Data-Driven Prediction of Athletes' Performance based on their Social Media Presence

Frank Dreyer, Jannik Greif, Kolja Günther, Myra Spiliopoulou, and Uli Niemann

Faculty of Computer Science, Otto von Guericke University, Magdeburg, Germany
{frank.dreyer, jannik.greif, kolja.guenther}@st.ovgu.de,
{myra, uli.niemann}@ovgu.de

Abstract. It is well known in the sports industry that the performance of athletes is strongly influenced by physiological and psychological factors. In recent years, many researchers have analysed whether athlete-generated social media content can be used as proxies for such performance factors, with some promising results. In this study, we investigated whether such proxies are useful features for a machine learning model to predict athletes' performance in subsequent competitions. We extracted millions of tweets that NBA basketball players posted themselves or were tagged in and derived features reflecting players' mood, social media behaviour, and sleep quality before games. Using these and other non-social media-related features, we performed statistical tests to examine whether the features significantly improve the accuracy of a random forest model for predicting players' BPM scores in upcoming games. The results show that, in particular, the number of tweets a player is tagged in prior to a game significantly improves the predictions of the model. Our findings provide insights for practitioners on the effects of social media on athlete performance that can be used prospectively for mental health awareness training and optimisation of pre-game routines.

Keywords: Machine learning · Athletic performance · Social media · Twitter · Sentiment analysis · Predictive significance

1 Introduction

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.

Athletes, who use social media not only to communicate with peers and fans but also to promote themselves, are no exception to this circumstance. To date there is plenty of anecdotal evidence that the media has the potential to affect the performance of athletes to a great extent. Ott and Puymbroeck [16] describe several cases where the performance of practitioners from different

sports dramatically changed after they were exposed to medial criticism. In their analysis they conclude that there is strong evidence that the media affects the performance of athletes.

The influence social media can have on an individual's mood are confirmed by athletes themselves. In an interview, 8-time NBA all-star Vince Carter explains how it is like to be constantly exposed to social media criticism [5]: *"It's an emotional rollercoaster [...]. We as athletes have social media at our fingertips at any time and of course if you're playing well, you go look at your mentions. If you're not playing well, you go look at your mentions and now you have diehard opponent fans saying whatever they want and sometimes we tend to get caught up in what's being said from these persons [...]."*

Such examples give rise to the question if social media content can be used to predict the performance of athletes in upcoming competitions. In many domains, social media content has become the new source of intelligence. Accordingly, it has been successfully used for a wide range of predictive modelling tasks, such as stock price prediction, election results forecasting and even disease outbreak prediction [18]. A model predicting athletic competition performance based on social media posts could also be of great value for various stakeholders in the sports industry. Consider coaches as an example. Their job is to prepare athletes for upcoming competitions so that they have the best possible chance of winning. This preparation is not only a simple matter of improving athletes' physical skills through training. Instead, and perhaps more importantly, it also involves strengthening athletes' belief in their own abilities and their mental resilience. To achieve the latter, coaches must be aware about all factors that affect athletes' psychological functioning in a positive way. This includes both intrapersonal factors (e.g. self-motivation) as well as interpersonal factors (e.g. social support) [11]. Social media content could be a great resource to uncover the satisfaction of such psychological performance factors. An according performance prediction model could then guide coaches if athletes need further mental or physical training to excel in the competition.

This paper addresses the following research question: *Do features derived from athlete-related social media posts lead to an improvement in accuracy of a machine learning model predicting athletes' performance in subsequent competitions?* To answer this research question, we gathered tweets NBA players posted themselves or were tagged in before games. From these tweets, we distilled various features that reflect athlete interaction on social media from different angles. We considered the quantity of the tweets, their temporal information as well as their sentiment. These features were then used in combination with other social media-unrelated features to study their predictive significance on athletic performance. This involved a permutation test that is based on random forests.

The remainder of this paper is organised as follows: In Section 2 we summarise existing research that analysed the relationship between social media and athletic performance. On this basis we describe our methodology to answer our proposed research question in Section 3. Our findings are presented

and discussed in Section 4 and 5, respectively. Finally, we conclude our paper in Section 6.

2 Related Work

To date there have been few studies that examined the relationship between athletes' social media interaction and their performance. In general researchers approach the topic in one of two ways. They either consider the content of social media posts produced by athletes as a proxy for their mood and behaviour and analyse the effects of this proxy on their performance or they consider social media activity as something that distracts athletes from focusing on performing well. In the following we will elaborate on both perspectives.

2.1 Social Media as a Mood and Behaviour Detection Framework

In the psychology field there is a consensus that an individual's ability to perform a certain task is greatly affected by his or her mood. While a positive mood is often associated with better concentration, motivation, creativity, and cooperation, a negative mood leads to the exact opposite, consuming many attention resources and recovery efforts [22]. Some researchers make use of this mood-performance relationship and consider social media posts as a way how athletes verbalise their feelings. To extract the mood expressed in the posts, sentiment analysis models are used to capture the polarity of a post in a single number. This sentiment score is then related to the performance of athletes in upcoming competitions. Xu and Yu [22] for example use sentiment analysis to capture the pre-game mood of NBA players from tweets they posted before games and show that there is a positive linear relationship between the approximated mood and the adjusted Plus/Minus game performance metric of the players. A similar approach is used by Grüttner et al. [7] who conduct a statistical test to compare the average first serve fault of ATP and WTA tennis athletes achieved between matches where they had a negative vs. a positive pre-match mood. In contrast to Xu and Yu [22] however they do not find a significant difference between the two groups of interest.

Lim et al. [13] go one step further. Backed by an extensive literature review they claim that there is a positive inverted U-shaped relationship between humility and athletic performance. To investigate this hypothesis, they train a linear regression model to predict NFL players Fantasy Football points in upcoming games based on how arrogant or humble the players appear before the game. Similar to the other mentioned researchers before they approximate humility by the social media content the athletes produce before games. Their results strongly suggest that there is indeed an inverted U-shaped relationship between humility and athletic performance.

2.2 Social Media as a Distraction Factor

In the social psychology field, the Distraction-Conflict Theory (DCT) [1] provides a theoretical attempt to explain the causes of impaired performance levels. According to DCT the mere presence of others can provoke an attentional conflict in an individual performing a certain task which in turn leads to elevated drive and probably impaired performance executing the task. Many researchers use DCT to explain the causes of performance drops among athletes by considering social media as a distractor for athletes. Hayes et al. [9] for example apply DCT by performing semi-structured interviews with elite Australian athletes to understand the elements of social media athletes perceive to be distracting during major sport events. The results of these interviews suggest that there are five distracting elements, including obligation to respond, susceptibility to unwanted commentary, pressure to build and maintain an athlete brand as well as competitor content and mood management. Another research by Grüttner et al. [7] use DCT to justify that high social media usage of athletes before a competition negatively impacts their performance in both a cognitive and motoric way. In their study they name two reasons why high social media usage represents a distractor for athletes. Firstly, the time and focus athletes spend on posting messages limit their capacity focusing on the preparation for the next competition. Secondly, the athletes' awareness that other social media users react to their produced content or post messages related to them may trigger internal distractions. To proof this theory, they conduct a statistical test to check if the difference in average first serve fault of tennis athletes between matches where they posted a large number of tweets before the match vs matches where they only posted a small number of tweets before the match is significant. Here, they assume that the quantity of posts an athlete generates is a good measure for his social media activity before a competition. Lim et al. [13] use the same connection between post quantity and social media activity in their regression analysis in the context of NFL stars. Similarly, Watkins et al. [20] use the iPhone screen time function to measure the number of hours college athletes spend on social media apps per week and relate the corresponding on-screen time to their competition performance after adjusting for confounding factors. All studies come to the same conclusion, that there is significant evidence that heavy social media usage hinders athletic performance.

Other researchers link social media activity to poor sleep quality. Watkins et al. [20] for example assign college athletes to moderate, active, or super active social media users based on their iPhone screen time and perform an ANCOVA to compare the difference in sleep quality among the three groups. Their results show that there is a significant difference between the groups and that sleep quality tends to decrease with increasing social media activity. Jones et al. [12] consider late-night tweeting, i.e. tweets posted in the middle of the night, as a proxy for sleep deprivation. In their study they use t-tests to assess how late night tweeting affects various next-day game statistics of NBA players, including shooting percentage, points scored and rebounds. According to their findings it

appears that late night tweeting significantly deteriorates NBA players next-day game performance.

3 Methodology

The studies discussed in the previous section show that social media content can be exploited in various ways to construct features that capture the mindset and well-being of athletes before competitions. In this study we assessed whether such features significantly contribute to the accuracy of a machine learning model predicting athletes' competition performance.

3.1 Data Selection

To answer our initially stated research question we focused our analysis on NBA basketball players. This choice was made since the NBA provides well established and easily accessible performance metrics and the fact that basketball was successfully studied by multiple researchers in the context of social media before [22,12,20]. Furthermore, we gathered social media posts from Twitter, a platform that is extensively used by NBA players to communicate with peers and fans. Figure 1 depicts the inclusion and exclusion criteria for the NBA dataset and the Twitter dataset.

NBA Data: All NBA-related data was collected from basketball-reference.com, a website providing historical basketball statistics from various US American and European leagues. We only considered NBA players that have a Twitter account and gathered various statistics from games they participated in between the seasons 2016-2019. To avoid a potential bias in our analysis results due to the COVID-19 pandemic, we excluded the more recent seasons 2019-2021 from our analysis. Additionally, to account for the effects of long-term injuries we only considered players that obtained playing time in at least 60% of the games each season. Because performance metrics tend to be unreliable when playing time is limited, we only included players who received an average of at least one quarter, i.e. 12 minutes of playing time and excluded all games in which the player was on court for less than 5 minutes. After applying these constraints we had a total of 108 players and 24,876 games in which these players actively participated.

Twitter Data: As social media interaction involves both generating and consuming content, we considered both tweets posted by the players themselves but also tweets produced by others in which the players were mentioned. Particularly, for each game we extracted all player-related tweets that were posted within 24 hours before tip-off. Since NBA league policies prohibit players and coaches from using social media from 45 minutes before a game starts until post-game interviews are completed, we set 45 minutes before tip-off as an upper limit for tweet extraction for each game. Furthermore, we only considered tweets in English and excluded retweets. These constraints resulted in a total of 8,018 tweets players posted by themselves and 1,920,901 tweets players were tagged in.

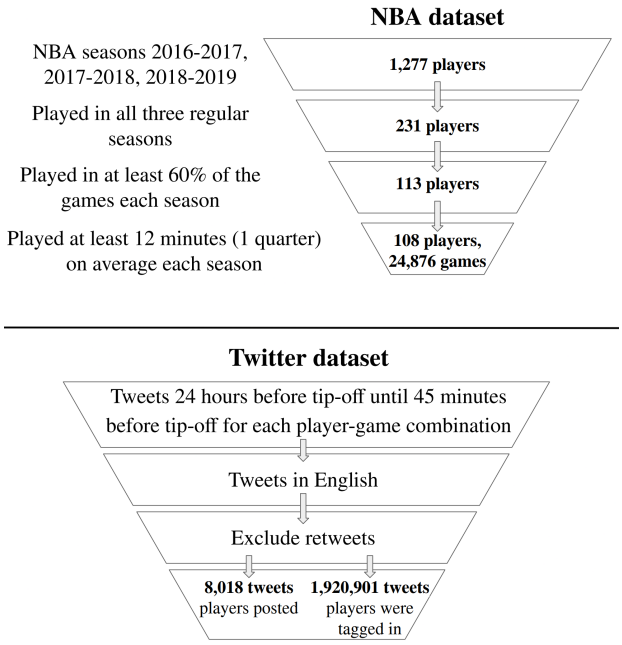

Fig. 1: Inclusion and exclusion criteria for the NBA dataset and the Twitter dataset.

3.2 Data Preparation

After extracting all relevant NBA and Twitter-related data, the next step was to prepare the data for predictive modelling. To that end we first preprocessed the textual information of the tweets to bring them into the desired format for a sentiment analysis model determining the polarity of the tweets. To this end, we used VADER [10], a lexicon and rule-based sentiment analysis model developed for social media texts. The polarity scores of the tweets were then used besides other information from the extracted data to create features for a final dataset ready for predictive modelling.

Tweet Preprocessing: To bring the tweets into the desired shape for the VADER sentiment analysis model we substituted all URLs, mentions and hashtags with placeholders. One thing to note is that VADER only considers empathic uppercasing (e.g. "AMAZING") to capture the sentiment amplification of words [10]. As word elongations (e.g. "amaaaaazing") as well as gaps between characters of a word (e.g. "D O P E") also intensify word sentiments [6] we decided to correct and uppercase such words in order for VADER to correctly identify the shift in sentiment intensity.

Tweet Sentiment Analysis: As the effectiveness of lexicon-based sentiment analysis models is highly context-specific and depends on the words that are present in the lexicon [6] we decided to extend the VADER sentiment lexicon with terms that are frequently used in basketball-related tweets. To that end we created a list of words that appeared in at least 0.05% of the tweets we collected and excluded all words that were already present in the VADER sentiment lexicon. We then manually traversed the list and excluded all terms that did not carry any sentiment or were ambiguous in terms of polarity. For the remaining 101 terms we let 10 annotators rate the polarity. Here we used the same ordinal scale ranging from -4 to 4 of the VADER lexicon. To ensure that participants evaluate each word in a basketball context, 10 randomly selected tweets were added to each word in which the word appeared and presented to the participants. We also familiarised the participants with the meaning of abbreviations like "MVP" (*Most Valuable Player*) or "GOAT" (*Greatest Of All Time*). Finally, to obtain a single sentiment score for each word, we took the mean of the corresponding ratings from the participants. Table 1 lists example words that were added to the VADER sentiment lexicon.

Table 1: Example words added to VADER lexicon.

Word	Sentiment Mean	SD
goat	3.8	0.42
dpoy	3.8	0.63
mvp	3.6	0.70
lit	3.0	0.67
dope	2.8	0.92
underrated	2.3	0.67
deserved	1.9	1.37
clutch	1.0	1.49
overrated	-2.6	0.70
clown	-2.8	1.03
punk	-3.0	0.82
garbage	-3.5	0.70

The VADER sentiment analysis model with its extended lexicon was then used to determine the polarity of the individual tweets in a range between -1 (very negative sentiment) to 1 (very positive sentiment). The VADER sentiment scores of some example tweets are displayed in Table 2.

Table 2: VADER sentiment scores of example tweets.

Tweet	Sentiment
@RealStevenAdams RESIGN !!	-0.456
Thank you, @swish41 you are my hero! @dallasmavs	0.750
@Pacers @yungsmoove21 Congratulations Thad!!! I hope you're a Pacer forever we love you!! @yungsmoove21	0.921

Final Dataset: With the extracted tweet sentiments we had all information needed to create a final dataset ready for predictive modelling. To allow for a better comparability about the predictive performance of variables we decided to include both social media-related as well as unrelated predictors to the dataset and only considered information that is present before a game starts.

With regard to the social media-related variables we mainly referred to the findings from previous researchers that dealt with the topic (see Section 2). Similar to Xu and Yu [22] and Grüttner et al. [7] we used the average sentiment of tweets a player posted before tip-off as a proxy for his mood before the game. Like Grüttner et al. [7] and Lim et al. [13] we used the number of tweets a player posted before tip-off as a measure for his social media activity. Furthermore, with regard to "late night tweeting" [12] we considered tweets players posted at night before the game as a potential indicator for sleep deprivation. To do so we created a binary variable and flagged all games in which a player posted a tweet during normal bedtime (11 p.m. to 7 a.m.) [12] within the time zone of the player's team.

To our knowledge there has not been any research so far that statistically analysed the influence of social media posts athletes were tagged in on their performance. Nevertheless, we believe that there is strong evidence that such posts may also be of use to predict athletic performance. As Hayes el al. [9] indicate, athletes are easily distracted by negative posts addressed to them as such posts give them undesired feelings [9]. For that reason, we included the proportion of negative tweets players were tagged in as a measure for the severity of negative feedback to our set of features. Hayes et al. [9] further note that athletes may also be distracted by the feeling of being compelled to respond to messages addressed to them and that athletes feel guilty if they cannot reply to all messages. As this feeling of guilt may become worse the more posts an athlete is tagged in prior to a competition, we included this variable as a measure for "obligation to respond" [9].

Besides social media-related and unrelated predictors we had to decide for a target that captures the overall performance of an NBA player for a particular game as accurately as possible, considering both offensive and defensive effort. We chose Box Plus Minus (BPM), a metric that uses a player's box score information, position, and the overall performance of the team to estimate the player's contribution in points above league average per 100 possessions played [15].

Table 3 summarises the variables that formed our final dataset.

3.3 Predictive Significance Analysis

The formed dataset provided all necessary data to assess whether the features derived from the tweets lead to a significant improvement in accuracy of a model predicting player's BPM score in upcoming games. Unlike other researchers before [22,13] we decided against a linear model in that regard, as its coefficient estimates are prone to be biased if the functional form is inappropriately chosen. Instead, we chose a random forest [3], that, in contrast to parametric models like linear regression, naturally adapts to non-linearities and interactions in the data without any prior knowledge about the data distribution. Besides this advantage, latest research found statistical properties of random forests that turned out to be of use for our analysis. As such, Mentch and Hooker [14] demonstrated that predictions from subsampled random forests can be viewed as incomplete,

Table 3: Variables of the final dataset.

Variable	Summary
(I) Social media-related features	
1 ***posted_count***: Number of tweets player posted within 24h before tip-off	$\bar{x}$: 0.3, s : 1.1
2 ***posted_sentiment***: Mean sentiment of tweets player posted within 24h before tip-off	$\bar{x}$: 0.3, s : 0.4, 83% missing
3 ***late_night_tweeting***: Flag if player posted a tweet during normal sleeping hours (11pm-7am) in the night before the game	T:2.6%, F:97.4%
4 ***tagged_count***: Number of tweets player was tagged within 24h before tip-off	$\bar{x}$: 77.1, s : 416.9
5 ***tagged_prop_negative***: Proportion of negative tweets player was tagged within 24h before tip-off	$\bar{x}$: 0.1, s : 0.2
(II) Social media-unrelated features	
6 ***player***: Twitter name of player	108 players
7 ***age***: Player age in years	$\bar{x}$: 27.4, s : 4.0
8 ***tenure***: Years past since player started playing for his current team	$\bar{x}$: 3.3, s : 2.7
9 ***salary***: Salary of player in Million USD	$\bar{x}$: 11.5, s : 9.0
10 ***position***: Position of player	SF:14%, PF:18%, PG:22%, SG:22%, C:23%
11 ***team***: Team of player	30 teams
12 ***opponent_team***: Opponent team	30 teams
13 ***homegame***: Flag if homegame for player's team	home:50.4%, away:49.6%
14 ***season_type***: Game in regular season or playoffs	regular: 92.5%, playoffs: 7.5%
15 ***missing_games***: Number of previous consecutive games player missed, e.g. due to injuries	$\bar{x}$: 0.1, s : 0.8
16 ***past_BPM***: Player's past 10-game exponential moving average BPM score	$\bar{x}$: 0.6, s : 4.2
17 ***past_win_percentage***: Team's past 10-game exponential moving average winning percentage	$\bar{x}$: 0.5, s : 0.2
(III) Target variable	
18 ***BPM***: Player's Box Plus Minus (BPM) score	$\bar{x}$: 0.7, s : 8.4

infinite-order U-statistics that are asymptotically normal so long as the subsample size grows slowly relative to the training set size. The authors made further use of these findings and developed a formal statistical test to assess whether a feature or a set of features make a significant contribution to the prediction for at least one test observation. This test, though valid, becomes computationally prohibitive for test set sizes N_t larger than 20-30 as the test statistic requires the estimation of an $N_t \times N_t$ covariance matrix. Recently, Coleman et al. [4] developed a permutation-style variant of this test that eliminates the need for

covariance estimation and thus retains the same computational complexity as the original random forest procedure regardless of the number of test points. The procedure estimates the predictive significance of a subset of variables X by training two random forests: One original forest RF_{orig} that is trained on all features and one reduced forest RF_{red} that is trained on all features where X is randomly permuted to remove any dependence of X to the target variable. The difference in mean squared error (MSE) between the two forests, i.e. $MSE(RF_{red}) - MSE(RF_{orig})$, is then evaluated on a test set as a measure for the importance of X for the prediction of RF_{orig}. To determine the significance of this difference, a permutation distribution is created to approximate a null distribution by repeatedly permuting the predictions between the forests and recomputing the MSE difference. The p-value is then estimated by evaluating the relative frequency of permutations that resulted in a difference as extreme as the observed MSE difference.

We adopted the testing framework by Coleman et al. [4] also in our analysis setting using a significance level of $\alpha = 0.05$ and performed two types of tests: One group test where we assessed the predictive significance of the set of social media derived features (cf. first category in Table 3) as a whole and one marginal test where we assessed the predictive significance of each of these features individually. For all tests we applied 1000 permutations and used 90% of the data for training and the remaining part for testing. To perform the marginal test for *posted_sentiment* we decided to exclude all records where players did not post any tweet before the corresponding game in order to correctly identify effects between player's approximated mood and their BPM score.

For all tests we used the same hyperparameter configuration to train the random forests. In doing so, we have chosen a setting that, in the best case, satisfies the constraints imposed by the test procedure and provides unbiased test results. Consequently, we chose a relatively small subsample size of $n^{0.6}$, where n corresponds to the training set size, as this value also provided robust test results in various experiments conducted by Coleman et al. [4]. By a similar reasoning we set the ensemble size to 500 trees. Since Strobl et al. [19] demonstrated that random forests based on CART trees tend to overestimate the importance of variables the more cut points they offer, we decided to use conditional inference trees as base learners to obtain unbiased predictive significance estimates. Furthermore, we set the minimal node size to 5, the default for regression problems. Finally, we tuned $mtry$, the number of randomly drawn candidate variables for each split, using 10-fold cross-validation with MSE as evaluation measure. This resulted in an optimal value of $mtry = 6$.

3.4 Implementation Details

To collect the tweets, we used *academictwitteR* [2], an R package that provides an interface to access the Twitter Academic Research Product Track v2 API endpoint. Preprocessing of the tweets was done using the R package *textclean* [17]. For efficiency reasons, the random forests needed for the predictive significance tests were trained with the R package *ranger* [21].

4 Results

In our tests, the RF_{orig} achieved a MSE of approximately 61.7 ($RMSE \approx 7.86$, $R^2 \approx 0.13$). The results of the group test, displayed in Figure 2 (a), suggest that the social media-related features (cf. the first five features in Table 3) make a significant contribution to the prediction of the random forest ($p < 0.001$).

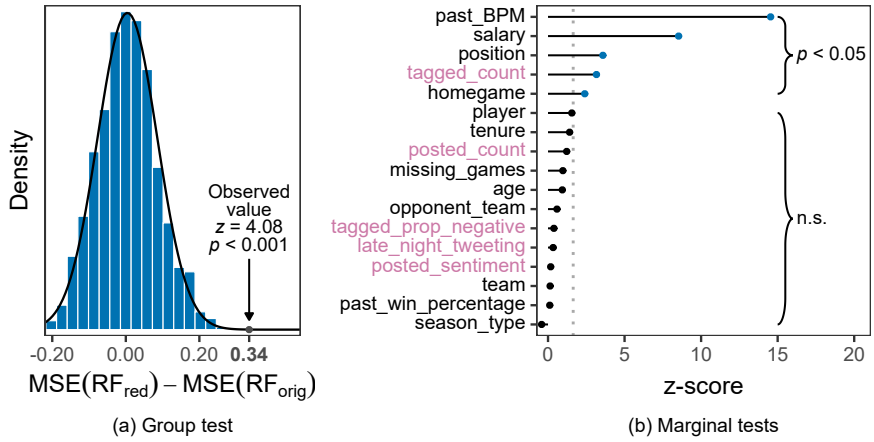

Fig. 2: Results of the group test (a) and marginal tests (b). Social media-related features are highlighted in pink.

By looking more closely at the marginal tests of the individual features (see Figure 2 (b)), only *tagged_count*, i.e. the number of tweets in which the player was tagged before tip-off, showed a significant MSE difference between the original and reduced forest ($p \approx 0.002$). The other social media-related features *posted_count* ($p \approx 0.104$), *tagged_prop_negative* ($p \approx 0.346$), *late_night_tweeting* ($p \approx 0.388$) as well as *posted_sentiment* ($p \approx 0.537$) did not significantly contribute to the predictions of the random forest.

In terms of social media-unrelated features the aggregated performance of the player from past games turned out to be the most important feature for the prediction overall, followed by the salary of the player and his position (all $p < 0.001$). Also, *homegame* significantly contributed to the prediction of the random forest ($p \approx 0.007$).

Figure 3 depicts the bivariate relationships of the significant continuous and discrete features to the BPM target. It should be noted that *tagged_count* is represented in logarithmic scale. Interestingly, the relationship between *tagged_count* and BPM appears to be positive. This contradicts our hypothesis that athletes' performance deteriorates the more messages they receive because they are distracted by the guilt of not being able to respond to all messages.

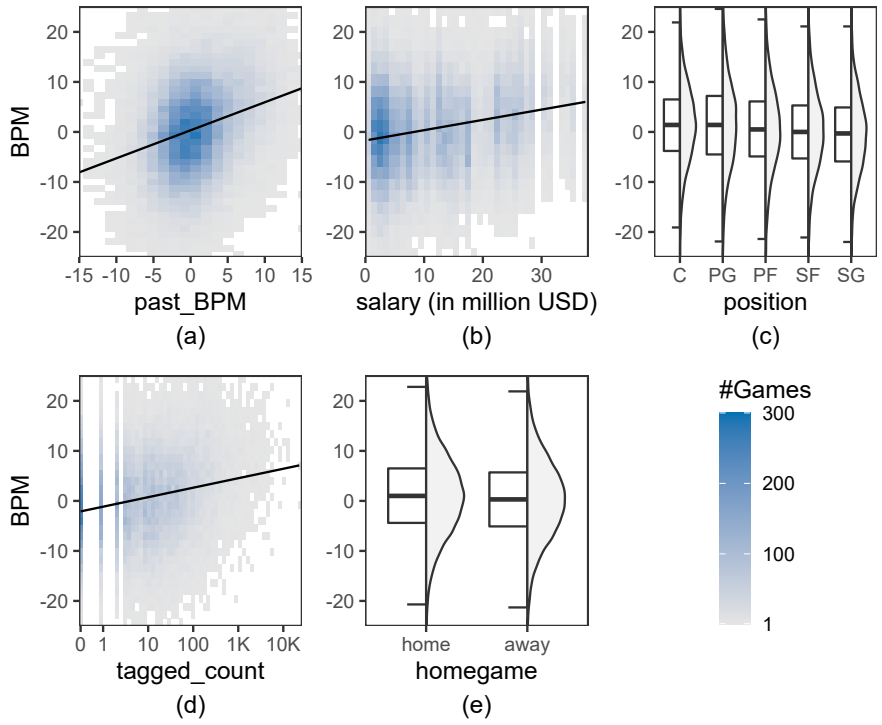

Fig. 3: Relationship between each significant feature (x-axis) and BPM (y-axis).

5 Discussion

The results of our study indicate that features derived from social media posts ultimately lead to a better performing random forest predicting the BPM score of NBA players in upcoming games. However, by having a closer look on the predictive significance of the individual features, only the number of tweets a player was mentioned in before games significantly reduced the MSE of the random forest. With our testing procedure we were unable to replicate the findings from other studies that support a positive relation between the average sentiment of tweets athletes posted and performance [22], a negative relation between the quantity of tweets athletes produced and performance [7,13,12] as well as a negative relation between night tweeting and performance [12]. Even if athletes' performance depends on their mood, social media activity as well as their sleep quality our findings suggest that the effects are either negligible or our approximations from tweets too inaccurate to predict athletic performance effectively. It should be noted, however, that the random forests in our testing procedure take into account potential confounding factors such as age, tenure, and past performance. Other studies, such as that of Grüttner et al. [7], compared only

means using t-tests without adjusting for confounders and thus may have more easily obtained significant results.

Nevertheless, there may be limitations in our study that could have led to inaccuracy of such proxies based on tweets. We believe that one source of error could come from the underlying sentiment analysis model we used to determine the polarity of the tweets. Like other studies before [22,7] we used a lexicon-based model in that regard. Such approaches have the advantage of being relatively easy to interpret and efficient to use, but do not consider that words can mean different things in different contexts. This inflexibility often leads to incorrect sentiment predictions. To give an example consider the word "killer". Without any context, one would naturally associate the word with a negative feeling, which is also reflected in the VADER sentiment Lexicon, which assigns the word a polarity of -3.3. Now consider the tweet: "@russwest44 should be the most respected NBA player of all time. This guy's game mentality is killer". In this case the sentiment of the word obviously flips to a positive meaning. However, due to the context-unspecific nature of lexicon-based approaches VADER assigns a negative sentiment to the entire tweet of -0.228.

Another limitation of our study was to assume that each tweet is equally important and free of bias. In terms of tweets the players posted themselves though, more recent tweets may better reflect the true mood of athletes before game time. It is also unclear whether athletes are actually responsible for all of their social media content as nowadays many celebrities employ agencies that maintain their social media profile for them. Similarly, it is unreasonable to believe that athletes actually read all of the up to thousands of tweets in which they are tagged before tip-off. Tweets from teammates, opponent players, family members and friends may have a higher chance of being actually seen by the athletes.

Given these limitations, our study provides many opportunities for future research. Firstly, it may be reasonable to repeat our analysis with a more sophisticated sentiment analysis model that is particularly built for basketball-related social media content and capable of detecting context-specific sentiment shifts. In that regard, future research could for example use "distant supervision" [6] to automate the sentiment annotation of tweets by making use of emojis or hashtags contained in the tweets. Following this a deep learning model could be trained on the annotated tweets that is capable of classifying the tweets as either positive or negative. Habimana et al. [8] provide an extensive overview how deep learning-based approaches can be used for sentiment classification. Secondly, it may also be interesting to investigate if the expressed emotions in social media text, such as joy, anger, excitement, and fear are useful predictors for athletic performance. Thirdly, future research could also incorporate the time-dependent aspect of social media posts into predictive modelling and train a sequence model to predict athletes' performance in upcoming competitions. Lastly, considering that the inclusion of the variable *tagged_count* significantly reduced the MSE of the random forest, it might be interesting to further investigate the influence of the amount of social media content in which athletes are tagged on their perfor-

mance. In this context, one could also include social media content from other platforms such as Facebook in the analysis.

6 Conclusion

In this study, we investigated the potential of athlete-related social media posts to predict athletes' performance in upcoming competitions. To do this, we extracted tweets NBA players posted themselves or were tagged by others. From these, we derived features that reflect athlete mood, social media behaviour, and sleep quality before games. Using these and other features, we performed a statistical test to investigate whether the MSE of a random forest predicting players' BPM score in upcoming games significantly decreases when the model can utilise the features. The results of this test show that, in particular, the number of tweets NBA players receive before games contributes significantly to the prediction. Contrary to some previous studies, our results neither support a relationship between the average sentiment or number of tweets posted by athletes and their performance nor a relationship between night tweeting and performance. Further research is needed to rule out the possibility that this is due to the inaccuracy of the sentiment analysis model or to limiting assumptions we made about athletes' social media interaction behaviour (e.g. athletes may not be responsible for content posted from their account).

References

1. Baron, R.S.: Distraction-Conflict Theory: Progress and problems. Advances in Experimental Social Psychology **19**, 1–40 (1986). https://doi.org/10.1016/S0065-2601(08)60211-7
2. Barrie, C., Chun-ting Ho, J.: academictwitteR: an R package to access the Twitter Academic Research Product Track v2 API endpoint. Journal of Open Source Software **6**(62), 3272 (2021). https://doi.org/10.21105/joss.03272
3. Breiman, L.: Random Forests. Machine learning **45**(1), 5–32 (2001). https://doi.org/10.1023/A:1010933404324
4. Coleman, T., Peng, W., Mentch, L.: Scalable and Efficient Hypothesis Testing with Random Forests (2019). https://doi.org/10.48550/arXiv.1904.07830
5. ESPN: Vince Carter Addresses the Negative Effects of Social Media on Athletes. `https://www.youtube.com/watch?v=1cX5_2YadU4` (2020), [Online, accessed 03.03.2022]
6. Giachanou, A., Crestani, F.: Like it or not: A Survey of Twitter Sentiment Analysis Methods. ACM Computing Surveys (CSUR) **49**(2), 1–41 (2016). https://doi.org/10.1145/2938640
7. Grüttner, A., Vitisvorakarn, M., Wambsganss, T., Rietsche, R., Back, A.: The New Window to Athletes' Soul–What Social Media Tells Us About Athletes' Performances. In: Proc. of Hawaii International Conference on System Sciences (HICSS). pp. 2479–2488 (2020). https://doi.org/10.24251/HICSS.2020.303
8. Habimana, O., Li, Y., Li, R., Gu, X., Yu, G.: Sentiment Analysis using Deep Learning Approaches: An Overview. Science China Information Sciences **63**(1), 1–36 (2020). https://doi.org/10.1007/s11432-018-9941-6

9. Hayes, M., Filo, K., Geurin, A., Riot, C.: An Exploration of the Distractions Inherent to Social Media Use Among Athletes. Sport Management Review **23**(5), 852–868 (2020). https://doi.org/10.1016/j.smr.2019.12.006
10. Hutto, C., Gilbert, E.: VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In: Proc. of AAAI Conf. on Web and Social Media. vol. 8, pp. 216–225 (2014), `https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109/8122`
11. Iso-Ahola, S.E.: Intrapersonal and Interpersonal Factors in Athletic Performance. Scandinavian Journal of Medicine & Science in Sports **5**(4), 191–199 (1995). https://doi.org/10.1111/j.1600-0838.1995.tb00035.x
12. Jones, J.J., Kirschen, G.W., Kancharla, S., Hale, L.: Association Between Late-Night Tweeting and Next-Day Game Performance Among Professional Basketball Players. Sleep Health **5**(1), 68–71 (2019)
13. Lim, J.H., Donovan, L.A.N., Kaufman, P., Ishida, C.: Professional Athletes' Social Media Use and Player Performance: Evidence From the National Football League. International Journal of Sport Communication **-1**, 1–27 (2020). https://doi.org/10.1123/ijsc.2020-0055
14. Mentch, L., Hooker, G.: Quantifying Uncertainty in Random Forests via Confidence Intervals and Hypothesis Tests. The Journal of Machine Learning Research **17**(1), 841–881 (2016)
15. Myers, D.: About Box Plus/Minus (BPM). `https://www.basketball-reference.com/about/bpm2.html` (2020), [Online, accessed 12.03.2022]
16. von Ott, K., Puymbroeck, M.V.: Does the Media Impact Athletic Performance. The Sport Journal **9** (2006)
17. Rinker, T.W.: textclean: Text Cleaning Tools. Buffalo, New York (2018), `https://github.com/trinker/textclean`, version 0.9.3
18. Rousidis, D., Koukaras, P., Tjortjis, C.: Social Media Prediction: A Literature Review. Multimedia Tools and Applications **79**(9), 6279–6311 (2020). https://doi.org/10.1007/s11042-019-08291-9
19. Strobl, C., Boulesteix, A.L., Zeileis, A., Hothorn, T.: Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. BMC Bioinformatics **8**(1), 1–21 (2007). https://doi.org/10.1186/1471-2105-8-25
20. Watkins, R.A., Sugimoto, D., Hunt, D.L., Oldham, J.R., Stracciolini, A.: The Impact of Social Media Use On Sleep Quality and Performance Among Collegiate Athletes. Orthopaedic Journal of Sports Medicine **9**(7_suppl3) (2021). https://doi.org/10.1177/2325967121S00087
21. Wright, M.N., Ziegler, A.: ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. arXiv preprint arXiv:1508.04409 (2015)
22. Xu, C., Yu, Y.: Measuring NBA Players' Mood by Mining Athlete-Generated Content. In: Proc. of Hawaii International Conference on System Sciences (HICSS). pp. 1706–1713. IEEE (2015). https://doi.org/10.1109/HICSS.2015.205