2021

Analysis of Geotagging Behavior: Do Geotagged Users Represent the Twitter Population?

Amir Karami
University of South Carolina, karami@mailbox.sc.edu

R. R. Kadari

L. Panati

H. Bheemreddy

B. Bozorgi

Publication Info

ISPRS International Journal of Geo-Information, ed. 10, Volume 6, 2021.

International Journal of Geo-Information

MDPI

Article

Analysis of Geotagging Behavior: Do Geotagged Users Represent the Twitter Population?

Amir Karami [1,*], Rachana Redd Kadari [2], Lekha Panati [2], Siva Prasad Nooli [2], Harshini Bheemreddy [2] and Parisa Bozorgi [3]

[1] School of Information Science, University of South Carolina, Columbia, SC 29208, USA
[2] Computer Science and Engineering Department, University of South Carolina, Columbia, SC 29208, USA; rkadari@email.sc.edu (R.R.K.); spanati@email.sc.edu (L.P.); snooli@email.sc.edu (S.P.N.); harshini@email.sc.edu (H.B.)
[3] Arnold School of Public Health, University of South Carolina, Columbia, SC 29208, USA; bozorgip@email.sc.edu
* Correspondence: karami@sc.edu

Abstract: Twitter's APIs are now the main data source for social media researchers. A large number of studies have utilized Twitter data for diverse research interests. Twitter users can share their precise real-time location, and Twitter APIs can provide this information as longitude and latitude. These geotagged Twitter data can help to study human activities and movements for different applications. Compared to the mostly small-scale data samples in different domains, such as social science, collecting geotagged data offers large samples. There is a fundamental question whether geotagged users can represent non-geotagged users. While some studies have investigated the question from different perspectives, they did not investigate profile information and the contents of tweets of geotagged and non-geotagged users. This empirical study addresses this limitation by applying text mining, statistical analysis, and machine learning techniques on Twitter data comprising more than 88,000 users and over 170 million tweets. Our findings show that there is a significant difference (p-value < 0.001) between geotagged and non-geotagged users based on 73% of the features obtained from the users' profiles and tweets. The features can also help to distinguish between geotagged and non-geotagged users with around 80% accuracy. This research illustrates that geotagged users do not represent the Twitter population.

Keywords: social media; Twitter; geotagging; text analysis; topic modeling; linguistic analysis; big data analytics

Citation: Karami, A.; Kadari, R.R.; Panati, L.; Nooli, S.P.; Bheemreddy, H.; Bozorgi, P. Analysis of Geotagging Behavior: Do Geotagged Users Represent the Twitter Population?. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 373. https://doi.org/10.3390/ijgi10060373

Academic Editors: Ourania Kounadi, Bernd Resch and Wolfgang Kainz

Received: 8 March 2021
Accepted: 28 May 2021
Published: 2 June 2021

1. Introduction

Social media has become a mainstream channel of communication where users share and exchange information. In the last decade, social media platforms have grown in popularity and are readily available on mobile devices to connect users to streams of information. Among social media sites, Twitter has more than 320 million users generating 500 million tweets per day [1]. In 2019, the number of monthly active US Twitter users reached 68 million [2]. Twitter comments are restricted to 280 characters and are mostly publicly available. Twitter users can participate in social media activities such as sharing their messages (tweets) or reposting previously published messages (retweets) [3].

Twitter Application Programming Interfaces (APIs) are, collectively, a tool for collecting user data. Twitter API data is now the main social media data source for researchers and policymakers [3]. In the last decade, there has been a marked increase in using Twitter data for research, and we expect to see more Twitter-related studies in the following years [3]. A large number of studies have utilized Twitter data for diverse research interests, such as analyzing the content of tweets related to health issues, such as happiness, diet, and physical activity [4], health disinformation [5], mental health [6], food [7], and LGBT health [8];

understanding Twitter discussions during natural disasters, such as Hurricane Sandy [9]; and mining public opinion during different events, such as the 2016 U.S. elections [10] and the #metoo movement representing sexual harassment experiences [11].

Current literature has utilized quantitative or qualitative methods to analyze Twitter data, such as utilizing machine learning classification methods for sentiment analysis [4], identifying self-expressions of mental illness diagnoses [6], and classifying sexual harassment experiences [11]; text mining techniques for disclosing the semantic of tweets regarding disinformation [5] and LGBT health [8]; and qualitative coding for examining information sharing strategies of social bot information [12].

Twitter users can use three methods to share their location [13]. In the first method, users share their location on their Twitter profile. In the second method, users mention their location in tweets. In the last one, users share their longitude and latitude information through a process called geotagging. The first and the second methods have limitations. They offer broad areas (e.g., city and state) and do not show the movement of users. However, the third method overcomes the limitations and shows precise real-time locations of users. These features have made the third method an attractive accurate format for research [13]. The focus of this research is on geotagging users who share at least one geotagged tweet.

It has been estimated that 4 million geotagged tweets are posted every day [13]. The location data is valuable for researchers to identify spatial patterns and link Twitter data to external datasets, such as the Behavioral Risk Factor Surveillance System (BRFSS) [4] to provide better perspectives on different issues. The location data can reveal different information, such as where users live and work [13]. Therefore, studying geotagged Twitter data can provide more research dimensions, such as comparing the content of tweets regarding different locations.

It has been shown that spatial analysis of Twitter data has attracted researchers [3]. A wide range of studies has been developed based on geotagged Twitter data, such as exploring the diversity of human mobility patterns among individuals and within/between cities [14] and during the COVID-19 pandemic [15,16]; identifying spatiotemporal patterns of tweets during floods and hurricanes [17,18]; analyzing tweets containing discussions on climate change [19]; recognizing health patterns regarding obesity [20–23], diabetes [20], happiness [4], diets and foods [4,24,25], physical activities [4], drug-related health problems [26], and Zika virus [27,28]; and studying geotagging behavior, such as the motivation behind geotagging [29] and patterns of geotagging [30].

The above studies assumed that geotagged users (GUs) can represent non-geotagged users (NGUs). Their reason is that most scientists optimize their data collection by obtaining data samples [31]. Compared to the mostly small-scale data samples in different domains, such as social science [32], collecting geotagged tweets offers large samples. While geotagged Twitter data offer a great opportunity for researchers to identify and study spatial patterns, less than 1% of Tweets are geotagged [33,34]. This small proportion of tweets imposes a negative impact on obtaining tweets containing specific terms [35]. There is a fundamental question whether geotagged tweets and users can represent non-geotagged ones [13,36]. In other words, do geotagged users constitute a random sample of the Twitter population or are they significantly different? This question has encouraged researchers to compare Twitter geotagged and non-geotagged users and tweets [13,36]. Comparing geotagged and non-geotagged tweets and users reveals whether geotagged tweets and users are representative of non-geotagged tweets and users. Two studies compared geotagged and non-geotagged users. The first study showed that there is a significant difference between the users who share and the users who do not share their location based on demographic characteristics, such as age and gender [13]. The second study found that there is a significant difference among geotagged and non-geotagged users based on types of device, country, and language. In addition, the users who share their location on their profile were found to be more likely to use a geotagging service and connect to users with similar geotagging behavior than the users who do not share their location [36].

While the two studies have provided valuable perspectives, they did not investigate profile information and the content of tweets of geotagged and non-geotagged users. Considering this limitation, we expanded the previous work by comparing geotagged users (GUs) and non-geotagged users (NGUs) based on profile information and the content of tweets. Specifically, this study aims to investigate Twitter activities and information sharing patterns of GUs and NGUs. To do this, this paper has collected and analyzed data of more than 88,000 Twitter users to address the following research questions.

RQ1: Is there a significant difference between geotagged and non-geotagged users regarding Twitter profile information?

RQ2: Is there a significant difference between geotagged and non-geotagged users based on the content of tweets?

RQ3: To what extent can the identified features in RQ1 and RQ2 distinguish between geotagged and non-geotagged users?

While addressing RQ3 discloses the prediction power of the features to distinguish between GUs and NGUs, investigating RQ1 and RQ2 reveals whether geotagged users can represent non-geotagged users based on the content features of Twitter profiles and tweets. If there are no discernible differences between GUs and NGUs regarding the features in RQ1 and RQ2, we can consider GUs as representative of NGUs. Otherwise, there is a need to consider methods for ameliorating or controlling.

2. Materials and Methods

This section provides details on data collection and pre-processing, feature extraction, statistical comparison, and prediction analysis.

2.1. Data

We followed the steps in Figure 1 to collect and prepare data for this study. We utilized Twitter's API (https://developer.twitter.com/en/docs/api-reference-index, accessed on 30 May 2021) to randomly sample 5000 real-time tweets per day for 30 days (1.5 million tweets in total) (https://developer.twitter.com/en/docs/twitter-api/v1/tweets/sample-realtime/overview, accessed on 30 May 2021) during 2020. Out of the 1.5 million tweets, we randomly selected 150,000 unique Twitter users (500 users per day for 30 days) and collected their profile information and tweets.

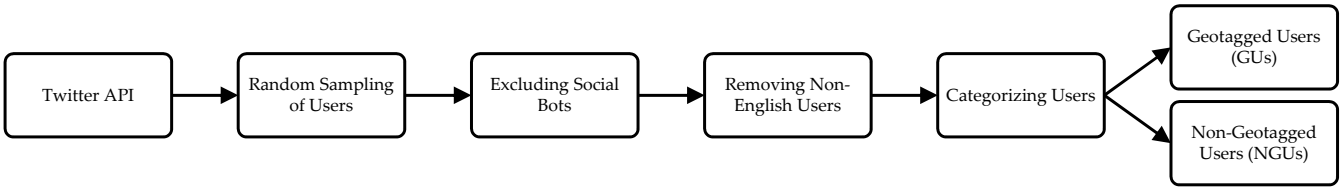

Figure 1. Data collection and preprocessing steps.

While relevant studies did not filter out bots [13,36], this study addresses this limitation. To remove bots and users who post tweets written in non-English languages, we utilized Botometer (https://botometer.osome.iu.edu/, accessed on 30 May 2021), which analyzes Twitter profile features (e.g., number of tweets) with machine learning. This tool assigns a score from 0 (human) to 5 (bot) to each of Twitter accounts and discloses the language of the users [37]. We removed the accounts that were assigned a Botometer's score between 4 and 5, which is the red zone defined by Botometer, and posted tweets written in non-English languages. This process provided 88,976 Twitter users. Due to the API limitation, we were able to collect up to 3200 tweets per user.

One approach to infer the location is identifying geotagged tweets [34]. This approach assumes that geotagged tweets can help to infer location information. This assumption

indicates that, if a user posts a tweet that does not have geotagging information, other geotagged tweets of the user can assist in inferring the location of users. Based on this assumption and similar to [13], we categorized the users into two groups. The first group (GU) included users who shared at least one geotagged tweet containing latitude and longitude coordinates, and the second group (NGU) included users who did not post any geotagged tweets. The final dataset was a list of users with a binary class, GU or NGU.

As we did not define a timeframe and collected tweets without any queries, our data is a representative sample, and we can apply statistical tests to investigate the research questions. Our dataset includes tweets posted in 2020 and before. For example, if one user created an account in 2015 and we collected her/his tweets in 2020, those tweets were posted between 2015 and 2020. In addition, more than 97.82% of accounts had been active for more than three years.

While Twitter stopped adding a precise geotag to tweets in 2019 (https://twitter.com/TwitterSupport/status/1141039841993355264?s=20, accessed on 30 May 2021), we were able to tag the precise location (latitude and longitude) of photographs and add our location to tweets via the service's integration with mapping services [38]. This means that, if a user takes a photo, tags the precise location in the photo using camera applications in her/his phone device, and shares the photo in a tweet, the Twitter API can find the precise location.

2.2. Feature Extraction

This research compared GUs and NGUs based on two types of features, including the profile features of Twitter users and the content features of tweets (Figure 2), to address RQ1–RQ3. The tweet features had three sub-features, including structural, linguistic, and semantic features.

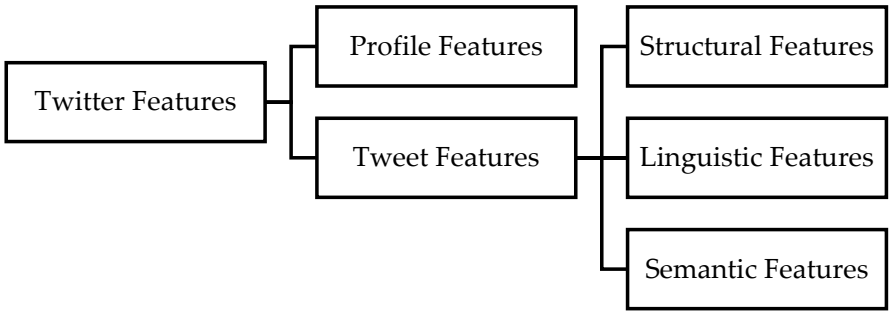

Figure 2. Categories of Twitter features.

To address RQ1, we obtained six structured features from Twitter profiles, including the number of followers, followings, tweets + retweets (RTs), words in bio, favorites, and the age of the account. The null hypothesis of RQ1 is that there is not a significant difference between GUS and NGUS regarding Twitter profile information.

To address RQ2, we obtained both structured and unstructured features from the content of tweets. The null hypothesis of RQ2 is that there is not a significant difference between GUS and NGUS based on structured and unstructured features of tweets. The structured features represent the structural features of tweets, such as the average number of hashtags per tweet. Those features are based on the frequency of tweets, RTs, URLs, hashtags (#), and at signs (@). These features illustrate how users shape and share their tweets. Table 1 shows definitions of Twitter terminology.

The unstructured features illustrate the content features of tweets. We analyzed the unstructured features based on linguistic and semantic features. We processed unstructured features in the content of tweets using two strategies.

Table 1. Twitter Terminology.

Term	Definition
At sign (@)	Tagging other user (s) in a tweet, such as @BarackObama
Hashtag (#)	Providing tweet context and connecting tweets to a specific topic
Followers	Twitter accounts follow updates of a Twitter account
Followings	Twitter accounts are followed by a Twitter account
RT (retweet)	Resharing tweets of other users
URL	Referring to a webpage in a tweet
Bio	A small public summary about a user displayed under Twitter profile picture
Favorite (Like)	Showing appreciation of a tweet by clicking on the likes tab

The first strategy obtained linguistic features using linguistic inquiry and word count (LIWC). This tool processes a document and counts the percentage of words that reflect different emotions, thinking styles, social concerns, and parts of speech within a group of built-in dictionaries [39]. This program was developed in the Java programming language with almost 6400 words, word stems, and selected emoticons [39]. Each dictionary word was mapped to one or multiple word categories using LIWC. For example, the word "cried" can be mapped to five categories: Sadness, Negative Emotion, Overall Affect, Verb, and Past Focus. LIWC has been used for different applications, such as opinion mining [40] and spam detection [41].

LIWC has also been applied in research utilizing geotagged tweets for difference purposes. One study investigated whether Twitter-derived linguistic variables are predictive of a county's health statistics [42]. This study predicted county-level health statistics of the top 100 most populous counties in the US using different features in the content of tweets, such as LIWC features. A similar study has used LIWC to predict county-wide obesity and diabetes [23].

The second strategy to understand unstructured tweets was utilizing topic modeling to disclose the hidden semantic layer of tweets. Among topic models, we chose latent Dirichlet allocation (LDA) [43], which has been considered an effective model [44]. LDA has been utilized for different applications, such as analyzing medical documents to understand clinical notes [45], exploring the content of research papers to identify trends and patterns [46–48], and mining social media to understand online discussions [49–52]. LDA is a generative model that assumes that there is an exchange between words and documents in a corpus represented by bag-of-words using the occurrence of words within a document to represent a corpus. LDA assigns semantically related words to a cluster called a topic. For example, LDA allocates "data", "number", and "computer" to a topic that is interpreted as a theme related to information technology [53].

In this paper, a document represents up to 3200 tweets from a single user. For n documents, m words, and t topics, the outputs of LDA are two matrices. The first one is the probability of each word given a topic, or $P(W_i \mid T_k)$, and the second one is the probability of each topic given a document, or $P(T_k \mid D_j)$ [54]:

$$\begin{array}{c} \\ \text{Words} \\ \\ \\ \\ \mathrm{P(W_i \mid T_k)} \end{array} \begin{array}{c} \text{Topics} \\ \begin{bmatrix} P(W_1|T_1) & \cdots & P(W_1|T_t) \\ \vdots & \ddots & \vdots \\ P(W_m|T_1) & \cdots & P(W_m|T_t) \end{bmatrix} \end{array} \quad \& \quad \begin{array}{c} \\ \text{Topics} \\ \\ \\ \\ \mathrm{P(T_k \mid D_j)} \end{array} \begin{array}{c} \text{Documents} \\ \begin{bmatrix} P(T_1|D_1) & \cdots & P(T_t|D_n) \\ \vdots & \ddots & \vdots \\ P(T_t|D_1) & \cdots & P(T_t|D_n) \end{bmatrix} \end{array}$$

We defined the retrieved tweets from one user as one document and obtained $P(T \mid D)$ to find the probability of topics given a document. Then, we utilized the Java-based MALLET [55] to apply LDA to the tweets. Using the Mallet, we removed stopwords (e.g., "the") and obtained $P(T \mid D)$ for three sets of topics, including 10, 25, and 50 topics. To define the LDA hyperparameters, we followed the literature [46,56] and selected β at 0.01 and α based on $\frac{50}{T}$, where T is the number of topics.

2.3. Statistical Comparison

We investigated 194 features, including six profile features and 10 structural, 93 linguistic, and 85 semantic features of tweets, obtained from Twitter profiles and the content of tweets. We developed statistical tests using the two-sample t-test developed in the R mosaic package [57] to compare GUs and NGUs based on the mean of features. The level of significance level should be set based on sample size [58] using $\frac{0.05}{\sqrt{\frac{N}{100}}}$ [59], where N is the number of users (88,976). So, the passing p-value would be 0.001. To minimize both false positives and false negatives, we adjusted p-values by controlling the false discovery rate (FDR) [60]. Regarding RQ1 and RQ2, our alternative hypothesis was that there was a significant difference between GUs and NGUs regarding the features.

2.4. Prediction Analysis

To address RQ3, we developed a quantitative evaluation based on the profile features and structured, linguistic, and semantic features of tweets. This experiment shows how accurate a classifier can distinguish between GUs and NGUs regarding the 194 features. We hypothesized empirically that the features could help to identify (predict) GUs and NGUs without obtaining location information. In this classification, a classifier assigns a binary class (GU or NGU) to each user. To develop the classifier, we followed the framework in Figure 3.

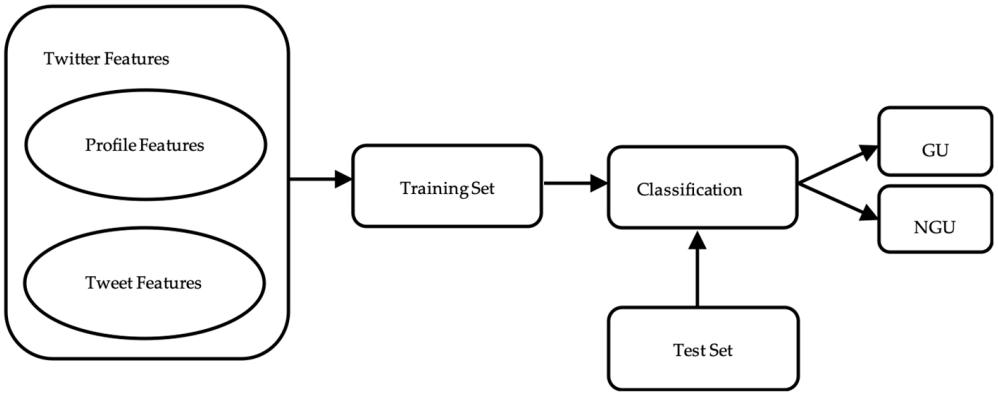

Figure 3. Classification framework.

In our classification formulation, there are n documents $U = \{u_1, \ldots, u_k\}$. Each document represents the tweets of each user. The classification task predicts whether u_i is GU or NGU utilizing a classifier (c) that assigns a label (l) to u_i:

$$c : u_i \rightarrow \{l_1, l_2, \ldots, l_m\}$$

The classifier uses a set of n features, $F = \{f_1, \ldots, f_n\}$, obtained from the profile and tweets of users [61]. To determine the class of each user, we utilized Random Forest, which is a high-performance classification method [62] developed in Weka (https://www.cs.waikato.ac.nz/ml/weka/, accessed on 30 May 2021). We used three-fold cross-validation, in which the data are broken into three subsets, and the holdout method is repeated three times. Each time, one of the three subsets is used as the test set and the other two subsets are used as the training set. To evaluate the classifier, we used the following confusion matrix [63]:

		Predicted	
		GU	**NGU**
Actual	**GU**	True Positive (TP)	False Positive (FP)
	NGU	False Negative (FN)	True Negative (TN)

We measured precision, recall, accuracy, F-measure, and area under the ROC curve (AUC) based on the following definitions:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FP} + \text{TN}}$$

$$\text{F} - \text{measure} = 2 \times \frac{\text{P} \times \text{R}}{\text{P} + \text{R}}$$

ROC shows the tradeoff between FP and TP by plotting FP on the *X*-axis and TP on the *Y*-axis. We provided the features (independent variables) for prediction and used all features, not just those features that satisfied the significance test, to find whether a user is a GU or NGU (dependent variable).

3. Results

We obtained more than 173 million tweets with more than 823 million tokens. Our data contained 18,116 GUs with more than 40 million tweets and 70,860 NGUs with more than 132 million tweets (Table 2). On average, we retrieved 1950.7 tweets per user. The proportion of GUs is much higher than the 1% estimation of geotagged tweets [33,34] because this research focuses on users rather than tweets.

Table 2. Data Summary.

	#Users	#Retrieved Tweets	#Retrieved Tweets/User
Geotagged Users (GUs)	18,116 (20.4%)	40,982,868 (23.6%)	2262.2
Non-Geotagged Users (NGUs)	70,860 (79.6%)	132,583,642 (76.4%)	1871.1
Total	88,976	173,566,510	1950.7

Figure 4 shows the distribution of the number of retrieved tweets per user. In our dataset, 99.09%, 91.43%, 67.67%, 53.38%, and 39.55% of the users have at least 10, 100, 1000, 2000, and 3000 retrieved tweets, respectively. This result indicates that most users have a considerable number of tweets in our dataset.

Considering GUs, the number of geotagged tweets per user is between 1 and 3163 with median 15 and mean 83.95. Our analysis shows that 15%, 27.2%, 37.5%, 16.9%, and 3.4% of GUs posted one, 2–10, 11–100, 101–500, and more than 500 geotagged tweet(s), respectively. This illustrates that 85% of GUs posted more than one geotagged tweet.

To address RQ1, we compared the profile information for the two groups and found that there is a very significant difference (adjusted p-value ≤ 0.001) between GUs and NGUs based on the mean of the number of followers, tweets + RTs, words in the bio, and the account age (Table 3). While this analysis shows that NGUs have a larger number of followers and post a larger number of tweets and retweets than GUs, the number of words in the bio and the account's age are larger for GUs than NGUs. These two groups have a similar number of followings and favorites.

For RQ2, we built our experiments on structural, linguistic, and semantic features. Considering the structural features of tweets, we standardized the features with the total number of tweets or retweets. Our comparison analysis showed that there was a significant difference (adjusted p-value ≤ 0.001) between GUs and NGUs based on seven out of ten features (Table 4). Out of the seven features, we found that the average weight of three features was higher for NGUs than GUs. For example, NGUs retweet more than GUs.

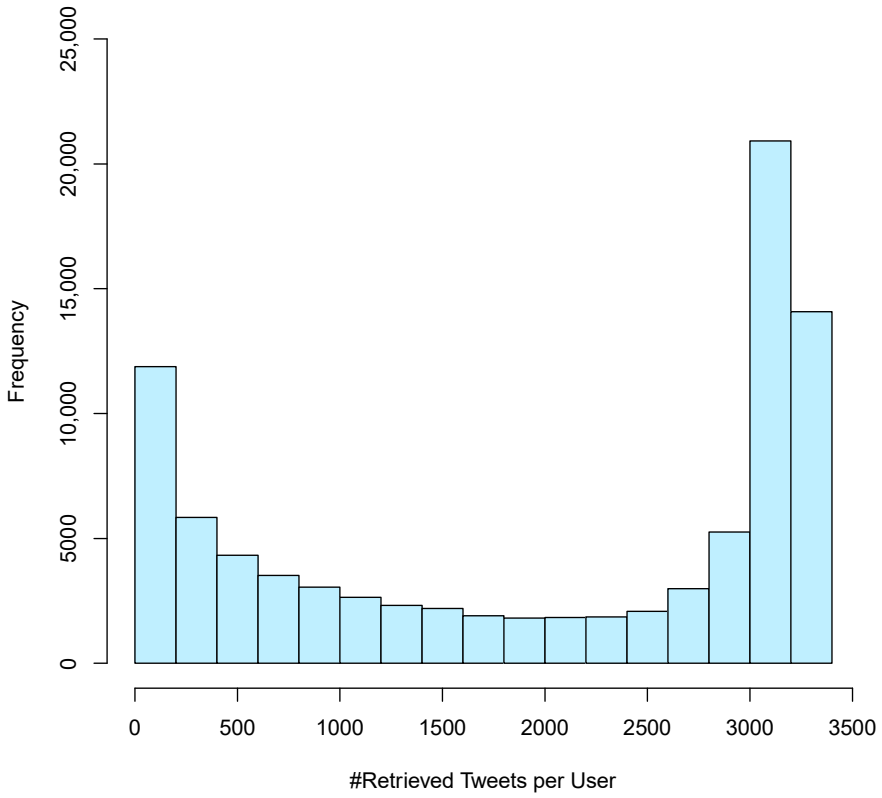

Figure 4. Distribution of the number of retrieved tweets per user.

Table 3. GUs vs. NGUs based on profile information.

Variable	GUs vs. NGUs	*p*-Value
#Followers	GUs < NGUs	*
#Followings	GUs = NGUs	ns
#(Tweets + RTs)	GUs < NGUs	*
#Words in Bio	GUs > NGUs	*
#Favorites	GUs = NGUs	ns
Age of Account	GUs > NGUs	*

(ns: adjusted p-value > 0.001; *: adjusted p-value $\leq$ 0.001).

Considering the linguistic features of tweets, LIWC helped this study to obtain 93 linguistic features of retrieved tweets for each user. This tool provided the total number of words (WC) in the tweets of a user. To standardize WC, we divided WC by the total number of retrieved tweets and retweets. Our analysis showed that there was a significant difference between GUs and NGUs based on 78 features (Table 5). For example, the average number of words per tweet and retweet was higher for NGUs, and GUs used more positive words than NGUs. This finding illustrates that GUs have different preferences than NGUs on choosing words (e.g., pronouns) and discussing issues (e.g., religious). The differences are reflected in clout (i.e., the authority of the author), authenticity (i.e., honest self-depiction), and tone (i.e., the emotional inclination of the author) [64]. These four variables represent a summary of some other variables [39]. For example, the clout category includes we, social words, I, negations (e.g., no, not), and swear words, illustrating social hierarchy [65,66]. For more information on each of the 93 features, refer to [39].

Table 4. GUs vs. NGUs based on the structure of tweets.

Variable	GUs vs. NGUs	*p*-Value
$\frac{\text{\#Retrieved RTs}}{\text{\#Retrieved (Tweets+RTs)}}$	GUs < NGUs	*
$\frac{\text{\#Retrieved Tweets}}{\text{\#Retrieved (Tweets+RTs)}}$	GUs > NGUs	*
$\frac{\text{\#Retrieved RTs With URL}}{\text{\#Retrieved RTs}}$	GUs = NGUs	ns
$\frac{\text{\#Retrieved RTs Without URL}}{\text{\#Retrieved RTs}}$	GUs > NGUs	*
$\frac{\text{\#Retrieved Tweets With URL}}{\text{\#Retrieved Tweets}}$	GUs = NGUs	ns
$\frac{\text{\#Retrieved Tweets Without URL}}{\text{\#Retrieved Tweets}}$	GUs = NGUs	ns
$\frac{\text{\#Retrieved (Tweets+RTs) With URLs}}{\text{\#Retrieved (Tweets+RTs)}}$	GUs < NGUs	*
$\frac{\text{\#Retrieved (Tweets+RTs) Without URLs}}{\text{\#Retrieved (Tweets+RTs)}}$	GUs > NGUs	*
$\frac{\text{\#Retrieved Hashtags}}{\text{\#Retrieved (Tweets+RTs)}}$	GUs > NGUs	*
$\frac{\text{\#Retrieved At Signs}}{\text{\#Retrieved (Tweets+RTs)}}$	GUs < NGUs	*

(ns: adjusted *p*-value > 0.001; *: adjusted *p*-value $\leq$ 0.001).

Table 5. GUs vs. NGUs based on the linguistic features of tweets.

Variable	GUs vs. NGUs	*p*-Value	Variable	GUs vs. NGUs	*p*-Value
$\frac{\text{WC}}{\text{\#Retrieved (Tweets+RTs)}}$	GUs < NGUs	*	Perpetual Processes	GUs > NGUs	*
Analytical Thinking	GUs > NGUs	*	Seeing	GUs > NGUs	*
Clout	GUs < NGUs	*	Hearing	GUs = NGUs	ns
Authentic	GUs > NGUs	*	Feeling	GUs > NGUs	*
Emotional Tone	GUs > NGUs	*	Biological Processes	GUs > NGUs	*
Words per sentence	GUs < NGUs	*	Body	GUs < NGUs	*
Words >6 letters	GUs < NGUs	*	Health/illness	GUs = NGUs	ns
Dictionary words	GUs = NGUs	ns	Sexuality	GUs < NGUs	*
Function Words	GUs < NGUs	*	Ingesting	GUs > NGUs	*
Total pronouns	GUs < NGUs	*	Core Drives and Needs	GUs < NGUs	*
Personal pronouns	GUs < NGUs	*	Affiliation	GUs > NGUs	*
1st pers singular	GUs > NGUs	*	Achievement	GUs > NGUs	*
1st pers plural	GUs < NGUs	*	Power	GUs < NGUs	*
2nd person	GUs < NGUs	*	Reward focus	GUs > NGUs	*
3rd pers singular	GUs < NGUs	*	Risk/prevention focus	GUs < NGUs	*
3rd pers plural	GUs < NGUs	*	Past focus	GUs < NGUs	*
Impersonal pronouns	GUs < NGUs	*	Present focus	GUs = NGUs	ns
Articles	GUs > NGUs	*	Future focus	GUs > NGUs	*
Prepositions	GUs > NGUs	*	Relativity	GUs > NGUs	*
Auxiliary verbs	GUs < NGUs	*	Motion	GUs > NGUs	*
Common adverbs	GUs > NGUs	*	Space	GUs > NGUs	*
Conjunctions	GUs < NGUs	*	Time	GUs > NGUs	*
Negations	GUs < NGUs	*	Work	GUs < NGUs	*

Table 5. *Cont.*

Variable	GUs vs. NGUs	*p*-Value	Variable	GUs vs. NGUs	*p*-Value
Regular verbs	GUs = NGUs	ns	Leisure	GUs > NGUs	*
Adjectives	GUs > NGUs	*	Home	GUs > NGUs	*
Comparatives	GUs = NGUs	ns	Money	GUs < NGUs	*
Interrogatives	GUs < NGUs	*	Religion	GUs < NGUs	*
Numbers	GUs > NGUs	*	Death	GUs < NGUs	*
Quantifiers	GUs = NGUs	ns	Informal Speech	GUs = NGUs	ns
Affect Words	GUs = NGUs	ns	Swear words	GUs < NGUs	*
Positive emotion	GUs > NGUs	*	Netspeak	GUs = NGUs	ns
Negative emotion	GUs < NGUs	*	Assent	GUs > NGUs	*
Anxiety	GUs < NGUs	*	Nonfluencies	GUs > NGUs	*
Anger	GUs < NGUs	*	Fillers	GUs = NGUs	ns
Sadness	GUs < NGUs	*	All Punctuation	GUs = NGUs	ns
Social Words	GUs < NGUs	*	Periods	GUs > NGUs	*
Family	GUs = NGUs	ns	Commas	GUs < NGUs	*
Friends	GUs > NGUs	*	Colons	GUs < NGUs	*
Female referents	GUs < NGUs	*	Semicolons	GUs = NGUs	ns
Male referents	GUs < NGUs	*	Question marks	GUs < NGUs	*
Cognitive Processes	GUs < NGUs	*	Exclamation marks	GUs > NGUs	*
Insight	GUs < NGUs	*	Dashes	GUs > NGUs	*
Cause	GUs < NGUs	*	Quotation marks	GUs > NGUs	*
Discrepancies	GUs < NGUs	*	Apostrophes	GUs > NGUs	*
Tentativeness	GUs < NGUs	*	Parentheses (pairs)	GUs > NGUs	*
Certainty	GUs < NGUs	*	Other punctuation	GUs = NGUs	ns
Differentiation	GUs < NGUs	*			

(ns: adjusted *p*-value > 0.001; *: adjusted *p*-value $\leq$ 0.001).

Considering the semantic features of tweets, we compared GUs and NGUs based on three sets of topics, including 10, 25, and 50 (total 85 topics), generated by LDA. We used the statistical tests to compare GUs and NGUs based on the average weight of each topic given the tweets of a user. Each topic represented a theme, such as an election. This analysis showed the preference of topics for GUs and NGUs. Our comparison analysis showed that the difference between GUs and NGUs was significant (adjusted *p*-value $\leq$ 0.001) in more than 60% of the topics (Table 6). We also found that the difference between GUs and NGUs expanded by increasing the number of topics from 10 to 50.

Table 6. GUs vs. NGUs based on three sets of topics.

#Topics	#Topics with Adjusted *p*-Value > 0.001	#Topics with Adjusted *p*-Value $\leq$ 0.001
10	5 (50%)	5 (50%)
25	10 (40%)	15 (60%)
50	16 (32%)	34 (68%)
85 (total)	31 (37%)	54 (63%)

For RQ3, we developed three models, including all features with a different number of topics. We applied Random Forest on each model. These classification models infer whether a user has at least one geotagged tweet. Table 7 summarizes the performance

metrics of the three models. Our binary classification experiments are based on the profile features (PR) and structured (ST), linguistic (LI), and semantic (SE) features of tweets. Our findings indicate that the features can show whether a user shares geotagged tweets with around 80% accuracy, and GUs and NGUs can be seen as two distinct groups (Table 7).

Table 7. The classification Pprformance of Random Forest.

Set	Features	Accuracy%	F-Measure	AUC
1	PR + ST + LI + SE (10 topics)	80.38	0.739	0.777
2	PR + ST + LI + SE (25 Topics)	80.49	0.742	0.778
3	PR + ST + LI + SE (50 Topics)	80.37	0.737	0.777

PR, ST, and LI include 6, 10, and 93 features, respectively.

The model with 25 topics offered the highest classification performance, including accuracy of 80.40%, F-measure of 0.742, and AUC of 0.778. Although we have thus far shown that there is a significant difference between GUs and NGUs regarding most PR, ST, LI, and SE features examined independently, the classification experiments illustrate that the combination of the features is different for GUs and NGUs. According to the *t*-test, the improvement of Random Forest over ZeroR, which relies on the target and ignores all predictors, is statistically significant. We also compared the three sets of features in Table 5 using the paired *t*-test and found that there is a significant difference among the three sets. The classification showed acceptable accuracy (Table 7) and precision (Table 8), but the recall was low for the GU class. While the goal of this paper is not to predict GUs and NGUs, future study can incorporate other features to improve the recall.

Table 8. The precision and recall of Random Forest.

Precision	Recall	Class	Features
0.652	0.092	GU	
0.808	0.987	NGU	PR + ST + LI + SE (10 topics)
0.776	0.804	Weighted Average	
0.654	0.096	GU	
0.809	0.987	NGU	PR + ST + LI + SE (25 Topics)
0.777	0.804	Weighted Average	
0.666	0.086	GU	
0.807	0.989	NGU	PR + ST + LI + SE (50 Topics)
0.779	0.804	Weighted Average	

4. Discussion

Twitter users can share their exact location when they post their tweets by using the geotagging feature. The location of users is a valuable data source and can be linked to external datasets (e.g., Census Bureau) to enrich social media data and provide a multiple-perspective analysis, such as exploring public opinion on a topic (e.g., election) across different locations. While the proportion of users who enable their geotagging is very low [67], four million geotagged tweets are posted per day [13]. Obtaining spatial data and developing spatial analysis on social media are interesting for not only researchers [3] but also companies for different business purposes, such as developing targeted real-time social media advertisements to enhance services and products [68].

In the literature, there is an assumption that users who geotag their tweets could represent all Twitter users [33,34]. Thus, they could not report any significant difference between geotagged and non-geotagged users. While there are few studies that compare these two groups, there is no research comparing GUs and NGUs based on the content of profile and tweets. This research addresses this limitation by comparing GUs and NGUs based on 194 variables (features) obtained from profiles and tweets of GUs and NGUs. This paper set out to address three research questions.

This study illustrates that there is a significant difference between GUs and NGUs based on 143 features (73% of the 194 features). Therefore, the assumption of no difference between the two groups is not accurate, and geotagged users may not represent non-geotagged users. Our data size indicates that we can be confident that our findings are not due to random chance.

This study indicates that information sharing and the behavior of GUs and NGUs are mostly different. For example, compared to NGUs, geotagged users have been on Twitter for a longer time and prefer to disclose more information about themselves using more words in their profile bio. However, NGUs have more followers and post a higher number of tweets and retweets. While NGUs like retweets, having URLs in their tweets and retweets, and mentioning other users in their posts, GUs prefer tweets (instead of retweets) and using more hashtags.

This paper also shows that GUs and NGUs use different linguistics patterns. For instance, NGUs use more words in their tweets and retweets and talk about work and money more than GUs. Geotagged users post tweets and prefer using pronouns such as "I" and positive words and share more information about home and leisure than NGUs. This is an important finding because, for example, if researchers only analyze geotagged users or tweets to understand public opinion regarding a candidate during an election, they might conclude Twitter users have a positive opinion with respect to that candidate. However, it might simply reflect the fact that GUs are willing to post positive tweets.

Our semantic analysis using topic modeling also shows that GUs and NGUs have different preferences in terms of topical themes, indicating that GUs and NGUs do not share similar interests in their Twitter posts. Suppose that GUs are more interested in topic x than topic y, but NGUs prefer topic y more than topic x; therefore, analyzing geotagged tweets can lead us to conclude that Twitter users are more interested in topic x than topic y. However, the reason behind this conclusion is that NGUs are excluded from the analysis. For example, we found that GUs shows more interest in talking about different foods than NGUs. In this case, diet experts might overestimate food consumption of Twitter users if they focus on geotagged users.

Previous studies have shown that there is a significant difference between GUs and NGUs based on demographics, types of device, country, language, profile location, and the geotagging preference of their friends. In brief, the findings of previous studies and our results indicate that the geotagged users are not representative of the Twitter population. To address this limitation, there is a need to utilize methods to not only use profile information but also infer the location of NGUs through utilizing profile information [34], the user's network of friends [69], time zones [34], the content of tweets [70], URL links [34], combination of different features such as profile information, the content of tweets, and place labelling in tweets [71,72].

This research concludes that GUs and NGUs have different preferences on topics of interest and creating and sharing social media posts. This study offers an empirical foundation to inform research on geotagged Twitter data, such as monitoring health issues and public opinion. The impact of our results will depend on the topic being studied. If researchers need to analyze geotagged users for a specific study, we suggest investigating whether there is a significant difference between GUs and NGUs on a data sample. Otherwise, their results would not be generalizable to the entire Twitter population. This suggestion helps to improve the generality power of Twitter studies investigating spatial factors by increasing the number of users whose location is identified. We also suggest researchers who are interested in linking Twitter data to external datasets to explore whether there is a significant correlation between Twitter data and external characteristics, such as socio-demographic factors, before further investigation [73].

While this study examined important research questions, there are some limitations. First, we were not aware of demographic information about users (e.g., gender). Second, the focus of this study was limited to tweets written in English. Third, this research was limited to geotagged users on the Twitter population, not other types of users and the

real-world population. Fourth, while we obtained a random sample of Twitter users, this study does not consider spatial and temporal factors. Fifth, we assumed GUs are the ones who share at least one geotagged tweet. However, we did not investigate whether there is a statistical difference between GUs with high and low number of geotagged tweets. Future work could address these limitations by utilizing methods to infer demographic information such as the gender of users (e.g., analyzing the name of users), analyzing non-English tweets, studying other types of location Twitter sharing strategies, such as place-tagging, investigating differences regarding different times and locations, exploring categories of GUs based on the number of geotagged tweets, and analyzing special and temporal factors such as spatiotemporal analysis of tweets [74]. It would also be interesting to expand this study by comparing users based on their county (e.g., GUs vs. NGUs in Canada or France). It was found that most geotagged tweets are from non-Twitter platforms, such as Instagram [36]. Therefore, comparing the geotagging behavior of users regarding different platform could be an interesting extension of this paper. In addition, while digging into the details of the detected differences is beyond the scope of this research, understanding the reasons behind the identified differences would be a fruitful future avenue of research.

To conclude, the contributions of this paper are five-fold. First, this is the first research that compares GUs and NGUs based on profile information and the content of tweets on Twitter. Second, this study shows GUs and NGUs mostly have different information sharing behavior. Third, GUs and NGUs have distinct linguistics choices and topics of interests. Fourth, this work offers new directions in studying location sharing behavior. Fifth, this paper proposes suggestions for controlling adverse effects of sampling geotagged users and tweets. The contributions of this paper can be used by studies utilizing geotagged Twitter data for studying the content of profile and tweets and their linguistics and semantic pattern.

Author Contributions: Conceptualization, Amir Karami; methodology, Amir Karami; software, Amir Karami, Rachana Redd Kadari, Lekha Panati, Siva Prasad Nooli, and Harshini Bheemreddy; validation, Amir Karami, Rachana Redd Kadari, Lekha Panati, Siva Prasad Nooli, and Harshini Bheemreddy; formal analysis, Amir Karami, Rachana Redd Kadari, Lekha Panati, Siva Prasad Nooli, Harshini Bheemreddy, and Parisa Bozorgi; investigation, Amir Karami, Rachana Redd Kadari, Lekha Panati, Siva Prasad Nooli, Harshini Bheemreddy, and Parisa Bozorgi; resources, Amir Karami; data curation, Amir Karami, Rachana Redd Kadari, Lekha Panati, Siva Prasad Nooli, and Harshini Bheemreddy; writing—original draft preparation, Amir Karami, Rachana Redd Kadari, Lekha Panati, Siva Prasad Nooli, Harshini Bheemreddy, and Parisa Bozorgi; writing—review and editing, Amir Karami, Rachana Redd Kadari, Lekha Panati, Siva Prasad Nooli, Harshini Bheemreddy, and Parisa Bozorgi; visualization, Amir Karami; supervision, Amir Karami; project administration, Amir Karami; funding acquisition, Amir Karami. All authors have read and agreed to the published version of the manuscript.

Funding: This work is partially supported by the Advanced Support for Innovative Research Excellence (ASPIRE) grant, the Big Data Health Science Center, and the Social Science Research grant at the University of South Carolina. All opinions, findings, conclusions, and recommendations in this paper are those of the authors and do not necessarily reflect the views of the funding agency.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Aslam, S. Twitter by the Numbers: Stats, Demographics & Fun Facts. Available online: https://www.omnicoreagency.com/twitter-statistics/#:~{}:text=Twitter%20Demographics&text=There%20are%20262%20million%20International,users%20have%20higher%20college%20degrees. (accessed on 11 February 2021).
2. Clement, J. Twitter: Number of Monthly Active U.S. Users 2010–2019. Available online: https://www.statista.com/statistics/274564/monthly-active-twitter-users-in-the-united-states/ (accessed on 7 July 2020).
3. Karami, A.; Lundy, M.; Webb, F.; Dwivedi, Y.K. Twitter and research: A systematic literature review through text mining. *IEEE Access* **2020**, *8*, 67698–67717. [CrossRef]

4. Nguyen, Q.C.; Li, D.; Meng, H.-W.; Kath, S.; Nsoesie, E.; Li, F.; Wen, M. Building a national neighborhood dataset from geotagged Twitter data for indicators of happiness, diet, and physical activity. *JMIR Public Health Surveill.* **2016**, *2*, e158. [CrossRef]
5. Karami, A.; Lundy, M.; Webb, F.; Turner-McGrievy, G.; McKeever, B.W.; McKeever, R. Identifying and Analyzing Health-Related Themes in Disinformation Shared by Conservative and Liberal Russian Trolls on Twitter. *Int. J. Environ. Res. Public Health* **2021**, *18*, 2159. [CrossRef]
6. Coppersmith, G.; Dredze, M.; Harman, C. Quantifying mental health signals in Twitter. In Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality; ACL: Baltimore, MD, USA, 2014; pp. 51–60.
7. Abbar, S.; Mejova, Y.; Weber, I. You tweet what you eat: Studying food consumption through twitter. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, Seoul, Korea, 18–23 April 2015; pp. 3197–3206.
8. Karami, A.; Webb, F. Analyzing health tweets of LGB and transgender individuals. *Proc. Assoc. Inf. Sci. Technol.* **2020**, *57*, e264. [CrossRef]
9. Pourebrahim, N.; Sultana, S.; Edwards, J.; Gochanour, A.; Mohanty, S. Understanding communication dynamics on Twitter during natural disasters: A case study of Hurricane Sandy. *Int. J. Disaster Risk Reduct.* **2019**, *37*, 101176. [CrossRef]
10. Fang, A.; Habel, P.; Ounis, I.; MacDonald, C. Votes on twitter: Assessing candidate preferences and topics of discussion during the 2016 US presidential election. *SAGE Open* **2019**, *9*, 2158244018791653. [CrossRef]
11. Modrek, S.; Chakalov, B. The #MeToo movement in the United States: Text analysis of early twitter conversations. *J. Med. Internet Res.* **2019**, *21*, e13837. [PubMed]
12. Kitzie, V.L.; Mohammadi, E.; Karami, A. "Life never matters in the DEMOCRATS MIND": Examining strategies of retweeted social bots during a mass shooting event. *Proc. Assoc. Inf. Sci. Technol.* **2018**, *55*, 254–263. [CrossRef]
13. Sloan, L.; Morgan, J. Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter. *PLoS ONE* **2015**, *10*, e0142209. [CrossRef]
14. Jurdak, R.; Zhao, K.; Liu, J.; AbouJaoude, M.; Cameron, M.; Newth, D. Understanding human mobility from Twitter. *PLoS ONE* **2015**, *10*, e0131469.
15. Huang, X.; Li, Z.; Jiang, Y.; Li, X.; Porter, D. Twitter reveals human mobility dynamics during the COVID-19 pandemic. *PLoS ONE* **2020**, *15*, e0241957. [CrossRef]
16. Xu, P.; Dredze, M.; Broniatowski, D.A. The Twitter Social Mobility Index: Measuring Social Distancing Practices with Geolocated Tweets. *J. Med. Internet Res.* **2020**, *22*, e21499. [CrossRef] [PubMed]
17. Li, Z.; Wang, C.; Emrich, C.T.; Guo, D. A novel approach to leveraging social media for rapid flood mapping: A case study of the 2015 South Carolina floods. *Cartogr. Geogr. Inf. Sci.* **2018**, *45*, 97–110. [CrossRef]
18. Martín, Y.; Li, Z.; Cutter, S.L. Leveraging Twitter to gauge evacuation compliance: Spatiotemporal analysis of Hurricane Matthew. *PLoS ONE* **2017**, *12*, e0181701. [CrossRef]
19. Dahal, B.; Kumar, S.A.; Li, Z. Topic modeling and sentiment analysis of global climate change tweets. *Soc. Netw. Anal. Min.* **2019**, *9*, 1–20. [CrossRef]
20. Nguyen, Q.C.; Brunisholz, K.D.; Yu, W.; McCullough, M.; Hanson, H.A.; Litchman, M.L.; Li, F.; Wan, Y.; VanDerslice, J.A.; Wen, M. Twitter-derived neighborhood characteristics associated with obesity and diabetes. *Sci. Rep.* **2017**, *7*, 1–10. [CrossRef]
21. Cesare, N.; Dwivedi, P.; Nguyen, Q.C.; Nsoesie, E.O. Use of social media, search queries, and demographic data to assess obesity prevalence in the United States. *Palgrave Commun.* **2019**, *5*, 1–9. [CrossRef]
22. Ghosh, D.; Guha, R. What are we 'tweeting'about obesity? Mapping tweets with topic modeling and Geographic Information System. *Cartogr. Geogr. Inf. Sci.* **2013**, *40*, 90–102. [CrossRef]
23. Gore, R.J.; Diallo, S.; Padilla, J. You are what you tweet: Connecting the geographic variation in america's obesity rate to Twitter content. *PLoS ONE* **2015**, *10*, e0133505. [CrossRef] [PubMed]
24. Widener, M.J.; Li, W. Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US. *Appl. Geogr.* **2014**, *54*, 189–197. [CrossRef]
25. Karami, A.; Dahl, A.A.; Shaw, G.; Valappil, S.P.; Turner-McGrievy, G.; Kharrazi, H.; Bozorgi, P. Analysis of Social Media Discussions on (#)Diet by Blue, Red, and Swing States in the U.S. *Healthcare* **2021**, *9*, 518.
26. Cao, Y.; Stewart, K.; Factor, J.; Billing, A.; Massey, E.; Artigiani, E.; Wagner, M.; Dezman, Z.; Wish, E. Using socially-sensed data to infer ZIP level characteristics for the spatiotemporal analysis of drug-related health problems in Maryland. *Health Place* **2020**, *63*, 102345. [CrossRef]
27. Farhadloo, M.; Winneg, K.; Chan, M.-P.S.; Jamieson, K.H.; Albarracin, D. Associations of topics of discussion on Twitter with survey measures of attitudes, knowledge, and behaviors related to Zika: Probabilistic study in the United States. *JMIR Public Health Surveill.* **2018**, *4*, e16. [CrossRef]
28. Daughton, A.R.; Pruss, D.; Arnot, B.; Szafir, D.A.; Paul, M.J. Characteristics of Zika Behavior Discourse on Twitter. In Proceedings of the SMM4H@ AMIA, Washington, DC, USA, 4 November 2017; pp. 27–31.
29. Tasse, D.; Liu, Z.; Sciuto, A.; Hong, J.I. State of the geotags: Motivations and recent changes. In Proceedings of the Eleventh International AAAI Conference on Web and Social Media, Montreal, QC, Canada, 15–18 May 2017.
30. Noulas, A.; Scellato, S.; Mascolo, C.; Pontil, M. An empirical study of geographic user activity patterns in foursquare. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, 17–21 July 2011.

31. de Marchi, S.; Page, S.E. Computational Social Science: Discovery and Prediction. *Perspect. Polit.* **2016**, *14*, 1169. [CrossRef]
32. Chang, R.M.; Kauffman, R.J.; Kwon, Y. Understanding the paradigm shift to computational social science in the presence of big data. *Decis. Support Syst.* **2014**, *63*, 67–80. [CrossRef]
33. Tweet Geospatial Metadata. Available online: https://developer.twitter.com/en/docs/tutorials/tweet-geo-metadata (accessed on 9 July 2020).
34. Ajao, O.; Hong, J.; Liu, W. A survey of location inference techniques on Twitter. *J. Inf. Sci.* **2015**, *41*, 855–864. [CrossRef]
35. Burnap, P.; Gibson, R.; Sloan, L.; Southern, R.; Williams, M. 140 characters to victory?: Using Twitter to predict the UK 2015 General Election. *Elect. Stud.* **2016**, *41*, 230–233. [CrossRef]
36. Huang, B.; Carley, K.M. A large-scale empirical study of geotagging behavior on twitter. In Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, East Lansing, MI, USA, 14–19 August 2019; pp. 365–373.
37. Yang, K.-C.; Varol, O.; Hui, P.-M.; Menczer, F. Scalable and generalizable social bot detection through data selection. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 1096–1103.
38. Khalid, A. Twitter Removes Precise Geo-Tagging Option from Tweets. Available online: https://www.engadget.com/2019-06-19-twitter-removes-precise-geo-tagging.html (accessed on 23 April 2021).
39. Pennebaker, J.W.; Boyd, R.L.; Jordan, K.; Blackburn, K. *The Development and Psychometric Properties of LIWC2015*; Pennebaker Conglomerates: Austin, TX, USA, 2015.
40. Karami, A.; Bennett, L.S.; He, X. Mining public opinion about economic issues: Twitter and the us presidential election. *Int. J. Strateg. Decis. Sci. IJSDS* **2018**, *9*, 18–28. [CrossRef]
41. Karami, A.; Zhou, B. Online Review Spam Detection by New Linguistic Features. In Proceedings of the iConference, Newport Beach, CA, USA, 24–27 March 2015.
42. Culotta, A. Estimating county health statistics with twitter. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Toronto, ON, Canada, 26 April–1 May 2014; pp. 1335–1344.
43. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
44. Karami, A. Fuzzy Topic Modeling for Medical Corpora. Ph.D. Thesis, University of Maryland, Baltimore County, MD, USA, 2015.
45. Wang, L.; Lakin, J.; Riley, C.; Korach, Z.; Frain, L.N.; Zhou, L. Disease trajectories and end-of-life care for dementias: Latent topic modeling and trend analysis using clinical notes. In Proceedings of the AMIA Annual Symposium Proceedings, San Francisco, CA, USA, 3–7 November 2018; Volume 2018, p. 1056.
46. Van Altena, A.J.; Moerland, P.D.; Zwinderman, A.H.; Olabarriaga, S.D. Understanding big data themes from scientific biomedical literature through topic modeling. *J. Big Data* **2016**, *3*, 23. [CrossRef]
47. Mohammadi, E.; Karami, A. Exploring research trends in big data across disciplines: A text mining analysis. *J. Inf. Sci.* **2020**, 0165551520932855. [CrossRef]
48. Karami, A.; Bookstaver, B.; Nolan, M.; Bozorgi, P. Investigating Diseases and Chemicals in COVID-19 Literature with Text Mining. *Int. J. Inf. Manag. Data Insights* **2021**, 100016. [CrossRef]
49. Money, V.; Karami, A.; Turner-McGrievy, B.; Kharrazi, H. Seasonal characterization of diet discussions on Reddit. *Proc. Assoc. Inf. Sci. Technol.* **2020**, *57*, e320. [CrossRef]
50. Anderson, M.; Karami, A.; Bozorgi, P. Social media and COVID-19: Can social distancing be quantified without measuring human movements? *Proc. Assoc. Inf. Sci. Technol.* **2020**, *57*, e378. [CrossRef] [PubMed]
51. Frank, W.; Karami, A.; Vanessa, K. Characterizing Diseases and Disorders in Gay Users' Tweets. In Proceedings of the Southern Association for Information Systems (SAIS), Atlanta, GA, USA, 23 March 2018.
52. Collins, M.; Karami, A. Social media analysis for organizations: Us northeastern public and state libraries case study. In Proceedings of the Southern Association for Information Systems (SAIS), Atlanta, GA, USA, 23 March 2018.
53. Blei, D.M. Probabilistic Topic Models. *Commun. ACM* **2012**, *55*, 77–84. [CrossRef]
54. Karami, A.; White, C.N.; Ford, K.; Swan, S.; Spinel, M.Y. Unwanted advances in higher education: Uncovering sexual harassment experiences in academia with text mining. *Inf. Process. Manag.* **2020**, *57*, 102167. [CrossRef]
55. McCallum, A.K. *MALLET: A Machine Learning for Language Toolkit*; University of Massachusetts: Amherst, MA, USA, 2002.
56. Steyvers, M.; Griffiths, T. Probabilistic topic models. In *Handbook of Latent Semantic Analysis*; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 2007; Volume 427, pp. 424–440.
57. Pruim, R.; Kaplan, D.; Horton, N. *Mosaic: Project MOSAIC Statistics and Mathematics Teaching Utilities*; R Package Version 06-2 HttpCRAN R-Proj. Orgpackage Mosaic Google Sch. *R J.* **2012**, *9*, 77–102. [CrossRef]
58. Kim, J.H.; Ji, P.I. Significance testing in empirical finance: A critical review and assessment. *J. Empir. Finance* **2015**, *34*, 1–14. [CrossRef]
59. Good, I.J. C140. Standardized tail-area prosabilities. *J. Stat. Comput. Simul.* **1982**, *16*, 65–66. [CrossRef]
60. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **1995**, *57*, 289–300. [CrossRef]
61. Karami, A.; Swan, S.; Moraes, M.F. Space identification of sexual harassment reports with text mining. *Proc. Assoc. Inf. Sci. Technol.* **2020**, *57*, e265. [CrossRef]
62. Statnikov, A.; Wang, L.; Aliferis, C.F. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinform.* **2008**, *9*, 1–10. [CrossRef]

63. Karami, A. Application of fuzzy clustering for text data dimensionality reduction. *Int. J. Knowl. Eng. Data Min.* **2019**, *6*, 289–306. [CrossRef]
64. Arenas, E.S. Exploring pornography in Widad Benmoussa's poetry using LIWC and corpus tools. *Sex. Cult.* **2018**, *22*, 1094–1111. [CrossRef]
65. Kacewicz, E.; Pennebaker, J.W.; Davis, M.; Jeon, M.; Graesser, A.C. Pronoun use reflects standings in social hierarchies. *J. Lang. Soc. Psychol.* **2014**, *33*, 125–143. [CrossRef]
66. Xu, W.W.; Zhang, C. Sentiment, richness, authority, and relevance model of information sharing during social Crises—The case of# MH370 tweets. *Comput. Hum. Behav.* **2018**, *89*, 199–206.
67. Sloan, L.; Morgan, J.; Housley, W.; Williams, M.; Edwards, A.; Burnap, P.; Rana, O. Knowing the Tweeters: Deriving sociologically relevant demographics from Twitter. *Sociol. Res. Online* **2013**, *18*, 74–84. [CrossRef]
68. Gigante, M.D. Why Location Data Matters: 4 Big Benefits for Marketers. Available online: https://www.mdgadvertising.com/marketing-insights/why-location-data-matters-4-big-benefits-for-marketers/ (accessed on 3 March 2021).
69. Rahimi, A.; Cohn, T.; Baldwin, T. Twitter User Geolocation Using a Unified Text and Network Prediction Model. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Beijing, China, 26–31 July 2015; pp. 630–636.
70. Han, B.; Cook, P.; Baldwin, T. Geolocation prediction in social media data by finding location indicative words. In Proceedings of the COLING 2012, Mumbai, India, 1 December 2012; pp. 1045–1062.
71. Laylavi, F.; Rajabifard, A.; Kalantari, M. A multi-element approach to location inference of twitter: A case for emergency response. *ISPRS Int. J. Geo-Inf.* **2016**, *5*, 56. [CrossRef]
72. Havas, C.; Resch, B.; Francalanci, C.; Pernici, B.; Scalia, G.; Fernandez-Marquez, J.L.; Van Achte, T.; Zeug, G.; Mondardini, M.R.R.; Grandoni, D. E2mc: Improving emergency management service practice through social media and crowdsourcing analysis in near real time. *Sensors* **2017**, *17*, 2766. [CrossRef]
73. Ostermann, F.O. Linking Geosocial Sensing with the Socio-Demographic Fabric of Smart Cities. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 52. [CrossRef]
74. Almatar, G.M.; Alazmi, H.S.; Li, L.; Fox, E.A. Applying GIS and Text Mining Methods to Twitter Data to Explore the Spatiotemporal Patterns of Topics of Interest in Kuwait. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 702. [CrossRef]