International Journal of Geo-Information

MDPI

Article

A Multi-Element Approach to Location Inference of Twitter: A Case for Emergency Response

Farhad Laylavi *, Abbas Rajabifard and Mohsen Kalantari

Centre for Disaster Management and Public Safety, Department of Infrastructure Engineering, The University of Melbourne, Parkville, Melbourne, VIC 3010, Australia; abbas.r@unimelb.edu.au (A.R.); mohsen.kalantari@unimelb.edu.au (M.K.)

* Correspondence: laylavi.f@unimelb.edu.au; Tel.: +61-3-9035-3723

Academic Editors: Georg Gartner, Haosheng Huang and Wolfgang Kainz

Received: 1 March 2016; Accepted: 21 April 2016; Published: 28 April 2016

Abstract: Since its inception, Twitter has played a major role in real-world events—especially in the aftermath of disasters and catastrophic incidents, and has been increasingly becoming the first point of contact for users wishing to provide or seek information about such situations. The use of Twitter in emergency response and disaster management opens up avenues of research concerning different aspects of Twitter data quality, usefulness and credibility. A real challenge that has attracted substantial attention in the Twitter research community exists in the location inference of twitter data. Considering that less than 2% of tweets are geotagged, finding location inference methods that can go beyond the geotagging capability is undoubtedly the priority research area. This is especially true in terms of emergency response, where spatial aspects of information play an important role. This paper introduces a multi-elemental location inference method that puts the geotagging aside and tries to predict the location of tweets by exploiting the other inherently attached data elements. In this regard, textual content, users' profile location and place labelling, as the main location-related elements, are taken into account. Location-name classes in three granularity levels are defined and employed to look up the location references from the location-associated elements. The inferred location of the finest granular level is assigned to a tweet, based on a novel location assignment rule. The location assigned by the location inference process is considered to be the inferred location of a tweet, and is compared with the geotagged coordinates as the ground truth of the study. The results show that this method is able to successfully infer the location of 87% of the tweets at the average distance error of 12.2 km and the median distance error of 4.5 km, which is a significant improvement compared with that of the current methods that can predict the location with much larger distance errors or at a city-level resolution at best.

Keywords: location inference; social media; twitter; emergency response

1. Introduction

Being ubiquitous and omnipresent, social media are becoming significant channels for information dissemination and communication among the general population. Such platforms are also changing the speed and nature with which people perceive and respond to emergency or unanticipated events. The first news about an emergency situation is now likely to appear on social media channels, such as Twitter, rather than conventional news sources. The most recent example of such an event is the Paris attacks that occurred on 13 November 2015, after which eyewitnesses posted on their social network accounts, being mainly Twitter, to warn others about what was happening [1]. The practical application of Twitter in emergency situations suggests new avenues for investigation regarding the effective use of social media platforms, such as Twitter, in catastrophic events and making them fit into the requirements of emergency response. Social media, in this context, can bridge the gap that

exists in the current emergency response systems regarding what South [2] describes as the lack of immediate flow of information from people at the scene towards authorities or those who can provide help. Supported by a number of studies [3,4], Twitter, among the social media platforms, has shown capability to be a valuable augmentation to the current emergency response systems. However, this is not a straightforward addition, as there are significant challenges that must first be overcome.

Up-to-date and spatially-referenced crisis information plays a vital role in emergency response [5–7]. In other words, emergency response necessitates information that is timely and from an identifiable location. Whilst Twitter is benefitting from a reliable timeline that meets the timeliness requirement of emergency response and makes Twitter a perfect fit for the time-sensitive contexts, identifying the location from where the information is being disseminated is still a non-trivial issue. It is important to note that timely information from an unknown location does not carry much value for emergency response. Since 2009, when Twitter started accommodating geotagging [8], tweets have been able to contain geographic coordinates attached by GPS enabled devices. However, despite the inherent real-time nature of tweets, geotagging is an "opt-in" service to be enabled at the user's discretion. Results of the experimental analysis conducted on over 300 thousand randomly collected tweets in the Centre for Disaster Management and Public Safety (CDMPS) in April 2015 show that about 2% of all tweets are geotagged and contain a precise location. In the related literature, this rate ranges from 0.42% [9] to 3.17% [10]. Systems or tools that extract the location of tweets by taking only geotagging into account can benefit from a small fraction of Twitter data, even though there is valuable information in the remaining non-geotagged chunk. Thus, finding methods to infer the location information from the other inherent capabilities of tweets, such as textual contents or user profile location, can be an essential alternative.

There is a growing research interest in the issues centred around the location inference of Twitter along with proposing methods to address them, which are discussed in Section 2. However, there are considerable gaps in the current state of knowledge. Firstly, the current methods utilise only one of the existing elements of Twitter data for inferring the location, whilst several potential elements exist for location inference (e.g., textual content, profile location and place labelling) that can be combined to improve the performance of location inference algorithms. Additionally, the current methods could reach a geographical location accuracy of up to the city level in the best case, which does not seem to be an adequate level of resolution for emergency response.

This paper introduces a novel method that, to the best of authors' knowledge, for the first time exploits all the potential sources of location information in a multi-elemental approach and achieves the average and median distance error of 12.2 km and 4.5 km, respectively, for 87% of the sample tweets. The introduced method makes use of all potential location information carriers for the inference of the location of Twitter data in the absence of geotagging. Some experiments that validate the proposed method are carried out. The experiments employ a dataset of tweets that have been collected between the 21 and 26 of April 2015, when severe weather conditions affected the Sydney area, causing the collapse of a number of warehouses in Western Sydney [11]. The main contribution of this research can be summarized in the following areas:

- Getting to know Twitter data, the potential elements of location information within a tweet, as well as dealing with the Twitter data collection and sampling
- Proposing a hybrid and multi-elemental approach towards the location inference on Twitter, which significantly improves the location accuracy of the current methods.

The rest of the paper is organized as follows. Section 2 describes and summarises the related work. Section 3 gives a brief introduction to Twitter data and investigates the potential location related elements. Section 4 explains the design of the method and explains essential preliminaries of the work, including data collection and sampling processes. Both the implementation and evaluation of the proposed method are discussed in Section 5. The paper is discussed and concluded in Section 6, with perspectives for future research.

2. Existing Approaches to Location Inference

Retrieving location information of Twitter data, known as "location inference", has received comparatively considerable attention in the literature. Location inference, in general, can be explained as the retrieval process of the location information from each of textual content, location-specific elements, or the user's social network. A number of studies have focused on the nature of geotagged tweets only, and how this capability can be used to track and analyse different subjects in domains, such as public health [12], societal events [13], political elections [14], tourist spots [15], and earthquakes [16]. However, as mentioned before, geotagged tweets form about 2% of all public tweets broadcasted by Twitter users. This poses a need for methods that can use the other components of tweets to extract location information and enhance the overall location reliability of Twitter data for emergency response. Different methods have been adopted from various fields, such as machine learning, statistics, probability and natural language processing to fulfill the need for more accurate and precise location inference methods [17].

Studies concerned with the textual content of tweets for determining location in the absence of geotagging predominantly focus on detecting and extracting the geographic references cited in the textual content. These references might be in the form of "location-indicative words" (LIWs) or gazetteer terms that can be geocoded using a spatial database. Eisenstein *et al.* [18] describe a model named "geographic topic model" and implement this model on US-based users to geolocate them based on their content. Their model obtains a median distance error of 494 km. This error rate is lowered by Wing and Baldridge [19], who get a median error of 479 km for their model. Cheng, Caverlee and Lee [9] offer an approach that analyses the content of geotagged tweets and provides the statistics of the most frequent words in each city. With their method, 51% of randomly sampled Twitter users are placed within 100 miles of their actual location. Watanabe *et al.* [20] propose a system called 'Jasmine', for detection of events on a local scale through extracting and analysing the co-occurring terms within the content of tweets. In an approach carried out by Dalvi *et al.* [21], users are located based on indirect spatial references found in the content, taking into account the restaurants as the target object of the study. Han *et al.* [22] introduce a geolocation prediction platform by detecting and analysing the "location-indicative words". Their method reduces the median prediction error distance by 209 km. In a method performed by Minot *et al.* [23], a combination of content analysis and assessment of the users' social interactions (user mentions in content) is used and city-level accuracy is observed for 60% of users in their sample.

Approaches also exist that go beyond the textual content of tweets for location inference purposes. For example, Hecht *et al.* [24] study users' profile locations through utilising a Multinomial Naïve Bayes model to classify user location with a regional focus and allocate users to their home states with an accuracy of up to 30%. Hecht *et al.* find that users, either knowingly or inadvertently, disclose location information in their tweets. Hiruta *et al.* [25] carry out a method to detect and classify tweets based on the possible correlation of users' profile locations with both textual content and geotagging in different categories. There exists no stated evidence of the achieved geographic granularity in their study, but it seems that the achieved geographic resolution of their work is not finer than city-level.

In a more related work, Schulz *et al.* [26] propose a location inference method through combining the potential sources of spatial indicators, such as tweet messages, profile location information, internet links and time zones using a polygon mapping technique which estimates the location of 54% of tweets within a 50 km radius. In overall, their method is able to create the location estimation of 92% of tweets with the average distance error of 1408 kilometres and the median distance error of 30 km by exploiting multiple external sources such as Geonames, DBPedia Spotlight, IPinfoDB, *etc.*, for inferring the location of tweets. Though, compared to the other studies, the method enhances the median distance error, the average distance error is still too coarse to be considered useful in emergency response context. In addition, utilising multiple external sources for estimating the location of tweets seems too time-consuming, complex and labour intensive to be employed in time-sensitive scenarios.

Much of the work conducted on the location inference of Twitter data exploits either tweet content or one of the location-specific elements to infer the location. The purpose of this study is to explore a possible combination of different elements to predict the location of tweets that are present in a dataset. The proposed method evaluates a tweet against each potential location-specific element, to investigate the level of responsiveness of the tweet to each element. The method eventually predicts the location of the tweet, based on the best-fit element. Moreover, in terms of average distance error, existing works achieve either the city-level granularity or the average distance error of over 200 km at best, which are too coarse and not sufficiently detailed for the emergency response domain. Thus, methods need to be developed to reach a more detailed and finer granular level. The proposed method in this study estimates the location of 87% of the sample tweets with the average distance error of 12.2 km and the median distance error of 4.5 km, which is considered to be a significant improvement compared to the current methods.

3. Twitter Data and Location-Specific Elements

Launched in 2006, Twitter is a free social networking and microblogging service that allows users to post real time messages, called tweets. Tweets are short messages that are restricted to 140 characters in length. However, a tweet is more than a short message. Tweets come bundled with a relatively rich set of metadata. Through the streaming API, subsets of public status descriptions can be retrieved based on user-defined criteria in JavaScript Object Notation (JSON) formatted data which is a lightweight and text-based data exchange format. Figure 1 shows a raw Twitter feed presented in indented JSON format to facilitate reading, as well as understanding thereof.

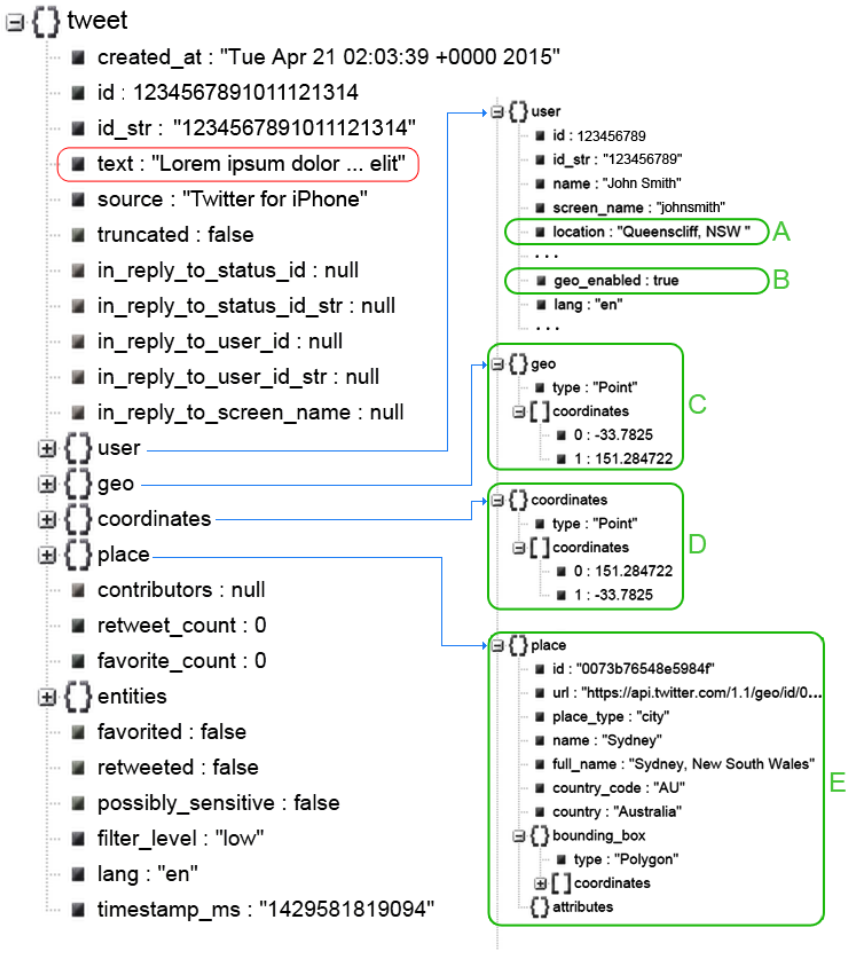

Figure 1. A tweet in indented JSON Format.

What is generally known as a tweet constitutes just one part of a whole feed and is accommodated within the "text" element. This element is shown within the red box in Figure 1. As it is clearly seen in the figure, there are a variety of elements accompanying the "text" element in a Twitter feed. It falls out of the scope of this article to describe all the elements, however, the location-related elements are introduced and discussed to address the main focus of the study. Based on what is shown in the figure above, apart from the "text" element, which may contain location references, there are location-specific elements that can have values of different types. These elements are highlighted in the green boxes labelled from A to E. The location-related elements and a brief description of each are listed in Table 1.

Table 1. List of location-related elements in a tweet.

Label	Element	Description
A	.\user\location	Nullable. The user-defined location for this account's profile. Not necessarily a location nor parsable.
B	.\user\geo_enabled	When true, indicates that the user has enabled the possibility of geotagging their Tweets. This field must be true for the current user to attach geographic data.
C	.\geo	Deprecated. Nullable. The "coordinates" field can be used instead.
D	.\coordinates	Nullable. Represents the geographic location of this Tweet as reported by the user or client application. The inner coordinates array is formatted as longitude first, then latitude.
E	.\place	Nullable. When present, indicates that the tweet is associated with (but not necessarily originated from) a Place.

Source: [27].

Among the location-related fields in Table 1, "geo" and "coordinates" correspond to geotagging, and both contain the same information [27]. Since the "coordinates" field is official and recommended by Twitter, this study uses the "coordinates" field where needed. There are also a few terms in Table 1 that need to be further clarified. "Nullable" means that a field does not necessarily contain a value and can be left blank. Most of the fields dealing with the user's settings are nullable fields, enabling the users to maintain some level of anonymity and privacy. Additionally, an "unparsable" field, like *user\location*, usually means that there might be unexpected entries in the field that are not compatible with the expected data type of the field. This is because there is no strict format for the *user\location*, and it can be anything that user writes down, for example "somewhere" or it might be null. Thus, if there is an entry, it is not necessarily a location name.

There is also another field within the "user" element called "geo_enabled". This field is the indication of whether a user has ever chosen to share any location information. If the "geo_enabled" field is true, it means that the user has agreed to turn on the location service at least once, but it does not necessarily indicate that the "coordinates" and "place" fields have values. This field is quite useful in location-related studies, and can be used to perform initial filtering of the tweets, even though it cannot provide any location information for inference purposes.

Users are also able to selectively attach a place name (such as a city or neighbourhood) of their choice to a tweet, by tapping the location marker and selecting the location they want to attach. "Places", from the Twitter data perspective, are specific and named locations with a few attributes that altogether are pushed into the "place" field and its immediate subfields. Tweets bound with places are not necessarily issued from that place, but are likely to be from within or around the place [27]. To investigate the current status of the Twitter data in relation to each of the location corresponding elements, an experimental study is carried out using a random sample of over 300 K tweets collected globally in April 2015. Figure 2 demonstrates the outcome of the analysis of the location-related elements.

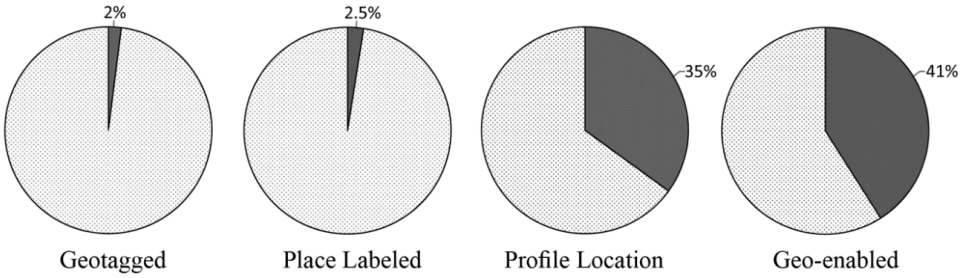

Figure 2. Current status of the location related elements of Twitter data.

Figure 2 shows that only 41% of users have agreed to share their location at least once, and 59% of users have never agreed or consented to share the location information in any way. It is revealed that only 35% of users have valid location information of various types, formats, geographic scales and languages in their profiles. In addition, 2.5% of all tweets are place-labelled, out of which 89% are at city-level granularity. Finally, as mentioned in the earlier section, only 2% of tweets are geotagged bundled with precise location coordinates. It seems needless to mention that the statistics provided here are on the average global scale, and the results may vary depending on the geographic resolution, time and how the data are collected.

4. Method Design and Development

Figure 3 outlines the design and architecture of the proposed method. It is seen in the figure below that the method is made up of three main components. The data preparation component mainly deals with the data collection and sampling processes. Following the data preparation phase, the event-related sample tweets go towards the location inference component that tries to predict the location from the potential location-related sources, which are explained in the previous section. The component needs location name classes as the input required for the location inference. The core function of this method is the location scoring and extraction function that assigns each tweet with the finest granular location extracted from the potential sources. The assignment of geocoordinates is the last step to be performed in the location inference processes. Finally, the result evaluation component, which is discussed in Section 5, compares the inferred location with the actual location of the sample tweets and calculates the distance error of the method. The rest of this section describes the data preparation and location inference components in detail.

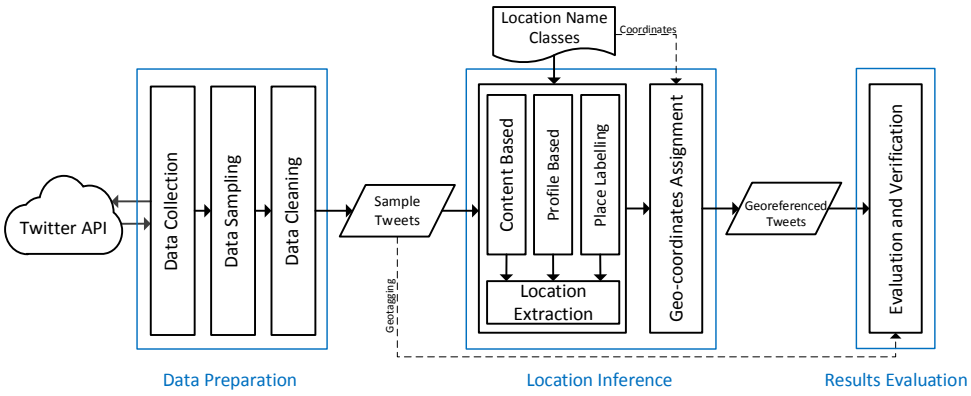

Figure 3. Overview of the method and its components.

4.1. Data Preparation

Before getting into the details of the location inference component, essential parts of the data preparation component are explained here. This includes the data collection and sampling processes

along with the data cleaning techniques and pre-processes that are important for performing the experiment smoothly.

4.1.1. Data Collection

In Twitter research, which can be generally characterised as data-driven, having access to appropriate Twitter datasets is crucial in order to validate theories and methods. Twitter data can be obtained either by purchasing from the commercial data vendors that Twitter partners with (e.g., Dataminr [28], Gnip [29] and Datasift [30], or cost-free collection through the Twitter Application Programming Interface (API), each with its own pros and cons. However, using freely available Twitter APIs seems more suitable for research purposes, where funds are strictly limited and multiple data collection efforts may be needed to gather the appropriate datasets. Among Twitter APIs, the streaming API, which provides low latency access to subsets of public tweets, is used to collect data from the area surrounded by a bounding box with the bottom-left corner at (35.00°S, 150.00°E), and the top-right corner at (32.00°S, 153.00°E), as shown in Figure 4.

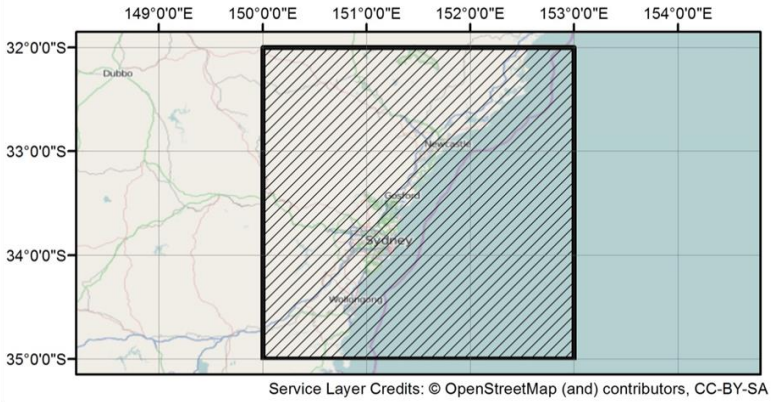

Figure 4. Data collection area.

The area includes Sydney, as well as the major regional centres of New South Wales. The data collection was carried out from 12:00 p.m., Tuesday, 21 April 2015, up to 11:59 p.m., Sunday, 26 April 2015 during which heavy rainfalls and occasional hailstorms struck Sydney and surrounding areas and caused dozens of floods across the region. During this period, 90,078 unique and non-retweeted tweets were collected and stored in a local database. These severe weather conditions are reflected in Bureau of Meteorology [11], the April 2015 issue of the Australia Monthly Weather.

4.1.2. Data Sampling

In order to create a reasonably sized dataset to examine the performance of the proposed method, a procedure is used to obtain a sample of tweets from the dataset of about 90 K collected tweets. In the very first step, non-English tweets are filtered out from the local database. This is because the method is designed to find the English location references within the location-related elements of a tweet and presence of the tweets in other languages (e.g., Arabic or Chinese) may result in impracticability of the method. The second step of the sampling is to find tweets that are assumed to be related to the observed severe weather conditions. To achieve this, a keyword search is performed on the contents of the tweets using the hailstorm and flood-related terms such as “storm”, “hail” and “flood”. As the result of the keyword filtering, over 3000 corresponding tweets are retrieved form the collected tweets.

There are many accounts from automated tweet ‘bots’ with tens of hourly tweets, most of them identical and high likely to be for business or marketing purposes, which should be taken out of the sample data. Thus, in the next step, the tweets that are less likely to be sent by real users are targeted. To conduct this, “source” field of tweets is taken into account and only tweets sent from handheld

mobile devices (mobile phones and tablets) and web-clients are extracted. The assumption behind this is that phones and tablets are normally used as personal devices and are unsuitable for mass tweet dissemination. Additionally, based on the information provided by Twitter, the source value of "web" is used for tweets that are directly sent from the Twitter website [27], which only allows users to read and write tweets through a web browser. Thus, its usage as a tweet bot seems very unlikely.

In the final step, the remaining tweets in which the "coordinates" field is non-null and has a value, are sent to the final sample dataset. The "coordinates" filed, as discussed in Section 3, represents geotagging information and is used for the evaluation and accuracy assessment of the proposed method in Section 6. Conducting the sampling procedure results in creation of the sample of this study, which contains 2409 unique and geotagged tweets in English, which are likely to be related to severe weather conditions and sent by real human users. Figure 5 shows the entire sampling procedure.

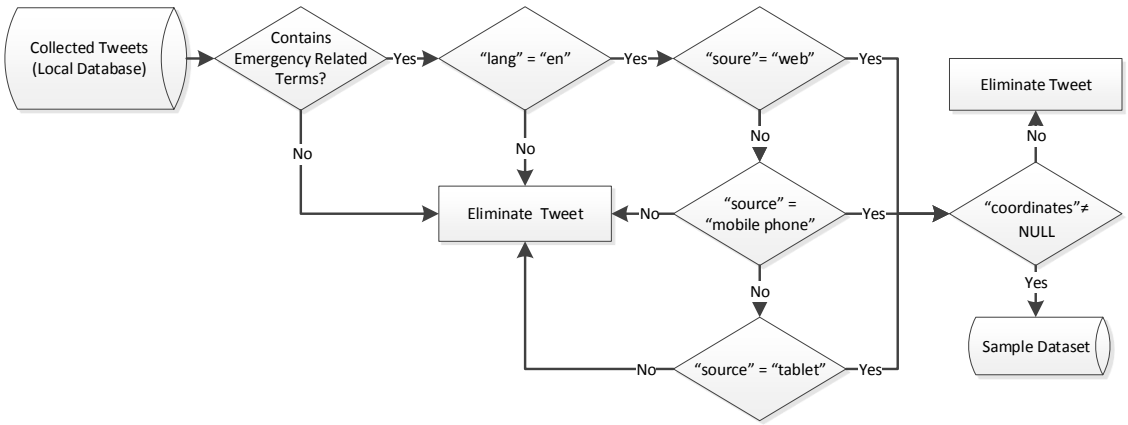

Figure 5. Data sampling procedure.

4.1.3. Data Cleaning

Some elements of Twitter data represent user-created information (e.g., text and user profile location) and are highly prone to different types of noise and redundancy. For example, there are huge numbers of emoticons, user mentions and Internet links within the text field, which may result in slow and inefficient performance of the method. A cleaning process, as the pre-processing step, should be performed to achieve a uniform textual content on user-created fields. In order to clean text and user profile location fields, all the following elements are first removed:

- Multiple dots " . . . " which people use in a variety of situations (replaced by a single space).
- User mentions (@somebody).
- Hashtag signs (#) from the beginning of all hashtag words.
- All the punctuation marks, numbers and Internet links (starting with "http://").

After removing the mentioned elements, all the characters are converted to lower case. The lower case conversion helps assessment of the location references carrying the same value with either upper or lower case forms. Following the lower case conversion, all probable multi-spaces are merged into a single space. The clean-up process is completed through standardising the text, by removing non-ASCII characters (like ä, £, 質). The cleaning process is applied on both text and user profile location fields within the sample dataset.

4.2. Location Inference

The location inference component deals with the extraction of predefined location references from each of three possible sources: textual content, user profile location and place labels. The predefined

sets of location name references are named "location name classes" and are described in the following subsection.

4.2.1. Location Name Class

To define the location name classes, this study partially uses the GIS shapefiles provided by the Australian Bureau of Statistics (ABS) [31], which are free and publicly accessible. Considering the availability of reliable data, location names are divided into three different levels of granularity. These levels, where each represents a class, are divided into three groups:

1. **Suburb level:** Suburbs that are partially or totally within the data collection zone are selected. To identify the suburbs, the suburbs polygon shapefile downloaded from the ABS website is intersected with the data collection zone (Figure 6). 1381 suburbs are selected and the name field of these suburbs represents the suburb-level name class (L_1). The geographic centroid of the selected suburbs is calculated in a GIS environment. The coordinates of the centroid are considered to be the location of the corresponding suburb.
2. **City level:** The main cities within the data collection zone are identified to constitute the city-level name class (L_2). The coordinates of these cities are extracted from Google Maps and attached to the related name class.
3. **Administrative level:** The names of large-scale administrative areas (state or country) in any possible forms (NSW, New South Wales, Australia, Aus and OZ) surrounding the data collection zone are considered to shape the administrative name class (L_3). As they are too large to be represented as a single location point, geographic coordinates at this level are not calculated.

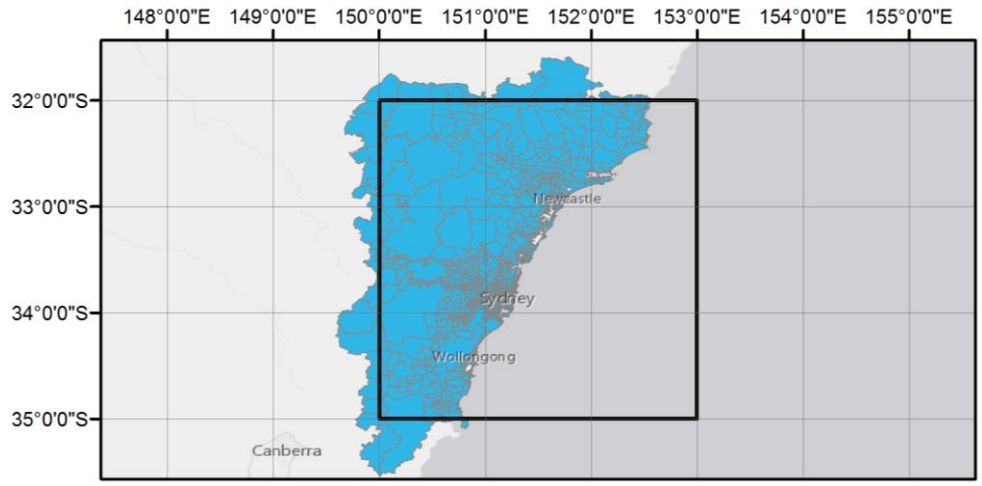

Figure 6. Suburbs intersected with the data collection zone.

4.2.2. Location Scoring and Assignment

As evident in Figure 3, the location inference component exploits three main sources: textual content, profile location and place labels. Each of the mentioned sources is checked against location name classes to investigate whether it corresponds to any location name within one of the location-name classes or not. To formulate this:

Let,

- $text.d_i$ be the textual content of a tweet d_i
- $profile.d_i$ be the profile location field of a tweet d_i
- $place.d_i$ be the place label field of a tweet d_i
- L_j be a location-name class

Then, a matrix representation of any relationship between the content of a tweet $text.d_i$ and class L_j can be shown as:

$$M_{con} = \begin{array}{c} \\ text.d_1 \\ text.d_2 \\ \vdots \\ text.d_i \end{array} \begin{array}{c} \begin{array}{ccc} L_1 & L_2 & L_3 \end{array} \\ \begin{bmatrix} f_{1,1} & f_{1,2} & f_{1,3} \\ f_{2,1} & f_{2,2} & f_{2,3} \\ \vdots & \vdots & \vdots \\ f_{i,1} & f_{i,2} & f_{i,3} \end{bmatrix} \end{array} \quad (1)$$

where $f_{i,j}$ is a location name, which is observed in both $text.d_i$ and L_j and can be defined as below:

$$f_{i,j} = \begin{cases} x, \ if \ \exists \ x \in text.d_i \mid x \in L_j \\ null, \ otherwise \end{cases} \quad (2)$$

In the equation above, when there are multiple instances of the location names in $text.d_i$, which belong to the same location name class (e.g., multiple suburb names), only the first instance will be assigned to x.

Having M_{con} constructed, the content-based location extraction is performed based on the following *IF* statement, which assigns the location name of the finest granularity as the content-based location of the tweet d_i through function F.

$$\begin{array}{l} IF \ \left(f_{i,1} \neq null\right) \ THEN \\ F\left(text.d_i\right) = f_{i,1} \\ ELSE \ IF \ \left(f_{i,2} \ \neq null\right) \ THEN \\ F\left(text.d_i\right) = f_{i,2} \\ ELSE \ IF \ \left(f_{i,3} \ \neq null\right) \ THEN \\ F\left(text.d_i\right) = f_{i,3} \\ ELSE \\ F\left(text.d_i\right) = null \\ END \ IF \end{array} \quad (3)$$

$F\left(text.d_i\right)$ can have a null value if there is no matching location name observed in the content. Exactly the same process is performed on the profile location field ($profile.d_i$), as well as place label field ($place.d_i$) and as a result, $F\left(profile.di\right)$ and $F\ \left(place.di\right)$, representing the finest granularity level of each field, are identified for all the sample tweets. As the output of this stage, each tweet is assigned new fields containing the values extracted for $\left(text.d_i\right)$, $F\left(profile.d_i\right)$ and $F\left(place.d_i\right)$ along with the class ID that the value belongs to. Following this step, each tweet should be assigned with one location only, and a decision should be made on which extracted field is the most suitable to be used as the final location of a tweet. For this purpose, a rule is defined as follows:

- Final location of a tweet is the extracted field that belongs to the finest granular level.
- If there is more than one field belonging to the same granular level, the final location is assigned based on the following order of importance:

    - Content-based location $F\left(text.d_i\right)$
    - Place-labelled based location $F\left(place.d_i\right)$
    - Profile-based location $F\left(profile.d_i\right)$

The reason behind the second rule is that the location references in both the text and place labelling are generated at the time of creation of a tweet, and are likely to be in connection with the topic of the tweet. They are also much more current than user profile location, which is likely to be generated at

the time of Twitter account opening. Moreover, the content-based location is considered to be more related and more detailed than place labelling, which is mostly used to assign broad and general place names (cities). After all, if the method is unable to find any location references that match the location name classes, or if there are no location references found within the location-related elements, it simply returns NA (Not Applicable) to indicate that the method is unable to infer the location of that specific tweet. Following the rule above, each tweet is assigned a location name from the corresponding location name class. After assigning each sample tweet with a location name, the coordinates of the centroid of the inferred location (calculated in Section 4.2.1) are allocated to that tweet. The next section reports the results of the implementation of the method.

5. Results and Evaluation

The method described in the previous section is applied to 2409 sample tweets. Table 2 shows a few examples of sample tweets after the execution of the method.

Table 2. Results of the application of the method on sample tweets.

No.	Tweet ID	Source	Location Name Class	Inferred Location			Actual Location		Distance Error (KM)
				Location Name	Latitude	Longitude	Latitude	Longitude	
1	590334736905572352	Place	L1	Brighton-Le-Sands	−33.9583	151.1536	−33.9697	151.1367	2.0105
2	590335052610936833	Text	L1	Manly	−33.8042	151.2905	−33.7825	151.2847	2.4746
6	590338256392323072	Profile Location	L1	Sunshine	−33.1121	151.5619	−32.9252	151.7733	28.6381
7	590338761805930498	Place	L2	Newcastle	−32.9167	151.7500	−32.9242	151.7470	0.8836
8	590338765140332544	-	-	NA	NA	NA	−33.9194	151.2526	und
9	590339563270184962	Text	L1	Sydenham	−33.9167	151.1680	−33.9482	151.1401	4.3454
10	590341916333629441	Text	L1	Broke	−32.7681	151.0883	−33.9174	151.2310	128.4835
11	590342183258968064	Place	L2	Central Coast	−33.2992	151.1922	−33.3722	151.4796	27.9043
14	590351290875518976	-	-	NA	NA	NA	−31.8964	152.4614	und
15	590351614130528256	Text	L1	Bulahdelah	−32.3868	152.1530	−32.0242	152.4728	50.3128
16	592169625951014912	Text	L1	Petersham	−33.8946	151.1549	−33.8963	151.1535	0.2267
17	592172876750409728	Text	L1	Wyong	−33.2778	151.4374	−33.2688	151.4343	1.0432
18	592184074103566338	Profile Location	L1	Manly	−33.8042	151.2905	−33.7744	151.2929	3.3290
20	592212076082405377	Text	L1	Petersham	−33.8946	151.1549	−33.8964	151.1532	0.2534
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2399	592217778284814338	Text	L1	Rhodes	−33.8292	151.0877	−33.8870	151.1791	10.6023
2400	592218784267567104	Profile Location	L1	The Entrance	−33.3450	151.4957	−33.3384	151.4958	0.7343
2401	592221934223368192	Text	L1	Manly	−33.8042	151.2905	−33.7679	151.1065	17.4738
2402	592204745135108096	Place	L3	New South Wales	NA	NA	−30.8144	152.5375	und
2403	592228305371172864	Place	L2	Sydney	−33.8651	151.2099	−33.7191	150.8924	33.5307
2404	592228457842544640	Place	L2	Newcastle	−32.9167	151.7500	−32.9340	151.7250	3.0236
2405	592263688632963072	Text	L1	Rooty Hill	−33.7733	150.8401	−33.8580	151.0340	20.2390
2406	592267839433637888	Place	L2	Sydney	−33.8651	151.2099	−33.8663	151.0465	15.0855
2407	592269376172109824	Text	L1	Rosebery	−33.9189	151.2048	−33.7890	151.0849	18.1999
2408	592281403053764608	Text	L1	Rhodes	−33.8292	151.0877	−33.8866	151.1787	10.5456
2409	592283790472445952	Text	L1	Maroubra	−33.9440	151.2443	−33.9556	151.2249	2.2034

In the table above, the "Tweet ID" field is the unique identifier of a tweet assigned by Twitter. The "source" field indicates the location-related element, which is determined as the suitable element for location inference by the method. The "Location Name Class" field indicates the corresponding location name class, from which a location name is assigned to each sample tweet. The "Inferred Location" field and its subfields ("Location Name", "Latitude" and "Longitude") show the name and geocoordinates of the inferred location. In addition, as mentioned in Section 4.1.2, only geotagged tweets are chosen to be in the sample dataset. This means that each sample tweet has the geotagging information (in the form of longitude and latitude) nested within the "coordinates" element of the tweet. The geotagged coordinates of the sample tweets are assumed as the actual location of the tweets and are shown in the "Actual Location" field. The "Distance Error" field, which is discussed later in this section, denotes the distance between the actual location and the inferred location of a tweet. This field is used as the evaluation metric to measure the accuracy of the results.

As it can be observed in the table above, the records highlighted in yellow show examples of the records in which the "Location Name", "Latitude" and "Longitude" subfields are marked as NA (Not Applicable). This means that the method was unable to allocate geocoordinates to those tweets, either

because there were no matching location names within location name classes or there were no location references cited within the potential sources. Moreover, as marked in cyan in Table 2, the method may return NA for only "Latitude" and "Longitude" subfields if the inferred location belongs to the administrative level (L_3), which is considered to be too large and thus inappropriate to be represented by an assigned point coordinates. A detailed analysis of the results shows that the method was unable to infer and assign the geocoordinates to 312 sample tweets. This implies that the method failed to infer the location coordinates of 312 (out of 2409) sample tweets due to the discussed reasons. For the rest of the sample dataset, which includes 2097 tweets, the proposed method was able to successfully infer the location name and allocate the matching geocoordinates to each tweet. This indicates a success rate of 87% for the proposed method in terms of inferring the location of the sample tweets.

In order to further evaluate the performance of the method within the location inferred tweets (87%), an evaluation metric is defined to measure the accuracy of the results. Accuracy in the location inference context is defined as the distance between the inferred location obtained from the localisation attempt and the actual location in the physical space [32]. Accuracy, from the perspective of location inference techniques, can be referred to as distance error. Zekavat and Buehrer [32] argue that the average distance error can be adopted as the performance metric for the evaluation of the location inference and localisation techniques.

To evaluate the accuracy of the method, the distance between the inferred geocoordinates and the geocoordinates of the actual location of the tweets is calculated using the "Haversine" formula [33]. This formula calculates the great-circle distance as the shortest distance between two points based on the given coordinates. For instance, let's assume that there are two points as $P_1 = (\phi_1, \lambda_1)$ and $P_2 = (\phi_2, \lambda_2)$, then the distance between these two points can be calculated using the following formula:

$$d = 2r\arcsin\left(\sqrt{sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1) \times \cos(\phi_1) \times sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right) \quad (4)$$

where r is the radius of the sphere, which is approximately equal to 6372 km.

Using Equation (4), the distance between the inferred and actual locations of the sample tweets is calculated and shown in the "Distance Error" field in Table 2. The function which applies the formula returns NA where the inferred geocoordinates have no valid values due to the aforementioned reason. The results indicate that the distance error ranges from as little as 0.11 km to as much as 177.6 km. Figure 7 shows the distance error for the sample tweets in 10 km intervals. The sample tweets for which the distance error is indeterminable are marked as NA in the figure.

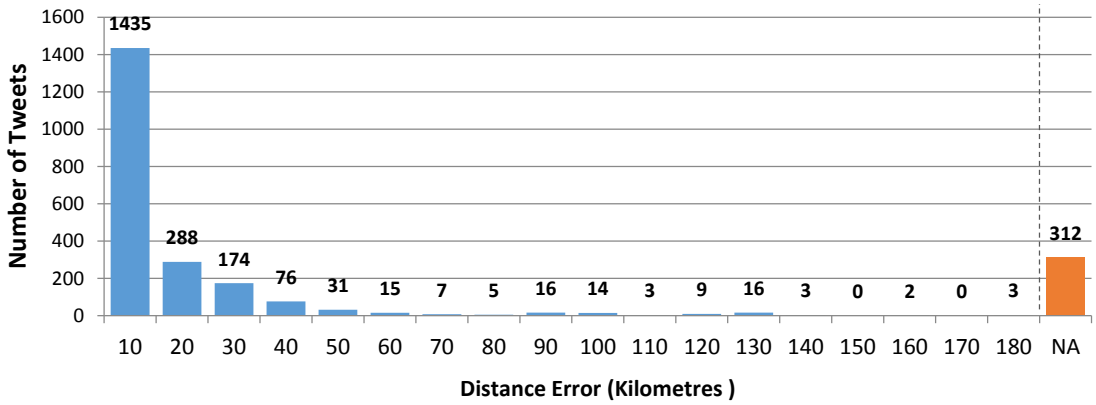

Figure 7. Distance error based distribution of the location inferred tweets.

It is evident from Figure 7, that for 1435 tweets (60%) out of 2409 sample tweets, the inferred location was at a distance equal to or smaller than 10 km from their actual location. In addition, it can be seen that 569 tweets were located within proximity of 10 to 50 km of their actual location.

Among the remaining tweets, the location of 93 tweets was inferred, with an accuracy of 50 to 180 km, and finally, the distance error of 312 tweets remains undeterminable due to the inability of the method to infer their location. Figure 8 shows the accuracy of the method based on the percentage of sample tweets falling within different ranges of the distance error (DE) metric.

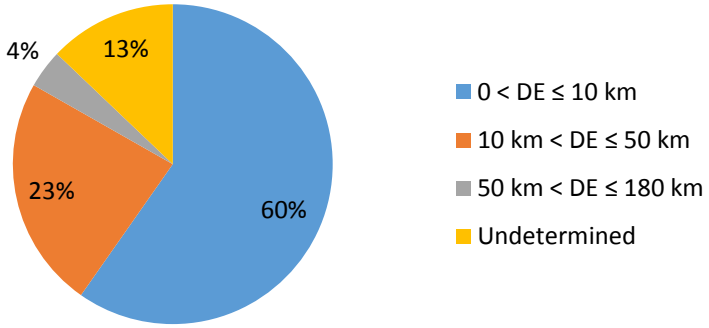

Figure 8. Accuracy of the location inference method based on distance error (DE).

To evaluate overall performance of the method, the average distance error can be calculated as the mean value of the calculated distance errors for 2097 tweets, for which the method was able to successfully perform the location inference. Putting the undetermined tweets aside, the method infers the location of 87% of the sample tweets, with the average distance error of 12.2 km, which, compared to the current state-of-the-art methods, can be viewed as a significant improvement over the current location inference methods.

6. Discussion, Conclusions and Future Work

Twitter has shown potential to be an effective tool in disseminating and obtaining up-to-the-minute information about real-world incidents. However, there are significant issues and problems in ensuring the quality and reliability of Twitter data for emergency response. Currently, being less than 2% geotagged, the location inference of Twitter data is one of the notable challenges. To give insight into Twitter data and to suggest possible solutions, the study provides a detailed investigation into the location-related elements of Twitter data. Getting to know the nature of Twitter data and utilising methods to deal with it, by itself, is an essential knowledge area. Therefore, a state-of-the-art description of location-related elements, as well as providing the overall current status of each element through practical studies, can be considered as the first contribution of this paper.

This study also proposes a multi-elemental location inference method, which uses three probable sources of location information and attempts to infer the location of tweets based on these elements. As far as authors are aware, the proposed location inference method is the first of its kind, which considers all the possible elements of a tweet through scoring and ranking algorithms, to achieve and predict the finest level of location granularity. In addition, in terms of the performance and accuracy of the proposed method, it was able to successfully infer the location of 87% of the sample tweets with an average distance error of 12.2 km and the median distance error of 4.5 km. This is a significant improvement compared with that of the current methods in the literature, which can predict the location either with a much larger average and median distance prediction error of 200 km and 30 km, respectively. This study, however, presents limitations that should be acknowledged. These limitations, at the current stage, include but may not be limited to the following:

- When there are multiple location references belonging to the same location name class within a location-related element (e.g., tweet text), the method only detects the first instance and ignores the others. A more detailed investigation of a selected number of tweets shows that about 1% of tweets may have multiple location references of the same class (e.g., multiple suburb

names), which are most likely to be neighbouring and adjacent. Even though this amount can be considered negligible without significantly affecting the performance and accuracy of the method, future developments of the method should include a more sophisticated handling of such cases.
- The method is not able to appropriately cope with the location references that might be found in the location-related element in a tweet but are not present in the location name classes. Resolving this issue in the future can increase the overall success rate of the method.
- The method is programmed to be applied to English tweets and may not be applicable on Non-English languages, especially the languages that use non-ASCII characters (e.g., Arabic and Chinese).

The study does not end here. Implementing the method for different types of datasets related to various kinds of incidents (e.g., bushfire, earthquake and terrorist attacks), along with the steps required for overcoming the above-mentioned limitations of the study, shape a future research direction that the authors wish to follow. Furthermore, a deeper investigation of results, focusing on the location-related elements with an aim to fine-tune the method, can be considered as other future work in this line.

Acknowledgments: This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors. The authors would like to take the opportunity to thank the CDMPS team members, who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations and conclusions of this paper.

Author Contributions: Farhad Laylavi designed the method, performed analysis, interpreted data, wrote manuscript and acted as corresponding author. Abbas Rajabifard and Mohsen Kalantari supervised development of work, reviewed and edited the manuscript and helped in data interpretation and method evaluation.

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

Abbreviations

The following abbreviations are used in this manuscript:

API	Application Programming Interface
ASCII	American Standard Code for Information Interchange
CDMPS	Centre for Disaster Management and Public Safety
GPS	Global Positioning System
JSON	JavaScript Object Notation

References

1. BBC. How the Paris Attacks Unfolded on Social Media. Available online: http://www.bbc.com/news/blogs-trending-348 36214 (accessed on 23 November 2015).
2. South, J.A. Interactive Emergency Information and Identification Systems and Methods. U.S. Patent 20,150,111,524, 23 April 2015.
3. Steiger, E.; Albuquerque, J.P.; Zipf, A. An advanced systematic literature review on spatiotemporal analyses of twitter data. In *Transactions in GIS*; Wiley Online Library: Hoboken, NJ, USA, 2015; pp. 809–834.
4. Williams, S.A.; Terras, M.M.; Warwick, C. What do people study when they study twitter? Classifying twitter related academic papers. *J. Doc.* **2013**, *69*, 384–410. [CrossRef]
5. Heinzelman, J.; Waters, C. *Crowdsourcing Crisis Information in Disaster-Affected Haiti*; US Institute of Peace Press: Washington, DC, USA, 2010.
6. Mansourian, A.; Rajabifard, A.; Valadan Zoej, M.J.; Williamson, I. Using SDI and web-based system to facilitate disaster management. *Comput. Geosci.* **2006**, *32*, 303–315. [CrossRef]
7. Poser, K.; Dransch, D. Volunteered geographic information for disaster management with application to rapid flood damage estimation. *Geomatica* **2010**, *64*, 89–98.
8. Twitter. Twitter Blog: Location, Location, Location. Available online: https://blog.twitter.com/2009/location-location-location (accessed on 12 October 2015).

9. Cheng, Z.; Caverlee, J.; Lee, K. You are where you tweet: A content-based approach to geo-locating twitter users. In Proceedings of the 19th ACM International Conference on Information and Knowledge Management, Toronto, ON, Canada, 26–30 October 2010; pp. 759–768.
10. Morstatter, F.; Pfeffer, J.; Liu, H.; Carley, K.M. *Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose*; Cornell University arXiv: Ithaca, NY, USA, 2013.
11. Bureau of Meteorology. Monthly Weather Review Australia April 2015. Available online: http://www.bom.gov.au/climat e/mwr/aus/mwr-aus-201504.pdf (accessed on 21 October 2015).
12. Paul, M.J.; Dredze, M. You are what you tweet: Analyzing twitter for public health. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, Spain, 17–21 July 2011.
13. Ciulla, F.; Mocanu, D.; Baronchelli, A.; Gonçalves, B.; Perra, N.; Vespignani, A. Beating the news using social media: The case study of American Idol. *EPJ Data Sci.* **2012**, *1*, 1–11. [CrossRef]
14. Skoric, M.; Poor, N.; Achananuparp, P.; Lim, E.-P.; Jiang, J. Tweets and votes: A study of the 2011 Singapore general election. In Proceedings of the 45th Hawaii International Conference on System Science (HICSS), Maui, HI, USA, 4–7 January 2012; pp. 2583–2591.
15. Oku, K.; Ueno, K.; Hattori, F. Mapping geotagged tweets to tourist spots for recommender systems. In Proceedings of the IIAI 3rd International Conference on Advanced Applied Informatics (IIAIAAI), Kitakyushu, Japan, 31 August–4 September 2014; pp. 789–794.
16. Sakaki, T.; Okazaki, M.; Matsuo, Y. Earthquake shakes twitter users: Real-Time event detection by social sensors. In Proceedings of the 19th International Conference on World Wide Web, Raleigh, NC, USA, 26–30 April 2010; pp. 851–860.
17. Ajao, O.; Hong, J.; Liu, W. A survey of location inference techniques on twitter. *J. Inf. Sci.* **2015**, *41*, 855–864. [CrossRef]
18. Eisenstein, J.; O'Connor, B.; Smith, N.A.; Xing, E.P. A latent variable model for geographic lexical variation. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA, 27–29 July 2010; pp. 1277–1287.
19. Wing, B.P.; Baldridge, J. Simple supervised document geolocation with geodesic grids. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 19–24 June 2011; pp. 955–964.
20. Watanabe, K.; Ochi, M.; Okabe, M.; Onai, R. Jasmine: A real-time local-event detection system based on geolocation information propagated to microblogs. In Proceedings of the 20th ACM International Conference on Information and Knowledge Management, Glasgow, Scotland, 24–28 October 2011; pp. 2541–2544.
21. Dalvi, N.; Kumar, R.; Pang, B. Object matching in tweets with spatial models. In Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, Seattle, WA, USA, 8–12 February 2012; pp. 43–52.
22. Han, B.; Cook, P.; Baldwin, T. Text-based twitter user geolocation prediction. *J. Artif. Intell. Res.* **2014**, *49*, 451–500.
23. Minot, A.S.; Heier, A.; King, D.; Simek, O.; Stanisha, N. Searching for twitter posts by location. In Proceedings of the 2015 International Conference on The Theory of Information Retrieval, Northampton, MA, USA, 27–30 September 2015; pp. 357–360.
24. Hecht, B.; Hong, L.; Suh, B.; Chi, E.H. Tweets from Justin Bieber's heart: The dynamics of the location field in user profiles. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Vancouver, BC, Canada, 7–12 May 2011; pp. 237–246.
25. Hiruta, S.; Yonezawa, T.; Jurmu, M.; Tokuda, H. Detection, classification and visualization of place-triggered geotagged tweets. In Proceedings of the 2012 ACM Conference on Ubiquitous Computing, Pittsburgh, PA, USA, 5–8 September 2012; pp. 956–963.
26. Schulz, A.; Hadjakos, A.; Paulheim, H.; Nachtwey, J.; Mühlhäuser, M. A multi-indicator approach for geolocalization of tweets. In Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, Cambridge, MA, USA, 8–11 July 2013.
27. Twitter. Twitter Developers Documentation. Available online: https://dev.twitter.com/overview/documentation (accessed on 21 October 2015).
28. Dataminr. Available online: https://www.dataminr.com/ (accessed on 16 January 2016).
29. GNIP. Available online: https://www.gnip.com/ (accessed on 16 January 2016).
30. DATASIFT. Available online: http://www.datasift.com/ (accessed on 16 January 2016).

31. Australian Bureau of Statistics. Available online: http://www.abs.gov.au/ (accessed on 16 January 2016).
32. Zekavat, R.; Buehrer, R.M. *Handbook of Position Location: Theory, Practice and Advances*; John Wiley & Sons: Hoboken, NJ, USA, 2011.
33. Rick, D. Deriving the Haversine Formula. Available online: http://mathforum.org/ library/drmath /view/51879.html (accessed on 16 January 2016).