# Scraping Big Data without Breaking the Bank: Using Twitter API for Academic Research

Hannah L. Ford
Department of Agricultural Education and Communications
Texas Tech University
Box 42131
Lubbock, TX 79409-2131
hannford@ttu.edu

# Scraping Big Data without Breaking the Bank: Using Twitter API for Academic Research

## Introduction

Since the 1989 birth of the information system known as the World Wide Web, the accumulation of social data available to researchers has rapidly increased (Manguri et al., 2020; McPherson, 2009). *Big data*, a term describing large volumes of data and the way it is strategically analyzed, is changing the way we manage and derive insight from online information (SAS, 2021). Social networking sites continuously generate information, but few are as chronologically stored and public as Twitter (Sohail et al., 2021). Reporting 199 million daily active users and over a trillion tweets as of 2019, Twitter meets all definitions of big data (Leetaru, 2019; Tankovska, 2021).

Big data is widespread, but the capability of coding experience needed for accessing it is not; valuable data is often obtained by researchers and universities paying expensive third-party social media "scraping" businesses, who employ programmers to code tools (Agrawal, 2019). Twitter is no exception, pricing external businesses up to $2,499 for monthly database access (Perez, 2017). Businesses such as Meltwater and Cision offer individual user access at up to $7,200, with limitations on data storage as low as 380,000 "responses" at a time (Donda, 2021). Individual researchers were left with either paying expensive third-party businesses or manually scraping Twitter data – one tweet at a time – which limited research opportunities and frustrated the coding community (Perez, 2017). Attempting to bridge this gap, Twitter implemented its *Application Programming Interface* (API) in 2017, allowing coders and developers to sift through large amounts of tweets by defining specific words, phrases, and hashtags (Franz, 2021; Perez, 2017). In late 2020, an updated version released allowing between 7 and 30 days of automated data scraping with higher levels of access for academic researchers (Franz, 2021).

Previously limited to 3,200 manual tweets for a platform-wide search (Barrios-O'Niell, 2020), the new academic research developer Twitter API v2 now grants individual researchers – with proper coding experience – access to the entire database archive, from the first tweet in 2006 until the present moment. API v2 has primarily been utilized in COVID research, with one study analyzing over 8 million tweets for sentiment analysis (Aguilar-Gallegos et al., 2020). Data are currently characterized globally as the "most valuable resource", and the agricultural industry is responding to this trend (Borrero & Zabalo, 2021, p. 809). Although relatively new, API has been used in a general study regarding the language surrounding food (Fried et al., 2014) and a few studies surrounding the agricultural industry in marketing and conservation (Borrero & Zabalo, 2021; Boyer et al., 2020; Franz et al., 2021). Upon building this literature review, no agricultural education and communication research studies implementing the API v2 were found.

## How it Works

Utilizing a coding language, developers create rules and coding functions to filter tweets through phrases, keywords, dates, hashtags, language, and more. Coding languages include, but are not limited to: Java, JavaScript, Python, R, RB, cURL, twURL and HTTP. Results may vary in structure depending on the language used. I used JavaScript functions in my exploration of the Twitter API, which provided *dehydrated tweets* (tweets simplified to full tweet text and a single identifying code, ex: 83993). Researchers can stop at this point and analyze text with a .csv file or engage additional, free coding applications to gain information, as stated in the results section.

## Program Phases

The developer application process for academic research access is similar to submitting a study to a university's Institutional Review Board. Through an existing Twitter account, the researcher applied through the developer portal on Twitter's website and was asked the following prompts:

1. What's your research project's name?
2. Does this project receive funding from outside your academic institution?
3. Describe your research project.
4. Describe how Twitter data and Twitter APIs will be used in your research project.
5. Will your research present Twitter data individually or in aggregate?
6. Describe your methodology for analyzing Twitter data, Tweets, etc.
7. Describe how you will share the outcomes of your research.
8. Will your analysis make Twitter content or data available to a government entity?

Upon approval, the user becomes a Twitter API developer and can utilize authorization keys (passwords used in coding) to gain access to the full archive of the database history, specifically on the academic researcher track. The process of retrieving data can be done by creating rules and functions through one of the approved coding languages and programs that support them. I invested approximately 60 hours to understand the Twitter API v2 process from start to finish, even with years of various coding experience. Every coding language listed was tested in dozens of application programs and the terminal of the computer was often utilized to track progress and eliminate pathways. An account has a limitation of 10,000,000 tweets per month, utilizing high-efficiency computers and SPSS or similar applications capable of big data cleaning and analysis.

## Results to Date

The final product of this process involved taking dehydrated tweets scraped from the Twitter API v2, processing it through several free coding applications and stages, and *rehydrating* tweets to include information such as full text, hashtags, engagement, the author's location, biography, followers, following, tweet counts, verified status, exact time and date of posting, and the exact coordinates of the author when the tweet was posted (if security permissions were bypassed by the Tweet author). Within 30 seconds, the API collected 498 tweets on the primary test of this process for the phrase "food desert", with rehydrated information added later through my code.

## Future Plans & Advice to Others

After being involved in several manual data scraping studies on Twitter, I was determined to utilize my past coding knowledge and find a better way to acquire Twitter data. I recommend that anyone who is interested in acquiring big data from the platform through API have coding experience. This process will be utilized in my dissertation research to analyze Twitter sentiment regarding food insecurity in university student using supervised and unsupervised machine learning AI analysis. I am also seeking opportunities to consult anyone interested in big data.

## Costs and Resources Needed

The biggest incentive for this method of data collection is cost. Provided the user has necessary previous coding experience, this method is free to access and can filter millions of tweets at a time. Potentially replacing the need for costly third-party programs that charge thousands of dollars, this free method provides more specific information and more big data. This process also saves time by allowing the use of machine learning analysis, which require big data to function.

**References**


Agrawal, H. (2019, February 18). *7 Social Media Scraping Tools for 2019*. Datahut. https://www.blog.datahut.co/post/social-media-scraping-tools

Aguilar-Gallegos, N., Romero-García, L. E., Martínez-González, E. G., García-Sánchez, E. I., & Aguilar-Ávilaa, J. (2020). Dataset on dynamics of Coronavirus on Twitter. *Data in Brief*, *30*. https://doi.org/10.1016/j.dib.2020.105684

Barrios-O'Neill, D. (2020). Focus and social contagion of environmental organization advocacy on Twitter. *Conservation Biology*, *35*(1), 307-315. https://doi.org/10.1111/cobi.13564

Borrero, J. D., & Zabalo, A. (2021). Identification and analysis of strawberries' consumer opinions on twitter for marketing purposes. *Agronomy*, *11*(4), 809. https://doi.org/10.3390/agronomy11040809

Boyer, A., Vuador, L., Lay, Y. L., & Marty, P. (2020). Building consensus? The production of a water conservation discourse through Twitter: The water use it wisely campaign in Arizona. *Environmental Communication*, *15*(3). https://doi.org/10.1080/17524032.2020.1821743

Donda, B. (2021). C*ision vs Meltwater vs Prowly – 2021 Feature & Pricing Comparison*. Prowly. https://prowly.com/magazine/cision-vs-meltwater/

Franz, A., Junirianto, E., & Suswanto. (2021). Web design and application programming interface (API) smart farming application. *TEPIAN*, *2*(1), 33-37. https://media.neliti.com/media/publications/344830-web-design-and-application-programming-i-cb69bbcf.pdf

Fried, D., Surdeanu, M., Koburov, S., Hingle, M., & Bell, D. (2014, October 30). *Analyzing the language of food on social media* [Paper presentation]. 2014 IEEE International Conference on Big Data (Big Data), Washington DC, USA. https://doi.org/10.1109/BigData.2014.7004305

Leetaru, K. (2019, February 11). *AI & Big Data: How Big is Social Media and Does it Really Count as 'Big Data'?*. Forbes. https://www.forbes.com/sites/kalevleetaru/2019/02/11/how-big-is-social-media-and-does-it-really-count-as-big-data

Manguri, K. H. , Ramadhan, R. N., & Amin, P.R. (2020). Twitter sentiment analysis on worldwide COVID-19 outbreaks. *Kurdistan Journal of Applied Research,* 5(3) 54-65. https://doi.org/10.24017/covid.8

McPherson, S. S. (2009). *Tim Berners-Lee: Inventor of the World Wide Web*. Twenty-First Century Books. https://archive.org/details/timbernerslee0000mcph

Perez, S. (2017, November 14). *Twitter launches lower-cost subscription access to its data through new premium APIs*. TechCruch. https://techcrunch.com/2017/11/14/twitter-launches-lower-cost-subscription-access-to-its-data-through-new-premium-apis/

SAS. (2021). *Big Data: What it is and Why it Matters*. https://www.sas.com/en_us/insights/big-data/what-is-big-data.html

Sohail, S. S., Khan, M. M., Arsalan, M., Khan, A., Siddiqui, J., Hasan, S. H., & Alam, M. A. (2021, May 22). Crawling Twitter data through API: A technical/legal perspective. ARXIV. https://arxiv.org/pdf/2105.10724.pdf

Tankovska, H. (2021, June 29). *Leading countries based on number of Twitter users as of April 2021*. Statista. https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/