Tracking Social Media Discourse About the COVID-19 Pandemic:Development of a Public Coronavirus Twitter Data Set

Abstract

Background: At the time of this writing, the coronavirus disease (COVID-19) pandemic outbreak has already put tremendousstrain on many countries' citizens, resources, and economies around the world. Social distancing measures, travel bans,self-quarantines, and business closures are changing the very fabric of societies worldwide. With people forced out of publicspaces, much of the conversation about these phenomena now occurs online on social media platforms like Twitter.Objective: In this paper, we describe a multilingual COVID-19 Twitter data set that we are making available to the researchcommunity via our COVID-19-TweetIDs GitHub repository.Methods: We started this ongoing data collection on January 28, 2020, leveraging Twitter's streaming application programminginterface (API) and Tweepy to follow certain keywords and accounts that were trending at the time data collection began. Weused Twitter's search API to query for past tweets, resulting in the earliest tweets in our collection dating back to January 21,2020.Results: Since the inception of our collection, we have actively maintained and updated our GitHub repository on a weeklybasis. We have published over 123 million tweets, with over 60% of the tweets in English. This paper also presents basic statisticsthat show that Twitter activity responds and reacts to COVID-19-related events.Conclusions: It is our hope that our contribution will enable the study of online conversation dynamics in the context of aplanetary-scale epidemic outbreak of unprecedented proportions and implications. This data set could also help trackCOVID-19-related misinformation and unverified rumors or enable the understanding of fear and panic—and undoubtedly more

(JMIR Public Health Surveill 2020;6(2):e19273) doi: 10.2196/19273