Data Article

Dataset of two decades of Tiger Woods press conferences and tournament performance

David Pastoriza*, Thierry Warin

HEC Montreal, Canada

A R T I C L E   I N F O

Article history:
Received 25 November 2021
Revised 19 January 2022
Accepted 9 February 2022
Available online 15 February 2022

Keywords:
Tiger Woods
Press conferences
Sentiment analysis
Natural language processing
Machine learning

A B S T R A C T

This data article describes a dataset that allows exploring the determinants of superstars’ sentiment in tournaments. It consists of 1,284 press conferences of Tiger Woods in the PGA Tour between 1996 and 2020. We used natural language processing, a form of artificial intelligence, to extract and encode in a quantitative form the sentiment in Tiger Woods press conferences both before the tournament and after the rounds played. Additionally, the dataset provides a series of variables that describe Tiger Woods’ scoring and performance momentum in each round and variables that describe health-related and off-the-course issues that could affect his performance on the course. This data can be useful to understand the sentiment that superstars go through before important tournaments, their sentiment following a major victory or defeat, how that sentiment evolves throughout their athletic career, and how sentiment is associated with performance momentum.

© 2022 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

* Corresponding author.
E-mail address: david.pastoriza@hec.ca (D. Pastoriza).

https://doi.org/10.1016/j.dib.2022.107955
2352-3409/© 2022 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
### Value of the Data

- The phenomenon of superstars, wherein a reduced number of individuals earn enormous amounts of money and dominate their field, has long attracted the interest of scholars. Thus far, research has primarily focused on sports superstars’ effects on their rivals, teammates, spectators, or companies whose products they endorse. This contrasts with the lack of scholarly work to understand the superstar athletes themselves, a situation partly caused by the small size of superstar athletes available, which has prevented scholars from using complex statistical approaches.

- This dataset details the press conferences of Tiger Woods in PGA Tour tournaments he entered throughout his career. The content of the press conference transcripts has been analyzed with natural language processing methods [1], allowing to extract the sentiment from his press conferences before the tournament and after the rounds he played. The dataset also includes highly detailed performance variables for every round he played.

- The dataset is useful for researchers who want to understand the sentiment that superstars go through before important tournaments, their sentiment following a major victory or defeat, or how that sentiment evolves throughout their athletic career. Also, since superstars are under constant scrutiny and face extra pressures (vis-à-vis non-superstar players), the dataset can be useful for researchers willing to understand the sentiment that superstars go through when they face off-the-course problems (i.e., health-related or personal issues).

- The dataset can also be used to examine how sentiment is associated with performance momentum, a concept also referred to in the literature as “hot hand” [3]. Thus far, researchers have primarily proxied positive (negative) momentum as positive (negative) streaks of results. Therefore, the sentiment variable provided in this dataset may allow for a more nuanced understanding of performance momentum.

- This dataset is useful for researchers who want to understand how superstars react to competing against their direct rivals [2] – i.e., competitors with whom the superstar has a subjective competitive relationship based on past competitive interactions that have increased the psychological stakes of competition.

### 1. Data Description

The phenomenon of superstars, wherein a reduced number of individuals earn enormous amounts of money and dominate their field, has long attracted the interest of scholars [4]. Sports superstar Tiger Woods has been particularly appealing for scholars from various disciplines to examine essential questions. In economics, Brown showed that the presence of Tiger Woods in a
tournament was associated with a reduced effort from his competitors [5]. In marketing, Chung et al. showed evidence of the positive impact of Tiger Woods endorsement on sales of Nike golf balls [6]. In finance, Knittel and Stango showed that Tiger Woods’ off-the-course issues had an adverse effect on the market value of the companies he endorses [7].

Despite the progress made to understand the phenomenon of sports superstars, thus far, research has primarily focused on the latter’s effects on its rivals, teammates, spectators, or companies whose products they endorse. This contrasts with the lack of scholarly work to understand the superstar athletes themselves, a situation that is partly caused by the methodological constraints associated with studying such a small population of eminent elite athletes [8] – i.e., small sample sizes have prevented the use of complex statistical approaches.

Tiger Woods is not the only superstar in the PGA Tour. There are three important reasons why this database focuses on Tiger Woods: First, no other superstar has gone through the media scrutiny that he went through during his entire career. As a result, for over a twenty-five-year period there is reliable information on his weekly health condition (i.e., minor and major injuries) and his weekly off-the-course personal issues. This allows for an in-depth examination of the factors that drive a superstar sentiment beyond on-the-course competitive dynamics, to include off-the-course issues, an important issue that has not been examined yet. Second, contrary to other PGA Tour players (including superstars), who generally attend the press conference only once the tournament started, Tiger Woods has always been requested to the press conference prior to the beginning of the tournaments in which participated. This allows for an examination of how his pre-tournament sentiment may influence his tournament performance. Third, the stardom of Tiger Woods is unparallel in golf – amongst the active players, he has been #1 ranked player in the Official World Golf Ranking for 683 weeks, very far from his follower, Dustin Johnson, with 135 weeks. Thus, such an extensive period of celebrity is extremely rare, allowing for interesting sentiment analysis before, during, and after celebrity.

The file format is RDS, which can be directly used in R. People using other languages will convert it easily in their preferred language with the appropriate translator.

In order to compute the sentiment score, the protocol was the following: First, we coded an R script to collect the 1065 speeches, using the rvest package. Then, we converted the speeches into a document-term matrix. This stage transformed the unstructured data into a structured dataframe. We organized the dataframe by sentences. The document-term matrix allowed us to do some necessary data cleaning, such as separating the reporter’s questions from Tiger Woods’ answers. We could also clean the data by removing the stopwords (articles, punctuation, etc.) and text formatting. This stage was necessary to homogenize the dataframe before running the sentiment analysis. We then created vectors of words for each speech. Using the bag-of-words approach in natural language processing, we then computed the positive and negative scores using Bing’s sentiment lexicon and techniques [9]. This approach has some well-known limitations (space and time complexity, and context), which may not apply in this very specific context (i.e., golf tournaments). It comes anyway with important benefits such as ease of use and, most importantly, computation speed. We could have used other techniques such as Latent Dirichlet Allocation, topic modeling or transformers for instance. The bag-of-words’ benefits outweighed the costs in our empirical setting. The sentiment score was computed by the subtraction between the positive and negative scores.

The database not only details the sentiment (i.e., the difference between positive and negative words) in Tiger Woods’ speech after the round but also those before the tournament. Additionally, it provides a series of variables that describe Tiger Woods’ scoring and momentum [10] in each round and variables that describe health-related and off-the-course issues that could affect his performance on the course. The following table describes the list of variables in the dataset:

Fig. 1 below represents the sentiment score of Tiger Woods speeches from 1996 until 2020. We compute the sentiment scores just on Tiger Woods’ answers during the interviews, regardless of the questions (be it with a positive or negative tone) from the journalists [11]. Output is detailed by year and by tournament.
The dataset contains Tiger Woods’ speeches and scoring on PGA Tour tournaments, the professional circuit in which Tiger Woods developed his career. Tiger Woods’ participation varied over the years; on a first stage that goes from 1996 until 2007, his tournament participations were frequent, while between 2008 and 2020, his tournament participations were less frequent, both because of injuries and off-the-course personal issues. Even though it is only a handful of PGA Tour players who are invited into the press room before the beginning of a tournament (generally restricted to former champions) and after each round (usually restricted to the provisional leader(s) and runner(s)-up), Tiger woods press conferences were very frequent since he was the star in the PGA Tour for the period covered in the database.

The fact that the number of Tiger Wood press conferences in our dataset increases over the years is explained by at least two factors. First, even though Tiger Woods was already considered by many as a high potential player in his first PGA Tour tournaments in the 1990s, its stardom took off in the early 2000s, when he became a social phenomenon and his mere presence in a tournament would increase fan and press attendance alike, thus making compulsory his presence in the press room. Second, he participated in a few minor tournaments (e.g., Bell Canadian Open, Walt Disney Classic) at the beginning of his career with less coverage.

In Table 2 below, we can see the descriptive statistics of the variables in the dataset.

2. Experimental Design, Materials and Methods

Part of the data was obtained from two databases. First, we use the ShotLink® database, obtained directly from the PGA TOUR, which provides detailed round-level information on players’ scoring in every tournament. This database allows to trace Tiger Woods’ tournament on every round of every tournament he played in the PGA TOUR (i.e., birdies, bogeys, driving distance, driving accuracy, scrambling, etc.). Importantly, based on the round-level results of this database, we coded and created several variables that reflect Tiger's momentum in the tournament (i.e.,
Table 1  
Description of the variables in the dataset.

| Variable                                      | Type       | Description                                                                 |
|-----------------------------------------------|------------|-----------------------------------------------------------------------------|
| Tournament_Year                               | Numeric    | Yearly season                                                               |
| Tournament_Order                             | Ordinal    | Chronological order of tournaments within a season (i.e., smaller numbers took place earlier in the season) |
| Permanent_Tournament_Number                  | Numeric    | Identification number that is unique to that tournament, regardless of the sponsor/name of the tournament that may change over the years |
| Course_Number                                | Numeric    | Course number that does not change over time (i.e., different tournaments may be played on the same course) |
| Player_Number                                | Numeric    | Player identification number. Does not change over the years                |
| Round_Number                                 | Numeric    | Round number. PGA Tour tournaments generally have four rounds               |
| Event_Name                                   | Numeric    | Name of the tournament                                                      |
| Course_Name                                  | Text       | Name of the course in which the tournament took place                       |
| Interview_Text                               | Text       | The integral text of the interview                                           |
| Number_Of_Answers                            | Numeric    | Number of answers provided by Tiger Woods during the Q&A section            |
| Link                                         | Text       | Link to the original document (redirecting on ASAP Transcription website)   |
| Response_Negative                            | Numeric    | The negative score computed on Tiger Woods’ responses                       |
| Response_Positive                            | Numeric    | The positive score computed on Tiger Woods’ responses                       |
| Response_Sentiment                           | Numeric    | The subtraction between the positive and the negative scores computed on Tiger Woods’ responses |
| Round_Score                                  | Numeric    | Number of strokes of the round                                              |
| End_of_Round_Pos_numeric_                    | Ordinal    | Player’s rank in the round (i.e., 1 means he is leading the tournament, 2 means he is the runner-up, etc.) |
| Total_Holes_Over_Par                         | Numeric    | Number of holes in the round in which the player scored bogey or worse in the round |
| Birdies                                      | Numeric    | Number of birdies in the round                                              |
| Birdies_Rank                                 | Ordinal    | Players’ rank in number of birdies in the round (i.e., 1 means he was the player with the highest number of birdies in the round) |
| Bogey_Avoidance_Rank                         | Ordinal    | Player’s rank in terms of the number of holes in which the player saved a situation of bogey in the round |
| Driving_Distance_Rank                        | Ordinal    | Player’s rank in driving distance in the round                              |
| Driving_Accuracy_Rank                        | Ordinal    | Player’s rank in driving accuracy in the round                              |
| GIR_Rank                                     | Ordinal    | Player’s rank in number of greens in regulation in the round               |
| Scrambling_Rank                              | Ordinal    | Player’s rank in scrambling in the round (i.e., ability to recover from difficult situations) |
| Distance_to_leader_strokes                   | Numeric    | Distance in strokes to the interim leader at the end of the round (i.e., if Tiger is trailing by two shots, it takes value 2; if Tiger is leading, the variable takes value 0) |
| Distance_to_leader_ranks                     | Numeric    | Distance in ranks to the interim leader at the end of the round (i.e., if Tiger is in rank 3, it takes value 2; if Tiger is leading, it takes value 0) |
| Ranks_gained                                 | Numeric    | Equal to the rank at the end of Round_n minus the rank at the end of Round_{n-1}. For instance, if Tiger had a position 3 at the end of Round_n and 5 at the end of Round_{n-1}, the variable’s value is −2. It takes missing values for round 1 |
| Strokes_gained_v_a_v_Leader                  | Numeric    | Equal to distance to the leader (in strokes) at the end of Round_n minus distance to the leader (in strokes) at the end of Round_{n-1}. It takes missing values for round 1 |
| Distance_to_runner_up_strokes                | Numeric    | Distance in strokes to the interim runner-up at the end of the round. For instance, if Tiger is leading by two shots, the variable should take value 2; if Tiger is co-leading, the variable should take value 0; if Tiger is neither leading nor co-leading, it takes missing value. |

(continued on next page)
## Table 1 (continued)

| Variable                        | Type       | Description                                                                                                                                 |
|---------------------------------|------------|---------------------------------------------------------------------------------------------------------------------------------------------|
| Strokes_gained_v_a_v_Runner_up  | Numeric    | Distance in strokes to runner-up at the end of Round $n$, minus distance to runner-up at the end of Round $n-1$. Note that this variable should have a missing value for observations of round 1. Note that this variable should have a missing value when Tiger is not leading. |
| Minor_injury                    | Categorical [1; 0] | Tiger had a minor injury when he entered the event                                                                                         |
| Major_injury_surgery            | Categorical [1; 0] | Tiger Woods had a major injury when he entered the event                                                                                   |
| Personal_issues                 | Categorical [1; 0] | Tiger Woods had personal issues when he entered the event                                                                                   |
| Major                            | Categorical [1; 0] | Takes value 1 if tournament is a major (i.e., prestigious)                                                                                   |
| Prize_Money                     | Numeric    | Prize money of the tournament                                                                                                               |
| SoF                             | Numeric    | Strength of the field of players in the tournament (i.e., the higher the number, the more competitive is the field)                          |
| OWGR                            | Numeric    | OWGR of Tiger at the moment of the observation                                                                                               |

1 In golf, par is the predetermined number of strokes that a proficient golfer should require to complete a hole. Bogey is a score of one stroke more than par. Birdie is a score of one stroke fewer than par. For instance, if in a par-4 hole a player scored 3, he birdied the hole; if he scored 4, he made par in the hole; if he scored 5, he bogeyed the hole.

2 A drive is the long-distance shot intended to move the ball a great distance down the fairway towards the green, in which the target hole is located.

3 The green is the area of specially prepared grass around the hole. A player hits a green “in regulation” if the ball is on the surface of the green and the number of strokes taken is at least two fewer than par (i.e., second stroke in a par-4 hole).

4 When a player misses the green in regulation, but still makes par or birdie on a hole.

...distance in strokes to the leader, distance in strokes to runner up, ranks gained, etc.). Additionally, from this database, we retrieved or computed the season in which the observation occurs, the chronological order in which the tournament took place within the season, and the tournament prize money. The second database was that of the Official World Golf Ranking database, which measures players’ ability. Additionally, based on the Official World Golf Ranking, we were able to compute the strength of the field (i.e., aggregated level of ability of the players who entered the tournament).

The data on health conditions and personal issues off the course were manually retrieved from the magazine Golf Digest and PGA TOUR’s Media Guides, which are booklets that the PGA Tour produces for the media containing players’ biographic information updated every season.

For those researchers who are relatively unfamiliar with golf, it is important to note that PGA Tour golfers are independent contractors who can decide whether they enter in a tournament (i.e., players may abstain due to fatigue or because the tournament prize is not large enough, amongst other reasons). On this respect, Tiger Woods is known to enter fewer tournaments than other players and to systematically chose the same tournament/courses. The researchers may want to consider these elements in their econometric model specifications.

### Ethics Statement

The present study did not involve any experiment using human subjects or animals. Thanks to the contractual agreement with the PGA Tour, the latter allows the authors to publish data computed from the ShotLink Database if the publication is in an academic, peer-reviewed journal.
Table 2
Descriptive statistics of the variables.

| Variable                        | Mean  | S.D.  | Min  | Max  |
|---------------------------------|-------|-------|------|------|
| Round_Score                     | 69.57 | 3.18  | 61   | 85   |
| End_of_Round_Pos_numeric_       | 21.19 | 26.24 | 1    | 152  |
| Total_Holes_Over_Par\(^1\)     | 2.53  | 1.67  | 0    | 9    |
| Birdies                         | 4.26  | 1.79  | 1    | 10   |
| Birdies_Rank                    | 25.37 | 27.18 | 1    | 145  |
| Bogey_Avoidance_Rank            | 25.86 | 27.88 | 1    | 152  |
| Driving_Distance_Rank\(^2\)     | 20.48 | 26.62 | 1    | 153  |
| Driving_Accuracy_Rank           | 43.08 | 34.56 | 1    | 175  |
| Response_Positive               | 13.6  | 9.85  | 0    | 75   |
| Response_Negative               | 4.364 | 3.77  | 0    | 24   |
| Response_Sentiment              | 9.235 | 8.42  | –9   | 60   |
| GIR_Rank\(^3\)                  | 28.96 | 29.64 | 1    | 147  |
| Scrambling_Rank\(^4\)           | 36.56 | 34.28 | 1    | 172  |
| Distance_to_leader_strokes      | 5.61  | 4.67  | 0    | 30   |
| Distance_to_leader_ranks        | 20.19 | 26.24 | 0    | 151  |
| Ranks_gained                    | –4.06 | 16.74 | –90  | 65   |
| Strokes_gained_v_a_v_Leader     | 0.69  | 3.17  | –9   | 16   |
| Distance_to_runner_up_strokes_  | 2.35  | 2.62  | 0    | 15   |
| Strokes_gained_v_a_v_Runner_up  | 1.11  | 2.30  | –4   | 7    |
| Minor_Injury                    | 0.04  | 0.19  | 0    | 1    |
| Major_Injury_Surgery            | 0.02  | 0.16  | 0    | 1    |
| Personal_Issues                 | 0.03  | 0.17  | 0    | 1    |
| Major                           | 0.24  | 0.42  | 0    | 1    |
| Prize_Money\(^0\)               | 5.69  | 2.46  | 1.0  | 12.7 |
| SoF                             | 552   | 188   | 69   | 849  |
| OWGR                            | 33    | 120   | 1    | 1199 |

\(^0\) Expressed in millions of dollars.
\(^1\) In golf, par is the predetermined number of strokes that a proficient golfer should require to complete a hole. Bogey is a score of one stroke more than par. Birdie is a score of one stroke fewer than par. For instance, if in a par-4 hole a player scored 3, he birdied the hole; if he scored 4, he made par in the hole; if he scored 5, he bogeyed the hole.
\(^2\) A drive is the long-distance shot intended to move the ball a great distance down the fairway towards the green, in which the target hole is located.
\(^3\) The green is the area of specially prepared grass around the hole. A player hits a green “in regulation” if the ball is on the surface of the green and the number of strokes taken is at least two fewer than par (i.e., second stroke in a par-4 hole).
\(^4\) When a player misses the green in regulation, but still makes par or birdie on a hole.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have or could be perceived to have influenced the work reported in this article.

CRediT Author Statement

**David Pastoriza:** Conceptualization, Data curation, Writing – original draft; **Thierry Warin:** Conceptualization, Writing – original draft, Methodology, Software.

Acknowledgments

We thank the PGA Tour for providing us access to ShotLink data. We are grateful to David Emond (Delta Statistique) for his coding.
References

[1] M. Gentzkow, B. Kelly, M. Taddy, A.A. Shaibu, Text as Data, J. Econ. Lit. (3) (2019) 535–574.
[2] G.J. Kilduff, H.A. Elfenbein, B.M. Staw, The psychology of rivalry: a relationally dependent analysis of competition, Acad. Manag. J. 53 (2010) 943–969.
[3] M. Bar-Eli, S. Avugos, M. Raab, Twenty years of “hot hand” research: review and critique, Psychol. Sport Exerc. 7 (6) (2006) 525–553.
[4] S. Rosen, The Economics of superstars, Am. Econ. Rev. 71 (1981) 845–859.
[5] J. Brown, Quitters never win: the (adverse) incentive effects of competing with superstars, J. Polit. Econ. 119 (2011) 982–1013.
[6] K.Y.C. Chung, T.P. Derdenger, K. Srinivasan, Economic value of celebrity endorsements: Tiger Woods’ impact on sales of Nike golf balls, Mark. Sci. 32 (2013) 271–293.
[7] C.R. Knittel, V. Stango, Celebrity endorsements, firm value, and reputation risk: evidence from the Tiger Woods scandal, Manag. Sci. 60 (2014) 21–37.
[8] J. Baker, J. Schorer, S. Lemez, N. Wattie, Understanding high achievement: the case for eminence, Front. Psychol. 10 (2019) 1–9.
[9] L. Bing, in: Sentiment Analysis, 2nd ed., Cambridge University Press, 2015, p. 448. pages.
[10] D. Pastoriza, I. Alegre, M. Canela, Conditioning the effect of prize on tournament self-selection, J. Econ. Psychol. 86 (2021) 1–19.
[11] W. Sanger, T. Warin, European central bank’s monetary policy decisions: a dataset of two decades of press conferences, Data Brief (2018) 794–798, doi:10.1016/j.dib.2018.08.061.