Journal of Computational Social Science (2024) 7:2457–2479
https://doi.org/10.1007/s42001-024-00311-5

RESEARCH ARTICLE

Comparing location-specific and location-open social media data: methodological lessons from a study of blaming of minorities on Twitter during the COVID-19 pandemic

Shiyi Zhang[1] · **Panayiota Tsatsou**[2] · **Lauren McLaren**[3] · **Yimei Zhu**[1]

Received: 7 December 2023 / Accepted: 8 July 2024 / Published online: 20 July 2024

Abstract

Social media platforms such as Twitter (currently X) have become important sites of public discourse and participation. Researchers have attempted to identify and collect Twitter data within a certain country or region in order to answer research questions within a particular locale. However, location information of tweets is limited. Tackling the case of public blaming of minorities on Twitter in the context of the COVID-19 pandemic in the UK, we present a method for identifying UK-based tweets and analyse two types of datasets that we collected and processed: (a) tweets with UK location-tags (labelled as *location-specific* data and referred to as *UK datasets*); and (b) tweets with UK location-tags and / or user profiles containing potential UK location information (labelled as *location-open* data and referred to as *ALL datasets*). The empirical results reveal that the overall sentiments in the two dataset types align in the same direction, but the location-specific datasets contain more extreme discourses (i.e., more positive and more negative sentiments and fewer neutral sentiments). Furthermore, in the location-specific datasets, the range of theme areas is narrower, although the themes still grasp the essence of the discussion about blaming minorities found in the larger dataset. The findings demonstrate strengths and limitations of the two dataset types and that the location-specific data can be suitable especially when the available research resources are insufficient for collecting or processing larger datasets. Nevertheless, we propose that future research may consider comparing smaller and bigger datasets to test differences between these for other topics for which specific locations may be of particular interest.

Keywords Location tracking · Computational thematic analysis · Sentiment analysis · Twitter analysis · Othering discourse · Blaming of minorities

Extended author information available on the last page of the article

Introduction

Social media was an important site of public discourse and participation during the COVID-19 pandemic. As the pandemic produced widespread feelings of threat, uncertainty and lack of control, ethnic and religious minorities in particular appeared to become the centre of critique and blame in the virtual space and especially on social media. As a result, researchers have attempted to identify and collect data from social media users based in specific geographic locations, aiming to answer questions about blame and hatred discourse online within certain countries or regions.

This paper draws upon the authors' study of online blaming of minorities in the UK during the major public health crisis of COVID-19 to test a methodological approach to the collection, selection and analysis of geographically specific online public discourse. Twitter-based discourse was selected for analysis, as Twitter (currently X) is the most prominent social media platform used for public discourse and one that, at the time of writing this paper, granted access to and analysis of secondary data.[1] Unlike Facebook, where users focus on relationship maintenance and emotional connection, Twitter is 'more able to reveal the structural inequality of pandemic experiences for racial communities' [1]. Relatedly, Chen et al. [2] have argued that the increasing usage of Twitter data for biopolitical research is due to Twitter data being multilevel and multifaceted and the fact that even plain text tweets can provide rich information such as the distribution of sentiments and themes, users' personality or other individual characteristics and information-sharing behaviours.

At the same time, studies on social media discourse analysis [3, 4] show that researchers struggle to provide a satisfactory method to collect Twitter data within a certain country or region because only 1–2% of tweets are estimated to have location tags [5]. Qazi et al. [6] suggest that the use of small-scale datasets may not adequately capture the breadth of public sentiment and discourse. Besides this tweet-level location information, which can vary in real time with users' movements, other types of location data include location mentioned in content of tweets and in user profiles. These are freeform values which might be fuzzy or invalid, posing challenges in accurately tracking location based on the values. This makes the methodological discussion of the location of Twitter data particularly important and timely. Researchers have proposed various approaches for identifying the location of tweets, but these methods have several challenges, such as significant time requirements, high computational demands, or low accuracy [3, 6–8]. Solving this methodological difficulty is substantively important because public discourse is likely to vary significantly by context. In the absence of accurate location information, the analysis of Twitter discourse becomes global in focus, and this is likely to mask important contextual differences in discourse.

[1] Twitter announced on 2nd February 2023 that free access to Twitter API was no longer allowed.

For our analysis, we selected the online discourse concerning blaming and racist or hate speech about the role and liability of ethnic and religious minorities[2] (i.e., BAME[3] communities) in the global public health crisis of COVID-19. This pandemic produced widespread feelings of threat, uncertainty and lack of control, prompting the blaming of groups believed to be liable, with ethnic and religious minorities being particularly targeted [11–13]. Indicative of this is that in the period of the pandemic, there was an increase of hate crimes directed at those perceived to be Chinese [14, 15]. The higher COVID-19-related mortality among ethnic minorities also meant these groups were likely to be blamed for the continued spread of the virus, while vaccine hesitancy [16] may have also contributed to further othering and blaming of victims for their own deaths [17].

Social media became a key site of discourse on blaming ethnic and religious minorities during COVID-19 [18]. In the UK, for instance—with over 57 million active social media users as of February 2022 (84 percent of the UK population) and social media penetration well above the global average [19]—there were prominent instances of blaming or shaming of ethnic and religious minorities for the spread of COVID-19 on social media. For instance, minorities were accused of ignoring social distancing rules and were said to still be allowed to attend religious services and festivals during national lockdowns [20].

To facilitate the study of Twitter and other sites of discourse in a particular geographical context—such as that related to minorities in times of COVID-19 in the UK—this paper presents and tests a method for Twitter data sampling, collection, selection and processes in order to capture location-based data. The paper reflects on the tested method through discussion of a series of findings derived from the comparative analysis of location-tagged (labelled as 'location-specific') and 'location-open' Twitter datasets and, hence, it draws methodological lessons from tackling the timely issue of blaming of ethnic and religious minorities in a particular locale during a public crisis.

The next section sets the research context of the study. Then, we present an overview of the design of the study and, after that, a detailed presentation of the method for collection of location-specific Twitter data. The penultimate section presents the empirical testing of the two dataset types, and, through presentation of a series of findings, it reflects on the methodological lessons derived from our findings. The paper closes with a series of conclusions and recommendations for the future study of online discourse of large numbers of users/members of the public (for instance, discourse in times of crisis) and within specific geographical contexts.

[2] 'Ethnic minorities' refers to all ethnic groups except the White British group [9] while 'religious minorities' refers to those identifying with any religion apart from the majority religion in the UK, Christianity [10].

[3] 'BAME' in the UK refers to 'Black, Asian and Minority Ethnic' [9].

Research background

Broadly speaking, studies of public opinion traditionally use quantitative surveys or qualitative interviews and focus groups to capture citizens' attitudes and experiences. In more recent years, the Internet and digital media platforms have opened up an opportunity for citizens to discuss social issues and develop civic culture [21]. Social media platforms such as Twitter and Facebook arguably provide the medium for online discourse of political, social and other public issues and thus constitute an online public sphere for citizens' democratic participation [22, 23]. Twitter, as a microblogging site, is often used by politicians and citizens to communicate and debate before election campaigns. Hence, more and more public opinion studies collect and analyse social media data which are available in the public domain. Based on a systematic literature review of 127 studies addressing the use of Twitter in election campaigns, 104 studies analysed digital trace data, while a small proportion collected data via surveys, experiments or interviews, thus confirming the importance of Twitter data in election, public opinion and public discourse analysis [24].

Twitter has also become an important site of public discourse during crises and major events, and this was especially the case during the COVID-19 pandemic. With the emergence and spread of COVID-19, the UK and many other countries instigated a lockdown policy whereby citizens were physically restricted to their homes and began to increasingly use social media platforms such as Twitter and TikTok to share information or experiences during the pandemic [25, 26]. Fear and anxiety are known as common responses to infectious outbreaks [27] and these emotions and the loneliness under lockdown had a negative impact on individuals' mental health [28]. Whilst Internet users sought support and socialised through social media in the emergence and spread of COVID-19, they were also exposed to misinformation and conspiracy on Twitter, Facebook and other platforms [29, 30].

Blaming ethnic and religious minorities during the pandemic also occurred over social media [31]. Croucher et al. [32] found that social media channels facilitated misinformation and biased attitudes towards Asian-Americans,[4] and that those consuming this information have, in turn, internalised it. In their study of social media discourses toward four ethnic communities in the United States—Asians, Blacks, Hispanics, and Native Americans—Lu and Liu [1] found that the Black community suffered structural inequality, racism, and discrimination and adopted an inward-oriented coping strategy during the pandemic, while other ethnic communities experienced a more implicit form of inequality and adopted an outward-oriented coping strategy that resulted in these communities developing coalitions with other communities to face discrimination during the pandemic. Looking into possible explanations, automated content analysis of the transcripts of 2,152 trending COVID-19-related YouTube videos in six countries conducted by Ng [34]—the United States, Brazil, Russia, Taiwan, Canada, and New Zealand—found that fear content vastly

[4] In the United States, 'Asian' is primarily used to refer to people of Far Eastern origin, while in the UK, 'Asian' often refers to people whose ancestry is from Pakistan, India, Bangladesh and Sri Lanka [33].

overshadowed efficacy content in all six countries and particularly in countries such as Russia and Brazil, where the government response to the pandemic was particularly inadequate.

As a result of these phenomena as well as the spread of information / misinformation online, a growing body of literature has examined public responses to COVID-19 on social media using computational social science methods [35]. For instance, Boon-Itt and Skunkan [36] studied public perceptions of the COVID-19 pandemic on Twitter using topic modelling and sentiment analysis, whilst Xie et al. [37] employed a similar approach to analyse public responses to COVID-19 on Weibo, a Chinese microblogging site. Lee et al. [38] studied public discourse on Twitter about COVID-19 in South Korea and Japan using Python for data collection and then word rank and frequency analysis, dividing the words into four categories: social distancing, prevention, issue, and emotion.

After reviewing a range of studies on social media public discourse analysis generally and related to the COVID-19 pandemic, it became evident that collecting tweets from a specific country or region to understand public discourse *within a particular geographic context* can be exceedingly difficult. An easy and commonly used approach is to collect data with location-tags [39, 40]. However, the availability of tweets containing location tags is limited, which raises questions about their representativeness of the wider discourse within a country or region. Our study presents and tests a data collection and selection method to identify larger—and potentially more representative—datasets from users in a specific location. Additionally, we examine whether smaller datasets, comprised of tweets with location-tags (labelled as 'location-specific' data), reflect or differ from the broader discourse found in the larger location-open datasets. Given that researchers no longer have free access to Twitter API for data collection, it is more practical for researchers to gather smaller volumes of data for their studies. This shift underscores the importance of determining whether a smaller dataset can effectively capture public discussion on specific issues and in defined contexts.

Focusing on the case of public blaming of minorities on Twitter in the context of the COVID-19 pandemic in the UK, this paper presents and tests a novel approach to collect and select what we label as 'location-open' data, namely tweets with user profiles containing potential UK location information in addition to tweets with UK-based location-tags. Since none of the previous studies of Twitter has applied a process for the comparison of location-tagged data with other approaches that capture much larger datasets of location-based data, this study reflects on and draws recommendations on how the parameter of 'location' can generate a different range and quality of findings from Twitter data with a case study of public discourse on ethnic and religious minorities during the pandemic in the UK. Based on sentiment analysis and computational thematic analysis, which have been demonstrated to yield a comprehensive understanding of public discourse [36, 37], we compare the smaller location-tagged data with the larger location-open data, employing an innovative approach to construct the latter dataset. Our comparison of the two datasets draws conclusions about the strengths and limitations of both dataset types, with implications for researchers studying online/social media data when research is geographically and context specific.

Hence, this paper explores the following research questions:

1. What are the main methodological steps and associated pitfalls in the collection, selection and comparative analysis of location-specific and location-open Twitter data?
2. What are the variations in the nuances of the findings regarding the themes and sentiments in Twitter discourses when comparing location-specific and location-open data, and what are the associated lessons for researchers?

In what follows, we overview the design of the study before presenting the method of location-open data collection and selection and the analytic comparison with the location-specific dataset, which address the research questions above.

Research design

Data sampling and collection

For the purpose of collecting tweets related to COVID-19, we drew upon recent studies on use of Twitter during COVID-19 generally [41] and use of Twitter to discuss minorities more specifically [20], and used the following keywords to retrieve relevant tweets:

#covid?19uk OR #covid?2019uk OR coronavirus OR covid OR #coronavirus OR #covid OR #covid19 OR #covid2019 OR #covid-19 OR #covid-2019 OR #covid_19 OR #covid_2019 OR #covid?19 OR #covid?2019 OR #coronavirusuk OR #coviduk OR #covid19uk OR #covid2019uk OR #covid-19uk OR #covid-2019uk OR #covid_19uk OR #covid_2019uk

The 'OR' logic was used in order to retrieve tweets that meet any of the criteria.

Another key parameter in the query was timestamp which was used to define the timeframe of the tweets to be collected. Through initial testing, we found that approximately 2 million tweets in each week matched the above keywords and it took approximately 15 h to collect the data for one random week of data within the entire period of the pandemic, indicating that the data volume during the whole pandemic would be huge and collecting all the data would be particularly time consuming. Thus, in our study, two one-week periods were selected for analysis, with the aim being to evaluate the ability of our methods to detect discourse about minorities. Our first one-week-period for analysis occurred when wider (non-Twitter) discourse was likely to include a strong focus on minorities; our second one-week-period occurred when wider (non-Twitter) discourse was less likely to include a strong focus on minorities.

The first one-week period selected was the seven-day period after Conservative Party MP Craig Whittaker stated that 'It is the BAME communities that are not taking this seriously enough' in an interview with LBC which was published on LBC's website on 31 July 2020, 8:33am [42] and widely publicised [43–45]. Prior to this

Table 1 Other parameters in search query

Parameters	Values
Language	English
Retweet	No
Reply	No
Location	UK/NA
Tweet fields	id, text, author_id, created_at, source, public_metrics
User fields	id, name, username, created_at, location, verified
Location fields	id, name, full_name, country_code

statement, Eid al-Fitr had been celebrated at home instead of communal prayers and gatherings due to COVID-19 restrictions. Consequently, BAME communities, particularly Muslims, were very likely to feel aggrieved by the statement. We were also interested in finding out whether non-BAME individuals would express agreement with Whittaker's statement. The data collection timeframe was set from 31 July 2020, 8:33am to 7 August 2020, 8:33am. July 2020 was also the month in which a record high number of racially or religiously aggravated offences (hate crimes) was recorded [46]. The expectation was that Whittaker's statement and its timing would generate a wealth of tweets/reactions in the social media realm of relevance to the purpose of the study. The second one-week time period selected was the end of the month that witnessed the lowest amount of hate crime (January 2021) and was a month for which we are unaware of any major events that would have been likely to trigger large amounts of Twitter discourse on minorities. The data collection timeframe for this one-week period was set from 31 January 2021, 00:00am to 7 February 2021, 00:00am. All in all, the tweets collected in the two weekly timeframes were those that allowed us to identify keywords, themes and sentiments in Twitter discourses that mentioned COVID-19, while they expectedly contained varying frequency and a range of tones in their references to ethnic and religious communities in the UK.

In the data collection process, we considered a few other parameters, such as the language of tweets (only tweets in English were collected), whether to collect retweets and replies of tweets or not (retweets and replies were excluded), location of the tweets (tweets with UK-based location tag and tweets without location tag were collected, see below), information about the tweets (i.e., tweet fields), author of the tweets (i.e., user fields), and location of the tweets (i.e., location fields) when the tweets had a location tag (see Table 1).

Data collection and pre-processing

Two queries were used to collect the data for each time period. The first was with the parameter 'location' with value "UK" in order to collect tweets with a UK-based location tag, and the second was without the parameter 'location' for the purpose of collecting all the eligible tweets regardless of whether the tweets have a location

tag or not. Python was used to automatically request and save all the collected data. After data collection and de-duplication, two datasets were created for each time period. For the time period related to Craig Whittaker's statement in July—August 2020, we collected 5775 UK-based tweets (labelled as UK-0 dataset) and 2,028,066 all eligible tweets (labelled as ALL-0 dataset). For the January—February 2021 period, we collected 5769 UK-based tweets (UK-0 dataset) and 1,446,506 all eligible tweets (ALL-0 dataset).

Transformer-based Natural Language Processing (NLP) is a state-of-the-art method of processing large amounts of natural language data. In the study, we planned to use NLP to conduct sentiment analysis and computational thematic analysis of the collected tweets. Computational thematic analysis combines computational topic modelling and qualitative thematic analysis, which can take the advantage of both methods [47]. However, NLP of tweets is challenging because of the variety of forms of information in Tweets such as URLs and emojis [48]. Therefore, data pre-processing was conducted with the purpose of facilitating the NLP. Two Python libraries were used to pre-process the collected Tweets. The first library is *emoji* which is used to convert emojis to emoji names (e.g., covert 👏 to 'clapping hands'), and the second one is *ekphrasis* [48] which is used to remove URL and e-mail, check and correct spelling errors, unpack hashtags and contractions, restore elongated words, and change the capital words to lowercase ones.

Method for identifying location of tweets

The aim of our original study was to understand UK public discourse surrounding minorities during the pandemic. The volume of UK-0 (location-tagged) datasets was much smaller than expected, raising concerns about the reliability of inferences about the public discourse, while ALL-0 (unprocessed location-open) datasets could not be used for understanding the public discourse in the UK due to containing data from all over the world. Thus, we devised a method to indirectly identify location of tweets (Fig. 1).

Specifically, we took consideration of previous studies and their methods of tracking tweets' locations. Several methods have been used in location tracking [49], with two of the most widely used methods analysing users' historical tweets and parsing location information in user's profile or mentioned in content using geo datasets or geocoding API. The first method has been used by Mahmud et al. [3], Lau et al. [7], Lamsal et al. [4], and Simanjuntak et al. [50]. These studies have achieved good performance in location tracking; however, the collection and analysis of historical tweets requires a considerable amount of time and computing power. It is indicative that the studies mentioned above analysed less than 1,000 users. Burton et al. [51], Ntompras et al. [8], Qazi et al. [6], and Nguyen et al. [52] have employed the second method, tracking location by matching location information in user profiles with other databases containing location information. However, the performance of this method in location tracking has not been satisfactory because the above-mentioned studies only matched location information in user profiles

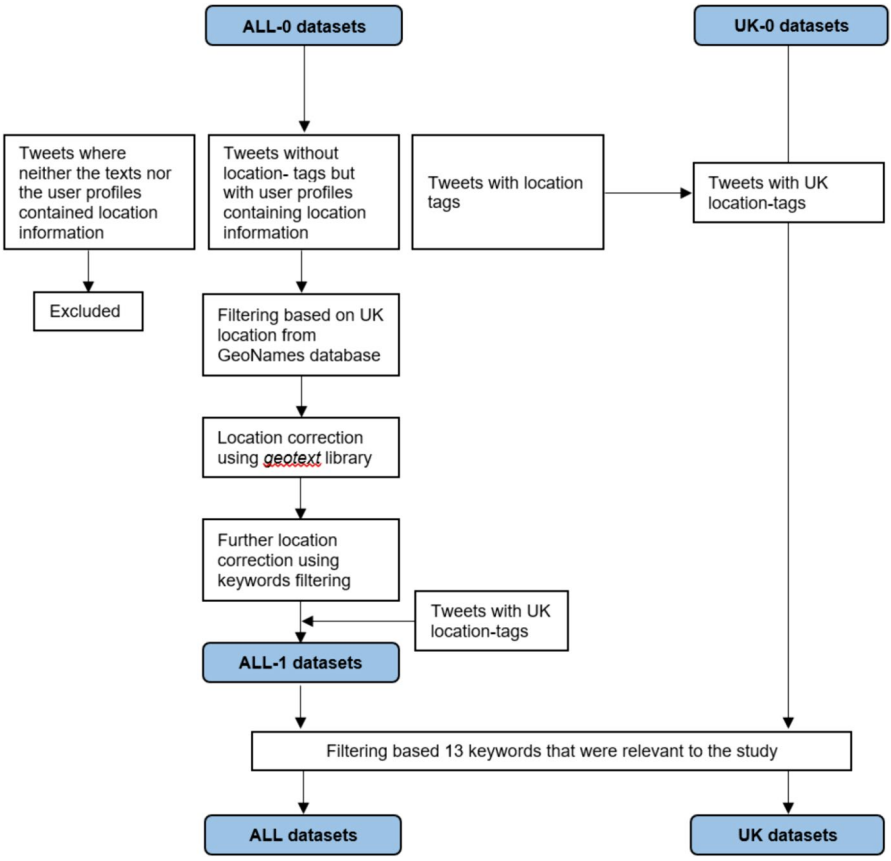

Fig. 1 Data processing steps

with other databases once, and some of the databases are not very comprehensive, which engendered several inaccuracies. Geocoding API, such as Google Maps API and OpenStreetMap Nominatim API, is also a commonly used approach in location tracking. It converts textual location data in user profiles or content into geographical coordinates, allowing automated processing of large amount of data efficiently and with promising precision [53–55]. However, this approach relies heavily on the accuracy and completeness of the location information provided by users, which may contain human errors of inaccurate geographic information [53, 54].

In our study, based on the location information in users' account profile in our ALL-0 datasets, we categorised the data into three types: (a) tweets without location-tag or information about location in user profiles, (b) tweets without location-tags but with user profiles containing location information, and (c) tweets with location-tags. The first group of tweets were excluded because no location information could

be used to identify where the tweets were tweeted from.[5] For the second group, the location information in user profiles was used as an indicator of the location of tweets. For the third group, tweets from the UK were selected directly by location-tag, which is the same as the data in our UK-0 datasets.

Given that the location information specified in the user profile is a freeform value which might be fuzzy or invalid, it was processed in order to filter out non-UK-based locations. More specifically, in this case, the location information was processed in three steps. In the first step, a database containing 69,771 location names in the UK, including administrative divisions (e.g., shires, cities), names of some institutions (e.g., universities, hospitals) and other types of locations, was found from GeoNames [56], which claims to 'cover all countries and contain over eleven million placenames'. This database also specifies which country (i.e., England, Wales, Scotland and Northern Ireland) each location belongs to. Based on this database, programming on Python was used to match the names of locations from the user profile in two[6] ALL-0 datasets (original location) with the names of locations from the GeoNames database. For each original location, if the match was successful, the name of location and which country this location belongs to would be returned. If the match was unsuccessful, the original locations would be split by comma or space because some of the original locations were too detailed to be successfully matched. For example, some original locations are 'London, England', while only 'London' or 'England' instead of 'London, England' exists in the GeoNames database. Therefore, unsuccessfully matched original locations were split and each part of the original location was matched with the name of locations from the GeoNames database. If the match was still unsuccessful, it meant that the original location does not belong to the UK, or the original location was not included in GeoNames database. This suggested that while aiming at an optimal degree of location matching, an absolute 100 percent matching was not possible. After this, two new datasets (for the 31 July—7 August 2020 and 31 January—7 February 2021 periods) containing all the successfully matched original locations were generated.

An additional problem with the above method was that some locations in other countries were categorised as belonging to the UK, such as 'New York'. The reason for this problem is that 'New York, USA' was split into 'USA' and 'New York', and 'New York' exists in GeoNames database for the UK. Thus, in a second step, a Python library, GeoText, which contain mainly major cities worldwide, was applied to filter out locations that are indeed in other countries.

For the original locations that were labelled as other countries, we found that some of them are in the UK but are categorised as other countries because the locations in the GeoText database are mainly major cities and while some

[5] Such tweets are expectedly excluded. The purpose of new location tracking methods is not to analyse / cover all the tweets, but to increase the proportion of tweets that can be location tracked to the maximum possible degree.

[6] Recall that we collected data during two one-week periods; these two datasets are analysed separately, given the very different contexts discussed above.

Table 2 Confusion matrix of two ways of location labelling

		Automatically labelled	
		UK	non-UK
Manually labelled	UK	42	20
	Non-UK	0	438
Accuracy = 0.96, Precision = 1.00, Recall = 0.67, F1-score = 0.81			

location names represent small cities in the UK, they represent major cities in other countries. For example, the original location 'Thornton, Cleveleys UK' was categorised as being in the US because 'Thornton' is the name of a village in the UK while it represents a small city in the US. In order to minimise such a problem, in a third step, the original locations that were categorised as other countries in the newly built datasets in the second step were filtered by the six keywords 'UK', 'United Kingdom', 'England', 'Wales', 'Scotland' and 'Northern Ireland'. If one original location was categorised as if it was located in other countries but contained one of these keywords, it was re-categorised as if located in the UK. After that, the tweets with UK location-tags, and the original locations labelled as 'UK', 'null', and other countries while containing one of the six above-mentioned keywords were merged, and the duplicates were removed.

After data screening, we ended up with 166,364 tweets (ALL-1 dataset) from the ALL-0 dataset (see Fig. 1) in the selected July—August 2020 week and 151,750 tweets (ALL-1 dataset) from the ALL-0 dataset in the selected January—February 2021 week. As a final step in the process, we tested the accuracy of location tracking. Specifically, we first randomly sampled 250 users from each of two ALL-1 datasets, and then manually labelled their locations. After that, a confusion matrix was used to compare the manually labelled and automatically labelled results (Table 2). Accuracy of the automatically labelled results was 0.96, indicating that the automatic approach of tracking Twitter location performed well. The confusion matrix also shows that some locations in the UK were labelled as a location in other countries, while none of the locations outside the UK was labelled as a location in the UK, indicating that all tweets included in our analysis were UK-based, despite the fact that some UK-based tweets were excluded.

Hence, the presented method achieved an accuracy of 96% in tracking Twitter location by: (a) tracking the location of Twitter users on the basis of the location information in their profile; (b) matching the location information with the names of locations from the GeoNames and GeoText databases successively, and (c) matching keywords ('UK', 'United Kingdom', 'England', 'Wales', 'Scotland' and 'Northern Ireland') to address the problems that arose when matching the location information with datasets. Based on these steps, we produced two ALL-1 datasets—which contain tweets with UK location-tags and tweets with user profiles providing potential UK location information—from the two ALL-0 datasets. In what follows, we empirically test the implications of our method for

location identification by analysing the ALL datasets (location-open data) and UK datasets and comparing the findings obtained from each of these.

Empirical testing of the method for location identification

We now turn to the question of whether identifying location matters for the insights obtained from social media data about public discourse on themes such as blaming minorities during a public crisis. That is, we assess the potential substantive value of the method of location identification presented above. We do this by analysing both location-specific and location-open datasets and identifying similar and different patterns of themes and sentiments in the analysed data. The purpose is to show whether the data sampling, collection and selection method for identifying the location of the tweets has resulted in the collection of two very different datasets or not and also to reach lessons for future researchers with limited resources regarding: (a) the strengths and limitations of using smaller datasets; and (b) the importance of identifying the location of online/social media data when research has a clearly defined geographical scope.

Techniques and processes

To conduct the data analysis for assessing the extent to which identification of location matters, we chose 13 keywords that were relevant to the study's focus on ethnic and religious minorities in the UK: *race*, *white*, *muslim*, *bame*, *eid*, *asian*, *leicester*, *bradford*, *china*, *chinese*, *india*, *africa*, *ethnic*. Based on this list of keywords, the following number of tweets from the ALL-1 and UK-0 datasets were selected for analysis: 241 in the UK dataset and 5763 in the ALL dataset for the July—August 2020 week; 125 in the UK dataset and 3,696 in the ALL dataset for the January—February 2021 week. The tweets in the UK datasets include location tags and are also part of the ALL datasets, and users of all the tweets have location information in their profiles. Therefore, about 3.73% of tweets have location tags in this study. Unlike surveys, which can ask the respondents about their demographic information, it is challenging to ascertain users' backgrounds using social media data, as user profiles are unlikely to contain information such as income which are often used to determine users' economic status in survey studies. Moreover, users may not be willing to disclose their authentic demographic details on their profiles and some social media accounts are bots or represent organisations rather than individuals. Consequently, it becomes difficult to determine who is more likely to use location tags in their tweets. This requires further investigation in future studies.

Transformer, the state-of-the-art machine learning model that is rarely used in social science, was used to conduct computational thematic analysis and sentiment analysis where appropriate (see below). As for computational thematic analysis, a sentence transformer model, Contextulized Topic Models (CTM) that uses pre-trained representations of language, was used to conduct computational topic modelling to identify N topics in the collected tweets [57].

Comparing with Latent Dirichlet allocation (LDA), the widely used method for topic modelling that takes input documents as Bag-of-Word and may neglect the inherent syntactic and semantic relations among the words in a document [58], CTM has proven to generate more coherent topics [57]. CTM was firstly used to generate different numbers of topics (*N* value) in both UK datasets and the number of topics with the highest coherence score was selected as the best *N* value. For the UK dataset in the July—August 2020 week, the best *N* value was 20 topics, and the best *N* value for the UK dataset in the January—February 2021 week was 10 topics. After that, CTM was used to generate 20 topics in the UK and ALL datasets (see Fig. 1) in the July—August 2020 week, and 10 topics in the UK and ALL datasets in the January—February 2021 week. For each topic, CTM returned the top 15 words that could best describe the topic. In order to better understand the theme of each topic, we retrieved the top 30 tweets that were most likely assigned to each topic based on the possibilities given by CTM. Each tweet was assigned to a topic based on the results of topic modelling, but the same tweet could appear in multiple topics. Subsequently, qualitative thematic analysis was conducted following Braun and Clarke's [59] guidelines and involving collaborative discussion among the four authors. The computationally derived topics and tweets were reviewed for the purpose of identifying emerging patterns, and the themes were generated in this process. Then, the themes were reviewed to investigate whether they were distinct from each other and whether they reflected all tweets included in the dataset. This resulted in the identification of 5 themes in the UK dataset and 14 themes in the ALL dataset for the July—August 2020 week, as well as 3 themes in the UK dataset and 8 themes in the ALL dataset for the January—February week.

As for sentiment analysis, pysentimiento [60], a Transformer-based model, was used, as computer science researchers have proved that this model achieves a good performance in sentiment analysis [61]. Previous studies conducted sentiment analysis using human coders or relying on sentiment dictionaries [62–64], but these techniques are time-consuming and may bias results due to subjectivity and the lack of context. The Transformer-based models have been proved to perform much better in sentiment analysis than such previous methods [65]. For each tweet, pysentimiento returned the possibility of the tweet belonging to different sentiments (i.e., positive, neutral, negative), and the sentiment with the highest probability was chosen as the sentiment of that tweet. Positive was coded as 1, neutral was coded as 0, and negative was coded as -1, and the sentiment of the four datasets was calculated based on the sentiment of each tweet in the respective dataset. In addition, the sentiment of each keyword in the four datasets was calculated based on the sentiment of the tweets containing the keyword, and the sentiment of each theme in the four datasets was calculated based on the sentiment of the tweets belonging to each theme.

What follows reports a series of findings obtained from our datasets as well as the commonalities and discrepancies from the datasets.

Table 3 Sentiments in ALL and UK datasets

	31 July—7 August 2020		31 January—7 February 2021	
	UK	ALL	UK	ALL
Positive	0.22	0.15	0.18	0.15
Neutral	0.3	0.44	0.48	0.54
Negative	0.48	0.41	0.34	0.31
Sentiment	− 0.26	− 0.3	− 0.18	− 0.21

The "Sentiment" row was calculated based on the sentiments of each tweet in the respective dataset

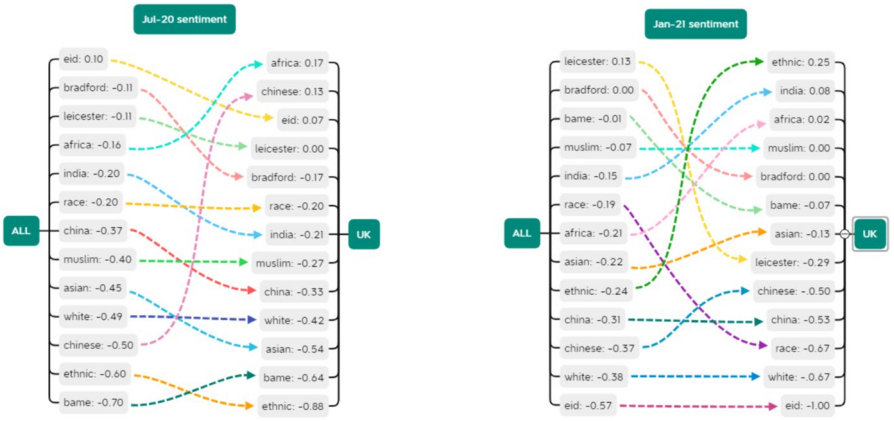

Fig. 2 Keyword sentiments

Findings

Sentiments in the datasets

Table 3 shows the overall sentiments in all four datasets and the possibility of each dataset belonging to different sentiments (i.e., sentiment distribution). Overall, sentiments in both the UK and ALL datasets each week align in the same direction, consistently showing negative sentiments (Table 3). Moreover, the sentiments in the UK datasets are slightly more positive than the corresponding ALL datasets. The sentiment distribution in the four datasets shows that the UK datasets contained slightly more positive and more negative sentiments compared to the corresponding ALL datasets, where the proportion of neutral sentiment discourses was higher than in the corresponding UK datasets.

Regarding sentiments of keywords[7] in the datasets, Fig. 2 shows that, for the July—August 2020 week, similar sentiments were observed for a range of keywords in both the UK and ALL datasets, while sentiments for specific ethnic groups and locations, especially for Africa, Leicester and Chinese, were more negative in the ALL dataset than in the UK dataset. For the January—February 2021 week, Fig. 2 shows that sentiments for most keywords are negative in both the UK and ALL datasets, while more negative sentiments against specific ethnic groups exist in the ALL dataset than in the UK dataset, with the only exception being the keywords China and Chinese. These results indicate that despite keyword frequency and sentiments sometimes being similar, in general, it might not be realistic to presume sentiment characteristics of particular keywords in location-specific datasets on the basis of those in larger, location-unfiltered datasets, and vice versa.

Themes in the datasets

Moving on to themes in the analysed data, for the July—August 2020 week, the computational thematic analysis identified 14 distinct themes within the ALL dataset and 5 within the UK dataset (Table 4). For the January—February 2021 week, the computational thematic analysis revealed 8 themes in the ALL dataset and 3 in the UK dataset (Table 4). Based on a qualitative assessment of the content and underlying semantics of each theme, the common themes in the UK and ALL datasets were identified (subset of Table 4, common themes).

The UK datasets in the two examined weeks appeared to contain less coherent content than the content of tweets in the corresponding ALL datasets. This aligns with the finding that there are not many repeated or heavily weighted tweets in the UK datasets within each theme area. Hence, the UK datasets exhibit less clarity and themes were more difficult to identify. In other words, the ALL datasets performed better in computational thematic analysis as the themes identified from the ALL datasets are more coherent and easier to understand.

At the same time, the range of theme areas contained in the UK datasets is narrower than that in the corresponding ALL datasets. Some of the themes identified in the ALL datasets were not found in the UK datasets. For example, we observed themes such as ‘Support for BAME communities during COVID-19 pandemic’ and ‘New lockdown in Northern England’ only in the ALL dataset of July—August 2020 week, and themes such as ‘Investigation in Wuhan’ and ‘Impacts of COVID-19 on India’ only in the ALL dataset of 31 January—7 February 2021 week (see Table 4). While both the UK and ALL datasets contained themes related to the UK and BAME communities in the UK, which was the actual focus of the study, the ALL datasets also contained other themes, with a global span, as well as themes other than BAME communities. These differences show that, contrary to smaller location-tagged UK datasets, the ALL datasets capture a global rather than a

[7] Negative sentiments of the keywords do not necessarily indicate negative sentiments towards minorities. Many tweets with negative sentiments aim to convey dissatisfaction with Craig Whittaker’s statement or the policies of the UK government in managing COVID-19.

Table 4 Similarities and differences in ALL and UK datasets

	31 July—7 August 2020	31 January—7 February 2021
Common themes in ALL and UK datasets	1. Celebrate Eid and keep safe	1. Call for BAME to get vaccinated
	2. BAME communities should not be blamed for the spread of COVID-19	2. More testing to be introduced in response to new variant
	3. Blame China	3. Discontent toward China
	4. BAME communities are more vulnerable to COVID-19	
	5. BAME communities affected disproportionally by COVID-19	
Themes only in ALL datasets	1. COVID-19 in India	1. BAME communities are more vulnerable to COVID-19
	2. COVID-19 in South Africa	2. New South Africa variant
	3. Support for BAME communities during COVID-19 pandemic	3. Investigation in Wuhan
	4. F1 contestants tested COVID-19 Positive	4. COVID-19 cases globally
	5. COVID-19 vaccine trial did not represent BAME well	5. Impacts of COVID-19 in India
	6. Impact of COVID-19 in India Economy	
	7. Antipathy to China surged	
	8. New lockdown in Northern England	
	9. COVID-19 outbreak is a warning to complacent white middle class	

UK-centric perspective, with some of the themes being less likely to be directly relevant to the context or region of focus.

Discussion and conclusion

This paper contributes to methodological understanding of Twitter data sampling, collection and selection in order to determine the location of online discourse without a location tag.

In our study, about 3.73% tweets were location-tagged. While this is higher than the 1–2% indicated by Twitter [5] and 0.072% by [6], the proportion remains notably low. Therefore, it is important that researchers use additional means to determine the likely location of tweets and compare with location-open tweets. This will inform other researchers with limited financial and human resources and/or capacity on: (a) the degree and importance but also traps in identifying the location of tweets when the research is region-specific and with a clearly defined geographical scope; (b) the similarities and differences in the results obtained in the comparative analysis of location-specific and location-open tweets for the purpose of tackling timely issues and (c) the strengths and limitations of using smaller and larger datasets, and the related selection of different types of datasets for analysis.

More specifically, this paper has presented a method for location-tracking of Twitter data, which can also be extrapolated to other digital platforms where location information is embedded within user profiles. Furthermore, this method can be instrumental in research oriented towards other countries because the databases used in this study (i.e., GeoNames and GeoText) encompass location information worldwide. Through empirical testing and by analysing sentiments and themes of two types of datasets, this paper also sheds light on understanding the strengths and limitations of both smaller and larger datasets on the basis of whether the tweets are location-tagged or not, hence providing insights into the selection of different types of datasets for analysis. This can be particularly useful in the context of the recently adopted X policy that does not allow free use of Twitter API.

Twitter (currently X) is used worldwide. However, it is difficult to identify tweets from a specific country or region for the purpose of research data collection. This has driven researchers to propose methodologies for collection and analysis of tweets from a specific location [3, 7, 8, 50, 51]. However, these approaches present various limitations, including being exceedingly time-consuming, costing considerable computing power, and low accuracy. Therefore, this paper presents and tests a methodology for location tracking in research of online discourses concerning a range of themes and discourses, such as discourses that have been generated by a large part of the public and are against or in favour of specific groups or communities in times of public crisis.

Thus, regarding the first question of this paper—What are the main methodological steps and associated pitfalls in the collection, selection and comparative analysis of location-specific and location-open Twitter data?—the method presented suggests the crucial role of screening and selection of tweets via suitable databases and machine learning tools as well as specific steps and

processes in the identification of location so as to be able to extract tweets from specific locales. At the same time, this method sheds light on some problems and lessons for future research. Specifically, because users' profiles are free-form values, it proved impossible to accurately match location information in users' profile to a city or country, even when three-step matching was used in the study. Nevertheless, for a large volume of data, this is still a workable way to identify locations of tweets when the research is region-specific and when tweets with a location tag are limited. Compared to previous studies analysing users' historical tweets and tweeting behaviours [3, 4], our method takes much less time and less computing power when analysing a large volume of data. Based on our confusion matrix, the precision of our method is 1.00, suggesting that the unneeded data in this study (i.e., non-UK-based tweets) was all excluded from data analysis, and the recall is 0.67, indicating that about one third of UK-based tweets were excluded. This could potentially have an impact on the results of our study, as it means that we may not have captured all the relevant data. While at least the unneeded data was excluded, the increase in the volume of data using our proposed processing techniques may increase its representativeness and coherence and help to compensate for the exclusion of needed data. Having said that, researchers must still be cognizant of the existence of potential traps (e.g., cities in different countries with the same name) and the lack of a method for ensuring 100 percent accuracy in the selection of location-specific data.

For empirically testing the method, we processed and compared tweets with location tagging with a larger set of tweets that included both location tagging and user profiles with a UK location, and reviewed the related insights and findings. On a general note, we support Chen et al. [2] in noting the increasing usage of Twitter data for biopolitical research due to Twitter data being multilevel and multifaceted and the fact that even plain tweet text can provide rich information, such as the distribution of sentiments and themes, personality or other individual characteristics, and information-sharing behaviours. On a more specific note, the results we obtained demonstrated the meaningfulness of comparison between location-specific and location-open Twitter data for the study of biopolitical phenomena, such as public feelings and discourses about minorities during a public health crisis.

Answering the second research question—What are the variations in the nuances of the findings regarding the themes and sentiments in Twitter discourses when comparing location-specific and location-open data, and what are the associated lessons for researchers?—the comparison between the two dataset types found that while keywords and sentiments in Twitter discourses were similar, they were mostly negative, with the location-tagged-only data containing more extreme discourses (i.e., more positive and negative and fewer neutral sentiments) than the larger datasets, as described above. As for the themes, the most discussed themes in both dataset types are BAME-related, such as calling on BAME communities to keep safe and get vaccinated, and the argument that BAME should not be blamed. Although common themes in both dataset types can only represent about one third of all themes, our analysis revealed that, while the location-specific datasets contain more narrow and less coherent themes, these still grasp the essence of the discussion about blaming minorities and may be utilised especially when resources are insufficient

for collecting, selecting or processing larger datasets. This is a potentially valuable insight for other scholars, given Twitter/X's revocation of free access to its archive of tweets and the related restriction for researchers to only access a small corpus of tweets due to financial and time resource limitations.

The findings of this study show that (a) overall sentiments in the two types of datasets align in the same direction, but the smaller datasets contain more extreme discourses, (b) the smaller datasets can grasp the essence of the discussion about blaming minorities, but contain a narrower and less coherent range of theme areas and (c) the larger datasets have a more coherent but broader focus, with some of their themes being likely to be less relevant to the focus of a geographically or other context specific study. The empirical testing in our study has demonstrated certain strengths and limitations of the two methods, and researchers can select either of these methods on the basis of an evaluation of the respective strengths and limitations vis-à-vis the needs of their study. Thus, based on our results, if a researcher aims to gauge sentiments without over-representation of extreme discourses and also needs to identify a wide range of coherent themes/topics, then a larger dataset would be preferable. On the other hand, if the aim is to locate and analyse extreme discourses, among others, as well as to identify a narrow set of topics/themes that are more likely to be context-relevant, then the smaller dataset is preferable.

Nevertheless, it is important to note that further empirical testing of the presented method is needed, as the results of our study might not be applicable to all other research due to the differences in the research topics and types of data. We thus propose that future research concerning specific locations or contexts also tests our proposition and compares smaller with larger datasets.

Funding Funding for this research was provided by the University of Leicester College of Social Sciences Research Development Fund.

Data availability The data used in this study is secondary data from Twitter. The data cannot be shared openly in order to protect the privacy of Twitter users.

Declarations

Conflict of interest The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

1. Lu, J., & Liu, J. (2023). Communicating concerns, emotional expressions, and disparities on ethnic communities on social media during the COVID-19 pandemic: a structural topic modeling approach. *American Behavioral Scientist*. https://doi.org/10.1177/00027642231164046
2. Chen, K., Duan, Z., & Yang, S. (2022). Twitter as research data: Tools, costs, skill sets, and lessons learned. *Politics and the Life Sciences, 41*, 114–130.
3. Mahmud, J., Nichols, J., & Drews, C. (2014). Home location identification of twitter users. *ACM Transactions on Intelligent Systems and Technology (TIST), 5*, 1–21.
4. Lamsal, R., Harwood, A., & Read, M.R. (2022). Where did you tweet from? Inferring the origin locations of tweets based on contextual information, *arXiv Preprint* https://arXiv.org/2211.16506.
5. Twitter. (2021). Advanced filtering with geo data.
6. Qazi, U., Imran, M., & Ofli, F. (2020). GeoCoV19: A dataset of hundreds of millions of multilingual COVID-19 tweets with location information. *SIGSPATIAL Special, 12*, 6–15.
7. Lau, J.H., Chi, L., Tran, K.-N., & Cohn, T. (2017). End-to-end network for twitter geolocation prediction and hashing, arXiv Preprint https://arXiv.org/1710.04802.
8. Ntompras, C., Drosatos, G., & Kaldoudi, E. (2021). A high-resolution temporal and geospatial content analysis of Twitter posts related to the COVID-19 pandemic. *Journal of Computational Social Science, 5*, 1–43.
9. GOV.UK. (2020). Writing about ethnicity.
10. ONS. (2020). Exploring religion in England and Wales: February 2020.
11. Dionne, K. Y., & Turkmen, F. F. (2020). The politics of pandemic othering: Putting COVID-19 in global and historical context. *International Organization, 74*, E213–E230.
12. Esses, V. M., & Hamilton, L. K. (2021). Xenophobia and anti-immigrant attitudes in the time of COVID-19. *Group Processes & Intergroup Relations, 24*, 253–259.
13. Xun, Z., & Gilman, S. L. (2021). *'I Know Who Caused COVID-19': Pandemics and Xenophobia*. Reaktion Books.
14. Gray, C., & Hansen, K. (2021). Did COVID-19 lead to an increase in hate crimes toward Chinese people in London? *Journal of Contemporary Criminal Justice, 37*, 569–588.
15. Tessler, H., Choi, M., & Kao, G. (2020). The anxiety of being Asian American: Hate crimes and negative biases during the COVID-19 pandemic. *American Journal of Criminal Justice, 45*, 636–646.
16. Kadambari, S., & Vanderslott, S. (2021). Lessons about COVID-19 vaccine hesitancy among minority ethnic people in the UK. *The Lancet Infectious Diseases, 21*, 1204–1206.
17. Dukes, K. N., & Gaither, S. E. (2017). Black racial stereotypes and victim blaming: Implications for media coverage and criminal proceedings in cases of police violence against racial and ethnic minorities. *Journal of Social Issues, 73*, 789–807.
18. Costello, M., Cheng, L., Luo, F., Hu, H., Liao, S., Vishwamitra, N., Li, M., & Okpala, E. (2021). COVID-19: a pandemic of anti-Asian cyberhate. *Journal of Hate Studies, 17*, 108–118.
19. Dixon, S. (2022). Active social media audience in the United Kingdom (UK) in February 2022.
20. Awan, I., Carter, P., Sutch, H., & Lally, H. (2021). COVID-19: What are the drivers of the islamophobic infodemic communications on social media?
21. Klinenberg, E. (2005). Convergence: News production in a digital age. *The Annals of the American Academy of Political and Social Science, 597*, 48–64.
22. Haro-de-Rosario, A., Sáez-Martín, A., & del Carmen-Caba-Pérez, M. (2018). Using social media to enhance citizen engagement with local government: Twitter or Facebook? *New Media & Society, 20*, 29–49.
23. Bisgin, N., Bisgin, H., Hummel, D., Zelner, J., & Needham, B. L. (2023). Did the public attribute the flint water crisis to racism as it was happening? Text analysis of twitter data to examine causal attributions to racism during a public health crisis. *Journal of Computational Social Science, 6*, 165–190.
24. Jungherr, A. (2016). Twitter use in election campaigns: A systematic literature review. *Journal of Information Technology & Politics, 13*, 72–91.
25. Li, Y., Guan, M., Hammond, P., & Berrey, L. E. (2021). Communicating COVID-19 information on TikTok: A content analysis of TikTok videos from official accounts featured in the COVID-19 information hub. *Health Education Research, 36*, 261–271.

26. Singh, L., Bansal, S., Bode, L., Budak, C., Chi, G., Kawintiranon, K., Padden, C., Vanarsdall, R., Vraga, E., & Wang, Y. (2020). A first look at COVID-19 information and misinformation sharing on Twitter, *arXiv Preprint* https://arXiv.org/2003.13907.
27. Usher, K., Durkin, J., & Bhullar, N. (2020). The COVID-19 pandemic and mental health impacts. *International Journal of Mental Health Nursing, 29*, 315.
28. Al-Dwaikat, T. N., Aldalaykeh, M., & Rababa, M. (2020). The relationship between social networking sites usage and psychological distress among undergraduate students during COVID-19 lockdown. *Heliyon, 6*, e05695.
29. Gruzd, A., & Mai, P. (2020). Going viral: How a single tweet spawned a COVID-19 conspiracy theory on Twitter. *Big Data & Society, 7*, 2053951720938405.
30. Saud, M., Mashud, M., & Ida, R. (2020). Usage of social media during the pandemic: Seeking support and awareness about COVID-19 through social media platforms. *Journal of Public Affairs, 20*, e2417.
31. Anand, D., & Hsu, L. (2020). COVID-19 and black lives matter: Examining anti-Asian racism and anti-Blackness in US education. *International Journal of Multidisciplinary Perspectives in Higher Education, 5*, 190–199.
32. Croucher, S. M., Nguyen, T., & Rahmani, D. (2020). Prejudice toward Asian Americans in the COVID-19 pandemic: The effects of social media use in the United States. *Frontiers in Communication, 5*, 39.
33. Bhopal, R. (2004). Glossary of terms relating to ethnicity and race: For reflection and debate. *Journal of Epidemiology and Community Health, 58*, 441–445. https://doi.org/10.1136/jech.2003.013466
34. Ng, Y. M. M. (2022). A cross-national study of fear appeal messages in YouTube trending videos about COVID-19. *American Behavioral Scientist*. https://doi.org/10.1177/00027642231155363
35. Charquero-Ballester, M., Walter, J. G., Nissen, I. A., & Bechmann, A. (2021). Different types of COVID-19 misinformation have different emotional valence on Twitter. *Big Data & Society, 8*, 20539517211041280.
36. Boon-Itt, S., & Skunkan, Y. (2020). Public perception of the COVID-19 pandemic on Twitter: Sentiment analysis and topic modeling study. *JMIR Public Health and Surveillance, 6*, e21978.
37. Xie, R., Chu, S. K. W., Chiu, D. K. W., & Wang, Y. (2021). Exploring public response to COVID-19 on Weibo with LDA topic modeling and sentiment analysis. *Data and Information Management, 5*, 86–99.
38. Lee, H., Noh, E. B., Choi, S. H., Zhao, B., & Nam, E. W. (2020). Determining public opinion of the COVID-19 pandemic in South Korea and Japan: Social network mining on twitter. *Healthcare Informatics Research, 26*, 335–343.
39. Yang, J.-A., Tsou, M.-H., Jung, C.-T., Allen, C., Spitzberg, B. H., Gawron, J. M., & Han, S.-Y. (2016). Social media analytics and research testbed (SMART): Exploring spatiotemporal patterns of human dynamics with geo-targeted social media messages. *Big Data & Society, 3*, 2053951716652914.
40. Arthur, R., & Williams, H. T. (2019). Scaling laws in geo-located Twitter data. *PLoS ONE, 14*, e0218454.
41. Kehoe, A. (2021).TRAC:COVID—Trust and Communication: A Coronavirus Online Visual Dashboard.
42. Sherling, A. (2020). Muslim and BAME communities not taking coronavirus pandemic seriously, Tory MP says.
43. BBC. (2020). Craig Whittaker: MP defends saying some Muslims not taking covid seriously,
44. Braddick, I. (2020). Tory MP Craig Whittaker claims Muslims and BAME community "not taking pandemic seriously".
45. Walker, P., Halliday, J. (2020). Tory MP condemned for claim BAME people breaching lockdown most.
46. Home Office. (2021). *Hate Crime, England and Wales, 2020/21*: Appendix Table 6, Appendix Table 6.
47. Delgosha, M. S., Hajiheydari, N., & Talafidaryani, M. (2022). Discovering IoT implications in business and management: A computational thematic analysis. *Technovation, 118*, 102236.
48. Baziotis, C., Pelekis, N., & Doulkeridis, C. (2017). Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis, *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pp 747–754.

49. Qarabash, N. A., & Qarabash, H. A. (2018). Twitter location-based data: Evaluating the methods of data collection provided By Twitter Api. *International Journal of Computing, 19*, 583–589.
50. Simanjuntak, L. F., Mahendra, R., & Yulianti, E. (2022). We know you are living in Bali: Location prediction of twitter users using BERT language model. *Big Data and Cognitive Computing, 6*, 77.
51. Burton, S. H., Tanner, K. W., Giraud-Carrier, C. G., West, J. H., & Barnes, M. D. (2012). "Right time, right place" health communication on Twitter: Value and accuracy of location information. *Journal of Medical Internet Research, 14*, e2121.
52. Nguyen, H. L., Tsolak, D., Karmann, A., Knauff, S., & Kühne, S. (2022). Efficient and reliable geocoding of German Twitter data to enable spatial data linkage to official statistics and other data sources. *Frontiers in Sociology, 7*, 910111.
53. Hoffmann, M., & Heft, A. (2020). Here, there and everywhere": Classifying location information in social media data–Possibilities and limitations. *Communication Methods and Measures, 14*(3), 184–203.
54. Stock, K. (2018). Mining location from social media: A systematic review. *Computers, Environment and Urban Systems, 71*, 209–240.
55. Ardon, S., Bagchi, A., Mahanti, A., Ruhela, A., Seth, A., Tripathy, R. M., & Triukose, S. (2011). Spatio-temporal analysis of topic popularity in twitter, *arXiv Preprint* https://arXiv.org/1111.2904.
56. GeoNames. (2022). GeoNames.
57. Bianchi, F., Terragni, S., Hovy, D. (2020). Pre-training is a hot topic: Contextualized document embeddings improve topic coherence, *arXiv Preprint* https://arXiv.org/2004.03974.
58. Wang, X., & Grimson, E. (2007). Spatial latent dirichlet allocation, *Advances in Neural Information Processing Systems* 20.
59. Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology, 3*(2), 77–101.
60. Pérez, J. M., Giudici, J. C., & Luque, F. (2021). pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks, *arXiv Preprint* https://arXiv.org/2106.09462.
61. Bayram, U., Benhiba, L. Emotionally-informed models for detecting moments of change and suicide risk levels in longitudinal social media data, *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pp. 219–225.
62. Xu, Z., & Guo, H. (2018). Using text mining to compare online pro-and anti-vaccine headlines: Word usage, sentiments, and online popularity. *Communication Studies, 69*, 103–122.
63. Gerts, D., Shelley, C. D., Parikh, N., Pitts, T., Ross, C. W., Fairchild, G., Chavez, N. Y. V., & Daughton, A. R. (2021). "Thought I'd Share First" and other conspiracy theory tweets from the COVID-19 infodemic: Exploratory study. *JMIR Public Health and Surveillance, 7*, e26527.
64. Thakur, N. (2023). Sentiment analysis and text analysis of the public discourse on Twitter about COVID-19 and MPox. *Big Data and Cognitive Computing, 7*, 116.
65. Gillioz, A., Casas, J., Mugellini, E., & Abou Khaled, O. (2020). Overview of the Transformer-based Models for NLP Tasks. In: *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, pp. 179–183.

Authors and Affiliations

Shiyi Zhang[1] · Panayiota Tsatsou[2] · Lauren McLaren[3] · Yimei Zhu[1]

✉ Yimei Zhu
yz411@le.ac.uk
https://scholar.google.com/citations?user=mrFjwxEAAAAJ&hl=en

Shiyi Zhang
https://scholar.google.com/citations?user=bCxipZUAAAAJ&hl=en

Panayiota Tsatsou
https://scholar.google.com/citations?user=RTo2-iEAAAAJ&hl=en

Lauren McLaren
https://scholar.google.co.uk/citations?user=y00Q_2UAAAAJ&hl=en

[1] School of Arts, Media and Communication, University of Leicester, Leicester LE1 7RH, UK

[2] Sir Lenny Henry Centre for Media Diversity Centre & College of English and Media, Birmingham City University, Birmingham, UK

[3] School of History, Politics and International Relations, University of Leicester, Leicester, UK