interpretation is that Topic 0 refers to Brooklyn Nets player Ben Simmons who failed to debut because of injury. Topic 1 possibly refers to the Philadelphia 76ers, known as the Sixers, who made it to the Eastern Conference semifinals. Topic 2 discusses the Lakers star James and Warriors sharpshooter Klay Thompson. During Game 1, Thompson overtook James to have the second-most three-pointers made in NBA postseason history (Evans, 2022). Topic 3 may indicate a league player offer. Topic 4 mentioned the contract and score, so it can imply Andrew Wiggins from the Golden State Warriors agreed to a 4-year contract extension (NBA.com, n.d.). Topic 5 shows the championship team Warriors and their MVP Stephen Curry in the 2022 finale. Topic 6 may refer to the Detroit Pistons rebound statistics. Topic 7 covers the Memphis Grizzlies and Los Angeles Clippers games, and how the Grizzlies beat the Clippers. Topic 8 involves the rumor that Kyrie Irvine was traded to the Dallas Mavericks. Topic 9 may discuss that the Lakers are frantically trying to get rid of Russell Westbrook (Fox Sports, 2022).

```
Topic 0: game thread celtic series heat tonight discssion defeat post bck
Topic 1: team championship time want think win lead wold trade draft
Topic 2: player time think leage history draft wold basketball role rank
Topic 3: playoff make time series rond miss history eliminate career laker
Topic 4: season laker record westbrook win time post start miss history
Topic 5: warrior final celtic state time steph conference win beat wiggin
Topic 6: play say want basketball tonight mint think simmon time net
Topic 7: year final today draft contract championship extension time sign deal
Topic 8: point rebond assist shoot score tonight career time steal lead
Topic 9: make sorce tell say pick trade laker coach draft leage
```

Figure B-7: Top 10 Topics for Non-matrix Factorization Model

We also performed a Non-negative Matrix Factorization (NMF) model using the top 10 topics. In comparison to the LDA model, the topic discussion is similar to the LDA model but with a different ranking. In Figure B7, Topic 0 discusses the Boston Celtics game thread on each game because the Celtics made it to the finals. Topic 1 shows that team championships can result in teams getting better draft trades. Topic 2 shows a league ranking history for each player. Topic 3 refers to the Lakers in the playoffs and their elimination before this year's playoff, so this may refer to the absence of James from the Lakers. Topic 4 is the same as Topic 9 in the LDA model, which refers to Russel Westbrook. Topic 5 is almost the same as Topic 5 in the LDA model, depicting the final games in 2022, which involved the Warriors and Celtics. Topic 6 is like Topic 1 in the LDA, discussing how Simmons failed to play in Game 4. Topic 7 discusses contract extensions but does not specify the player and team. It may refer to Wiggins's contract extension like Topic 4 in the LDA model. Topic 8 shows the keyword skills statistics such as rebound, assist, shoot, and steal. Topic 8 may refer to NBA player box scores. Topic 9 discusses the Lakers trades and coach, so topic 9 may refer to Westbrook again.

NMF models are more difficult to interpret than LDA models, and NMF models store words that are approximately similar in meaning, but it makes it harder for human interpretation for exact events. However, the LDA stores events that happen closely related to each other and includes team and player names. Especially when data is new, we can make better guess-topics using the LDA model.

Challenges

Data cleaning was the most challenging part of analyzing the Reddit data. We began by cleaning all the common errors such as spaces, NaN, and symbols, which is especially important because users have personal preferences for syntax and characters when typing. There is no unique identification for certain typos, so we had to clean text manually, sentence by sentence. In addition, NBA posts often involve many players' names and team names, and some users use nicknames. This meant we were obligated to research which players or team names the nicknames were referring to so we could verify the correct spelling. In addition, users often made spelling mistakes, such as writing "Stephen Curry" as "Stephen Crry." To address misspellings, we added bigrams and trigrams to clean and remove spelling errors.

The second challenge was the large data size, so we had to figure out how to reduce the data size by shrinking the tokenization lists by including only important words such as nouns and verbs. In addition, we also used models such as spaCy tool's en_core_web_sm, which is an English pipeline that helps optimize the CPU to run the lemmatization.

Failure Analysis

When we started analyzing this report, we chose several API to start with the testing code, and only a few worked. The first API we chose was PRAW. This is a Reddit API Wrapper with a built-in function to extract subreddit posts. However, this API has an extraction limit of only 1,000 posts on each authorized server, so we later switched to Pushshift API.

In topic modeling, we tried several methods to calculate the coherence score to test the hyperparameter to test the optimal number of topics. We failed to use the Gensim library LDA method. Unlike the GridSearchCV, the Gensim library contains a tool CoherenceModel that can automatically calculate the coherence value. However, we could give the model a large value that could return several smaller topics, but the iteration times took too long, and the CPU ran out of memory.

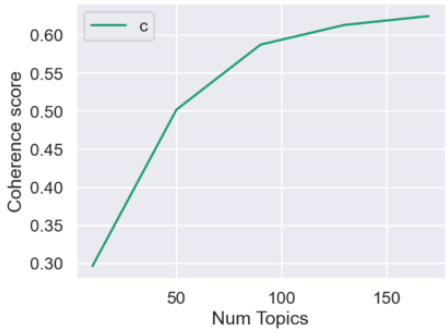


Figure B-8: Coherence Score for Different Numbers of Topics

In Figure B8, we set a limitation of 200 topics, and, with a step of 40, the five models that we built had an increasing climbing coherence score, where the higher the coherence score, the better the topics. However, we expected to see a decrease in coherence score when the number of topics climbed too high. These methods took a longer process time and made the topics difficult to interpret. Eventually, we switched to the GridSearchCV to manually update the tuning.

Part C. Unsupervised Learning: NBA Text Analysis through Twitter

Motivation

Twitter is one of the most widely used methods of social-media communication that currently exists. Over 200 million users actively use Twitter daily, including celebrities and professional athletes. Emotions and motivations derived from peer and societal pressures can potentially have an impact on the performance of professional athletes. In this text-analysis, extracted Twitter text data was tested using sentiment-analysis and sarcasm-detection to observe potentially different mood levels present amongst the different NBA teams. A Word2Vec Model, Latent Semantic Indexing (LSI), Cosine Similarity, LDA, and NMF were utilized to observe natural language features present in the Twitter text data. These natural language features

permitted better insight into the moods present on Twitter during the past year, and can provide insight into what types of influences social media has on professional basketball players.

Data Sources

Data Source 1: Twitter API

Twitter API

- **Original source:** Keys provided by a Twitter API allowed extraction of text data directly from Twitter to the Jupyter lab notebooks. https://developer.twitter.com/en/docs/twitter-api
- **Format:** Dataframes that could be converted to Excel spreadsheets using Python
- **Count of records retrieved:** See Table 1
- **Time period covered:** 08-06-2021 to 09-12-2022
- **Important variables:** The natural language features in the dataset were the fifteen teams from the NBA Western Conference and the fifteen teams from the NBA Eastern Conference. These teams, and the amount of Twitter data extracted for each team, can be observed in Table C-1 (see Appendix Table C-1).

Unsupervised Learning Methods and Features

After cleaning the data for each of the NBA Western Conference teams and Eastern Conference teams, the feature selected for the sentiment analysis was the Twitter text data for each of the 30 individual teams. The features selected for the word-analyses were the Twitter text data from the Golden State Warriors and the Boston Celtics. The features selected for the sarcasm-detection analysis were the Twitter text data and the Twitter source data from the Golden State Warriors and the Boston Celtics. The feature variables focused-on the most in this research were the Golden State Warriors and the Boston Celtics.

Multiple types of models were selected to help provide more data for analytical comparisons. Sentiment-analysis was conducted using both TextBlob and sentimentVader (Vader) models. Vader is already well-known as a more efficient method when conducting sentiment-analysis. Vader was compared to TextBlob in this experiment to allow better potential observation of the increased effectiveness Vader has when compared to TextBlob. The two teams that made it to the previous year's NBA finals, the Golden State Warriors and Boston Celtics, had the highest number of tweets amongst all the teams and were selected for further analysis. Sarcasm detection was initially selected as a potentially interesting measurement when compared with sentiment-analysis. Due to project time-constraints, sarcasm-detection was not able to be performed for all 30 NBA teams, and was instead used as another type of individual text-data observation. A sarcasm-detection was conducted for the Warriors and Celtics starting with splitting the data and converting the training and test corpuses to TFIDF feature vectors. These feature vectors were then placed in Logistic Regression (LR) and Random Forest Classifier (RFC) models after GridSearchCV was utilized to find the optimal parameter measurements for each model. A dummy-classifier was developed to allow a comparison-model and better accuracy while creating the predictions of sarcasm levels Twitter users may have while discussing these teams in the future. Classifier-pipelines were used to run these machine-learning models and prevent data leakage. LimeTextExplainer were then utilized to develop and chart predictive probabilities from the levels of sarcasm contained in the Twitter text data.

Machine-learning models were only utilized for the sarcasm-detection portion of this experiment, for they were not needed in the sentiment-analysis and semantic-analysis portions. Before conducting further word-analysis, NLTK and RegEx tokenization methods were tested to determine which would be most effective for the NBA Twitter data. NLTK provided more unique-tokens compared to RegEx, and was used for the rest of this project's analyses. The Word2Vec model was applied to some of the more commonly used words and word-groups discovered during the Warriors' sentiment-analysis to observe other word-usage on Twitter that has high levels of similarity with these terms. For further analysis, Bag-of-Words model was accessed to convert the NBA Twitter text data into numeric document-term matrices containing shapes of (4714, 8195) for the Warriors and (6045, 4973) for the Celtics. Corpus term information and tf-idf weights were then applied to these matrices to prepare the data for LSI. LSI assisted with the indexing of the matrix data in preparation for Cosine Similarity testing and LDA testing. Cosine Similarity was used to produce LSI scoring amongst the data. The LSI scoring was then applied to the term-matrices to develop a list of the top twenty Twitter tweet LSI-related terms. LDA was next applied to help answer the question "What is the probability of observing a given multinomial distribution over 'k' categories?" A list of Warriors and Celtics

topic-lists were developed, and the document-weights per topic were calculated to observe the top ten most popular topics discussed through Warriors and Celtics Twitter text. This semantic analysis allowed observation of syntactic structure relationships existing amongst the words contained in the Twitter data discussing the Golden State Warriors and the Boston Celtics. To better visualize this data, a word-cloud was created to view the most common terms and phrases used on Twitter discussions.

A pyLDAvis interactive panel was also developed to permit interactive observation of the term-frequencies and the marginal topic distributions observed in the Warriors and Celtics related Twitter data. To provide comparison for the LDA analysis, a NMF was created using the same text data used in the LDA analysis. NMF typically is useful when working with numerical data and having knowledge that some of the numerical data contains zeros or unknown values. NMF decomposes the numerical data into two smaller matrices: the document-topic matrix and the topic-term matrix. NMF was considered as an alternative method to use for comparison alongside LDA after the text data had already been converted to numerical form. The top topics were again observed and visualized using a bar-chart and a pyLDAvis interactive chart. Utilization of both LDA and NMF providing insight into the strengths of each model while analyzing text data.

Challenges

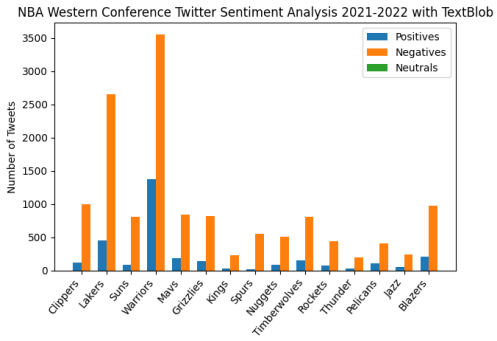
When attempting to extract the NBA correlated data from Twitter, a list of key-words had to be created that the Twitter server could use to search through its tweet-library. We put some effort into thinking and checking to see what names and what terms could be interpreted as relating to other topics outside of basketball; but really a whole month, and multiple scientists, could have been devoted to this task to provide a longer list of more accurate keywords to be used in this type of experiment. Tweet extractions from Twitter also required 4-6 hours of time due to the limits provided by Twitter per extraction. Some of the libraries initially imported did not work with the most up-to-date desktop version of Jupyter Notebook, and would only run through the Coursera Jupyter notebooks. There were smaller quantities of tweets for some teams than preferred, and more time available to develop more key-words to help increase the tweet-count for all the teams having lower-tweet levels could have provided better accuracy for the entire experiment. A stronger computer processor may have helped with permitting testing of more model-parameters while using GridSearchCV to determine the optimal parameter measurements for each machine-learning model used during sarcasm-detection. Coding used for printing LSI-related terms after performing LSI, and the coding used for developing pyLDAvis visualizations following application of NMF, worked during the first few weeks of the project, but in following weeks stopped working and had to be excluded. Time limits did not permit detailed correlations between sentiment analysis of text data and specific performances of NBA players to be fully studied, but can possibly be studied in future research.

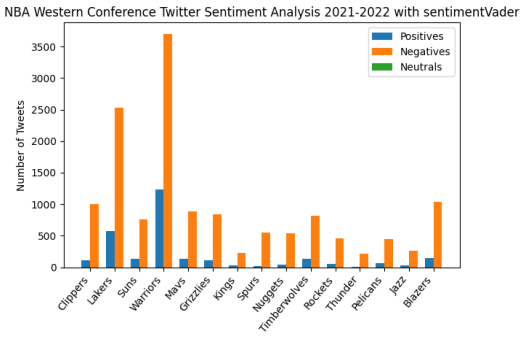
Unsupervised Evaluation

Sentiment analysis conducted for the Eastern and Western conference NBA teams developed similar patterns when comparing the utilization of TextBlob to sentimentVader. The difference between the two models had minor differences in this experiment. The sentimentVader model is already known to be more precise and efficient than TextBlob. The lack of contrast in this experiment may have been due to lack of data (~200 tweets) for some of the teams. As can be seen in Figures C-1 to C-4, negative discussion about the teams highly outweighed the positive discussion. The top four teams people talked about on Twitter during the past year were the Warriors, Lakers, Celtics, and Cavs.

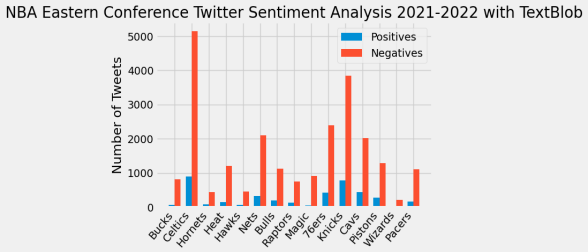
Interpretations derived from Figures C-1 to C-4:

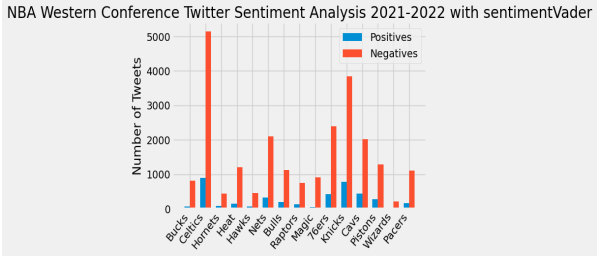
- The Warriors and Celtics held the highest numbers of tweets during the past year since they were both in the 2022 NBA Finals.
- The Lakers had a significant number of tweets because Lebron James is coming close to achieving the highest total-points-scored record in the history of the NBA
- The Cavs and Knicks had a significant number of tweets due to the significant players they acquired during Summer 2022.
- Increasing the quantity of text data acquired from the Twitter data extraction may have helped develop more observable differences between the TextBlob and vaderSentiment sentiment-analysis models.
- Twitter tweets had higher counts for teams that acquired significant players through trade.





Figures C-1 & C-2: The sentiment analysis for Western conference teams displays negative tweets overpopulating positive tweets, and most of Twitter discussing the most recent NBA champions more than any other teams.





Figures C-3 & C-4: The sentiment analysis for Eastern conference teams displays negative tweets overpopulating positive tweets, and most of Twitter discussing the recent NBA finalists and teams that performed significant trades during Summer 2022.

The top two teams, Warriors and Celtics, were selected for further analysis. The Warriors and Celtics text data was tested for sarcasm-detection. Sarcasm levels overall were observed to be at very low levels. Table C-2 displays the sarcasm prediction probabilities produced through LR and RFC models. Highest levels of sarcasm were observed when referring to ‘warriors’ or ‘celtics’ in-general.

Interpretations derived from Table C-2 (see Appendix Table C-2)**:**

- Low levels of sarcasm detected in this text data could be due to the brevity of communications found on Twitter.

Further semantic analysis was conducted using LSI and LDA. The LSI provided an indexing of the levels of correlation amongst the most commonly used terms and phrases, as can be seen in Figure C-5. The top 3 most commonly indexed terms using LSI for the Warriors were GoldenStateWarriors, Warriors, and DubNation.

Interpretations derived from Figure C-5 (see Appendix Figure C-5)**:**

- DubNation is still a popular method of referring to the the Golden State Warriors
- The full name GoldenStateWarriors is used on Twitter more than the abbreviation GSW possibly as a sign of respect.

LDA was applied next to create lists (Figure C-10 and Figure C-11) of the top terms with the top coherence scores, and to apply weights to Twitter comments that received more attention.

Interpretations derived from Figures C-6 & C-7 (see Appendix Figure C-6 & C-7)**:**

- ‘warriors’ , ‘gsw’, and ‘dubnation’ are some of the most commonly used references on Twitter when discussing the Golden State Warriors
- Discussion of Lebron James and Kevin Durant are still very common when conversing about the Golden State Warriors on Twitter.
- ‘teamceltics’, ‘greenrunsdeep’, and ‘celticstalk’ are commonly used references on Twitter when discussing the Boston Celtics

Attention was measured by the number of tweet replies given to each comment about that NBA team. The word-clouds developed (Figure C-8 and Figure C-9) displayed that common topics in discussion about the Golden State Warriors this past year were Jordan Poole, Draymond Green, and Let’s Go Warriors; and

common topics in discussion about the Boston Celtics during the past year were celtics-talk, team-celtics, and Blake Griffin.

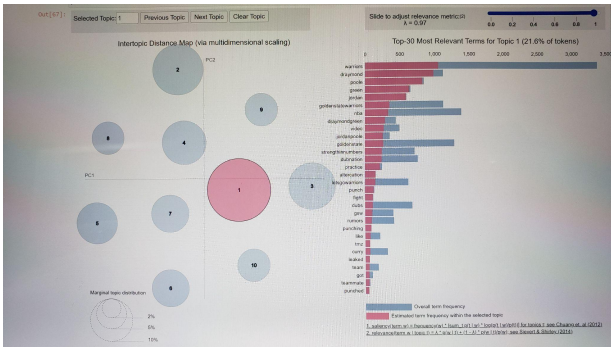
Figures C8 & C9: "draymond green" and "jordan poole" seem to be some of the more common and recent discussion topics when people discuss the Golden State Warriors on Twitter. Cheers such as "letsgowarriors", "strengthinnumber", and "dubnation" also seem to be common amongst Warriors tweets. Terms such as "teamceltics", "celticstalk", and "greenrunsdeep" seem to be common amongst Celtics fans on Twitter.

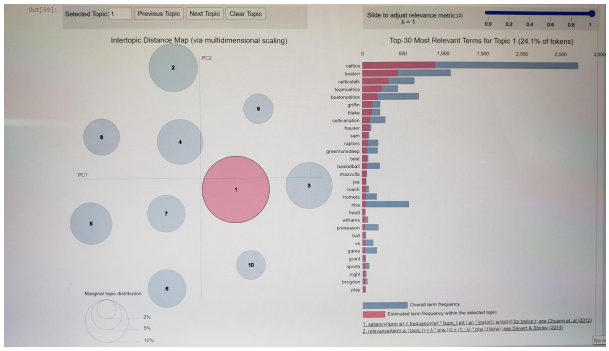
The relationships amongst text data observed through LDA were graphed, as can be seen in Figures C-10 and C-11. The data from Figure C-10 was derived from the Topic 2 list of terms developed during LDA. The term frequency when a term is compared to a specific selected topic, or when a term is compared to the entire text being analyzed, can be observed in Figures C-10 & C-11.

Adjusting the lamda levels of the relevance metric in these figures differentiates the terms' relevance-rankings.

Interpretations derived from Figures C-10 & C-11:

- Term relevance levels differentiate when viewing different topic clusters in the Intertopic Distance Map.
- Analysis of a term to a specific cluster of terms, or to the entire Twitter text dataset, can be observed.
- General terms, such as 'celtics', have more semantic relationships with other words in-contrast with more specific words.
- More specific terms, such as 'draymondgreen', are ranked more frequently when decreasing the relevance metric.





Figures C-10 & C-11: Interactive graphs were developed using LDA analysis to present the levels of topic-relevance amongst terms commonly used while discussing the Golden State Warriors and Boston Celtics on Twitter.

The general term 'warriors' held the highest term frequency amongst all of the Warriors text topics. The general term 'celtics' held the highest term frequency amongst all of the Celtics text topics. The NMF analysis, as can be seen in Figure C-12 and Figure C-13 (see Appendix Figure C-12 & C-13), displayed the topics with highest weight levels for the Warriors this past year to be 'letsgowarriors' and 'strengthinnumbers.' The topics with the highest weight levels for the Celtics this past year, according to NMF analysis, were 'greenrunsdeep' and 'celticsnation.' Another topic-relevance interactive map was created to observe and compare the term-relevance levels observed using NMF in-place of LDA (see Appendix Figure C-14).

Failure Analysis

Choosing to limit the analysis, and Twitter data extraction, to the past year permitted avoiding the significant changes that can occur in an NBA team in only 2-3 years. This limitation also prevented acquisition of high levels of text data for every NBA team, which could have limited some of the accuracy of the analyses conducted. Text data quantities, as can be seen in Table C1, range from 250 up to 6045. Development of a function or for-loop that could have more quickly developed the multiple data extraction codes I input for each team also could have been more time-effective. Lack of any result accuracies may or may not have resulted from more data-cleaning needed. The printing of LSI related terms, that had developed Table C-3 (see Appendix Table C-3) in the Warriors_NLP_Word_Analysis notebook, worked the first week I tried using it, but stopped working during the following weeks and had to be excluded from the Celtics word-analysis. The pyLDAvis visualization (see Appendix Figure C-14) worked for the NMF analysis on the Warriors text data, but encountered problems processing the Celtics text data during the NMF analysis. The initial semantic-similarity conducted on the Celtics and Warriors text data produced similarity levels of 0.0 and 0.0. These values are highly questionable due to other values appearing in other semantic-similarity testing (cosine similarity, LSI). Diagnosis of the conflict for this specific semantic-similarity testing was undetermined during the scope of this project. The levels of sarcasm detected in the August 2021 to September 2022 Twitter NBA text data were so small, the level of accuracy for that data could have been argued. Increasing the timespan from one-year to three-to-five years may have helped improve the sarcastic detection data analyses. Table C-3 (see Appendix Table C-3) was created when data-cleaning had accidentally been forgotten, and after adding data-cleaning to this project the coding for Table C-3 (see Appendix Table C-3) ran into errors. Improved proficiencies in python-programming can help prevent this project fault in the future. Some of the topics weighted do not seem to be correlated with NBA basketball. Adding more text-cleaning may help improve this result error.

Discussion

For the supervised learning portion (Part A), we learned that it can be relatively easy to overfit and have poor predictions if you don't use the right features. The results of our model were relatively good, but our model doesn't factor in player trades, injuries, and other features that are more granular than the team level data we used for prediction. If we had more time and resources, we would absolutely work on a model that does player level predictions per game that gets aggregated at the team level to predict whether a home team beats the visitor team. We'd also consider more advanced statistics that weren't readily available through the API we used. Nonetheless, we were impressed and excited that we were able to develop a predictive model entirely from scratch.

In part B, we learned that real-world analyzing is more difficult than expected. There were many unexpected failures that we will attend to. It is surprising that LDA performed better than NMF in topic modeling. Usually, NMF provided better interpretative results that have similar words in combination. However, LDA performed better with news, which implies LDA does better when there is a factual event. We can easily distinguish each event by using an LDA model, especially that LDA will not skip players names and teams. Because NMF struggled to find similar words to these names, the NMF results eliminated those important names from the results and made it hard to interpret. If we had more time and resources for this project, we could use a larger data set and test more hyperparameters. In addition, we could extend our data set by including more words such as adjectives and adverbs and obtain more informative information from the topic modeling.

For part C, the sentiment analysis (Figures C-1 & C-4), using both TextBlob and sentimentVader models, discovered that most of the NBA related tweets were negative. This is not surprising since the majority of moods in the news-media online, and in newspapers, is also negative. The levels of positive tweets for each NBA team seemed to exist proportionally with the existing levels of negative tweets, at much lower levels. Aiming to have at least 1000 tweets for each team could possibly help improve some of the analyses, but would have required more data past the one-year time-span and would have developed other influential factors, such as player-trades, to consider. Another potential solution for this data limitation could have been utilizing Generative Adversarial Networks (GANs) to develop random replicas of the data and increase the data quantities existing. Developing charts of the word-counts and the word-vector similarity ratings for future presentation was also considered. The sarcasm-detection analysis displayed the highest levels of sarcasm in the same terms and phrases that were observed during semantic analysis: 'draymondgreen', 'jordanpoole', 'goldenstatewarriors'. Developing more word-vector similarity calculations, and creating a more extensive analysis using only word-vector similarities, could have provided more insight about communications developed on Twitter during the past year about the Warriors and Celtics. Topic-indexing provided some surprising names not related to the Golden State Warriors such as 'bensimmons',

‘kevindurant’, and ‘lebronjames.’ Topic-indexing provided some interesting, yet not surprising, names related to the Boston Celtics such as ‘udoka’, ‘celticsnation’, and ‘greenrunsdeep.’ LDA provided Twitter tweets discussing each team ranked by topic-weight. Some common phrases on Twitter not all Warriors fans may be familiar with are ‘Dubnation’, ‘strengthinnumber’, and ‘letsgowarriors.’ Some common phrases on Twitter not all Celtics fans may be familiar with are ‘teamceltics’, ‘celticstalk’, and ‘celticsnation.’

The pyLDAvis visualizations provided a thorough and interactive way of studying the topic correlations existing amongst the most common topics discussed on Twitter and their relations to the other text data. This topic correlation can be very useful for determining which NBA topics are most popular and to utilize this popular-topic information for more effective advertising. Words that are commonly used already through social media will stay in people’s minds longer than words that are not as commonly used through social media discussions. LDA’s most relevant terms for Topic 2 (Figures C-10 & C-11) seemed to provide more NBA players not on the Golden State Warriors roster than the NMF’s relevant terms for Topic 2. Both models produced some general terms also that had no specific relations to the individual teams. NMF seemed to provide the more accurate results in this experiment compared to LDA. In LDA, independent components are normalized whereas in NMF they are not. This removes the independence of the LDA data, and could have played a role in why more NBA players not on the Warriors team were listed as related terms during LDA analysis. Having more time to develop more correlation between the supervised-learning and unsupervised-learning portions of this project also would have provided overall improvement.

Ethical Considerations

For the supervised learning portion (Part A), it would be unfortunate if someone believed our model accuracy was true and used the predictions to place bets on teams. Our predictions can potentially be completely incorrect and result in large financial damage for any individual or institution that uses them.

For the unsupervised learning portion with Reddit data (Part B), the biggest limitation within our project was time. The information collected online is a limited resource. For the Reddit data, we limited ourselves to posts with high scores, but this does not imply that posts with low scores do not contain common or interesting topics. This means our sampling method introduced selection bias that cannot ultimately be eliminated. This problem could be avoided by randomizing the posts without using the sort method, but this may increase the risk of potential data leakage. In addition, we extracted only a certain number of submissions from each query, which means there is missing and incomplete information in our data set. Therefore, our results cannot be used as an analysis of how a team might improve its branding strategy. Although our limited data can provide a team with a report based on users’ personal interests, there is uncertainty about these results. If our research is used for any commercial uses, such as athletic branding, we will address the implications of limited data resources and advocate for the use of complete data sets for future analysis.

Ethical issues that could possibly arise during this type of Twitter NBA text data analysis (Part C) would involve unintentionally offending an athlete after publishing his/her name in the data. It is important to help the public understand that data science is solely used to search for ways to help improve sports and other parts of the world, and that none of these analyses are meant to search for faults on the team. Respect of personal privacy is another issue that needs to be remembered as data scientists continue to conduct research about other people. It requires less time to remove a single piece of data after someone requests more respect of his/her privacy versus having to face legal accusations for defamation. Explaining the purpose of the research project, and the positive goals the project is aiming to achieve, always helps the public further understand the intentions of most data scientists are for the common good. Inappropriate language can also sometimes appear in extracted Twitter text data, and warnings should always be provided to public viewers that extraction of this type of data is unintentional and in no way meant to offend anyone. Notifications that the research being conducted is in no way intended to offend others, and is solely being studied to help support those who utilize social media, can help prevent some of these ethical situations.

Statement of Work

Every team member contributed equally in this project. Corbin covered the supervised learning using NBA numerical data (Part A); Levina (Minxin) covered the unsupervised learning, utilizing natural language processing (NLP) and Reddit data (Part B); Tom covered the unsupervised learning, utilizing NLP and Twitter data. Slightly different styles were applied while conducting the unsupervised learning (Part C).

References

1. Champs or Chumps (2022). *Best NBA Regular Seasons.* https://champsorchumps.us/records/best-nba-regular-season-records.
2. Dizikes, P. (2018, March 8). *Study: On Twitter, false news travels faster than true stories.* MIT News. https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308
3. Evans, A. (2022, June 3). *Klay Thompson passes LeBron James on 1 all-time NBA playoffs list.* Larry Brown Sports. https://larrybrownsports.com/basketball/klay-thompson-lebron-james-all-time-playoff-3-pointers-list/598112
4. Fox Sports. (2022, April 19). *The one thing Lakers will NOT do to end $63m NBA nightmare*. Fox Sports. https://www.foxsports.com.au/basketball/nba/nba-playoffs-2022-russell-westrbook-future-with-lakers-trade-options-charlotte-hornets-gordon-hayward-news-updates/news-story/e7477012827bf8742fe5ebc2a501549f
5. Jasser, J., Garibay, I., Scheinert, S., & Mantzaris, A. V. (2021). Controversial information spreads faster and further than non-controversial information in Reddit. *Journal of Computational Social Science, 5,* 111–122. https://doi.org/10.1007/s42001-021-00121-z.
6. Karlis, Michael (2022, September 6). *Coach Gregg Popovich warns San Antonio Spurs fans not to plan on a championship season.* San Antonio Current. https://www.sacurrent.com/arts/coach-gregg-popovich-warns-san-antonio-spurs-fans-not-to-plan-on-a-championship-season-29946283.
7. Kennedy, A. (2022, January 28). *An inside look at how the NBA became a social-media juggernaut.* BasketballNews.com.https://www.basketballnews.com/stories/an-inside-look-at-how-the-nba-became-a-socialmedia-juggernaut
8. NBA.com. (n.d.). *Andrew Wiggins, Warriors agree on 4-year contract extension.* NBA.com. https://www.nba.com/news/andrew-wiggins-warriors-agree-4-year-contract-extension
9. Palka, K. *Athletes and personal branding: The power of social media.* (2022, April 19). EasyPromos.com. https://www.easypromosapp.com/blog/en/2021/02/personal-branding-athletes-bewolfish/
10. Park, Danny (2019). *Balldontlie API*. https://www.balldontlie.io/.
11. Basketball Reference (2022). https://www.basketball-reference.com/leagues/NBA_2023_games.html.
12. Polacek, S. (2022, April 6). *LeBron James, Lakers eliminated from 2022 NBA playoff race after loss to Suns.* Bleacher Report. https://bleacherreport.com/articles/10031659-lebron-james-lakers-eliminated-from-2022-nba-playoff-race-after-loss-to-suns
13. Rutto, C.J. & Gilbert, E.E. (2014). *Vader: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.
14. *The one thing Lakers will NOT do to end $63m NBA nightmare*. (2022, April 19). Fox Sports. https://www.foxsports.com.au/basketball/nba/nba-playoffs-2022-russell-westrbook-future-with-lakers-trade-options-charlotte-hornets-gordon-hayward-news-updates/news-story/e7477012827bf8742fe5ebc2a501549f
15. Yadav, P. *Difference between Reddit and Twitter.* (2022, September 29). Askanydifference.com. https://askanydifference.com/difference-between-reddit-and-twitte/

Appendices

Table C-1: Number of Twitter Tweets Extract for Each NBA Team

Western Conference Team	Number of Tweets Extracted	Eastern Conference Team	Number of Tweets Extracted
Clippers	1111	Bucks	874
Lakers	3104	Celtics	6045
Suns	895	Hornets	526
Warriors	4927	Heat	1362
Mavs	1025	Hawks	518
Grizzlies	956	Nets	2429
Kings	258	Bulls	1316
Spurs	570	Raptors	888
Nuggets	588	Magic	960
Timberwolves	947	76ers	2822
Rockets	510	Knicks	4638
Thunder	228	Cavs	2466
Pelicans	514	Pistons	1559
Jazz	292	Wizards	250
Blazers	1187	Pacers	1284

Table C-1: Number of Twitter tweets extracted for each of the 30 NBA professional basketball teams while staying within the timespan of August 6, 2021 to September 12, 2022 to help avoid larger team-roster changes.

Table C-2

Model	Percent Sarcasm Detected for Warriors Text	Percent Sarcasm Detected for Celtics Text
Logistic Regression	0.072	0.05
Random Forest Classifier	0.003	0

Table C-2: Sarcasm detection percent levels measured using Logistic Regression and Random Forest Classifier. Very low levels of sarcasm were detected in the Twitter data possibly due to the limit of text permitted in each Twitter tweet.

Figure C-5

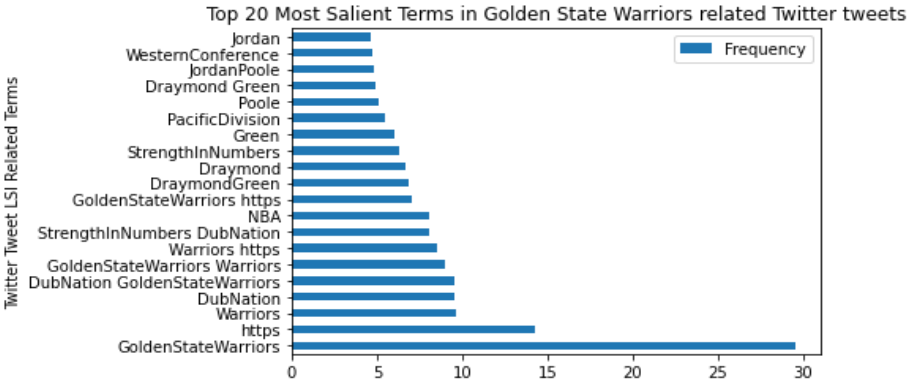


Figure C-5: The 20 most salient terms rated using LSI-ranking.

Figure C-6

topic 0: ['warriors', 'nba', 'vs', 'golden', 'state', 'wizards', 'video', 'new']
topic 1: ['warriors', 'goldenstatewarriors', 'dubnation', 'goldenstate', 'nba', 'basketball', 'best', 'gsw']
topic 2: ['klaythompson', 'vs', 'lebronjames', 'stephencurry', 'kevindurant', 'warriors', 'bensimmons', 'damianlillard']
topic 3: ['strengthinnumbers', 'letsgowarriors', 'goldenstate', 'rumors', 'warriors', 'preseason', 'dubs', 'thompson']
topic 4: ['warriors', 'draymond', 'poole', 'green', 'jordan', 'goldenstatewarriors', 'nba', 'draymondgreen']
topic 5: ['warriors', 'lana', 'stevekerr', 'lunafreaks', 'fridaymorning', 'postseason', 'draymond', 'fridayfeeling']

Figure C-7

topic 0 ['boston teamceltics', 'teamceltics', 'teamceltics celtics', 'boston', 'celticstalk', 'celtics celticstalk', 'celtics', 'celtics boston']
topic 1 ['greenrunsdeep', 'celticsnation', 'greenrunsdeep bostonceltics', 'celticsnation greenrunsdeep', 'celtics celticsnation', 'bostonceltics', '22 celtics', '10 22']
topic 2 ['celtics', 'game', 'love', 'season', 'let', 'bleedgreen', 'preseason', 'celtics bleedgreen']
topic 3 ['blake', 'griffin', 'blake griffin', 'debut', 'griffin celtics', 'celtics', 'solid', 'blakegriffin']
topic 4 ['hauser', 'sam', 'sam hauser', 'role', 'hauser celtics', 'role celtics', 'going', 'making']
topic 5 ['tatum', 'jayson', 'jayson tatum', 'tatum au', 'au', 'au tarif', 'tarif', 'vert']

Figures C-6 & C-7: Word-lists, sorted by levels of coherence in relation to the extracted NBA Twitter text data, were developed through Latent Semantic Indexing. The lower the topic-number, the higher the levels of coherence with other terms.

Figure C-12

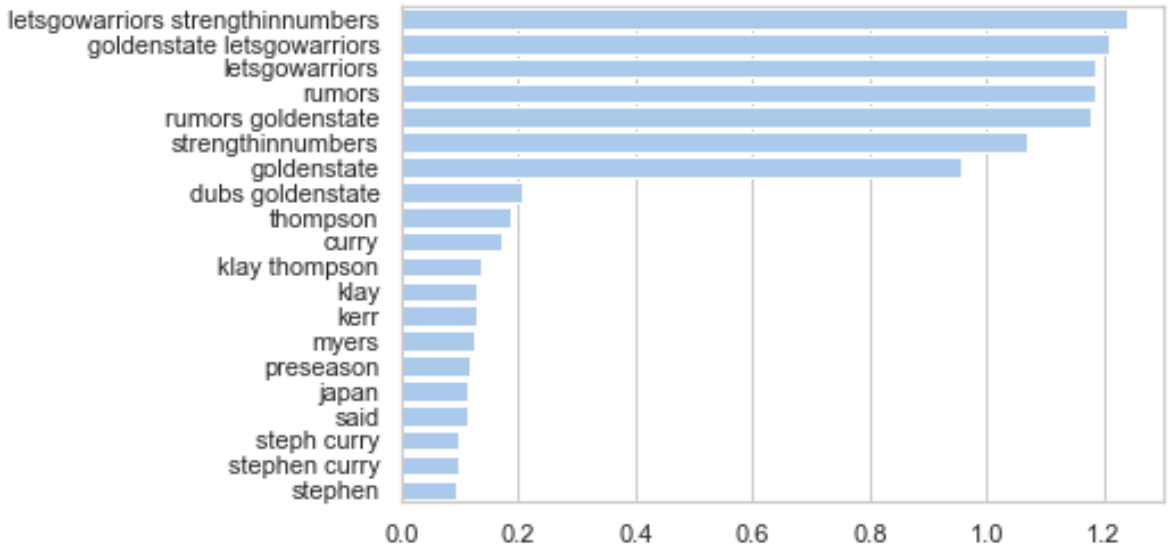


Figure C-12: The weight levels from topic 1 while utilizing Non-negative Matrix Factorization on the Warriors Twitter text data.

Figure C-13

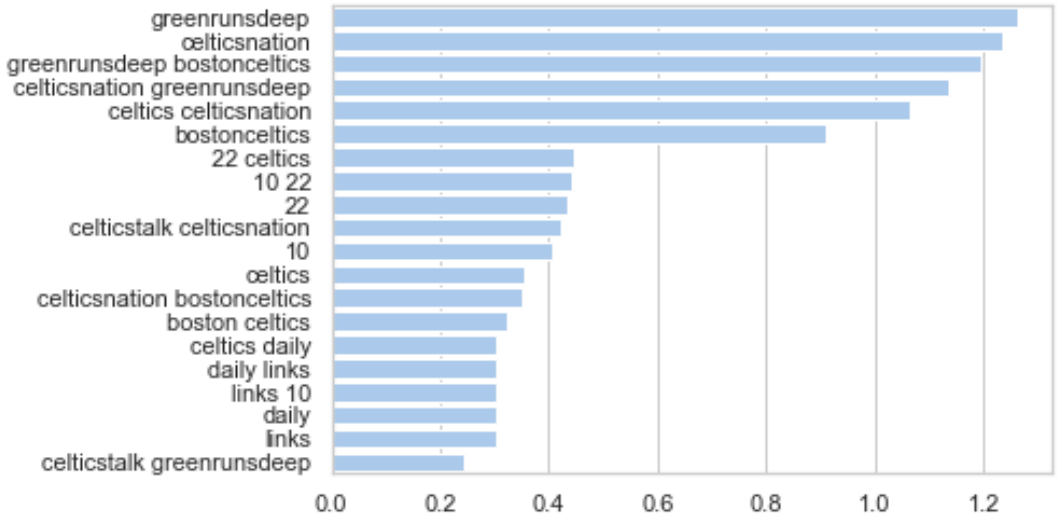


Figure C-13: The weight levels from topic 1 while utilizing Non-negative Matrix Factorization on the Celtics Twitter text data.

Figure C-14

Figure C-14: An interactive graph was developed using NMF analysis to present the levels of topic-relevance amongst terms commonly used while discussing the Golden State Warriors on Twitter

Table C-3

	Twitter Tweet LSI Related Terms	Frequency
0	GoldenStateWarriors	29.53
1	https	14.29
2	Warriors	9.61
3	DubNation	9.58
4	DubNation GoldenStateWarriors	9.52
5	GoldenStateWarriors Warriors	9.00
6	Warriors https	8.54
7	StrengthInNumbers DubNation	8.09
8	NBA	8.04
9	GoldenStateWarriors https	7.04
10	DraymondGreen	6.85
11	Draymond	6.67
12	StrengthInNumbers	6.31
13	Green	6.07
14	PacificDivision	5.49
15	Poole	5.13
16	Draymond Green	4.95
17	JordanPoole	4.81
18	WesternConference	4.70
19	Jordan	4.66

Table C-3: LSI-Related Term Frequency using Golden State Warriors Twitter data