Tracking Gauteng thunderstorms using Crowdsourced Twitter data between Soweto and Pretoria

L Butgereit[1]

Abstract

Summer thunderstorms in Gauteng are often dramatic, noisy, wet events. They can appear suddenly on exceptionally hot sunny days travelling fast across the province. With such dramatic arrivals, people often flock to social media sites such as Twitter to comment on the rain, wind, hail, lightning and thunder. This paper investigates the possibility of mapping the track of Gauteng thunderstorms by using crowdsourced data from Twitter. This paper describes a model (entitled the ThunderChatter Model) and instantiation of that model which extracts data from Twitter, analyses the textual information for thunderstorm information and plots the appropriate data on a map. For evaluation purposes, these generated maps are then compared against lightning-stroke maps provided by the South African Weather Service. The maps are visually compared by independent people using Content Analysis techniques ensuring unbiased and reproducible results. The results of this research are mixed. For thunderstorms which traverse the strip of land between Soweto and Pretoria more or less correlated to the N1 highway (and representing the most heavily populated area of Gauteng and the area with the highest percentage of home Internet facilities), the results are excellent. However, in outlying areas of Gauteng such as Carletonville, Heidelberg, Hammanskraal and Bronkhorstspruit, the thunderstorms are only trackable using crowdsourced Twitter data in the case of extreme storms which damage property. The results imply that data obtained from social media could be used in some cases to supplement geographical data obtained from traditional sources.

Keywords:

Introduction

With evidence of hominid habitation stretching back millions of years, Gauteng province situated in the north-eastern part of South Africa accommodates over 10 million people (Mabin, 2013) and can justifiably be regarded as the economic heartland of South Africa (Kok, 1998). Gauteng is situated on what is called the *Highveld* of South Africa which is an inland plateau ranging from 1500 meters above sea level to 2100 meters above sea level.

Major natural disasters such as earthquakes, volcanoes and tsunamis do not affect Gauteng. Heat induced thunderstorms, on the other hand, are a common occurrence during the summer months (Archer et al., 2010). These thunderstorms are often dramatic with high

[1]. Dr Laurie Butgereit is a research associate at Nelson Mandela Metropolitan University where she investigates the uses of mobile technology in society. Email: Laurie.Butgereit@nmmu.ac.za

gusting wind (Kruger et al., 2010) accompanying frequent lightning, thunder (Bhavika, 2007) and hail (Carte & Held, 1978).

Although Gauteng only comprises 1.5% of South Africa's land area (Kok, 1998), it is home to over 20% of South Africa's population (Lehohla, 2012). With such a high population density, one would expect that these dramatic thunderstorms would receive comment in social media.

People discuss the weather for various reasons. It is a safe conversation topic which does not involve personal details or sensitive topics (Harley, 2003). In societies around the world, people talk about the current weather, the past weather and the future weather (Strauss & Orlove, 2003). Severe weather events decades past are often discussed. One could argue that the story of Noah and the Great Flood which appears in Judaism, Christianity and Islam is an example of society still talking about a weather event which occurred thousands of years ago.

Weather occasionally plays a role in helping people organize their autobiographical memory (Harley, 2003). Autobiographical memory can be defined as the memories a person has (Robinson & Rubin, 1986). The author can anecdotally confirm this idea of weather assisting in organizing autobiographical memory with personal memories of the “Christmas when it rained so much that the roads were all flooded and we had problems going Christmas shopping.”

This research initially investigated the possibility of monitoring these thunderstorms using crowdsourced data (known as *tweets*) from Twitter. People comment on Twitter as these dramatic thunderstorms pass through their neighbourhoods providing more up-to-date weather information than traditional weather news broadcasts. A model was created by this study which the author called the ThunderChatter Model. Thunderstorm tweets were extracted from the Twitter feed using a Twitter API (Application Programming Interface) and were plotted on a map of Gauteng. The map was then visually compared against lightning-stroke maps compiled by the South African Weather Service. During the course of this research, it became clear that some Gauteng thunderstorms were very easy to map using Twitter data and some were not easy to map. The reasons for this were investigated and are also presented in this paper.

This article will first provide some background about crowdsourcing and background about the types of thunderstorms which appear in Gauteng. After this background material is presented, a model is provided for mapping thunderstorm activity using crowdsourced Twitter data. The instantiation of the model is also described along with the model. The results were analysed using techniques of Content Analysis as defined by Krippendorf (1980) producing reproducible results free from author bias.

Previous attempts by the author (along with co-authors) to track other types of environmental phenomena such as veld fires and building fires in South Africa using Twitter was not particularly successful (Butgereit et al, 2014). This was due to the fact that the author and co-authors in that previous study did not take into account previous social research which shows people's willingness to talk about the weather conditions (as opposed to other types of environmental conditions) with total strangers. Additional previous research by the author in attempting to track general weather conditions in the city of Pretoria using Twitter was also not particularly successful (Butgereit, 2014). This was due to the fact that continuous, unchanging, boring weather (such as a week of hot weather) does not drive people to comment on social media. In that previous research the author did not take into account the nature of the different types of weather conditions and the dramatic nature of Gauteng thunderstorms.

As will be shown, this research spans a number of disciplines. Interdisciplinary research tries to integrate multiple perspectives in a coherent whole. Transdisciplinary research goes beyond interdisciplinary research and applies new insights and knowledge to the underlying problem (Kroeze & Van Zyl, 2014). This research is transdisciplinary in nature including aspects of social behaviour, weather sciences and information technology. It solves an underlying problem of mapping fast-moving weather conditions using social media.

Crowdsourcing

The coining of the term *crowdsourcing* is attributed to Jeff Howe, contributing editor to Wired Magazine, in his 2006 article entitiled *The Rise of Crowdsourcing* (Howe, 2006). The term was used to describe a phenomenon where companies which grew up in the Internet Age and were designed to take advantage of the networked world published an open request for content, assistance or work to be done. The request was open to the crowd of people on the Internet. Although Howe did not provide a specific definition, he paraphrased that it was *outsourcing to the masses*.

The definition has been refined by a number of authors. In her blog *What is Crowdsourcing?*, Jennifer Alsever explained that crowdsourcing has four distinct components: 1) the public at large 2) provides information 3) free of charge 4) to an organisation or company. Crowdsourcing taps the collective intelligence of the public at large to complete a task that would normally be done by a specific agent (Alsever, 2007).

Crowdsourcing websites are common. Websites such as Wikipedia and YouTube attract people and encourage them to create content using a crowdsourcing model. The Volunteer Match website (Volunteer Match, 2013) and Mobile Volunteering website (Mobile Volunteering, 2013) attract volunteers using a crowdsourcing model and encourage them to help solve social problems by volunteering their time and energy.

Twitter is an example of a crowdsourcing site. Twitter is a microblogging website where participants are free to publish 140 character status updates or *tweets* (Java et al., 2007; Kwak et al., 2010). People tweet about the activities in their personal lives. They also tweet about events they witness or experience including social unrest, sporting events, political transitions, weather emergencies and celebrity secrets. When events are changing quickly, the status updates on Twitter are often more up-to-date than traditional news broadcasts. For example, in Japan 96% of all earthquakes which are stronger than 3 on the JMA (Japan Meterological Agency) seismic intensity scale can be detected using Twitter (Sakaki et al., 2010). Even during times of on-going natural disasters, the nature of the public's tweets provide insight into the disaster as the disaster moves from preparation or warning stages (as with flooding with a long anticipation period as water levels rise) through to impact stages (Vieweg et al., 2010).

Why do people contribute to Crowdsourcing sites?

The reasons people contribute to crowdsourcing sites are often as varied as the sites themselves. Sites from which it is possible to crowdsource data are very different. Three common such sites are Wikipedia, Twitter (as was used in this research) and Facebook. The reasons people contribute to these three different sites are also different.

Wikipedia is an online encyclopedia which has been extremely successful in recent years. It is maintained by a cohort of volunteer *wikipedians* who update and edit content on the encylopedia (O'Sullivan, 2012). In a general study on why people volunteer, researchers have identified six motivational categories: values (volunteering gives people an opportunity to express their values), social (volunteering provides outlets for social activity), understanding (volunteering provides better understanding to situations), career (volunteering allows people to better their careers), protection (volunteering gives people the opportunity to protect people, ideas, etc) and enhancement (volunteering allows people to enhance situations) (Clary et al., 1998). To this group of six motivational factors, Nov (2007) adds two more motivational factors specifically related to volunteers at Wikipedia: fun and ideology. People contribute to Wikipedia because the contribution itself is fun and allows people to align with their ideology. Other researchers have added reasons such as altruism, reciprocity, community, reputation and autonomy (Kuznetsov, 2006).

However, the reasons people contribute to Facebook are dramatically different. Whereas contributors to Wikipedia are writing about other people and events, contributors to Facebook are typically writing about themselves. Previous research into why students contribute to Facebook have found that one of the primary reasons is to keep in touch with friends (Cheung et al., 2011). Personality traits such as extraversion and openness to experiences are also drivers behind Facebook usage (Ross et al., 2009).

And, the reasons people contribute to Twitter are also different from the two preceding examples. The research of Java et al. (2007) has shown that there are four different types of *tweets* which people post: daily chatter (general personal news), conversations (retweeting and replying), information sharing and reporting news. In addition, Java et al have found that a single user may have multiple intentions or reasons when posting tweets.

This research paper specifically deals with information drawn from Twitter, and, in general, the tweets fall into three of the four categories defined by Java et al (2007): daily chatter, information sharing and reporting news.

Thunderstorms

Heat induced thunderstorms start with convective ascent of air. As surface air is heated, it expands thereby becoming less dense than the surrounding air and rises. The upward motion of the air results in cooling and the eventual condensation of water vapour with associated formation of billowing clouds with extreme vertical height. When the upward motion of hot air can no longer support the weight of the condensed water droplets, the droplets fall to the earth as rain. In some cases, the falling droplets are lifted up even higher on the column of rising warm air and freeze eventually falling to the earth as hail. The turbulent nature inside the storm results in the water droplets having different electrical charges. As the droplets rise, they typically lose electrons resulting in positively charged upper regions of the storm. The larger particles and droplets in the lower portions of the storm become negatively charged. This results in lightning and thunder. The thunderstorms that ensue may be of a variety of types ranging from single-cell, to multi-cell, to super-cell storms (Tyson & Preston-Whyte, 2000).

Single-cell thunderstorms are typically 5-10 kilometres in horizontal extent, are short-lived (commonly less than 60 minutes) and change markedly over time. In the highveld area (which includes Gauteng), single-cell thunderstorms are generally 18-30 minutes in duration. The storms customarily travel at speeds of about 25 kilometres per hour from the south-west to the north-east (Tyson & Preston-Whyte, 2000).

Multi-cell thunderstorms consist of a sequence of evolving single-cells each of which goes through the life cycle of a single-cell thunderstorm. Multi-cell thunderstorms are typically 30-50 kilometres in horizontal extent. Multi-cell thunderstorms are common over Gauteng moving generally from the south-west to the north-east with new single-cells usually joining on the front side of the storm and old single-cells dissipating at the trailing edge of the storm. One documented storm travelled for 180 kilometres over a four hour period. Up to 30 individual cells may take part in a multi-cell thunderstorm (Tyson & Preston-Whyte, 2000).

Super-cell thunderstorms are not as frequent as single-cell and multi-cell storms. When they do occur, however, they usually cause havoc and devastation. They may extend 20-30 kilometres in horizontal extent and 12-15 kilometres in vertical dimension. One documented super-cell storm lasted for more than four hours and tracked for over 100 kilometres (Tyson & Preston-Whyte, 2000).

Gauteng province experiences these heat induced thunderstorms during the summer months (Archer et al., 2010). The storms normally start to form in the early afternoon and the resulting storm is in the later afternoon or evening (Tyson & Preston-Whyte, 2000). In the Johannesburg-Pretoria areas of Gauteng, hail occurs on average 69 days of the year ranging in size from one cm to, in extreme cases, the size of tennis balls (Tyson & Preston-Whyte, 2000). Rain can often fall at rates higher than 10mm per hour. Thunder from the associated lightning can rattle loose objects far from the actual storm.

Research methodology

A Design Science Research methodology was initially adopted for this project. Design Science Research (DSR) has as one of its fundamental pillars the requirement that an innovative artefact must be created to solve an important problem in an innovative manner (Vaishnavi & Kuechler, 2007). Hevner and Chatterjee define Design Science Research as a research paradigm in which the designer "...answers questions relevant to human problems via the creation of innovative artifacts".

There are five possible outputs from Design Science Research. The created artefact can be a construct, a model, a method, an instantiation, or a better design theory (Hevner & Chatterjee, 2010). Constructs define the conceptual vocabulary of a problem domain. Models are propositions or statements about the inter-relationship among constructs. Methods are sets of steps or algorithms used to perform a task. Instantiations are the realisations of constructs, models or methods in the real world. Better design theories, the fifth possible output of Design Science Research, can be created when an artefact exposes relationships between its elements which were previously unknown.

The General Design Cycle (GDC) of DSR is defined as having five steps (Vaishnavi & Kuechler, 2007; Oates, 2006):

- Awareness – the recognition of a problem and the statement thereof.
- Suggestion – the offering of tentative ideas of how to solve the problem.
- Development – the implementation of the aforementioned suggestions.
- Evaluation – the examination of the developed artefacts.
- Conclusion – the consolidation of results.

Each of these five steps has a defined output. The awareness step creates as an output a research proposal. The suggestion step creates as an output a tentative design. The development step creates as an output an artefact. The evaluation step creates output performance measurements. The conclusion step creates as an output research results (Vaishnavi & Kuechler, 2007).

The General Design Cycle is appropriately called a *cycle* which is iterated over a number of times. In the case of this research, the GDC had the following steps:

1. Awareness – Awareness that often Twitter has more up-to-date weather information online than the official weather reports and that the geolocation of these tweets could be used to map the weather information. This step created a proposal for the remaining research.
2. Suggestion – Suggestions on how these tweets and their geolocations could be analysed using various text processing techniques and subsequently mapped. This step created a tentative design for the artefact.
3. Development – The suggestions were implemented in an iterative manner with steps #2, #3 and #4 being cycled through numerous times. Appropriate tweets were accessed and the weather summary maps were generated. This step created the DSR artefact.
4. Evaluation – The weather summary maps generated in #3 above were evaluated. This evaluation was a comparison with actual lightning-stroke data provided by the South African Weather Service. Until such time that the summaries were satisfactory, steps #2, #3 and #4 were cycled through again. This step created results of the evaluation of the artefact.
5. Conclusion – The final step created the final results which were summarised for the scope of this research.

This research involved the creation of a model (called the ThunderChatter Model) to map the track of Gauteng thunderstorms and the instantiation of that model as a prototype artefact.

The ThunderChatter model

The ThunderChatter Model is specifically designed to attempt to recognise thunderstorms using Twitter data and to plot that information on a map. In order to attempt to map thunderstorms using crowdsourced data from Twitter, the ThunderChatter Model has three major steps. These three steps are:

- Extract information from Twitter by using numerous query searches with and without geolocation information. It is noted that some of the results of the query searches will be unrelated to thunderstorms and will be filtered out in the next step.
- Analyse the extracted data using various text processing utilities to determine if they indicate thunderstorms.
- Map the data depending on the geolocation of the search queries or of the results.

Each of these steps will be discussed in a separate section along with the actual instantiation of that particular step.

ThunderChatter: Step one – Extract information

The first step in the ThunderChatter Model requires that sufficient information is extracted from Twitter. People do not conveniently use a #thunderstorm tag on tweets. Having said that, however, there has been previous research where weather services specifically asked people to tweet weather reports with a specific hashtag (Brice & Pieper, 2009). The research reported by Brice and Pieper is specifically different from the current research reported in this paper. The current research does not restrict users in this respect. During steps #2, #3 and #4 of the General Design Cycle various keywords were suggested as search keys, the searches were implemented and the results were evaluated. Search terms such as "highveld storm", #heat, #storm, "storm is coming", "crazy storm", "hail", "thunder" and "lightning" were used in an attempt to find the keywords which produced the best results.

During these experimental stages (steps #2, #3 and #4 of the General Design Cycle), it was noted that very few of the tweets included the geolocation of the person sending the tweet. The search algorithm was modified to include various city names, suburb names and area names within the Gauteng province along with alternate spellings. An example of this included a search term of

(pretoria OR pta OR tshwane)

Approximately 50 cities, suburbs, or areas in Gauteng were compiled along with their geolocations. Various alternate spellings of these names were also equated. These locations covered the broad area of Gauteng from Carletonville in the south-west, to Hammanskraal in the north, to Bronkhorstspruit in the east and to Heidelberg in the south-east. This list was compiled during iterations of the GDC. Twitter was searched for specific locations and combinations of specific words which indicated thunderstorms such as hail, lightning, thunder, rain, etc.

ThunderChatter: Step two – Analyse tweets

The second step in the ThunderChatter Model involved analysing the text portion of the tweets. For this step, the μ Model (pronounced "mu" and representing the phrase "microtext understander") (Butgereit, 2012) was used. The μ Model was specifically designed to classify short microtext messages (such as tweets) into predefined sets of categories. It was originally implemented as a way to classify synchronous XMPP (eXtensible Messaging and Presence Protocol) based conversations extracted from a mathematics tutoring platform into mathematical topics.

The term *microtext* is defined as very short utterances. Ellen (2011) provided three possible characteristics of microtext:

- Microtexts are very brief utterances consisting of as little as a single word or symbol
- Microtexts are generally informal and unstructured
- Microtexts have a "minute level" time stamp and an author

Examples of microtext include SMS (Short Message System), instant messaging, voicemail transcriptions and microblogs such as Twitter.

The µ Model is a topic spotting model specifically for microtext and consists of four steps (Butgereit, 2012):

- Removal of stopwords
- Stemming (removing suffixes)
- Correcting misspellings
- Topic determination

These four steps of the µ Model were incorporated into the ThunderChatter Model.

Stopwords are words which are not relevant to the topic under discussion (Wilbur & Sirotkin, 1992). These are words such as *the*, *a* and *an*. A tweet such as

Watching my car get pelted with hail #Centurion #Storm

would be reduced to just

pelted hail storm

depending on the stopword list. For this specific instantiation, over six hundred stop words were implemented.

Stemming removes suffixes from words leaving just the stem of the word (Hatcher, & Gospodnetic, 2004). Words such as *pelting, pelted* and *pelts* would all be reduced to simply *pelt* by stemming.

The third step is to correct any misspelled words if possible. A tweet such as

Huge lighting storm heading over #Roodepoort now.

Would be transformed into

lighting storm

after stopword removal. After stemming, the tweet would look like

light storm

and then with spelling correction, the tweet would look like

lightn storm

The last step in the µ Model determines whether or not these two words indicate a thunderstorm. This was done using N-gram comparisons to a specific vocabulary list of words which indicate thunderstorms. N-grams leverage the fact that conversations on specific topics typically have some words which occur more often than other words (Cavnar & Trenkle, 1984). This implies that there is always a set of words, which dominate most other words in such conversations. Cavnar and Trenkle assert that when texts about two similar topics are compared, they will have similar N-gram profiles.

Once it was determined that the tweet referred to a thunderstorm, the data was analysed a second time for geolocation data. In some cases, the tweet received from Twitter did hold the geolocation of the author of the tweet. In other cases, however, the tweet text was analysed for suburb and city names such as Pretoria, Jozi, Sandton, and Midrand. If location strings were found in the text of the tweet, then the appropriate geolocation was added to the tweet.

ThunderChatter: Step three – Plot Tweets on map

The final step in the ThunderChatter Model is to count and plot the tweets on a map using the quantity of tweets and their geolocation as input data. Plotting the data on a map requires either geolocation data on the tweet itself or a location name on the tweet (such as Pretoria or Johannesburg). In the case where tweets did not have specific longitude and latitude attached to it and where the geolocation was added by means of natural language processing, the actual longitude and latitude of the city or suburb were modified slightly to spread the tweets out over the relevant city or suburb ensuring that each tweet could be seen on a map.

Evaluation

In order to evaluate the efficacy of the ThunderChatter Model and Instantiation actual data from thunderstorms was obtained from the South African Weather Service. The South African Weather Service provides information to entities such as aviation companies and insurance companies. For this research, lightning-stroke maps were obtained from the South African Weather Service and visually compared to the maps generated by ThunderChatter. Because of the commercial value of these lightning-stroke maps, the actual dates are not given in this paper. The dates will be denoted by letters of the alphabet. Because of funding issues, the actual hard data of the lightning-strokes were not available to the author. Only a map was made available to the author given the limited budget of this research.

For this reason, a mechanism needed to be put in place to visually compare two maps without the accurate underlying geolocations of the markers on the maps. In order to do this, an exercise in Content Analysis as defined by Krippendorf (1980) was undertaken. Content analysis is a research technique for making replicable and valid inferences from data to the surrounding context (Krippendorff, 1980, p. 21).

Krippendorff (1980) argues that content analysis is a research technique which involves specialised procedures for processing scientific data. If the results of any research technique were to be valid, the results would need to be reliable. Reliability determines or assesses if the research design and its results are free from extraneous circumstances of measurement, hidden idiosyncrasies and surreptitious biases. In order to test for reliability, some duplication of effort is essential.

Krippendorf defines three different types or reliability: stability, reproducibility and accuracy.

Stability is the degree to which a process is unchangeable or constant over time. Using stability as a measurement of reliability can be done when there is, perhaps, only one person to evaluate a set of data. In order to reduce bias, the person could evaluate the data, wait a certain amount of time so that the person forgets the first set of results and then re-evaluate the data. If the results do not change over that period of time, then the results could be considered to be stable. Krippendorff argues that stability is the weakest form of reliability and should not be trusted as the only indicator thereof (Krippendorff, 1980, p. 130).

Reproducibility is the degree to which a process can be recreated under different circumstances using different people in different locations. Using reproducibility as a measurement of reliability can be done where more than one person evaluates the same set of data using the same set of criteria. If the results from these two or more evaluations by different people are similar, then the results could be considered to be reproducible (Krippendorff, 1980, p. 131).

Accuracy is the degree to which a process conforms to a known standard. Using accuracy as a measurement of reliability can be done where there is one person evaluating a set of data and the evaluation is then compared to a known standard. Accuracy is the strongest test of reliability (Krippendorff, 1980, p. 131)

Without the actual geolocations of each lightning-stroke (which was beyond the budget of this research), it was not possible to *accurately* evaluate this research using Krippendorf's definition of *accuracy*. It was possible, however, to evaluate this research in a way that was reproducible and free of author bias.

The following steps were taken to evaluate this research.

- During the summer period from November, 2013, to April, 2014, the software implementation of the ThunderChatter Model was kept operational with a 15-minute polling interval at Twitter.
- When a typical Gauteng thunderstorm crossed Gauteng, the author would telephonically order a lightning-stroke map from the South African Weather Service.
- The lightning-stroke maps from the South African Weather Service and the lightning intensity map generated by the ThunderChatter implementation were divided into 6x6 grids.
- Two students from the University of Pretoria were approached independently to compare the two maps and mark which of the 36 grid cells were similarly populated between the two maps.

By using two students working independently, it was possible to remove author bias. From the results of these students, it was found that the thunderstorms were always noted on Twitter when they crossed the strip of land from Soweto, through Johannesburg, Midrand, Centurion and onto Pretoria. Only in extremely rare occasions when there was actual property damage in areas outside this strip of land (such as in Hammanskraal) were tweets outside this area received.

The left most map in Figure 1 shows the lightning-stroke map for date A over that area of Gauteng which spans from Soweto, through Johannesburg, Midrand, Centurion and on to Pretoria. Cities were indicated by number with 5 indicating Johannesburg, 6 indicating Pretoria and 7 indicating Soweto. Although there were lightning-strokes outside of this area, there were no tweets collected outside this area. It is this particular strip of Gauteng which is of primary interest. Date A was a Wednesday and the storm occurred in the afternoon. Although the map does not indicate times, the storm moved from the south-west to the north-east as is typical of Gauteng thunderstorms (Tyson & Preston-Whyte, 2000). Although the entire lightning-stroke map is not included in this paper, there was a high level of lightning strokes in the south-west around Carletonville, another cluster of lightning strokes just south of Pretoria (as can be seen in this extract) and a final cluster of lightning strokes to the north-east of Bronkhorstspruit (which is not included in this extract).

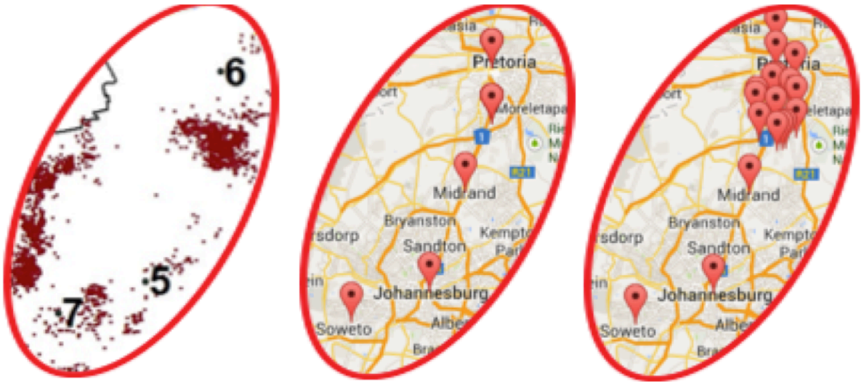

Figure 1: Date A

The middle map and far right map in Figure 1 are the summary maps of the tweets extracted from Twitter. The middle map merely plots the path of the storm with only one marker displayed in any particular location. The right map shows the intensity of the storm with each marker indicating one tweet over a very short period of time. All of the tweets which were extracted came from the strip of Gauteng running from Soweto, through Johannesburg, Midrand, Centurion and onto Pretoria. The fact that there were no tweets collected about the thunderstorm outside of this area will be dealt with in detail in the conclusion of this paper.

Data collected from Twitter include:

> *My soweto peeps yall feeling the rain*
>
> *Eish Jozi weather sucks…hate rain at this time of the day*
>
> *rain clouds gather in Midrand*
>
> *Watching my car get pelted with hail :(#Centurion #Storm*

Visually comparing the maps in Figure 1 shows high activity just south of Pretoria. This correlates with very dramatic tweets coming from Centurion on date A:

> *Centurion second hailstorm in 3 days (pea size and smaller). The weathergods are treating you hail fanatics*
>
> *Centurion heavy rain + heal hi way heavy busy*
>
> *Is anyone seeing this fucking storm in centurion?????*
>
> *Crazy storm in #Centurion - be safe on the roads*
>
> *The storm&rain in Centurion…MOER!#HECTIC!!!!*
>
> *The way it goes from clear skies to rain storm in a matter of minutes here in Centurion*

A second example can be seen in Figure 2. Date B was a Monday and the storm was in the afternoon. The three maps in Figure 2 show the lightning-strokes, the storm path as seen by monitoring tweets and the storm intensity.

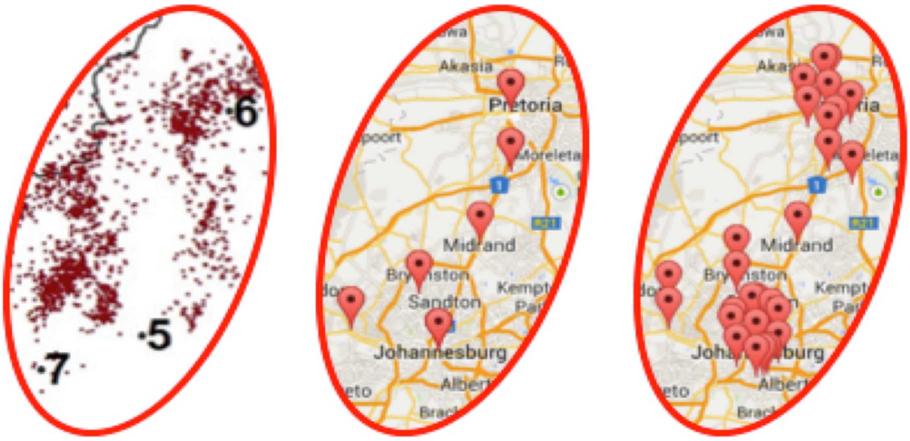

Figure 2: Date B

Visually comparing the maps in Figure 2 shows again that the thunderstorm activity between Soweto and Pretoria is very clearly captured using Twitter data.

This strip of area between Soweto and Pretoria coincidently matches the path of the N1 highway. These two examples show that thunderstorms in the vicinity of the N1 highway between Soweto and Pretoria can be tracked by crowdsourced Twitter data.

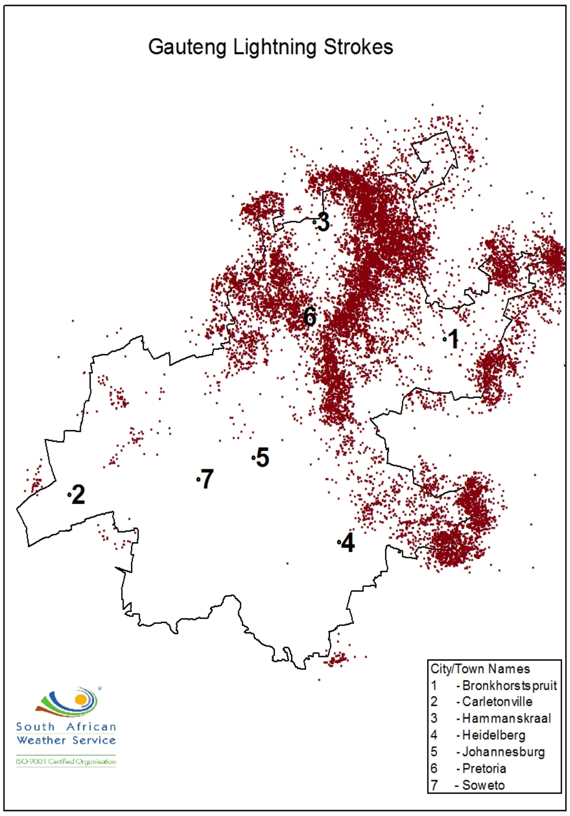

Figure 3 Lightning-stroke Date C

Figure 3 shows an entire lightning-stroke map on date C. The thunderstorm shown in Figure 3 was centred in the east and north-east of Gauteng. There were many tweets from the

Pretoria/Mamelodi area and nothing received from farther north such as from Hammanskraal.

> *mother of a storm ! #Pretoria*
>
> *huge thunder storm over parts of Pta and Pta east*
>
> *Stormy Pretoria hail and thunders!*

This characteristic of thunderstorms needing to be in the vicinity of the N1 highway between Soweto and Pretoria before its path is mappable using crowdsourced Twitter data correlates with the population density of Gauteng and the percentage of households which indicate they have access to the Internet as seen in Figure 4.

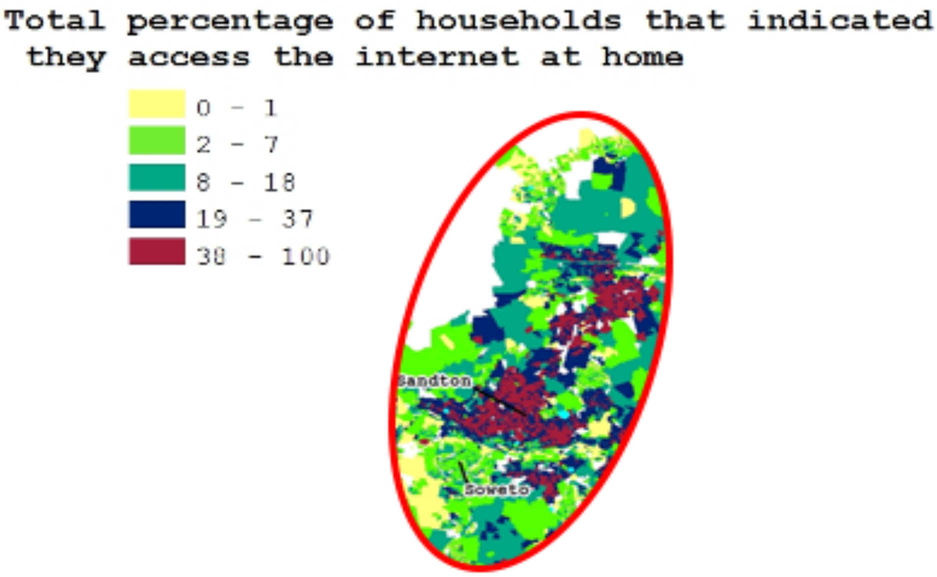
Total percentage of households that indicated they access the internet at home
0 - 1
2 - 7
8 - 18
19 - 37
38 - 100
Sandton
Soweto

Figure 4 Percentage households that have Internet (Image Credit GCRO)

This is not to say that areas such as Hammanskraal are not mentioned in tweets. But this typically comes in the form of retweets from news agencies when the storm has been of such an intensity that there has been damage to buildings and vehicles.

> *Soshanguve too @unisaradio : Just in: Houses have been damaged and roads flooded by a hail storm in Mamelodi and Hammanskraal outside Pretoria*
>
> *@SAfmnews Houses have been damaged and roads flooded by a hail storm in Mamelodi and Hammanskraal outside Pretoria.*
>
> *@TrafficSA Storm damage in Mamelodi Pretoria. 80 houses destroyed on a construction site.*
>
> *Storms and hail assaulted Pretoria North Akasia and Soshanguve few*

minutes ago. Zinc roofs and broken glass everywhere.

As more thunderstorms were monitored and compared with the crowdsourced Twitter data, it became clear that only thunderstorms which traversed certain areas could be mapped using Twitter data. Those areas were more-or-less defined as a strip of densely populated areas from Soweto at the south to Pretoria in the north. This strip of populated area coincidently corresponds to the N1 highway.

Conclusion

This research attempted to track Gauteng thunderstorms by using crowdsourced Twitter data. The resulting information was displayed on maps of Gauteng. The maps were then compared with lightning-stroke maps supplied by the South African Weather Service to determine the accuracy of the thunderstorm tracks based on the crowdsourced Twitter data.

The results of this research were positive for thunderstorms which crossed the strip of populated areas between Soweto and Pretoria. Thunderstorms which came from the south-west and traveled to the north-east (which were the most common) were easily tracked from Soweto, through Johannesburg, Midrand, Centurion and onto Pretoria. Thunderstorms which did not follow this path but still cut across this strip of populated area were still mapped over the appropriate populated area. The Twitter comments for these thunderstorms were more up-to-date than traditional weather news broadcasts.

For outlying areas such as Carletonville, Heildelberg, Bronkhorstspruit and Hammanskraal, the results of this research were not positive. Thunderstorms were occasionally mapped in these areas but that was only in extreme cases where there was damage to property and the Twitter data was from news agencies.

It is a well documented and researched fact that people like to talk about the weather. The weather is a safe topic of conversation which complete strangers can safely discuss. It is not a personal topic of conversation. Researchers have shown that weather events help people organise their autobiographical memory. As Internet connectivity reaches more and more households in Gauteng, more people will use the Internet for conversations with other people using various Internet protocols. Many of these conversations will naturally be about the weather. As more and more people take part in Twitter, it is suggested that it will become easier to track thunderstorms over the entire extent of Gauteng instead of just the band of area between Soweto and Pretoria.

This research shows that thunderstorms travelling over populated areas can be easily mapped using data from Twitter. This research could be augmented in future with attempts to predict the future course of these thunderstorms and, thereby, provide warnings to people in the predicted pathway of the storm.

Acknowledgements

The lightning-stroke maps in Figures 1, 2 and 3 were provided by the South African Weather Service. The maps containing the crowdsourced Twitter data in Figures 1 and 2 were generated using Google's Maps API. The map of Gauteng in Figure 4 is from the Gauteng Region-City Observatory.

References

ALSEVER, J., 2007. What is Crowdsourcing? *BNET.com, March,* 7.

ARCHER, E., ENGELBRECHT, F., LANDMAN, W., LE ROUX, A., VAN HUYSSTEEN, E., FATTI, C., VOGEL, C., AKOON, I., MASERUMULE, R. and COLVIN, C., 2010. *South African Risk and Vulnerability Atlas.* Department of Science and Technology.

BHAVIKA, B., 2007. Master of Science dissertation. The Influence of Terrain Elevation on Lightning Density in South Africa: University of Johannesburg.

BRICE, T. and PIEPER, C., 2009. Using Twitter to Receive Storm Reports. *Available online at: ams.confex.com/ams/pdfpapers/163543.pdf.*

BUTGEREIT, L., 2012. *A Model for Automated Topic Spotting in a Mobile Chat Based Mathematics Tutoring Environment.* PhD Thesis. South Africa: Nelson Mandela Metropolitan University.

BUTGEREIT, L., MOONSAMY, S., THOMSON, T., VAN ZYL, T. and MCFERREN, G., 2014. Fire Hazard Notifications via Satellite, Twitter, Citizen Reports, and Android Apps, J.F. VAN NIEKERK, ed. In: *Proceedings of the African Cyber Citizenship Conference*, November 5-6, 2014, Nelson Mandela Metropolitan University, pp. 122-128.

BUTGEREIT, L., 2014. Crowdsourced Weather Reports: An Implementation of the µ Model for Spotting Weather Information in Twitter, *IST-Africa 2014 Conference Proceedings* 2014.

CARTE, A. and HELD, G., 1978. Variability of Hailstorms on the South African Plateau. *Journal of Applied Meteorology,* 17, pp. 365-373.

CAVNAR, W.B. and TRENKLE, J.M., 1994. N-gram-based Text Categorization. *Ann Arbor MI,* 48113, pp. 4001.

CHEUNG, C.M., CHIU, P. and LEE, M.K., 2011. Online Social Networks: Why do Students use Facebook? *Computers in Human Behavior,* 27(4), pp. 1337-1343.

CLARY, E.G., SNYDER, M., RIDGE, R.D., COPELAND, J., STUKAS, A.A., HAUGEN, J. and MIENE, P., 1998. Understanding and Assessing the Motivations of Volunteers: A Functional Approach. *Journal of Personality and Social Psychology,* 74(6), p. 1516.

ELLEN, J., 2011. All about Microtext-A Working Definition and a Survey of Current Microtext Research within Artificial Intelligence and Natural Language Processing. *ICAART (1)* 2011, pp. 329-336.

HARLEY, T.A., 2003. Nice Weather for the time of Year: The British Obsession with the Weather. In: S. STAUSS and B. ORLOVE, eds, *Weather, Climate, Culture.* Oxford: Berg Publiishers, pp. 103-120.

HATCHER, E. and GOSPODNETIC, O., 2004. *Lucene in Action.* Greenwich, Connecticut: Manning Publications.

HEVNER, A.R. and CHATTERJEE, S., 2010. *Design Research in Information Systems: Theory and Practice.* New York: Springer Verlag.

HOWE, J., 2006. The Rise of Crowdsourcing. *Wired Magazine,* 14(6), pp. 1-4.

JAVA, A., SONG, X., FININ, T. and TSENG, B., 2007. *Why we Twitter: Understanding Microblogging Usage and Communities.* ACM New York, NY, USA.

KOK, P., ed, 1998. *South Africa's Magnifying Glass. A Profile of Gauteng.* South Africa: Human Sciences Research Council.

KRIPPENDORFF, K., 1980. *Content Analysis: An Introduction to its Methodology.* Thousand Oaks, California. Sage Publications, Inc.

KROEZE, J.H. and VAN ZYL, I., 2014. Transdisciplinarity in Information Systems: Extended Reflections, *AMCIS 2014 Proceedings*, Aug 7-9, 2014, pp. 1-10.

KRUGER, A., GOLIGER, A., RETIEF, J. and SEKELE, S., 2010. Strong Wind Climatic Zones in South Africa. Korea: Techno Press.

KUZNETSOV, S., 2006. Motivations of Contributors to Wikipedia. *ACM SIGCAS Computers and Society,* 36(2), pp. 1.

KWAK, H., LEE, C., PARK, H. and MOON, S., 2010. *What is Twitter, a Social Network or a News Media?* ACM.

LEHOHLA, P., 2012. *Census 2011 Census in Brief.* South Africa: Statistics South Africa.

MABIN, A., 2013. *The Map of Gauteng: Evolution of a City-Region in Concept and Plan.* Johannesburg: Gauteng City-Region Observatory.

Mobile Volunteering, 2013. Available: www.mobilevolunteering.co.uk Accessed Jan 7, 2013.

NOV, O., 2007. What Motivates Wikipedians? *Communications of the ACM,* 50(11), pp. 60-64.

OATES, B.J., 2006. *Researching Information Systems and Computing.* London: Sage Publications Ltd.

O'SULLIVAN, M.D., 2012. *Wikipedia: a New Community of Practice?* London: Ashgate Publishing, Ltd.

ROBINSON, J.A. and RUBIN, D., 1986. Autobiographical Memory: A Historical Prologue. In: D.C. RUBIN, ed, *Autobiographical Memory.* pp. 19-24.

ROSS, C., ORR, E.S., SISIC, M., ARSENEAULT, J.M., SIMMERING, M.G. and ORR, R.R., 2009. Personality and Motivations Associated with Facebook use. *Computers in Human Behavior,* 25(2), pp. 578-586.

SAKAKI, T., OKAZAKI, M. and MATSUO, Y., 2010. Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors, *Proceedings of the 19th International Conference on World Wide We*b 2010, ACM, pp. 851-860.

STRAUSS, S. and ORLOVE, B., 2003. Up in the Air: the Anthropology of Weather and Climate. *Weather, Climate, Culture. Berg, Oxford,* pp. 3-16.

TYSON, P.D. and PRESTON-WHYTE, R.A., 2000. *The Weather and Climate of Southern Africa.* Second Edition. Cape Town: Oxford University Press.

VAISHNAVI, V. and KUECHLER, W., 2007. *Design Science Research Methods and Patterns: Innovating Information and Communication Technology*. Boca Raton: Auerbach Publications.

VIEWEG, S., HUGHES, A.L., STARBIRD, K. and PALEN, L., 2010. Microblogging during Two Natural Hazards Events: What Twitter may Contribute to Situational Awareness, *Proceedings of the SIGCHI Conference on Human Factors in Computing System*s 2010, ACM, pp. 1079-1088.

Volunteer Match, 2013. Available: www.volunteermatch.org Accessed Jan 7, 2013.

WILBUR, W.J. and SIROTKIN, K., 1992. The Automatic Identification of Stop Words. *Journal of Information Science*, 18(1), pp. 45. London: Sage Publications.