

Article

An Exploratory Study of Tweets about the SARS-CoV-2 Omicron Variant: Insights from Sentiment Analysis, Language Interpretation, Source Tracking, Type Classification, and Embedded URL Detection

Nirmalya Thakur *  and Chia Y. Han

Department of Electrical Engineering and Computer Science, University of Cincinnati, Cincinnati, OH 45221-0030, USA; han@ucmail.uc.edu

* Correspondence: thakurna@mail.uc.edu

Abstract: This paper presents the findings of an exploratory study on the continuously generating Big Data on Twitter related to the sharing of information, news, views, opinions, ideas, knowledge, feedback, and experiences about the COVID-19 pandemic, with a specific focus on the Omicron variant, which is the globally dominant variant of SARS-CoV-2 at this time. A total of 12,028 tweets about the Omicron variant were studied, and the specific characteristics of the tweets that were analyzed include sentiment, language, source, type, and embedded URLs. The findings of this study are manifold. First, from sentiment analysis, it was observed that 50.5% of tweets had a ‘neutral’ emotion. The other emotions—‘bad’, ‘good’, ‘terrible’, and ‘great’—were found in 15.6%, 14.0%, 12.5%, and 7.5% of the tweets, respectively. Second, the findings of language interpretation showed that 65.9% of the tweets were posted in English. It was followed by Spanish or Castilian, French, Italian, Japanese, and other languages, which were found in 10.5%, 5.1%, 3.3%, 2.5%, and <2% of the tweets, respectively. Third, the findings from source tracking showed that “Twitter for Android” was associated with 35.2% of tweets. It was followed by “Twitter Web App”, “Twitter for iPhone”, “Twitter for iPad”, “TweetDeck”, and all other sources that accounted for 29.2%, 25.8%, 3.8%, 1.6%, and <1% of the tweets, respectively. Fourth, studying the type of tweets revealed that retweets accounted for 60.8% of the tweets, it was followed by original tweets and replies that accounted for 19.8% and 19.4% of the tweets, respectively. Fifth, in terms of embedded URL analysis, the most common domain embedded in the tweets was found to be twitter.com, which was followed by biorxiv.org, nature.com, wapo.st, nzherald.co.nz, recvprofits.com, science.org, and other domains. Finally, to support research and development in this field, we have developed an open-access Twitter dataset that comprises Tweet IDs of more than 500,000 tweets about the Omicron variant, posted on Twitter since the first detected case of this variant on 24 November 2021.

Keywords: COVID-19; SARS-CoV-2; Omicron; Twitter; tweets; sentiment analysis; Big Data; Natural Language Processing; Data Science; Data Analysis



Citation: Thakur, N.; Han, C.Y. An Exploratory Study of Tweets about the SARS-CoV-2 Omicron Variant: Insights from Sentiment Analysis, Language Interpretation, Source Tracking, Type Classification, and Embedded URL Detection. *COVID* **2022**, *2*, 1026–1049. <https://doi.org/10.3390/covid2080076>

Academic Editors: Sandra Costanzo and Giuseppe Novelli

Received: 17 May 2022

Accepted: 20 July 2022

Published: 25 July 2022

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

An outbreak of an unknown respiratory disease started in December 2019 in the seafood market in Wuhan, China, infecting about 66% of the people at the market. Very soon, more people in different parts of China got infected by the same disease. This prompted an investigation from the healthcare and medical sectors, and very soon, it was concluded that a novel coronavirus was responsible for this disease. This novel coronavirus was named severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2, 2019-nCoV) as it was observed to have a high homology of about 80% with SARS-CoV [1]. The disease that humans suffer from after getting infected by this virus is known as COVID-19 [2]. The Chinese government implemented multiple measures to reduce the spread of the

virus [3]. However, the virus rapidly spread across China and soon started spreading in different countries of the world. COVID-19 was declared a pandemic by the World Health Organization (WHO) on 11 March 2021 [4].

At the time of writing this paper, globally, there have been 549,667,293 cases of COVID-19 with 6,352,025 deaths [5]. The SARS-CoV-2 virus mainly attacks the respiratory system of humans, although infections in other organs of the body have also been reported in some cases. The symptoms, as reported from the initial studies of the cases from Wuhan, China, include fever, dry cough, dyspnea, headache, dizziness, exhaustion, vomiting, and diarrhea. The report also mentions that not everyone has the same symptoms, and the nature and intensity of the symptoms vary from person to person [6]. When the genetic sequence of a virus changes, it is said to have mutated. Genomes of a virus that are different from each other in terms of their genetic sequence are called variants. Variants that differ in terms of phenotype are known as strains [7]. On 10 January 2020, the genome sequences of SARS-CoV-2 were made publicly available by the Global Influenza Surveillance and Response System (GISAID), which is a primary source on a global scale for open access to the genomic data of influenza viruses [8]. Since then, the database of GISAID has made more than 5 million genetic sequences of SARS-CoV-2 from 194 countries and territories publicly available for research [9]. In an attempt to prioritize research related to COVID-19, the WHO has made a conscious effort to classify the variants of SARS-CoV-2 into three categories. These include variants of concern (VOCs), variants of interest (VOIs), and variants under monitoring (VUMs).

Since the initial cases in December 2019, the SARS-CoV-2 virus has undergone multiple mutations, and as a result, several variants have been detected in different parts of the world. Some of these include Alpha (B.1.1.7), Beta (B.1.351), Gamma (P.1), Delta (B.1.617.2), Epsilon (B.1.427 and B.1.429), Eta (B.1.525), Iota (B.1.526), Kappa (B.1.617.1), Zeta (P.2), Mu (B.1.621 and B.1.621.1), and Omicron (B.1.1.529, BA.1, BA.1.1, BA.2, BA.3, BA.4, and BA.5) [10]. Out of all these variants, the Omicron variant, first detected on 24 November 2021, was classified as a VOC by WHO on 26 November 2021 [11]. Soon thereafter, the Omicron variant became the globally dominant form of SARS-CoV-2 [12]. The Omicron spike protein contains 30 mutations and has been reported to be the most immune-evasive VOC of SARS-CoV-2 and is highly resistant against antibody-mediated neutralization [13,14]. Despite the development of vaccines [15] and various forms of treatments [16], a recent report from WHO states that the Omicron variant accounts for 86% of the infections caused by SARS-CoV-2 on a global scale [17]. In another recent report, WHO mentioned that the cases due to the Omicron variant were “off the charts” as global infections due to the SARS-CoV-2 variant set new records [18]. Currently, some of the countries that have recorded the most cases due to the SARS-CoV-2 Omicron variant include: United Kingdom (1,138,814 cases), USA (945,470 cases), Germany (245,120 cases), Denmark (218,106 cases), France (110,959 cases), Canada (92,341 cases), Japan (71,056 cases), India (56,125 cases), Australia (46,576 cases), Sweden (43,400 cases), Israel (39,908 cases), Poland (33,436 cases), and Brazil (32,880 cases) [19].

In today's world, where the internet virtually connects people in different geographic regions, social media has been considered an “integral vehicle” of people's lives and an “online community” for communication, information, news, views, opinions, perspectives, ideas, knowledge, feedback, and experiences related to pandemics, epidemics, viruses, and diseases [20–24]. Out of all the social media platforms, Twitter is popular amongst all age groups [25], and there are about 192 million daily active users on Twitter [26].

Twitter has been highly popular amongst healthcare researchers, epidemiologists, medical practitioners, data scientists, and computer science researchers for studying, analyzing, modeling, and interpreting social media communications related to pandemics, epidemics, viruses, and diseases such as Ebola [27], E-Coli [28], Dengue [29], Human papillomavirus (HPV) [30], Middle East Respiratory Syndrome (MERS) [31], Measles [32], Zika virus [33], H1N1 [34], influenza-like illness [35], swine flu [36], flu [37], Cholera [38],

Listeriosis [39], cancer [40], Liver Disease [41], Inflammatory Bowel Disease [42], kidney disease [43], lupus [44], Parkinson's [45], Diphtheria [46], and West Nile virus [47].

A comprehensive analysis of multimodal aspects of tweets posted on Twitter has been of significant interest to the scientific community during similar epidemics and virus outbreaks in the past. For instance, during the outbreak of the Ebola Virus, researchers studied tweets posted during the outbreak to perform sentiment analysis [48], embedded URL detection [49], tweet content investigation [50], and analysis of the tweet type, such as retweets [51]. Similarly, during the outbreak of the Zika virus, researchers studied tweets posted during the outbreak to perform sentiment analysis [52], investigate retweet characteristics [53], detect embedded URLs in tweets [54], understand the source of the tweets [55], and detect the language of the tweets [56]. A third example can be seen from the research works related to studying the multimodal aspects of tweets that focused on the flu outbreak. Relevant tweets posted both during and just prior to the outbreak were studied for performing sentiment analysis [57], analyzing tweet content [58], investigating retweet characteristics [59], and detecting embedded words [60].

The outbreak of the COVID-19 pandemic has served as a “catalyst” towards increasing the usage of Twitter [61–64] for conversations on a wide range of topics in this regard, resulting in the generation of tremendous amounts of Big Data. Some of the most popular use cases of Twitter during this pandemic, as reported in recent works in this field, include the following:

1. Sharing of symptoms, information, and experiences as reported by frontline workers and people who were infected with the virus [65].
2. Providing suggestions, opinions, and recommendations to reduce the spread of the virus [66].
3. Communicating updates on vaccine development, clinical trials, and other forms of treatment [67].
4. Sharing guidelines mandated by various policy-making bodies, such as mask mandate, social distancing, etc., that were required to be followed by members in specific geographic regions of the world under the authority of the associated policy-making bodies [68].
5. Dissemination of misinformation such as the use of certain drugs or forms of treatment that have not been tested or have not undergone clinical trials [69].
6. Creating and spreading conspiracy theories such as considering 5G technologies responsible for the spread of COVID-19, which eventually led to multiple 5G towers being burnt down in the United Kingdom [70].
7. Studying public opposition to available vaccines in different parts of the world [71].

In addition to these works related to the use of Twitter during COVID-19, there have been some other works in this field (as discussed in detail in Section 2) that further uphold the extensive increase in Twitter posts for information seeking and sharing during the ongoing surge of COVID-19 cases due to the Omicron variant. However, none of these works in this field took a comprehensive approach towards drawing insights from such tweets—such as performing sentiment analysis, tweet type detection, embedded URL tracking, tweet source tracking, and language detection—that were performed during the outbreaks of viruses in the past, such as during the outbreaks of Ebola Virus [48–51], Zika virus [52–56], and the flu [57–60]. To address this research challenge, in this work, a total of 12,028 tweets containing views, expressions, opinions, perspectives, attitudes, news, information, and related themes about the SARS-CoV-2 Omicron variant posted publicly on Twitter between 5 May 2022, to 12 May 2022 were studied, analyzed, and interpreted to perform a comprehensive analysis. In summary, the findings of this study are as follows:

1. The results from sentiment analysis showed that a majority of the tweets (50.5%) had a ‘neutral’ emotion, which was followed by the emotional states of ‘bad’, ‘good’, ‘terrible’, and ‘great’ that were found in 15.6%, 14.0%, 12.5%, and 7.5% of the tweets, respectively.

2. The results from tweet source tracking showed that 35.2% of the Tweets were posted from an Android source, which was followed by the Twitter Web App, iPhone, iPad, TweetDeck, and other sources that accounted for 29.2%, 25.8%, 3.8%, 1.6%, and <1% of the tweets, respectively.
3. The results from tweet language interpretation showed that 65.9% of the tweets were posted in English, which was followed by Spanish or Castillian (10.5%), French (5.1%), Italian (3.3%), Japanese (2.5%), and other languages that accounted for <2% of the tweets.
4. The results from tweet type classification showed that the majority of the tweets (60.8%) were retweets which was followed by original tweets (19.8%) and replies (19.4%).
5. The results from embedded URL analysis showed that the most common domain embedded in the tweets was twitter.com, which was followed by biorxiv.org, nature.com, wapo.st, nzherald.co.nz, recvprofits.com, science.org, and a few other domains.

In addition to the above, to further support research and development in this field, we present an open-access dataset of 522,886 Tweet IDs of the same number of tweets about the COVID-19 Omicron variant, which have been posted publicly on Twitter since the first detected case of this variant on 24 November 2021. The development of Twitter datasets has been of significant interest to the scientific community, as can be seen from the recent Twitter datasets on the 2020 U.S. Presidential Elections [72], 2022 Russia–Ukraine war [73], climate change [74], natural hazards [75], European Migration Crisis [76], movies [77], toxic behavior amongst adolescents [78], music [79], civil unrest [80], drug safety [81], and Inflammatory Bowel Disease [82]. Since the outbreak of COVID-19, there have been a few works that focused on the development of Twitter datasets. These include an Arabic Twitter dataset [83], a Twitter dataset for vaccine misinformation [84], a Twitter dataset on misleading information about COVID-19 [85], a Twitter dataset for COVID-19-related misinformation [86], and a dataset on COVID-19 rumors [87]. However, none of these works [83–87] have focused on the development of a dataset that comprises tweets about the Omicron variant that were posted on Twitter since the first detected case of this variant. Therefore, the development of this dataset further helps to uphold the scientific contributions of this paper. This open-access dataset, publicly available at <https://doi.org/10.5281/zenodo.6893676>, is compliant with the privacy policy, developer agreement, and guidelines for content redistribution of Twitter, as well as with the FAIR principles (Findability, Accessibility, Interoperability, and Reusability) principles for scientific data management. The rest of this paper is presented as follows. In Section 2, a review of recent works in this field is presented. The methodology is discussed in Section 3. Section 4 discusses the results. The conclusion is presented in Section 5, which summarizes the scientific contributions of this study and outlines the scope for future work in this field. It is followed by the references section.

2. Literature Review

In this section, a review is presented of the recent works in this field that focused on studying social media behavior with a specific focus on Twitter since the outbreak of SARS-CoV-2. Shamrat et al. [88] developed a k-nearest neighbors (kNN)-based machine learning classifier to classify tweets related to COVID-19 into three classes—‘positive’, ‘negative’, and ‘neutral’. The study focused on filtering tweets related to COVID-19 vaccines, and thereafter, the kNN algorithm was applied to the filtered tweets for further analysis. Sontayasara et al. [89] used the support vector machine (SVM) classifier to develop an algorithm for sentiment analysis. The algorithm was tested on tweets where people communicated their plans to visit Bangkok during the pandemic and how those plans were affected. This classifier was able to classify the tweets into three classes of sentiments—‘positive’, ‘negative’, and ‘neutral’. Asgari-Chenaghlu et al. [90] developed an approach to detect the trending topics and major concerns related to the COVID-19 pandemic as expressed by people on Twitter. Amen et al. [91] proposed a framework that applied a directed acyclic graph model on tweets related to COVID-19 to detect any anomaly events. The work of Liu

et al. [92] involved developing an approach to study tweets about COVID-19 that involved the Centers for Disease Control and Prevention (C.D.C.). The objective of this study was to detect public perceptions, such as concerns, attention, expectations, etc., related to the guidelines of the C.D.C. regarding COVID-19. Al-Ramahi et al. [93] developed a methodology to filter and study tweets posted between 1 January 2020, and 27 October 2020, where people expressed their opposing views towards wearing masks to reduce the spread of COVID-19. Jain et al. [94] proposed a methodology to analyze tweets related to COVID-19 that could assign an influence score to the associated users who posted these tweets. The objective of this study was to identify influential users on Twitter who posted about COVID-19. Madani et al. [95] developed a random-forest-based classifier to detect tweets about COVID-19 that contained fake news. The classifier achieved a performance accuracy of 79%.

Shokoohyar et al. [96] proposed a system to study tweets where people expressed their opinions regarding the lockdown in the United States on account of COVID-19. Chehal et al. [97] developed a software using Python and R to analyze the mindset of Indians as expressed in their tweets during the two nationwide lockdowns that were implemented by the Indian government on account of COVID-19. Glowacki et al. [98] developed a systemic approach to identify and study tweets related to COVID-19 where Twitter users discussed addiction issues. Selman et al. [99]’s study focused on studying tweets where Twitter users reported their relative, friend, or acquaintance passing away from COVID-19. The study specifically focused on patients who were reported to have been alone at the time of their death. The work of Koh et al. [100] aimed to identify tweets using specific keywords where Twitter users communicated about feelings of loneliness during COVID-19. The authors tested their approach on a total of 4492 tweets. Mackey et al. [101]’s work focused on filtering and investigating tweets related to COVID-19 where people self-reported their symptoms, access to testing sites, and recovery status. In [102], the authors focused on studying tweets related to COVID-19 to understand the anxiety and panic-buying behavior with a specific focus on buying toilet paper during this pandemic. The work involved specific inclusion criteria for the tweets, and a total of 4081 tweets were studied.

As can be seen from these works involving studying social media behavior and user characteristics on Twitter during COVID-19, while there have been several innovations and advancements made in this field, the following limitations exist in these works:

1. Most of these works used approaches to detect tweets that contained one or more keywords, hashtags, or phrases such as “COVID-19”, “coronavirus”, “SARS-CoV-2”, “covid”, “corona,” etc., but none of these works focused on including one or more keywords directly related to the SARS-CoV-2 Omicron variant to include the associated tweets. As the SARS-CoV-2 Omicron variant is now responsible for most of the COVID-19 cases globally, the need in this context is to filter tweets that contain one or more keywords, hashtags, or phrases related to this variant.
2. The works on sentiment analysis [88,89] focused on the proposal of new approaches to detect the sentiment associated with tweets; however, the categories for classification of the associated sentiment were only ‘positive’, ‘negative’, and ‘neutral’. In a realistic scenario, there can be different kinds of ‘positive’ emotions, such as ‘good’ and ‘great’. Similarly, there can be different kinds of ‘negative’ emotions, such as ‘bad’ and ‘terrible’. The existing works cannot differentiate between these kinds of positive or negative emotions. Therefore, the need in this context is to expand the levels of sentiment classification to include the different kinds of positive and negative emotions.
3. While there have been multiple innovations in this field of Twitter data analysis, such as detecting trending topics [90], anomaly events [91], public perceptions towards C.D.C. [92], and views towards not wearing masks [93], just to name a few, there has been minimal work related to quantifying and ranking the associated insights.
4. The number of tweets that were included in previous studies (such as 4081 tweets in [102] and 4492 tweets in [100]) comprises a very small percentage of the total number of tweets that have been posted related to COVID-19 since the beginning

of the outbreak. Therefore, the need in this context is to include more tweets in the studies.

The work proposed in this paper aims to explore the intersections of Big Data, Natural Language Processing, Data Science, Information Retrieval, Machine Learning, and their related areas to address the abovementioned needs. The methodology is outlined in the next section.

3. Materials and Methods

This section is divided into multiple parts. Section 3.1 provides an overview of how this work is compliant with the privacy policy [103] and developer agreement [104] of Twitter. Section 3.2 provides a brief overview of the research tool that was used for this study. Section 3.3 presents the methodology of the work that was performed. Section 3.4 outlines the approach that was followed for the development of the associated Twitter dataset.

3.1. Compliance with Twitter Policies

The work primarily involves drawing insights by mining tweets. Therefore, the privacy policy [103] and developer agreement and policy [104] of Twitter were studied at first. The privacy policy of Twitter [103] states, “Twitter is public and Tweets are immediately viewable and searchable by anyone around the world”. It also states, “Most activity on Twitter is public, including your profile information, your display language, when you created your account, and your Tweets and certain information about your Tweets like the date, time, and application and version of Twitter you tweeted from The lists you create, people you follow and who follow you, and tweets you Like or Retweet are also public By publicly posting content, you are directing us to disclose that information as broadly as possible, including through our APIs, and directing those accessing the information through our APIs to do the same.” To add, the Twitter developer agreement and policy [104] defines tweets as “public data”. Therefore, based on the terms and conditions mentioned in these two policies, it can be concluded that performing this data analysis and drawing insights from tweets by connecting with the Twitter API is compliant and adheres to both these policies.

3.2. Overview of Social Bearing

The work was performed by using Social Bearing, a research tool for performing Twitter research [105]. It was developed by Tom Elliott, and the tool was made available to the public starting on 1 January 2015. The tool uses multiple JavaScript, text processing, and text analysis libraries and algorithms in combination, as well as in standalone form, for performing Big Data mining, Data Analysis, Information Processing, and Information Retrieval on tweets (obtained based on keyword or hashtag searches from Twitter). The specific JavaScript libraries used by this research tool include: jquery-ui.min.js [106], jquery.tinysort.min.js [107], masonry.pkgd.min.js [108], d3.layout.cloud.js [109], d3.min.js [110], analytics.js [111], and loader.js [112].

This is a popular tool amongst academic researchers in this field and has been used for several interdisciplinary applications that focused on the studying and analysis of tweets in the last few years [113–121]; these include the analysis of the literacy of the Twitter metaverse [113], preparing Twitter data for a Multivariate Analysis of Covariance (MANCOVA) Test [114], studying the usage of Twitter to reduce drunk driving on New Year’s Eve [115], tracking fake news related to COVID-19 vaccinations [116], studying polarized politics communicated via Twitter in India [117], assessing social support and stress in autism-focused virtual communities on Twitter [118], tracking online propaganda by Twitter usage [119], investigating the proliferation of Twitter accounts in a Higher Education setting [120], and studying market identity claims on Twitter [121].

There are several advantages of this research tool as compared to several other research tools, which were used in prior works in this field. These are outlined as follows:

1. The tool works by complying with the privacy policy, developer agreement, and developer policies of Twitter [103,104] and does not constitute any illegal, unethical, or unauthorized calling, posting, or scraping performed on the Twitter API [122].
2. The traditional approach of searching tweets based on keyword search involves a series of steps, which include setting up a Twitter developer account, obtaining the GET oauth/authorizes code, obtaining the GET oauth/authenticate code, obtaining the Bearer Token, obtaining the Twitter API Access Token and Secret, and manually entering the same in a given application (such as a Python program) to connect with the Twitter API. As this research tool is already set up to work in accordance with the privacy policy, developer agreement, and developer policies of Twitter, therefore, the manual entry of all the above-mentioned codes and tokens is not required, and the process of connecting with the Twitter API to search tweets is simplified; it can be performed by just clicking the Twitter Sign-In button on the visual interface, and then the user is directed to sign into Twitter with an active Twitter account, thereafter, the tool is set up to work by searching tweets based on the keyword(s) or the hashtag(s).
3. The tool uses the in-built JavaScript, text processing and text analysis libraries and algorithms to extract characteristics from the relevant tweets and displays the same on the visual interface in the form of results, thereby reducing the time and complexity of developing these algorithms and functions from scratch as well as for developing the approaches for visualizing the results.
4. While displaying the results, the tool shows a list of users (with their Twitter usernames), who posted the tweets, in the “All Contributors” section. It allows removing one or more users from this list (if the usernames indicate that the Twitter profiles are bots) so that the new results are obtained based on the tweets posted by the rest of the users.

In addition to the above characteristic features that uphold the applicability of this research tool for different use cases, another reason for the popularity of this research tool can be attributed to the fact that there is no software development, API development, or library development necessary when the tool is used for a study such as this one. The users of this tool are only required to be familiar with the functionalities of the tool and how to use the same for collecting, studying, and analyzing Twitter data. As this research tool was used for this study, this paper does not report any code, program, software, library, or API as a supplementary item on any open-source repository such as GitHub.

3.3. Methodology

As can be seen from Figure 1, the first step was to perform Big Data mining to collect these tweets from Twitter. The Twitter API standard search has a 7-day limit on the tweets that can be searched [123]; in other words, tweets posted more than 7-days ago cannot be searched and studied. Given this, a total of 12,028 tweets about the SARS-CoV-2 Omicron variant that were posted on Twitter between 12 May 2022 (the most recent date at the time of submission of this paper) and 5 May 2022 (the date up to which tweets could be searched in compliance with the Twitter standard search API guidelines) were studied. This data mining process was performed by filtering tweets from Twitter in this date range that contained the “Omicron” keyword or hashtag. This was performed to capture tweets that comprised views, expressions, opinions, perspectives, attitudes, news, information, and similar conversation themes related to the SARS-CoV-2 Omicron variant, where, other than the “Omicron” keyword, the rest of the words used in the respective tweets would be associated with such information.

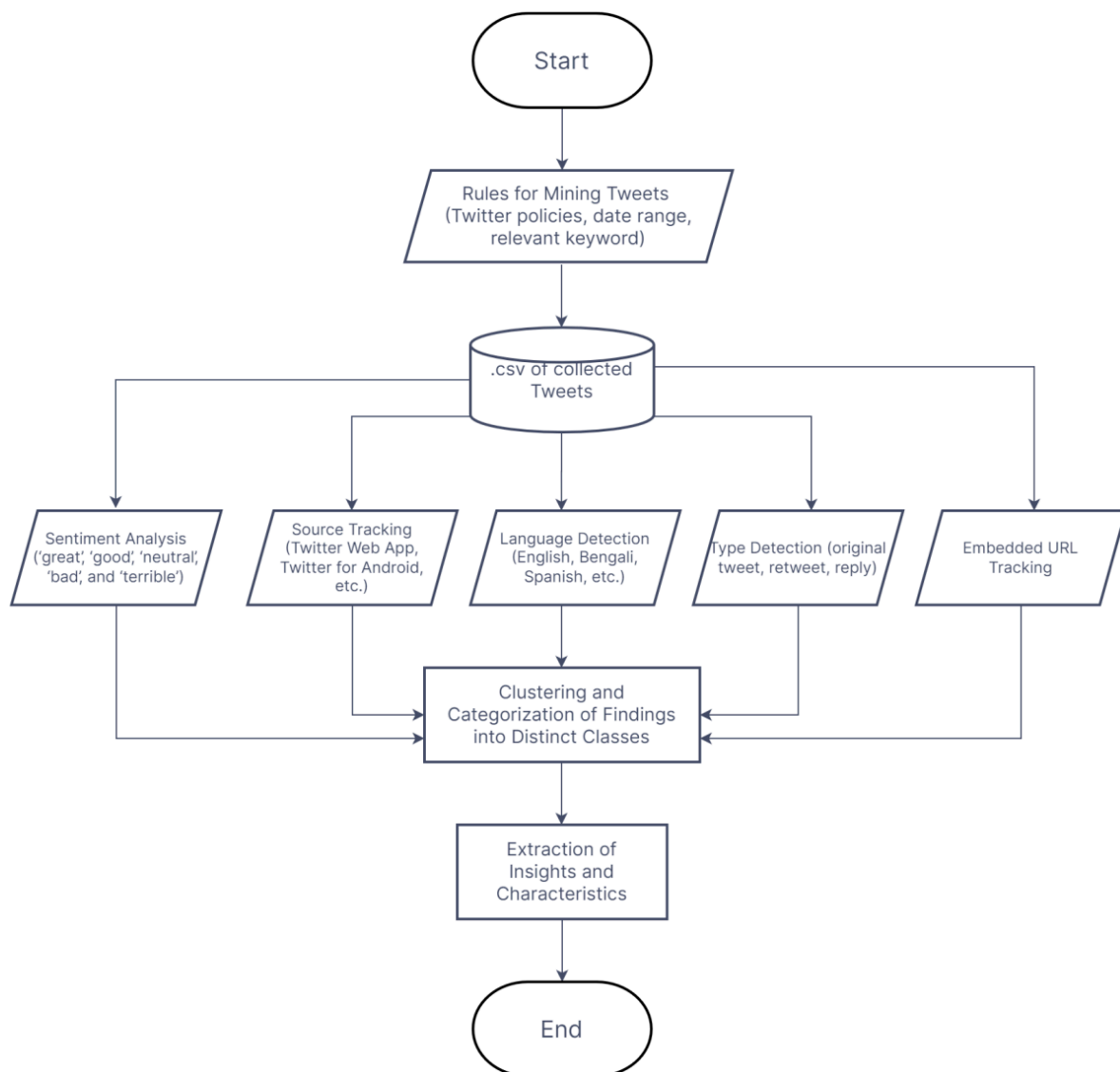


Figure 1. Flowchart-based representation of the methodology that was followed.

Multiple random subsamples of the data (consisting of 7 tweets in each subsample) were studied to evaluate the commonly used phrases to refer to the SARS-CoV-2 Omicron variant. A few examples of these phrases were: “first case of omicron variant”, “approaching its Omicron peak”, “Omicron is a stone cold killer”, “case of Omicron sub-variant”, “new omicron sub-variants”, “the highly transmissible omicron variant”, and “sinus pain of omicron”. Even though these phrases are different, the keyword “Omicron” is present in all these phrases. This helps to uphold the relevance of performing the data mining and data collection of relevant tweets by using this keyword. The process of data mining that was followed was not case sensitive, so the tweets containing the “Omicron” keyword (or hashtag) as “omicron” or “Omicron” or “OMICRON” or any other order of capitalization of these alphabets were also mined, and these specific cases did not need to be specifically mentioned as separate or different keywords (or hashtags).

For all these tweets (available as a .csv file after the data mining process was completed), the specific insights that were studied, computed, interpreted, analyzed, quantified, and ranked in the next steps included the sentiment, source tracking, detecting the language of the tweet, inferring the type of tweet, and detecting the URLs embedded in the tweets. The Social Bearing tool’s visual interface shows all the Twitter user profiles whose tweets are being considered for a specific use-case scenario for which the tool is being used. The

Twitter usernames of these profiles are listed in the “All contributors” section. Prior to performing this comprehensive analysis, the “All contributors” section of the tool was thoroughly reviewed and the accounts that had Twitter handles or usernames that indicated that those accounts are likely to be bots were manually unselected. As a result, the tweets from these accounts were not included for the analysis. In addition to this, in the “Other Options” section on Social Bearing’s visual interface, there is a checkbox—“Remove Duplicate Tweets”; we had checked this checkbox during the study and the tool automatically removed the duplicate tweets from the results. In other words, duplicate tweets did not play a role in the findings that are presented. The process of sentiment analysis involved the tokenization and lemmatization of the text of the tweet and classifying it into various sentiments such as ‘great’, ‘good’, ‘neutral’, ‘bad’, and ‘terrible’. While there can be multiple ways of performing sentiment analysis, this process of sentiment analysis was selected as multiple sentiment classes can be created, and the resultant classification is neither a binary classification between ‘positive’ and ‘negative’ sentiments nor is it a three-level classification between ‘positive’, ‘negative’, and ‘neutral’ sentiment classes, which have been the limitations in prior works in this field (discussed in Section 2). The source tracking aspect of the study involved tracking the publicly available “source label” [124] that Twitter associates with each tweet. This source level has different categories, which include Twitter Web App, Twitter for Android, Twitter for iPhone, and several more. Table 1 explains the condition for the assignment of four such “source labels” by Twitter.

Table 1. Four tweet Source Labels and the conditions for assignment of these labels by Twitter.

| Tweet Source Label | Condition for Assignment |
|---------------------------|--|
| Twitter Web App | The tweet was posted by visiting the official website of Twitter [125] |
| Twitter for Android | The tweet was posted using the Twitter app for Android operating systems, which is available for free download on the Google Playstore [126] |
| Twitter for iPhone | The tweet was posted using the Twitter app for iPhone, which is available for free download on the App Store [127] |
| TweetDeck | The tweet was posted by using TweetDeck, a social media dashboard application for the management of Twitter accounts [128] |

The Twitter platform allows its users to tweet in any of the 34 languages supported by Twitter [129]. Each of these languages is assigned a unique two-letter code in the Twitter developer API, which helps to uniquely identify the associated language. Some examples include “en” for English, “ar” for Arabic, “bn” for Bengali, “cz” for Czech, “da” for Danish, “de” for German, “el” for Greek, “es” for Spanish, “fa” for Persian, and so on.

Inferring the type of tweet involved tracking the publicly available information about a tweet on Twitter that mentions whether it is an original tweet or a retweet (an original tweet that has been re-posted) or a reply (response to an original tweet or a retweet). Finally, detecting the embedded URL in a tweet involved processing the text associated with a tweet to detect any domain(s) that may have been included in the tweet.

After detecting these characteristics and features from the set of 12,028 tweets, in the next step, they were grouped together into distinct classes for quantification, categorization, and ranking. This helped to deduce multiple insights from each category of the results. For instance, in the tweet type category, this analysis helped to deduce the percentage of original tweets, retweets, and replies, which further helped in ranking these specific classes of results in the tweet-type category. The ranking process helped to determine which of these respective classes constituted the maximum occurrence in each category. The above-mentioned functionalities for data mining, sentiment analysis, source tracking, language interpretation, type detection, and embedded URL analysis of tweets were performed in a collective manner per the flowchart shown in Figure 1 by using Social Bearing, as mentioned earlier in this section. The results that were obtained are discussed in Section 4.


3.4. Data Availability

As described in Section 3.3, this work uses the Social Bearing research tool that works by complying with the Twitter standard search policies. The policies of Twitter standard search state —“Keep in mind that the search index has a 7-day limit. In other words, no tweets will be found for a date older than one week” [130]. Therefore, all the tweets that were analyzed in this study were posted in the range of 5 May 2022 to 12 May 2022 (the most recent date at the time of submission of this paper). For supporting similar works such as — the replication of this study, the repetition of this study on a bigger scale (by including tweets that were posted before 5 May 2022), and for the investigation of similar research questions based on studying tweets about the Omicron variant, we have developed a dataset that contains more than 500,000 Tweet IDs of the same number of tweets posted about the Omicron variant since the first detected case of this variant on 24 November 2021.

This data collection was performed by the newly added Advanced Search feature provided by the Twitter platform [131], which allows searching of Tweets between any two given dates (the dates can be more than a week(s) or a month(s) apart) based on keyword search. As this is a feature developed by Twitter to support research and development using Twitter [132], it can be safely concluded that the usage of Twitter Advanced Search follows Twitter privacy policy and Twitter developer guidelines [103,104]. Furthermore, it can also be concluded that the process by which Twitter Advanced Search fetches tweets doesn't constitute any illegal, unethical, or unauthorized calling, posting, or scraping performed on the Twitter API [122].

The usage of Twitter Advanced Search involves logging into Twitter and then entering the relevant keyword(s) and date range to develop a RegEx or regular expression. This RegEx is then used by the Twitter Advanced Search to display the relevant tweets as per the rules mentioned in the regular expression. This RegEx, along with the different parts of the same, that were used for the development of this dataset is shown in Figure 2.

https://twitter.com/search?q=%22%22omicron%22%22%20until%3A2022-05-12%20since%3A2021-11-24&src=typed_query&f=live



The diagram illustrates the components of the URL: `https://twitter.com/search?q=%22%22omicron%22%22%20until%3A2022-05-12%20since%3A2021-11-24&src=typed_query&f=live`. Brackets below the URL identify five parts: `twitter.com` (domain), `search` (search keyword), `q=%22%22omicron%22%22%20until%3A2022-05-12%20since%3A2021-11-24` (end date), `&src=typed_query` (start date), and `&f=live` (source).

Figure 2. Representation of different parts of the RegEx that was used in Twitter Advanced Search.

As can be seen from Figure 2, this RegEx is comprised of five parts. From left to right, in Figure 2, these parts represent the `twitter.com` domain, the search keyword, the end date, the start date, and the source. The search keyword was “Omicron”, the end date was selected to be 12 May 2022—as this is the most recent date at the time of submission of this paper—and the start date was selected to be 24 November 2021—as, on this date, the first case of the Omicron variant was detected. The source parameter set as “typed_query” helped in the elimination of potential bot-related content, where the underlining bots shared anything on Twitter that was not typed (such as images, advertisements, spam links, etc.). The results from the Twitter Advanced Search API were collected, compiled, and exported to develop this dataset. The dataset developed contains only Tweet IDs in compliance with the privacy policy, developer agreement, and guidelines for content redistribution of Twitter [103,104]. This section is divided into four subsections. The compliance of this dataset with the privacy policy, developer agreement, and guidelines for content redistribution of Twitter is described in Section 3.4.1. The compliance of this dataset with the FAIR (Findability, Accessibility, Interoperability, and Reusability) principles for scientific data management [132] is outlined in Section 3.4.2. Section 3.4.3 presents the data description and provides the URL that can be used to access and download the dataset. The instruction for using this dataset is presented in Section 3.4.4. It is worth mentioning here that Twitter Advanced Search does not return an exhaustive list of Tweets posted between two specific dates. So, it is possible that multiple tweets posted between 24 November 2021

and 12 May 2022 were not returned by Twitter Advanced Search when the data collection was performed. Therefore, these tweets are not a part of this dataset. In addition to this, Twitter allows users the option to delete a tweet, which would mean that there would be no retrievable Tweet text and other related information (upon hydration, which is explained in Section 3.4.4) for a Tweet ID of a deleted tweet. All the Tweet IDs available in this dataset correspond to tweets that have not been deleted at the time of writing this paper.

3.4.1. Compliance with Guidelines for Twitter Content Redistribution

The privacy policy of Twitter [103] states, “Twitter is public and tweets are immediately viewable and searchable by anyone around the world”. The guidelines for Twitter content redistribution [104] state, “If you provide Twitter Content to third parties, including downloadable datasets or via an API, you may only distribute tweet IDs, Direct Message IDs, and/or User IDs (except as described below)”. It also states, “We also grant special permissions to academic researchers sharing tweet IDs and User IDs for non-commercial research purposes. Academic researchers are permitted to distribute an unlimited number of tweet IDs and/or User IDs if they are doing so on behalf of an academic institution and for the sole purpose of non-commercial research.” Therefore, it may be concluded that mining relevant tweets from Twitter to develop a dataset (comprising only Tweet IDs) to share the same is in compliance with the privacy policy, developer agreement, and content redistribution guidelines of Twitter.

3.4.2. Compliance with FAIR

The FAIR principles for scientific data management [132] state that a dataset should have findability, accessibility, interoperability, and reusability. The dataset is findable as it has a unique and permanent DOI. The dataset is accessible online. It is interoperable due to the use of .txt files for data representation that can be downloaded, read, and analyzed across different computer systems, devices, and applications. The dataset is reusable as the associated tweets and related information, such as User ID, username, retweet count, etc., for all the Tweet IDs can be obtained by the process of hydration, in compliance with Twitter policies (Section 3.4.4), for data analysis and interpretation.

3.4.3. Data Description

This section presents the description of the dataset that is publicly available at <https://doi.org/10.5281/zenodo.6893676>. The dataset consists of a total of 522,886 Tweet IDs of the same number of tweets about the Omicron variant of COVID-19 that were posted on Twitter from 24 November 2021 to 12 May 2022. The Tweet IDs are presented in seven different .txt files based on the timelines of the associated tweets. Table 2 provides the details of these dataset files.

Table 2. Description of all the files present in this dataset.

| Filename | No. of Tweet IDs | Date Range of the Tweet IDs |
|-----------------------|------------------|--------------------------------------|
| TweetIDs_November.txt | 16471 | 24 November 2021 to 30 November 2021 |
| TweetIDs_December.txt | 99288 | 1 December 2021 to 31 December 2021 |
| TweetIDs_January.txt | 92860 | 1 January 2022 to 31 January 2022 |
| TweetIDs_February.txt | 89080 | 1 February 2022 to 28 February 2022 |
| TweetIDs_March.txt | 97844 | 1 March 2022 to 31 March 2022 |
| TweetIDs_April.txt | 91587 | 1 April 2022 to 20 April 2022 |
| TweetIDs_May.txt | 35756 | 1 May 2022 to 12 May 2022 |

To comply with the privacy policy, developer agreement, and guidelines for content redistribution of Twitter [103,104], only the Tweet IDs associated with these 522,886 tweets are presented in this dataset. To obtain the detailed information associated with each of these tweets, such as the tweet text, username, User ID, timestamp, retweet count, etc., these Tweet IDs need to be hydrated. There are several applications, such as the Hydrator app [133], Social Media Mining Toolkit [134], and Twarc [135], that work by complying with

Twitter policies and may be used for hydrating the Tweet IDs in this dataset. A step-by-step process for using one of these applications, the Hydrator app for hydrating the files in this dataset, is presented in Section 3.4.4.

3.4.4. Instructions for Using the Dataset

The following is the step-by-step process for using one of these applications, the Hydrator app [133], to hydrate this dataset or, in other words, to obtain the text of the tweet, User ID, username, retweet count, language, tweet URL, source, and other public information related to all the Tweet IDs present in this dataset. The Hydrator app works in compliance with the policies for accessing and calling the Twitter API.

1. Download and install the desktop version of the Hydrator app from <https://github.com/DocNow/hydrator/releases> (accessed on 14 May 2022).
2. Click on the “Link Twitter Account” button on the Hydrator app to connect the app to an active Twitter account.
3. Click on the “Add” button to upload one of the dataset files (such as Tweet IDs_November.txt). This process adds a dataset file to the Hydrator app.
4. If the file upload is successful, the Hydrator app will show the total number of Tweet IDs present in the file. For instance, for the file—“TweetIDs_November.txt”, the app would show the number of Tweet IDs as 16,471.
5. Provide details for the respective fields: Title, Creator, Publisher, and URL in the app, and click on “Add Dataset” to add this dataset to the app.
6. The app will automatically redirect to the “Datasets” tab. Click on the “Start” button to start hydrating the Tweet IDs. During the hydration process, the progress indicator will increase, indicating the number of Tweet IDs that have been successfully hydrated and the number of Tweet IDs that are pending hydration.
7. After the hydration process ends, a .jsonl file will be generated by the app that the user can choose to save on the local storage.
8. The app would also display a “CSV” button in place of the “Start” button. Clicking on this “CSV” button would generate a .csv file with detailed information about the tweets, which would include the text of the tweet, User ID, username, retweet count, language, tweet URL, source, and other public information related to the tweet.
9. Repeat steps 3–8 for hydrating all the files of this dataset.

4. Results and Discussions

This section presents and discusses the results obtained upon the development and implementation of the proposed methodology (Section 3.3) on the corpus of 12,028 tweets about the SARS-CoV-2 Omicron variant posted on Twitter from 5 May 2022 to 12 May 2022. The results are shown in Figures 3–7. Figure 3 shows the results of the sentiment analysis. The specific categories into which the sentiment of a tweet was classified comprised ‘great’, ‘good’, ‘neutral’, ‘bad’, and ‘terrible’. As can be seen from Figure 3, the ‘neutral’ emotion was present in a majority of the tweets (50.5% of the total tweets). It was followed by tweets that had the ‘bad’ (15.5% of the total tweets) and ‘good’ (14.0% of the total tweets) emotions associated with them. These respective sentiment categories were followed by the sentiment categories of ‘terrible’ (12.5% of the total tweets) and ‘great’ (7.5% of the total tweets).

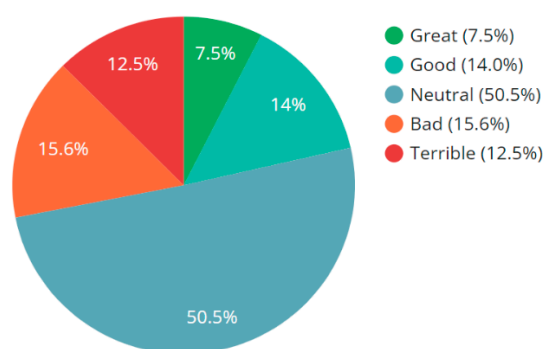


Figure 3. Results of sentiment analysis performed on tweets about the SARS-CoV-2 Omicron variant.

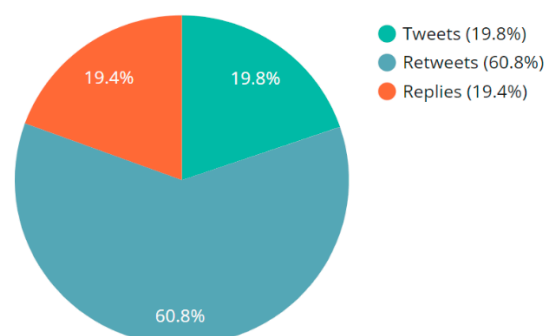


Figure 4. Results of tweet type detection performed on tweets about the SARS-CoV-2 Omicron variant.

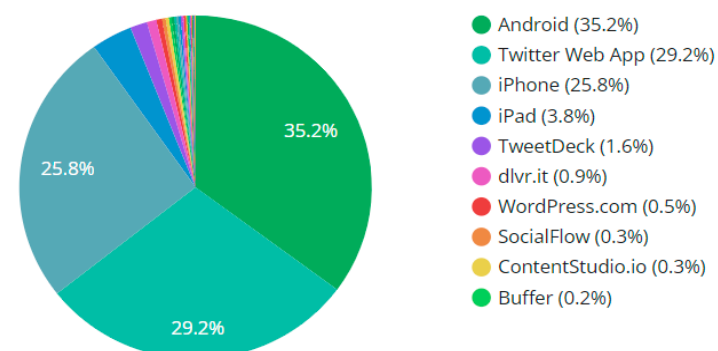


Figure 5. Results of tweet source detection performed on tweets about the SARS-CoV-2 Omicron variant.

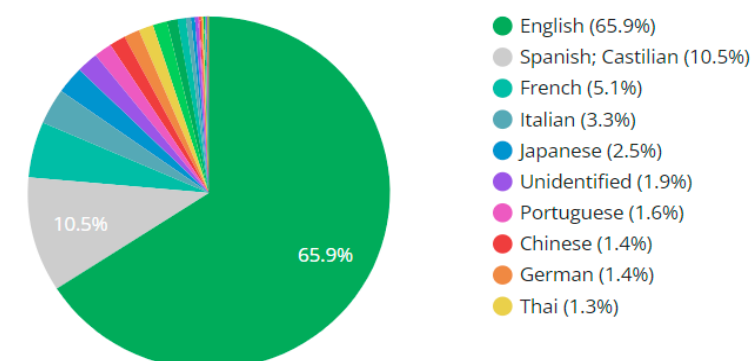


Figure 6. Results of language detection performed on tweets about the SARS-CoV-2 Omicron variant.

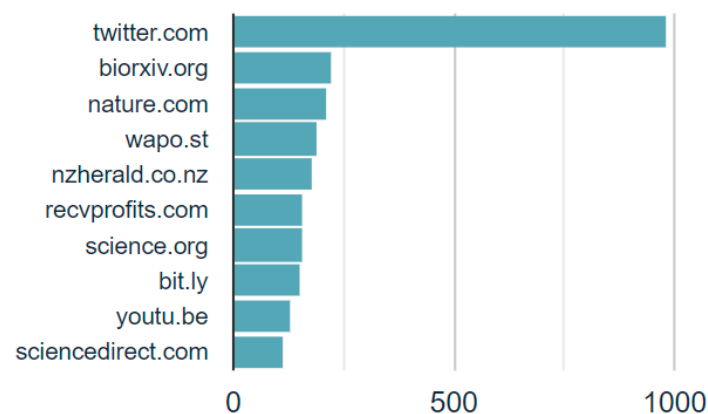


Figure 7. Results of embedded URL detection performed on tweets about the SARS-CoV-2 Omicron variant.

The results of the tweet type detection are shown in Figure 4. This process involved the detection, categorization, and ranking of the tweets into original tweets, retweets, and replies. In the index shown in this Figure, original tweets are referred to as “tweets”. As can be seen from this Figure, retweets comprised a majority (60.8%) of all the tweets about the SARS-CoV-2 Omicron variant. It was followed by original tweets (19.8% of the total tweets) and replies (19.4% of the total tweets).

Figure 5 shows the results of the tweet source detection as publicly shown by the Twitter platform. Here, “Android” refers to “Twitter for Android”, “iPhone” refers to “Twitter for iPhone”, and “iPad” refers to “Twitter for iPad” (meanings of these categories are mentioned in Table 1). The total number of sources in this corpus of 12,028 tweets was observed to be more than 10. So, only the top 10 sources are listed in the index provided in this Figure for clarity and readability. As can be seen from this Figure, “Twitter for Android” accounted for the most number of tweets (35.2% of the total tweets). It was followed by “Twitter Web App” (29.2% of the total tweets), “Twitter for iPhone” (25.8% of the total tweets), “Twitter for iPad” (3.8% of the total tweets), and “TweetDeck” (1.6% of the total tweets). “TweetDeck” was followed by a few other sources, which accounted for less than 1% of the total tweets.

The results of the language detection and ranking of the used languages are shown in Figure 6. The total number of languages detected from all these 12,028 tweets was observed to be more than 10. So, only the top 10 languages are listed in the index provided in this Figure for clarity and readability. As can be seen from this Figure, more than a majority (65.9%) of the tweets were written in English. In second place was Spanish or Castilian (10.5% of the total tweets), which was followed by French (5.1% of the total tweets), Italian (3.3% of the total tweets), Japanese (2.5% of the total tweets), and a few other sources that accounted for less than 2% of the total tweets.

The results of the detection of the embedded URLs and the ranking of the associated clusters (domains) are shown in Figure 7. The total number of different domains in this corpus of 12,028 tweets was observed to be more than 10. So, only the top 10 embedded domains are listed in the index provided in this Figure for clarity and readability. As can be seen from Figure 7, the domain, twitter.com, comprised the highest count. This can be attributed to the fact that retweets comprised a significant percentage of the tweets about the SARS-CoV-2 Omicron variant. It was followed by the domains: biorxiv.org, nature.com, wapo.st, nzherald.co.nz, recvprofits.com, science.org, bit.ly, YouTube (youtu.be is a shortened version of YouTube URLs [136]), and sciencedirect.com.

In addition to the above, a keyword-based word cloud analysis and a hashtag-based word cloud analysis were performed on this corpus of 12,028 tweets about the SARS-CoV-2 Omicron variant. The results are shown in Figures 8 and 9, respectively. In each of these word clouds, a different font color has been used to identify a keyword and hashtag,

respectively. The font size of each of these words and hashtags is directly proportional to the frequency of the same. In other words, the keyword (or hashtag) that has the highest frequency has the largest font size, and the keyword (or hashtag) that has the lowest frequency has the smallest font size. Those keywords (or hashtags) that had a very low frequency and would subsequently have an extremely small font size were not included in the results shown here to avoid cluttering and to enhance the readability of these Figures. As can be seen from both these figures, the “omicron” keyword was the most frequent keyword, and similarly, the “#omicron” hashtag was the most frequent hashtag across all these tweets. This observation further helps to support the relevance of the proposed approach for searching tweets by using “omicron” as the keyword or hashtag to be searched, which was outlined at the beginning of Section 3.3.

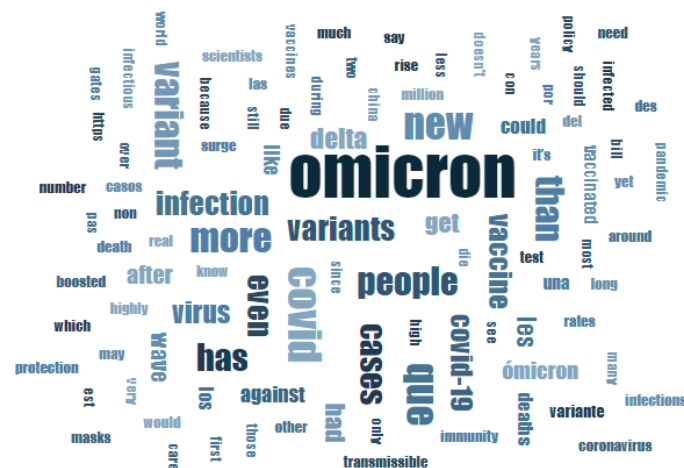


Figure 8. Results of word cloud analysis of the keywords present in the tweets about the SARS-CoV-2 Omicron variant.

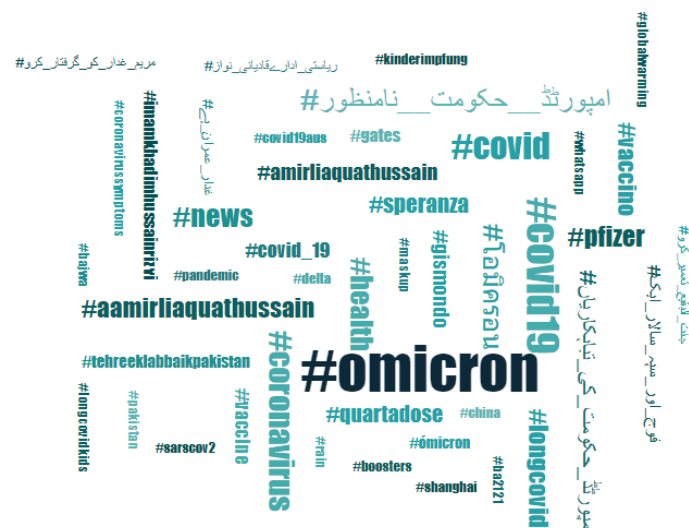


Figure 9. Results of word cloud analysis of the hashtags present in the tweets about the SARS-CoV-2 Omicron variant.

In addition to these findings, a discussion is presented next that upholds how this work addresses the multiple distinct needs in this field of research that were identified after a comprehensive review of recent works (presented in Section 2).

1. The previous works in this field proposed approaches to filter tweets based on one or more keywords, hashtags, or phrases such as “COVID-19”, “coronavirus”, “SARS-CoV-2”, “covid”, “corona” but did not contain any keyword or phrase speci-

cally related to the Omicron variant. Given this, those approaches for tweet searching or tweet filtering might not be applicable to collect the tweets posted about the Omicron variant unless the Twitter user specifically mentions something like “COVID-19 omicron variant” or “SARS-CoV-2 Omicron variant” in their tweets. As discussed in Section 3, there were multiple instances when the Twitter users did not use keywords, hashtags, or phrases such as “COVID-19”, “coronavirus”, “SARS-CoV-2”, “covid”, “corona” along with the keyword or hashtag “Omicron”. Thus, the need is to develop an approach to specifically mine tweets posted about the Omicron variant. This work addresses this need by proposing a methodology that searches tweets based on the presence of “Omicron” either as a keyword or as a hashtag. The effectiveness of this approach is justified by the word clouds presented in Figures 8 and 9.

2. The prior works [88,89] on sentiment analysis of tweets about COVID-19 focused on developing approaches for classifying the sentiment only into three classes—‘positive’, ‘negative’, and ‘neutral’. In a realistic scenario, there can be different kinds of ‘positive’ emotions, such as ‘good’ and ‘great’. Similarly, there can be different kinds of ‘negative’ emotions, such as ‘bad’ and ‘terrible’. The existing works cannot differentiate between these kinds of positive or negative emotions. To address the need in this context, associated with increasing the number of classes for classification of the sentiment of the tweet, this work proposes an approach that classifies tweets into five sentiment classes: ‘great’, ‘good’, ‘neutral’, ‘bad’, and ‘terrible’ (Figure 3).
3. The emerging works in this field, for instance, related to detecting trending topics [90], anomaly events [91], public perceptions towards C.D.C. [92], and views towards not wearing masks [93], focused on the development of new frameworks and methodologies without focusing on quantifying the multimodal components of the characteristics of the tweets and ranking these characteristics to infer insights about social media activity on Twitter due to COVID-19. This work addresses this need. The results from sentiment analysis, type detection, source tracking, language interpretation, and embedded URL observation were categorized into distinct categories, and these categories were ranked in terms of the associated characteristics to infer meaningful and relevant insights about social media activity on Twitter related to tweets posted about the SARS-CoV-2 Omicron variant (Figures 3–7). For instance, for tweet type analysis, the findings of this study show that “Twitter for Android” accounted for the most number of tweets (35.2% of the total tweets), which was followed by “Twitter Web App” (29.2% of the total tweets), “Twitter for iPhone” (25.8% of the total tweets), and other sources.
4. The previous works centered around performing data analysis on tweets related to COVID-19, included a small corpus of tweets, such as 4081 tweets in [102] and 4492 tweets in [100]. In view of the number of active Twitter users and the number of tweets posted each day, there is a need to include more tweets in the data analysis process. This work addresses this need by considering a total of 12,028 relevant tweets that had a combined reach of 149,500,959, with 226,603,833 impressions, 1,053,869 retweets, and 3,427,976 favorites.
5. The development of Twitter datasets has been of significant importance and interest to the scientific community in the areas of Big Data mining, Data Analysis, and Data Science. This is evident from the recent Twitter datasets on 2020 U.S. Presidential Elections [72], 2022 Russia–Ukraine war [73], climate change [74], natural hazards [75], European Migration Crisis [76], movies [77], toxic behavior amongst adolescents [78], music [79], civil unrest [80], drug safety [81], and Inflammatory Bowel Disease [82]. Twitter datasets help to serve as a data resource for a wide range of applications and use cases. For instance, the Twitter dataset on music [79] has helped in the development of a context-aware music recommendation system [137], next-track music recommendations as per user personalization [138], session-based music recommendation algorithms [139], music recommendation systems based on the use of affective hashtags [140], music chart predictions [141], user-curated playlists [142], sentiment

analysis of music [143], listener engagement with popular songs [144], culture aware music recommendation systems [145], mining of user personalities [146], and several other applications. The works related to the development of Twitter datasets on COVID-19 [83–87] in the last few months did not focus on the development of a Twitter dataset comprising tweets about the Omicron variant of COVID-19 since the first detected case of this variant. To address this research gap, we developed a Twitter dataset (Section 3.4) that comprises 522,886 Tweet IDs of the same number of tweets about the Omicron variant of COVID-19 since the first detected case of this variant on 24 November 2021.

As per the best knowledge of the authors, such a comprehensive analysis of tweets about the Omicron variant of COVID-19 has not been done before. The results presented and discussed support the novelty and relevance of this research work. However, the paper has a few limitations. First, the research tool, Social Bearing, that was used utilizes several JavaScript libraries, which include jquery-ui.min.js [106], jquery.tinysort.min.js [107], masonry.pkgd.min.js [108], d3.layout.cloud.js [109], d3.min.js [110], analytics.js [111], and loader.js [112], for performing a multimodal analysis of tweets as presented in this paper. We did not develop any algorithm or software solution of our own to perform the same analysis and/or verify the results presented by this research tool. If any algorithm or software solution is developed that uses the same methodology of integrating the functionality of these JavaScript libraries as the Social Bearing research tool, then it is likely that the results would be identical. However, use of other methodologies such as Bidirectional Long Short-Term Memory (Bi-LSTM), Bidirectional Encoder Representations from Transformers (BERT), Word Vector Analysis, Topic Modeling, Lexicon-based analysis, Dialogue Bidirectional Encoder Representations from Transformers (DialBERT), and Deep Learning, in such an algorithm or software solution, might provide slightly different results in terms of the number of entities classified into each bin during the classification process. For instance, our approach shows that 50.5% of the tweets had a ‘neutral’ emotion. Any of these alternate approaches might provide a number very close to 50.5% but not exactly equal to 50.5%. To address this limitation, we plan to develop, implement, and test all these alternative approaches on this dataset to identify the best approach for performing such a comprehensive analysis on tweets about the SARS-CoV-2 Omicron variant. Second, as the SARS-CoV-2 Omicron variant continues to spread across the globe, rapid advances are being made related to Omicron-specific vaccines [147,148], and new studies related to this variant are also getting published [149,150]. These advances, studies, new findings, and the nature of public reaction, views, opinion, feedback, thoughts, and perspectives towards the same may impact one or more of the characteristic features of social media behavior on Twitter that were investigated and analyzed in this study. To address this limitation, a follow-up study will be conducted in the future. In that follow-up study, the aim would be to include all the relevant tweets in between the dates when the first case of Omicron was recorded and when the last case of Omicron would be recorded to perform this exploratory analysis once again to compare the associated findings and to comment on any similarities and dissimilarities in the insights that might be observed. Third, the in-built feature of Social Bearing that was used to remove potential bot-related content might not be the most optimal approach for the removal of all bot-related tweets, as certain advanced bot accounts that exactly mimic real user Twitter accounts in terms of Twitter usernames and share/post content on Twitter after randomized time intervals (as opposed to certain bots which share/post content after a fixed time interval) might be difficult to identify using this approach. Emerging works in the field of Natural Language Processing, such as TweepzBot [151], Bot-DenseNet [152], Bot2Vec [153], Botter [154], and GlowWorm-based Generalized Regression [155], could be used for identifying such advanced bot accounts on Twitter. We could not implement any of these emerging works in our study due to the limited integration options provided by the Social Bearing research tool. We plan to address this limitation in the future by developing our own algorithm and software solution that would not have any such limited integration options.

5. Conclusions

Since the initial outbreak in Wuhan, China, in December 2019, the SARS-CoV-2 virus has resulted in a total of 549,667,293 cases and 6,352,025 deaths on a global scale. The virus has undergone multiple mutations, and as a result, several variants have been detected, such as Alpha, Beta, Gamma, Delta, Epsilon, Eta, Iota, Kappa, Zeta, Mu, and Omicron, in different parts of the world. Out of these variants, the Omicron variant, a variant of concern (VOC) as per WHO, is currently the globally dominant variant and has been reported to be the most immune-evasive VOC of SARS-CoV-2 and is also considered to be highly resistant against antibody-mediated neutralization. The number of cases and deaths due to Omicron in different parts of the world is on a constant rise.

Research conducted during pandemics in the past suggests that people extensively use social media platforms for sharing information, news, views, opinions, ideas, knowledge, feedback, and experiences related to the pandemic they are facing. Twitter, one such social media platform, is popular amongst all age groups. Therefore, this work took a comprehensive approach to identify, study, and analyze tweets related to the SARS-CoV-2 Omicron variant to understand, categorize, and interpret the associated dynamics and characteristic features of social media behavior. A total of 12,028 tweets about the SARS-CoV-2 Omicron variant were mined from Twitter, and the associated sentiment ('great', 'good', 'neutral', 'bad', and 'terrible'), type (original tweet, retweet, or reply), source (such as "Twitter for Android", "Twitter Web App", "Twitter for iPhone", etc.), language (such as English, Spanish, Bengali, etc.), and embedded URLs (URLs that were included in the tweet text) were analyzed. The findings from this exploratory study are manifold. First, a majority of the tweets had a 'neutral' emotion (50.5% of the total tweets), which was followed by 'bad' (15.6% of the total tweets), 'good' (14.0% of the total tweets), 'terrible' (12.5% of the total tweets), and 'great' (7.5% of the total tweets). Second, 35.2% of the tweets had "Twitter for Android" as their source. It was followed by the "Twitter Web App" (29.2% of the total tweets), "Twitter for iPhone" (25.8% of the total tweets), "Twitter for iPad" (3.8% of the total tweets), "TweetDeck" (1.6% of the total tweets), and other sources that accounted for less than 1% of the total tweets. Third, a majority of the tweets (65.9%) were posted in English, which was followed by Spanish or Castilian (10.5% of the total tweets), French (5.1% of the total tweets), Italian (3.3% of the total tweets), Japanese (2.5% of the total tweets), and other languages that accounted for less than 2% of the tweets. Fourth, the majority of the tweets (60.8%) in terms of tweet type were retweets, which were followed by original tweets (19.8% of the total tweets) and replies (19.4% of the total tweets). Fifth, in terms of embedded domains, the most common domain embedded in the tweets was found to be twitter.com, which was followed by biorxiv.org, nature.com, wapo.st, nzherald.co.nz, recvprofits.com, science.org, and a few other sources. Finally, to support research and development in this field, this work presents an open-access Twitter dataset of 522,886 Tweet IDs of the same number of tweets about the SARS-CoV-2 Omicron variant that was posted on Twitter since the first detected case of this variant on 24 November 2021. Future work would involve addressing the limitations of this study, as discussed at the end of the previous section. In addition to this, the dataset associated with this work will be updated on a routine basis so that the research community has access to the most recent data in this regard.

Author Contributions: Conceptualization, N.T.; methodology, N.T.; software, N.T.; validation, N.T.; formal analysis, N.T.; investigation, N.T.; resources, N.T.; data curation, N.T.; visualization, N.T.; data analysis and results, N.T.; writing—original draft preparation, N.T.; writing—review and editing, N.T.; supervision, Not Applicable; project administration, C.Y.H.; funding acquisition, Not Applicable. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are publicly available at <https://doi.org/10.5281/zenodo.6893676>.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Ksiazek, T.G.; Erdman, D.; Goldsmith, C.S.; Zaki, S.R.; Peret, T.; Emery, S.; Tong, S.; Urbani, C.; Comer, J.A.; Lim, W.; et al. A Novel Coronavirus Associated with Severe Acute Respiratory Syndrome. *N. Engl. J. Med.* **2003**, *348*, 1953–1966. [CrossRef] [PubMed]
2. Fauci, A.S.; Lane, H.C.; Redfield, R.R. Covid-19—Navigating the Uncharted. *N. Engl. J. Med.* **2020**, *382*, 1268–1269. [CrossRef] [PubMed]
3. Wilder-Smith, A.; Chiew, C.J.; Lee, V.J. Can we contain the COVID-19 outbreak with the same measures as for SARS? *Lancet Infect. Dis.* **2020**, *20*, e102–e107. [CrossRef]
4. Cucinotta, D.; Vanelli, M. WHO declares COVID-19 a pandemic. *Acta Biomed.* **2020**, *91*, 157–160. [CrossRef]
5. COVID Live—Coronavirus Statistics—Worldometer. Available online: <https://www.worldometers.info/coronavirus/> (accessed on 14 May 2022).
6. Huang, C.; Wang, Y.; Li, X.; Ren, L.; Zhao, J.; Hu, Y.; Zhang, L.; Fan, G.; Xu, J.; Gu, X.; et al. Clinical Features of Patients Infected with 2019 Novel Coronavirus in Wuhan, China. *Lancet* **2020**, *395*, 497–506. [CrossRef]
7. Luring, A.S.; Hodcroft, E.B. Genetic Variants of SARS-CoV-2—What Do They Mean? *JAMA* **2021**, *325*, 529–531. [CrossRef] [PubMed]
8. Khare, S.; Gurry, C.; Freitas, L.; Schultz, M.B.; Bach, G.; Diallo, A.; Akite, N.; Ho, J.; Lee, R.T.; Yeo, W.; et al. GISAID’s Role in Pandemic Response. *China CDC Wkly.* **2021**, *3*, 1049–1051. [CrossRef]
9. Global Initiative on Sharing All Influenza Data. GISAID—Initiative. Available online: <https://www.gisaid.org/> (accessed on 14 May 2022).
10. Centers for Disease Control and Prevention. SARS-CoV-2 Variant Classifications and Definitions. Available online: <https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html> (accessed on 14 May 2022).
11. World Health Organization. Classification of Omicron (B.1.1.529): SARS-CoV-2 Variant of Concern. Available online: [https://www.who.int/news/item/26-11-2021-classification-of-omicron-\(b.1.1.529\)-sars-cov-2-variant-of-concern](https://www.who.int/news/item/26-11-2021-classification-of-omicron-(b.1.1.529)-sars-cov-2-variant-of-concern) (accessed on 14 May 2022).
12. Gobeil, S.M.-C.; Henderson, R.; Stalls, V.; Janowska, K.; Huang, X.; May, A.; Speakman, M.; Beaudoin, E.; Manne, K.; Li, D.; et al. Structural diversity of the SARS-CoV-2 Omicron spike. *Mol. Cell* **2022**, *82*, 2050–2068.e6. [CrossRef]
13. Schmidt, F.; Muecksch, F.; Weisblum, Y.; Da Silva, J.; Bednarski, E.; Cho, A.; Wang, Z.; Gaebler, C.; Caskey, M.; Nussenzweig, M.C.; et al. Plasma Neutralization of the SARS-CoV-2 Omicron Variant. *N. Engl. J. Med.* **2022**, *386*, 599–601. [CrossRef]
14. Hoffmann, M.; Krüger, N.; Schulz, S.; Cossmann, A.; Rocha, C.; Kempf, A.; Nehlmeier, I.; Graichen, L.; Moldenhauer, A.-S.; Winkler, M.S.; et al. The Omicron variant is highly resistant against antibody-mediated neutralization: Implications for control of the COVID-19 pandemic. *Cell* **2022**, *185*, 447–456.e11. [CrossRef]
15. Francis, A.I.; Ghany, S.; Gilkes, T.; Umakanthan, S. Review of COVID-19 vaccine subtypes, efficacy and geographical distributions. *Postgrad. Med. J.* **2022**, *98*, 389–394. [CrossRef] [PubMed]
16. Gavriatopoulou, M.; Ntanasis-Stathopoulos, I.; Korompoki, E.; Fotiou, D.; Migkou, M.; Tzanninis, I.-G.; Psaltopoulou, T.; Kastiris, E.; Terpos, E.; Dimopoulos, M.A. Emerging treatment strategies for COVID-19 infection. *Clin. Exp. Med.* **2021**, *21*, 167–179. [CrossRef] [PubMed]
17. World Health Organization. Weekly Epidemiological Update on COVID-19—22 March 2022. Available online: <https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---22-march-2022> (accessed on 14 May 2022).
18. Feiner, L. WHO Says Omicron Cases Are “Off the Charts” as Global Infections Set New Records. Available online: <https://www.cnbc.com/2022/01/12/who-says-omicron-cases-are-off-the-charts-as-global-infections-set-new-records.html> (accessed on 14 May 2022).
19. SARS-CoV-2 Omicron Variant Cases Worldwide 2022. Available online: <https://www.statista.com/statistics/1279100/number-omicron-variant-worldwide-by-country/> (accessed on 14 May 2022).
20. Katz, M.; Nandi, N. Social Media and Medical Education in the Context of the COVID-19 Pandemic: Scoping Review. *JMIR Med. Educ.* **2021**, *7*, e25892. [CrossRef]
21. Sharma, M.; Yadav, K.; Yadav, N.; Ferdinand, K.C. Zika virus pandemic—analysis of Facebook as a social media health information platform. *Am. J. Infect. Control* **2017**, *45*, 301–302. [CrossRef]
22. Wiederhold, B.K. Social Media and Social Organizing: From Pandemic to Protests. *Cyberpsychol. Behav. Soc. Netw.* **2020**, *23*, 579–580. [CrossRef]
23. Fung, I.C.-H.; Duke, C.H.; Finch, K.C.; Snook, K.R.; Tseng, P.-L.; Hernandez, A.C.; Gambhir, M.; Fu, K.-W.; Tse, Z.T.H. Ebola virus disease and social media: A systematic review. *Am. J. Infect. Control* **2016**, *44*, 1660–1671. [CrossRef] [PubMed]
24. Ding, H. Social Media and Participatory Risk Communication during the H1N1 Flu Epidemic: A Comparative Study of the United States and China. *China Media Res.* **2010**, *6*, 80–91.
25. Longley, P.A.; Adnan, M.; Lansley, G. The Geotemporal Demographics of Twitter Usage. *Environ. Plan. A* **2015**, *47*, 465–484. [CrossRef]

26. Data Reportal. Twitter Statistics and Trends. Available online: <https://datareportal.com/essential-twitter-stats> (accessed on 14 May 2022).
27. Lazard, A.J.; Scheinfeld, E.; Bernhardt, J.M.; Wilcox, G.B.; Suran, M. Detecting themes of public concern: A text mining analysis of the Centers for Disease Control and Prevention's Ebola live Twitter chat. *Am. J. Infect. Control* **2015**, *43*, 1109–1111. [\[CrossRef\]](#)
28. Bolotova, Y.V.; Lou, J.; Safro, I. Detecting and Monitoring Foodborne Illness Outbreaks: Twitter Communications and the 2015 U.S. Salmonella Outbreak Linked to Imported Cucumbers. *arXiv* **2017**, arXiv:1708.07534.
29. Gomide, J.; Veloso, A.; Meira, W., Jr.; Almeida, V.; Benevenuto, F.; Ferraz, F.; Teixeira, M. Dengue Surveillance Based on a Computational Model of Spatio-Temporal Locality of Twitter. In Proceedings of the 3rd International Web Science Conference on—WebSci '11, Koblenz Germany, 15–17 June 2011; ACM Press: New York, NY, USA, 2011.
30. Tomaszewski, T.; Morales, A.; Lourentzou, I.; Caskey, R.; Liu, B.; Schwartz, A.; Chin, J. Identifying False Human Papillomavirus (HPV) Vaccine Information and Corresponding Risk Perceptions from Twitter: Advanced Predictive Models. *J. Med. Internet Res.* **2021**, *23*, e30451. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Do, H.J.; Lim, C.-G.; Kim, Y.J.; Choi, H.-J. Analyzing Emotions in Twitter during a Crisis: A Case Study of the 2015 Middle East Respiratory Syndrome Outbreak in Korea. In Proceedings of the 2016 International Conference on Big Data and Smart Computing (BigComp), Hong Kong, China, 18–20 June 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 415–418.
32. Radzikowski, J.; Stefanidis, A.; Jacobsen, K.H.; Croitoru, A.; Crooks, A.; Delamater, P.L. The Measles Vaccination Narrative in Twitter: A Quantitative Analysis. *JMIR Public Health Surveill.* **2016**, *2*, e5059. [\[CrossRef\]](#) [\[PubMed\]](#)
33. Fu, K.-W.; Liang, H.; Saroha, N.; Tse, Z.T.H.; Ip, P.; Fung, I.C.-H. How people react to Zika virus outbreaks on Twitter? A computational content analysis. *Am. J. Infect. Control* **2016**, *44*, 1700–1702. [\[CrossRef\]](#) [\[PubMed\]](#)
34. Signorini, A.; Segre, A.M.; Polgreen, P.M. The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. *PLoS ONE* **2011**, *6*, e19467. [\[CrossRef\]](#) [\[PubMed\]](#)
35. Gesualdo, F.; Stilo, G.; Agricola, E.; Gonfiantini, M.V.; Pandolfi, E.; Velardi, P.; Tozzi, A.E. Influenza-Like Illness Surveillance on Twitter through Automated Learning of Naïve Language. *PLoS ONE* **2013**, *8*, e82489. [\[CrossRef\]](#) [\[PubMed\]](#)
36. Szomszor, M.; Kostkova, P.; de Quincey, E. #swineflu: Twitter Predicts Swine Flu Outbreak in 2009. In *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 18–26, ISBN 9783642236341.
37. Alessa, A.; Faezipour, M. Flu Outbreak Prediction Using Twitter Posts Classification and Linear Regression with Historical Centers for Disease Control and Prevention Reports: Prediction Framework Study. *JMIR Public Health. Surveill.* **2019**, *5*, e12383. [\[CrossRef\]](#)
38. Hirschfeld, D. Twitter data accurately tracked Haiti cholera outbreak. *Nature* **2012**. [\[CrossRef\]](#)
39. Van Der Vyver, A.G. The Listeriosis Outbreak in South Africa: A Twitter Analysis of Public Reaction. Available online: <http://www.icmis.net/icmis18/ICMIS18CD/pdf/S198-final.pdf> (accessed on 14 May 2022).
40. Thackeray, R.; Burton, S.H.; Giraud-Carrier, C.; Rollins, S.; Draper, C.R. Using Twitter for breast cancer prevention: An analysis of breast cancer awareness month. *BMC Cancer* **2013**, *13*, 508. [\[CrossRef\]](#)
41. Da, B.L.; Surana, P.; Schueler, S.A.; Jalaly, N.Y.; Kamal, N.; Taneja, S.; Vittal, A.; Gilman, C.L.; Heller, T.; Koh, C. Twitter as a Noninvasive Bio-Marker for Trends in Liver Disease. *Hepatol. Commun.* **2019**, *3*, 1271–1280. [\[CrossRef\]](#)
42. Khan, A.; Silverman, A.; Rowe, A.; Rowe, S.; Tick, M.; Testa, S.; Dodds, K.; Alabbas, B.; Borum, M.L. Who Is Saying What about Inflammatory Bowel Disease on Twitter? In Proceedings of the G. W. Research Days 2016–2020, Washington, DC, USA, 2018.
43. McLean, R.; Shirazian, S. Women and Kidney Disease: A Twitter Conversation for One and All. *Kidney Int. Rep.* **2018**, *3*, 767–768. [\[CrossRef\]](#)
44. Stens, O.; Weisman, M.H.; Simard, J.; Reuter, K. Insights from Twitter Conversations on Lupus and Reproductive Health: Protocol for a Content Analysis. *JMIR Res. Protoc.* **2020**, *9*, e15623. [\[CrossRef\]](#) [\[PubMed\]](#)
45. Cevik, F.; Kilimci, Z.H. Analysis of Parkinson's Disease using Deep Learning and Word Embedding Models. *Acad. Perspect. Procedia* **2019**, *2*, 786–797. [\[CrossRef\]](#)
46. Porat, T.; Garaizar, P.; Ferrero, M.; Jones, H.; Ashworth, M.; Vadillo, M.A. Content and Source Analysis of Popular Tweets Following a Recent Case of Diphtheria in Spain. *Eur. J. Public Health* **2019**, *29*, 117–122. [\[CrossRef\]](#) [\[PubMed\]](#)
47. Sugumaran, R.; Voss, J. Real-Time Spatio-Temporal Analysis of West Nile Virus Using Twitter Data. In Proceedings of the 3rd International Conference on Computing for Geospatial Research and Applications—C.O.M.Geo '12, Reston, VA, USA, 1–3 July 2012; ACM Press: New York, NY, USA, 2012.
48. Kim, E.H.-J.; Jeong, Y.K.; Kim, Y.; Kang, K.Y.; Song, M. Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news. *J. Inf. Sci.* **2016**, *42*, 763–781. [\[CrossRef\]](#)
49. Tully, M.; Dalrymple, K.E.; Young, R. Contextualizing Nonprofits' Use of Links on Twitter During the West African Ebola Virus Epidemic. *Commun. Stud.* **2019**, *70*, 313–331. [\[CrossRef\]](#)
50. Odum, M.; Yoon, S. What can we learn about the Ebola outbreak from tweets? *Am. J. Infect. Control* **2015**, *43*, 563–571. [\[CrossRef\]](#)
51. Su, C.-J.; Yon, J.A.Q. Sentiment Analysis and Information Diffusion on Social Media: The Case of the Zika Virus. *Int. J. Inf. Educ. Technol.* **2018**, *8*, 685–692. [\[CrossRef\]](#)
52. Wood, M.J. Propagating and Debunking Conspiracy Theories on Twitter During the 2015–2016 Zika Virus Outbreak. *Cyberpsychol. Behav. Soc. Netw.* **2018**, *21*, 485–490. [\[CrossRef\]](#)

53. Ghenai, A.; Mejova, Y. Catching Zika Fever: Application of Crowdsourcing and Machine Learning for Tracking Health Misinformation on Twitter. *arXiv* **2017**, arXiv:1707.03778.
54. Yang, J.-A.J. Spatial-Temporal Analysis of Information Diffusion Patterns with User-Generated Geo-Social Contents from Social Media. Ph.D. Thesis, San Diego State University, San Diego, CA, USA, 2017.
55. Barata, G.; Shores, K.; Alperin, J.P. Local chatter or international buzz? Language differences on posts about Zika research on Twitter and Facebook. *PLoS ONE* **2018**, *13*, e0190482. [\[CrossRef\]](#)
56. Maci, S.M.; Sala, M. *Corpus Linguistics and Translation Tools for Digital Humanities: Research Methods and Applications*; Maci, S.M., Sala, M., Eds.; Bloomsbury Academic: London, UK, 2022; ISBN 9781350275232.
57. Alessa, A.; Faezipour, M. Tweet Classification Using Sentiment Analysis Features and TF-IDF Weighting for Improved Flu Trend Detection. In *Machine Learning and Data Mining in Pattern Recognition*; Springer International: Cham, Switzerland, 2018; pp. 174–186, ISBN 9783319961354.
58. Lamb, A.; Paul, M.J.; Dredze, M. Separating Fact from Fear: Tracking Flu Infections on Twitter. Available online: <https://aclanthology.org/N13-1097.pdf> (accessed on 6 July 2022).
59. Lee, K.; Mahmud, J.; Chen, J.; Zhou, M.; Nichols, J. Who Will Retweet This?: Automatically Identifying and Engaging Strangers on Twitter to Spread Information. In Proceedings of the 19th International Conference on Intelligent User Interfaces, New York, NY, USA, 24–27 February 2014; ACM: New York, NY, USA, 2014.
60. Dai, X.; Bikdash, M.; Meyer, B. From Social Media to Public Health Surveillance: Word Embedding Based Clustering Method for Twitter Classification. In Proceedings of the SoutheastCon 2017, Concord, NC, USA, 20 March–2 April 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–7.
61. Rahmanian, V.; Jahanbin, K. Using twitter and web news mining to predict COVID-19 outbreak. *Asian Pac. J. Trop. Med.* **2020**, *13*, 378. [\[CrossRef\]](#)
62. Rosenberg, H.; Syed, S.; Rezaie, S. The Twitter pandemic: The critical role of Twitter in the dissemination of medical information and misinformation during the COVID-19 pandemic. *Can. J. Emerg. Med.* **2020**, *22*, 418–421. [\[CrossRef\]](#) [\[PubMed\]](#)
63. Haman, M. The use of Twitter by state leaders and its impact on the public during the COVID-19 pandemic. *Heliyon* **2020**, *6*, e05540. [\[CrossRef\]](#) [\[PubMed\]](#)
64. Alhayan, F.; Pennington, D.; Ayouni, S. Twitter Use by the Dementia Community during COVID-19: A User Classification and Social Network Analysis. *Online Inf. Rev.* **2022**, ahead-of-print. [\[CrossRef\]](#)
65. Guo, J.; Radloff, C.L.; Wawrzynski, S.E.; Cloyes, K.G. Mining twitter to explore the emergence of COVID-19 symptoms. *Public Health Nurs.* **2020**, *37*, 934–940. [\[CrossRef\]](#)
66. Roy, S.; Ghosh, P. A Comparative Study on Distancing, Mask and Vaccine Adoption Rates from Global Twitter Trends. *Healthcare* **2021**, *9*, 488. [\[CrossRef\]](#)
67. Yousefinaghani, S.; Dara, R.; Mubareka, S.; Papadopoulos, A.; Sharif, S. An analysis of COVID-19 vaccine sentiments and opinions on Twitter. *Int. J. Infect. Dis.* **2021**, *108*, 256–262. [\[CrossRef\]](#)
68. Wang, S.; Schraagen, M.; Sang, E.T.K.; Dastani, M. Dutch General Public Reaction on Governmental COVID-19 Measures and Announcements in Twitter Data. *arXiv* **2020**, arXiv:2006.07283.
69. Krittanawong, C.; Narasimhan, B.; Virk, H.U.H.; Narasimhan, H.; Hahn, J.; Wang, Z.; Tang, W.W. Misinformation Dissemination in Twitter in the COVID-19 Era. *Am. J. Med.* **2020**, *133*, 1367–1369. [\[CrossRef\]](#)
70. Ahmed, W.; Vidal-Alaball, J.; Downing, J.; Seguí, F.L. COVID-19 and the 5G Conspiracy Theory: Social Network Analysis of Twitter Data. *J. Med. Internet Res.* **2020**, *22*, e19458. [\[CrossRef\]](#)
71. Bonnevie, E.; Gallegos-Jeffrey, A.; Goldbarg, J.; Byrd, B.; Smyser, J. Quantifying the rise of vaccine opposition on Twitter during the COVID-19 pandemic. *J. Commun. Healthc.* **2021**, *14*, 12–19. [\[CrossRef\]](#)
72. Chen, E.; Deb, A.; Ferrara, E. #Election2020: The first public Twitter dataset on the 2020 US Presidential election. *J. Comput. Soc. Sci.* **2021**, *5*, 1–18. [\[CrossRef\]](#) [\[PubMed\]](#)
73. Haq, E.-U.; Tyson, G.; Lee, L.-H.; Braud, T.; Hui, P. Twitter Dataset for 2022 Russo-Ukrainian Crisis. *arXiv* **2022**, arXiv:2203.02955.
74. Effrosynidis, D.; Karasakalidis, A.I.; Sylaios, G.; Arampatzis, A. The Climate Change Twitter Dataset. *Expert Syst. Appl.* **2022**, *204*, 117541. [\[CrossRef\]](#)
75. Meng, L.; Dong, Z.S. Natural Hazards Twitter Dataset. *arXiv* **2020**, arXiv:2004.14456.
76. Urchs, S.; Wendlinger, L.; Mitrovic, J.; Granitzer, M. MMoveT15: A Twitter Dataset for Extracting and Analysing Migration-Movement Data of the European Migration Crisis 2015. In Proceedings of the 2019 IEEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), Napoli, Italy, 12–14 June 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 146–149.
77. Dooms, S.; De Pessemier, T.; Martens, L. MovieTweatings: A Movie Rating Dataset Collected from Twitter. In Proceedings of the Workshop on Crowdsourcing and Human Computation for Recommender Systems (CrowdRec 2013), Held in Conjunction with the 7th ACM Conference on Recommender Systems (RecSys 2013), Hong Kong, China, 12 October 2013.
78. Wijesiriwardene, T.; Inan, H.; Kursuncu, U.; Gaur, M.; Shalin, V.L.; Thirunarayan, K.; Sheth, A.; Arpinar, I.B. ALONE: A Dataset for Toxic Behavior among Adolescents on Twitter. In *Lecture Notes in Computer Science*; Springer International: Cham, Switzerland, 2020; pp. 427–439, ISBN 9783030609740.

79. Zangerle, E.; Pichl, M.; Gassler, W.; Specht, G. #nowplaying Music Dataset: Extracting Listening Behavior from Twitter. In Proceedings of the First International Workshop on Internet-Scale Multimedia Management—WISMM '14, New York, NY, USA, 7 November 2014; ACM Press: New York, NY, USA, 2014.
80. Sech, J.; DeLucia, A.; Buczak, A.L.; Dredze, M. Civil Unrest on Twitter (CUT): A Dataset of Tweets to Support Research on Civil Unrest. In Proceedings of the Sixth Workshop on Noisy User-Generated Text (W-NUT 2020), November 2020; Association for Computational Linguistics, Stroudsburg, PA, USA; 2020; pp. 215–221.
81. Tekumalla, R.; Banda, J.M. A Large-Scale Twitter Dataset for Drug Safety Applications Mined from Publicly Existing Resources. *arXiv* **2020**, arXiv:2003.13900.
82. Stemmer, M.; Parmet, Y.; Ravid, G. What Are IBD Patients Talking about on Twitter? In *ICT for Health, Accessibility and Wellbeing*; Springer International: Cham, Switzerland, 2021; pp. 206–220, ISBN 9783030942083.
83. Alqurashi, S.; Alhindi, A.; Alanazi, E. Large Arabic Twitter Dataset on COVID-19. *arXiv* **2020**, arXiv:2004.04315.
84. Hayawi, K.; Shahriar, S.; Serhani, M.; Taleb, I.; Mathew, S. ANTi-Vax: A novel Twitter dataset for COVID-19 vaccine misinformation detection. *Public Health* **2022**, *203*, 23–30. [\[CrossRef\]](#)
85. Elhadad, M.K.; Li, K.F.; Gebali, F. COVID-19-FAKES: A Twitter (Arabic/English) Dataset for Detecting Misleading Information on COVID-19. In *Advances in Intelligent Networking and Collaborative Systems*; Springer International: Cham, Switzerland, 2021; pp. 256–268, ISBN 9783030577957.
86. Haouari, F.; Hasanain, M.; Suwaileh, R.; Elsayed, T. ArCOV19-Rumors: Arabic COVID-19 Twitter Dataset for Misinformation Detection. *arXiv* **2020**, arXiv:2010.08768.
87. Cheng, M.; Wang, S.; Yan, X.; Yang, T.; Wang, W.; Huang, Z.; Xiao, X.; Nazarian, S.; Bogdan, P. A COVID-19 Rumor Dataset. *Front. Psychol.* **2021**, *12*, 644801. [\[CrossRef\]](#)
88. Shamrat, F.M.J.M.; Chakraborty, S.; Imran, M.M.; Muna, J.N.; Billah, M.; Das, P.; Rahman, O. Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm. *Indones. J. Electr. Eng. Comput. Sci.* **2021**, *23*, 463–470. [\[CrossRef\]](#)
89. Sontayasara, T.; Jariyapongpaiboon, S.; Promjun, A.; Seelpipat, N.; Saengtabtim, K.; Tang, J.; Leelawat, N. Twitter Sentiment Analysis of Bangkok Tourism During COVID-19 Pandemic Using Support Vector Machine Algorithm. *J. Disaster Res.* **2021**, *16*, 24–30. [\[CrossRef\]](#)
90. Asgari-Chenaghlu, M.; Nikzad-Khasmakhi, N.; Minaee, S. Covid-Transformer: Detecting COVID-19 Trending Topics on Twitter Using Universal Sentence Encoder. *arXiv* **2020**, arXiv:2009.03947.
91. Amen, B.; Faiz, S.; Do, T.-T. Big data directed acyclic graph model for real-time COVID-19 twitter stream detection. *Pattern Recognit.* **2022**, *123*, 108404. [\[CrossRef\]](#) [\[PubMed\]](#)
92. Lyu, J.C.; Luli, G.K. Understanding the Public Discussion About the Centers for Disease Control and Prevention During the COVID-19 Pandemic Using Twitter Data: Text Mining Analysis Study. *J. Med. Internet Res.* **2021**, *23*, e25108. [\[CrossRef\]](#)
93. Al-Ramahi, M.; Elnoshokaty, A.; El-Gayar, O.; Nasrallah, T.; Wahbeh, A. Public Discourse Against Masks in the COVID-19 Era: Infodemiology Study of Twitter Data. *JMIR Public Health Surveill.* **2021**, *7*, e26780. [\[CrossRef\]](#)
94. Jain, S.; Sinha, A. Identification of Influential Users on Twitter: A Novel Weighted Correlated Influence Measure for COVID-19. *Chaos Solitons Fractals* **2020**, *139*, 110037. [\[CrossRef\]](#)
95. Madani, Y.; Erritali, M.; Bouikhalene, B. Using artificial intelligence techniques for detecting Covid-19 epidemic fake news in Moroccan tweets. *Results Phys.* **2021**, *25*, 104266. [\[CrossRef\]](#)
96. Shokoohyar, S.; Berenji, H.R.; Dang, J. Exploring the heated debate over reopening for economy or continuing lockdown for public health safety concerns about COVID-19 in Twitter. *Int. J. Bus. Syst. Res.* **2021**, *15*, 650. [\[CrossRef\]](#)
97. Chehal, D.; Gupta, P.; Gulati, P. COVID-19 pandemic lockdown: An emotional health perspective of Indians on Twitter. *Int. J. Soc. Psychiatry* **2021**, *67*, 64–72. [\[CrossRef\]](#)
98. Glowacki, E.M.; Wilcox, G.B.; Glowacki, J.B. Identifying #addiction concerns on twitter during the COVID-19 pandemic: A text mining analysis. *Subst. Abus.* **2021**, *42*, 39–46. [\[CrossRef\]](#) [\[PubMed\]](#)
99. Selman, L.E.; Chamberlain, C.; Sowden, R.; Chao, D.; Selman, D.; Taubert, M.; Braude, P. Sadness, despair and anger when a patient dies alone from COVID-19: A thematic content analysis of Twitter data from bereaved family members and friends. *Palliat. Med.* **2021**, *35*, 1267–1276. [\[CrossRef\]](#) [\[PubMed\]](#)
100. Koh, J.X.; Liew, T.M. How loneliness is talked about in social media during COVID-19 pandemic: Text mining of 4,492 Twitter feeds. *J. Psychiatr. Res.* **2022**, *145*, 317–324. [\[CrossRef\]](#) [\[PubMed\]](#)
101. Mackey, T.; Purushothaman, V.; Li, J.; Shah, N.; Nali, M.; Bardier, C.; Liang, B.; Cai, M.; Cuomo, R. Machine Learning to Detect Self-Reporting of Symptoms, Testing Access, and Recovery Associated with COVID-19 on Twitter: Retrospective Big Data Intelligence Study. *JMIR Public Health Surveill.* **2020**, *6*, e19509. [\[CrossRef\]](#)
102. Leung, J.; Chung, J.; Tisdale, C.; Chiu, V.; Lim, C.; Chan, G. Anxiety and Panic Buying Behaviour during COVID-19 Pandemic-A Qualitative Analysis of Toilet Paper Hoarding Contents on Twitter. *Int. J. Environ. Res. Public Health* **2021**, *18*, 1127. [\[CrossRef\]](#)
103. Privacy Policy of Twitter. Available online: https://twitter.com/en/privacy/previous/version_15 (accessed on 15 May 2022).
104. Twitter Developer Agreement and Policy. Available online: <https://developer.twitter.com/en/developer-terms/agreement-and-policy> (accessed on 15 May 2022).
105. Social Bearing Research Tool. Available online: <https://socialbearing.com/> (accessed on 15 May 2022).
106. JQuery. JQuery UI 1.11.2. Available online: <https://blog.jqueryui.com/2014/10/jquery-ui-1-11-2/> (accessed on 6 July 2022).

107. Jquery-Tinysort-Min.Js. Available online: <https://searchcode.com/codesearch/view/33978492/> (accessed on 6 July 2022).
108. De Sandro, D. Masonry: Cascading Grid Layout Plugin. Available online: <https://github.com/desandro/masonry> (accessed on 6 July 2022).
109. Npm. D3.Layout.Cloud. Available online: <https://www.npmjs.com/package/d3.layout.cloud> (accessed on 6 July 2022).
110. Bostock, M. Data-Driven Documents. Available online: <https://d3js.org/> (accessed on 6 July 2022).
111. Analytics.Js 2.0 Source. Available online: <https://segment.com/docs/connections/sources/catalog/libraries/website/javascript/> (accessed on 6 July 2022).
112. Loader.Js. Available online: <https://github.com/ember-cli/loader.js> (accessed on 6 July 2022).
113. Tunca, S.; Sezen, B.; Balcioglu, Y.S. Twitter Analysis for Metaverse Literacy. In Proceedings of the New York Academic Research Congress, New York, NY, USA, 16 January 2022.
114. Shaw, A. *Preparing Your Social Media Data for a MANCOVA Test Using Social Bearing*; SAGE: New York, NY, USA, 2022. [CrossRef]
115. Shaw, A. Promoting Social Change—Assessing How Twitter Was Used to Reduce Drunk Driving Behaviours Over New Year’s Eve. *J. Promot. Manag.* **2021**, *27*, 441–463. [CrossRef]
116. Maci, S. Discourse Strategies of Fake News in the Anti-Vax Campaign. *Lang. Cult. Mediat. (LCM J.)* **2019**, *6*, 15–43. [CrossRef]
117. Neyazi, T.A. Digital propaganda, political bots and polarized politics in India. *Asian J. Commun.* **2020**, *30*, 39–57. [CrossRef]
118. Saha, A.; Agarwal, N. *Assessing Social Support and Stress in Autism-Focused Virtual Communities: Emerging Research and Opportunities*; Information Science Reference: Hershey, PA, USA, 2018; ISBN 9781522540212.
119. Záhová, K. Propaganda on Social Media: The Case of Geert Wilders. Available online: <https://dspace.cuni.cz/handle/20.500.11956/99767> (accessed on 9 June 2022).
120. Almurayh, A.; Alahmadi, A. The Proliferation of Twitter Accounts in a Higher Education Institution, *IAENG International Journal of Computer Science*, 49:1. Available online: http://www.iaeng.org/IJCS/issues_v49/issue_1/IJCS_49_1_19.pdf (accessed on 9 June 2022).
121. Forgues, B.; May, T. Message in a Bottle: Multiple Modes and Multiple Media in Market Identity Claims. In *Research in the Sociology of Organizations*; Emerald: West Yorkshire, UK, 2017; pp. 179–202.
122. Chiauuzzi, E.; Wicks, P. Digital Trespass: Ethical and Terms-of-Use Violations by Researchers Accessing Data from an Online Patient Community. *J. Med. Internet Res.* **2019**, *21*, e11985. [CrossRef]
123. Search Tweets: Standard v1.1. Available online: <https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/overview> (accessed on 21 July 2022).
124. How to Tweet. Available online: <https://help.twitter.com/en/using-twitter/how-to-tweet> (accessed on 15 May 2022).
125. Twitter Official Website. Available online: <https://twitter.com/> (accessed on 15 May 2022).
126. Twitter Android Application. Available online: https://play.google.com/store/apps/details?id=com.twitter.android&hl=en_US&gl=US (accessed on 15 May 2022).
127. Twitter for iPhone. Available online: <https://apps.apple.com/in/app/twitter/id333903271> (accessed on 15 May 2022).
128. Wikipedia Contributors TweetDeck. Available online: <https://en.wikipedia.org/w/index.php?title=TweetDeck&oldid=1056092943> (accessed on 15 May 2022).
129. Supported Languages on Twitter. Available online: <https://developer.twitter.com/en/docs/twitter-for-websites/supported-languages> (accessed on 15 May 2022).
130. Standard Search API. Available online: <https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/api-reference/get-search-tweets> (accessed on 6 July 2022).
131. How to Use Advanced Search. Available online: <https://help.twitter.com/en/using-twitter/twitter-advanced-search> (accessed on 6 July 2022).
132. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018. [CrossRef]
133. Hydrator: Turn Tweet IDs into Twitter JSON & CSV from Your Desktop! Available online: <https://github.com/DocNow/hydrator> (accessed on 6 July 2022).
134. Tekumalla, R.; Banda, J.M. Social Media Mining Toolkit (SMMT). *Genom. Inform.* **2020**, *18*, e16. [CrossRef]
135. Twarc: A Command Line Tool (and Python Library) for Archiving Twitter JSON. Available online: <https://github.com/DocNow/twarc> (accessed on 6 July 2022).
136. Vijay Karunamurthy Make Way for Youtu.Be Links. Available online: <https://blog.youtube/news-and-events/make-way-for-youtube-links/> (accessed on 15 May 2022).
137. Pichl, M.; Zangerle, E.; Specht, G. Towards a Context-Aware Music Recommendation Approach: What Is Hidden in the Playlist Name? In Proceedings of the 2015 IEEE International Conference on Data Mining Workshop (ICDMW), Atlantic City, NJ, USA, 14–17 November 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 1360–1365.
138. Jannach, D.; Kamehkhosh, I.; Lerche, L. Leveraging Multi-Dimensional User Models for Personalized next-Track Music Recommendation. In Proceedings of the Symposium on Applied Computing—SAC ’17, New York, NY, USA, 3–7 April 2017; ACM Press: New York, NY, USA, 2017.
139. Ludewig, M.; Jannach, D. Evaluation of session-based recommendation algorithms. *User Model. User-Adapted Interact.* **2018**, *28*, 331–390. [CrossRef]

140. Zangerle, E.; Chen, C.-M.; Tsai, M.-F.; Yang, Y.-H. Leveraging Affective Hashtags for Ranking Music Recommendations. *IEEE Trans. Affect. Comput.* **2021**, *12*, 78–91. [\[CrossRef\]](#)
141. Zangerle, E.; Pichl, M.; Hupfauf, B.; Specht, G. Can Microblogs Predict Music Charts? An Analysis of the Relationship between #Nowplaying Tweets and Music Charts. Available online: http://m.mr-pc.org/ismir16/website/articles/039_Paper.pdf (accessed on 1 July 2022).
142. Pichl, M.; Zangerle, E.; Specht, G. Understanding user-curated playlists on Spotify: A machine learning approach. *Int. J. Multimed. Data Eng. Manag.* **2017**, *8*, 44–59. [\[CrossRef\]](#)
143. Abboud, R.; Tekli, J. Integration of nonparametric fuzzy classification with an evolutionary-developmental framework to perform music sentiment-based analysis and composition. *Soft Comput.* **2020**, *24*, 9875–9925. [\[CrossRef\]](#)
144. Kaneshiro, B.; Ruan, F.; Baker, C.W.; Berger, J. Characterizing Listener Engagement with Popular Songs Using Large-Scale Music Discovery Data. *Front. Psychol.* **2017**, *8*, 416. [\[CrossRef\]](#) [\[PubMed\]](#)
145. Zangerle, E.; Pichl, M.; Schedl, M. User Models for Culture-Aware Music Recommendation: Fusing Acoustic and Cultural Cues. *Trans. Int. Soc. Music Inf. Retr.* **2020**, *3*, 1–16. [\[CrossRef\]](#)
146. Hridi, A.P. *Mining User Personality from Music Listening Behavior in Online Platforms Using Audio Attributes*; Clemson University: Clemson, SC, USA, 2021.
147. Moderna Begins Next Phase of Omicron-Specific Booster Trial as Study Finds That Antibodies Remain Durable despite 6-Fold Drop over 6 Months. Available online: <https://www.cnn.com/2022/01/26/health/moderna-omicron-antibodies-booster/index.html> (accessed on 16 May 2022).
148. Pfizer and BioNTech Initiate Study to Evaluate Omicron-Based COVID-19 Vaccine in Adults 18 to 55 Years of Age. Available online: <https://www.pfizer.com/news/press-release/press-release-detail/pfizer-and-biontech-initiate-study-evaluate-omicron-based> (accessed on 16 May 2022).
149. Strasser, Z.; Hadavand, A.; Murphy, S.; Estiri, H. SARS-CoV-2 Omicron Variant Is as Deadly as Previous Waves after Adjusting for Vaccinations, Demographics, and Comorbidities. *Res. Sq.* **2022**. [\[CrossRef\]](#)
150. Bar-On, Y.M.; Goldberg, Y.; Mandel, M.; Bodenheimer, O.; Amir, O.; Freedman, L.; Alroy-Preis, S.; Ash, N.; Huppert, A.; Milo, R. Protection by a Fourth Dose of BNT162b2 against Omicron in Israel. *N. Engl. J. Med.* **2022**, *386*, 1712–1720. [\[CrossRef\]](#)
151. Shukla, R.; Sinha, A.; Chaudhary, A. *TweezBot*: An AI-Driven Online Media Bot Identification Algorithm for Twitter Social Networks. *Electronics* **2022**, *11*, 743. [\[CrossRef\]](#)
152. Martin-Gutierrez, D.; Hernandez-Penaloza, G.; Hernandez, A.B.; Lozano-Diez, A.; Alvarez, F. A Deep Learning Approach for Robust Detection of Bots in Twitter Using Transformers. *IEEE Access* **2021**, *9*, 54591–54601. [\[CrossRef\]](#)
153. Pham, P.; Nguyen, L.T.; Vo, B.; Yun, U. Bot2Vec: A general approach of intra-community oriented representation learning for bot detection in different types of social networks. *Inf. Syst.* **2022**, *103*, 101771. [\[CrossRef\]](#)
154. Pastor-Galindo, J.; Mármol, F.G.; Pérez, G.M. BOTTER: A Framework to Analyze Social Bots in Twitter. *arXiv* **2021**, arXiv:2106.15543.
155. Praveena, A.; Smys, S. Effective Spam Bot Detection Using Glow Worm-Based Generalized Regression Neural Network. In *Mobile Computing and Sustainable Informatics*; Springer: Singapore, 2022; pp. 469–487, ISBN 9789811618659.