infectious disease reports

MDPI

Communication

MonkeyPox2022Tweets: A Large-Scale Twitter Dataset on the 2022 Monkeypox Outbreak, Findings from Analysis of Tweets, and Open Research Questions

Nirmalya Thakur

Department of Computer Science, Emory University, Atlanta, GA 30322, USA; nirmalya.thakur@emory.edu

Abstract: The mining of Tweets to develop datasets on recent issues, global challenges, pandemics, virus outbreaks, emerging technologies, and trending matters has been of significant interest to the scientific community in the recent past, as such datasets serve as a rich data resource for the investigation of different research questions. Furthermore, the virus outbreaks of the past, such as COVID-19, Ebola, Zika virus, and flu, just to name a few, were associated with various works related to the analysis of the multimodal components of Tweets to infer the different characteristics of conversations on Twitter related to these respective outbreaks. The ongoing outbreak of the monkeypox virus, declared a Global Public Health Emergency (GPHE) by the World Health Organization (WHO), has resulted in a surge of conversations about this outbreak on Twitter, which is resulting in the generation of tremendous amounts of Big Data. There has been no prior work in this field thus far that has focused on mining such conversations to develop a Twitter dataset. Furthermore, no prior work has focused on performing a comprehensive analysis of Tweets about this ongoing outbreak. To address these challenges, this work makes three scientific contributions to this field. First, it presents an open-access dataset of 556,427 Tweets about monkeypox that have been posted on Twitter since the first detected case of this outbreak. A comparative study is also presented that compares this dataset with 36 prior works in this field that focused on the development of Twitter datasets to further uphold the novelty, relevance, and usefulness of this dataset. Second, the paper reports the results of a comprehensive analysis of the Tweets of this dataset. This analysis presents several novel findings; for instance, out of all the 34 languages supported by Twitter, English has been the most used language to post Tweets about monkeypox, about 40,000 Tweets related to monkeypox were posted on the day WHO declared monkeypox as a GPHE, a total of 5470 distinct hashtags have been used on Twitter about this outbreak out of which #monkeypox is the most used hashtag, and Twitter for iPhone has been the leading source of Tweets about the outbreak. The sentiment analysis of the Tweets was also performed, and the results show that despite a lot of discussions, debate, opinions, information, and misinformation, on Twitter on various topics in this regard, such as monkeypox and the LGBTQI+ community, monkeypox and COVID-19, vaccines for monkeypox, etc., "neutral" sentiment was present in most of the Tweets. It was followed by "negative" and "positive" sentiments, respectively. Finally, to support research and development in this field, the paper presents a list of 50 open research questions related to the outbreak in the areas of Big Data, Data Mining, Natural Language Processing, and Machine Learning that may be investigated based on this dataset.

Keywords: monkeypox; twitter; dataset; tweets; social media; big data; data mining; data analysis; natural language processing; machine learning

Citation: Thakur, N. MonkeyPox2022Tweets: A Large-Scale Twitter Dataset on the 2022 Monkeypox Outbreak, Findings from Analysis of Tweets, and Open Research Questions. *Infect. Dis. Rep.* **2022**, *14*, 855–883. https://doi.org/10.3390/idr14060087

Academic Editor: Nicola Petrosillo

Received: 23 August 2022
Accepted: 8 November 2022
Published: 14 November 2022

1. Introduction

Monkeypox, caused by the monkeypox virus, which belongs to the Poxviridae family, Chordopoxvirinae subfamily, and Orthopoxvirus genus [1], is a re-emerging zoonotic disease. The monkeypox virus was initially discovered in monkeys in 1958 [2], and the

first case of human monkeypox was detected in the Democratic Republic of the Congo (DRC) in a nine-month-old boy in 1970 [3]. The monkeypox virus is closely related to the variola virus (smallpox virus) and results in a smallpox-like disease. The incubation period of monkeypox is 5–21 days, and common symptoms include fever (between 38.5 °C and 40.5 °C), headache, and myalgia. A distinguishing feature of the monkeypox infection is the presence of swelling at the maxillary, cervical or inguinal lymph nodes (lymphadenopathy) [4,5]. A recent study found that during the ongoing outbreak of monkeypox, inguinal lymphadenopathy was more common than cervical and axillary lymphadenopathy [6]. In individuals infected with the monkeypox virus, rashes appear following the onset of fever, beginning on the face, tongue, and oral cavity before spreading across the body. In the later stages of the infection, lesions in the oral cavity may make it challenging for the patients to eat and drink [5]. However, during the ongoing outbreak, multiple atypical clinical observations have been reported as compared to the prior outbreaks [7,8]. The severity of the infection is usually determined by the lesion count, as there is a direct correlation between high lesion counts and severe health-related complications [5]. Studies have shown that patients with severe complications may experience respiratory and gastrointestinal issues [9], septicemia [9,10], encephalitis [5], and ocular infections [11].

The monkeypox virus had been endemic in the DRC and a few African countries for a very long time, and a few cases outside these geographic regions were recorded only twice—first in 2003 [12] and then in 2018–2019 [13,14]. However, at the time of writing this paper, the world is experiencing a global outbreak of the monkeypox virus with 71,096 cases, of which 70,377 cases have been reported in locations that have not historically reported any monkeypox infections [15]. Some of the countries that have recorded the greatest number of monkeypox cases so far include the United States (26,577 cases), Brazil (8207 cases), Spain (7209 cases), France (4043 cases), the United Kingdom (3654 cases), Germany (3645 cases), Peru (2587 cases), Colombia (2453 cases), Mexico (1968 cases), Canada (1411 cases), and the Netherlands (1221 cases).

The first case of this 2022 global monkeypox outbreak was confirmed in the United Kingdom on 7 May 2022 [16]. On 19 May 2022, the first draft genome sequence of the monkeypox virus was performed by scientists in Portugal [17]. The genomic data related to this outbreak that has been studied so far indicate that this outbreak is caused by the West African clade [18]. On 20 May 2022, the World Health Organization (WHO) called an "emergency meeting" [19] to discuss the global concerns centered around the rising cases of the monkeypox virus. Since then, WHO was considering whether the outbreak should be assessed as a "potential public health emergency of international concern" or PHEIC, as was done for the COVID-19 and Ebola outbreaks in the past [20]. On 6 June 2022, the Center for Disease Control (CDC) in the United States raised its monkeypox alert to "Level 2" following the rapid increase in cases [21]. On 23 July 2022, following another meeting, the WHO declared monkeypox a Global Public Health Emergency (GPHE) [22]. There have been several reports and findings related to the spread of Monkeypox. In a recent report, the CDC said, "monkeypox eradication unlikely in the U.S. as virus could spread indefinitely" [23]. In a report by the *New Scientist*, it was discussed that a dangerous monkeypox variant circulating in the DRC could go global [24]. According to a recent article published in *Nature* [25], monkeypox could become impossible to contain if wild animal spread continues.

As per the CDC, "currently there is no treatment approved specifically for monkeypox virus infections" [26]. However, recently, a vaccine for monkeypox has been approved by the Food and Drug Association (FDA). The vaccine, previously used for smallpox, is called JYNNEOS and was developed by Bavarian Nordic, a Danish biotechnology firm [27]. The JYNNEOS vaccine has been the primary vaccine being used in the United States during this outbreak [28]. The ACAM2000 vaccine is an alternative to JYNNEOS. It is also approved to help protect against smallpox and monkeypox [29]. In addition to vaccines, in the United States, as per the CDC, several antivirals, such as Tecovirimat (also known as TPOXX, ST-246), Vaccinia Immune Globulin Intravenous (VIGIV), Cidofovir (also known as Vistide),

and Brincidofovir (also known as CMX001 or Tembexa), are currently available from the Strategic National Stockpile (SNS) as options for the treatment of monkeypox [26].

As the cases surge, countries all over the world are taking various forms of preparations, initiatives, and measures to reduce the spread of the virus. These include a lockdown in Belgium [30], the United States ordering 500,000 doses of the JYNNEOS vaccine [31], Canada offering vaccination to high-risk groups [32], health authorities in France and Denmark suggesting a vaccine rollout to adults infected by the virus [33], Germany recommending vaccinations for high-risk groups [34], and the United Kingdom advising self-isolation for everyone infected with the virus [35], just to name a few.

The rising cases of monkeypox and the associated recommendations, initiatives, and measures by various countries have led to the public engaging in conversations for information seeking and sharing related to monkeypox. The Internet of Everything lifestyle of today's living is centered around people engaging in online conversations via the internet, specifically social media platforms, and spending a lot more time on the internet than ever before [36]. As a result, there has been a tremendous increase in the use of social media platforms in the recent past [37,38]. Conversations on social media include a wide range of topics, such as recent issues, global challenges, pandemics, emerging technologies, news, current events, politics, family, relationships, trending topics, and career opportunities [39]. Twitter, one such social media platform, is used by people of almost all age groups from different parts of the world [40,41]. At present, there are about 450 million monthly active users on Twitter [42]. In view of the surge in Tweets about monkeypox since the beginning of the outbreak, Twitter recently added a link for accurate information on monkeypox [43]. A recent press release reported—"medical experts are building brands as monkeypox influencers and thought leaders, using their credentials and controversial posts to gain Twitter clout as mounting anxiety over the virus continues to spread" [44]. In addition to this, several other Tweets about monkeypox have also been discussed and debated in press releases in the last few days [45–47].

Mining social media conversations, for instance, Tweets, to develop datasets has been of significant interest to the scientific community in the last few years, as can be seen from several recent works in this field (Section 2.1). Such Twitter datasets serve as a data resource for a wide range of applications and use-case scenarios related to studying the associated conversation paradigms as well as for investigating the patterns of the underlying information-seeking and sharing behavior on Twitter. Some of the recent virus outbreaks, such as COVID-19, Ebola, Zika virus, and flu, were followed by the scientific community developing Twitter datasets, performing a comprehensive analysis of the multimodal components of the Tweets (such as hashtags, language, retweets, studying the source of the Tweet, etc.), and analyzing the sentiments of these Tweets. The recent outbreak of monkeypox has also led to an increase in research and development in this field in the last few weeks (Section 2.2). However, none of these prior works focused on mining Tweets about the 2022 monkeypox outbreak to develop a dataset. Neither did any of these prior works focus on performing a comprehensive analysis of the Tweets about this outbreak. Furthermore, there has been no work conducted in this field thus far that has focused on outlining open research questions or research directions to advance knowledge, innovation, and discovery in this field. This paper aims to address these challenges. In summary, it makes the following scientific contributions to this field:

1. It presents an open-access dataset of 556,427 Tweet IDs of the same number of Tweets about monkeypox that were posted on Twitter from 7 May 2022 to 9 October 2022. The dataset is available at https://doi.org/10.7910/DVN/CR7T5E. The earliest date was selected as 7 May 2022, as the first case of the 2022 monkeypox outbreak was recorded on this date. 9 October 2022 was the most recent date at the time of resubmission of this paper after the second review round. The dataset 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. A comparative study is also

presented that compares this dataset with 36 prior works in this field that focused on the development of Twitter datasets to further uphold the novelty, relevance, and usefulness of this dataset.

2. It presents the findings from a comprehensive content analysis of these Tweets. The findings show that:
    a. All the 34 languages supported by the Twitter API have been used to post Tweets about the outbreak. However, English has been the most used language.
    b. The day WHO declared monkeypox as a GPHE, about 40,000 Tweets related to monkeypox were posted in a span of just 24 h.
    c. A total of 5470 distinct hashtags have been used in Tweets about this outbreak, of which #monkeypox is the most used hashtag as compared to all other variations of the spelling in terms of use of uppercase or lowercase characters, such as #MonkeyPox, #monkeyPox, #MONKEYPOX, etc.
    d. Twitter for iPhone has been the leading source that has been used to post Tweets about monkeypox since the first case of this outbreak. It is followed by Twitter for Android, the Twitter Web App, and other sources.
3. The paper also presents the findings of sentiment analysis of the Tweets of this dataset. The findings of this study show that despite a lot of discussions, debate, opinions, information, and misinformation on Twitter on various topics in this regard, such as monkeypox and the LGBTQI+ community, monkeypox and COVID-19, vaccines for monkeypox, etc., a "neutral" sentiment is present in most of the Tweets. It is followed by "negative" and "positive" sentiments, respectively.
4. Finally, to support research and development in this field, a list of 50 open research questions in the areas of Big Data, Data Mining, Machine Learning, Natural Language Processing, and Information Retrieval with a specific focus on this outbreak is presented that may be studied, analyzed, and investigated using this dataset.

The rest of the paper is organized as follows. Section 2 presents the literature review. The methodologies that were followed for the development of this dataset, content analysis of the Tweets, and the sentiment analysis of Tweets are presented in Section 3. Section 3 also outlines how the dataset 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) for scientific data management. Section 4 presents the results of this work. In Section 4.1, a detailed description of the dataset files is presented. It also presents step-by-step instructions on how to use this dataset. A comprehensive comparative study with prior works in this field that focused on the development of Twitter datasets is also presented in Section 4.1 to uphold the novelty, relevance, and usefulness of this dataset. Section 4.2 presents the results of the content analysis of the Tweets. It is followed by Section 4.3, where the results of the sentiment analysis of the Tweets are presented and discussed. A list of 50 open research questions that may be investigated using this dataset is presented in Section 4.4. It is followed by the conclusion and scope for future work in Section 5, which is followed by references.

2. Literature Review

This section presents an overview of recent works in this field. It is divided into three parts. Section 2.1 outlines the recent works that focused on the development of Twitter datasets on global challenges, arising matters, public needs, virus outbreaks, pandemics, trending topics, and related areas in the last few years. Section 2.2 presents an overview of the recent works related to the 2022 monkeypox outbreak. Section 2.3 discusses prior works in this field that focused on the analysis of the multimodal components of Tweets in the context of virus outbreaks, pandemics, and epidemics.

2.1. Works on the Development of Twitter Datasets and Use-Cases

The mining of social media conversations, for instance, Tweets, to develop datasets has been of significant interest to the scientific community in the fields of Big Data, Data Mining, and Natural Language Processing in the last few years, as such datasets serve as a data resource for a wide range of applications related to studying the associated conversation paradigms as well as for investigating the patterns of the underlying online information-seeking and sharing behavior on Twitter. In this section, a review of such works is presented. The use cases supported by a couple of recent Twitter datasets are also outlined as examples to discuss the applicability and usefulness of Twitter datasets for the investigation of different research questions.

Some of the recent works in this field include Twitter datasets on hate speech and abusive language [48], the European migration crisis [49], natural hazards [50], misogynistic language [51], offensive language [52], civil unrest [53], exoskeletons [54], the efficacy of hydroxychloroquine as a treatment for COVID-19 [55], pregnancy outcomes [56], drug-related knowledge [57], the public opinion of people in Indonesia on different matters [58], a severe storm and F1 tornado that struck Central Pennsylvania [59], online learning during the COVID-19 Omicron wave [60], multi-ideology or white supremacy [61], Sundanese (the second-largest tribe in Indonesia) [62], vaccines [63], BlackLivesMatter movement [64], the Omicron variant of COVID-19 [65], hazardous events at the Baths of Diocletian site in Rome [66], memes from Black Twitter [67], and the Arabic language [68]. In addition to this, the outbreak of COVID-19 was associated with the development of multiple Twitter datasets, such as Twitter datasets on conversations about COVID-19 in Spanish [69], Bengali [70], and English [71]. Furthermore, Twitter datasets have also been developed based on trending hashtags and phrases such as #IndonesiaHumanRightsSOS [72], #Blackwomanhood [73], #MarchForBlackWomen [74], #BlackTheory [75], #DuragFest [76], #BringBackOurInternet [77], #WOCAffirmation [78], #AskTimothy [79], #WITBragDay [80], #preuambicio [81], #MiPrimerRecuerdoFeminista [82], and "I Voted For Trump" [83], just to name a few.

All these datasets have been used for multiple use-case scenarios. For instance, the Twitter dataset on drug-related knowledge [57] was used for detecting medication mentions on Twitter [84], region-specific monitoring and characterization of opioid-related social media chatter [85], tracking birth defect-related conversations on Twitter [86], detection of the self-reports of prescription medication abuse from Twitter [87], development of a methodology for automatic detection of breast cancer cohort from Tweets [88], development of a methodology to identify mentions of specific drugs on Twitter [89], and identifying conversations on Twitter related to the adverse drug reactions (ADRs) of marketed drugs [90]. Similarly, the Twitter dataset on conversations on Twitter about the efficacy of Hydroxychloroquine as a treatment for COVID-19 was used for stance detection in Tweets related to COVID-19 [91], misinformation detection on Twitter [92], detection of fake news related to COVID-19 [93], studying the public perceptions of approved versus off-label use for COVID-19-related medications [94], understanding public opinion on using hydroxychloroquine for COVID-19 treatments [95], stance detection towards vaccination for COVID-19 [96], and a few other applications.

2.2. Works related to the 2022 Monkeypox Outbreak

This section outlines the recent works related to the ongoing monkeypox outbreak. The work of Miura et al. [97] involved estimating the incubation period of the 2022 monkeypox outbreak. The authors focused on the reported cases in the Netherlands and found that the incubation period was 21 days. Bragazzi et al. [98] studied the confirmed cases in 13 countries to discuss how stigmatization of the LGBTQI+ community should be avoided as the virus infects more people. The work by Dashraath et al. [99] presented a set of guidelines to be followed by pregnant individuals with monkeypox exposure. Kampf [100] studied the efficacy of biocidal agents and disinfectants against the monkeypox virus and other orthopoxviruses. The work involved an extensive review of the literature to

summarize and discuss the findings from prior works that presented results about the inactivation of any orthopoxvirus by different kinds of disinfectants.

Nörz et al. [101] examined the surfaces of two hospital rooms of monkeypox patients in Germany to discuss the different ways by which this virus spreads with a specific focus on surface contamination. The study by Abbas et al. [102] presented a list of response strategies to control the spread of the monkeypox virus. The authors also discussed specific guidelines for risk communication and community engagement. According to the findings presented by Mungmunpuntipantip et al. [103], diarrhea was a symptom reported in 5.9% of patients who were infected with the monkeypox virus. Sallam et al. [104] performed a comprehensive study to investigate the level of conspiracy theories about monkeypox in students in Jordanian Health schools. The study involved 615 students. The findings showed that only 26.2% of the students knew about vaccination for monkeypox. The study also showed that increased age and non-medical backgrounds were among the user diversity characteristics that were associated with harboring conspiracy theories about the virus. Ahsan et al. [105] presented a collection of 1905 images about the monkeypox virus, which were collected from different sources, such as websites, newspapers, and online portals. The work by Malik et al. [106] aimed to study the attitudes of the general population of the United States toward monkeypox. The findings showed that 47% of the respondents felt that their knowledge about the monkeypox virus was poor or very poor. Furthermore, the study also showed that people vaccinated against COVID-19 were more likely to receive the monkeypox vaccine if the same were recommended. In the work by Sypsa et al. [107], the focus was to study the transmission potential of monkeypox in mass gatherings. The authors estimated that, on average, more than one secondary case of monkeypox could be expected per infectious person in a mass gathering if they have a high number (more than 30) of group contacts or more than eight close contacts.

Based on the works reviewed in Section 2.1, it can be concluded that Twitter datasets have been developed on a wide range of topics in the past, such as global challenges, arising matters, public needs, and virus outbreaks. Such datasets have helped in the investigation of a wide range of research questions relevant to advancing timely knowledge, innovation, and discovery in the respective domains. From the review of recent works related to this outbreak in Section 2.2, it can be concluded that none of the prior works in this field have focused on the development of such a dataset. This upholds the need to develop a Twitter dataset on the ongoing 2022 outbreak of monkeypox. To address this need, this work presents an open-access dataset of 556,427 Tweets about the monkeypox outbreak. The methodology and results related to the development of this dataset are discussed in Sections 3.1 and 4.1, respectively.

2.3. Works on the Analysis of Tweets Related to Virus Outbreaks, Pandemics, and Epidemics

Performing a comprehensive analysis of Tweets related to virus outbreaks, pandemics, and epidemics has been of significant interest to researchers in this field in the recent past. In [65], the authors presented a study on Tweets posted about the Omicron variant of COVID-19. The specific characteristics of Tweets that were studied included sentiment, language usage, Tweet source, Tweet types (retweets, original Tweets, and replies), and embedded URLs. Tweets posted about the outbreak of Ebola have been studied by researchers to perform sentiment analysis [108] and Tweet content investigation [109]. Researchers in this field have also studied Tweets posted about the outbreak of the Zika virus to perform sentiment analysis [110], to detect the language of the Tweets [111], and to understand the source of the Tweets [112]. Similarly, Tweets about the flu outbreak have also been studied to perform sentiment analysis [113] and for analysis of the used hashtags [114].

In addition to these works, there has been a keen interest in the scientific community to perform sentiment analysis of Tweets related to virus outbreaks and associated matters in the recent past. There were several works that focused on sentiment analysis of Tweets posted during the COVID-19 pandemic. Kaushik et al. [115] used k-means and hierarchical clustering to perform sentiment analysis of Tweets about the COVID-19 pandemic. Jain

et al. [116] used deep learning to address the same research challenge. Marec et al. [117] used concepts of sentiment analysis to detect public sentiments about specific COVID-19 vaccines, such as AstraZeneca/Oxford, Pfizer/BioNTech, and Moderna. The works of Nezhad et al. [118], Agustiningsih et al. [119], and Ponmani et al. [120] presented the results of performing sentiment analysis of Tweets about COVID-19 vaccines from Iran, Indonesia, and India, respectively. In addition to this, the sentiment analysis of relevant Tweets during the COVID-19 pandemic was performed to detect the sentiments of people towards remote work [121], online education [122], social distancing [123], wearing masks [124], and vaccine boosters [125]. This helps to illustrate the keen interest related to performing a comprehensive analysis of the content of Tweets as well as sentiment analysis of Tweets related to virus outbreaks, pandemics, and epidemics in the recent past. As can be seen from this review, none of the prior works in this field focused on analyzing multimodal components of Tweets posted about the ongoing monkeypox outbreak. To address this need, this work presents the findings of a comprehensive content analysis as well as a sentiment analysis of Tweets posted about the ongoing monkeypox outbreak. The methodology that was followed to perform the content analysis and sentiment analysis are outlined in Sections 3.2 and 3.3, respectively. The results and findings of the same are discussed in Sections 4.2 and 4.3, respectively.

3. Methodology

This section is divided into three parts. Section 3.1 presents the specific steps that were followed for the development of this dataset. This section also outlines how this dataset is compliant with the privacy policy, developer agreement, and guidelines for content redistribution of Twitter. Section 3.1 also upholds the compliance of this dataset with the FAIR principles (Findability, Accessibility, Interoperability, and Reusability) for scientific data management. Section 3.2 presents the methodology that was followed for the content analysis of the Tweets. The steps that were used to perform sentiment analysis of the Tweets are presented in Section 3.3.

3.1. Steps for the Development of this Dataset

The dataset was developed by searching Tweets that comprised the keyword(s) "monkeypox" or "monkey pox," posted from 7 May 2022 to 9 October 2022 (the most recent date at the time of resubmission of this paper after the second review round). This search and the associated mining of Tweets were performed as per Twitter API's standard search policies [126] and by using the Advanced Search feature of the Twitter API [127].

In terms of Twitter API's standard search, there are various tools and applications available that comply with these policies and help to search Tweets based on one or more keywords. The specific tool that was used for this work is RapidMiner [128]. RapidMiner was used because of its easy-to-use integrated development environment that allows the development of a range of Big Data and Data Mining-based applications using a combination of both built-in and user-defined functionalities. These built-in functionalities are available in the form of "operators" that can be customized as well as integrated for developing a working application on the RapidMiner platform, known as a "process." The platform also allows the user to develop an "operator" from scratch and bundle the same with other built-in or user-defined "operators" to develop a "process."

For this work, RapidMiner studio, version 9.9.002, was downloaded and installed on a Computer with Microsoft Windows 10 Pro operating system (Version 10.0.19043 Build 19043) comprising of Intel(R) Core(TM) i7-7600U CPU @ 2.80GHz, 2904 Mhz, 2 Core(s), and 4 Logical Processor(s). The specific functionality that was required for this work was searching Tweets based on the matching keyword(s) within a date range. This functionality is already available in RapidMiner Studio 9.9.002 as a built-in "operator" called the Search Twitter "operator" [129] that works by connecting with the Twitter API and by complying with the Twitter API's standard search policies for searching relevant Tweets. Here, relevant Tweets are defined as those Tweets which contain the keyword(s) that are entered as input

to this "operator." So, a "process" was developed in RapidMiner that comprised only the Search Twitter "operator," and it was used to search Tweets that contained either "monkeypox" or "monkey pox" posted on Twitter in the date range of 7 May 2022 to 9 October 2022. This process was run multiple times on a routine basis in this date range to collect the relevant Tweets in compliance with the rate limits of accessing the Twitter API. The screenshot of the "process" that was developed in RapidMiner for the development of this dataset is shown in Figure 1. Table 1 outlines the functionality of the "operators" of this RapidMiner "process".

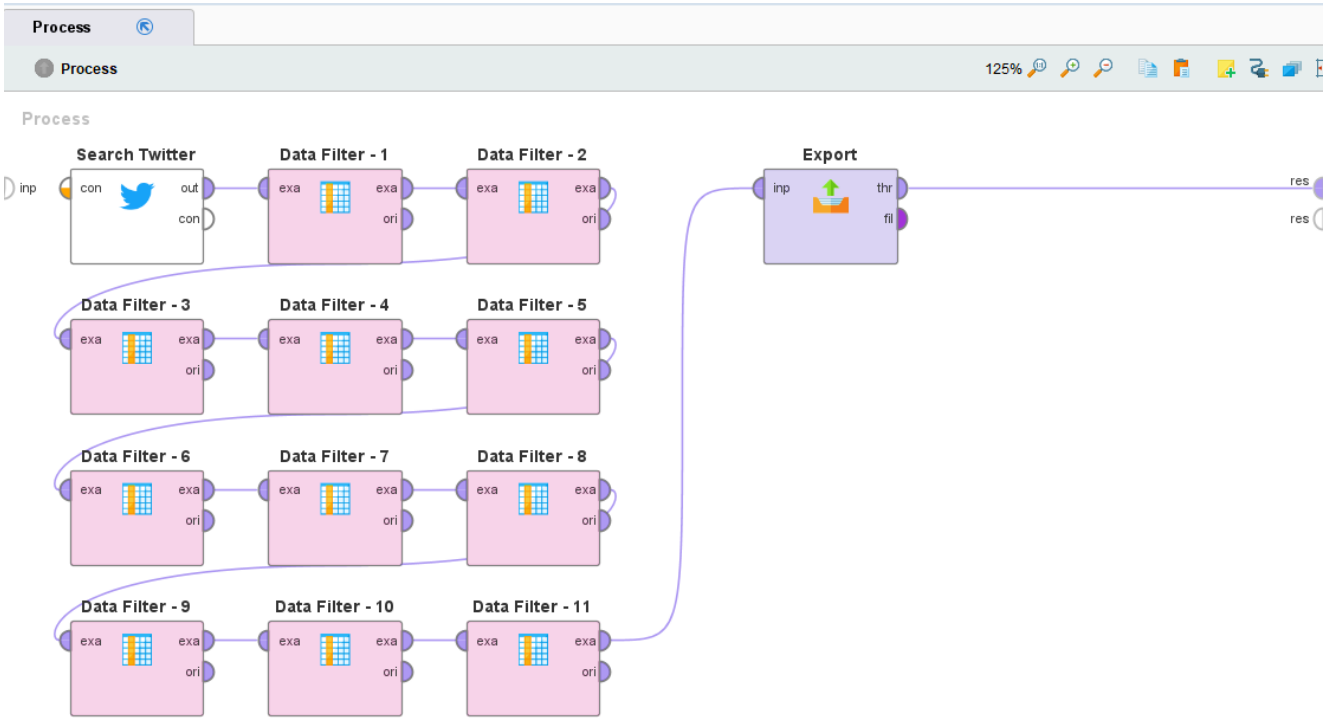

Figure 1. The RapidMiner "process" that was used for the development of this dataset.

The result of this RapidMiner "process" comprised multiple attributes—"Row no", "Id", "Created-At", "From-User", "From-User-Id", "To-User", "To-User-Id", "Language", "Source Text", "Geo-Location-Latitude", "Geo-Location-Longitude", and "Retweet Count". These refer to the row number of the results, the Tweet ID of the obtained Tweet, the date and time when the Tweet was posted, the username of the Twitter user who posted the Tweet, the user ID of the Twitter user who posted the Tweet, Twitter username of the user whose Tweet was replied to (if the Tweet was a reply) in the current Tweet, Twitter user ID of the user whose Tweet was replied to (if the Tweet was a reply) in the current Tweet, the language of the Tweet, source of the Tweet to determine if the Tweet was posted from an Android source, iPhone, Twitter website, etc., the complete text of the Tweet, including embedded URLs, geo-location (latitude) of the user posting the Tweet, geo-location (longitude) of the user posting the Tweet, and retweet count of the Tweet. To comply with the privacy policy, developer agreement, and guidelines for content redistribution of Twitter [130,131], multiple data filters were introduced in the RapidMiner "process" to remove all the attributes from the results other than the "Id" attribute. Thereafter, the results from multiple runs of this "process" were exported.

Table 1. Description of the "operators" of the RapidMiner "process" that was used for the development of this dataset.

Operator Name	Description
Search Twitter	Searches relevant tweets from Twitter by connecting with the Twitter API and by complying with the Twitter API's standard search policies
Data Filter-1	Removes the attribute that contains the date and time when the Tweet m
Data Filter-2	Removes the attribute that contains the Twitter username of the user who posted the Tweet
Data Filter-3	Removes the attribute that contains the Twitter User ID of the user who posted the Tweet
Data Filter-4	Removes the attribute that contains the Twitter username of the user whose Tweet was replied to (if the tweet was a reply) in the current tweet
Data Filter-5	Removes the attribute that contains the Twitter user ID of the user whose Tweet was replied to (if the tweet was a reply) in the current Tweet
Data Filter-6	Removes the attribute that contains the language of the Tweet
Data Filter-7	Removes the attribute that contains the source of the tweet, such as an Android source, Twitter website, etc.
Data Filter-8	Removes the attribute that contains the complete text of the Tweet, including embedded URLs
Data Filter-9	Removes the attribute that contains the geo-location (latitude) of the user posting the Tweet
Data Filter-10	Removes the attribute that contains the geo-location (longitude) of the user posting the Tweet
Data Filter-11	Removes the attribute that contains the retweet count of the Tweet
Export	Exports the result as a .csv file on the local computer

The Advanced Search feature of the Twitter API [127] is available to a user when they are logged in to twitter.com. It allows the user to tailor search results to specific date ranges, people, and more. Specifically, the Advanced Search feature of the Twitter API allows several inputs to be provided, which include specifications to search for Tweets containing all specified keywords in any position, Tweets containing an exact phrase(s), Tweets containing any of the specified keywords, Tweets excluding specific keywords, Tweets with a specific hashtag, and Tweets in a specific language. It also allows the searching of Tweets from a specific account, Tweets sent as replies to a specific account, and Tweets that mention a specific account. This makes it easier to find specific Tweets posted during specific date ranges based on the values of one or more of these inputs. Figure 2 shows two screenshots taken from the Advanced Search feature of the Twitter API that represent the keywords, date range, and other settings that were used to obtain the relevant Tweets. The specific RegEx that was run by the Advanced Search feature of the Twitter API is presented in Figure 3. The results from the Advanced Search feature of the Twitter API were exported, and all the attributes from the Tweets were deleted other than the Tweet IDs to comply with the privacy policy, developer agreement, and guidelines for the content redistribution of Twitter [130,131]. Thereafter, the set of Tweet IDs obtained as a result of the RapidMiner "process" was merged with the set of Tweet IDs obtained as a result of the Advanced Search feature of the Twitter API, and duplicate Tweet IDs were removed to develop this dataset. It is relevant to mention here that neither the results of Twitter API's standard search nor the results of the Advanced Search feature of the Twitter API return an exhaustive list of Tweets posted within a date range. Furthermore, Twitter users are allowed to delete a Tweet they have posted in the past. For a deleted Tweet, there will be no retrievable Tweet text and other related information upon hydration (Section 4.1.1) of that Tweet ID. The description of the dataset files, usage instructions, details for accessing the dataset, and a comparative study to uphold the novelty, relevance, and usefulness of this dataset as compared to prior works on Twitter dataset development (reviewed in Section 2.1) are presented in Section 4.1.

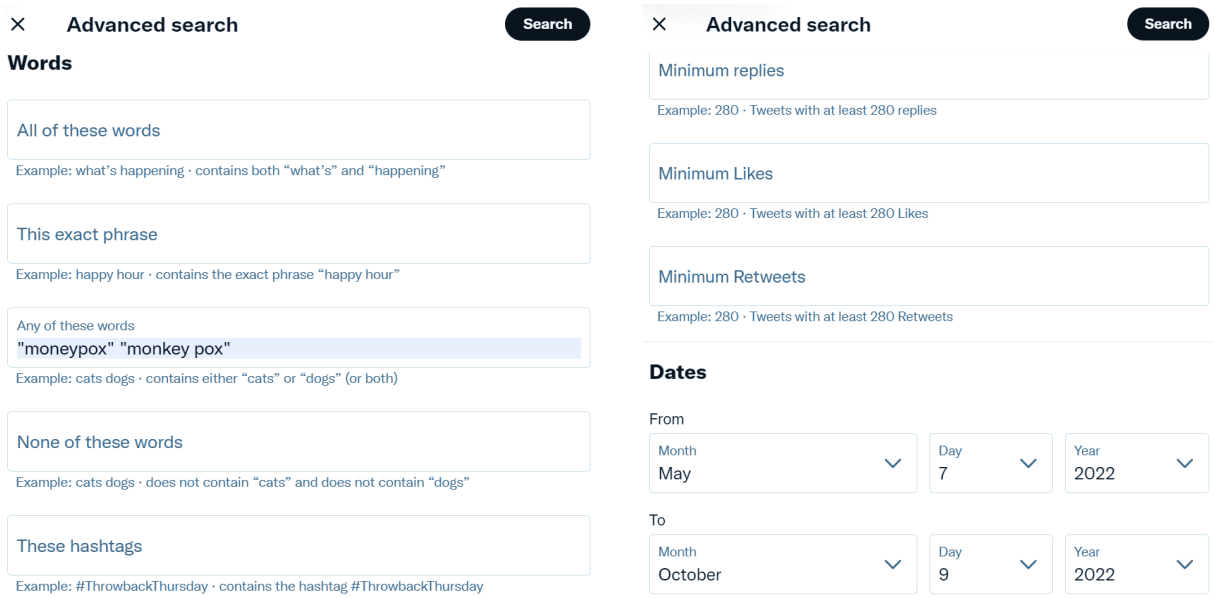

Figure 2. Screenshots from the Advanced Search feature of the Twitter API showing the specific settings that were used to obtain the relevant Tweets in this date range.

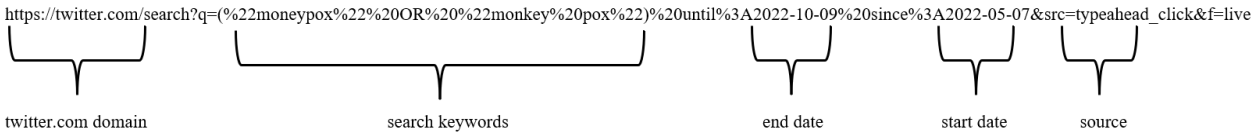

Figure 3. The RegEx that was used by the Advanced Search feature of the Twitter API to obtain the relevant Tweets in this date range.

3.1.1. Compliance with Twitter Policies

According to the privacy policy of Twitter [130]—"Twitter is public, and Tweets are immediately viewable and searchable by anyone around the world." As per the guidelines for Twitter content re-distribution [131]—"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." 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) is in compliance with the privacy policy, developer agreement, and content redistribution guidelines of Twitter.

3.1.2. Compliance with Fair Policies for Scientific Data Management

For a dataset to be compliant with the FAIR principles for scientific data management [132], it should have these four characteristics—Findability, Accessibility, Interoperability, and Reusability. The open-access dataset presented in this paper has a permanent and unique DOI (Section 4.1). The dataset is, therefore, findable and accessible online. The dataset files comprise only .txt files. The .txt files can be downloaded, opened, processed, and interpreted by almost all operating systems and frameworks, such as Windows, Linux, Ubuntu, Android, IOS, and so on, thereby upholding its interoperability. The dataset files present Tweet IDs. These Tweet IDs can be hydrated (Section 4.1.1) to obtain the associated Tweet texts, user IDs, timestamps, retweet count, etc., in compliance with Twitter policies. This information can then be used for multiple use cases and the investigation of different

research questions without the need to perform any other operations on the dataset files. This helps to justify that this dataset also meets the conditions of reusability.

3.2. Steps for Performing Content Analysis of the Tweets of this Dataset

Based on the review of prior works in this field presented in Section 2.3, it can be concluded that performing a comprehensive analysis of the content of the Tweets related to virus outbreaks, epidemics, and pandemics has been of significant interest to researchers in this field in the recent past. Therefore, a comprehensive content analysis of all the Tweets in this dataset was performed. The specific characteristics of the Tweets that were studied include distinct dates when the Tweets were posted, the date when the maximum number of Tweets were posted, distinct languages in which the Tweets are available, the most common language used for posting the Tweets, the total number of different hashtags present in all the Tweets, most commonly used hashtag, the percentage of Tweets posted using an iPhone (Twitter for iPhone), the percentage of Tweets posted using an Android phone (Twitter for Android), and the percentage of Tweets posted using the Twitter website (Twitter Web App). This analysis was performed using RapidMiner [128] using its data analysis features after hydrating the Tweet IDs. The Tweet IDs presented in this dataset must be hydrated prior to using them for the investigation of any research question, application, or use-case scenario. So, the step-by-step process for the hydration of the Tweet IDs is presented as a sub-section in Section 4.1.1. The results of the content analysis of these Tweets are presented and discussed in Section 4.2.

3.3. Steps for Performing Sentiment Analysis of the Tweets of this Dataset

The review of recent works in this field, presented in Section 2.3, also shows that there has been a keen interest in the scientific community to perform sentiment analysis of Tweets related to virus outbreaks and associated matters in the recent past. In view of this fact, as well as to discuss and demonstrate the applicability and effectiveness of this dataset for investigation of different research questions, sentiment analysis of the Tweets was performed. The methodology that was followed for this purpose is outlined in this section.

This study was performed using RapidMiner [128] and its inbuilt "Extract Sentiment" "operator." This "operator" uses the VADER (Valence Aware Dictionary and sEntiment Reasoner) methodology [133] to detect the positive and negative sentiments in Tweets as well as the intensity of the same. It is a lexicon-based sentiment analysis approach that has a time complexity of O(N). The time complexity is the computational complexity that describes the amount of time it takes to run an algorithm [134]. Time complexity is commonly estimated by counting the number of elementary operations performed by the algorithm, supposing that each elementary operation takes a fixed amount of time. Algorithmic complexities are usually represented using the big O notation.

The VADER approach [133] assigns the intensity of sentiments on a scale of −4 to +4. A score of −4 for a negative sentiment means that the associated Tweet is extremely negative. Similarly, a score of 4 for a Tweet means that the associated Tweet is extremely positive. A score of 0 assigned by this approach refers to a neutral sentiment for the associated Tweet. The "process" that was developed in RapidMiner is shown in Figure 4.

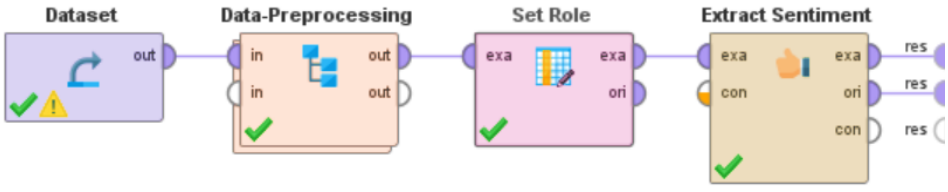

Figure 4. System developed in RapidMiner to perform sentiment analysis of the Tweets.

The "process" worked by detecting the sentiment of each Tweet (in terms of positive, negative, or neutral sentiments) and computing the intensity of the sentiments on a scale

of −4 to +4 using the VADER methodology. The results obtained from this RapidMiner "process" and the associated discussions are presented in Section 4.3.

4. Results and Discussions

This section presents the results and findings of this work. It is divided into three parts. Section 4.1 presents the description of the dataset files. It also discusses step-by-step instructions on how to use this dataset. Furthermore, a comprehensive comparative study of this dataset with 36-prior works in this field that focused on the development of Twitter datasets (Section 2.1) is also presented in this Section. The comparison study shows that the number of Tweet IDs present in this dataset is more as compared to the Tweet IDs present in all these prior works in this field. Section 4.2 presents the findings of a comprehensive content analysis of the Tweets of this dataset. The results of performing sentiment analysis on all the Tweets of this dataset are presented and discussed in Section 4.3. Finally, to support research and development in this field, a list of 50 open research questions related to the fields of Big Data, Data Mining, Natural Language Processing, Machine Learning, and Information Retrieval that may be analyzed and investigated using this dataset is presented in Section 4.3.

4.1. Description of the Dataset, Usage Instructions, and Comparison with Other Twitter Datasets

This open-access dataset is available at https://doi.org/10.7910/DVN/CR7T5E. The dataset consists of a total of 556,427 Tweet IDs of Tweets about monkeypox that were posted on Twitter from 7 May 2022 to 9 October 2022 (the most recent date as per the time of resubmission of this paper after the second review round). At the time of uploading the first version of this dataset on Zenodo [135] and the corresponding preprint of the paper on the preprints.org platform [136], it was the first public Twitter dataset on the 2022 monkeypox outbreak. The Tweet IDs in this dataset are presented in 11 different .txt files based on the timelines of the associated Tweets. Table 2 provides the details of these dataset files. To comply with the privacy policy, developer agreement, and guidelines for the content redistribution of Twitter [130,131], only the Tweet IDs associated with these 556,427 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 [137], Social Media Mining Toolkit [138], and Twarc [139], that work by complying with Twitter policies and the associated rate limits for accessing the Twitter API. Any of these applications 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 4.1.1.

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

Filename	No. of Tweet IDs	Date Range of the Tweet IDs
TweetIDs_Part1.txt	13,926	7 May 2022 to 21 May 2022
TweetIDs_Part2.txt	17,705	21 May 2022 to 27 May 2022
TweetIDs_Part3.txt	17,585	27 May 2022 to 5 June 2022
TweetIDs_Part4.txt	19,718	5 June 2022 to 11 June 2022
TweetIDs_Part5.txt	46,718	12 June 2022 to 30 June 2022
TweetIDs_Part6.txt	138,711	1 July 2022 to 23 July 2022
TweetIDs_Part7.txt	105,890	24 July 2022 to 31 July 2022
TweetIDs_Part8.txt	93,959	1 August 2022 to 9 August 2022
TweetIDs_Part9.txt	50,832	10 August 2022 to 24 August 2022
TweetIDs_Part10.txt	39,042	25 August 2022 to 19 September 2022
TweetIDs_Part11.txt	12,341	20 September 2022 to 9 October 2022

4.1.1. Usage Instructions

This section presents the step-by-step instructions to hydrate this dataset using the Hydrator app:

1. Installation: The desktop version of Hydrator [140] should be downloaded and installed.
2. Twitter Connection: The Hydrator app should be connected to an active Twitter account. This can be performed by clicking on the "Link Twitter Account" button on the app's interface.
3. Dataset File Upload: This step involves uploading a dataset file to the Hydrator app for hydration. Only one file can be added at a time. This can be performed by clicking the "Add" button on the Hydrator app's interface and then selecting one of the dataset files (for example, TweetIDs_Part3.txt) from the local computer. Upon successful file upload, the Hydrator app will show the exact number of Tweet IDs present in the uploaded file. In this case (for TweetIDs_Part3.txt), it will show 17,585.
4. Inputting Dataset Information: This step involves providing certain information about the uploaded dataset file (such as Title, Creator, Publisher, and URL) to the Hydrator app.
5. Completion of Dataset Upload: After completing Step 4, to complete the process of uploading the dataset to the app, the "Add Dataset" button on the app's interface should be clicked.
6. Start Hydration: After successful completion of Step 5, the Hydrator app will automatically redirect to the "Datasets" tab. In this tab, the "Start" button should be clicked to initiate the process of hydrating all the Tweet IDs present in the dataset file.
7. Export Results: The progress indicator on the "Datasets" tab would indicate the successful completion of the hydration of all the Tweet IDs after the process has been completed. Thereafter, the Hydrator app allows the results to be saved in the form of either a. jsonl or .CSV file on the local computer.

As mentioned in Step 3, the Hydrator app allows uploading only one file each time. Therefore, to hydrate all Tweet IDs of this dataset all the files may be merged to form a single .txt file which can then be uploaded to the app. Alternatively, Steps 3 to 7 may be repeated for all the files present in the dataset. After hydrating all the Tweet IDs that are present in the dataset, the Tweets may be used for the investigation and analysis of any of the open research questions mentioned in Section 4.4 or for any similar applications or use-case scenarios or studies.

4.1.2. Comparison with Other Twitter Datasets

As outlined in the review of prior works (Section 2.1) that focused on the development of Twitter datasets on recent issues, global challenges, pandemics, virus outbreaks, emerging topics, current events, politics, and trending topics, just to name a few; there has been no prior work in this field that has focused on the development of a Twitter dataset on the ongoing monkeypox outbreak. The fact that this dataset focuses on the 2022 monkeypox outbreak helps to uphold its novelty.

Recent studies [141–143] have shown that "large-scale" datasets are more helpful for the advancement of research, for improving the quality of innovation, and for supporting better investigation for research questions, as compared to datasets that are not "large-scale" or in other words, datasets that do not consist of a significant amount or quantity of relevant data. Therefore, to further uphold the novelty, relevance, and usefulness of this dataset, a comparative study was performed. The comparative study was characterized by a comparison of the number of Tweet IDs in this dataset with the Tweet IDs of all the datasets associated with prior works in this field. This is summarized in Table 3.

Table 3. Comparison of the number of Tweet IDs present in this dataset with the number of Tweet IDs in 36 prior works in this field that focused on the development of Twitter datasets.

Description of the Twitter Dataset	Number of Tweet IDs
Tweets with #preuambicio [81]	643
Tweets with #Blackwomanhood [73]	919
Tweets with #MiPrimerRecuerdoFeminista [82]	1238
Tweets with #BlackTheory [75]	1430
Tweets with #DuragFest [76]	1705
Tweets about Sundanese (the second-largest tribe in Indonesia) [62]	2518
Tweets about the European migration crisis [49]	3275
Tweets about civil unrest [53]	4381
Tweets involving offensive language [52]	5000
Tweets with #AskTimothy [79]	5680
Tweets involving hate speech and abusive language [48]	5846
Tweets reporting adverse pregnancy outcomes [56]	6487
Tweets involving misogynistic language [51]	6550
Tweets involving opinions of the Indonesian public on different matters [58]	7080
Tweets about BlackLivesMatter [64]	9165
Tweets about the efficacy of hydroxychloroquine as a treatment for COVID-19 [55]	14,374
Tweets with #MarchForBlackWomen [74]	18,646
Tweets about COVID-19 (posted in Spanish) [69]	18,958
Tweets about a severe storm and F1 tornado that struck Central Pennsylvania [59]	22,706
Tweets about COVID-19 (posted in Bengali) [70]	36,117
Tweets containing Multi-Ideology ISIS/Jihadist White Supremacy-based content [61]	40,000
Tweets in the Arabic language [68]	40,000
Tweets about natural hazards [50]	49,816
Tweets with #WITBragDay [80]	52,457
Tweets about Online Learning during the COVID-19 Omicron wave [60]	52,984
Tweets with #WOCAffirmation [78]	80,339
Tweets with #BringBackOurInternet [77]	81,419
Tweets with #IndonesiaHumanRightsSOS [72]	106,903
Tweets about exoskeletons [54]	138,584
Tweets containing the phrase—"I Voted For Trump" [83]	140,000
Tweets containing the word "vaccine" [63]	220,085
Tweets about COVID-19 (posted in English) [71]	226,668
Tweets containing drug-related knowledge [57]	267,215
Tweets about hazardous events at the Baths of Diocletian site in Rome [66]	276,865
Tweets about memes from Black Twitter [67]	402,650
Tweets about the COVID-19 Omicron variant [65]	522,886
Twitter Dataset on the 2022 Monkey Outbreak [this work]	556,427

As can be seen from Table 3, the number of Tweet IDs present in this dataset is more than the number of Tweet IDs in 36 prior works in this field (reviewed in Section 2.1) that were associated with the development of Twitter datasets on recent issues, global

challenges, pandemics, emerging technologies, news, current events, politics, and trending topics in the last few years. It is worth mentioning here that the number of Tweet IDs for some of these works (as mentioned in Table 3) are the numbers that are stated in the associated publications that have been cited. At present, these numbers may be slightly different in some cases, depending on whether the dataset files were updated by the authors of these respective datasets to remove irrelevant Tweets/deleted Tweets and/or to add more recent Tweets after the publications of the associated papers.

4.2. Results of Content Analysis of the Tweets in this Dataset

The findings from content analysis of the Tweets of this dataset are presented and discussed in this Section. First, the Tweet IDs present in this dataset were hydrated by using the Hydrator app as per the steps outlined in Section 4.1.1. Figure 5 is a screenshot of the Hydrator app after the successful completion of the Hydration process that was performed in compliance with Twitter policies and the associated rate limits for accessing the Twitter API. As the Hydrator app allows uploading only one file at a time (Step 3 in Section 4.1.1), so all the .txt files of this dataset were merged to form a single .txt file comprising all the Tweet IDs, which was uploaded to the Hydrator app for performing hydration. It is worth mentioning here that this analysis was performed just prior to the time of the initial submission of this paper using the most recent version of this dataset at that time. The version of the dataset [144] that was used for this analysis contained 254,363 Tweet IDs. That version of the dataset contained Tweet IDs of Tweets about monkeypox posted between 7 May 2022 and 23 July 2022. After the hydration process was completed, the hydrated dataset was uploaded to RapidMiner [128] to perform the data analysis. The free version of RapidMiner allows the analysis of up to 10,000 rows of data. As this dataset has 254,363 rows, the academic license (available to academic researchers) of RapidMiner was applied for, obtained, and downloaded. With the academic license, there is no limit to the number of rows that can be analyzed by using RapidMiner. Prior to performing the analysis, the preprocessing of the data was performed.

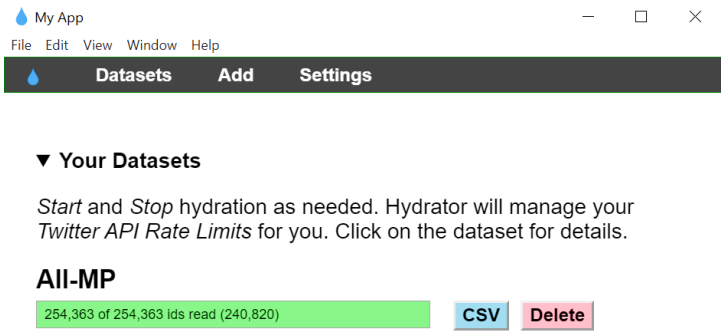

Figure 5. Screenshot from the Hydrator app after the successful Hydration of the entire dataset.

During the data preprocessing stage, it was observed that there were a few Tweets about monkeypox present in this dataset that was originally posted before 7 May 2022 but retweeted on or after 7 May 2022. These Tweets are present in this dataset because even though they were posted before 7 May 2022, their content, such as information about monkeypox, prophecies, conspiracy theories, policies to reduce the spread of the virus, etc., was found to be relevant by Twitter users during this outbreak; as a result, these Tweets were retweeted on or after 7 May 2022. These Tweets were identified and removed by using the data filtration operator in RapidMiner.

The analysis of the timestamp of the Tweets showed that Tweets were posted every day between 7 May 2022 and 23 July 2022. This analysis is shown in Figure 6. In this Figure, the *X*-axis represents the dates, and the *Y*-axis represents the number of Tweets about monkeypox posted on each of these dates. During this analysis, it was observed that while Tweets were posted every day in this date range, on certain dates, a high

number of Tweets were posted. Specifically, on 23 July 2022, the maximum number of Tweets (38,417 Tweets) about monkeypox were posted over a 24-h period. This global interest in tweeting about monkeypox on 23 July 2022 can be attributed to the fact that the WHO declared monkeypox a global public health emergency on this date. Thereafter, the language interpretation of these Tweets was performed. The Tweets were found to have been posted in all 34 languages supported by Twitter [145]. The Twitter developer portal [145] follows a language code for each of these languages. The language code is a code to represent a language in a two- or three-letter format. For instance, "en" stands for English, "ar" stands for Arabic, "bn" stands for Bengali, "cs" stands for Czech, and so on. The percentage of Tweets posted in each of these languages was computed, and the same is represented in Figure 7.

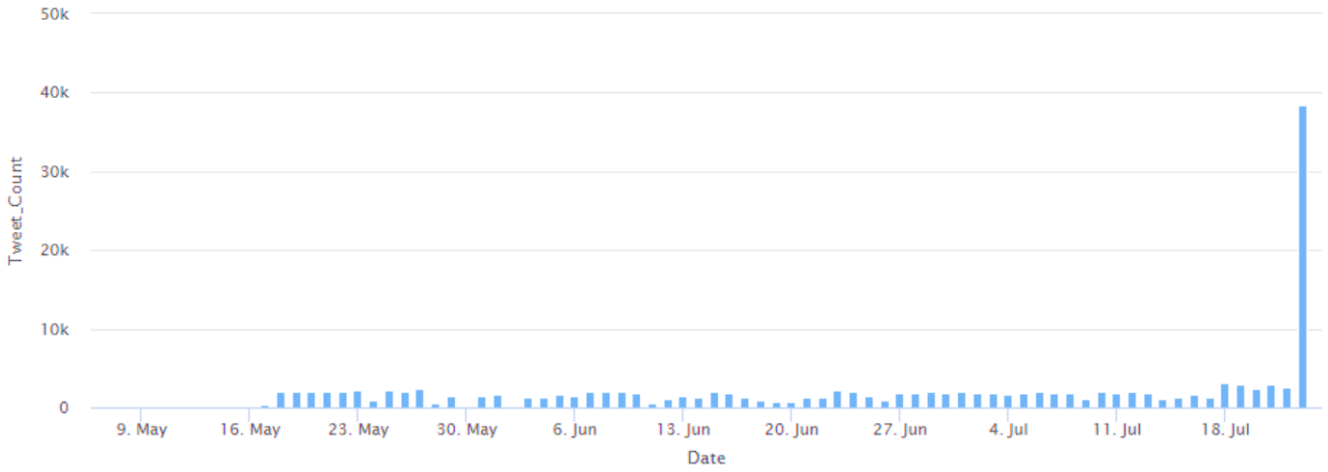

Figure 6. Representation of the number of Tweets posted about monkeypox on each day between 7 May 2022 to 23 July 2022.

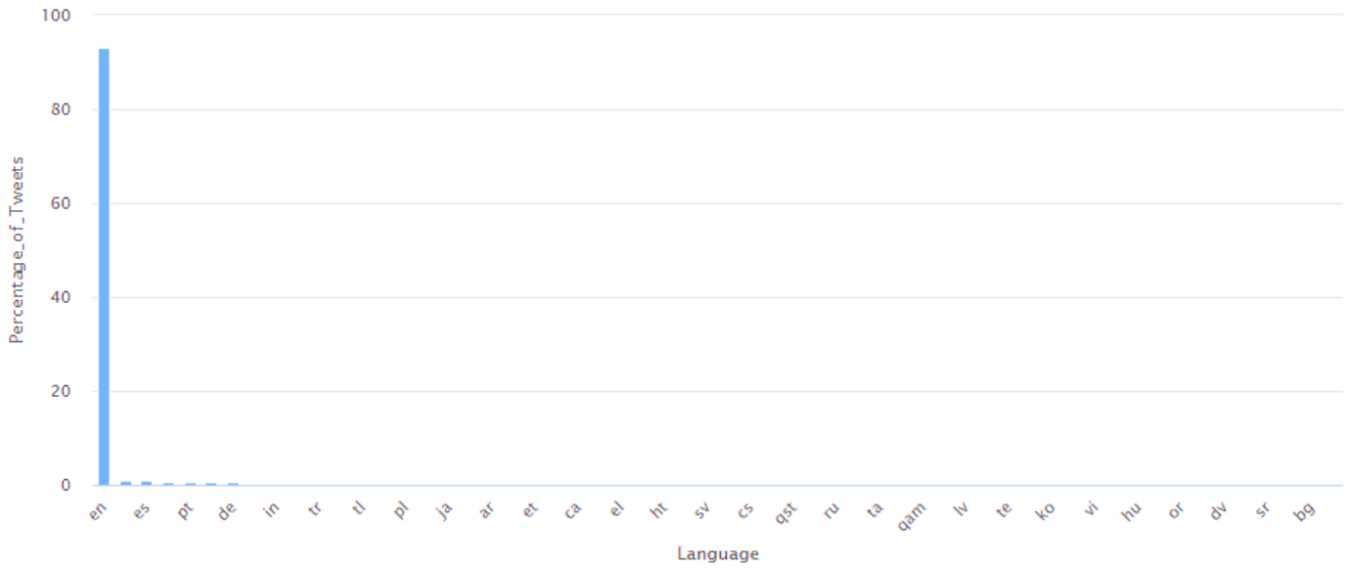

Figure 7. Representation of the different languages (in terms of percentage) in which the Tweets are present in this dataset.

In this Figure, the *X*-axis represents the language, and the *Y*-axis represents the percentage of Tweets posted in that language. As can be seen from this Figure, Tweets posted in English accounted for 93.2% of the Tweets, which was followed by Tweets posted in other languages. Those languages which did not constitute a high percentage of the Tweets

are not shown on the *X*-axis of Figure 4 to enhance its readability. These characteristic features of the Tweets of this dataset, along with other features that were obtained from the analysis, are summarized in Table 4. An aspect of this analysis also included studying the hashtags and the frequencies that were used in these Tweets. There were several Tweets that did not comprise any hashtags. By taking the Tweets that comprised one or more hashtags into consideration, the usage of a total of 5470 distinct hashtags was observed. Out of all these hashtags, "#monkeypox" was found to be the most used hashtag, present in 27.91% of the Tweets. Several other variations of "#monkeypox" in terms of different ordering of uppercase and lowercase characters in the spelling (such as "#Monkeypox", "#MonkeyPox", "#monkeyPox", and "#MONKEYPOX") as well as the use of this phrase with other phrases to create related but different hashtags (such as "#MonkeypoxVirus", "#OMSVarioleDuSingemonkeypox", and "#monkeypoxCOVID19"), were amongst the frequently used hashtags.

Table 4. Characteristic Features of the Tweets present in this dataset.

Characteristic Feature	Statistics
Distinct dates when the Tweets were posted	78
Date when the maximum number of Tweets were posted	23 July 2022
Number of Tweets posted on 23 July 2022	38.417
Distinct languages in which the Tweets are available	34
Most common language used for posting the Tweets	English
Total number of different hashtags present in all the Tweets	5470
Most commonly used hashtag	#monkeypox
Percentage of Tweets posted using an iPhone (Twitter for iPhone)	46.2%
Percentage of Tweets posted using an Android Phone (Twitter for Android)	22.4%
Percentage of Tweets posted using the Twitter Website (Twitter Web App)	20.0%

The Twitter API tracks each Tweet to detect the source that was used to post the Tweet. This is public information as per Twitter policies [130,131] and is displayed as a label with each Tweet on the Twitter platform. The label displays the source, such as Twitter for iPhone (if the Twitter app available for iPhones on the IOS Appstore [146] was used to post the Tweet), Twitter for Android (if the Twitter app available for Android operating systems on the Google Playstore [147] was used to post the Tweet), Twitter for Web (if the twitter.com website [148] was used to post the Tweet), and so on. It was observed that Twitter for iPhone accounted for 46.2% of all the Tweets present in this dataset, Twitter for Android was the source of 22.4% of the Tweets, and Twitter for Web accounted for 20% of the Tweets. The other sources included Tweetdeck [149] and a few similar platforms. Based on this analysis, it can be concluded that most of the Tweets about monkeypox were posted using iPhones as compared to other sources, such as Android Phones, Android Tablets, iPads, the Twitter website, and other platforms.

4.3. Results of Sentiment Analysis of the Tweets in This Dataset

As discussed in Section 3.3, this study was performed using RapidMiner [79] and its inbuilt "Extract Sentiment" "operator," which uses the VADER (Valence Aware Dictionary and sEntiment Reasoner) methodology [133] to detect sentiments (in terms of positive, negative, or neutral sentiments) in Tweets as well as the intensity of the same. The approach assigns the intensity of sentiments on a scale of −4 to +4. The RapidMiner "process" (Figure 4) worked by detecting the sentiment of each Tweet (in terms of positive, negative, or neutral sentiments) and computing the intensity of the sentiments on a scale of −4 to +4 using the VADER methodology. Figure 8 is a screenshot of the results that were computed by RapidMiner after the execution of this "process." From left to right, the attributes

represented in this screenshot are Row No., ID, Date, Score, Scoring String, Negativity, Positivity, Uncovered Tokens, and Total Tokens. These attributes refer to the row number of the results, Tweet ID, truncated date (by removing the time) when the Tweet was posted, the overall score of the Tweet as per the VADER approach, the phrase(s) that contributed towards the score of the Tweet, the intensity of negative sentiment (on a scale of −4 to +4) in the Tweet, the intensity of positive sentiment (on a scale of −4 to +4) in the Tweet, the total number of uncovered tokens in the Tweet, and the total number of tokens present in the Tweet, respectively. This figure shows a sub-sample of the results (from row number 39 to 51) to enhance readability and avoid potential redundancy via the presentation of 254,363 rows of data. Similar to Section 4.2, it is worth mentioning here that this analysis was performed just prior to the time of the initial submission of this paper using the most recent version of this dataset at that time. The version of the dataset [144] that was used for this analysis contained 254,363 Tweet IDs. That version of the dataset contained Tweet IDs of Tweets about monkeypox posted between 7 May 2022 to 23 July 2022. A number of Tweets were observed that did not contain any text and contained images, videos, news articles, and so on. A sentiment score was not computed for such content. Figure 9 shows the classification of these Tweets into positive, negative, and neutral sentiment categories.

Row No.	id	Date	Score	Scoring String	Negativity	Positivity	Uncovered Tokens	Total Tokens
39	1523000000...	5/7/2022	0.436	definitely (0.44)	0	0.436	12	13
40	1523070000...	5/7/2022	-0.282	hope (0.49) racist (-0.77)	0.769	0.487	42	44
41	1523000000...	5/7/2022	0		0	0	34	34
42	1523000000...	5/7/2022	0		0	0	15	15
43	1523010000...	5/7/2022	0		0	0	26	26
44	1523090000...	5/7/2022	0		0	0	4	4
45	1523000000...	5/7/2022	0.462	lucky (0.46)	0	0.462	16	17
46	1523010000...	5/7/2022	-0.718	wtf (-0.72)	0.718	0	5	6
47	1523050000...	5/7/2022	-0.538	cry (-0.54)	0.538	0	13	14
48	1523000000...	5/7/2022	0.077	assured (0.38) worried (-0.31)	0.308	0.385	38	40
49	1523050000...	5/7/2022	0		0	0	22	22
50	1523070000...	5/7/2022	-1.385	adverse (-0.38) serious (-0.08...	2.205	0.821	40	46
51	1522980000...	5/7/2022	0		0	0	8	8

Figure 8. Results of the RapidMiner process for sentiment analysis of the Tweets. This figure shows a sub-sample of the results (from row number 39 to 51) to enhance readability.

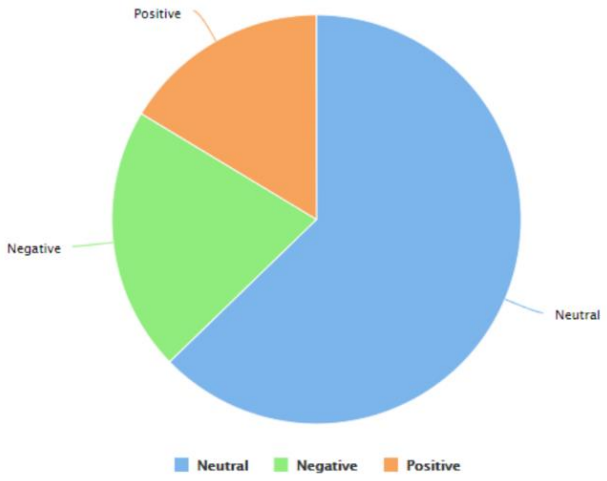

Figure 9. Classification of the Tweets into positive, negative, and neutral classes.

A total of 139,796 Tweets were observed to be neutral, 46,586 Tweets had a negative sentiment, and 36,413 Tweets had a positive sentiment. Therefore, it can be concluded that despite a lot of discussions, debate, opinions, information, and misinformation on Twitter on various topics in this regard, such as monkeypox and the LGBTQI+ community, monkeypox and COVID-19, vaccines for monkeypox, etc., the neutral sentiment is the most common sentiment that has been associated with Tweets about the 2022 monkeypox outbreak thus far. It is worth mentioning here that Tweets posted on Twitter and their associated sentiment are quite often based on recent developments and/or events [150]. For instance, as this paper reports, about 40,000 Tweets related to monkeypox were posted on the day WHO declared monkeypox as GPHE. The monkeypox virus is spreading at a rapid rate while governments in different parts of the world and other policy-making bodies are working to develop policies to reduce the spread of the virus. In the near future, it is possible that specific policies may be recommended by certain governments and/or policy-making bodies which Twitter users of those geographic regions might be in strong disagreement or agreement with. As a result, an influx of Tweets with negative sentiments (for strong disagreement with the policies) or positive sentiments (for strong agreement with the policies) from those geographic regions could be observed on Twitter (recent examples of such Twitter activity include people in certain geographic regions using Twitter to strongly oppose the mask mandates during the early days of COVID-19 [151,152]), which might impact the overall sentiment associated with Tweets about the 2022 outbreak of monkeypox as reported in this study. To address this limitation, when the outbreak ends, this study will be repeated by including all the Tweets about monkeypox that were posted during the entire duration of the outbreak.

In addition to the above, this study on sentiment analysis of Tweets is presented as a potential use-case of this dataset to discuss the applicability of the same for the investigation of the research questions mentioned in Section 4.4 and similar ones. There are several other approaches for sentiment analysis that have been developed in the last couple of years, such as the Bidirectional Long Short-Term Memory (Bi-LSTM) [153], Bidirectional Encoder Representations from Transformers (BERT) [154], and Dialogue Bidirectional Encoder Representations from Transformers (DialBERT) [155] that may also be used for performing sentiment analysis. Furthermore, instead of using RapidMiner as the application development framework, several other options, such as certain libraries in Python and R, may also be used. However, the objective of this work is not to deduce the most optimal and efficient approach for the sentiment analysis of Tweets about the 2022 monkeypox outbreak. Therefore, it does not focus on exploring all these alternative approaches for investigating this research question.

4.4. Open Research Questions

The recent works in the fields of Big Data, Data Mining, Natural Language Processing, Machine Learning, and Information Retrieval related to Twitter data analysis and the development of Twitter datasets, as discussed in Section 2, uphold the fact that Twitter datasets serve as a rich data resource for the investigation of research questions on a wide range of topics as well as for different use case scenarios. Therefore, to further support research and development in this field during the ongoing outbreak, the following is a compilation of 50 open research questions for researchers to study, analyze, evaluate, ideate, and investigate based on this dataset:

1. What is the overall sentiment (positive, negative, or neutral) of the general public related to the outbreak as expressed on Twitter?
2. Which machine learning classifier (such as Random Forest, Decision Trees, Naïve Bayes, etc.) or methodology [156] or approach [153–155] would achieve the best performance accuracy for the sentiment analysis of Tweets related to monkeypox?
3. Are there any specific aspects or subject matters related to the outbreak (such as vaccines, treatments, and protocols to reduce the spread) that are consistently associated with a positive (or negative) sentiment on Twitter?

4. Is there any correlation between the word counts of Tweets about monkeypox and the associated sentiment?
5. What are some of the commonly used hashtags and trends in the same related to Tweets about the outbreak?
6. Are any of the commonly used hashtags in Tweets about the outbreak associated with a specific sentiment?
7. Have there been any trending discussions on Twitter related to one or more matters (such as new protocols to reduce the spread or treatments) concerning the outbreak?
8. Has Twitter played a role in the development and spread of any conspiracy theories about monkeypox?
9. Are any political leaders or popular personalities using Twitter to spread misinformation or fake news related to monkeypox?
10. How is Twitter being used by news organizations, including regional media, local media, national media, and broadcast news agencies, in the dissemination of the latest developments related to the outbreak?
11. What were the specific characteristics of the Tweets (character count, embedded URLs, date, time stamp, etc.) about monkeypox that was retweeted the most?
12. Can the Tweets be analyzed to develop a machine learning classifier that would indicate the accuracy of information about monkeypox expressed in these Tweets from different sources?
13. What are some of the concerns or needs, or complaints about the outbreak expressed by people on Twitter from different geographic regions?
14. Is there any pattern of emoji usage in the Tweets about monkeypox since the beginning of the outbreak?
15. Is there any correlation between the number of Tweets about monkeypox from a geographic region and the number of cases in the same region?
16. What is the best time to Tweet (in terms of highest user engagement and impressions) about a new policy, measure, protocol, or news about monkeypox?
17. Can the content of the Tweets be studied to investigate any potential online stigmatization, discrimination, and/or hate faced by any diversity group, such as the LGBTQI+ community?
18. Do the Tweets reveal any form of panic behavior (such as the panic buying of certain products, as was observed during COVID-19) in regions with a high number of cases?
19. Is there any feedback that individuals infected with the virus have communicated on Twitter related to the treatment they received?
20. Can the Tweets be studied to infer stress or anxiety in individuals tweeting about the virus who are experiencing one or more symptoms after getting infected?
21. What are some of the most popular news outlets from which news has been shared the most on Twitter in the context of the sharing and exchange of information about monkeypox?
22. Can the Tweets be analyzed to develop different user personas in terms of the underlining views, opinions, and perspectives about monkeypox expressed in the Tweets?
23. Can the Point of Interest (POI) of the Tweets [157] be studied to track high-level location information about a place to understand the location-specific opinions, perspectives, or attitudes of the public towards monkeypox?
24. What are the global [158] and region-specific [159] reasons/drives for posting Tweets about monkeypox?
25. How can important Tweets [160] about monkeypox be identified and classified in real-time?
26. Have verified accounts on Twitter played any role in disseminating relevant or irrelevant information [161] about monkeypox since the beginning of the outbreak?
27. Have user diversities, such as gender differences [162,163], played a role in the Tweeting patterns as well as the content of Tweets about monkeypox?

28. Can the gratification theory [164] be applied to these Tweets to deduce any factors or information about the outbreak that gratify Twitter users as expressed in their Tweets?
29. Can the specific information about monkeypox expressed in the Tweets (such as medical opinion, treatment advice, etc.) be studied to determine the profession [165] of the Twitter users who posted those Tweets?
30. Can the Latent Dirichlet Allocation (LDA) model [166] be used to develop an approach that can be applied to Tweets about the outbreak to deduce the credibility of information expressed in every Tweet?
31. Can the Self-Exciting Point Process Model for Predicting Tweet Popularity (SEISMIC) [167] be used on this dataset of Tweets to develop an approach to predict the popularity of Tweets about monkeypox?
32. How can spam accounts and bot accounts be detected that might be responsible for posting spam or incorrect information related to the outbreak?
33. Is there any correlation between posting and/or retweeting research papers [168] about monkeypox and the citations of these papers?
34. What are some of the most common domains (such as biorxiv.org, nature.com, science.org, etc.) that are associated with research papers on monkeypox that have been retweeted the most?
35. Is there any correlation between tagging users while tweeting any new information [169] about the outbreak with the dissemination of that information?
36. What is the overall stance of the general public, as expressed on Twitter, towards the recent developments related to vaccines and treatments for monkeypox?
37. Can a classifier be developed to classify the Tweets into useful and useless suggestions and/or recommendations on factors or topics (such as reducing the spread of the virus) related to the outbreak?
38. Can the iFACT framework [170] be applied to the Tweets to identify, assess, and evaluate the underlying factual information about monkeypox?
39. What are the kinds of "events" [171] in the context of the outbreak that has been expressed in Tweets?
40. Has there been any form of deception (both positive and negative deception) [172] in the context of sharing information related to the outbreak on Twitter?
41. What are some of the trending topics [173] on Twitter about the outbreak?
42. What are some of the "alarming" and "reassuring" information [174] about monkeypox that has been tweeted so far?
43. Can a machine learning-based classifier be developed to detect instances of euphoria or delusion [175] in the context of information seeking and sharing on Twitter related to the outbreak?
44. What are some of the common perceptions [176] of the public related to the recommended vaccines or treatments for monkeypox?
45. Have any Twitter users posted a "regrettable" Tweet [177] about monkeypox that might cause any harm or damage to their reputation?
46. Can concepts of topic extraction and sentiment analysis of Tweets be used to develop a followee recommendation model [178] for Twitter users actively involved in communicating and sharing information about the outbreak?
47. What are the values of different Tweets [179] that have been posted about the outbreak so far?
48. Is there any correlation between the degree of readability of Tweets [180] about the outbreak and the number of comments and/or retweets of those respective Tweets?
49. What is the age group of Twitter users who have posted the most Tweets about monkeypox?
50. What are some of the fake news trends on Twitter related to the outbreak?

5. Conclusions and Scope of Future Work

Twitter datasets serve as a rich data resource for the investigation of different research questions for the timely advancement of knowledge, innovation, and discovery in different fields. Therefore, scientists in this field have focused on developing Twitter datasets on recent issues, global challenges, pandemics, virus outbreaks, emerging technologies, and trending matters in the last few years. In addition to the development of Twitter datasets, analysis of multimodal components of Tweets, specifically Tweets about virus outbreaks, has been of significant interest to the scientific community, as can be seen from several works that focused on analyzing different characteristics of Tweets posted about some of the recent virus outbreaks, such as COVID-19, Ebola, Zika virus, and the flu. The world is currently experiencing an outbreak of the monkeypox virus. A total of 71,096 cases have been reported so far, out of which 70,377 cases have been reported in locations that have not historically reported any monkeypox infections. The World Health Organization (WHO) has declared monkeypox to be a Global Public Health Emergency. This has resulted in a tremendous increase in different types of conversations on Twitter related to monkeypox. None of the prior works in this field have focused on mining these conversations to develop a Twitter dataset. Furthermore, no prior work has analyzed multiple components of these conversations about monkeypox on Twitter. The work presented in this paper aims to address these research challenges. First, it presents an open-access dataset of 556,427 Tweets about monkeypox that were posted on Twitter since the first detected case of this outbreak. Second, the paper reports the results of a comprehensive content analysis of the Tweets of this dataset. This analysis presents several novel findings such as − English has been the most used language (out of all the 34 languages supported by Twitter) to post Tweets about monkeypox, about 40,000 Tweets related to monkeypox were posted on the day WHO declared monkeypox as a GPHE, a total of 5470 distinct hashtags have been used on Twitter about this outbreak out of which #monkeypox is the most used hashtag, and Twitter for iPhone has been the leading source of Tweets about the outbreak. The sentiment analysis of the Tweets was also performed, and the results show that despite a lot of discussions, debate, opinions, information, and misinformation on Twitter on various topics in this regard, such as monkeypox and the LGBTQI+ community, monkeypox and COVID-19, vaccines for monkeypox, etc., "neutral" sentiment was present in most of the Tweets. It was followed by "negative" and "positive" sentiments, respectively. Finally, to support research and development in this field, the paper presents a list of 50 open research questions related to the outbreak in the areas of Big Data, Data Mining, Natural Language Processing, and Machine Learning that may be investigated based on this dataset. Future work on this research project would involve updating the dataset with more recent Tweets on a routine basis to ensure that the scientific community has access to the most recent data in this regard.

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/10.7910/DVN/CR7T5E.

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

References

1. McCollum, A.M.; Damon, I.K. Human Monkeypox. *Clin. Infect. Dis.* **2014**, *58*, 260–267. [CrossRef] [PubMed]
2. von Magnus, P.; Andersen, E.K.; Petersen, K.B.; Birch-Andersen, A. A Pox-like Disease in Cynomolgus Monkeys. *Acta Pathol. Microbiol. Scand.* **2009**, *46*, 156–176. [CrossRef]
3. Breman, J.G.; Kalisa-Ruti; Steniowski, M.V.; Zanotto, E.; Gromyko, A.I.; Arita, I. Human Monkeypox, 1970–1979. *Bull. World Health Organ.* **1980**, *58*, 165–182. [PubMed]

4. Charniga, K.; Masters, N.B.; Slayton, R.B.; Gosdin, L.; Minhaj, F.S.; Philpott, D.; Smith, D.; Gearhart, S.; Alvarado-Ramy, F.; Brown, C.; et al. Estimating the Incubation Period of Monkeypox Virus during the 2022 Multi-National Outbreak. *medRxiv* **2022**, arXiv:2022.06.22.22276713.
5. Jezek, Z.; Szczeniowski, M.; Paluku, K.M.; Mutombo, M. Human Monkeypox: Clinical Features of 282 Patients. *J. Infect. Dis.* **1987**, *156*, 293–298. [CrossRef] [PubMed]
6. Perez Duque, M.; Ribeiro, S.; Martins, J.V.; Casaca, P.; Leite, P.P.; Tavares, M.; Mansinho, K.; Duque, L.M.; Fernandes, C.; Cordeiro, R.; et al. Ongoing Monkeypox Virus Outbreak, Portugal, 29 April to 23 May 2022. *Euro Surveill.* **2022**, *27*, 2200424. [CrossRef]
7. Antinori, A.; Mazzotta, V.; Vita, S.; Carletti, F.; Tacconi, D.; Lapini, L.E.; D'Abramo, A.; Cicalini, S.; Lapa, D.; Pittalis, S.; et al. Epidemiological, Clinical and Virological Characteristics of Four Cases of Monkeypox Support Transmission through Sexual Contact, Italy, May 2022. *Euro Surveill.* **2022**, *27*, 2200421. [CrossRef]
8. Hammerschlag, Y.; MacLeod, G.; Papadakis, G.; Adan Sanchez, A.; Druce, J.; Taiaroa, G.; Savic, I.; Mumford, J.; Roberts, J.; Caly, L.; et al. Monkeypox Infection Presenting as Genital Rash, Australia, May 2022. *Euro Surveill.* **2022**, *27*, 2200411. [CrossRef]
9. Huhn, G.D.; Bauer, A.M.; Yorita, K.; Graham, M.B.; Sejvar, J.; Likos, A.; Damon, I.K.; Reynolds, M.G.; Kuehnert, M.J. Clinical Characteristics of Human Monkeypox, and Risk Factors for Severe Disease. *Clin. Infect. Dis.* **2005**, *41*, 1742–1751. [CrossRef]
10. Adler, H.; Gould, S.; Hine, P.; Snell, L.B.; Wong, W.; Houlihan, C.F.; Osborne, J.C.; Rampling, T.; Beadsworth, M.B.; Duncan, C.J.; et al. Clinical Features and Management of Human Monkeypox: A Retrospective Observational Study in the UK. *Lancet Infect. Dis.* **2022**, *22*, 1153–1162. [CrossRef]
11. Learned, L.A.; Reynolds, M.G.; Wassa, D.W.; Li, Y.; Olson, V.A.; Karem, K.; Stempora, L.L.; Braden, Z.H.; Kline, R.; Likos, A.; et al. Extended Interhuman Transmission of Monkeypox in a Hospital Community in the Republic of the Congo, 2003. *Am. J. Trop. Med. Hyg.* **2005**, *73*, 428–434. [CrossRef] [PubMed]
12. Centers for Disease Control and Prevention (CDC) Multistate Outbreak of Monkeypox–Illinois, Indiana, and Wisconsin, 2003. *MMWR Morb. Mortal. Wkly. Rep.* **2003**, *52*, 537–540.
13. Public Health England. Monkeypox Case Confirmed in England. 2019. Available online: https://www.gov.uk/government/news/monkeypox-case-confirmed-in-england (accessed on 30 July 2022).
14. Erez, N.; Achdout, H.; Milrot, E.; Schwartz, Y.; Wiener-Well, Y.; Paran, N.; Politi, B.; Tamir, H.; Israely, T.; Weiss, S.; et al. Diagnosis of Imported Monkeypox, Israel, 2018. *Emerg. Infect. Dis.* **2019**, *25*, 980–983. [CrossRef] [PubMed]
15. CDC. 2022 Monkeypox Outbreak Global Map. Available online: https://www.cdc.gov/poxvirus/monkeypox/response/2022/world-map.html (accessed on 22 August 2022).
16. Saxena, S.K.; Ansari, S.; Maurya, V.K.; Kumar, S.; Jain, A.; Paweska, J.T.; Tripathi, A.K.; Abdel-Moneim, A.S. Re-Emerging Human Monkeypox: A Major Public-Health Debacle. *J. Med. Virol.* **2022**, *2*, 902. [CrossRef]
17. Isidro, J.; Borges, V.; Pinto, M.; Ferreira, R.; Sobral, D.; Nunes, A.; Snatos, D.; Borrego, M.J.; Nuncio, S.; Pelerito, A.; et al. First Draft Genome Sequence of Monkeypox Virus Associated with the Suspected Multi-Country Outbreak, May 2022 (Confirmed Case in Portugal). Available online: https://virological.org/t/first-draft-genome-sequence-of-monkeypox-virus-associated-with-the-suspected-multi-country-outbreak-may-2022-confirmed-case-in-portugal/799 (accessed on 30 July 2022).
18. Mauldin, M.R.; McCollum, A.M.; Nakazawa, Y.J.; Mandra, A.; Whitehouse, E.R.; Davidson, W.; Zhao, H.; Gao, J.; Li, Y.; Doty, J.; et al. Exportation of Monkeypox Virus from the African Continent. *J. Infect. Dis.* **2022**, *225*, 1367–1376. [CrossRef]
19. Grover, N.; Rigby, J. WHO Calls Emergency Meeting as Monkeypox Cases Top 100 in Europe. *Reuters* **2022**.
20. Grover, N. Unlikely Monkeypox Outbreak Will Lead to Pandemic, WHO Says. *Reuters* **2022**.
21. Kelleher, S.R. CDC Raises Monkeypox Travel Alert to Level 2. Available online: https://www.forbes.com/sites/suzannerowankelleher/2022/06/07/cdc-raises-monkeypox-travel-alert-to-level-2/?sh=269eee1e3f93 (accessed on 30 July 2022).
22. Kozlov, M. Monkeypox Declared a Global Emergency: Will It Help Contain the Outbreak? *Nature* **2022**. [CrossRef]
23. Kimball, S. Monkeypox Eradication Unlikely in the U.S. as Virus Could Spread Indefinitely, CDC Says. Available online: https://www.cnbc.com/2022/10/01/monkeypox-unlikely-to-be-eliminated-in-the-us-cdc-says.html (accessed on 22 August 2022).
24. Murugesu, J.A. Monkeypox emergency. *New Sci.* **2022**, *255*, 7. [CrossRef]
25. Mega, E.R. Why Scientists Fear Monkeypox Spreading in Wild Animals. *Nature* **2022**. [CrossRef]
26. CDC. Treatment Information for Healthcare Professionals. Available online: https://www.cdc.gov/poxvirus/monkeypox/clinicians/treatment.html (accessed on 30 July 2022).
27. Available online: https://assets.publishing.service.gov.uk/government (accessed on 30 July 2022).
28. SIGA. Receives Approval from the FDA for Intravenous (IV) Formulation of TPOXX® (Tecovirimat). Available online: https://investor.siga.com/news-releases/news-release-details/siga-receives-approval-fda-intravenous-iv-formulation-tpoxxr (accessed on 30 July 2022).
29. CDC. Vaccines. Available online: https://www.cdc.gov/poxvirus/monkeypox/vaccines/index.html (accessed on 22 August 2022).
30. Gilchrist, K. Belgium Becomes First Country to Introduce Mandatory Monkeypox Quarantine as Global Cases Rise. Available online: https://www.cnbc.com/2022/05/23/belgium-introduces-mandatory-monkeypox-quarantine-as-global-cases-rise.html (accessed on 30 July 2022).
31. Bahl, R. Monkeypox Vaccine: U.S. Orders 500,000 Jynneos Doses as Cases Rise. Available online: https://www.healthline.com/health-news/monkeypox-vaccine-existing-vaccines-provide-strong-protection-one-fda-approved (accessed on 30 July 2022).

32. MSN. Toronto to Offer Monkeypox Vaccine Clinics Targeting High-Risk Communities. Available online: https://www.msn.com/en-ca/news/canada/toronto-to-offer-monkeypox-vaccine-clinics-targeting-high-risk-communities/ar-AAYIYhH (accessed on 30 July 2022).
33. With, A.M. Push for Targeted Monkeypox Vaccine Rollout in France, Denmark. Available online: https://www.rfi.fr/en/europe/20220525-push-for-targeted-monkeypox-vaccine-rollout-in-france-denmark (accessed on 30 July 2022).
34. Monkeypox: German Panel Recommends Vaccine for Risk Groups. Available online: https://www.dw.com/en/monkeypox-german-panel-recommends-vaccine-for-risk-groups/a-62084728 (accessed on 30 July 2022).
35. UK Health Authority Advises Self-Isolation for Monkeypox Infections. Available online: https://www.thesundaily.my/world/uk-health-authority-advises-self-isolation-for-monkeypox-infections-NE9314681 (accessed on 30 July 2022).
36. da Costa, V.C.F.; Oliveira, L.; de Souza, J. Internet of Everything (IoE) Taxonomies: A Survey and a Novel Knowledge-Based Taxonomy. *Sensors* **2021**, *21*, 568. [CrossRef] [PubMed]
37. Perrin, A.; Smith, A.; Duggan, M.; Greenwood, S.; Porteus, M.; Page, D. Social Media Usage. Available online: https://www.secretintelligenceservice.org/wp-content/uploads/2016/02/PI_2015-10-08_Social-Networking-Usage-2005-2015_FINAL.pdf (accessed on 30 July 2022).
38. Noor Al-Deen, H.S.; Hendricks, J.A. *Social Media: Usage and Impact*; Lexington Books: Laham, MD, USA, 2011; ISBN 9780739167304.
39. Kavada, A. Social Media as Conversation: A Manifesto. *Soc. Media Soc.* **2015**, *1*, 205630511558079. [CrossRef]
40. Smith, A.; Brenner, J. Twitter Use 2012. Available online: https://www.pewinternet.org/wp-content/uploads/sites/9/media/Files/Reports/2012/PIP_Twitter_Use_2012.pdf (accessed on 30 July 2022).
41. Morgan-Lopez, A.A.; Kim, A.E.; Chew, R.F.; Ruddle, P. Predicting Age Groups of Twitter Users Based on Language and Metadata Features. *PLoS ONE* **2017**, *12*, e0183537. [CrossRef] [PubMed]
42. Iqbal, M. Twitter Revenue and Usage Statistics. 2022. Available online: https://www.businessofapps.com/data/twitter-statistics (accessed on 30 July 2022).
43. Ring, T. Twitter Adds Link for Accurate Info on Monkeypox. Available online: https://www.advocate.com/health/2022/8/16/twitter-adds-link-accurate-info-monkeypox (accessed on 22 August 2022).
44. Cao, S. Medical Experts Are Becoming Influencers amid All the Anxiety over Monkeypox. Available online: https://www.buzzfeednews.com/article/stefficao/monkeypox-influencers-medical-expert-hysteria (accessed on 22 August 2022).
45. Wiggins, C. Rep. Marjorie Taylor Greene Tweets Monkeypox Disinformation. Available online: https://www.advocate.com/news/2022/7/25/rep-marjorie-taylor-greene-tweets-monkeypox-disinformation (accessed on 22 August 2022).
46. McGee, K. UT-Dallas Is Investigating a Professor's Homophobic Tweet with Misinformation about Monkeypox. The Texas Tribune 2022. Available online: https://www.texastribune.org/2022/07/20/ut-dallas-monkeypox-lgbtq/ (accessed on 22 August 2022).
47. Niemietz, B. Prominent Medical Writer's Typo Warns Sex with 'Me' Can Lead to Monkeypox. Daily News. 2022. Available online: https://www.nydailynews.com/news/national/ny-medical-writer-benjamin-ryan-monkeypox-20220721-lyzp26f7obhgtc2nnplvxiumtm-story.html (accessed on 22 August 2022).
48. Mulki, H.; Haddad, H.; Bechikh Ali, C.; Alshabani, H. L-HSAB: A Levantine Twitter Dataset for Hate Speech and Abusive Language. In Proceedings of the Third Workshop on Abusive Language Online, Florence, Italy, 1 August 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 111–118.
49. 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; pp. 146–149.
50. Meng, L.; Dong, Z.S. Natural Hazards Twitter Dataset. *arXiv* **2020**, arXiv:2004.14456.
51. Mulki, H.; Ghanem, B. Let-Mi: An Arabic Levantine Twitter Dataset for Misogynistic Language. *arXiv* **2021**, arXiv:2103.10195.
52. Manolescu, M.; University of Tübingen, Germany; Çöltekin, Ç. ROFF—A Romanian Twitter Dataset for Offensive Language. In Proceedings of the Conference Recent Advances in Natural Language Processing—Deep Learning for Natural Language Processing Methods and Applications, Varna, Bulgaria, 1–3 September 2021.
53. 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), https://aclanthology.org/volumes/2020.wnut-1/, Online, 19 November 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 215–221.
54. Thakur, N. Twitter Big Data as a Resource for Exoskeleton Research: A Large-Scale Dataset of about 140,000 Tweets from 2017–2022 and 100 Research Questions. *Analytics* **2022**, *1*, 72–97. [CrossRef]
55. Mutlu, E.C.; Oghaz, T.; Jasser, J.; Tutunculer, E.; Rajabi, A.; Tayebi, A.; Ozmen, O.; Garibay, I. A Stance Data Set on Polarized Conversations on Twitter about the Efficacy of Hydroxychloroquine as a Treatment for COVID-19. *Data Brief* **2020**, *33*, 106401. [CrossRef]
56. Klein, A.Z.; Gonzalez-Hernandez, G. An Annotated Data Set for Identifying Women Reporting Adverse Pregnancy Outcomes on Twitter. *Data Brief* **2020**, *32*, 106249. [CrossRef]
57. Sarker, A.; Gonzalez, G. A Corpus for Mining Drug-Related Knowledge from Twitter Chatter: Language Models and Their Utilities. *Data Brief* **2017**, *10*, 122–131. [CrossRef]
58. Riccosan; Saputra, K.E.; Pratama, G.D.; Chowanda, A. Emotion Dataset from Indonesian Public Opinion. *Data Brief* **2022**, *43*, 108465. [CrossRef]

59. Grace, R. Crisis Social Media Data Labeled for Storm-Related Information and Toponym Usage. *Data Brief* **2020**, *30*, 105595. [CrossRef] [PubMed]
60. Thakur, N. A Large-Scale Dataset of Twitter Chatter about Online Learning during the Current COVID-19 Omicron Wave. *Data* **2022**, *7*, 109. [CrossRef]
61. Gaikwad, M.; Ahirrao, S.; Phansalkar, S.; Kotecha, K. Multi-Ideology ISIS/Jihadist White Supremacist (MIWS) Dataset for Multi-Class Extremism Text Classification. *Data* **2021**, *6*, 117. [CrossRef]
62. Putra, O.V.; Wasmanson, F.M.; Harmini, T.; Utama, S.N. Sundanese Twitter Dataset for Emotion Classification. In Proceedings of the 2020 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM), Surabaya, Indonesia, 17–18 November 2020.
63. Averza, A.A. Twitter Dataset—Over 200,000 Tweets Containing the Word "Vaccine" for Research Porpuses. 2022. Available online: https://ieee-dataport.org/documents/twitter-dataset-over-200000-tweets-containing-word-vaccine-research-porpuses (accessed on 22 August 2022).
64. Pranesh, R.R.; Kumar, S.; Shekhar, A. TweetBLM: A Hate Speech Dataset and Analysis of BlackLivesMatter-Related Microblogs on Twitter. 2020. Available online: https://zenodo.org/record/4000539#.Y0W3F3bMLEY.
65. 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*, 76. [CrossRef]
66. RESILTECH s.r.l. Dataset of Tweets, Used to Detect Hazardous Events at the Baths of Diocletian Site in Rome. 2019. Available online: https://zenodo.org/record/3258416#.Y0W4anbMLEY.
67. Jules, B. BestofThrowBackBlackTwitter. 2019. Available online: https://zenodo.org/record/4976950#.Y0W6SXbMLEY.
68. Kora, R.; Mohammed, A. Corpus on Arabic Egyptian Tweets. 2019. Available online: https://doi.org/10.7910/DVN/LBXV9O. [CrossRef]
69. Garain, A. COVID-19 Tweets Dataset for Spanish Language. 2020. Available online: https://ieee-dataport.org/open-access/covid-19-tweets-dataset-spanish-language (accessed on 22 August 2022).
70. Garain, A. COVID-19 Tweets Dataset for Bengali Language. 2020. Available online: https://ieee-dataport.org/open-access/covid-19-tweets-dataset-bengali-language (accessed on 22 August 2022).
71. Garain, A. English Language Tweets Dataset for COVID-19. 2020. Available online: https://ieee-dataport.org/open-access/english-language-tweets-dataset-covid-19 (accessed on 22 August 2022).
72. Nawwar, A. #IndonesiaHumanRightsSOS Twitter Hashtag Tweets Dataset. 2020. Available online: https://zenodo.org/record/4362505#.Y0W1XnbMLEY.
73. Jules, B. Dataset of Tweets with #Blackwomanhood. 2018. Available online: https://zenodo.org/record/4944545#.Y0W4yXbMLEY.
74. Jules, B. Dataset of Tweets with #MarchForBlackWomen. 2017. Available online: https://zenodo.org/record/5018193#.Y0W5TXbMLEY.
75. Jules, B. Dataset of Tweets with #BlackTheory. 2017. Available online: https://zenodo.org/record/4950437#.Y0W7SHbMLEY.
76. Jules, B. Dataset of Tweets with #DuragFest. 2018. Available online: https://zenodo.org/record/4938042#.Y0W7SnbMLEY.
77. Jules, B. Dataset of Tweets with #BringBackOurInternet. 2017. Available online: https://zenodo.org/record/4973415#.Y0W70nbMLEY.
78. Jules, B. Dataset of Tweets with #WOCAffirmation. 2017. Available online: https://zenodo.org/record/4993283#.Y0W8t3bMLEY.
79. Jules, B. Dataset of Tweets with #AskTimothy. 2018. Available online: https://zenodo.org/record/4958263#.Y0W-zHbMLEY.
80. Wrubel, Laura (George Washington University) WITBragDay Tweet Ids. 2017. Available online: https://doi.org/10.7910/DVN/IRNS5Z. [CrossRef]
81. Maria, A. Dataset of Tweets with #preuambicio 2021/03/04 to 2021/05/21. 2021. Available online: https://doi.org/10.7910/DVN/DVXTCX. [CrossRef]
82. Maria, A. Dataset of Tweets with #MiPrimerRecuerdoFeminista 2020.03.06–2020.03.11. 2020. Available online: https://doi.org/10.7910/DVN/3GAZGD. [CrossRef]
83. Jules, B. Dataset of Tweets with the phrase—"I Voted For Trump". 2017. Available online: https://zenodo.org/record/4940956#.Y0W9eHbMLEY.
84. Weissenbacher, D.; Sarker, A.; Klein, A.; O'Connor, K.; Magge, A.; Gonzalez-Hernandez, G. Deep Neural Networks Ensemble for Detecting Medication Mentions in Tweets. *J. Am. Med. Inform. Assoc.* **2019**, *26*, 1618–1626. [CrossRef]
85. Sarker, A.; Gonzalez-Hernandez, G.; Ruan, Y.; Perrone, J. Machine Learning and Natural Language Processing for Geolocation-Centric Monitoring and Characterization of Opioid-Related Social Media Chatter. *JAMA Netw. Open* **2019**, *2*, e1914672. [CrossRef]
86. Klein, A.Z.; Sarker, A.; Cai, H.; Weissenbacher, D.; Gonzalez-Hernandez, G. Social Media Mining for Birth Defects Research: A Rule-Based, Bootstrapping Approach to Collecting Data for Rare Health-Related Events on Twitter. *J. Biomed. Inform.* **2018**, *87*, 68–78. [CrossRef]
87. Al-Garadi, M.A.; Yang, Y.-C.; Cai, H.; Ruan, Y.; O'Connor, K.; Graciela, G.-H.; Perrone, J.; Sarker, A. Text Classification Models for the Automatic Detection of Nonmedical Prescription Medication Use from Social Media. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 27. [CrossRef]

88. Al-Garadi, M.A.; Yang, Y.-C.; Lakamana, S.; Lin, J.; Li, S.; Xie, A.; Hogg-Bremer, W.; Torres, M.; Banerjee, I.; Sarker, A. Automatic Breast Cancer Cohort Detection from Social Media for Studying Factors Affecting Patient-Centered Outcomes. In *Artificial Intelligence in Medicine*; Springer International Publishing: Cham, Germany, 2020; pp. 100–110, ISBN 9783030591366.
89. Tekumalla, R.; Banda, J.M. Using Weak Supervision to Generate Training Datasets from Social Media Data: A Proof of Concept to Identify Drug Mentions. *Neural Comput. Appl.* **2021**, 1–9. [CrossRef]
90. Farooq, H.; Naveed, H. GPADRlex: Grouped Phrasal Adverse Drug Reaction Lexicon. In Proceedings of the 2019 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), Dalian, China, 20–23 September 2019.
91. Glandt, K.; Khanal, S.; Li, Y.; Caragea, D.; Caragea, C. Stance Detection in COVID-19 Tweets. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Online, 1–6 August 2021; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 1596–1611.
92. Kumari, R.; Ashok, N.; Ghosal, T.; Ekbal, A. Misinformation Detection Using Multitask Learning with Mutual Learning for Novelty Detection and Emotion Recognition. *Inf. Process. Manag.* **2021**, *58*, 102631. [CrossRef]
93. Kumari, R.; Ashok, N.; Ghosal, T.; Ekbal, A. A Multitask Learning Approach for Fake News Detection: Novelty, Emotion, and Sentiment Lend a Helping Hand. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzen, China, 18–22 September 2021; pp. 1–8.
94. Hua, Y.; Jiang, H.; Lin, S.; Yang, J.; Plasek, J.M.; Bates, D.W.; Zhou, L. Using Twitter Data to Understand Public Perceptions of Approved versus Off-Label Use for COVID-19-Related Medications. *J. Am. Med. Inform. Assoc.* **2022**, *29*, 1668–1678. [CrossRef] [PubMed]
95. Do, T.T.; Nguyen, D.; Le, A.; Nguyen, A.; Nguyen, D.; Hoang, N.; Le, U.; Tran, T. Understanding Public Opinion on Using Hydroxychloroquine for COVID-19 Treatment via Social Media. *arXiv* **2022**, arXiv:2201.00237.
96. Ceslov, R. Detecting Stance on COVID-19 Vaccine in a Polarized Media. 2021. Available online: https://academicworks.cuny.edu/gc_etds/4616/ (accessed on 22 August 2022).
97. Miura, F.; van Ewijk, C.E.; Backer, J.A.; Xiridou, M.; Franz, E.; Op de Coul, E.; Brandwagt, D.; van Cleef, B.; van Rijckevorsel, G.; Swaan, C.; et al. Estimated Incubation Period for Monkeypox Cases Confirmed in the Netherlands, May 2022. *Euro Surveill.* **2022**, *27*, 2200448. [CrossRef] [PubMed]
98. Bragazzi, N.L.; Khamisy-Farah, R.; Tsigalou, C.; Mahroum, N.; Converti, M. Attaching a Stigma to the LGBTQI+ Community Should Be Avoided during the Monkeypox Epidemic. *J. Med. Virol.* **2022**. [CrossRef]
99. Dashraath, P.; Nielsen-Saines, K.; Mattar, C.; Musso, D.; Tambyah, P.; Baud, D. Guidelines for Pregnant Individuals with Monkeypox Virus Exposure. *Lancet* **2022**, *400*, 21–22. [CrossRef]
100. Kampf, G. Efficacy of Biocidal Agents and Disinfectants against the Monkeypox Virus and Other Orthopoxviruses. *J. Hosp. Infect.* **2022**, *127*, 101–110. [CrossRef]
101. Nörz, D.; Pfefferle, S.; Brehm, T.T.; Franke, G.; Grewe, I.; Knobling, B.; Aepfelbacher, M.; Huber, S.; Klupp, E.M.; Jordan, S.; et al. Evidence of Surface Contamination in Hospital Rooms Occupied by Patients Infected with Monkeypox, Germany, June 2022. *Euro Surveill.* **2022**, *27*, 2200477. [CrossRef]
102. Abbas, S.; Karam, S.; Schmidt-Sane, M.; Palmer, J. *Social Considerations for Monkeypox Response*; Institute of Development Studies: Brighton, UK, 2022.
103. Mungmunpuntipantip, R.; Wiwanitkit, V. Diarrhea and Monkeypox: A Consideration. Rev. *Esp. Enferm. Dig.* **2022**, *in press*. [CrossRef]
104. Sallam, M.; Al-Mahzoum, K.; Dardas, L.A.; Al-Tammemi, A.B.; Al-Majali, L.; Al-Naimat, H.; Jardaneh, L.; AlHadidi, F.; Al-Salahat, K.; Al-Ajlouni, E.; et al. Knowledge of Human Monkeypox and Its Relation to Conspiracy Beliefs among Students in Jordanian Health Schools: Filling the Knowledge Gap on Emerging Zoonotic Viruses. *Medicina* **2022**, *58*, 924. [CrossRef]
105. Md, A.P.; Ahsan, M.; Ramiz Uddin, M.; Luna, S.A. Monkeypox Image Data Collection. *arXiv* **2022**, arXiv:2206.01774.
106. Malik, A.A.; Winters, M.S.; Omer, S.B. Attitudes of the US General Public towards Monkeypox. *bioRxiv* **2022**, arXiv: 2022.06.20.22276527.
107. Sypsa, V.; Mameletzis, I.; Tsiodras, S. Transmission Potential of Human Monkeypox in Mass Gatherings. *bioRxiv* **2022**, arXiv:2022.06.21.22276684. [CrossRef] [PubMed]
108. 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]
109. Odlum, M.; Yoon, S. What Can We Learn about the Ebola Outbreak from Tweets? *Am. J. Infect. Contro.* **2015**, *43*, 563–571. [CrossRef]
110. 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]
111. 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]
112. Yang, J.-A.J. Spatial-Temporal Analysis of Information Diffusion Patterns with User-Generated Geo-Social Contents from Social Media. 2017. Available online: https://www.proquest.com/openview/34ee11be7a87469d16f0ae25d1bc99aa/ (accessed on 22 August 2022).

113. 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 Publishing: Cham, Switzerland, 2018; pp. 174–186, ISBN 9783319961354.
114. Hellsten, I.; Jacobs, S.; Wonneberger, A. Active and Passive Stakeholders in Issue Arenas: A Communication Network Approach to the Bird Flu Debate on Twitter. *Public Relat. Rev.* **2019**, *45*, 35–48. [CrossRef]
115. Kaushik, N.; Bhatia, M.K. Twitter Sentiment Analysis Using K-Means and Hierarchical Clustering on COVID Pandemic. In *Advances in Intelligent Systems and Computing*; Springer Singapore: Singapore, 2022; pp. 757–769, ISBN 9789811625930.
116. Jain, R.; Bawa, S.; Sharma, S. Sentiment Analysis of COVID-19 Tweets by Machine Learning and Deep Learning Classifiers. In *Advances in Data and Information Sciences*; Springer Singapore: Singapore, 2022; pp. 329–339, ISBN 9789811656880.
117. Marcec, R.; Likic, R. Using Twitter for Sentiment Analysis towards AstraZeneca/Oxford, Pfizer/BioNTech and Moderna COVID-19 Vaccines. *Postgrad. Med. J.* **2022**, *98*, 544–550. [CrossRef]
118. Bokaee Nezhad, Z.; Deihimi, M.A. Twitter Sentiment Analysis from Iran about COVID 19 Vaccine. *Diabetes Metab. Syndr.* **2022**, *16*, 102367. [CrossRef]
119. Agustiningsih, K.K.; Utami, E.; Alsyaibani, M.A. Sentiment Analysis of COVID-19 Vaccines in Indonesia on Twitter Using Pre-Trained and Self-Training Word Embeddings. *J. Ilmu Komput. Dan Inf.* **2022**, *15*, 39–46. [CrossRef]
120. Ponmani, K.; Thangaraj, M. Clustering Based Sentiment Analysis on Twitter Data for COVID-19 Vaccines in India. *Int. J. Health Sci.* **2022**, 4732–4748. [CrossRef]
121. Parameshwar Hegde, N.; Vikkurty, S.; Kandukuri, G.; Musunuru, S.; Hegde, G.P. Employee Sentiment Analysis towards Remote Work during COVID-19 Using Twitter Data. *Int. J. Intell. Eng. Syst.* **2022**, *15*, 75–84. [CrossRef]
122. Waheeb, S.A.; Khan, N.A.; Shang, X. Topic Modeling and Sentiment Analysis of Online Education in the COVID-19 Era Using Social Networks Based Datasets. *Electronics* **2022**, *11*, 715. [CrossRef]
123. Jyothsna; Rohini; Paulose, J. Sentiment Analysis on COVID-19 Related Social Distancing across the Globe Using Twitter Data. *ECS Trans.* **2022**, *107*, 3995–4001. [CrossRef]
124. Bahekar, K.B.; Gautam, P.; Sharma, S. Sentiment Analysis on Wearing Mask during COVID-19 Pandemic in India: A Case Study on Twitter. *ECS Trans.* **2022**, *107*, 7165–7178. [CrossRef]
125. Hikmah, K.; Fauzan, A.C.; Harliana, H. Sentiment Analysis of Vaccine Booster during Covid-19: Indonesian Netizen Perspective Based on Twitter Dataset. *J. Teknol. Komput. Dan Sist. Inf.* **2022**, *5*, 102–106. [CrossRef]
126. Standard Search API. Available online: https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/api-reference/get-search-tweets (accessed on 31 July 2022).
127. How to Use Advanced Search. Available online: https://help.twitter.com/en/using-twitter/twitter-advanced-search (accessed on 22 August 2022).
128. Mierswa, I.; Wurst, M.; Klinkenberg, R.; Scholz, M.; Euler, T. YALE: Rapid Prototyping for Complex Data Mining Tasks. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD '06, Philadelphia, PA, USA, 20–23 August 2006; ACM Press: New York, NY, USA, 2006.
129. RapidMiner GmbH Search Twitter. RapidMiner Documentation. Available online: https://docs.rapidminer.com/latest/studio/operators/data_access/applications/twitter/search_twitter.html (accessed on 31 July 2022).
130. Privacy Policy. Available online: https://twitter.com/en/privacy/previous/version_15 (accessed on 31 July 2022).
131. Developer Agreement and Policy. Available online: https://developer.twitter.com/en/developer-terms/agreement-and-policy (accessed on 31 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. Hutto, C.; Gilbert, E. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *ICWSM* **2014**, *8*, 216–225. [CrossRef]
134. Woeginger, G.J. Space and Time Complexity of Exact Algorithms: Some Open Problems. In *Parameterized and Exact Computation*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 281–290, ISBN 9783540230717.
135. Thakur, N. MonkeyPox2022Tweets: A Large-Scale Twitter Dataset on the 2022 MonkeyPox Out-break, Findings from Analysis of Tweets, and Open Research Questions, version 1. 2022. Available online: https://zenodo.org/record/6635559#.Y0X_J3bMLEY.
136. Thakur, N. MonkeyPox2022Tweets: The First Public Twitter Dataset on the 2022 MonkeyPox Outbreak. *Preprints* **2022**.
137. Hydrator: Turn Tweet IDs into Twitter JSON & CSV from Your Desktop. Available online: https://github.com/DocNow/hydrator (accessed on 31 July 2022).
138. Tekumalla, R.; Banda, J.M. Social Media Mining Toolkit (SMMT). *Genomics Inform.* **2020**, *18*, e16. [CrossRef]
139. Twarc: A Command Line Tool (and Python Library) for Archiving Twitter JSON. Available online: https://github.com/DocNow/twarc (accessed on 31 July 2022).
140. Hydrator. Available online: https://github.com/DocNow/hydrator/releases (accessed on 31 July 2022).
141. Warren, E. Strengthening Research through Data Sharing. *N. Engl. J. Med.* **2016**, *375*, 401–403. [CrossRef]
142. Fecher, B.; Friesike, S.; Hebing, M. What Drives Academic Data Sharing? *PLoS ONE* **2015**, *10*, e0118053. [CrossRef] [PubMed]
143. Logan, J.A.R.; Hart, S.A.; Schatschneider, C. Data Sharing in Education Science. *AERA Open* **2021**, *7*, 233285842110064. [CrossRef]

144. Thakur, N. MonkeyPox2022Tweets: MonkeyPox2022Tweets: A Large-Scale Twitter Dataset on the 2022 MonkeyPox Outbreak, Findings from Analysis of Tweets, and Open Research Questions, version 3. Available online: https://zenodo.org/record/6898178#.Y0YgcnbMLEY.
145. Supported Languages and Browsers. Available online: https://developer.twitter.com/en/docs/twitter-for-websites/supported-languages (accessed on 31 July 2022).
146. Twitter for IOS. Available online: https://apps.apple.com/us/app/twitter/id333903271 (accessed on 31 July 2022).
147. Twitter for Android. Available online: https://play.google.com/store/apps/details?id=com.twitter.android&hl=en_US&gl=US (accessed on 31 July 2022).
148. Twitter Website. Available online: https://twitter.com/home (accessed on 31 July 2022).
149. TweetDeck Website. Available online: https://tweetdeck.twitter.com/ (accessed on 22 August 2022).
150. He, L.; He, C.; Reynolds, T.L.; Bai, Q.; Huang, Y.; Li, C.; Zheng, K.; Chen, Y. Why Do People Oppose Mask Wearing? A Comprehensive Analysis of U.S. Tweets during the COVID-19 Pandemic. *J. Am. Med. Inform. Assoc.* **2021**, *28*, 1564–1573. [CrossRef] [PubMed]
151. Al-Ramahi, M.; Elnoshokaty, A.; El-Gayar, O.; Nasralah, 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] [PubMed]
152. Nagel, A.C.; Tsou, M.-H.; Spitzberg, B.H.; An, L.; Gawron, J.M.; Gupta, D.K.; Yang, J.-A.; Han, S.; Peddecord, K.M.; Lindsay, S.; et al. The Complex Relationship of Realspace Events and Messages in Cyberspace: Case Study of Influenza and Pertussis Using Tweets. *J. Med. Internet Res.* **2013**, *15*, e237. [CrossRef]
153. Zhang, K.; Song, W.; Liu, L.; Zhao, X.; Du, C. Bidirectional Long Short-Term Memory for Sentiment Analysis of Chinese Product Reviews. In Proceedings of the 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC), Beijing, China, 12–14 July 2019; pp. 1–4.
154. Deepa, M.D.; Al, E. Bidirectional Encoder Representations from Transformers (BERT) Language Model for Sentiment Analysis Task: Review. *Turk. J. Comput. Math. Educ.* **2021**, *12*, 1708–1721. [CrossRef]
155. Li, T.; Gu, J.-C.; Zhu, X.; Liu, Q.; Ling, Z.-H.; Su, Z.; Wei, S. DialBERT: A Hierarchical Pre-Trained Model for Conversation Disentanglement. *arXiv* **2020**, arXiv:2004.03760.
156. Kotsiantis, S.B.; Zaharakis, I.D.; Pintelas, P.E. Machine Learning: A Review of Classification and Combining Techniques. *Artif. Intell. Rev.* **2006**, *26*, 159–190. [CrossRef]
157. Li, W.; Serdyukov, P.; de Vries, A.P.; Eickhoff, C.; Larson, M. The Where in the Tweet. In Proceedings of the 20th ACM International Conference on Information and Knowledge Management—CIKM '11, Glasgow, UK, 24–28 October 2011; ACM Press: New York, NY, USA, 2011.
158. Liu, I.L.B.; Cheung, C.M.K.; Lee, M.K.O. Understanding Twitter Usage: What Drive People Continue to Tweet. In Proceedings of the PACIS 2010—14th Pacific Asia Conference on Information Systems, Taipei, Taiwan, 9–12 July 2010; pp. 928–939.
159. Cheng, Z.; Caverlee, J.; Lee, K. You Are Where You Tweet: A Content-Based Approach to Geo-Locating Twitter Users. In Proceedings of the 19th ACM international conference on Information and knowledge management—CIKM '10, Toronto, ON, Canada, 26–30 October 2010; ACM Press: New York, NY, USA, 2010.
160. Uysal, I.; Croft, W.B. User Oriented Tweet Ranking: A Filtering Approach to Microblogs. In Proceedings of the 20th ACM international conference on Information and knowledge management—CIKM '11, Glasgow, UK, 24–28 October 2011; ACM Press: New York, NY, USA, 2011.
161. Tao, K.; Abel, F.; Hauff, C.; Houben, G.-J. What Makes a Tweet Relevant for a Topic? Available online: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.309.8507&rep=rep1&type=pdf (accessed on 22 August 2022).
162. Pujazon-Zazik, M.; Park, M.J. To Tweet, or Not to Tweet: Gender Differences and Potential Positive and Negative Health Outcomes of Adolescents' Social Internet Use. *Am. J. Mens. Health* **2010**, *4*, 77–85. [CrossRef]
163. Merler, M.; Cao, L.; Smith, J.R. You Are What You Tweet . . . pic! Gender Prediction Based on Semantic Analysis of Social Media Images. In Proceedings of the 2015 IEEE International Conference on Multimedia and Expo (ICME), Turin, Italy, 29 June–3 July 2015; pp. 1–6.
164. Han, S.; Min, J.; Lee, H. Antecedents of Social Presence and Gratification of Social Connection Needs in SNS: A Study of Twitter Users and Their Mobile and Non-Mobile Usage. *Int. J. Inf. Manage.* **2015**, *35*, 459–471. [CrossRef]
165. Hu, T.; Xiao, H.; Nguyen, T.-V.T.; Luo, J. What the Language You Tweet Says about Your Occupation. *arXiv* **2017**, arXiv:1701.06233. [CrossRef]
166. Ito, J.; Song, J.; Toda, H.; Koike, Y.; Oyama, S. Assessment of Tweet Credibility with LDA Features. In Proceedings of the 24th International Conference on World Wide Web—WWW '15 Companion, Florence, Italy, 18–22 May 2015; ACM Press: New York, NY, USA, 2015.
167. Zhao, Q.; Erdogdu, M.A.; He, H.Y.; Rajaraman, A.; Leskovec, J. SEISMIC: A Self-Exciting Point Process Model for Predicting Tweet Popularity. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD '15, Sydney, Australia, 10–13 August 2015; ACM Press: New York, NY, USA, 2015.
168. Tonia, T.; Van Oyen, H.; Berger, A.; Schindler, C.; Künzli, N. If I Tweet Will You Cite? The Effect of Social Media Exposure of Articles on Downloads and Citations. *Int. J. Public Health* **2016**, *61*, 513–520. [CrossRef]
169. Haugh, B.R.; Watkins, B. Tag Me, Tweet Me If You Want to Reach Me: An Investigation into How Sports Fans Use Social Media. *Int. J. Sport Communication* **2016**, *9*, 278–293. [CrossRef]

170. Lim, W.Y.; Lee, M.L.; Hsu, W. IFACT: An Interactive Framework to Assess Claims from Tweets. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Singapore, 6–10 November 2017; ACM: New York, NY, USA, 2017.
171. Morabia, K.; Bhanu Murthy, N.L.; Malapati, A.; Samant, S. SEDTWIk: Segmentation-Based Event Detection from Tweets Using WIkipedia. In Proceedings of the 2019 Conference of the North, Minneapolis, Minnesota, 3–5 June 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 77–85.
172. Alowibdi, J.S.; Buy, U.A.; Yu, P.S.; Ghani, S.; Mokbel, M. Deception Detection in Twitter. *Soc. Netw. Anal. Min.* **2015**, *5*, 32. [CrossRef]
173. Classification and Ranking of Trending Topics in Twitter Using Tweets Text. *J. Crit. Rev.* **2020**, *7*, 895–899. [CrossRef]
174. Vemprala, N.; Akello, P.; Valecha, R.; Rao, H.R. An Exploratory Analysis of Alarming and Reassuring Messages in Twitterverse during the Coronavirus Epidemic. In Proceedings of the American Conference on Information Systems 2020, Salt Lake City, UT, USA, 10–14 August 2020.
175. Akpojivi, U. Euphoria and Delusion of Digital Activism: Case Study of #ZumaMustFall. In *Advances in Social Networking and Online Communities*; IGI Global: Hershey, PA, USA, 2018; pp. 179–202, ISBN 9781522528548.
176. Culotta, A.; Cutler, J. Mining Brand Perceptions from Twitter Social Networks. *Mark. Sci.* **2016**, *35*, 343–362. [CrossRef]
177. Zhou, L.; Wang, W.; Chen, K. Identifying Regrettable Messages from Tweets. In Proceedings of the 24th International Conference on World Wide Web—WWW '15 Companion, Florence, Italy, 18–22 May 2015; ACM Press: New York, NY, USA, 2015.
178. Yamamoto, Y.; Kumamoto, T.; Nadamoto, A. Followee Recommendation Based on Topic Extraction and Sentiment Analysis from Tweets. In Proceedings of the 17th International Conference on Information Integration and Web-based Applications & Services, Brussels, Belgium, 11–13 December 2015; ACM: New York, NY, USA, 2015.
179. Yan, J.L.S.; Kaziunas, E. What Is a Tweet Worth? Measuring the Value of Social Media for an Academic Institution. In Proceedings of the 2012 iConference on—iConference '12, Toronto, ON, Canada, 7–10 February 2012; ACM Press: New York, NY, USA, 2012.
180. Davis, S.W.; Horváth, C.; Gretry, A.; Belei, N. Say What? How the Interplay of Tweet Readability and Brand Hedonism Affects Consumer Engagement. *J. Bus. Res.* **2019**, *100*, 150–164. [CrossRef]