Rachael Jackson
Victoria Owens

Ben Schneider
Sofi Sinozich

Assignment #2: Using the Twitter API and Sentiment Analysis of Tweets[1]

The purpose of this project was to analyze the sentiment of tweets related to the University of Maryland's men's basketball team during the final games of the regular season. This report outlines the methods our team used to gather and analyze the tweets and summarizes the most important results.

Data Collection and Preparation

Selecting Keywords

Keywords were determined based on group members' understanding of relevant terms after some initial qualitative research. These included references to the specific UMD men's basketball Twitter account (@TerrapinHoops), terms specific to Maryland sports (terp nation, go terps, #feartheturtle, umd turgeon), and generic descriptions (maryland basketball, umd bball, college park basketball). We included emoji in our tweet keywords as well (🐢🏀⛹️‍♂️) both in combination with relevant words and in two emoji-only combinations (🐢🏀 and its reverse). Our final keywords file is available here.

Tweet Streaming Process

Tweets that included the selected keywords were streamed near-continuously over several periods between February 29th and March 13th using *stream_tweets* from *rtweet* (all times Eastern, earliest and latest tweet from each stream session recorded):

- February 29, 5:10p to March 1, 11:08p
- March 5, 9:52a to March 7, 5:33a
- March 8, 9:42a to March 13, 12:42p

We thus achieved coverage of the February 29 game against Michigan State, the March 8 game against Michigan, and the cancellation of the Big Ten and NCAA tournaments on March 12. There were occasional drop-offs within these periods due to API disconnects out of our control. For example, no tweets were recorded on March 11, likely due to reconnection issues, and the March 13 cutoff was due to a disconnection, after which it was judged that sufficient data had

[1] All of the code for this assignment can be found in our repo on GitHub. Full-sized plots can be found at our accompanying page and are linked at each figure title.

been collected. In this report, we clearly denote time periods for which streaming was interrupted (as in Figure 1).

The keyword list changed slightly across these periods due to qualitative examination of incoming tweets. For example, an email blast promoting #FearTheTurtle on March 8 prompted us to add that hashtag and related phrases to the list on that day. (All changes to our keywords can be viewed in the commit history for our keywords file.)

Tweets from all these periods were then concatenated into one master tweets dataset and duplicates due to overlapping streams (matching on tweet status_id) were removed prior to further cleaning and analysis.

Tweet Cleaning

To clean the dataset, each tweet was first tokenized and all tokens were converted to lowercase. The team then identified irrelevant tweets from two likely sources of irrelevancy: different gender/age and different sport. When creating the original list of keywords, the team intentionally omitted words related to gender and age (e.g. high school boys basketball) because we believed such terms might cause us to miss relevant tweets. Therefore, our first consideration when removing irrelevant tweets was to identify tweets related only to UMD's women's basketball team or high school basketball in Maryland. We identified common words related to women's and boys' basketball and removed tweets that included any of those terms *but not* any relevant men's basketball terms. Next, the team identified common words related to other UMD sports and removed tweets that included any of those terms *but not* any relevant basketball terms. (See Table 1 for list of relevant and irrelevant terms related to gender and sports.)

Table 1. List of Relevant and Irrelevant Terms

	Gender			**Sports**		
Irrelevant Words	high	girls/girls'/ girl's/girl	boy's/ boys'	baseball	football	soccer
	ladies	brenda	frese	lacrosse	wlax	mlax
	njcaa	wbb	wbball			
	ashley	owusu	stephanie			
Relevant Words	men/ men's	mbb		basketball	bball	hoops

Creating the list of irrelevant terms was an iterative task. We began by brainstorming an initial list of irrelevant terms and removing tweets containing those terms *but not* containing any of the relevant terms. The team then ran frequency analyses using the qdap package to identify the most common 100 terms in the remaining list. This revealed additional terms that we had not thought of (e.g. the names of individual female basketball players and coaches). We scanned the dataset to see if tweets with these words did, indeed, appear to be irrelevant. Based on this, we added

irrelevant words to the list of irrelevant terms, re-cleaned the dataset, and re-ran the frequency analyses until we stopped seeing irrelevant terms in the list.

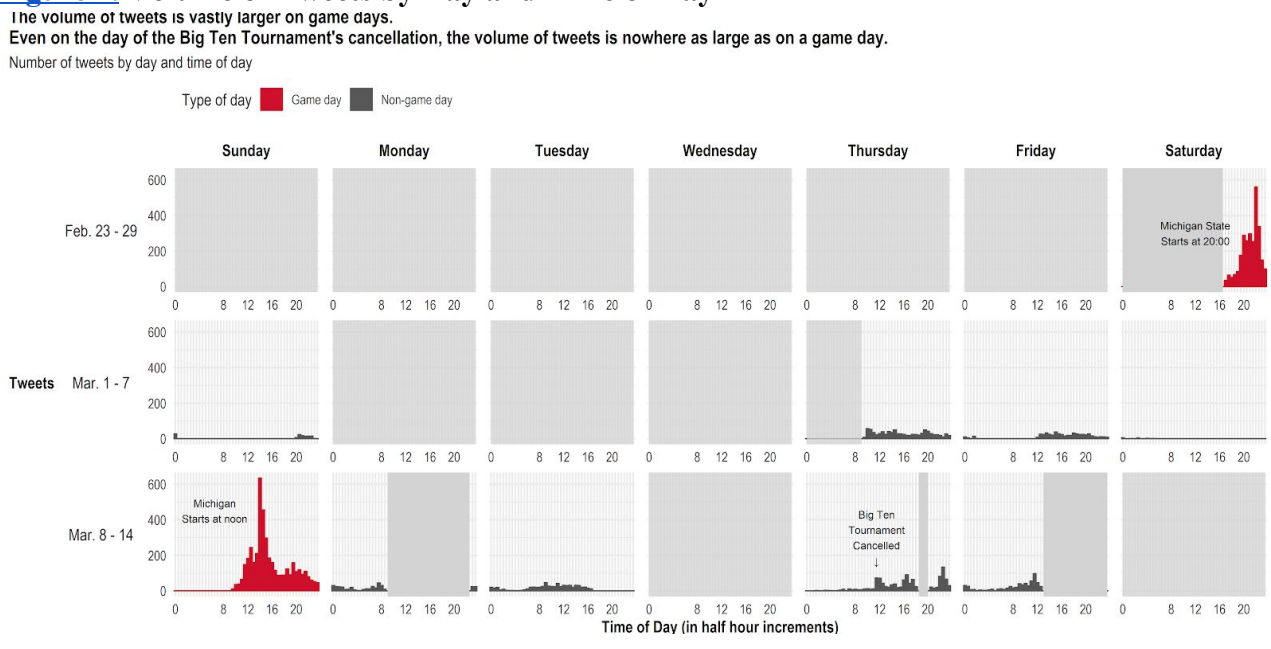
Descriptive Analysis of Tweets

Analysis of Trends in Tweet Volume

To understand how the volume of tweets changes across days, we counted the total number of relevant tweets produced during each half hour block covered by our streaming. The most striking pattern in tweet volume is the vast difference between tweets produced on game days compared to days where no UMD men's basketball games took place. (See Figure 1.) We found that 62% of the relevant tweets produced in the covered period came from days on which games played, despite the fact that games were held on only two of the ten days in which streaming occurred. Outside of game days, the greatest volume of tweets occurred on the day when the Big Ten Conference announced that the Big Ten Men's Basketball Tournament would be cancelled due to the coronavirus pandemic. Nonetheless, the number of tweets streamed from the day of that announcement (1,110) is only one quarter of the number of tweets from the day of the game against the University of Michigan four days prior (4,247).[2]

On each game day, the volume of tweets increased substantially in the couple of hours before the game began and increased even more sharply once the game got underway. In the period covered by our streaming, there was no clear, consistent pattern to tweet volume throughout the course of a day on which games were not held. The time at which tweet volume picked up varied between about 8:00am and 12:00pm, and did not seem to diminish for the night at a consistent time.

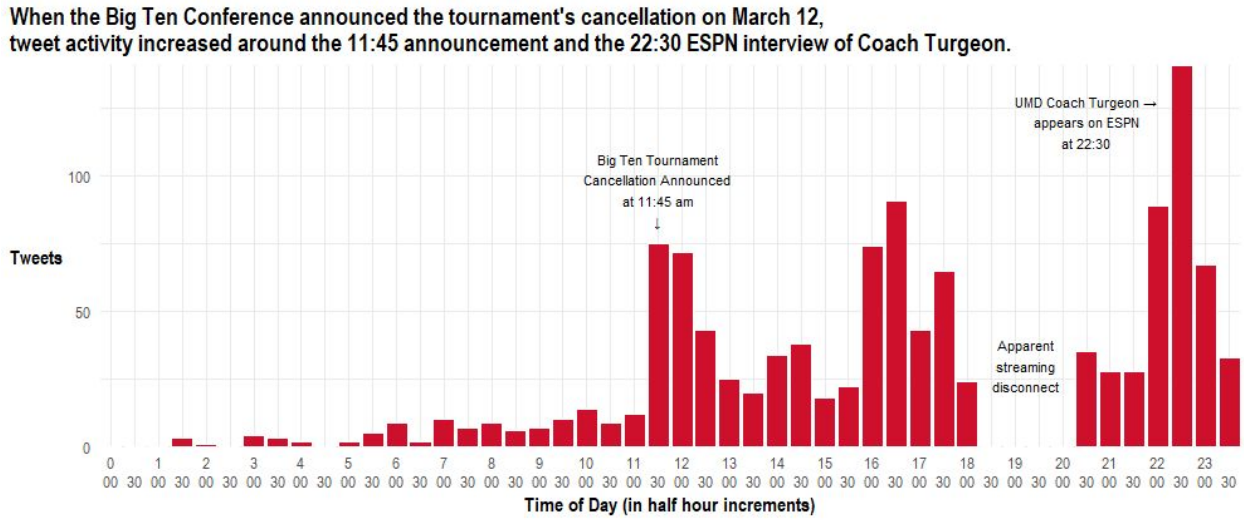
Figure 1. Volume of Tweets by Day and Time of Day



[2] Unfortunately, there was a streaming disconnect that lasted about two hours on the evening of March 12, so the total number of tweets from that day is not known precisely.

Interestingly, on March 12, when the Big Ten Tournament was cancelled, tweeting activity immediately increased after the 11:45am press release by the Big Ten Conference.[3] However, tweeting activity increased even more when the UMD men's basketball head coach, Mark Turgeon, appeared on ESPN for an interview to discuss the announcement.[4] (See Figure 2.)

Figure 2. Volume of Tweets by Time of Day on March 12



Analysis of Most Common Words

Using the qdap package in R, the team ran frequency analyses to identify the 30 most common words in the tokenized tweets. Figure 3 provides a list of these terms and their frequency. The majority of the identified terms were stop words, such as "the", "to", and "in", with minimal relevant meaning; however, a few relevant words, such as "terrapinhoops", "terps", "maryland", and "basketball" were still high on the list.

The team then removed stop words and other common meaningless terms (e.g., "https" and individual letters) and reran the frequency analyses. Figure 4 provides a list of the 30 most common words (and their counts) after meaningless terms were removed. These words are far more meaningful. Most words are related to UMD and basketball, but we also see the names of star players and coaches, as well as terms related to the teams that UMD played against (namely, Michigan and Michigan State). At least one meaningful term, "big", was a stop word that got removed through this process. "Big" was primarily seen in tweets referring to the "Big Ten", UMD's conference. But after stop words were removed, only the term "ten" remains in the list of most common words.

[3] Press release available as of March 23 at the following location: https://bigten.org/news/2020/3/12/mens-basketball-big-ten-conference-statement.aspx

[4] Interview available as of March 23 at the following location: https://www.nbcsports.com/washington/ncaa/mark-turgeon-ncaa-canceling-tournament-its-devastating-we-knew-it-was-going-way

Figure 3. 30 Most Common Terms in Raw Tokenized Tweets

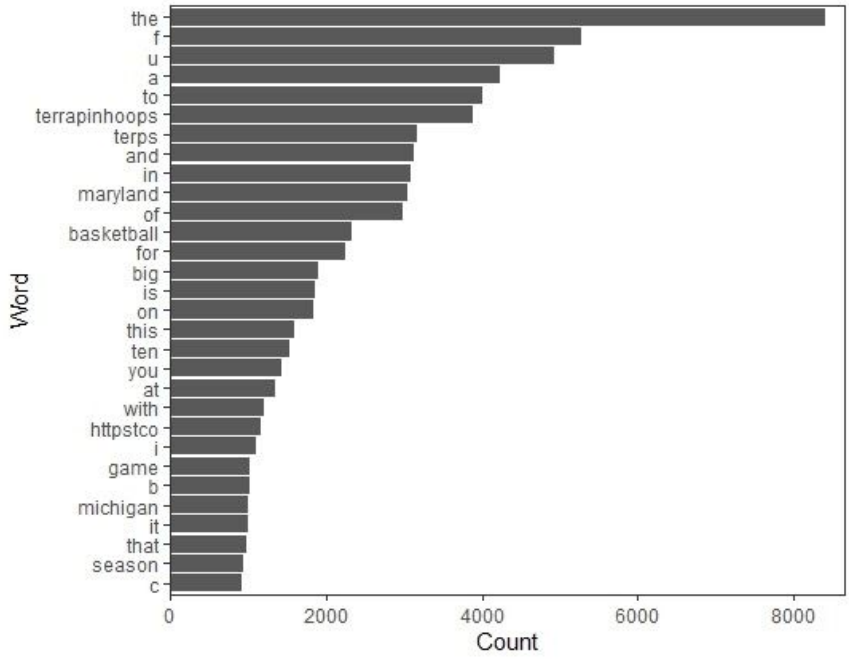
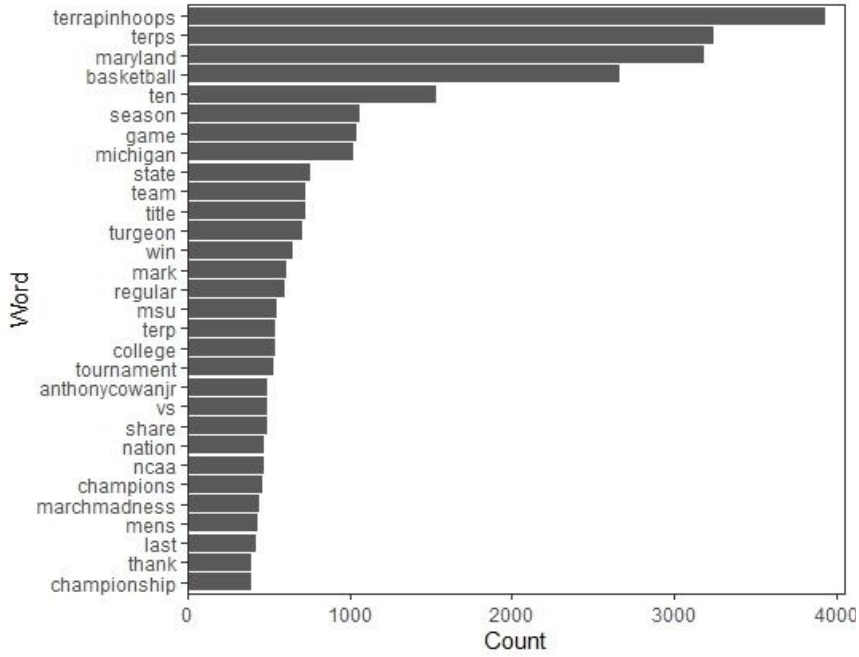


Figure 4. 30 Most Common Terms in Scrubbed Tokenized Tweets



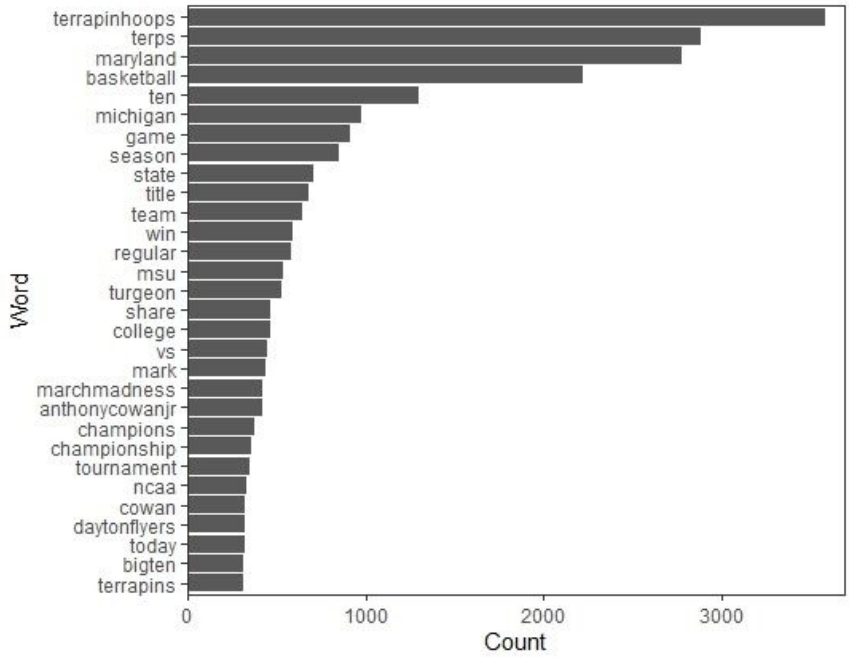
Before and After NCAA Cancellation

The team was curious to see if there was a difference in the most common words before and after the NCAA cancelled the Big Ten Tournament on March 12. Thus, we split the dataset of tokenized tweets by date--tweets submitted before March 12 and those submitted on or after

March 12. Figures 5 and 6 provide the list of common words and counts for each of these two datasets.

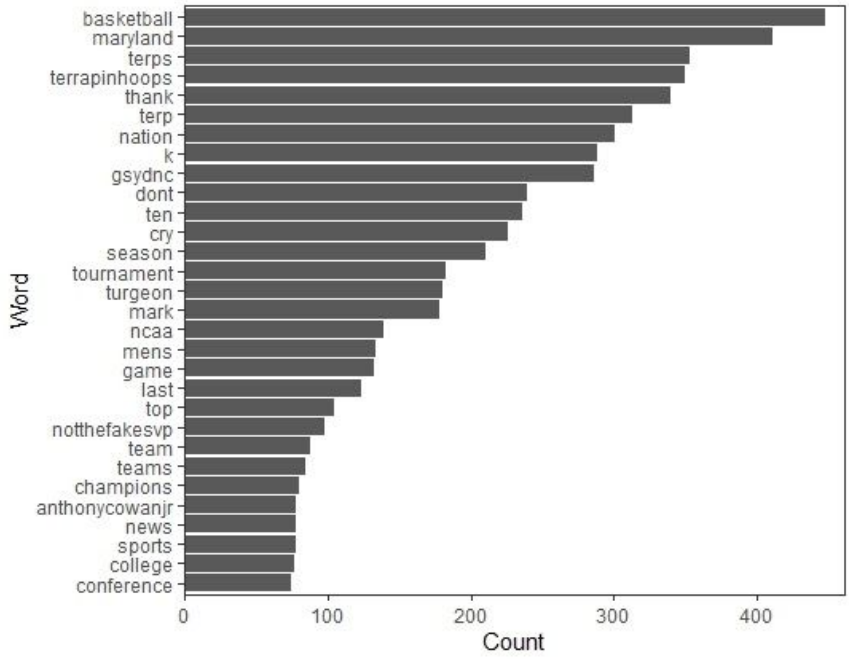
The “before NCAA cancellation” dataset consisted of 9,491 tweets. This represented the majority of all tweets collected, so the list of most common words in this dataset is quite similar to the list of most common words from the entire dataset. However, there is some variation in the order of the most common words.

Figure 5. 30 Most Common Terms in Tokenized Tweets Before Tournament Canceled



The “after NCAA cancellation” dataset consisted of 1,684 tweets. The list of most common words in this dataset includes some words that overlap with the previous lists, but there are also some new words (e.g., “thank”, “don’t”, “cry”, “news”, “sports”). The term “thank” was in a tweet from @TerrapinHoops that was retweeted 358 times. The message said simply, "Thank you, Terp Nation". Terms like “cry” are indicative of a significant sentiment shift after the tournament was cancelled.

Figure 6. 30 Most Common Terms in Tokenized Tweets After Tournament Canceled



Sentiment Analysis of Tweets

Sentiment Calculation

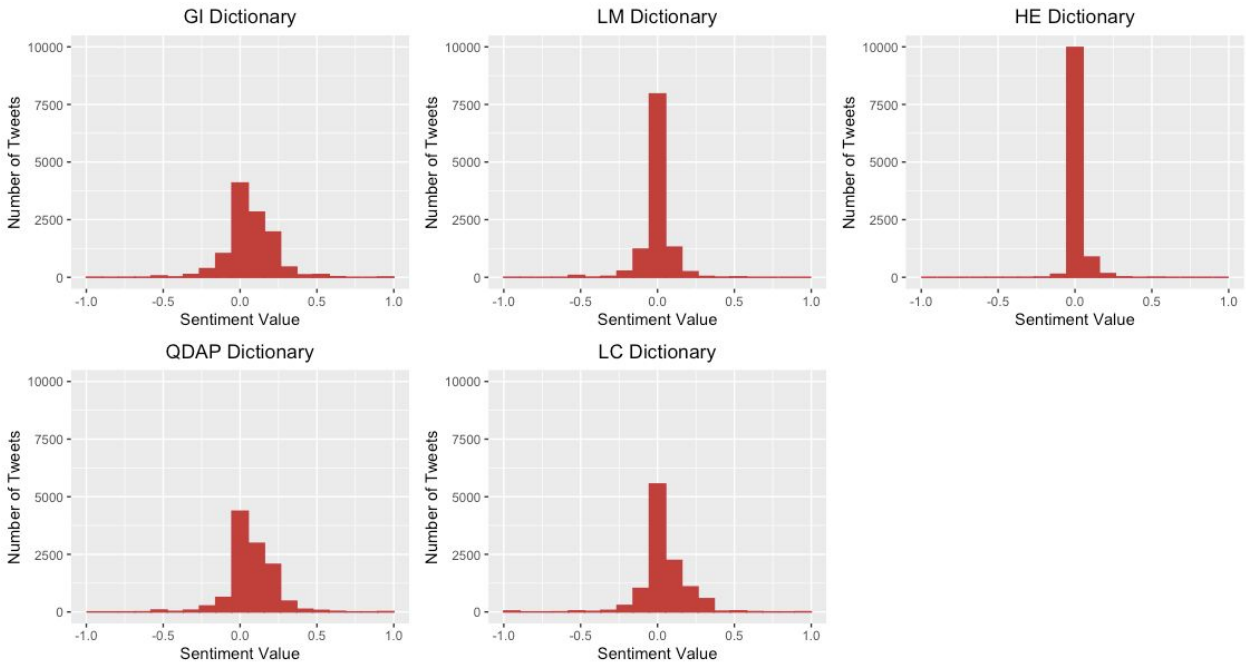
Each tweet was assigned sentiment scores using *analyzeSentiment* (from *SentimentAnalysis*) or with custom functions applied after using *tokens_lookup* (from *quanteda*). All default dictionaries from *SentimentAnalysis* were considered, in addition to Lexicoder and an emoji dictionary derived from the Novak et al. (2015) dataset. We chose to maintain the default sentiment calculation equation ((positives-negatives)/total) for all sentiment calculations.

Sentiment Selection

For our final analyses, we chose to use the GI dictionary. Major considerations in dictionary selection were dictionary construction, size, and (most importantly) performance in predicting sentiment in our dataset. A general-purpose dictionary was expected to better capture sentiment in this situation than a dictionary specifically constructed for financial purposes (LM, HE, and QDAP), which may include more irrelevant words or words with unintended meanings. For example, several words present in the financial dictionaries like “embezzlement” would not be relevant. Additionally, the GI dictionary was much larger than the HE and LM dictionaries (3,642 terms compared to 2,709 and 190, respectively) but smaller than QDAP (4,232 terms). The consequences of this are likely visible in the distributions of sentiments given by each dictionary (Figure 7). Dictionaries of similar size gave similar distributions to each other, with smaller dictionaries more hesitant to assign a non-neutral sentiment (having fewer words to work with), and larger dictionaries assigning stronger sentiments more liberally, though not always accurately.

We did not have sufficient time to fully evaluate the emoji dictionary, and the age of the source dataset in addition to difficulties with tokenizing continuous strings of emoji led us to eliminate it as a serious contender. That said, in limited testing, the Lexicoder+emoji combined dictionary did seem to sometimes provide useful improvements over Lexicoder by itself and other word-only methods, and could be a productive avenue if further research were to be done.

Figure 7. Distributions of Sentiments by Dictionary

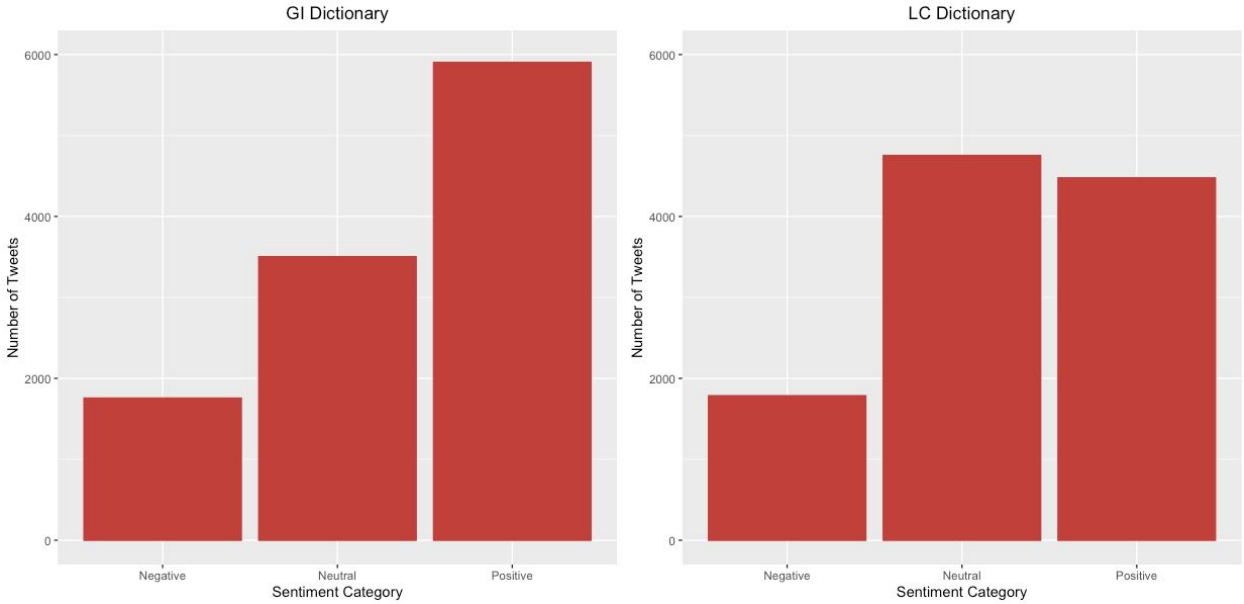


Comparing Two Dictionaries

To further compare two dictionaries specifically, we more closely examined the GI and Lexicoder (LC) dictionaries. We were interested in choosing between these two for the final analysis because of their advantages in hosting a general vocabulary and accounting for negated negatives and positives.

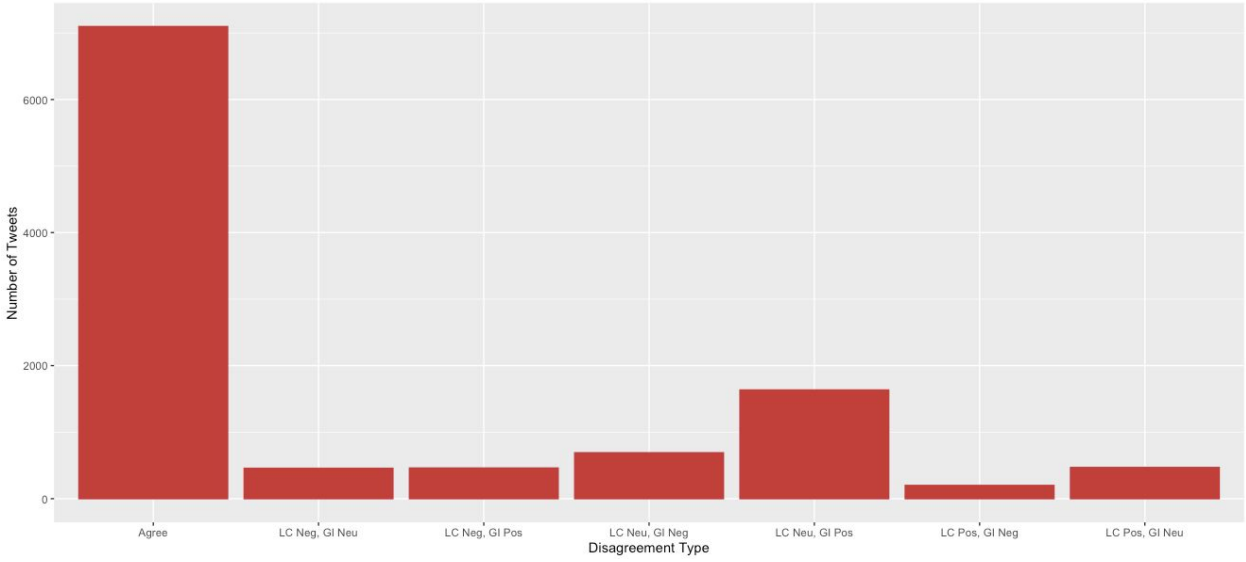
The average GI dictionary sentiment was about 0.066, while the average LC dictionary sentiment was about 0.038. Figure 8 shows the distributions of positive, negative, and neutral sentiments assigned by each dictionary. Note that the LC dictionary tended to assign neutral sentiments and the GI dictionary tended to assign positive sentiments.

Figure 8. Comparison of Sentiment Categories between GI and LC Dictionaries



The two dictionaries agreed on 7,248 of 11,379 tweets, or about 64% of the time (Figure 9). On tweets on which they did not agree, most cases were neutral assignments by the LC dictionary and either positive or negative by the GI dictionary. Based on manual inspection, usually the LC dictionary failed to detect an existing sentiment, although on occasion, the GI dictionary detected sentiments where there were none. Among discordant cases where the LC dictionary detected positive sentiment, the LC dictionary tended to be correct. Similarly, among discordant cases where LC detected negative sentiment, the LC dictionary also tended to be correct. As the first category is the largest, contested sentiments overall were usually best resolved by the GI dictionary's assignment.

Figure 9. Sentiment Disagreements between GI and LC Dictionaries



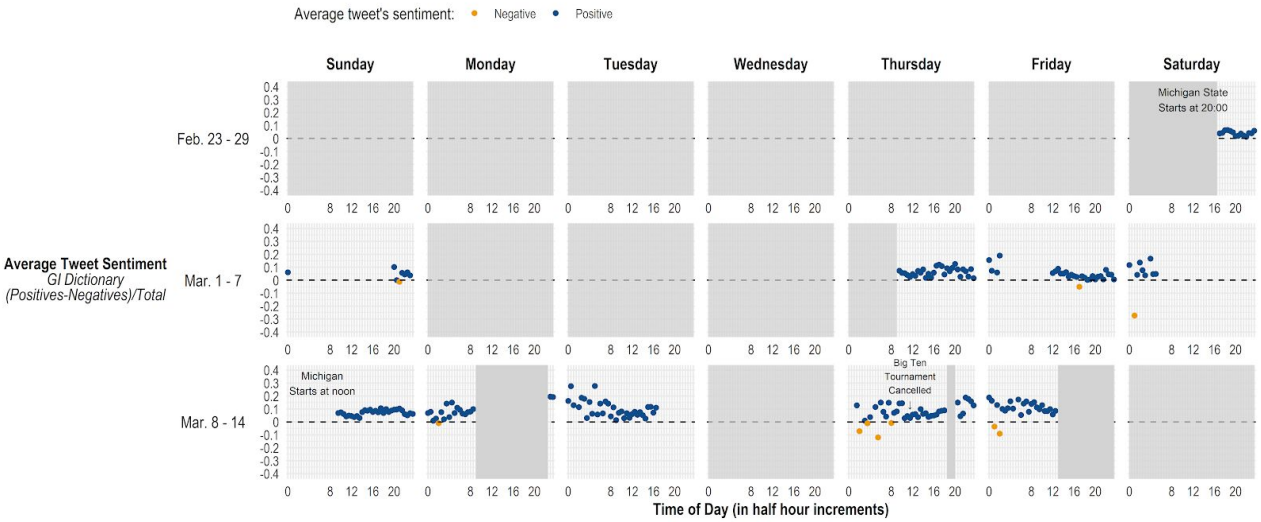
Analyzing Sentiment over Time

To understand how sentiment changed over the period of interest, we calculate the average tweet sentiment in each half-hour period covered by our streaming. For this analysis, we employed the GI dictionary and calculated each tweet's sentiment as the share of positive terms minus the share of negative terms in a tweet.

As can be seen in the chart below, sentiment was generally more positive than negative during the period of streaming. The only half-hour blocks where the average tweet's sentiment was negative occurred in early morning periods when few tweets were created; in short, the unusually negative sentiment of these periods can be explained as aberrations that arise due to the small number of tweets in those periods. The period of our streaming covered two days which we expected to have particularly marked changes in sentiment: March 8, when Maryland defeated Michigan in its final regular season game, and March 12, when the Big Ten Tournament was cancelled. The sentiment from these days can be seen in Figure 10 below, but we summarize these days in greater detail in the figures that follow.

Figure 10. Sentiment over the Period of Streaming

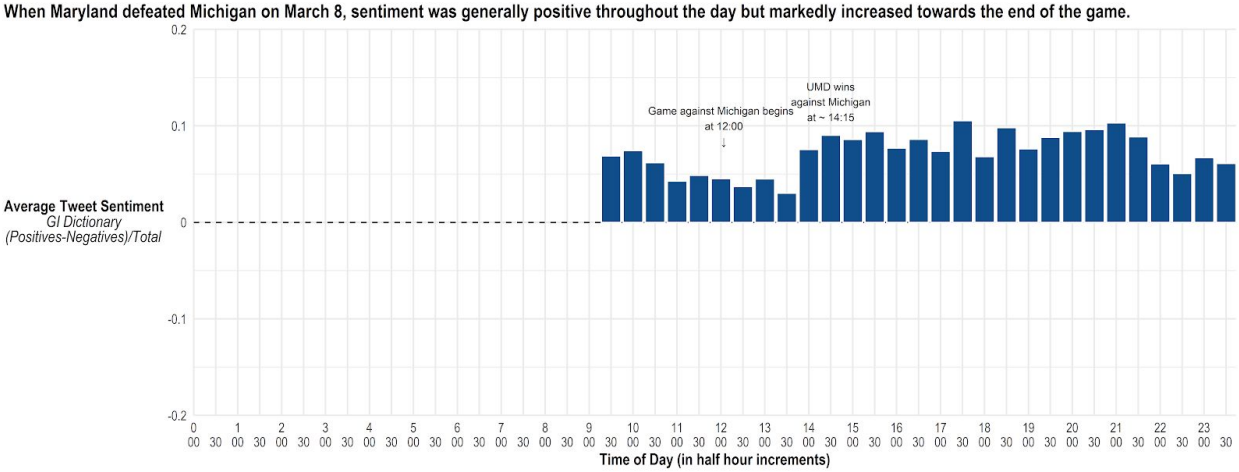
Over the period of streaming, sentiment was generally more positive than negative, even when Maryland lost to Michigan State and after the Big Ten tournament was cancelled.



On the final day of the regular season (March 8), sentiment was generally positive throughout the day. However, sentiment markedly increased in the final minutes of that day's game against Michigan and remained elevated for several hours following the team's victory in the afternoon at approximately 14:15. (See Figure 11.)

Figure 11. Sentiment on March 8, the Final Day of the Regular Season

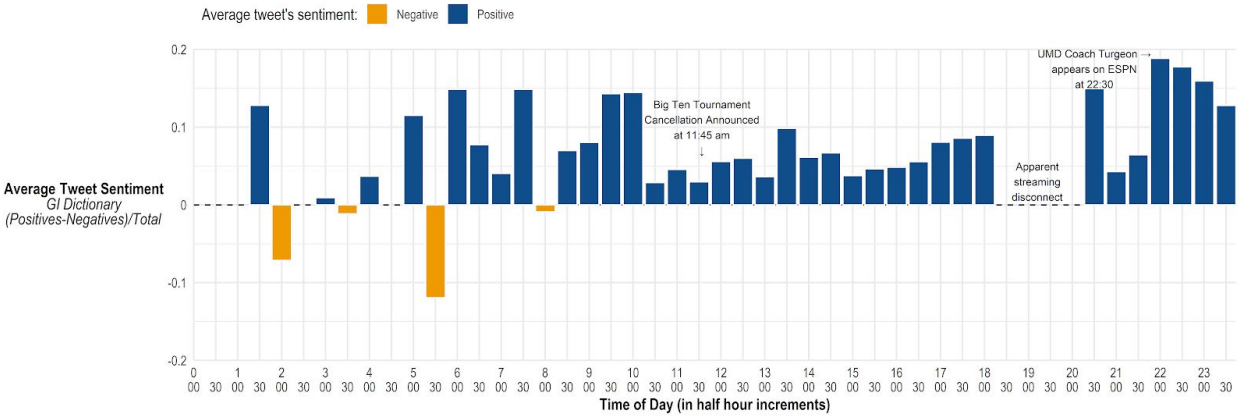
When Maryland defeated Michigan on March 8, sentiment was generally positive throughout the day but markedly increased towards the end of the game.



On the day when the Big Ten Tournament was cancelled, sentiment was largely positive throughout the day, even in the hours following the announcement. As in other days covered by our streaming, the only periods where the average tweet's sentiment was negative occurred in early morning periods when few tweets were created. Sentiment on March 12 peaked in the hour surrounding Coach Turgeon's interview on ESPN that evening. (See Figure 12.)

Figure 12. Sentiment on March 12, the Day when the Big Ten Tournament Was Cancelled

On the day that the Big Ten Conference announced the tournament's cancellation, sentiment peaked in the period surrounding Coach Turgeon's ESPN interview. Sentiment was generally more positive than negative, apart from a handful of early-morning periods with a small number of tweets.

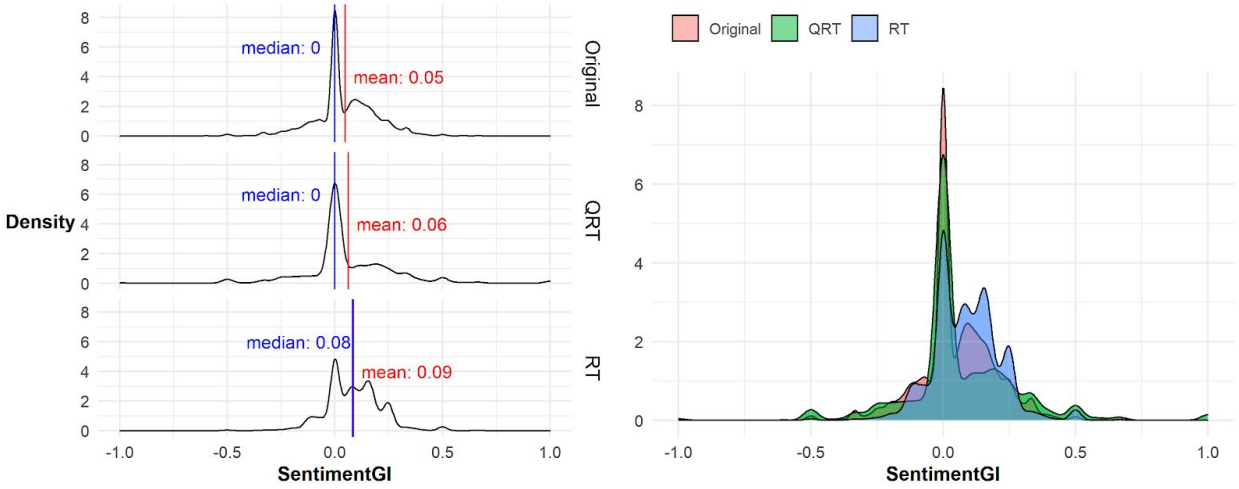


Comparing Tweet Types

We also considered sentiment across the three types of tweets in the dataset: original tweets (50% of all tweets in the data), retweets (40%), and quote retweets (10%). Retweets of quoted retweets were considered retweets and not quote retweets, as the retweeting user did not contribute any additional content in those cases. For quote retweets, we refer to the sentiment of the commentary on the quoted tweet, rather than the content of the quote itself, except where mentioned.

On average, straight retweets were the most positive-leaning (mean 0.09), followed by quote retweets (0.06), and finally original tweets (0.05). Compared with quote retweets and original tweets, retweets had the flattest distribution of sentiment, with a smaller spike at the neutral 0 point. (See Figure 13.)

Figure 13. Distributions of Sentiment by Tweet Type

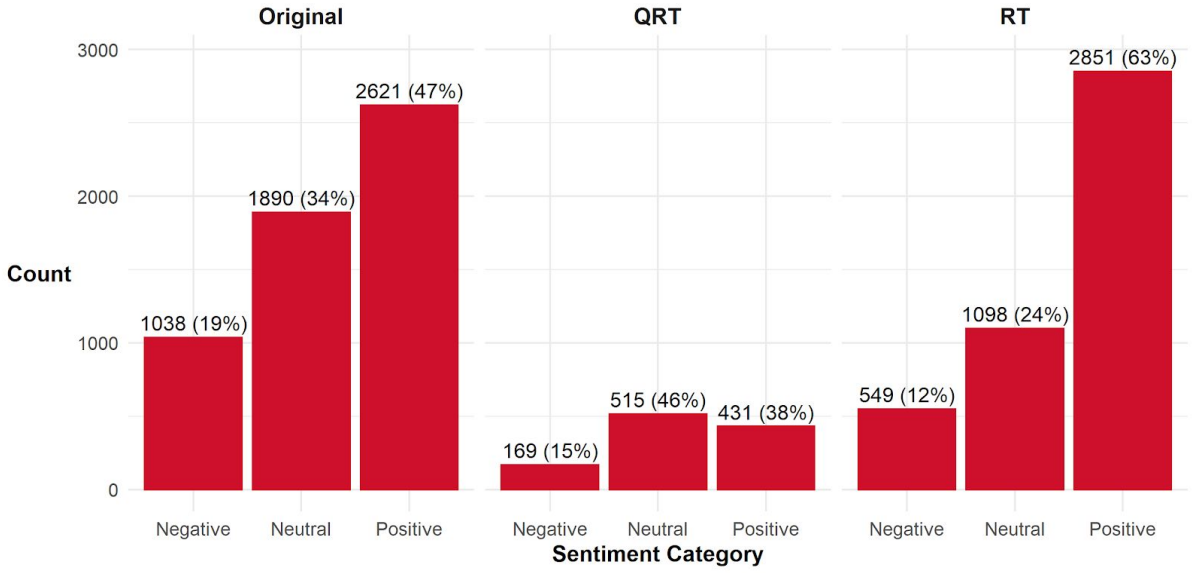


Splitting sentiment up into negative (<0), neutral (0), and positive (>0) categories, we see that a majority of retweets are positive, compared with just under half of original tweets, and only four in 10 quote retweets. (See Figure 14.)

Based on this, it is unsurprising that the most retweeted users in our data are all pro-Maryland or Maryland-centric sources, including @TerrapinHoops (used as a keyword in our streaming), @barstoolUMD, @MarylandonBTN, and @testudotimes. Even using a very crude mechanism, tweets originating from accounts with "Terps", "Terrapin", "UMD", "Maryland" or "Testudo" in their handles made up 31 percent of all retweets in our dataset.

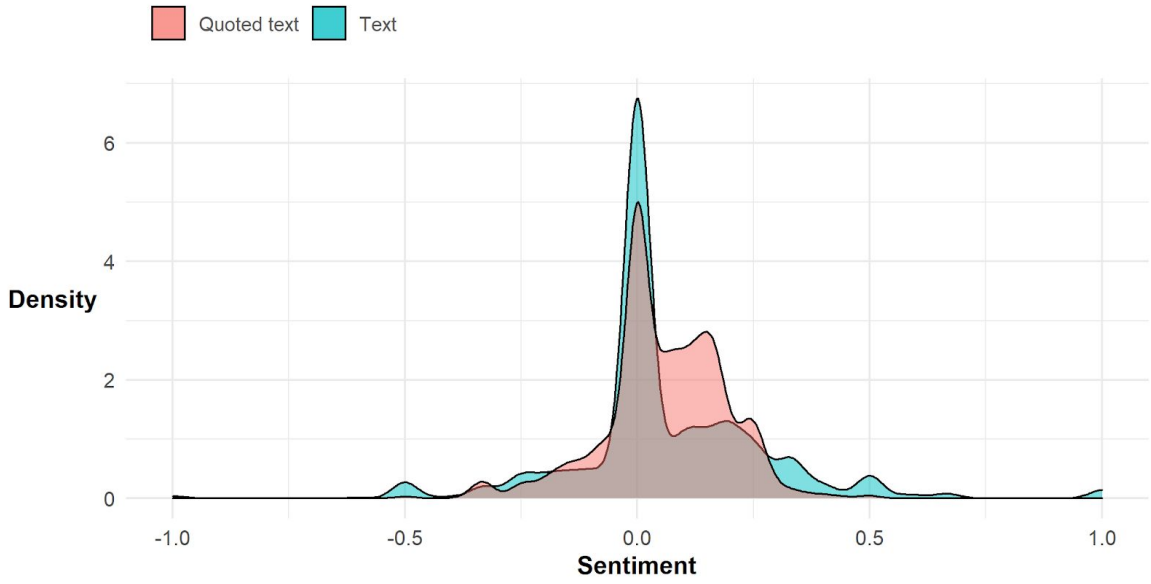
That said, the most retweeted tweets, rather than users, were somewhat less pro-Maryland. For example, the third-most-retweeted tweet was from @BarstoolMSU making fun of @BarstoolUMD following the MSU win over UMD. But because the entirety of this tweet (aside from the quote) was a video clip, sentiment was calculated at 0 - a major downside of all text-based sentiment analyses.

Figure 14. Frequency of Sentiment Categories by Tweet Type



As noted, for quote retweets, we focused on analyzing the main "reaction" content provided by the user doing the quote retweets, rather than on the quoted text. But we were interested in whether there was any relationship between the sentiment of the main tweet and the quoted tweet. Anecdotally, we saw many tweets quoting neutral, public announcement-type tweets and then reacting to them, but also considered that quoted tweets with a positive or negative sentiment could be amplified or negated with commentary.

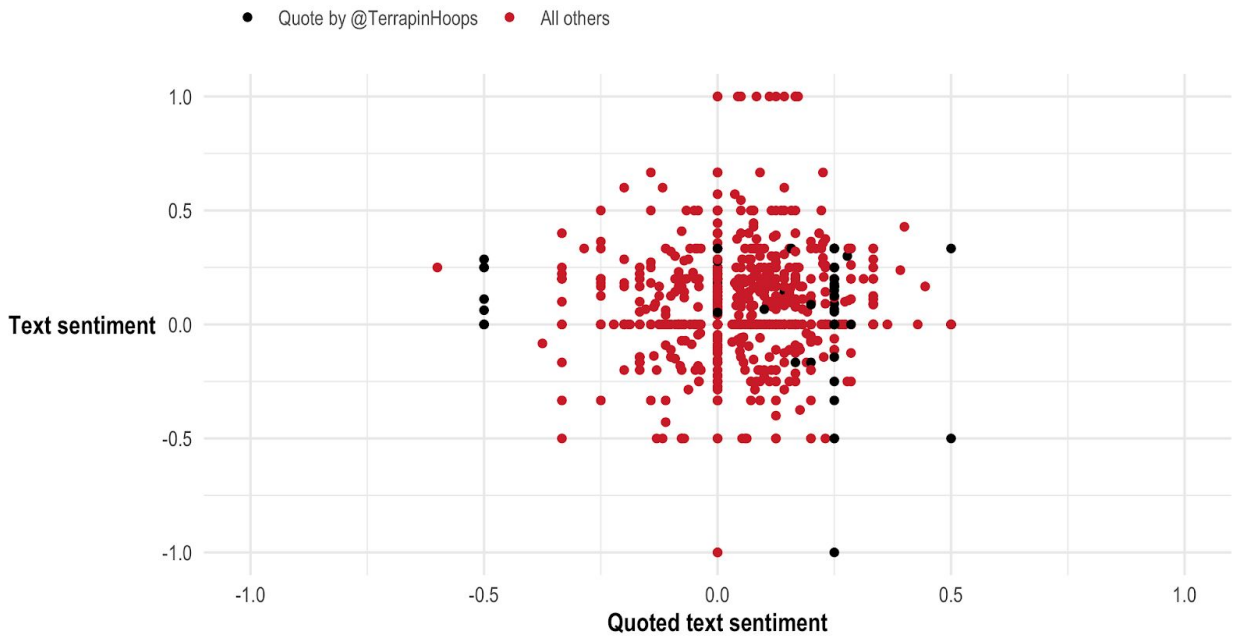
Figure 15. Sentiment Distributions of Text and Quoted Text



We found that on average the quoted tweets were more likely to be positive than the reaction text (Figure 15), and also that there appeared to be little relationship between the sentiment of a

quoted tweet and the additional commentary provided by quoting users (Figure 16). Figure 16 also identifies tweets quoting our most-quoted user, @TerrapinHoops, to emphasize the lack of trend even for a relatively popular, Maryland-centric account.

Figure 16. Sentiment of Quoted Text vs. Sentiment of Text



Why do we see higher positivity in retweets in our data? We hypothesize that this is likely due to the period in which we sampled, which covered highly publicized times of success for the Terps, including a win over Michigan and clinching a top spot in the Big 10 rankings.

As retweeting is a one-click operation, users wishing to simply endorse or rebroadcast these positive sentiments are more likely to do so than post an original tweet regarding the win or ranking. This is particularly true of those less engaged in college basketball or Maryland sports generally - these accounts would be unlikely to post anything related to these topics outside of these circumstances. (We could similarly guess that users who did make the effort to do a quote retweet would be unlikely to be adding equally positive reactions, as there's minimal return on the extra effort, thus explaining the positivity being more in line with original tweets.)

Had we begun our streaming earlier to catch more lead-up and reaction to the MSU loss, or sampled during a longer period generally, it's possible we would have seen less of a positive bent in our retweets. That'd also be true if the Maryland team had performed less well - we would assume there would be less enthusiasm and likely fewer "bandwagon"-type retweets.

Limitations

Limitations in Measurement

This project was impacted by a few limitations related to measurement. First, none of the dictionaries were capable of adequately analyzing the sentiment of our keywords and other colloquialisms in the tweets. For example, “fear the turtle” has positive sentiment about the team when used by UMD fans, but the word “fear” has negative sentiment in the GI Dictionary. Similarly, “Hell of a season!” has positive sentiment, but the word “hell” has negative sentiment in the GI Dictionary. Thus, using these generic dictionaries to assess the sentiment of sports discussions is likely to be insufficient.

Furthermore, the sentiment analysis misses a lot of context. A tweet from a rival team’s fan that taunts the UMD basketball team and a tweet from a UMD fan bemoaning the NCAA tournament cancellation may both involve negative sentiment, but there is a stark difference in the type and meaning of that negative sentiment. Such nuances are completely missed by this analysis. The use of images or video clips is also necessarily missing from the text-based analysis.

Finally, although our team attempted to improve the sentiment analysis by using an emoji dictionary, we were ultimately unable to do so. This was in part due to the fact that the emoji dictionary was out-of-date and seemed to be missing important emoji and in part due to lack of time and difficulties tokenizing continuous strings of emojis. However, based on our limited testing, analysis with the emoji dictionary in conjunction with Lexicoder seemed to improve results. Further research is needed in this area.

Limitations in Coverage

In addition, our research was limited by two key problems with implications for coverage error. First, our ability to collect tweets over the period of interest was hampered by connectivity issues. During time spans when our team programmed a computer to stream tweets, there were periods when streaming was inexplicably interrupted. For example, streaming was interrupted for all of March 11th, and on the crucial date of March 12, there was an approximately two-hour period in the afternoon when streaming was interrupted. Unfortunately, the *rtweet* package provided no diagnostic tools to indicate precisely when and for how long streaming was interrupted, and thus it is possible that some periods which we have interpreted to have an absence or low number of tweets in fact merely had an absence of streaming. This fact also implies that there may be measurement error in the form of downwardly biased estimates of the volume of tweets during periods when streaming interruptions occurred.

A second key limitation is that our classifications of whether tweets are related to the University of Maryland men’s basketball are unavoidably made with error. It seems plausible that there are many tweets about the team which were not captured by our list of keywords. For example, tweets about a specific team member would certainly be relevant but were only included in our streaming if they used particular subsets of the many broader keywords used to determine

relevance (‘terps’, ‘basketball’, etc.) Similarly, there are undoubtedly uncaptured tweets which feature only oblique or implicit references to the Maryland men’s basketball team which are difficult to capture using keywords (e.g. a tweet from an opposing team’s fan saying “I can’t believe they would play unfairly like that”).

Conversely, despite our best efforts to avoid streaming irrelevant tweets and to remove irrelevant tweets from our corpus, it seems almost certain that our corpus contains at least some irrelevant tweets. This would be due to similar issues as above (misspellings, slang, oblique references), including misspecification of relevant and irrelevant word combinations on our part.

Conclusions

Overall, sentiment in tweets leaned positive throughout the time period covered by our data collection. Positive sentiment remained even through negative experiences, such as the loss to Michigan State and the cancellation of the NCAA tournament. However, the analysis of most common words and a quick review of tweets suggests that the sentiment analysis may be missing important context. Although the sentiment analysis reveals overall positive sentiment after the Big Ten tournament was cancelled, the list of most common words includes the word “cry” and the most frequently retweeted statement (@TerrapinHoops’ “Thank you, Terp Nation”) was more bittersweet. Thus, this project demonstrates some of the strengths and the weaknesses of dictionary-based sentiment analysis: the method allows researchers to quickly determine the overall sentiment of a corpus of text and facilitates in-depth analysis of trends and sub-domain comparisons which would be prohibitively expensive to conduct using manual sentiment classification. Nonetheless, the expanded analytical scale enabled by the dictionary-based method comes with distinctive challenges of measurement.