New York University

# Classifying Tom Tango’s Tweets Using a Dictionary-Based Model:

## A study on Twitter Usage During the MLB Lockout

Jean An & Victoria Xie

DS-GA 1015: Text As Data

Prof. Andrew Halterman

May 11, 2022

## Introduction

On December 2, 2021, with the expiration of the 2016 Collective Bargaining Agreement (CBA), Major League Baseball (MLB) underwent a lockout, which marked the ninth work stoppage in the league's century-plus-long history. ("2021–22") This lockout—which will be referred to as "The Lockout" for the remainder of this paper—lasted 99 days and ended on March 10, 2022, when MLB and the Major League Baseball Player Association (MLBPA) finally agreed and signed a new CBA. With the previous MLB work stoppage occurring back in 1994 and 1995 due to a player's strike, The Lockout marked the first disruption to labor piece in MLB history that took place after the Internet had become prevalent, and it led to some interesting fallouts that did not exist in previous work stoppages. One of these fallouts is the usage of Twitter, which is the most widely used social media platform in baseball and other sports in the United States.

When MLB announced the institution of The Lockout, a significant change was made on its official website, MLB.com. All the names, images, and stories related to any player that is either on the 40-man roster of any of the thirty teams or a Major League free agent—essentially, those that are represented by the MLBPA—were immediately removed from the site. What remained were mainly stories about retired players, as well as a letter from MLB Commissioner Rob Manfred that detailed the reasons behind MLB's decision to implement The Lockout. According to reports, this action was taken "at the advice of legal counsel per the National Labor Relations Act." (Nesbitt, et al.) In addition, although not stated explicitly, one quickly realized in the upcoming days that everyone employed by MLB or other companies owned by MLB was no longer allowed to talk about any active players in any public setting, including on Twitter.

The list of people that were barred from discussing active players included beat writers and columnists of MLB.com, anchors of MLB Network, and those that worked for MLB Advanced Media (MLBAM), which powers the Statcast system that collects and distributes data related to the on-field gameplay of MLB games. One of these people is "Tom Tango"—also known as "TangoTiger"—who currently serves as the Senior Data Architect of MLBAM according to his own Twitter biography. (Note that "Tom Tango" is believed to be an alias and not the real name of said person, whose real name is not publicly known, but for the remainder of this paper he will be referred to as if Tom Tango is his real name.) Tango is a very active user of Twitter under the handle @tangotiger, which is a verified account. His tweets will be the source of data that is studied in this piece.

Although his real identity is not publicly known, Tango is a very well-known figure on the Internet for his ground-breaking work in sports analytics, particularly in baseball and ice hockey. It is believed that he has consulted for several MLB teams and is currently still a consultant for several teams in the National Hockey League (NHL). ("Tom Tango") Therefore, the majority of his Tweets are either on the topics of baseball or ice hockey. However, due to The Lockout, he was barred from tweeting about active players for those 99 days. This paper will use a dictionary-based method to classify a portion of his Tweets into three classes: Tweets about active baseball players, Tweets about retired baseball players, and Tweets about ice hockey. From there, a trend in his Tweets over the period of The Lockout can clearly be observed.

## Data

### 1. Data Source

The data was acquired using the Twitter API with Elevated Access and the *rtweet* package in R. Using the `get_timeline()` function from *rtweet*, a total of 3,248 of the most recent Tweets from @tangotiger were successfully acquired and stored in a data frame with 90 variables. After filtering out all the rows that were pure Retweets—meaning that none of the texts were written by Tango himself—there are 3,009 Tweets remaining, including 656 Quote Tweets. Note that these also include replies, whether it is part of his own Tweet thread or a reply to a Tweet from another Twitter account. The texts of these 3,009 Tweets are the data used for this paper.

### 2. Data Processing

In order to classify these Tweets into the three classifications mentioned at the end of the Introduction section, dictionaries for each classification must first be constructed. With the understanding that a major distinction among these classes is player names, and with prior knowledge that Tango always properly capitalizes all names and proper nouns, a decision was made to use dictionaries of names as the tool for classification. This part of the work relied on the spaCy package in Python, with the assistance of Professor Andrew Halterman. Using the trained pipeline `en_core_web_lg` from spaCy, all names from the texts were quickly identified and collected. This produced a dataset with 702 unique names—including first name-last name bigrams—after removing all hyperlinks, which were also identified as names.

The next step of data processing required the author to manually label these 702 unique names relying on domain knowledge on the subject and searches on Google. All names were given one of the three labels—“baseball”, “hockey”, and “neither”—or simply left blank because it is impossible to tell which of the three a given name belongs to. All names labeled as “baseball” were then further assigned a class of “active”, “retired”, or left blank if it is impossible to tell the difference or does not apply. A few examples are below:

- The most commonly mentioned name is “Gretzky”, which appeared 53 times, with an additional thirteen counts of “Wayne Gretzky”; these are clearly in reference to the legendary ice hockey player, and therefore is labeled as a “hockey” name.
- The name “Jeter” appeared eleven times, with an additional count of “Derek Jeter” and another count of “Derek Jeter’s”; these are clearly in reference to the legendary baseball player who played shortstop for the New York Yankees, and is therefore labeled as a “baseball” name and given a class of “retired”.
- The name “Ohtani” appeared five times; these are clearly in reference to Shohei Ohtani, the reigning American League Most Valuable Player who plays for the Los Angeles Angels, and is therefore labeled as a “baseball” name and given a class of “active”.
- The name “Keaton” appeared five times; these are clearly in reference to the renowned actor Michael Keaton, and is therefore labeled as “neither”.

While labeling, some terms that are clearly misidentified as names were also manually removed from the dataset, which resulted in 629 remaining rows. Further filtering out all names that were not labeled in any of the three classifications produced a dataset with 467 rows, each with a labeled name.

## Method

### 1. Model Selection

Due to the nature of this dictionary-based classification method, a decision was made to use the simplest method of prediction: using the function `str_detect()` from the *stringr* package in R. All existing supervised models that required some form of training set were quickly eliminated from consideration, as there isn't a set of Tweets from Tom Tango or any other user that were pre-labeled with the classifications needed in order to train a model to make predictions. Unsupervised models such as The Structural Topic Model or Latent Dirichlet Allocation were briefly considered, but the produced outcome would not necessarily meet the specific desired classifications.

### 2. Pre-Processing

With the data relying heavily on correctly capitalized names and the aforementioned selection of `str_detect()` as the prediction method, no pre-processing steps were needed. It is unnecessary to remove punctuations, numbers, or any other string that is unrelated to our labeled names, because the `str_detect()` function would simply ignore them. In addition, it is preferred to not change words to lowercase or apply any stemming, as these pre-processing steps might affect the model's ability to correctly identify names in the Tweets. Note that in this study, the classification data is collected from the same source which the prediction is applied to; thus, any further pre-processing made to the text source would most likely only deteriorate the accuracy of the predictions and would not provide any benefit.

### 3. Predictions

The first stage of predictions was made with three dictionaries that were constructed using the three different labels: "baseball", "hockey", and "neither". After applying these to the data frame with the 3,009 Tweets from Tom Tango, the model predicted 447 Tweets as "baseball" Tweets, 240 Tweets as "hockey" Tweets, and 150 Tweets as "neither" Tweets. Note that this means more than 2,000 Tweets were unclassified based on the predictions. This number should not be that surprising, however, as one would expect that the majority of his Tweets do not include any names at all, especially when many of these are simply replies to Tweets from other Twitter accounts and could consist of as few as one word.

The second stage of predictions was made with dictionaries that were constructed using the two different classes within the "baseball" label: "active" and "retired". After applying these to the data frame with the 3,009 Tweets from Tom Tango, the model predicted 122 Tweets as "active" and 221 Tweets as "retired". Note that the two numbers add up to 343 Tweets, which is less than the 447 total Tweets that were predicted to be "baseball" Tweets in the previous stage. This is logical, as all names that were assigned classes had to be labeled as "baseball", but not all names labeled as "baseball" were given a class. This means that these dictionaries must be subsets of the baseball dictionary used in the previous stage; hence, the predictions must also be subsets.

## Results

1. First Stage Predictions

The purpose of this study is to examine the distribution of Tom Tango's Tweets based on the classifications over time. Note that the time period covered by the acquired Tweets begins on November 11, 2021, and ends on May 9, 2022, which is a span of 180 days. Moreover, comparison between the time period covered by The Lockout (December 2, 2021, to March 9, 2022) and time periods outside of The Lockout would be particularly interesting. Figure 1 shows the aggregated counts of Tango's Tweets over the entire time period, grouped by the three labels of predictions performed in the first stage:

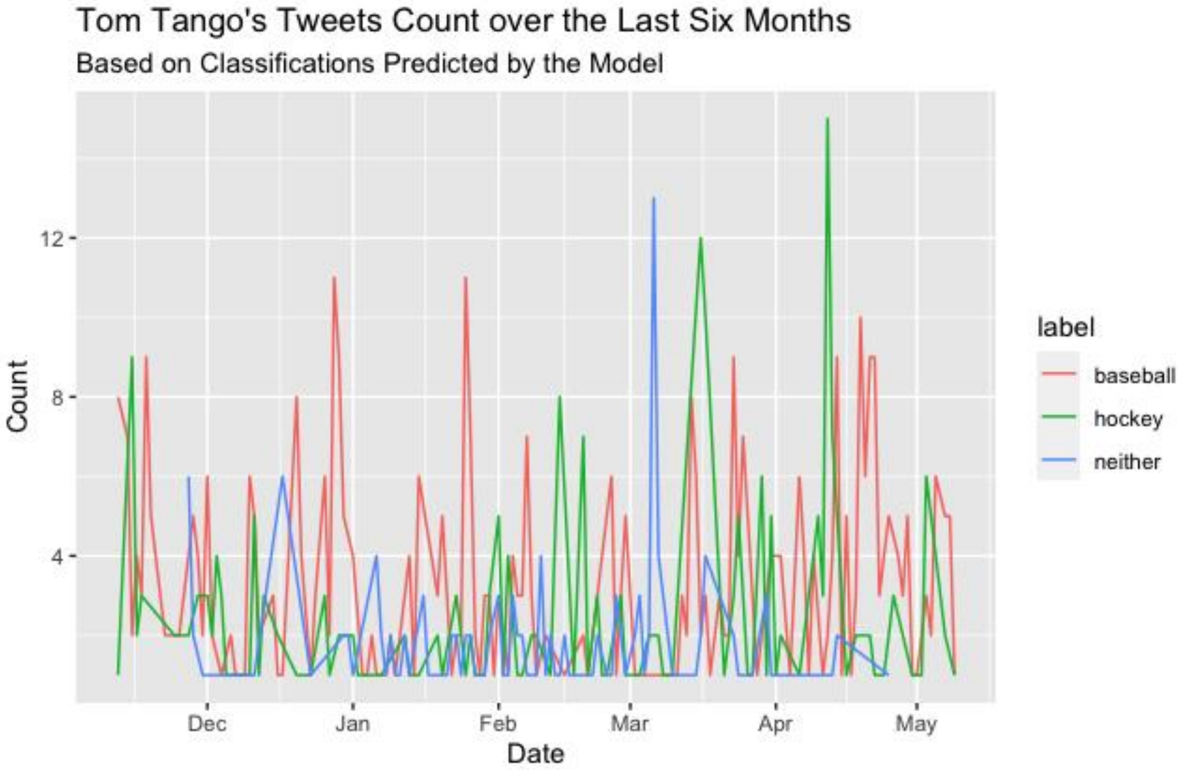


Figure 1

From the plot, some peaks can be observed, but the general trend is not very clear. It seems like Tango tweeted less about hockey and more about baseball in January, and gradually increased the number of hockey Tweets entering February, until a few noticeable peaks for hockey Tweets emerged in March and April. This is somewhat interesting, as the current NHL season started back in mid-October last year, and will continue into this upcoming June. In other words, the NHL season covers the entire span of this time period, so it is strange that the counts of hockey Tweets from Tango would vary so much from month to month.

Next, when focused on the period of The Lockout, when Tango should be barred from tweeting anything about active baseball players, one can observe an interesting outcome in Figure 2, particularly around New Year's and in January, that was mentioned in the previous paragraph: Tango actually tweeted more about baseball than hockey in this time period.

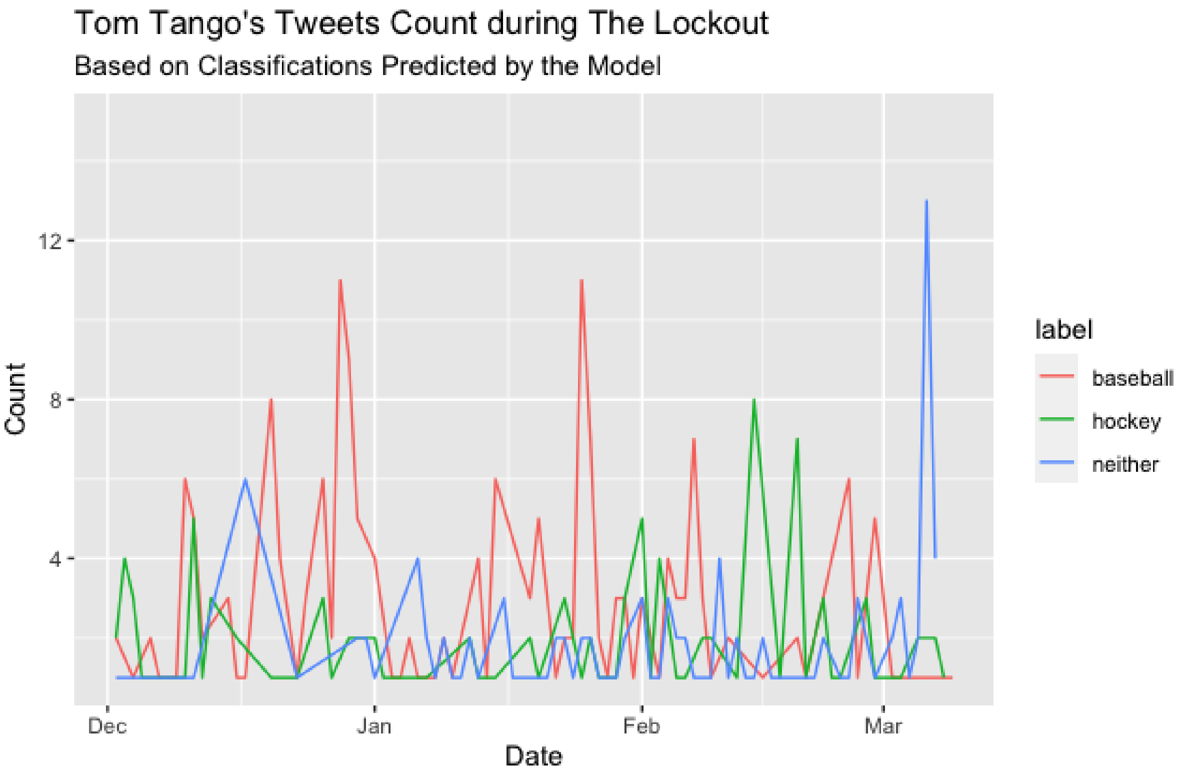


Figure 2

During a period where Tom Tango cannot tweet about active baseball players, which is supposed to be a major source of tweet-worthy topics given his position, he not only did not replace the void with more "hockey" Tweets, but he essentially focused on tweeting about baseball entirely. Given the knowledge of this time frame and the premise of what he is and isn't allowed to tweet about, however, one should be able to make a bold prediction: Tom Tango tweeted about retired baseball players.

2. Second Stage Predictions

Further looking into the baseball Tweets using the "active" and "retired" classifications in the prediction model yields results that confirm this bold prediction. The Baseball Hall of Fame announced the voting results of the Baseball Writers' Association of America (BBWAA) ballot for the Class of 2022 Hall of Fame nominees on January 25, 2022. In addition, various information regarding the voting was released on Twitter from the start of December until the results were announced and legendary slugger David Ortiz of the Boston Red Sox became the only player inducted into the Hall of Fame via the BBWAA ballot this year. (Castrovince) Figure 3 shows the aggregated counts of Tango's "baseball" Tweets over the entire time period, grouped by the two labels of predictions performed in the second stage:

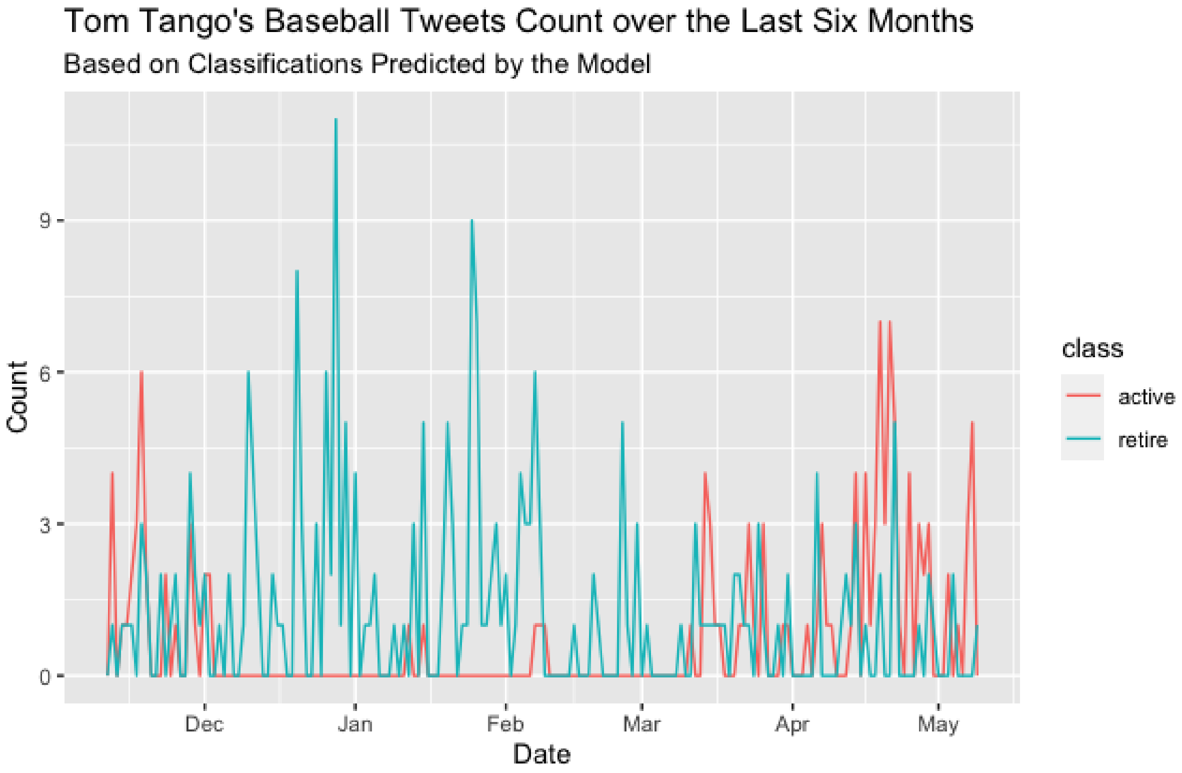


Figure 3

From the plot, one can certainly observe the clear trends demonstrated by the two different classifications. First, it is clear that Tom Tango did tweet a lot about retired baseball players in December and January, which matched the time frame associated with the Baseball Hall of Fame announcements discussed in the previous paragraph. Second, it is clear that Tom Tango rarely—if at all—tweeted about active baseball players during the time period of The Lockout, from the start of December to early March. This can be confirmed in Figure 4, which shows the same distribution of predictions but is limited to only the period of The Lockout. Lastly, in time periods outside of The Lockout (i.e., prior to December and beginning mid-March), Tom Tango actually tweeted about active and retired baseball players fairly evenly, and the distribution of the counts is clearly similar.

While Tango almost exclusively tweeted about retired players in his baseball Tweets during The Lockout, we still see that he occasionally tweeted about active players during this period. This is strange, as prior knowledge on the subject suggests that he is not allowed to tweet about active baseball players at all during this time period. However, since there are only very few of these Tweets, it allows the ability to simply examine these occurrences one by one.

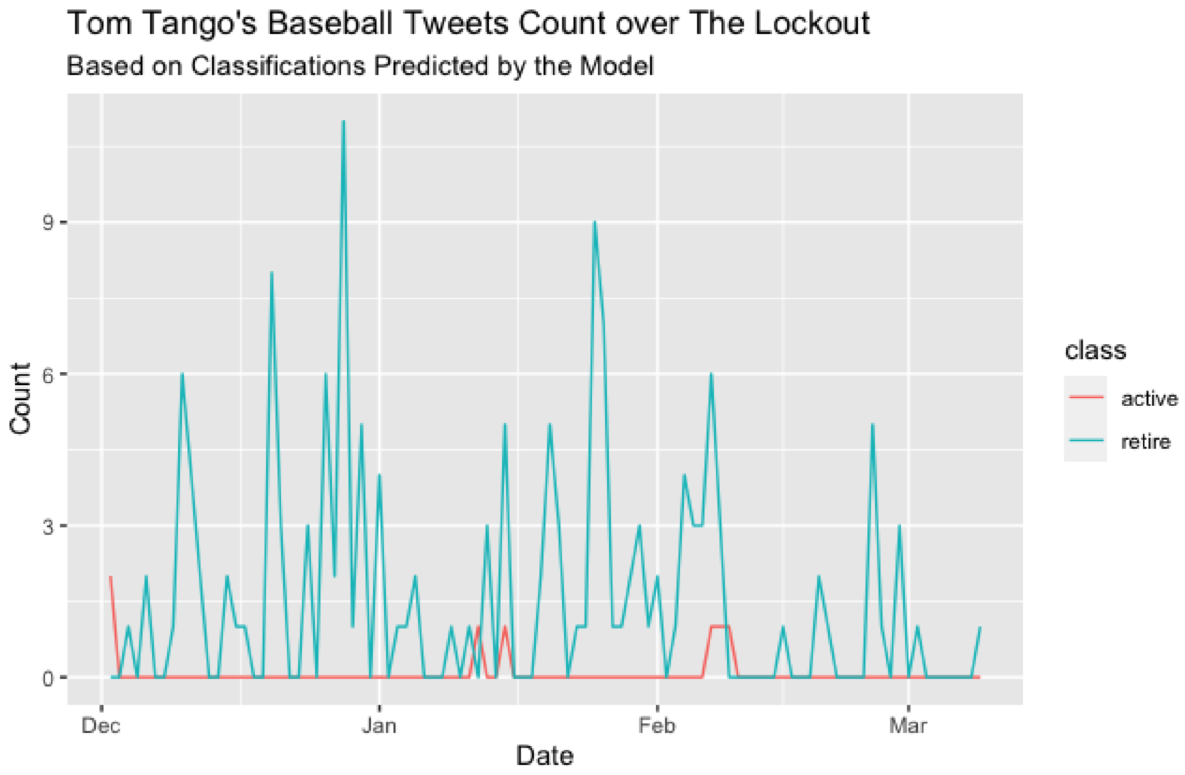
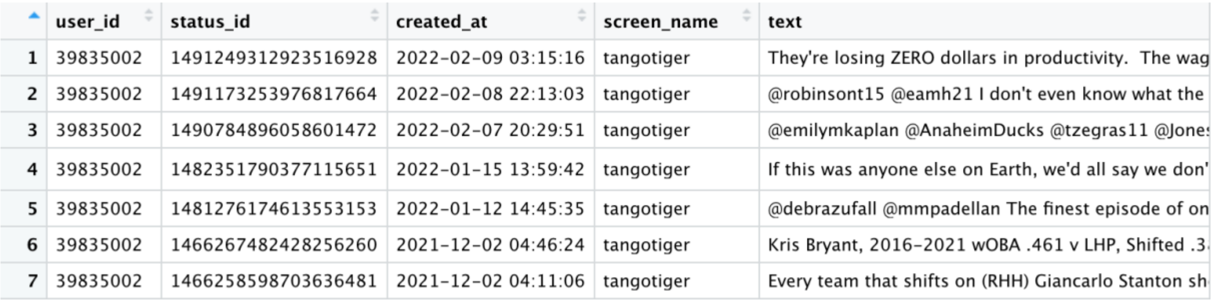


Figure 4

3. <u>Incorrect Predictions</u>

Looking at all the Tweets with a predicted "active" class and within the date range of The Lockout, seven Tweets were identified. Figure 5 shows a partially visible screenshot of these seven Tweets, presented in a data frame in RStudio.

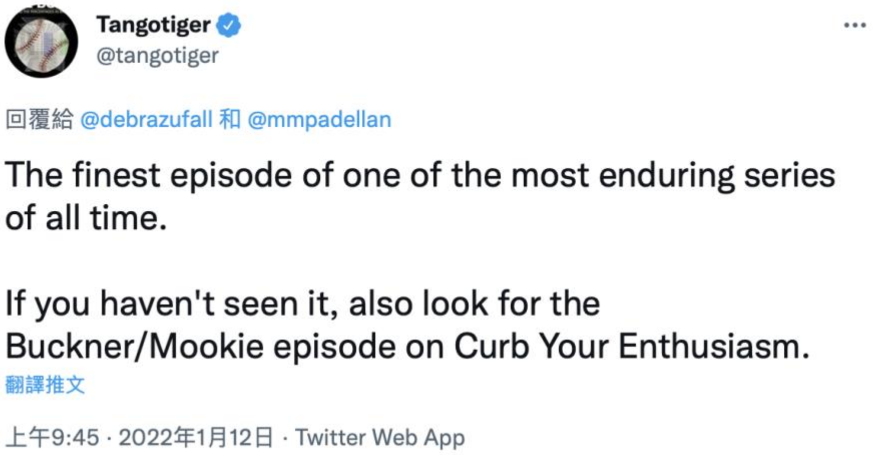


| | user_id | status_id | created_at | screen_name | text |
|---|---|---|---|---|---|
| 1 | 39835002 | 1491249312923516928 | 2022-02-09 03:15:16 | tangotiger | They're losing ZERO dollars in productivity. The wag |
| 2 | 39835002 | 1491173253976817664 | 2022-02-08 22:13:03 | tangotiger | @robinsont15 @eamh21 I don't even know what the |
| 3 | 39835002 | 1490784896058601472 | 2022-02-07 20:29:51 | tangotiger | @emilymkaplan @AnaheimDucks @tzegras11 @Jones |
| 4 | 39835002 | 1482351790377115651 | 2022-01-15 13:59:42 | tangotiger | If this was anyone else on Earth, we'd all say we don' |
| 5 | 39835002 | 1481276174613553153 | 2022-01-12 14:45:35 | tangotiger | @debrazufall @mmpadellan The finest episode of on |
| 6 | 39835002 | 1466267482428256260 | 2021-12-02 04:46:24 | tangotiger | Kris Bryant, 2016-2021 wOBA .461 v LHP, Shifted .3 |
| 7 | 39835002 | 1466258598703636481 | 2021-12-02 04:11:06 | tangotiger | Every team that shifts on (RHH) Giancarlo Stanton sh |

Figure 5

According to the `created_at` column produced directly from the Twitter API and collected using the rtweet package, the timestamps in which these seven Tweets are created do match the requirement. Two of these Tweets were created on December 2, 2021; two more were created in January; and the three remaining Tweets were created in February. From Figure 5, it is clear that the two Tweets created on December 2 mentioned Giancarlo Stanton and Kris Bryant, both active MLB players. However, when one visits these two Tweets directly on Twitter using the `status_id`, one would quickly realize these Tweets were posted on December 1, 2021, at 23:11 and 23:46 Eastern Standard Time, respectively.

The observed five-hour time difference is exactly the time difference between Eastern Standard Time (EST) and Coordinated Universal Time (UTC), which is the timezone used by the Twitter API when returning datetime values. ("Timezones") The adjusted timezone eliminates these two Tweets from the time period of The Lockout, therefore are not considered incorrect predictions. The five remaining Tweets require further examinations. First, the Tweet created on January 12, 2022:



Tangotiger
@tangotiger

回覆給 @debrazufall 和 @mmpadellan

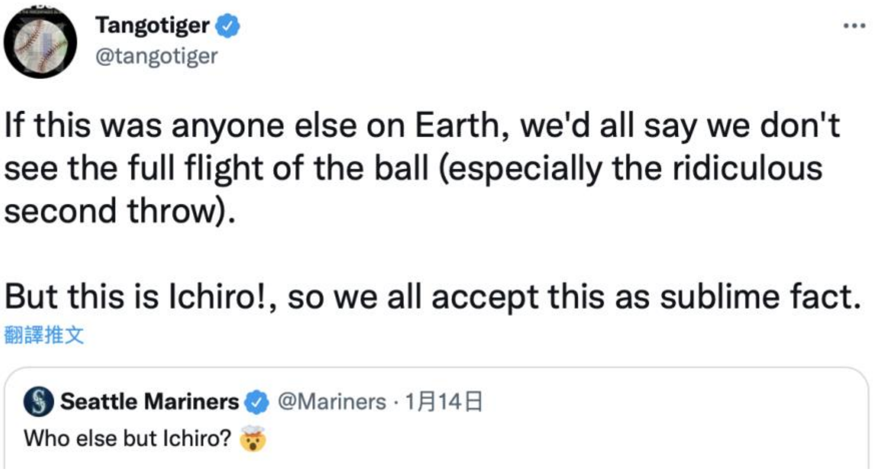
The finest episode of one of the most enduring series of all time.

If you haven't seen it, also look for the Buckner/Mookie episode on Curb Your Enthusiasm.

翻譯推文

上午9:45 · 2022年1月12日 · Twitter Web App

In this Tweet, Tango mentions the name "Mookie", which is a name labeled as "baseball" and given the class "active" because of Mookie Betts, the right fielder of the Los Angeles Dodgers. However, when one further examines this Tweets thread, one can clearly tell that this is a thread about Seinfeld, the beloved comedy TV series, and the mentioning of the name "Mookie" here is actually in reference to former New York Mets outfielder Mookie Wilson, instead of Mookie Betts. Clearly, this Tweet is mis-predicted due to a name shared by a famous active baseball player, and a lesser-known retired baseball player. Next, the Tweet created on January 15, 2022:



Tangotiger
@tangotiger

If this was anyone else on Earth, we'd all say we don't see the full flight of the ball (especially the ridiculous second throw).

But this is Ichiro!, so we all accept this as sublime fact.

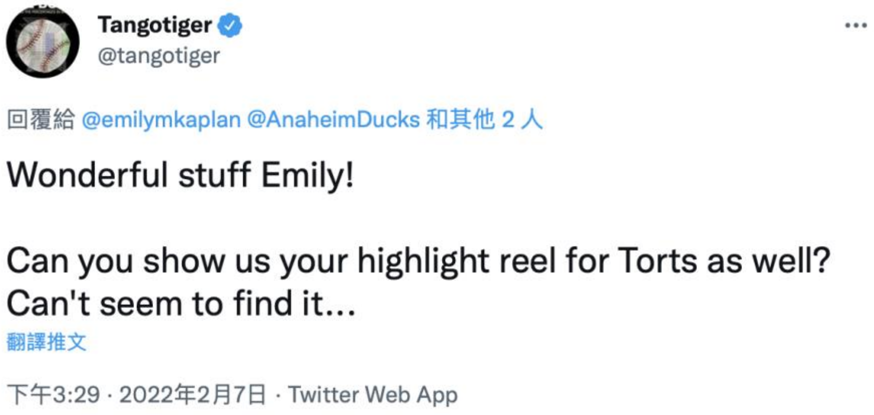
翻譯推文

Seattle Mariners @Mariners · 1月14日

Who else but Ichiro? 🤯

The only name mentioned in this Quote Tweet is Ichiro, who is a retired baseball player. This makes it strange that this Tweet was predicted as an "active" class by the model. However, after going back and looking at the labels, the author quickly realized that the name "Ichiro" was indeed incorrectly given the class "active" instead of "retired". It turns out that this particular instance was caused by a simple mistake made when manually labeling the names, and not a poor performance by the prediction model.

Next, the following Tweet was created on February 7, 2022:



Tangotiger
@tangotiger

回覆給 @emilymkaplan @AnaheimDucks 和其他 2 人

Wonderful stuff Emily!

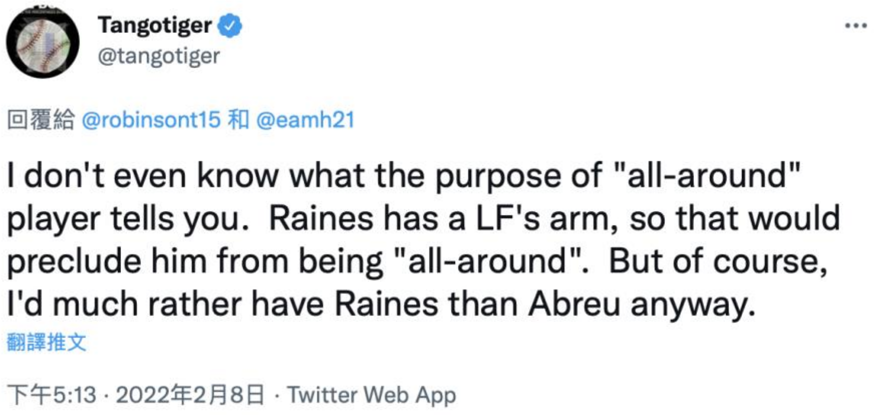
Can you show us your highlight reel for Torts as well? Can't seem to find it...

翻譯推文

下午3:29 · 2022年2月7日 · Twitter Web App

On the surface, it is difficult to understand how this was misclassified as an active baseball Tweet. The names mentioned in the Tweet include "Emily" and "Torts", but neither are labeled names within the classification data. In addition, since Emily Kaplan (@emilymkaplan) is a hockey reporter of ESPN, and the Anaheim Ducks (@AnaheimDucks) is one of the teams in the NHL, this particular Tweet by Tango was certainly a hockey Tweet.

Upon further examination, it appears that Max Jones (@Jones_Max19), a player of the Ducks, is also mentioned in this Tweet thread. However, the name "Max" is labeled as an active baseball name in reference to Max Scherzer, an active pitcher who has won multiple Cy Young Awards. Clearly, this Tweet is mis-predicted due to a name shared by a famous active baseball player, and a lesser-known hockey player. Next, the following Tweet was created on February 8, 2022:



Tangotiger
@tangotiger

回覆給 @robinsont15 和 @eamh21

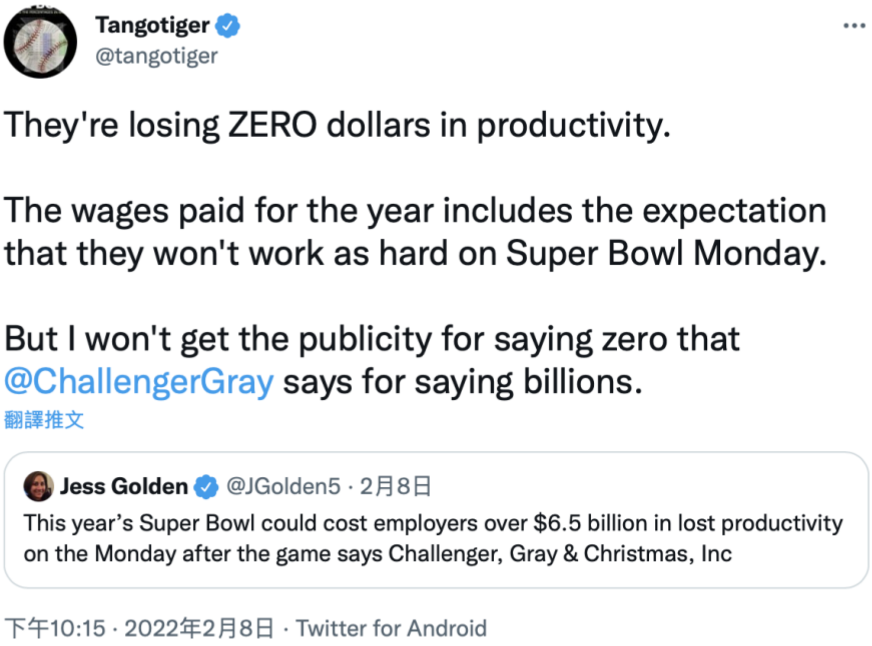
I don't even know what the purpose of "all-around" player tells you.  Raines has a LF's arm, so that would preclude him from being "all-around".  But of course, I'd much rather have Raines than Abreu anyway.

翻譯推文

下午5:13 · 2022年2月8日 · Twitter Web App

This Tweet mentions two names, Raines and Abreu. The name Raines is clearly in reference to Hall of Fame outfielder Tim Raines—whom Tom Tango has many times mentioned to be his favorite baseball player growing up—while the name Abreu is referring to retired outfielder Bobby Abreu. However, the name "Abreu" was given a class "active" due to current Chicago White Sox first baseman Jose Abreu. Clearly, this Tweet is mis-predicted due to a last name shared by a famous active baseball player and a famous retired baseball player. This error can partially be charged to the manual label mistake by the author.

Lastly, the following Tweet was created on February 9, 2022:



Tangotiger @tangotiger

They're losing ZERO dollars in productivity.

The wages paid for the year includes the expectation that they won't work as hard on Super Bowl Monday.

But I won't get the publicity for saying zero that @ChallengerGray says for saying billions.

翻譯推文

Jess Golden @JGolden5 · 2月8日

This year's Super Bowl could cost employers over $6.5 billion in lost productivity on the Monday after the game says Challenger, Gray & Christmas, Inc

下午10:15 · 2022年2月8日 · Twitter for Android

In this Quote Tweet, Tango did not mention any names; however, he did tag the account @ChallengerGray, which belongs to the executive outplacement firm Challenger, Gray & Christmas. ("Challenger") This explained the misclassification, as the name "Gray" is labeled as an "active" class baseball name because there are multiple current MLB starting pitchers with that last name: Sonny Gray, Jon Gray, and Josiah Gray. Clearly, this Tweet was not about baseball at all; it is simply an instance where there is an overlap between a last name of multiple MLB pitchers and the name of a firm.

After carefully examining these five Tweets that were misclassified as "active" baseball Tweets, it can be concluded that none of these were actually about active baseball players. Tom Tango did in fact follow the orders and did not tweet anything about active players during The Lockout. However, he did enjoy tweeting about retired baseball players over this time frame, especially given the discussion surrounding the Baseball Hall of Fame announcements. A simple dictionary-based model has successfully captured this trend, which is easily observed in Figure 3 and Figure 4.

## Further Discussions

### 1. What could have been

The original intent of this study is to examine Tom Tango's Tweets distribution during The Lockout in comparison to the same data range (December 2 to March 9) from past winters. This is because the author believes that studying the same time frame of a year would better capture the natural distributions of Tweets, given certain events such as the Baseball Hall of Fame announcement occurring around the same time each year. The MLB season and the NHL season are also generally aligned to certain months in a year, so one would expect there to be some fluctuation in Tweet class distribution given these preconditions.

However, this decision was not achievable due to the limited access to the Twitter API. The author was able to apply and attain the Twitter API with Elevated Access, which allowed the acquisition of the data necessary to perform the tasks described in this study. To obtain more Tweets from Tom Tango's timeline and going further back would require the Twitter API with Academic Research Access, which the author does not possess. The author has also attempted to scrape Tweets from Tom Tango's timeline using the TWINT package in Python, but the returned results were actually less ideal than the data acquired through the Twitter API with Elevated Access.

In addition, some other methods of constructing the classification dictionaries were also considered. In this study, the names that are used to construct the dictionaries were taken directly from Tom Tango's Tweets; one could essentially interpret this as using the same dataset for both training and testing purposes, which would obviously lead to high accuracy. A way to generate a dictionary of names that is more representative of these classifications would be to scrape rosters from MLB.com and NHL.com to obtain the names of all active baseball and hockey players. One could also scrape from BaseballHall.org to obtain names of all Hall of Famers, but it is worth noting that Hall of Famers are only a small subset of all retired players. For example, Ichiro and Bobby Abreu, who have been mentioned in the Results section, are retired players but have not (yet) been elected to the Hall of Fame. A similar problem is also present for retired hockey players, many of which are often mentioned in Tango's Tweets.

Another method considered is to simply scrape articles from MLB.com, BaseballHall.org, and NHL.com to construct dictionaries by collecting all words used in these pieces. The sets of words would have to be disjoint, and do not have to be limited to just names. This method was not chosen in the end due to the complexity of having to scrape many webpages that aren't labeled with sequential, numerical IDs that could be easily parsed.

### 2. What could be done

With Academic Research Access, one should be able to complete a study based on the original intent described in the previous section. Otherwise, based on this current study, a simple way to expand the research and increase the classification count and accuracy would be to extend the classification dictionaries to include more than just names. In addition to names, there are many more terms that are relatively unique to baseball or ice hockey that could aid the classification power of the model. For example, simply adding the word "baseball" and "hockey" would probably increase the counts of predicted Tweets under the current model. Furthermore, terms like "home run", "pitcher", "fastball"…etc. are common baseball words, while terms like "goaltender", "puck", "Power Play" are common hockey words.

Although the addition of these terms would undoubtedly increase the accuracy and power of the model's predictiveness compared to the current model, which is based solely on names, it would also lead to requirements of additional pre-processing steps. The current model relies on the fact that names and other proper nouns have capitalizations, which Tom Tango follows properly in his Tweets. However, the additional terms mentioned in the previous paragraph are mostly not proper nouns, which means they could be either capitalized or not, especially when used at the start of a sentence. Usually, this would be handled by simply converting all tokens in the dictionary and in the Tweets to lowercase.

Converting to lowercase, however, would create a problem for names and proper nouns. For example, there is a legendary Hall of Fame slugger named Babe Ruth, who should be known by most people in the United States, even those that do not follow baseball. Due to the uniqueness of his first name, he is often referred to as "The Babe." Converting that to lowercase would likely create confusion with the term "babe" which is commonly used in song lyrics, for instance. Another example is Boston Red Sox second baseman Trevor Story, whose last name converted to lowercase might create confusion with many stories that aren't about baseball.

A counter method for this problem would be to keep everything in its original cases, and when adding terms into the dictionaries, simply add both versions—one with capitalization and one without. For example, "fastball" and "Fastball" should both be added to the baseball dictionary, and "puck" and "Puck" should both the added to the hockey dictionary. While achievable with small volumes, this method quickly becomes complicated and tedious when the list of terms to be added is increased. It also does not solve the issue that may appear in occasions where the user chooses to "shout" a word or sentence using all caps (e.g., "HOME RUN!").

Lastly, it must also be noted, while there are many terms relatively unique to baseball and ice hockey that could be used to help the model better identify these two labels, there are not many terms that could be used to separate and identify baseball Tweets between the two classes, namely active and retired baseball players. Perhaps the term "Hall of Fame" could be used to identify Tweets about retired players, but that is not always the case. Sometimes people might refer to active players who have already had substantial achievements as a "Future Hall of Famer". Therefore, using additional terms to classify baseball Tweets into "active" and "retired" may be difficult, and additions beyond players' names may not substantially increase the accuracy in the predictions.

## 3. What has been done

Dictionary-based methods have been vastly used in previous studies on text analysis incorporating tweets. In the case of sentiment analysis, it is particularly beneficial to build a collection of known and precompiled sentiment terms known as sentiment lexicon, and it proves to capture the underlying sentiments fairly well when a high-quality dictionary is available that is of interest to the researchers (Okango and Mwambi). Comparably, in this study, customized dictionaries consisting of name entities were curated and employed in order to capture the topics discussed in each document, in the hopes that domain knowledge could help with the performance of our dictionary-based model.

## Conclusion

Major League Baseball implemented The Lockout from December 2, 2021, to March 10, 2022. During this period, employees of MLB were not allowed to publicly discuss any active baseball players represented by the MLBPA. One of these employees is Tom Tango, who is a very active user of Twitter under the handle @tangotiger. He often shares his opinions, ideas, and research on the social media platform, and the two major topics are baseball and ice hockey. This study uses a simple dictionary-based method with names and proper nouns to identify and classify Tango's Tweets from November 11, 2021, to May 9, 2022, and discusses the observed trends within these time frame based on the model's predictions.

The model first labels Tweets into three buckets: "baseball", "hockey", and "neither", while discarding Tweets that do not fall into any of these three buckets. Based on the prediction counts by group and relative to date, some trends were observed but were not very conclusive. By further breaking down all "baseball" Tweets into two classes—"active" and "retired"—a clear trend is observed around the period of The Lockout, and one can clearly tell Tom Tango followed the orders and did not tweet about any active baseball players. The same model could theoretically be applied to other Twitter users that are employed by MLB, such as MLB.com analyst Mike Petriello (@mike_petriello).

## Bibliography

- “2021–22 Major League Baseball Lockout.” *Wikipedia: The Free Encyclopedia*, Wikimedia Foundation, 10 Apr. 2022, https://en.wikipedia.org/wiki/2021–22_Major_League_Baseball_lockout.
- Castrovince, Anthony. “Big Papi elected to Hall on 1st ballot.” *MLB.com*, MLB Advanced Media, 25 Jan. 2022, https://www.mlb.com/news/david-ortiz-hall-of-fame-voting.
- “Challenger, Gray & Christmas.” *Wikipedia: The Free Encyclopedia*, Wikimedia Foundation, 9 Apr. 2022, https://en.wikipedia.org/wiki/Challenger,_Gray_&_Christmas.
- Nesbitt, Stephen J., et al. “On MLB-Owned Media, the Players Now Barely Exist. What's behind That Decision?” *The Athletic*, The Athletic Media Company, 7 Dec. 2021, https://theathletic.com/3002495.
- Okango, Elphas, and Mwambi, Henry. “Dictionary Based Global Twitter Sentiment Analysis of Coronavirus (COVID-19) Effects and Response.” *Annals of Data Science*, Vol. 9, Jan. 2022, pp 175–86, https://doi.org/10.1007/s40745-021-00358-5.
- “Tom Tango.” *Wikipedia: The Free Encyclopedia*, Wikimedia Foundation, 13 Jan. 2022, https://en.wikipedia.org/wiki/Tom_Tango.