Predicting Twitter Engagement for Leading Sports Apparel Brands

Boishin Iva, Deleval Justine, Dafni Krystallidou

February 14, 2019

1 Abstract

This report presents our research approach to predicting users' online engagement with famous Sports Brands Twitter posts using machine learning models. 2746 tweets were extracted from the top 5 Sports companies (Nike, Adidas, UnderArmour, Columbia, Reebok). We conducted fine-grained analyses to extract components such as topics, sentiments, and keywords which are important contributors to a tweet's content. The results from random-forest and logistic regression methods were contrasted. The models were assessed using the AUC as well as precision and classification accuracy metrics. The normalised logistic regression performed better on engagement prediction (likes and retweets) across the different evaluation metrics. Predictor variables that showed to be important for predicting engagement included the presence of video, is_retweet, presence of hashtag, photo, questions, or urls. The findings are consistent with previous literature confirming the importance of the aforementioned features for online engagement. We thus, propose a normal logistic regression model for Twitter engagement prediction with Sports Apparel Brands. Finally, limitations and avenues for future research are discussed.

2 Introduction

Given the rise of importance of social media marketing, companies are looking for ways to create the most relevant social media content to get the best return on their investment. Despite the rapidly expanding focus on social media analytics for leveraging business, there is no comprehensive model that allows companies to determine whether or not their tweets will perform well (Lassen et al., 2014). In addition, the majority of studies rely on a single type of variables and specifically that of volume. To determine the aspects that increase the likelihood that a Twitter post resonates with social media users and earns the highest engagement (i.e. "favorites" and "retweets"), we created a model that considers all the details of the tweets and used an algorithm to select the most important predictors and create the model.

In particular, we decided to predict Twitter Engagement for Sports Apparel Brands in an effort to extend the extant research on predictive analytics to a new domain. Although, Twitter data has been extensively used for predictive analysis in many application domains, from social issues to public health and politics, it has not been implemented in the context of the sports industry.

This report aims to explain our approach for predicting online engagement for 5 major sports apparel brands using Twitter data. First, we will present the database, data preprocessing methods and variables. Then, the findings will be presented and explained. Finally, we will conclude with a discussion and directions for future research.

2.1 Data Preparation and Preprocessing

Using the rtweet package, we accessed the Twitter API to extract the most recent tweets of the top five sports apparel brands. Only the tweets from profiles set to English were kept to allow for accurate text mining later in the process. We decided to do this because the number of tweets from profiles not in English was very minimal and would have required a lot of extra processing for only very marginal information. Then, we explored the data to get a better idea of what type of information we can extract from the data gathered.

While there was a media type column, the only values in that column were “photo” and “NA”. This meant that we had to do extra processing to check whether the media type was actually a photo or whether it was a video. Noticing that “video” appears in either one or two of the two media url columns, we used grepl to determine which tweets had an embedded video. Then, the rest of the tweets having a media url were images. This step was necessary as there is quite a bit of literature suggesting that photos and videos gather more attention from social media users, who are consistently overloaded with information and are looking for a simple way of getting information.

Next, we created predictor variables for time metadata of the tweets. We created dummy variables for the weekday as well as for the month of the year. Moreover, we extracted the hour of the tweet to determine the time of day that the tweet was sent and create more dummy variables for those. We were curious to see whether there was a particular time of day, week or year that was a particular popular Twitter usage time and that yielded higher engagement.

Afterwards, we used regex to count the number of elements specific to tweets: mentions, hashtags and emojis. We also created a dummy variable for the presence of each of those in the tweets. The rationale behind having created a dummy variable on top of having a frequency variable is that several studies claim that binary regression models and classification models tend to get the best results from variables that have zeros and ones instead of a large range of numbers. We also created a dummy variable for tweets that include mentions of the top athletes or not and for tweets that include questions or not. The idea behind these potential predictor variables is that questions could encourage users to retweet or comment while mentions of popular athletes could encourage favorites simply due to the loyalty of users for that athlete in particular.

Finally, for the non-binary numeric columns that we planned to use as variables or targets, we checked and corrected the skewness through various approaches. We first tried to winsorize the data to see if by reducing the effect of outliers, the skewness might reduce. Through research, we also saw that taking the square root or the log of a column could normalize the data within the column. Given that the mentions and emoji variables were not as skewed as favorites and retweets, those former columns were normalized using the square root approach. For favorites and retweets, we needed to use the more drastic approach as the data were very skewed. The goal behind the normalization was to allow for a more balanced separation of targets and a model with more predictive power at the end.

The last step was to convert the data to binary for our binary regression. We used the winsorized target variables and the normalized target variables to create measures of high and low engagement. Our threshold for this was the mean of the data. We kept the winsorized data because it highlights the tweets that truly stood out in terms of engagement from the other tweets. However, we were aware that this might result in a model that is does not have high predictive power as it might simply predict all values as zero due to the low incidence rate. As a result, we wanted to create some models on the normalized data, which has a split closer to 50-50.

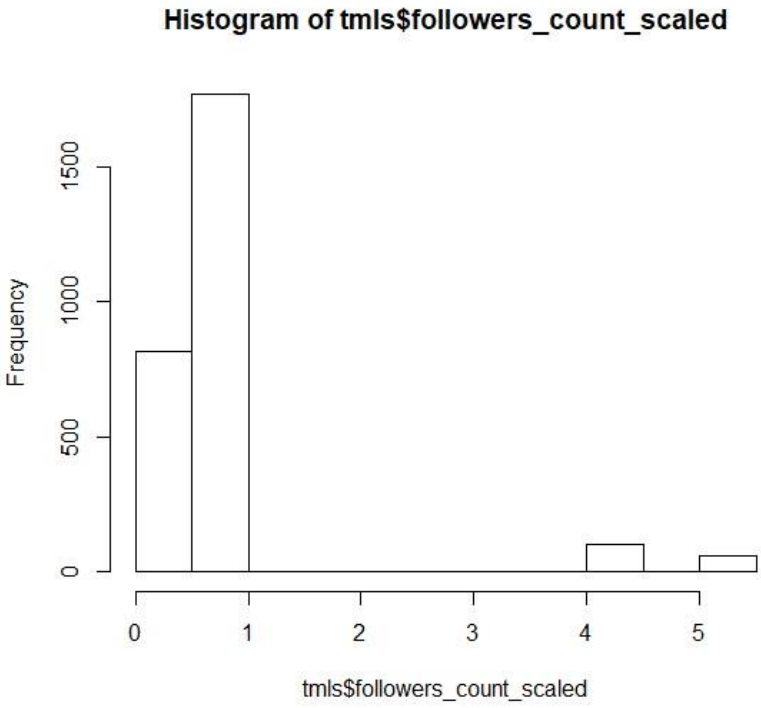
2.1.1 Points of Consideration

Initially, we planned on running this analysis for a specific company: Nike. However, due to the Twitter API rate limit and the prevalence of “reply” tweets, we were only able to extract less than 200 relevant tweets. As a result, we expanded our model to take the official accounts of the top brands in the sports apparel industry. Depending on the company, it might be possible to run this analysis on one company by looking at the various country accounts of one particular company and take a global view to the process. Some difficulties that might arise are problems with text processing due to the presence of languages other than English.

Additionally, we want to mention that we were planning on running the model for the number of comments as an engagement metric. Yet, the free API does not give access to that type of data. As a result, for individuals and companies with access to the premium Twitter API, it would be possibly to extend this code to include a model for comments engagement.

Finally, after running our initial forward-stepwise logistic regression and random forest, we noticed that the API was unrealistically high at around 0.96. Looking at the scaled followers count histogram (below) and comparing it to the tweets with the highest engagement, we noticed a near perfect correlation between engagement and brand (essentially, since the followers count variable is the same for all tweets from the same brand). It would stand to reason that brands with loyal customers will have

higher engagement. This however is very intuitive and we were more interested in the specific aspects of a tweet that increased engagement. So, we excluded that variable from the set of predictor variables as it was a leak.



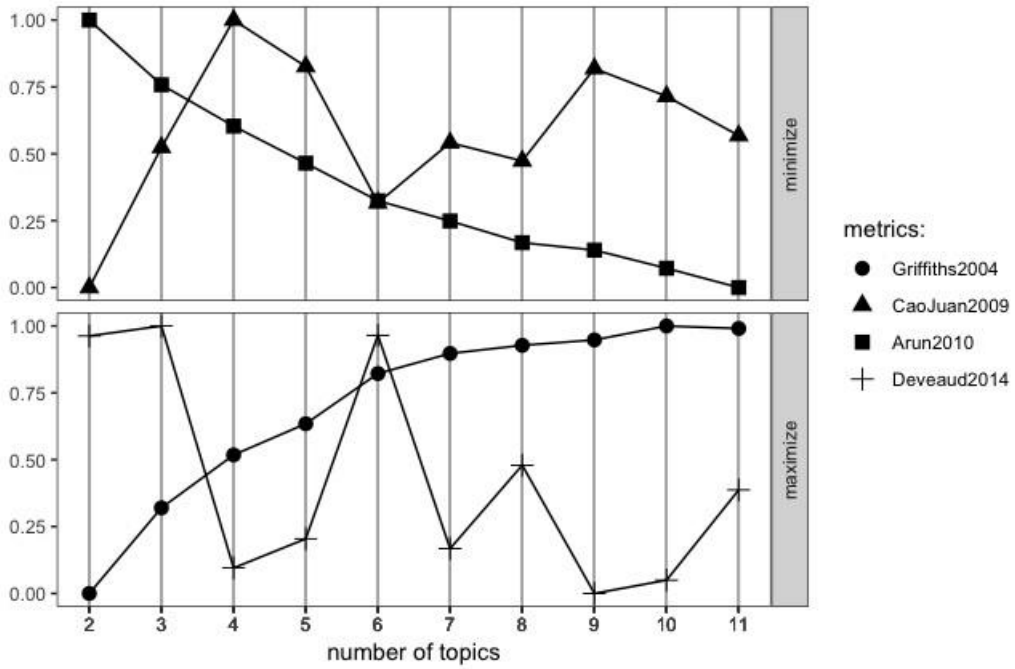
3 Creation of Additional Predictor Variables

To better predict twitter engagement, we tried to include important predictor variables. On this basis we conducted 3 different fine grained text analyses before building our model: sentiment, topic analysis, and keyword extraction. Conducting these types of analysis ensured that different types of content-related variables would be considered including volume-based, semantic and sentimental variables.

3.1 Topic Analysis

We first performed the topic analysis. We opted for an unsupervised approach using the Lateral Dirichlet Allocation (LDA) technique. The text of the tweets were analyzed. The steps were performed in the following order:

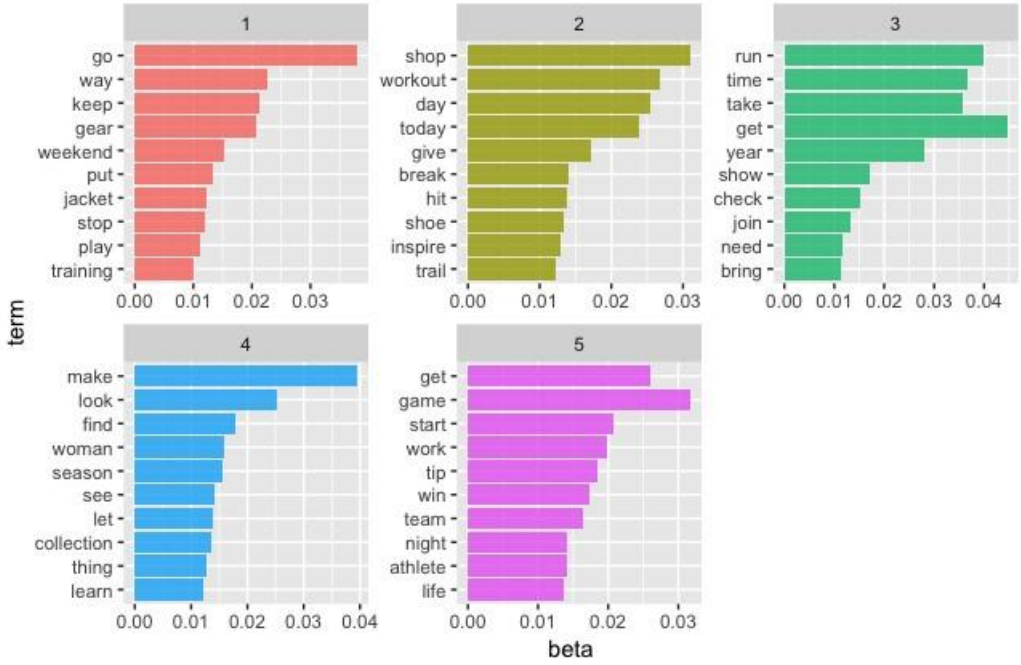
1. The encoding of the users' comments was converted from "utf8"" to "ascii" format.
2. Corpus Generation from the character vector of the 'text' variable.
3. Text Cleansing. Transformer functions were applied to clean the text for the analysis. We removed mentions, hashtags, links, punctuation, numbers as well as stripped whitespace.
4. The Corpus was then converted into a data frame to perform word annotation using the udpipe package. Only tokens tagged as nouns and verbs were included in the topic analysis.
5. Stop-words removal.
6. Creation of a Document-Term-Matrix (dtm).
7. Identifying the optimal number of topics.



In order to determine the ideal number of topics for the LDA mode, 5 different techniques were used from the lda tuning package (Murzintcev, 2016). The graph below demonstrates that 5 topic is the best number of topics for our analysis.

8. LDA model Building with 5 topics

The LDA model was built with 5 topics and 10 words per topic. The results of the topic modelling are presented in the graph below.



A possible interpretation of the topics is described below:

Topic 1: The words team, athletes, world, and woman led us to infer that the topic is about promotion of athletes, team and woman athletes.

Topic 2: The words game, day, winning, today led us to infer that the topic is about announcement and reminders of game days for teams or athletes.

Topic 3: The words gear, shoe, season led us to infer that the topic is about promotion, advertisement of gear and/or Equipments.

Topic 4: The words inspiration, feel, look, and want led us to infer that the topic is about inspirational tweets promoting positive aspect of sports.

Topic 5: The words workout, run, and meet led us to infer that the topic is about specific workout and running related tweets.

Finally, The topics were converted into dummy variables and were added as columns in the original basetable as follows: “topic 1: Athletes”, “topic 2: Game Day”, “topic 3: Equipments”, “topic 4: Inspirational”, “topic 5: Workouts”. Tweets were assigned to topics with highest gamma value.

Tweets having same number of gamma values for two different topics were assigned a value of 1 for both topics.

3.1.1 Points of Consideration:

In the beginning we tried to perform topic analysis using only noun tokens. However, one of the challenges we encountered was that approximately 130 tweets did not contain any noun tokens and were subsequently excluded from the analysis as they were not assigned any topic. For this reason, we included both nouns and verbs in the topic modelling. The resulting $\hat{I}^2$ values were low, indicating the absence of strong associations between the words and topics. Similarly, the gamma values were low showing a weak mapping of tweets into topics. This could have resulted from a lack of consideration of ‘the context’ of the words. Implementation of word embeddings could powerfully prevent this issue.

As illustrated in the graph some of the words (e.g. ‘work’) are present in more than one topics, hampering the topics distinctiveness. Topics to be interpretable and convey enough information need to be distinct (Wallach et al., 2009). A possible solution would be to implement a topic-weak correlated LDA to reduce the overlapping between the topic-word distributions. This would in turn, result in more semantically meaningful topics that could be better predictors of twitter engagement (Tan & Ou, 2015).

3.1.1.1 Sentiment Analysis

The extant literature has shown the predictive power of sentiment-related variables for users’ online engagement (Kalampokis et al., 2013). On this basis, we conducted a dictionary-based sentiment analysis to get the emotional valence of the tweets. The following steps were taken:

1. The English sentiment dictionary was loaded
2. Emoji Dictionary was found online, web scrapped and recoded for the lookup
3. The mentions and tabs were removed from the tweet text to ensure that the emojis were standalone
4. A sentiment analysis lookup loop was run for each tweet

a. Original tweet texts were tokenized and searched for emojis

b. Tweet lemma sentences created during the Topic Analysis were tokenized and searched

c. Word valences for tokens preceded by a negator were reversed

d. Emoji and lemma valences were combined and averaged

e. Tweets containing exclamation points were escalated by 20% to convey intense emoji

5. Emotional scores were added in the basetable in the "sentiment_analysis" variable. The variable has values from -5 to 5.

3.1.2 Points of Consideration

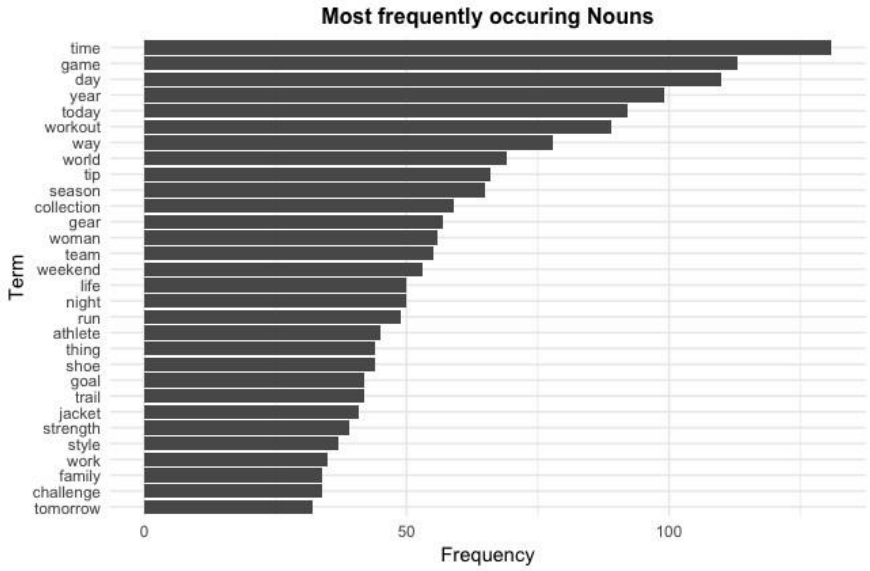
The main difficulties with this text mining was finding and properly encoding an emoji dictionary. The problem was caused due to the fact that the online dictionary found was encoded in unicode codepoint while the tweets were encoded in UTF-8. Once we understood the different types of encodings, we were able to find a function–intToUtf8–to translate the unicode codepoint emojis in the dictionary to UTF-8. Even though the emoji encoding finally looked like the one in the tweet text, they were still not getting matched when the sentiment analysis was run. After experimenting a bit with the 38th tweet, we realized that R was not able to understand the value in the dictionary. As a result, we used the function enc2native to convert the way that the emojis are stored.

One thing to mention is that the emoji dictionary takes into account the type of emoji but it does not take into account the skin color of emoji. For this analysis, the skin tone color was irrelevant but if it is necessary, the six skin tone colors can be added to the dictionary and they can then be combined with the code for the specific emoji type through a bigram token approach.

3.1.2.1 Keyword Extraction

Keyword extraction was performed to identify commonly used keywords and create relevant variables for our basetable. Two different methodological approaches were employed:

In the first approach, we utilized the Part-Of-Speech (POS) taggings of the tokens to identify the most frequently occurring words. We focused only on nouns. The 30 most frequently occurring words are presented in the graph below. Although this approach provides some interesting insights into the most frequently occurring nouns, it assumes that the most frequent nouns are also the most relevant, which is not always the case. Additionally, it ignores complex relationships and dependencies between tokens. To counteract this issue, we employed an intricate extractive summarization method using on the TextRank Package.



The TextRank algorithm enabled us to assess the relationship between the words. In effect, the central tenet of the TextRank algorithm is to understand how words relate to each other and uncover links between them. Relevant words that follow one another are first identified, linked and subsequently used to construct a word network. The Google Pagerank algorithm is then applied on the word network to rank words in order of importance. Links between words occurring frequently together are assigned a higher weight. Finally, keywords are extracted by combining together the most relevant words that follow one another. Using this approach, 20 additional keywords were extracted.

The keywords constructed using the two approaches were combined, forming a character vector of 40 keywords. This was incorporated in our basetable in three different forms: one variable “keyword_count” denoting the count of keywords in each tweet, a second variable keyword_freq displaying the frequency of keywords (i.e. keywords/total number of words) and a third dummy variable named “keyword_dummy”" indicating the presence or absence of keywords (0 if not containing the keyword, 1 if contain at least 1).

3.1.3 Variable Description

The final basetable contained 66 variables:

1. ***X***: an index for each tweet - Type: integer
2. ***created-at***: datetime on which the tweet was posted - Type: character
3. ***screen_name***: name of the sports Brand (Adidas, Nike, Columbia1938, UnderArmour, Reebok)- Type: character
4. ***text***: the comment entailed in the tweet. This variable was used for text mining - Type: character
5. ***display_text_width***: number of characters contained in each tweet - Type: character
6. ***is_quote***: a binary variable indicating whether the brand's tweet is a quote (i.e. retweet with additional comments or mentions). A value of 0 means that the tweet is not a quote, a value of 1 indicates a quote - Type: integer
7. ***is_retweet***: a binary variable indicating whether the brand's tweet is a retweet. Retweets are assigned a value of 1, non-retweets are assigned a value of 0 - Type: integer
8. ***favorite_count***: number of user likes for each brand's tweet - Type: integer
9. ***retweet_count***: number of times the brand's tweet was retweeted - Type: integer
10. ***url_present***: a binary variable indicating whether the brand's tweet contains a url. If the tweet contains a url, it is assigned a value of 1 otherwise 0. - Type: integer
11. ***video***: a binary variable indicating whether the brand's tweet contains a video. If a video is present then the tweet is assigned a value of 1 otherwise 0. - Type: integer
12. ***photo***: a binary variable indicating whether the brand's tweet contains a photo. If a photo is present then the tweet is assigned a value of 1

otherwise 0. - Type: integer

13. ***WD***: a variable indicating the weekday of the brand's tweet. - Type: character

14-19. ***Actual Weekday of the tweet***: Binary variables for each Day of the Week ranging from Monday to Friday - Type: integer

20. ***year***: the year the brand's tweet was made - Type: integer.
21. ***month*** : the month in which the brand's tweet was made - Type: character.

22-32. ***Actual Month of the tweet***: Binary variables for each Month ranging from January to November - Type: integer

33. ***hour***: the hour at which each brand tweet was made. - Type: integer
34. ***morning***: Binary variable indicating whether the tweet was made between in the morning (i.e. between 4 to 11:59 o'clock) - Type: integer
35. ***afternoon***: Binary variable indicating whether the tweet was made between in the afternoon (i.e. between 12 to 19:59 o'clock). - Type:
    integer
36. ***mentions_count***: the number of mentions in each brand's tweet. - Type: integer
37. ***mentions_presence***: Binary variable indicating the presence of mentions in each brand's tweet. If mention is present, then assigned a value of 1 otherwise 0. - Type: integer
38. ***hashtag_count***: the number of hashtags in each brand's tweet. - Type: integer
39. ***hashtag_presence***: Binary variable indicating the presence of hashtags in each brand's tweet. If hashtag is present, then assigned a value of 1 otherwise 0. - Type: integer
40. ***emoji_count***: the number of emojis per Brand's tweet. - Type: integer
41. ***emoji_presence***: Binary variable indicating the presence of emojis in a brand's tweet. If emoji is present, then assigned a value of 1 otherwise 0. - Type: integer
42. ***mentions_sqrt***: the adjusted mentions score using adjusting square root function.- Type: numeric
43. ***emoji_sqrt***: the adjusted emojis score using adjusting square root function. - Type: integer
44. ***favorite_log***: the adjusted favorites score (i.e. likes) using logarithmic transformation. - Type: numeric
45. ***retweet_log***: the adjusted retweet score using logarithmic transformation. - Type: numeric

46. ***target_fav_norm***: first binary target variable for favorites (i.e. likes) created using normalised data. Values under the mean were classified as 'low engagement', while values over mean were classified as 'high engagement'. - Type: integer
47. ***target_RT_norm***: second binary target variable for retweets created using normalised data. Values under the mean were classified as 'low engagement', while values over mean were classified as 'high engagement'. - Type: integer
48. ***target_fav***: binary target variable for favorites (i.e. likes) using winsorized data. Values under the mean are low engagement while values over zero are high engagement - Type: integer.
49. ***target_RT***: binary target variable for retweet using winsorized data. Values under the mean are low engagement while values over zero are high engagement - Type: integer.
50. ***tweet_length_scaled***: standard score (z-score) for the length of tweets. - Type: numeric
51. ***followers_count_scaled***: scaled score for the followers count per Brand. Scaling was performed by assigning values to outliers equal to 3 sds. This variable was not used to construct the model to avoid multicollinearity issues. - Type: numeric
52. ***question***: binary variable indicating the presence of a question mark. Tweets containing questions might open the scene for discussions and thus might increase users' engagement. If the tweet contained a question mark it was assigned a value of 1, otherwise 0. - Type: integer
53. ***athletes_mentions***: binary variable indicating the mention of an athlete. In order to create this variable, a character vector 'dictionary' was constructed containing the twitter account names of famous athletes mentioned by the 5 Sports brands of interests. The dictionary included only account names of athletes that had been mentioned more than one time. - Type: integer
54. ***text_lemma***: the lemmatised text following annotation.- Type: character
55. ***word_total***: total number of words per brand tweet. - Type: integer
56. ***noun_freq***: frequency of noun per brand tweet. - Type: numeric
57. ***verb_freq***: frequency of verbs per brand tweet. - Type: numeric

58-62. ***topics 1 -topic 5***: the topic dummy variables indicating the mapping of a tweet to one of the 5 topics. - Type: numeric

63. ***sentiment_analysis***: the overall sentimental score per Brand tweet. Type: numeric
64. ***keyword_count***: number of keywords per brand's tweet. Type: numeric
65. ***keyword_freq***: frequency of keywords per brand's tweet. This was calculated by dividing the keyword count by the word_total.Type: numeric
66. ***keyword_dummy***: a binary variable indicating the presence of keywords in a brand's tweet. Type: numeric

4 Model Building

We ran a forward-stepwise logistic regression and a random forest to create several models and determine their predictive power. This feature selection algorithm allowed us to select the best predictors by adjusting the number of variables and checking the AUCs of the respective models. In both the "favorite" engagement target and the "retweet" engagement target, the forward-stepwise logistic regression with the normalize targets seemed to have the highest performance. Performance details are shown in the shiny application.

5 Evaluation

Both models using logistic regression & random forest algorithms were evaluated for the two different targets namely, favorites and targets. The robustness of the models was assessed in terms of prediction and classification accuracy as shown in the graphs of the shiny application.

5.1 AUC Performance Measure

The AUC of the models was evaluated as an aggregate measure of performance. For all models the AUC was above 0.7, with the exception of the random forest for the favorites target. This demonstrates that the models were good in classifying high and low engagement. The details of the AUC performance are presented in the shiny application.

5.2 Note

Since the random forest model that we used was of a classification type, we couldn't use the MSE and RMSE metrics of prediction accuracy as they are not relevant. For these reasons we decided to look at classification accuracy.

5.3 Classification Accuracy

In terms of classification accuracy, we used the confusion matrix technique. The findings indicated that although the non-target (no engagement) was classified correctly all the times our model was very 'lenient' in that target was often classified as non-engagement as evidenced by the increased number of false positives. On this basis, we tried to adjust the cut-off threshold. Although the result slightly improved, we still had a considerable classification error (7.59%). A possible partial explanation could be attributed to the reduced number of targets in our dataset. Despite the considerable efforts to improve the effectiveness of our model, by stratifying the target variable in train and test datasets, the very low target incidence might have hampered the classification accuracy of our model.

We further evaluated the F1 of the models. F1 is a harmonic measure of accuracy representing the weighted mean between Recall and Precision. This renders it a good indication of both exactness and completeness as it takes into account False Positives and False Negatives. The results showed that the normalised logistic regression consistently achieved higher classification accuracies than its non-scaled counterpart.

To conclude, normalised logistic regression proved to be a good predictor of Twitter engagement with Sports Apparel Brands. In particular, the optimal prediction model (as measured by likes) encompassed 7 prediction variables including: the presence of video, is_retweet, hashtag, photo, question mark, url presence and the display text_width. Interestingly, the retweet, question and url features negatively correlated with engagement, resulting in reduced user interaction. The aggregate performance of the model as measured by the AUC was 0.88 for the train and 0.86 for the test dataset, indicating almost no overfitting. Importantly, the model scored high in classification accuracy, classifying 71% of high-engagement cases correctly.

6 Limitations & Avenues for Future Improvement

Despite the considerable efficiency of our model, there are some important limitations that need to be addressed. These limitations mostly relate to availability of Twitter user demographics, location-based and profiling metrics that have been shown to be good engagement predictors. Second, the findings of this study are exclusive to the sports industry making it hard to generalize the model in different settings and domains. Moreover, our project did not consider more novel approaches for topic modelling using deep learning methods or fine-grained latent semantic indexing.

Furthermore, in terms of data visualization and the Shiny application it would be interesting to create an application where the user could input the Sports Brand name, the model was run up until the point of topic modelling. The user could then input the number of topics and see how the model would differ accordingly. Additionally, the AUC graphs could be presented in a more interactive way where the user could input the number of predictor variables and the respective AUC graphs would be plotted.

Based on our findings the following directions for future research have been formulated. First, we suggest that future studies focus on predicting online engagement with Sports Brand over a longer time-span. This will result in a more representative model, where the effect of predictors such as year and month can be better evaluated. Second, future research should take into account variables relating to followers' profiling characteristics, location and observed behaviours. The number of clicks per brand tweet, interests of users interacting with the brands and referral channel could prove important predictors of Twitter engagement. However, we did not have access to the aforementioned information at the time of the analysis as this information is restricted to Twitter for business users. Furthermore, it would be interesting to explore the relationship between the sports company's growth rate and user's engagement. These features could be inputted as predictor variables in our model, and could potentially improve its accuracy.

What is more, the sentiment and topics variables were not important predictor variables in our model. This could be partly attributed to the informal language and jargon often found in Twitter texts. However, an additional explanation could be the use of a lexicon-based approach. The extant literature questions the appropriateness of lexicon-based methods for sentimental analysis of social media texts. Hence, the need for a shift towards machine learning techniques for sentiment text classification is imperative.

Similarly, for topic modelling word-embedding methods could be used to ensure the generation of relevant, context-dependent topics. The LDA, though a useful topic-modelling technique, ignores the context of words. The context should not be ignored as its importance for social media prediction has been shown to be pivotal. Finally, a further suggestion would be to conduct the analysis based on hashtags tokens. Hashtags serve to categorise a tweet's topics (Kursuncu et al., 2018). Thus, it would be interesting to see if topics generated purely on hashtags would be a significant predictive feature of our model.

7 References

Kalampokis, E., Tambouris, E., & Tarabanis, K. (2013). Understanding the predictive power of social media. Internet Research, 23(5), 544-559.

Kursuncu, Ugur & Gaur, Manas & Lokala, Naga Usha Gayathri & Thirunarayan, Krishnaprasad & Sheth, Amit & Arpinar, Ismailcem. (2018).
Predictive Analysis on Twitter: Techniques and Applications.

Niels Buus Lassen, Rene Madsen, and Ravi Vatrapu. 2014. Predicting iPhone Sales from iPhone Tweets. In Proceedings of the 2014 IEEE 18th International Enterprise Distributed Object Computing Conference (EDOC '14). IEEE Computer Society, Washington, DC, USA, 81-90.
DOI=http://dx.doi.org/10.1109/EDOC.2014.20

P. K. Novak, J. Smailovi Ìc, B. Sluban, and I. MozetiËc, âSentiment of Emojis,â PLOS One, 2015.

Tan, Yimin & Ou, Zhijian. (2011). Topic-weak-correlated Latent Dirichlet allocation. 224 - 228. 10.1109/ISCSLP.2010.5684906.

Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09). ACM, New York, NY, USA, 1105-1112. DOI:
http://dx.doi.org/10.1145/1553374.1553515