NBA Regular Season MVP Prediction

Submitted by Team 13

Can Song	Chenghao Zhan	Jing Zhao	Jinil Kim	Yaqing Luo
A0206478X	A0206448A	A0206460N	A0206472J	A0206690E

Abstract

Every May to June, basketball fans all over the world tensely wait for an announcement, National Basketball Association (NBA) regular season Most Valuable Player (MVP). It is perceived as the player whose performance is the most prominent over the season, and determined by media associates' votes. In this project, to predict 2019/2020 regular season MVP, we used two approaches, a statistical approach that uses player performance statistics to perform regression and a text-based approach that uses tweets of the voters and sentiment analysis. Different machine learning algorithms are applied in our analyses, and parameters are tuned to obtain higher accuracy. Through the project, two hypotheses are tested: 1. Performance statistics can capture players' value using machine learning methods, 2. Voter's preferences are reflected in their tweets, and are observed from tweets sentiment.

1. Introduction

The National Basketball Association (NBA) Most Valuable Player (MVP) Award is an annual award given to the best player of the NBA regular season. It is one of the most anticipated and speculated awards of NBA by the players, fans, and media. Not only is the MVP award a great honor for the winner, it also significantly boosts the commercial value of both the player and the team. For example, the annual salary of Giannis Antetokounmpo, the 2019 NBA MVP, is expected to jump from USD 25M to over USD 40M in 2021 when his current contract ends. And the total franchise value of all NBA teams in season 2019/2020 is USD 63.69 billion, which is almost 17% of GDP of Malaysia in 2020.

Starting in 1980/1981 season, the MVP has been decided by a penal of media associates working closely with NBA. There have been variations in the number, location, and agencies of appointed voters. With the latest rules of NBA, the MVP voting poll is comprised of one hundred media voters, who must be independent of teams, and one public fan vote. Each vote includes five ranked nominations, with each place worth 10, 7, 5, 3, and 1 points respectively from the first to fifth places. The player with the highest total points wins the MVP of the year. In this project, we propose to predict the 2020 MVP award using two approaches:

• Applying machine learning algorithms on structured player performance data (statistical approach)

• Applying natural language processing (NLP) techniques on media voters' Twitter feeds and commentaries (text-based approach).

Through the first approach, we test the hypothesis that the term "most valuable" is reflected in players' performance statistics, although it is relatively well explored by many data scientists. Use cases for the second approach are quite rare. While statistics of player performance limits the potential candidates down significantly, in case there is a close tie in performance among multiple players, we hypothesize that voters' personal preferences play critical roles in deciding their final votes, which can be captured by the sentiment of their tweets.

In the analysis, we assumed that MVP of the season is highly likely to be selected as the All-star in the same season, so All-star player name lists of each season from NBA official website are used to narrow down the scope of analysis.

The next sections will be covered in the order of: 2. Related work, 3. Statistical approach with relevant data description, pre-processing, modeling methods, model refinement, and evaluations, 4. Text-based approach with the same workflow as in part 3 along with sentiment training, 5. Result interpretation and comparison, 6. Conclusion and future works.

2. Related Work

As a popular sports topic, plenty of related researches have been achieved previously to predict MVP regular season, MVP regular season, winning team, so on and so forth.

Github Directory: https://github.com/zhao0872/Projects.git

However, most of them are not published studies, but rather personal posts on blogs. These predictive analyses usually use different machine learning methods, including regression methods such as multiple linear regression, classification methods like logistic classifiers, and neural network models, and player performance statistics, including features like count of three-point and time on court, are mostly used. Besides, some of the most reputed researches that are published are reviewed below. In the prediction of regular season result by Yuanhao Yang, the multiple linear regression model is fitted on self-defined features using player performance and team statistics, namely team PER and win ratio (2015). Team PER is defined from personal PER by multiplying the minutes that the top five players (in terms of minutes played) are active on court. Win ratio is determined by the count of winning games of the team divided by the total count of games for the team. In the study, these two features are proved to be practical in predicting regular season results as they capture the fundamental rule of how teams are selected into the regular season. However, it failed in detecting an extremely small difference in win ratio across teams. That is, the model is restricted by a variety of features included for predictive analysis. Another relevant research is using neural network model to predict regular season MVP by Yuefei Chen et al (2019). Different datasets are used, including performance statistics which is composed of 17 features such as total field goal, three-point field goal of regular season and a mixed dataset with partial performance data and some further defined features like win ratio. And the framework of neural network is relatively basic, which has three layers and used mini-batch gradient descent algorithm. The activation function is not specified in the paper. They successfully predicted the regular season MVP of 2009/2010 and 2016/2017 seasons.

3. Statistical Approach

3.1 Data Source and Collection

Three parts of data are used in the statistical approach, including player performance data, all-star player name list , and MVP voting results.

Player Performance Data: This group of data contains all the statistics for every NBA player in each season, including their basic box-score statistics per game and other advanced performance measurements. We scrape the data from 'Basketball Reference', a website which transcribes all the basketball records, and the final dataset we obtain has 16,841 rows and 31 columns.

All-star Player Name List: Based on the assumption that MVP of the season is highly likely to be selected as the all-star in the same season, we obtain All-star player name list of each season from NBA official website. There are 956 players (includes the same player in different seasons) in the all-star player name list since 1980-81 season.

MVP Voting Result: From NBA official website, we obtain the rank and award share for every MVP candidate in each season.

3.2 Data Description

After combining the three datasets we mentioned above, we get our final dataset for prediction with 933 rows and 32 columns. All features can be classified into the following four categories:

Personal Statistics: We mainly focus on three personal statistics: position, age, team. For position, due to changes in game rules and game styles, point guards and small forwards who take more offensive duties have gotten attention and appear more than centers in all-star and MVP candidate lists for the recent 10 years. For age, we observe a positive correlation between a player's age and his possibility of receiving votes in MVP voting. This is mainly because an MVP candidate will stay competitive in the following few seasons and receive MVP votes again. For team, we assume that two players from the same team will not both receive high MVP votes, as voters may take their excellent performance as granted.

Voting Statistics: We take the award share of each MVP candidate as the y-variable for our prediction. Here award share is defined as the proportion of points from votes a candidate received over the maximum possible points a candidate could receive.

$$Award\ Share = \frac{Points\ Received}{Maximum\ Possible\ Points}$$

This measurement eliminates the effects of the variance of voting panel size in different seasons.

Box-score Statistics: Box-score includes points, rebounds, assists, steals, and blocks a player gets in each game, which directly reflects the performance of a player. Field goal percentage (FG%) and turnovers are also key features which can reflect the player's scoring efficiency.

Table 1. Features used for Statistical Approach from 4 categories.

PERSONAL	VOTING	BOX-SCORE	ADVANCED
Position	Share	PTS	EFG
Age		TRB	PER
Team		AST	TS
		STL	USG
		BLK	WS
		TOV	WS/48
		FG%	BPM
		3P%	VORP
		FT%	

Advanced Statistics: Advanced statistics are statistics that measure players' performance from a higher level, mainly focus on players' influences on games and contribution to the team. For example, value over replacement player (VORP) estimates the points per 100 possessions that a player can contribute above a replacement-level player, which is a key feature for identifying outstanding players.

3.3 Data Preprocessing

Before we put all the features into our models, several preprocessing steps are conducted as follows.

- One-hot encoding is taken for categorical variables, position, and team.
- Highly correlated features (correlation exceeds 80%) are removed.
- Cube and cube root transformed features are created for all numerical features.
- Normalization is performed for all numerical features, including cube and cube root terms derived from the previous step.

3.4 Models

3.4.1 Approach Description

To predict the MVP Award winner, the first approach (Approach 1) we use is to run regression models on award share and MVP player predicted will be the one has largest predicted award share.

But the award share is actually highly skewed, with around half of the observations having 0 award share, which reflects the nature of MVP election: top players receive most votes while a lot more players have zero or little vote. To solve this problem, one method is to convert the award share into rank and perform regression models on rank instead of award share (Approach 2). Another method is to classify the players into 3 groups (high vote, little vote, zero vote) and only perform regression models on players in group 1 and 2. To classify players, we can treat classes as different parallel categories and perform classification models (Approach 3). Moreover, ordinal regression models can be used as there is implicit order among groups (Approach 4).

Above 4 approaches cover Regression, Classification and Ordinal Regression Models as shown in Table 2. The models are selected to cover models from different theory basis to increase the probability of getting a better prediction result.

Table 2. Models

	Model
Regression	Linear Regression, Ridge Regression, Lasso Regression, Elastic-Net, SVR, Random Forest, Gradient Boost, Ada-Boost, Multi-Layer Perceptron(MLP)
Classification	Logistics Regression, KNN, Decision Tree, Random Forest, SVC, XGBoost, Gradient Boost, Ada-Boost, MLP
ordinal Regression	LogisticIT, LogisticAT, Ordinal Ridge, LAD

3.4.2 Model Evaluation

For the regression result, we start with the evaluation methods of both award-share-based analysis including MAE, MSE, MAPE and ranking-based analysis including

- MVP Average: whether MVP is correctly predicted
- MVP in Top 3 Average: whether predicted top 3 players include the actual MVP
- MVP in Top 5 Average: whether predicted top 5 players include the actual MVP

As our goal is to predict MVP, we hypothesize that the evaluation based on the ranking of predicted award share is more meaningful than the ones directly based on award share. So, we choose MVP Average, MVP in Top 3 Average and MVP in Top 5 Average as our evaluation indicators.

As stated before, our dataset contains yearly data from season 1980/1981 to season 2019/2020. For training and validation, we use leave-one-out cross validation to train 38 years and validate the result for 1 year. And based on cross validation result on 3 indicators we choose, we use the best model after hyperparameter tuning to predict 2019/2020 season MVP

3.4.3 Approach 1

By visualizing the award share data, we find that data is highly skewed with most entities having 0 vote.

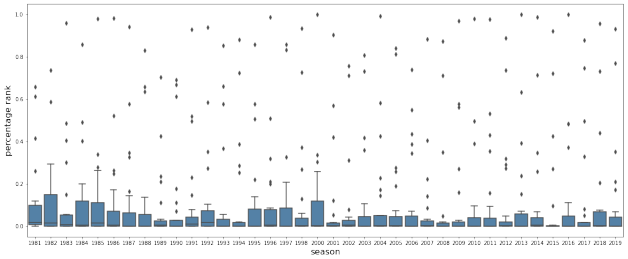

Figure 1. BoxPlot for Award Share Data across Years

For Approach 1, the best model is MLP which gives the MVP_Average of 60.5%, which means the percentage of accurately predicted MVP is 60.5%. The full result of Approach 1 is listed in Table 3 below.

Table 3. Hyperparameter Tuning Result for Approach 1

RANK	MODEL	HYPER PARAMETER	MVP AVG (%)	MVP IN TOP 3 AVG (%)	MVP IN TOP 5 AVG (%)
1	MLP	alpha: 0.0001, hidden_layer_size : (300,)	60.5	84.2	92.1
2	SVR	kernel: rbf, C: 1	57.9	84.2	92.1
3	Gradient Boost	learning_rate: 0.01, n_estimators: 200	57.9	81.6	97.4
4	Ada-Boost	learning_rate: 0.1, n_estimators: 100	57.9	86.8	94.7
5	Ridge	alpha: 1e-05	55.3	92.1	92.1

3.4.4 APPROACH 2

By converting y-variable from award share to the rank of award share, the data is less skewed as shown in Figure 2. The rank adopted in Approach 2 is dense rank to make the ranking consecutive. Besides, as we have different number of players for every year, percentage rank is more reasonable and is applied.

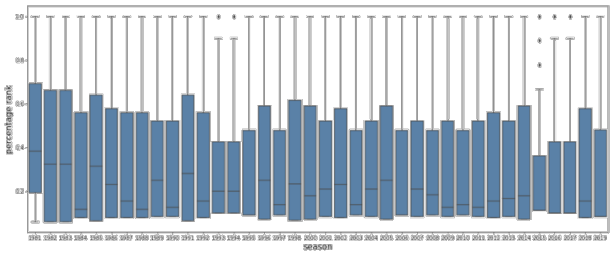

Figure 2. BoxPlot for Award Share Rank across Years

Comparing the result of Approach 2 to that of Approach 1, we find that the MVP_Average doesn't improve. But MVP in Top 3 Average and MVP in Top 5 Average increase to 90% and 95% separately.

From the coefficient plot for linear SVR (Figure 3), we find that the features have larger coefficient magnitude are more likely to be crucial features to evaluate players' value.

Table 4. Hyperparameter Tuning Result for Approach 2

RANK	MODEL	HYPER PARAMETER	MVP AVG (%)	MVP IN TOP 3 AVG (%)	MVP IN TOP 5 AVG (%)
1	SVR	kernel: linear, C: 100	60.5	89.5	94.7
2	Linear	None	57.9	89.5	94.7
3	Ridge	alpha: 1e-05	57.9	89.5	94.7
4	Lasso	alpha: 0.01	57.9	89.5	94.7
5	Elastic-Net	alpha: 0.1, l1_ratio: 0.1	57.9	92.1	94.7

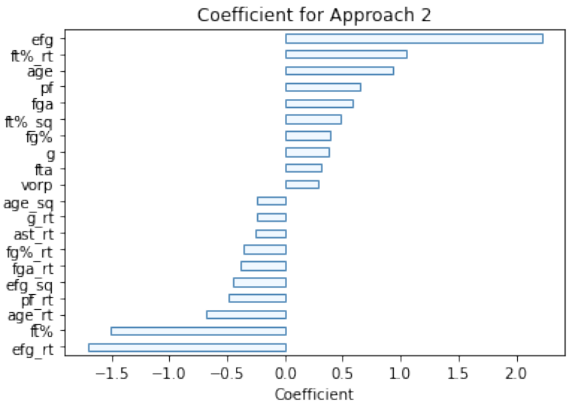

Figure 3. SVR Feature Coefficient for Approach 2

3.4.5 APPROACH 3

We classify the data into 3 groups, class 1 stands for the data with vote larger than the median vote of the year, class 3 stands for players with 0 vote of that year and left data is labeled as class 2.

After hyperparameter tuning, we find that the predicted class 1 classified by Gradient Boost Classifier has 100% probability of predicting MVP player, and 95.8% probability of predicting top 5 players.

To be more conservative, we still include both predicted class 1 and class 2 observations as the dataset for the regression model. As shown in Figure 4, the skewed y-variable problem has been relieved.

Table 5. Hyperparameter Tuning Result for Classification stage of Approach 3

RANK	MODEL	HYPER PARAMETER	ACCURACY (%)	TOP 5 IN PRED CLASS 1(%)	MVP IN PRED CLASS 1 (%)
1	Gradient Boost	learning_rate: 0.1, n_estimators: 100	75.0	95.8	100.0
2	XGB	n_estimators: 500, max_depth: 5, subsample: 0.6	74.5	95.8	100.0

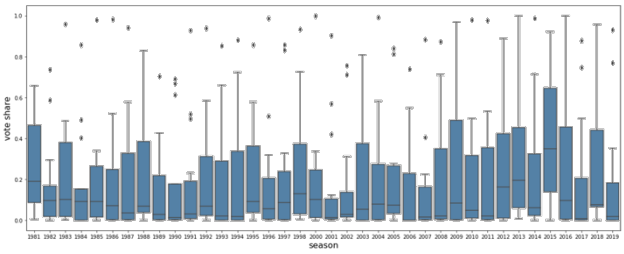

Figure 4. BoxPlot for Award Share Rank across Years (Approach 3)

From Table 6, we find that the MVP Average given by Ridge Regression increases to 65.8% compared to previous approaches.

Table 6. Hyperparameter Tuning Result for Approach 3

RANK	MODEL	HYPER PARAMETER	MVP AVG (%)	MVP IN TOP 3 AVG (%)	MVP IN TOP 5 AVG (%)
1	Ridge	alpha: 0.1	65.8	89.5	92.1
2	Lasso	alpha: 0.001	63.2	92.1	92.1
3	Elastic-Net	alpha: 0.001, l1_ratio: 0.9	63.2	92.1	92.1

3.4.6 APPROACH 4

As class 1, 2 and 3 contain order information, which is not captured by the previous classification models. We perform ordinal regression models using python library *mord*. Comparing the result with classification, we can see that ordinal regression has lower accuracy (Table 7). Besides, the probability of top 5 players in predicted class 1 dropped to 90.5%, so to be more conservative, we include both predicted class 1 and 2 as the dataset for regression.

Moreover, from table 8, regression result for approach 4 is similar to that of approach 3, with 65.8% MVP Average (better than 60% of approach 1 and 2) and 92.1% MVP in Top 3 Average.

When we plot the coefficient for best model, it's clear that the team dummies carry some prediction power now (Figure 5).

Table 7. Hyperparameter Tuning Result for Approach 4 (Ordinal Regression)

RANK	MODEL	HYPER PARAMETER	ACCURACY (%)	TOP 5 IN PRED CLASS 1(%)	MVP IN PRED CLASS 1 (%)
1	Logistic IT	'alpha': 1	65.9	90.5	97.4
2	Logistic AT	'alpha': 0.01	65.2	88.9	97.4

Table 8. Hyperparameter Tuning Result for Approach 4

RANK	MODEL	HYPER PARAMETER	MVP AVG (%)	MVP IN TOP 3 AVG (%)	MVP IN TOP 5 AVG (%)
1	Lass	alpha: 0.001	65.8	92.1	92.1
2	Elastic-Net	alpha: 0.001, l1_ratio: 0.1	65.8	92.1	92.1
3	SVR	kernel: linear, C: 1	65.8	86.8	92.1

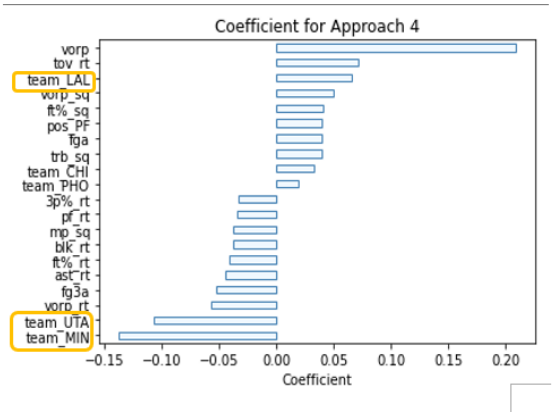

Figure 5. Lasso Feature Coefficient for Approach 4

3.5 Modeling Result and Analysis

The 2020 predictions given by the best model after hyperparameter tuning of 4 approaches are described in Table 9.

Table 9. Predicted Top 5 Players

	APPROACH 1	APPROACH 2	APPROACH 3	APPROACH 4
1	Giannis Antetokounmpo	Giannis Antetokounmpo	Giannis Antetokounmpo	Giannis Antetokounmpo
	(1.02)	(0.79)	(0.74)	(0.65)
2	LeBron James	LeBron James	LeBron James	LeBron James
	(0.73)	(0.75)	(0.59)	(0.53)
3	Anthony Davis	James Harden	James Harden	James Harden
	(0.36)	(0.68)	(0.34)	(0.41)
4	Luka Doncic	Anthony Davis	Anthony Davis	Anthony Davis
	(0.34)	(0.53)	(0.31)	(0.30)
5	Domantas Sabonis	Luka Doncic	Kawhi Leonard	Luka Doncic
	(0.31)	(0.51)	(0.27)	(0.29)

For all 4 approaches, the MVP is predicted to be Giannis Antetokounmpo and the player with second most vote is predicted to be LeBron James.

Comparing 4 approaches, Approach 3 and Approach 4 give better accuracy in MVP Average as 65.8% compared to 60.5% MVP Average given by Approach 1 and Approach 2, which indicates that classifying the players first and then running regression for predicted high-vote players may be a good way to solve the skewed data problem.

4. Text-based Approach

In order to make predictions that reflect real-world perceptions toward players, we adopt a text-based approach, where we predict MVP based on sentiment scores given to tweets of the MVP voter candidates. Our approach is mainly composed of four parts: data acquisition, data pre-processing, sentiment training, modeling, and prediction.

4.1 Data Collection and Description

Three types of datasets are retrieved for our text-based approach:

- Tweet data of voter candidates;
- Historical vote data (our final target variable);
- Text dataset for sentiment training including NLTK Tweets[2], Sentiment140[3], and self-scraped tweets with positive and negative emoticons

Tweets data is retrieved using "TWINT", an advanced Twitter scraping tool from which we obtain tweets of 73 media voter candidates[1], excluding 22 new, 2 non-tweeting, and 3 non-English-speaking voters from the original list of 100, with dates extending from 11 May 2016 to 8 March 2020. For every voter, Twitter handle, used for scraping of the data, is manually obtained via web search, while non-English-only speakers are excluded as our language processing is based on methods specific to English only. The attributes of the dataset include *user_id*, *username*, *date*, *time*, and *tweet* data over 610,578 datapoints. The collected data is altogether "unprocessed".

Historical vote data is also for past years, which is required for training of MVP prediction models.

There are 6000 rows (tweets) in NLTK Tweets dataset, 1.6 million in Sentiment140 dataset and 6000 in tweets with emoticons. The tweets with emoticons are scraped based on two lists of emoticons described in Table 10.

Table 10. Emoticons

POSITIVES	:)	:-)	:)	:D	=)
Negatives	:(	:-(	: (		

4.2 Data Pre-processing

Social media text is usually much less structured compared to articles or papers, hence requiring a different set of pre-processing to accurately extract the meaning. The following steps are implemented to process all of the aforementioned datasets.

- Tokenization
- Conversion of all twitter mentions of the players into a word[2] (stripping off the @ sign)

1. The list of media voters for the MVP award is disclosed upon the announcement of the MVP winner around June every year, which prevented us from inferring exactly who will be the voters for the year of 2020. However, the All-Star Starter Voting has a media panel as well, in which about 90% of the voters also vote for the MVP award at the end of the regular season. Since All-Star Starter voting panel is already announced, we used this list of voters as an alternative choice for voter candidates.

2. The handles are included in the nickname list for player identification later on.

- Removal of all numbers, spaces, punctuations, links, and other user mentions
- Removal of none-letters characters
- Lemmatization
- Conversion of exaggerated spelling (e.g. “happppyyy” to “happyy")

Note that stopwords are kept on purpose, as tweets are already short and the typical stopwords are likely to contain meanings that should not be ignored. The processed tokens are used for the sentiment analysis.

4.3 Sentiment Training

In order to assign sentiment to voter’s tweets we obtained and further fit the sentiment to the actual vote, training on labelled datasets is required. However, depending on the training data, the scores assigned can vary significantly. Additionally, depending on the context the choice of words differs dramatically. For example, the word “killer” is considered a negative term in a common context, but is a positive term in sports-related context. Therefore, multiple score assignment methods are adopted, including one method that implements languages specific to the context of NBA. All scores are scaled to -1 to +1 range, corresponding to negative and positive sentiments, respectively. Description for each score assignment method is as follows.

TextBlob: TextBlob is an open source Python package with pre-trained functions to give sentiment polarity scores based on the sequence of words. This in-built function is used, without any additional training required, to assign sentiment scores to tweets used for prediction. However, the training done for the in-built function does not incorporate NBA context.

NLTK Tweets: NLTK, an open source Python package specifically designed for natural language processing, also provides a training set that can be used for sentiment score assignment. Once training datasets, which is well-balanced with 3,000 positive and 3,000 negative tweets, are loaded, “Positive” and “Negative” labels are transformed into numeric values and trained on NLTK’s in-built Naïve-Bayes binary classification model, which gave a validation accuracy of 0.895. The advantage of this method is its convenience in training and applying to give scores, similar to TextBlob. Binary scores, which is the default output of this in-built Naïve-Bayes model, are given to tweets used for prediction.

Sentiment140: Sentiment140 is an open source dataset that contains 1.6 million labelled text data, well-balanced with 0.8 million non-negative (positive and neutral) and 0.8 negative data. The original labels are translated into 1 and 0, respectively, using LabelEncoder of Scikit-learn[2]. An LSTM model of Keras[3], suggested by the data providing party, along with Gensim’s Word2Vec [4] output implemented in a word embedding layer, is used for training of the labelled data, with validation accuracy of 0.781. The main advantage of this method is that the data is large that it can incorporate almost all possible sequence of words. However, due to the same reason it takes days for the training alone. Scores are assigned using the trained model for tweets used for prediction.

Tweets with Emoticons: Although the three methods mentioned above are widely used sentiment score assignment methods, they are incapable of conveying the context of NBA by themselves. Therefore, it is necessary for us to create a training dataset that reflects the language used in NBA context. 3,000 positive and 3,000 negative tweets are labelled based on emoticons contained in the sequence of words, are applied with TF-IDF vectorizer to extract text features and are analyzed using 5 different machine learning modelling methods. Among Naïve-Bayes classifier, logistic regression, decision tree classifier, random forest classifier, and XGBoost classifier, logistic regression showed the best cross-validation accuracy of 0.742. Using this model we assigned sentiment scores to the tweets used for prediction.

Once sentiment scores are given for every tweet, we perform a feasibility validation of the four methods of sentiment score assignment by manually labelling[5] 100 NBA-context tweets and comparing the scores. The result can be found in *Table 11*. While not all methods show satisfactory performances from the comparison, we move onto modelling using all four methods as predictors, as combinations of these methods can yield to improved predictions.

Table 11. Comparison of the four sentiment score assigning methods. We assume manually labelled score, Baseline, to be the ground truth.

Method	MSE	MAE
Baseline	0	0
TextBlob	68.92	76.25
NLTK Tweets	135.00	75.00
Sentiment140	141.82	103.77
Tweets with Emoticons	147.61	110.10

[2] Scikit-learn is an open source Python library providing both supervised and unsupervised machine learning algorithms.

[3] Keras is an open source Python library providing functionalities regarding neural networks.

[4] Gensim is an open source Python library for unsupervised topic modelling and natural language processing.

[5] Labelling was performed by personal with deepest understanding of NBA context.

4.4 Models

Two different approaches are developed to explore how well the tweets reflect the voter's preferences and whether they link to the actual votes. The first approach focuses on the ranking of the voter's tweet metric data towards different players. The second approach maps out the tweet metric data of the top 5 players in the actual MVP vote, and attempts to develop a link from the metric to the rank.

For player-sentiment mapping, a dictionary of player nicknames is created for 32 most commonly mentioned players and thus most likely MVP candidates including 24 All-Star players. Nicknames are drawn from use cases in NBA-related tweets along with articles referring to players in a variety of nicknames. Since our modelling requires total sentiment score given to each player per voter, this dictionary is used to generate total sentiment score more accurately and efficiently.

All four sentiment scores obtained from the previous part are used for this analysis part so as not to discriminate any model upfront based on accuracy scores alone. While sentiment score is undoubtedly the focus of this study, tweet count is also used as a metric to evaluate the interest of the voter. As the saying goes, 'there is no such thing as bad publicity'.

For the accuracy of metric mapping for each year, the 'voting year' for the analysis is defined to start on the day after the previous year's MVP award, and end the day before this year's MVP award. For example, the 2019 voting year tweets include all tweet published between Jun 26, 2018 and Jun 23, 2019.

4.4.1 Method 1 – Metric Ranking

Given that the data is available for the different metrics for each tweet and for the corresponding player linked to it, this model tries to simulate the voters' preferences based on the individual tweet metrics.

After assigning the metric score of the tweet to the mentioned player(s), we group the data by the voting year and voter, and sum up the metric score of all players for each voter in the year. The summed metric scores are then sorted in descending order, and the top 5 scores among the players mentioned by the voter is assumed to be the top choices of the voter. The filtration of only choosing mentioned players is crucial considering some top scores might be negative and smaller than the default 0, or some voters simply mentioned less than 5 players in the year.

The top rank list is compared to the actual votes of the voters. Without considering the rank, the NTLK Tweets sentiment score generates an 18.28% match with the actual vote, and the tweet count 16%. With rank considered, the numbers are significantly lower unsurprisingly.

Table 12. Summary of different metrics ranking matches

Metric	Textblob	NLTK	Sentiment 140	Tweets w/ Emoticon	Tweet Count
Top 5	14.38%	18.38%	12.00%	7.63%	16.00%
Top 5 ranked	2.25%	2.00%	2.00%	1.63%	3.75%
Top 3	3.75%	4.50%	4.38%	4.25%	2.88%
Top 3 ranked	1.63%	1.38%	1.13%	0.88%	2.38%

4.4.2 Method 2 – Top Player Mapping

The MVP votes can be added up based on their corresponding points to rank the players. The second method identifies the top 5 players for each year based on the actual vote, and map the corresponding score metrics in order to identify trends. If successful, the top 5 prediction generated from the statistics could be ranked in the same way.

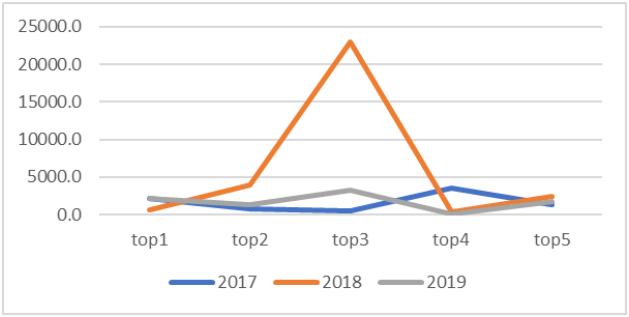

Figure 6. Total voter tweet counts for the top 5 players

The above figure draws out the total tweet counts from voters for the top players each year. There is no clear pattern observed.

4.5 Modeling Results Analysis

Both methods for mapping the tweet metrics yield little trace of the voter's actual MVP vote. Not entirely surprised, we take a further look into the tweets, together with the EDA conducted earlier, and identifies these top reasons behind this.

a. Unable to tell game statistics tweets apart from opinions: Many tweets labeled for players are simply a live statement of the game instead of a subjective comment on their performance. These tweets are still labeled with sentiment regardless and skews the actual sentiment.

b. Excessive amount of contextual words. As the voters are all media professional focused on basketball, there are many specific lingos used by

the group that does not appear as frequently in the tweet training data.

c. Neutral stand point of the voters. Voters might try to avoid showing strong opinion towards a player to avoid issue from fans or the players. They are also likely to cast their votes first based on player performance and then personal preference.

d. Shortage of tweet from some voters. Not all voters are fully fluent with twitter and seldom tweets. A few of the voters even have no twitter account as mentioned earlier.

Considering the above, the voter twitter sentiment analysis should not be used for MVP prediction, at least for now. With more years of tweets and more twitter-active voters in the future, the topic could be revisited and different results might be observed.

5. Result Comparison

As the statistical model has better predicted accuracy compared to the text-based approach, we decided to predict the 2020 MVP players mainly based on the result from the statistical approach. As 4 approaches all give top 2 players prediction as Giannis Antetokounmpo and Lebron James, we finalize our MVP prediction as Giannis Antetokounmpo.

The statistical approach utilizes data from a much longer period of time, which allows more accurate and extended means for prediction compared to the text-based approach.

Regarding the predictors used for the two approaches, the statistical approach mainly takes into account the profile and performance data, which is more objective compared to sentiment scores retrieved in text-based approach.

6. Conclusion and Future Works

In our project, to predict 2019/2020 regular season NBA MVP, we used a statistical approach and a text-based approach to fit player's performance and profile statistics, as well as MVP voter's tweets as text data to perform sentiment analysis. It is found that statistical analysis predicts with higher accuracy in terms of a list of top 3 to 5 players. One crucial reason can be the player statistics are more objective. And the amount of text data available to be trained to obtain sentiment of MVP voters is restricted.

And as for limitations, first, since 2019/2020 season has been suspended due to the COVID-19 outbreak, the ground truth of our work is not known at the moment. Thus, we cannot tell whether we obtain a valuable prediction to NBA associates, fans and coaches. Second, the sentiment model does not differentiate the sentiment if multiple players are mentioned within the same tweet. Although we can label such tweets with multiple player flags, we cannot really distinguish which player can the semantics be assigned to. Third, many voters are not active on Twitter and thus their semantics cannot be captured, critically limiting the prediction accuracy of our text-based approach. Lastly, only three years of Twitter data was used, limiting options for modelling methods used for mapping the sentiment scores to votes. If we have a longer period of tweets, machine learning methods can be applied to fit MVP voter's sentiment on their final vote.

References

Chen, Y., Dai J., Zhang C. A Neural Network Model of the NBA Most Valued Player Selection Prediction. *PRAI '19: Proceedings of the 2019 the International Conference on Pattern Recognition and Artificial Intelligence.* pp.16-20. 2019

Yang, Y. Predicting Regular Season Results of NBA Teams Based on Regression Analysis of Common Basketball Statistics. PhD thesis, Department of Statistics, University of California, Berkeley. 2015