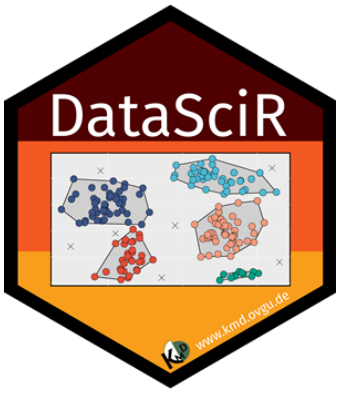


The Impact of NBA player-related Social Media Posts on their on-court Performance - An Analysis

DataSciR - Project Proposal

Frank Dreyer, Kolja Günther, Jannik Greif

20.05.2021

Github Repository

The project is documented here: https://github.com/jannikgreif/DataSciR_2021

Team Member

Name	Course of Studies	Mail
Jannik Greif	M.Sc. Wirtschaftsinformatik	jannik.greif@st.ovgu.de
Kolja Günther	M.Sc. Data and Knowledge Engineering	kolja.guenther@st.ovgu.de
Frank Dreyer	M.Sc. Data and Knowledge Engineering	frank.dreyer@st.ovgu.de

Overview

The project aims to discover a significant impact of social media posts addressed to NBA players before matches with respect to their influence on these players' game performance. For this purpose, we consider NBA players that are highly active on Twitter and extract tweets that are addressed to them within a short period of time before matches via the Twitter API. A sentiment analysis indicates the attitude of the posts and with the resulting sentiment polarity scores we test if there is a correlation between social media posts and players' on-court performance.

Background and Motivation

With the growing presence of social media in all areas of life, allowing people from around the world to react to current events in real-time, an increasingly controversial discussion can be noticed. Today more than ever, public figures are exposed to the reactions of millions of people, observing and commenting on every step in their life that becomes public. The resulting negative impact that extensive social media usage can have on users' behavior and mental state is subject to different scientific studies [1], [2].

Sports athletes, who use social media not only to communicate with peers and fans but also to promote themselves, are no exception to these issues [3]. Among researchers in the sports field, there is a consensus that the mental state of an athlete can have a significant impact on his or her performance [4]. However, only little research has been conducted in order to analyze if and how social media usage of athletes directly influences their performance. Xu and Yu [4] tried to capture the mood of basketball players in the NBA from the tweets they posted just before a match, using sentiment analysis, to analyze how the predicted mood influenced their performance on court. Gruettner, Vitisvorakarn and Wambsganss [5] used a similar approach on tennis players and additionally analyzed the relationship between the number of tweets they posted before matches and their performance within the match. Even though both contributions show that athletes with a bad predicted mood tend to perform worse on-court, they suffer from two limitations:

- The number of tweets an athlete posts per day is rather limited
- The predicted moods are not free of bias since an athlete might only post tweets how he or she wants to be seen on Twitter (also indicated in [5])

Both of these limiting factors may lead to an inaccurate prediction of the mental state of athletes.
We believe that the attitude of social media posts an athlete receives from peers and fans is also a good predictor for his or her performance. Ott and Puymbroeck support this claim [3]. In their article they list cases where athletic performance appeared to be immediately influenced by the media and conclude that the media has the potential to change the performance of an athlete in a negative as well as positive way. In this analysis we aim to assess this relationship more closely by analyzing how social media posts addressed to NBA players affect their game performance.

Project Objectives

This project aims to answer the following research question:
Does the attitude of social media posts addressed to NBA players affect their performance in games?

To answer the specified research question the following objectives can be formulated:

Objective 1: Dataset Creation
Acquire game statistics of NBA players that are highly active on Twitter and the tweets they received from peers and fans in an appropriate time window before games. The game statistics should include an appropriate metric that describes how the player performed within a corresponding game. The tweets need to be preprocessed accordingly to have them in an appropriate format in order to use them for further analysis steps. The attitude of the extracted posts should be captured by assigning a sentiment score to them. The sentiment scores of the tweets a player received in the corresponding time window before a game should be aggregated accordingly and linked to the respective game. As a result, this should end in a data set in which each record contains the game statistics of a player for a specific game and the aggregated sentiment information of the tweets that were addressed to the player before the game.

Objective 2: Exploratory Data Analysis
Analyze the association between the aggregated polarity scores of the tweets a player received before games and the performance of the player within the games using appropriate performance metrics. Additionally, the strength and significance of the correlation should be evaluated.

Objective 3: Presentation
Document the implementation and the results of the analysis in a document. Additionally, the results are summarized in a screencast and a web application.

With this, we can refine our initial research question as follows:
Can we find a correlation between negative/positive Social Media posts related to a specific NBA player and his on-court performance in the following matches?

Datasets

Twitter API/ rtweet:
In order to access the Twitter API, it is a prerequisite to create a Twitter account and apply for a developer account. After the acceptance, it is required to register an app to generate the API keys [6]. To access the Twitter data in R, we are using the package rtweet [7]. Important functionalities are:

Functionality	Explanation
create_token	Creating Twitter authorization token(s)
get_mentions	Get mentions for the authenticating user
get_timeline	Get one or more user timelines (tweets posted by target user(s))
lookup_users	Get Twitter users data for given users (user IDs or screen names)
parse_stream	Converts Twitter stream data (JSON file) into parsed data frame
plain_tweets	Clean up character vector (tweets) to more of a plain text
search_tweets	Get tweets data on statuses identified via search query
stopwordslangs	Twitter stop words in multiple languages data
write_as_csv	Save Twitter data as a comma separated value file

The following example illustrates the different variables that are saved using search_tweets().

```
options(width = 100)
library(rtweet)
names(search_tweets("#NBA", n = 10))
```

```
##  [1] "user_id"                "status_id"              "created_at"
##  [4] "screen_name"            "text"                   "source"
##  [7] "display_text_width"     "reply_to_status_id"     "reply_to_user_id"
## [10] "reply_to_screen_name"   "is_quote"               "is_retweet"
## [13] "favorite_count"         "retweet_count"          "quote_count"
## [16] "reply_count"            "hashtags"               "symbols"
## [19] "urls_url"               "urls_t.co"              "urls_expanded_url"
## [22] "media_url"              "media_t.co"             "media_expanded_url"
## [25] "media_type"             "ext_media_url"          "ext_media_t.co"
## [28] "ext_media_expanded_url" "ext_media_type"         "mentions_user_id"
## [31] "mentions_screen_name"   "lang"                   "quoted_status_id"
## [34] "quoted_text"            "quoted_created_at"      "quoted_source"
## [37] "quoted_favorite_count"  "quoted_retweet_count"   "quoted_user_id"
## [40] "quoted_screen_name"     "quoted_name"            "quoted_followers_count"
## [43] "quoted_friends_count"   "quoted_statuses_count"  "quoted_location"
## [46] "quoted_description"     "quoted_verified"        "retweet_status_id"
## [49] "retweet_text"           "retweet_created_at"     "retweet_source"
## [52] "retweet_favorite_count" "retweet_retweet_count"  "retweet_user_id"
## [55] "retweet_screen_name"    "retweet_name"           "retweet_followers_count"
## [58] "retweet_friends_count"  "retweet_statuses_count" "retweet_location"
## [61] "retweet_description"    "retweet_verified"       "place_url"
## [64] "place_name"             "place_full_name"        "place_type"
## [67] "country"                "country_code"           "geo_coords"
## [70] "coords_coords"          "bbox_coords"            "status_url"
## [73] "name"                   "location"               "description"
## [76] "url"                    "protected"              "followers_count"
## [79] "friends_count"          "listed_count"           "statuses_count"
## [82] "favourites_count"       "account_created_at"     "verified"
## [85] "profile_url"            "profile_expanded_url"   "account_lang"
## [88] "profile_banner_url"     "profile_background_url" "profile_image_url"
```

NBA players on-court performance dataset and Twitter accounts:
basketball-reference.com provides historical basketball statistics about different players and teams in various US-American and European leagues. From this site, we aim to extract game statistics from NBA players that are active on Twitter. Additionally, basketball-reference offers a list of Twitter accounts from various players that we want to incorporate in this process. The statistics of a specific game of a player can be reached by clicking on the players' name (e.g. from the Twitter account list), selecting a season the player was active (regular season or playoffs), and finally selecting a game within this season by clicking on its date. Then a table shows up that shows the player statistics of the selected game. This table provides the following information:

Attribute	Data Type	Description
Starters / Reserves	String	Name of player (separated in starters and reserves)
MP	Timediff	Minutes Played
FG	Int	Field Goals: number made shots (excluding free throws)
FGA	Int	Field Goal Attempts = number of shot attempts (excluding free throws)
FG%	Float	Field Goal Percentage: fraction of field goal attempts (FG/FGA)
3P	Int	3-Point Field Goals: number of made 3-point shots
3PA	Int	3-Point Field Goal Attempts: number of 3-point shot attempts
3P%	Float	3-Point Field Goal Percentage: fraction of three point shot attempts (3P/3PA)
FT	Int	Free Throws: number of free throw shots
FTA	Int	Free Throw Attempts: number of free throw shot attempts
FT%	Float	Free Throw Percentage: fraction of free throw attempts (FT/FTA)
ORB	Int	Offensive Rebounds
DRB	Int	Defensive Rebounds
TRB	Int	Total Rebounds (ORB+TRB)
AST	Int	Assists
STL	Int	Steals
BLK	Int	Blocks
TOV	Int	Turnovers
PF	Int	Personal Fouls
PTS	Int	Points made
+/-	Int	Estimates the players' contribution to the team when the player is on the court

Design Overview

To answer the specified research question we will use the statistical programming language R including the following packages:

- tidyverse (includes packages like ggplot2 for visualization, dplyr for data manipulation, etc.) [8]
- rvest [9]
- rtweet [7]
- tidytext [10]
- Shiny [11]

We plan to address the project objectives as follows:

Objective 1: Dataset Creation

To extract the tweets we will use rtweet which provides access to the Twitter API. To extract the game statistics of NBA players from basketball-reference.com we will make use of the web scraping package rvest. Since the measured performance of a player is rather unreliable, if he only gets a small amount of playing time we will only consider players that on average get at least 20 minutes of playing time. The performance

of players is rather unstable over many seasons. To address this issue we will only consider games in the span of two seasons for a player. To make sure that the players selected for our analysis actually read tweets they receive from peers and fans we will analyze the number of daily received tweets to which they react (e.g. received tweets they like/dislike, answer, etc.) and select a subset of the top twitter-active NBA players according to this measure. Since the NBA prohibits social media usage of players and coaches beginning 45 minutes before tip-off [12], we will consider all tweets that a player receives before that time in a 12 hour window to be able to analyze the immediate effects of the tweets a player receives on the same day before a match on his game performance. To preprocess the tweets in order to bring them into a term-document representation (tokenization, stemming, stopword removal, etc.) we will use tidytext. tidytext provides a set of sentiment lexicons which we will use to assign sentiments to tweets in the following way:

AFINN from Finn Arup Nielsen [13] provides numerical sentiment scores for a list of words in a range between -5 for strongly negative and +5 for strongly positive statements. We will assign the average AFINN-sentiment score of the terms a respective tweet contains as its sentiment score. NRC from Saif Mohammad and Peter Turney [14] provides categorical sentiments (positive, negative) as well as binary association to emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust) to a set of words. We will use the emotions and assign the proportion of terms that correspond to each emotion to a tweet.

The sentiment scores and emotion proportions of the tweets a player received in the 12-hour window before a game will then be averaged and linked to the respective game statistics of the player which builds our dataset for the analysis.

Objective 2: Exploratory Data Analysis

For data visualization and to manipulate the data frame with respect to the analysis we will use ggplot2 and dplyr, respectively (both included in tidyverse) To measure the performance of a player in a game we will use his Plus/Minus score (see above) since it considers the overall contribution of the player to the team when he is on-court (also includes his defensive effort and generally the contribution to his team to score points). To assess the relationship between the aggregated sentiment score of tweets a player received on game days and his performance in the games (indicated by the Plus/Minus score) we will perform a regression analysis. In case there is no clear linear relationship we may also consider the Predictive Power Score [15] to assess if there is a non-linear relationship between the two variables. We may also distinguish between game days where players received highly positive tweets (e.g. averaged sentiment score above 4) and game days where they received highly negative tweets (e.g. averaged sentiment score below -4) to test if there is a significant difference in game-day performance between these two groups (e.g. by using a two-tailed paired-samples t-test).

Objective 3: Presentation

The final project documentation will be written as RMarkdown notebook containing all previously done work and our results. The plots that we created with ggplot2 and dplyr get integrated into the file to underpin our findings. A project website in which we present our main results as general overview is planned to be written most likely using blogdown combined with an integration of shiny apps, which provides us a good method to embed our results into the web presence. With OBS Studio as a screen capturing tool and DaVinci Resolve as state-of-the-art video editing software, we plan to record the slideshow that tells a 2-minute summary of the previously mentioned RMarkdown notebook. The project presentation will then be designed with Microsoft PowerPoint or Prezi which is a powerful tool to create interactive and exciting slideshows.

Time Plan

Phase 1: Literature Research about related works, methods and approaches/ Initial project setup

Finding relevant literature of related works is significant to identify, whether and to which extent research was already done that is close to our project. On the one hand, this helps to avoid doing research that was already done, on the other hand, we can use the perceptions of this literature as foundation for our work. Therefore one scope of this phase is to identify related literature and refine our project based on these

findings. Furthermore, we then can set up the methods and approaches to use for the coming phases like methods for a well functioning data preparation pipeline, well-suited data analysis approaches or evaluation criteria.

Steps:

- Finding a research subject that is both, novel and not yet too heavily exploited by other research groups
- Review of related works, methods and approaches that fit our subject and can provide background information, additional insights and useful findings as well as help to sharpen the initial idea of the project
- Retrieval of relevant data that is needed to investigate our research question and checking for suitability (relevant features, data types, how much transformation is needed)
- Formulation of the project proposal

Milestones/Results:

- Project Proposal Submission
- Project Proposal Feedback

Phase 2: Data Cleaning, Transformation and Integration
Before diving deep into the main task of exploratory data analysis, the relevant data first needs to be extracted, cleaned, transformed and integrated. The main challenge in this phase will be to identify relevant Twitter posts, to prepare them for sentiment analysis and finally extract the attitude of each post. Additionally, the NBA-dataset has to be transformed and integrated to fit with the sentiment polarity scores created beforehand.

Steps:

- Identification, extraction and cleaning (tokenization, stopword removal, stemming/lemmatization) of relevant Twitter posts via Twitter API
- Sentiment analysis of the extracted posts to generate polarity scores
- Cleaning and transformation of the NBA dataset
- Integration of the preprocessed data sets

Milestones/Results:

- Final dataset for analysis

Phase 3: Exploratory Data Analysis
During the main phase of our project, the analysis for possibly existing positive/negative correlations between the extracted Twitter posts and the NBA player performance data is to be done. At first, a descriptive analysis, containing frequencies of variables, min. and max. values, means, etc., shall give insights into the dataset. Then, the data needs to be assessed for relevant variables, expressing the behaviors we want to observe and based on this, the actual analysis can be done. During the whole process, the results of our research get plotted in expressive figures which then find their way into the final project presentation, website and notebook.

Steps:

- Descriptive data analysis: Gaining insights into the datasets by frequency analysis, distributions, means, single feature comparison, etc.
- Choice of meaningful predictor and outcome variables that hold the information needed for our correlation analysis
- Identification of possible correlations based on these features
- Creation of meaningful plots to underpin our findings

Milestones/Results:

- Final Correlation analysis

Phase 4: Evaluation of the Results
To test whether our initial research question holds true, hypothesis testing needs to be done. For this purpose, the correlations found in the phase before need to be checked for significance by using one of several possible approaches. Assessing the proper significance test among Chi-Square test, Regression, T-test, ANOVA test, etc. (to just mention some) is one of the challenges which have to be tackled in this phase of our project.

Steps:

- Testing if and which of the hypotheses that we proclaimed beforehand holds, based on our correlation analysis results
- Assessing a proper significance test for the given variables
- Statistical testing of the results' significance

Milestones/Results:

- Final research results, suitable for expressive documentation & presentation of the project

Phase 5: Paperwork and Finalization
Finally, our research results need to be put in the right form. A Rmarkdown project notebook will contain all relevant information about the project, including an overview and motivation, related work, a detailed description of the data used, all relevant steps of the exploratory data analysis and a final justification of and personal feedback to the project. Additionally a project website and short screencast will wrap up the main findings in an interactive fashion to allow a quick glimpse into our work.

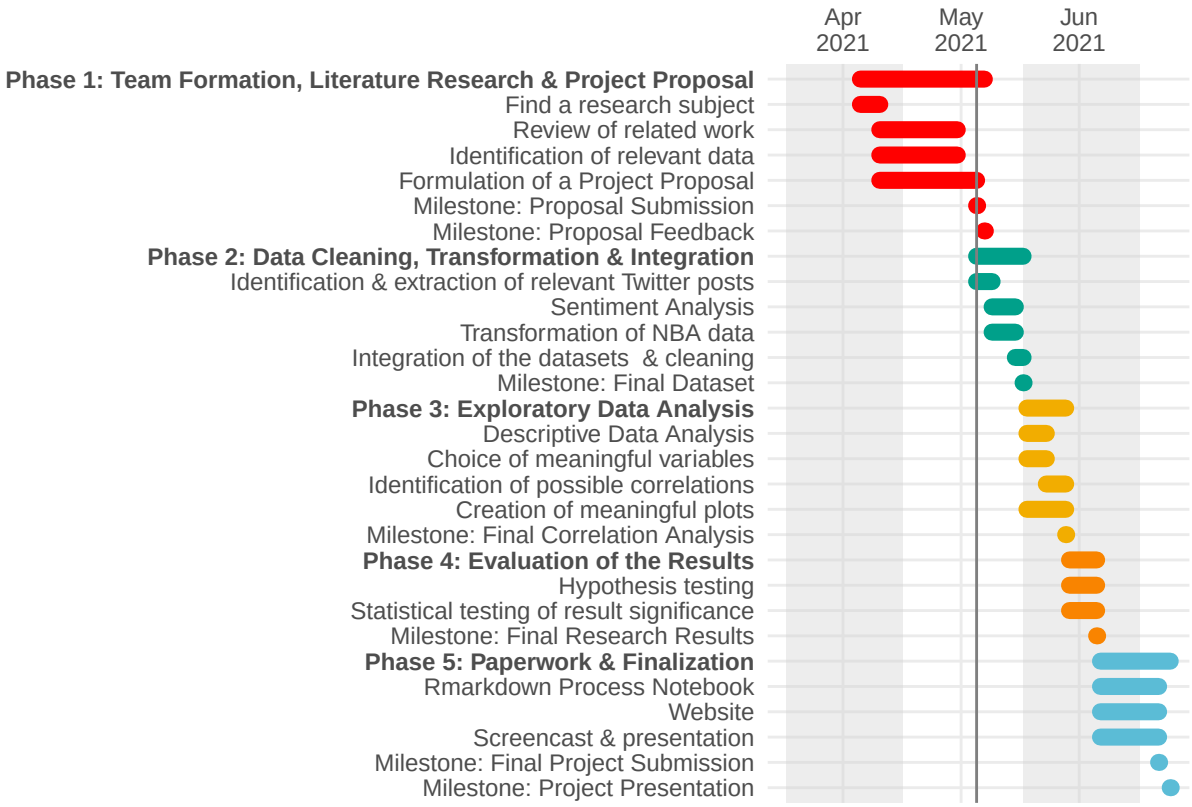
Steps:

- Creation of a Rmarkdown process notebook as documentation of the whole project process, beginning from the general idea, up to the discussion of final results and personal feedback about the overall workflow
- Setting up a project website that represents our research results in an interactive fashion
- Generation of project screencast & presentation which tell a story about our project, underlined with meaningful plots and explanatory results

Milestones/Results:

- Final project submission
- Project presentation

Gantt Diagram



Responsibilities

Task	Responsibility
Find a research subject	Team
Review of related work	Team
Identification of relevant data	Team
Formulation of a Project Proposal	Team
Identification & extraction of relevant Twitter posts	Jannik
Sentiment Analysis	Jannik, Kolja
Transformation of NBA data	Frank
Integration of the datasets & cleaning	Kolja
Descriptive Data Analysis	Team
Choice of meaningful variables	Team
Identification of possible correlations	Team
Creation of meaningful plots	Team
Hypothesis testing	Team
Statistical testing of result significance	Team
Rmarkdown Process Notebook	Frank
Website	Jannik
Screencast & presentation	Kolja

Literature

[1] K. K. Kapoor, K. Tamilmani, N. P. Rana, P. Patil, Y. K. Dwivedi, and S. Nerur, "Advances in social media research: Past, present and future," *Inf Syst Front*, vol. 20, no. 3, pp. 531–558, Jun. 2018, doi: 10.1007/s10796-017-9810-y.

[2] C. Berryman, C. J. Ferguson, and C. Negy, "Social media use and mental health among young adults," *Psychiatr Q*, vol. 89, no. 2, pp. 307–314, Jun. 2018, doi: 10.1007/s11126-017-9535-6.

[3] U. S. S. Academy, "Does the media impact athletic performance? The sport journal," Mar. 14, 2008. https://thesportjournal.org/article/does-the-media-impact-athletic-performance/ (accessed May 16, 2021).

[4] C. Xu and Y. Yu, "Measuring NBA players' mood by mining athlete-generated content," in *2015 48th hawaii international conference on system sciences*, Jan. 2015, pp. 1706–1713, doi: 10.1109/HICSS.2015.205.

[5] A. Grüttner, M. Vitisvorakarn, T. Wambsganß, R. Rietsche, and A. Back, "The new window to athletes' soul - what social media tells us about athletes' performances," Jan. 2020, doi: 10.24251/HICSS.2020.303.

[6] "Twitter API documentation." https://developer.twitter.com/en/docs/twitter-api (accessed May 16, 2021).

[7] M. W. Kearney, "Collecting twitter data [r package rtweet version 0.7.0]," Jan. 08, 2020. https://CRAN.R-project.org/package=rtweet (accessed May 19, 2021).

[8] H. Wickham *et al.*, "Welcome to the tidyverse," *JOSS*, vol. 4, no. 43, p. 1686, Nov. 2019, doi: 10.21105/joss.01686.

[9] "Easily harvest (scrape) web pages." https://rvest.tidyverse.org/ (accessed May 19, 2021).

[10] J. Silge and D. Robinson, "Tidytext: Text mining and analysis using tidy data principles in r," *JOSS*, vol. 1, no. 3, 2016, doi: 10.21105/joss.00037.

[11] "Shiny." https://shiny.rstudio.com/ (accessed May 19, 2021).

[12] "NBA dunks tweeting, social media during games. ESPN.com," Sep. 30, 2009. https://www.espn.com/nba/news/story?id=4520907 (accessed May 19, 2021).

[13] F. Å. Nielsen, "AFINN." Informatics; Mathematical Modelling, Technical University of Denmark, Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, Mar. 2011, [Online]. Available: http://www2.compute.dtu.dk/pubdb/pubs/6010-full.html.

[14] "NRC emotion lexicon." http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm (accessed May 20, 2021).

[15] F. Wetschoreck, "RIP correlation. Introducing the predictive power score. Medium," May 04, 2020. https://towardsdatascience.com/rip-correlation-introducing-the-predictive-power-score-3d90808b9598 (accessed May 20, 2021).