Usability of small crisis data sets in the absence of big data

Avijit Paul1, Axel Bruns1

1 ARC Centre of Excellence for Creative Industries and Innovation, Queensland University of Technology, Brisbane, Australia
alpaul / abruns @ qut.edu.au – cdtaavijit / snurb_dot_info – http://mappingonlinepublics.net/

Abstract

Big data is big news in almost every sector including crisis communication. However, not everyone has access to big data and even if we have access to big data, we often do not have necessary tools to analyze and cross reference such a large data set. Therefore this paper looks at patterns in small data sets that we have ability to collect with our current tools to understand if we can find actionable information from what we already have. We have analyzed 164390 tweets collected during 2011 earthquake to find out what type of location specific information people mention in their tweet and when do they talk about that. Based on our analysis we find that even a small data set that has far less data than a big data set can be useful to find priority disaster specific areas quickly.

Keywords: Crisis Communication, Big Data, Twitter, Location

1. Introduction

A large body of present social network research in crisis communication uses certain methods such as Twitter hashtags to identify messages related to a specific natural disaster and find meaningful information out of that (3). The biggest limitation of this approach is, it only tracks tweets that has the particular hashtag. Therefore various research (2), (3) suggests use of Big data sets to in order to eliminate the issue and analyse all possible data available. However access to Big Data is often difficult due to many limitations such as financial and programming ability in analysing real time big data (1). Therefore this paper looks into existing data sets that are not as large as Big Data sets and are gathered with the existing limitations to identify if it is possible to utilise the existing datasets as an assistive method in crisis situation.

In order to understand the difference between the two data sets, first we define what is the Social Media Big Data and what is considered as small data sets. 

2. Difference Between Big Data and Small data

Before we explain further it is first important to understand what are the difference between Big data and small data and if the difference is extremely critical.

By definition, Big Data is a large data sets that has huge amount of data. A small data sets has a smaller number. To simplify the concept, if Big Data has millions of rows of data, a small data set will have thousands. However, as mentioned by Boyd and Crawford, (1) Big Data is not just large datasets, it is also the way we conduct research. By having the access to big data, a researcher can gain access to almost every conversation hap-
pening around the world in that medium. And with such access (5), a data scientist may have a higher probability to come up with an answer that is more statistically significant than the other who may not have such access.

However, the aim of this paper is to analyse a small dataset and find out if it can be useful enough at the time of crisis and therefore can be used to create a framework that can potentially be used in a big data set.

3. Contextual location mention

For the purpose of this research we will look at location information mentioned in tweets that has been archived at ARC Centre of Excellence for Creative Industries and Innovation, Queensland University of Technology during February 2011 New Zealand Earthquake. About 164390 tweets were collected following “eqnz” hashtag and we want to find out if people start to talk about the affected areas early enough in order to use that by disaster and relief authorities. Since the first 24 hours is the most crucial according to DCS (4), we will use that as the timeframe for this research.

As the research is done on archived data and the earthquake has already happened, the list of affected areas are well known. Therefore we took the dataset, divided them into minutes and hours and reverse searched the areas for the number of time a certain location has been mentioned in that particular hour or minute. We have then added the numbers to find out how frequently they were being mentioned.

As our datasets starts when the hashtag was already known, we see a large volume of tweets at the very beginning with a gradual decrease time passed by.

As it can be seen in figure 1, there is a sharp decrease in the amount of time most places were mentioned. And if we are to count the places that get most mentions, we find that in the first six hours (Fig. 1), most notable mentions are about the country (New Zealand and NZ as short form) as well the broad location area (Christchurch). This is mostly because a large chunk of messages are retweets of news media reports or messages that has already been in circulation. And according to this graph, in first 6 hours the highest mention was about the place where the earthquake hit - Christchurch and it is followed by NZ, short form of New Zealand.

However, if we take the first six hours (Fig. 2), we can see that people talked about the larger area first as they were just getting to know about the crisis and
they may or may not be familiar with smaller area such as Christchurch.

However within first hour, people have become more familiar with the word christchurch and started to tweet about that. Now in order for the data to be useful for disaster recovery, we may need to look for smaller location entries such as a certain road, hospital etc.

We can see that (Fig. 3) the Cathedral was mentioned the most at the beginning and it was one of the first places that was affected. The word hospital was mentioned heavily between 2 to 2.5 hours. There may be two reasons for such mentions. One is, people are looking for hospital to go to or suggestions that certain hospital is open or closed or is over capacity. Another is, a hospital is also hit at that hour. For the case of this earthquake, Christchurch hospital was getting partly evacuated at that hour due to damage in some areas.

We also see mention of airport in various tweets at different time. By reading the tweets we can find that most of the mentions were either because people were getting unconfirmed news that the airport is hit as well and later they found that airport was not damaged and can be used.

If we now eliminate the top two mentioned areas among the specific areas, we find that CTV or Canterbury Television Building was mentioned heavily after 18 hours of the earthquake. It also had around 200 mentioned in the first 6 hours.

Since 94 of 168 causalities recorded in Christchurch earthquake was from this building let us focus on this a little more.

From the collected tweets we can see that first half hour there is only one mention about the building (Fig. 4). The number of mentioned did not increase to a noticeable amount for next 2 hours. However starting from third hour more and more information about CTV building starts to appear in tweet stream.

4. Discussion

Based on the limited data collected, we can see that it is possible to identify at least bigger areas really quickly as they get mentioned very frequently. For example, the Cathedral is at the heart of Christchurch and therefore it was extremely well known for people in that area and was heavily mentioned. Furthermore, early footages (images and videos) also contained clips of destroyed Cathedral, which was then retweeted many times.

Furthermore, by locating the bigger area, one can identify that the specific mentioned area that may share common name with another place falls into the disaster
area and not in another city or country. For example, the keyword, Cathedral was mentioned 801 times in the first 6 hours. Although New Zealand has several other places with the same “Cathedral” name such as Cathedral Place in Auckland, Cathedral Court in Hahei, Cathedral Cove in Waikato, the mention of the Cathedral was for Cathedral Square in Christchurch. Therefore by using the frequent mentioned area as a filter we can pinpoint conversations related to smaller areas inside the crisis area.

However, the more specific or smaller areas are mentioned less frequently. Although if one observes the repeated mention of a certain location or specific areas, one can find that that is a potentially disaster stricken area – which was the case for CTV building.

However, further research is needed in this area to identify if there are other keywords that indicates location information such as use of “at” or “in” or other preposition as location names will not be available while the disaster is in progress. By using various other combinations it is potentially possible to find the mention of a location even if it is not reported in other medium.

5. Conclusion

Based on the analysis, we can suggest that in the absence of access to large data sets, if we are only looking for location information to find out which area requires more help, we can still find out names of the places that were hard hit during disaster. Although the small dataset we have at CCI were set up after the hashtag became popular, and therefore missed certain amount of information, it is still appears to be quite useful for location identification from tweets gathered using existing method.

6. References


