NBA Twitter Correlation – Top Scorers

Andres Fandino, Josh Harkness, Mina Sonbol

**Introduction to the Problem**

Prediction methodologies using machine learning increased drastically in the last few years. Several studies were done where public mood was used to predict economic markets and other important public events such as public office elections. This study identified basketball player performance as an area where twitter feed sentiment analysis could be used as a predictor. Unlike other similar studies,this study will use the players’ tweets to analyze the mood of the tweet and predict the players performance.

Identifying the most relevant parameters in the sentiment analysis will perhaps be the most relevant feature of this problem. This study aims to predict one or more basketball statistics collected for each player. Compared to other team sports, basketball provides a wealth of statistics collected for each player. The study envisions a model where FGM (field goals made) positively correlates with positive sentiment analysis but negatively when the tweet sentiment is negative. This assumption is based on other studies that demonstrated negative sentiment was a better predictor of downward moves in a firm value.

**Importance of the Problem**

This study is important for basketball team organizations and fans. Coaches can use the analysis to support starting lineup decisions and fans can decide whether to attend a game and pay the high-ticket price for a good player performance or stay home and watch the game on the television when the performance is predicted to be poor. In the aggregate these types of predictions can give insight to basketball organizations during contract negotiations and possibly obtain a better ROI on the player. It also furthers the understanding of athletic performance as it relates to the athlete’s mood. Players in team sports can gain insight on upcoming matches and in the case of basketball give input to coaching staff on how many minutes should be played. Lastly, in geographies where sports gambling is legal this tool can be valuable to gamblers in placing better bets.

**Related Work**

The idea for this project came from research done by J. Bollen, H. Mao, and X. Zeng, titled “Twitter Mood Predicts the StockMarket”, and published in the Journal of Computational Science in 2011. The researchers of this study wanted to find if there was a correlation between sentiments collected from large twitter feeds and the stock market as measured by Dow Jones Industrial.

There are two major differences between our project and the aforementioned research. To begin with, there is a considerable difference in data size. While the researchers used 9.7 million tweets collected from 2.7 million users, this study will only focus on the tweets made by the top players in a given season. For example, in the 2019-20 season the top 15 players by total points score had a total of 1500 tweets between them.

Another difference is scope. The researchers focused only on tweets that explicitly mentioned their author’s mood, as this aligned with what they were trying to achieve; to find whether public mood can predict the performance of the stock market. This study includes all tweets made by the top scoring players. The focus is on whether the tweets can shed a light on their mood in the days leading up to any given game and whether that has any discernible effect on their performances.

**Techniques Used**

The study will use several tools to solve the problem. For data collection, the nba-api will collect performance statistics, and the snscrape api to collect tweets.

For sentiment analysis, Microsoft service will be used to get positive, negative, or neutral sentiments and the NRC lexicon will capture other emotions, i.e. fear, anger, etc. These sentiments will be the main features, but the study may look to add other features such as emojis, mentions, and hashtags depending on the remaining time.

Once the dataset is collected, different models will be used to test their predictive power. The target variable will be the total points scored per game above the average for that player. Currently, the specific models are unknown but since this essentially a regression problem, linear regression, decision trees for regression, and regression neural networks are being considered.

**Analysis Evaluation**

The evaluations will reflect the study’s primary focus of how the sentiment of tweets correlates with the points scored by a top player. Each analysis will compare the tweets leading up to a game with the points scored by the player in that game. Simply comparing the scores would not be enough though, because even amongst the best players, each person averages a different point total for each game.

Instead, the study will take each player’s season points scored, obtain the average, and compare it against each individual game. The tweets’ sentiments will then be analyzed alongside the point differential with the average to determine if a correlation exists. The expected result is a positive correlation will exist where positive sentiments will lead to more games with points scored above the average.

If enough time remains, the study will also take shot percentage into account. A player scoring less points may not have to do with them, but rather the team they are up against or the players teammates not getting them the ball. To account for this potential offset, the shot percentages could be used instead of points. The idea for this is even if the player doesn’t get the ball much, if they still make the most out of their shots (shot percentage is at or above average), a low scoring game should not count as a negative result.

**Contribution Breakdown**

The GitHub collaboration tools will be used for this project to allow each developer to work on the code at the same time. The ticket tracker, GitHub Actions, will be used to ensure each developer performs an equal variety of tasks.

Andres Fandino – Code contributor, Team Organizer.

Josh Harkness – Code contributor, Git Engineer, Project Outliner.

Mina Sonbol – Code contributor, Research Field Specialist.

Andres brings the most professional experience to the team and has used it well to help keep meetings directed towards the goal. Josh has the most experience with collaboration tools and his software engineering experience helps him design the project in a flexible way. Mina has the most knowledge in with the NBA, Twitter, and studies done in sentiment analysis. All three developers appear to have differing, but well suited strengths for the study.