# *Sentiment analysis of tweets about debut sports during the Tokyo 2020 Olympics*

Corinne Medeiros
August 2021
https://corinnemedeiros.github.io/

## Executive Summary

This study examines tweets related to new sports included in the 2020 Tokyo Olympics and aims to provide insight for customized marketing plans, scheduling, and potential sponsorships. The domain these data come from is the sports industry, with a focus on social media. Six debut sports, along with additional new events for existing sports, have been added to the Tokyo Olympics schedule including surfing, skateboarding, climbing, karate, and more (Smith, 2021). Fans on Twitter express their feelings towards these new events, providing important insight for the International Olympic Committee and sports sponsors through sentiment analysis.

To analyze the data, I used Python to filter tweets by a selection of new sports, create exploratory visualizations for understanding patterns, and process subsets of the data to prepare for text analysis. Then, I analyzed the sentiment of the relevant tweets and produced supporting visualizations to illustrate the trends.

## Intro and Background of the Problem

The Olympics, as well as any other major media sporting event, requires an extensive amount of planning and financial support. The number of events as well as the number of sports in the Olympics has grown over the years, and Tokyo is the largest yet with 339 events and 41 different sports (Day, 2021). Sponsorship plays an important role in the success of these events, making it possible for athletes to compete, increasing brand awareness, and in turn impacting consumer preferences when it comes to brand loyalty (Constantin Răzvan et al., 2021).

With the ups and downs amidst the global Covid-19 pandemic, the International Olympic Committee, sports teams, individual athletes, and athletic brands are faced with many uncertainties. On top of the continually evolving statistics and safety measures, there are mixed feelings towards large scale events, making planning and budgeting a challenge. Insight on general sentiment towards the newly added sports could be valuable for brands looking to invest in particular athletes and events. For example, Nike could sponsor and focus on clothing for skateboarders or sport climbers, depending on which sport garnered the most attention on Twitter. These insights would also allow for optimal program scheduling. Olympic Committee members could decide to include surfing on the schedule for the next Olympic games if it received positive sentiment from fans on Twitter, and broadcasting networks could make sure it receives a prime spot on the schedule. Additionally, depending on where tweets are coming from, targeted marketing could be applied, and schedules could be organized with respective time zones in mind to optimize viewing.

This dataset from Kaggle contains one csv file with over 150,000 tweets pulled from Twitter using the topic #Tokyo2020. Additional data about each tweet include username, user location, user description, hashtags, date, and more (see Appendix for a full list of available attributes). The data are collected using the Twitter API and the Tweepy Python library. The most updated pull comes from July 28, 2021, which is the version used for this project.

In this analysis I am exploring the following research questions:

Which new sports, like surfing and skateboarding, are most tweeted about?

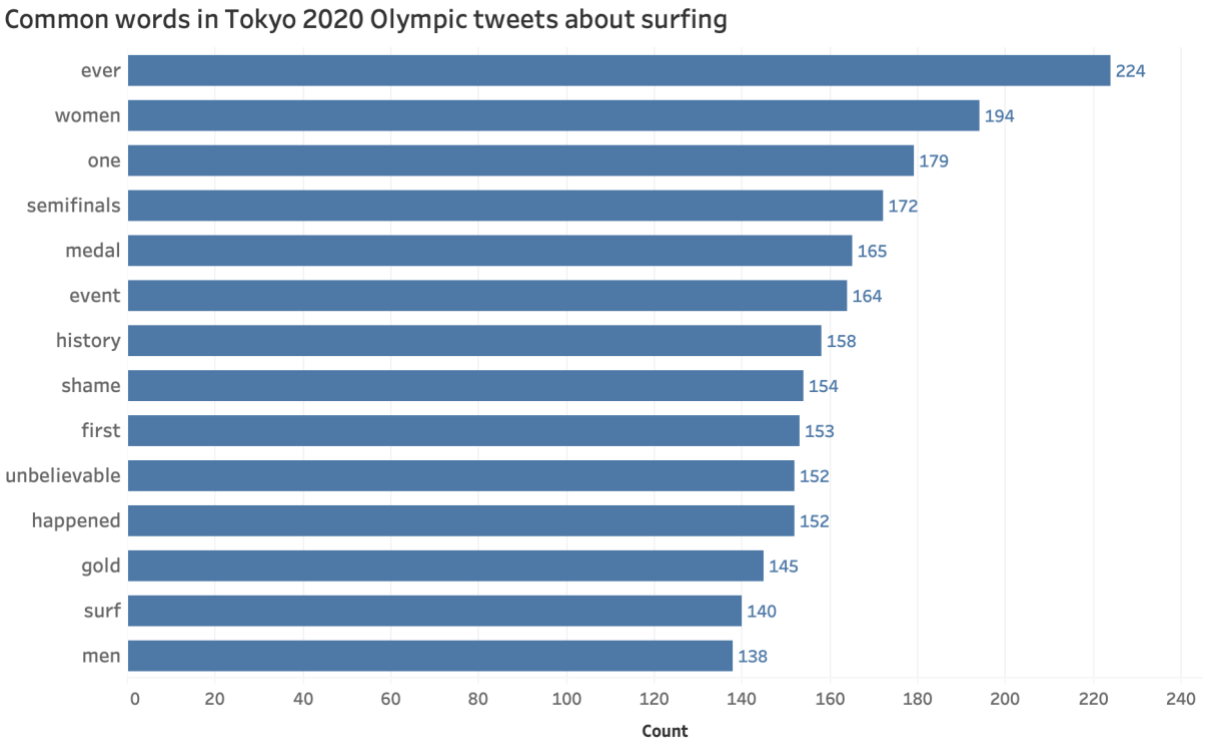
What is the sentiment of the tweets about these new sports?

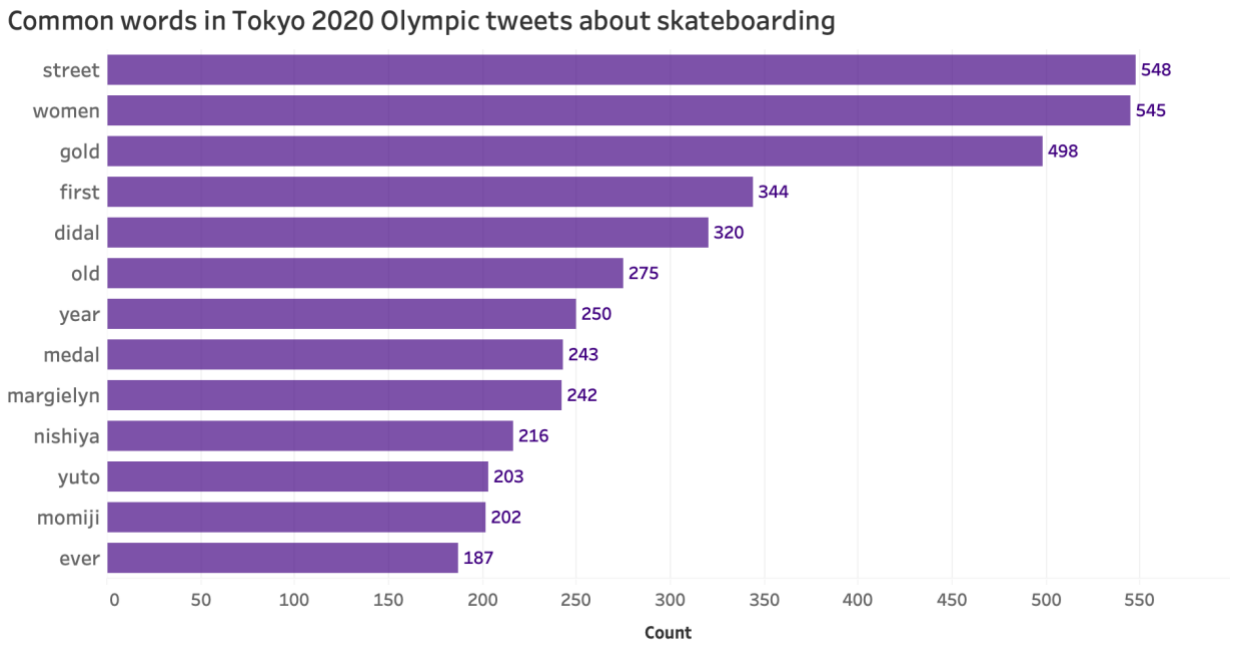
Which new sports or athletes should brands focus on investing in?

Which geographical regions are the majority of these tweets about new sports coming from?

## Methods

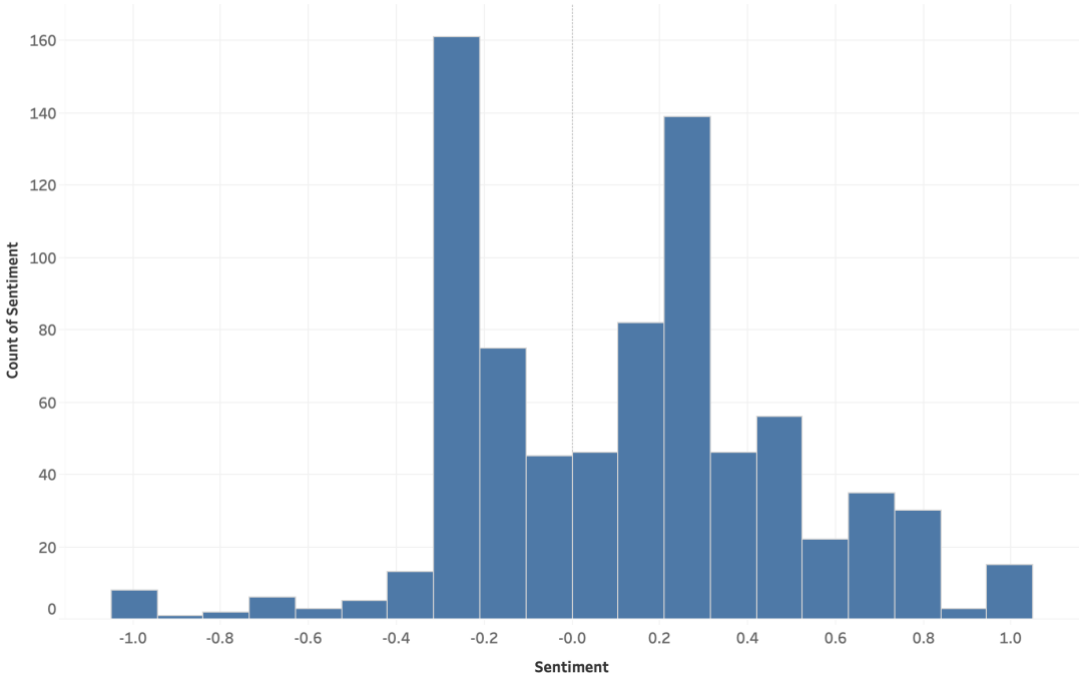
Data cleaning, processing, initial visualizations, and sentiment analysis took place with Python in Jupyter Notebook. I cleaned up the tweets by changing the text to lowercase, and removing special characters, punctuation, and stop words. Once cleaned up, I filtered the text data by searching for tweets containing the words surfing and then skateboarding. Surfing showed up in 1,758 tweets and skateboarding appeared in 3,420 tweets. From there I created a list of all the words in each of the tweets in order to calculate word frequencies. I removed the “surf” and “skateboard” search terms from the word frequencies results, as well as “Tokyo” and “Olympics” to limit excess noise. With the corresponding list of most common words for each sport, I plotted the word counts using the following bar graphs to visualize the trends.



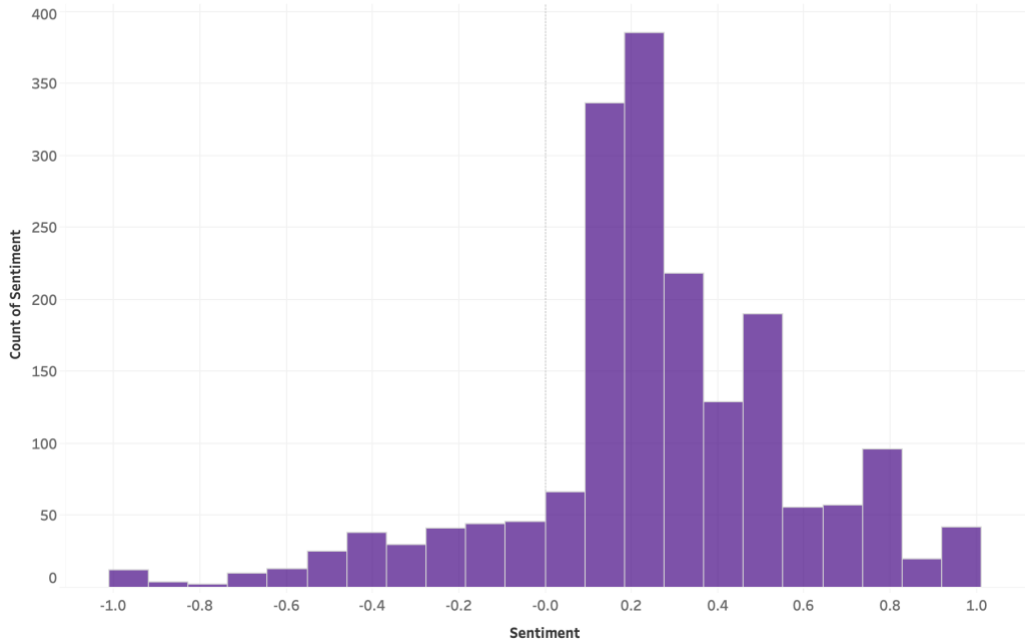



For sentiment analysis, I used the TextBlob package to produce polarity values for each tweet and store them in a new column. As a rule-based model, TextBlob uses a pre-defined library of categorized words to draw from when interpreting the sentiment of a word (Kuzminykh, 2020). Polarity values fall between -1 and 1, with values closer to -1 meaning negative sentiment and values closer to 1 meaning positive sentiment. Polarity values of 0 represent neutral statements, or they result from words not being available in TextBlob's library and therefore ignored. Taking the values and plotting them on a histogram revealed the distribution of negative and positive values. For a more interpretable visual, I removed the observations with a 0-polarity value and added a break at zero to more easily distinguish between the positive and negative sides.

Tokyo 2020 Olympics:
Sentiments from tweets about surfing




Tokyo 2020 Olympics:
Sentiments from tweets about skateboarding




I further explored the data using Tableau to create additional exploratory and final visualizations. To see where the tweets originated from, I created bar charts for each sport using the user location

column. The area with the most tweets related to surfing is Serra da Saudade in Brazil, followed by La Jolla, California. The other common regions of origin are cities near the water, which makes sense for surfing popularity. For skateboarding, the main locations the tweets came from are the Philippines, London, and Canada.

## Results

The main criteria I based my conclusions on are the word frequencies and the polarity values. More positive values mean more positive sentiment. Visually, the sentiment graphs show that the majority of tweets for surfing and skateboarding fall on the positive side. The word frequencies plots reveal that people are discussing women in both of these sports, and skateboarding fans frequently mention athlete Margielyn Didal.

Based on the information given, I had to make some assumptions about the data. From the data collection method described on Kaggle, I assumed it included a wide variety of locations around the globe and didn't limit any regions. Based on the information in the Data Explorer tool in Kaggle, I also assumed that there were no retweets included since it showed 0% true values.

With these results, sponsors can decide to concentrate on creating clothing for women skateboarders to generate interest among women consumers and fans, and they could partner with Margielyn Didal to further promote brand awareness. If a woman watching skateboarding becomes inspired to be the next Didal, she is more likely to pay attention to what Didal is wearing and then purchase or click on relevant advertising. This benefits the athletes with financial support from sponsorship, the consumers with brand choices, and the brands with successful marketing. Schedule-wise, broadcast networks, along with the Olympic Committee, could use the location insights to make sure surfing events are accessible for viewers in Serra da Saudade, and skateboarding events are easy to access in the Philippines and London.

## Discussion and Conclusion

Challenges with this project include data updates, text data, language bias, and changing external circumstances. Unforeseen evolving sentiment from outside factors like Covid-19 could create unexpected turns in how the public is feeling towards the events. Sentiment could shift as more events take place and different wins occur, and controversies could unfold. Since I am analyzing the text in English, I'm losing insight from foreign language tweets, which creates bias. In terms of current data, the dataset I'm using claims to be updated frequently but it has not been updated since I first downloaded it so it's difficult to know if there were any major changes in attitudes after more events took place. Due to the global nature of these tweets, there could also be limited representation from non-English speakers and countries without the same access to watch certain events or comment on Twitter. Finally, the locations data are messy and would require extensive cleanup in order to create proper and accurate visuals for additional insights.

To take this project to the next level, I could supplement the data with newer tweets from other datasets or pull the tweets myself using the Twitter API to get a sense of more current sentiment. I could work on cleaning up the user location column to get rid of duplicates and see a more accurate representation of regions. I could analyze additional sports to compare popularity trends. To offset the language bias, a multilingual transformer model could be applied, but unfortunately the nature of Twitter syntax makes this difficult (Barriere & Balahur, 2020). Furthermore, I could look at trends using the date column to plot the number of tweets over time to see when viewers were most positively engaged. Finally, taking a closer look at what specifically was mentioned in both negative and positive tweets would be very useful in understanding the context of where fans are coming from.

Gaining insight on how the world is feeling about particular sports during the Covid-19 pandemic is important for understanding what to focus on in a marketing sense globally. In the sports industry, there is an expectation for results when it comes to sponsorship. Sponsors need to know they'll have a good chance on return on investment, especially with tighter budgets brought on by the pandemic. Knowing

which newly introduced Olympic sports generate positive reactions and overall popular interest can help brands leverage current trends and feel good about where they're spending their money and who they're partnering with. Looking at word frequencies in the tweets gives us a general idea about what audiences are discussing in relation to each sport. Analyzing the characteristics and sentiment of these tweets also provides a framework to analyze future data.

**References**

Adwan, O., Al-Tawil, M., Huneiti, A., Shahin, R., Zayed, A. & Al-Dibsi, R. (2020). Twitter sentiment analysis approaches: A survey. International Journal of Emerging Technologies in Learning, 15, 79–93. https://doi-org.ezproxy.bellevue.edu/10.3991/ijet.v15i15.14467

Barriere, V. & Balahur, A. (2020). Improving sentiment analysis over non-English tweets using multilingual transformers and automatic translation for data-augmentation. In Proceedings of the 28th International Conference on Computational Linguistics, pages 266–271, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Constantin Răzvan, B., Bogdan, B., Laurenţiu, D., Cătălin, P., Daniel, P. & Paula, A. (2021). The role of social media on sponsorship activation. Studia Universitatis Babes-Bolyai, Educatio Artis Gymnasticae, 66(1), 111–126.

Day, M. (2021, July 23). The New Olympic Events Debuting in Tokyo for 2021. NBC Chicago. Retrieved from https://www.nbcchicago.com/news/sports/tokyo-summer-olympics/tokyo-olympics-new-sports-and-events-debuting-in-2021/2563173/

Kennedy, M., Fadel, L. & Goldman, T. (2021). Olympic opening ceremony is a delicate mix of celebration and solemnity. NPR. Retrieved from https://www.npr.org/sections/tokyo-olympics-live-updates/2021/07/23/1019622003/tokyo-olympics-opening-ceremony

Kuzminykh, N. (2020, October 23). Sentiment analysis in Python with TextBlob. Stack Abuse. Retrieved from https://stackabuse.com/sentiment-analysis-in-python-with-textblob/

Morrissey, M., Wasser, L., Diaz, J. & Palomino, J. (2020, September 11). Analyze word frequency counts using Twitter data and Tweepy in Python. Earth Lab. Retrieved from https://www.earthdatascience.org/courses/use-data-open-source-python/intro-to-apis/calculate-tweet-word-frequencies-in-python/

Preda, G. (2021). Tokyo Olympics 2020 Tweets. Kaggle. Retrieved from https://www.kaggle.com/gpreda/tokyo-olympics-2020-tweets

Rivenburgh, N. K. (2002). The Olympic games: Twenty-first century challenges as a global media event. Culture, Sport, Society, 5(3), 31. https://doi-org.ezproxy.bellevue.edu/10.1080/911094208

Seilsepour, A., Ravanmehr, R. & Sima, H. (2019). 2016 Olympic games on Twitter: Sentiment analysis of sports fans tweets using big data framework. Journal of Advances in Computer Engineering and Technology, 5(3), 143–160.

Smith, S. (2021, April 6). Tokyo Olympics 101: What are the new sports? NBC. Retrieved from https://www.nbcolympics.com/news/tokyo-olympics-101-new-sports

Zimmerman, M. & Huang, G. (2021, August 1). Tokyo Olympics: Tone shifts in Japan as medals outshine Covid-19 concerns. Stuff Limited. Retrieved from https://www.stuff.co.nz/sport/olympics/300371436/tokyo-olympics-tone-shifts-in-japan-as-medals-outshine-covid19-concerns

Zuccarini, D. (2020, July 21). Now, more than ever, the sponsorship industry must demand data. SportsPro. Retrieved from https://www.sportspromedia.com/opinion/sports-sponsorship-analytics-strategy-liverpool-manchester-city

# Appendix

***Tokyo Olympics 2020 Tweets Dataset Attributes:***

id

user_name

user_location

user_description

user_created

user_followers

user_friends

user_favourites

user_verified

date

text

hashtags

source

retweets

favorites

is retweet

## Questions

1. What can we learn about sports consumers using these data?
2. What additional data details could help improve this analysis?
3. Will these insights be useful outside of the surfing and skateboarding realm?
4. What potential do these data provide in terms of future predictions?
5. What measures of success do brands base their sponsorship decisions on?
6. Can this analysis be tailored to specific athletes or teams?
7. Can we implement this analysis on newer data?
8. What further visualizations and details can you provide?
9. How can the Olympic Committee compare these sports with other sports?
10. What other text sources besides Twitter data might offer supplemental insight?