IST-Africa 2014 Conference Proceedings
Paul Cunningham and Miriam Cunningham (Eds)
IIMC International Information Management Corporation, 2014
ISBN: 978-1-905824-43-4

CrowdSourced Weather Reports: An Implementation of the μ Model for Spotting Weather Information in Twitter

Laurie BUTGEREIT[1, 2]
[1]*Nelson Mandela Metropolitan University, Port Elizabeth, South Africa*
[2]*CSIR Meraka Institute, Pretoria, South Africa, Tel: +27 83 453 7039*
Email: laurie.butgereit(at)nmmu.ac.za, lbutgereit(at)meraka.org.za

Abstract: Twitter is a microblogging facility that allows people to post 140 character status updates about various topics. In times of special events (such as extreme weather, emergencies, sporting goals, etc), status updates on Twitter often give people a better view of the event than traditional news operations or weather services. This paper describes a project in monitoring Twitter for weather status updates for a specific city and being able to automatically determine the current weather by analysing those tweets.

Keywords: Twitter, Weather, Micro-blogging

1. Introduction

Twitter is a microblogging facility that allows people to freely publish 140 character status updates or *tweets* [1, 2]. When events are changing quickly, such as during natural disasters, the status updates on Twitter are often more up-to-date than traditional news broadcasts. Research has shown that in Japan 96% of all earthquakes, which were stronger than 3 on the JMA (Japan Meteorological Agency) seismic intensity scale can be detected using Twitter [3].

Sourcing information from Twitter can be considered to be a type of *crowdsourcing.* Definitions of the term *crowdsourcing* vary but underlying most definitions include the following four components [4]:

1. An organised or unorganised group of people...
2. provide information (such as knowledge or photos or assistance)...
3. usually free of charge...
4. to another person or entity

When groups of people post information on Twitter which another person or entity extracts and consolidates, this can be considered to be a type of crowdsourcing.

This paper investigates the use of Twitter to produce one- and two-word descriptions of the weather in a general location by monitoring tweets that pertain to that area. The tweets are either tagged with a location hash tag or are geolocated within a certain distance from the location. The algorithms involved take into account certain spelling conventions that are present in Twitter.

2. Crowdsourcing and Twitter

Crowdsourcing is tapping into the collective intelligence of the public to complete a task that would usually be done by a specific agent [4]. The term crowdsourcing was coined by

Copyright © 2014 The authors

Jeff Howe in an article in Wired Magazine in 2006 [5] to describe a situation where people use their "spare cycles to create content, solve problems, even do corporate R&D".

Crowdsourcing websites are common. Sites such as Wikipedia and YouTube encourage people to create content. Sites such as Volunteer Match [6] and Mobile Volunteering [7] encourage people to solve problems. Sites such as InnoCentive and iStockPhoto encourage corporate R&D using a crowdsourced model [5].

Twitter provides a wealth of data from people around the world. People tweet about their personal lives but they also tweet about events which they witness or experience – social unrest, weather emergencies, sporting events, political transitions, celebrity secrets, etc. Being able to extract information from Twitter to create helpful intelligence is important.

3. Research Methodology

Information Systems research can be divided into two paradigms: behavioural science and design science. The behavioural science paradigm has its roots in natural science and attempts to describe reality by developing and justifying theories. The design science paradigm has its roots in engineering and attempts to solve problems by developing artifacts [8].

Design Science Research (DSR) has as one of its fundamental pillars the requirement that an innovative artifact must be created to solve an important problem in an innovative manner. The artifact can be a construct, a model, a method, or an instantiation [9-11]. The General Design Cycle (GDC) of DSR is defined as having five steps [12]:

1. Awareness – the recognition of a problem and the statement thereof
2. Suggestion – offering of tentative ideas of how to solve the problem
3. Development – implementation of the aforementioned suggestions
4. Evaluation – examination of the developed artifacts
5. Conclusion – consolidation of results

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 had more up-to-date weather information online than the official weather reports.
2. Suggestion – Suggestions on how these tweets could be analysed using various text processing techniques and a one- or two-word weather summary could be generated.
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 summaries were generated.
4. Evaluation – The weather summaries generated in #3 above were evaluated. Until such time that the summaries were satisfactory, steps #2, #3, and #4 were cycled through again.
5. Conclusion – The final results were summarised for the scope of this research.

This research is based on a model that had already been developed using Design Science Research. This research involved the instantiation of that model and the development of a prototype artifact.

4. Pretoria

For the implementation of this research, the algorithms attempted to determine the weather in the vicinity of Pretoria, South Africa.

Pretoria is situated at an altitude of above 1,250 m above sea-level and experiences hot summers and cold winters. Winter rainfall is sparse and winter is characterised by sunny days, clear skies, and cold nights. On occasion, a cold front from the southern cape of South Africa may intrude north and bring snow. Summer is the most important rainfall season and is also characterised by heat induced thunderstorms [13].

5. Extraction of Tweets

Tweets were extracted daily from Twitter using the twitter4j Java library. During the early part of the project, tweets were only extracted once per day just before midnight. Two queries were executed.

The first query was to search for tweets that were tagged in a specific manner. These tagged searches included searching for #ptaweather @PretoriaZA #pretoria #pta and #weather.

The second query was based on the longitude and latitude of Pretoria and looked for specific search terms that were not tagged but were common in weather related tweets. Examples of these weather related terms included storm, rain, cold, hot, sun, etc. Due to the iterative nature of Design Science Research, the list of search terms evolved over the course of the project.

Included in the tweets, were official weather reports from reputable organisations. These tweets were separated from the public tweets and were treated as a type of "ground truth" against which the algorithms could be tested. The term "ground truth" can refer to different things in different industries. However, Sikdar, Kang, and Adah specifically speak of "ground truth" with respect to Twitter credibility [14]. To support a credible "ground truth" against which our algorithms could be tested, we chose an official weather report Twitter feed with numerical measurements.

An example of an official weather report which was separated into the "ground truth" category is:

> Monday ☘ 10|20°C ☂ 66%• Tuesday ☘ ☀ 9|23°C ☂ 20%• Wednesday ☘ ☀ 10|21°C ☂10%• Thursday ☘☀11|23°C ☂10% #PTAWeather

It is understood, however, that many of these "ground truth" tweets are, in fact, predictions of the weather and not actually weather reports. In addition to testing algorithms against these "ground truth" tweets, an exercise of Content Analysis as defined by Krippendorf [15] would be done with two human coders (working independently) interpreting the tweets and manually classifying them. The algorithms could also be tested against the human interpretation of the tweets.

Unofficial weather tweets from the public were extremely casual in nature. For example:

- Good Morning it's such a beautiful day
- Pretoria is crazy these days with these BRT road constructions rain and robots that are not working.
- Gloomy weather today pta
- Cold misty morning in Pretoria

After the tweets were separated into "ground truth" category and public category, the public category of tweets were processed by the various text processing algorithms used in this research. The outcome was manually compared against the "ground truth" category of tweets and against the classifications done by the human coders. Depending on the outcomes, steps #2, #3, and #4 were iterated again.

All processing was done in a case insensitive manner.

6. MicroText

Ellen [16] has defined microtext as utterances which have three possible characteristics:

1. Very brief consisting of as little as a single word or symbol.
2. Generally informal and unstructured.
3. Has a "minute level" timestamp and an author.

Examples of microtext include:

1. SMS (Short Message System)
2. Instant Messaging
3. Voicemail transcriptions
4. Microblogs such as Twitter.

7. The μ Model for Topic Spotting

The μ Model (pronounced "mu" and representing the phrase "microtext understander") is a model for spotting predefined topics in microtext [17]. The model was originally instantiated to spot mathematics topics in instant messaging conversations between school pupils and mathematics tutors.

The model is based on an archive of historical data. In an iterative research methodology, incoming data is used to test the implementation of the model. After the testing, the data is then incorporated into the historical data in order to improve the instantiation.

The μ Model consists of four steps:

1. Removal of stopwords
2. Stemming words (removing suffixes)
3. Correcting any misspellings where possible
4. Topic determination

The first step is to remove stopwords. Stopwords are defined as words that have the same likelihood of occurring in documents or conversations that are not relevant to a topic as in documents or conversations that are relevant to a specific topic [18]. For example, in a tweet such as

- Its a really beautiful day out here in the Capital city

words such as *its, a, out, here, in* and *the* can be considered stopwords. They add no meaning to the sentence and can be safely removed without altering the topic of the tweet:

- really beautiful day Capital city

The second step is to stem the word. A stemmer is a utility that removes suffixes (and, in some cases, prefixes) from the ends (and, perhaps, beginnings) of words leaving just the root stems. Stemmers are often used in search engines and other information retrieval systems [19].

Consider the following tweets:

- Just got rained on at Pta
- Its raining
- Loving the rain in Pretoria

By removing stops words, the tweets are reduced to

- rained pta
- raining

- rain Pretoria

The stemming process then removes the suffixes -ed and -ing leaving the tweets as

- rain pta
- rain
- rain pretoria

The third step is to correct any obviously misspelled words taking into account spelling conventions in microtext. For example, in the following two tweets

- wow what lightning
- this lightening is scary

stopword removal would change the tweets to

- lightning
- lightening

Stemming would then change them to

- lightn
- lighten

During the third step, the second tweet which now consists of only the word *lighten* would be changed to the word *lightn.* This was done through the use of N-grams and a Jaccard similarity calculation.

The fourth step of the µ model is to now compare these keywords (which have had stopwords removed, which have been stemmed, and which have had their spelling corrected when possible) against lists of categorised weather terms. These lists were created manually after analysing the historical archive of tweets. During the development and evaluation phases, these lists were manually edited. Again, this was done through the use of N-grams and a Jaccard similarity calculation.

For example, the word list for the topic cloudy weather included the stemmed words such as dark, scatt (from scattered), cloud, overcast, part (from partly), grey, gray, and gloom (from gloomy).

8. The µ Instantiation for Weather Topics

The µ model was instantiated specifically for weather topics. In addition, only weather situations that have arisen in Pretoria over recent history were catered for. That means although snow as a weather phenomenon could be recognised, blizzard would not be recognised.

As the project iterated over steps #2, #3, and #4 of the General Design Cycle, the list of stopwords was manually edited. Future implementations of the µ model could attempt to automate this process. At the time of writing this paper, the stopword list contains just under one thousand words.

The stemmer from a previous instantiation of the µ model was used for this instantiation. It handled common English language suffixes such as -ing, -ies, -ied, etc, and also catered for a few new suffixes which are more common in microtext. One such suffix is the -a suffix which replaces the -er, -ir, -or, -ar suffix as in the word *afta* (meaning *after*).

- Did u hear dat thunda?

On the original implementation of this project, 15 different weather types were categorised. For each weather type, words which were commonly used to describe that weather condition were listed in their stemmed format. One weather event is only recognised in South African English. A “monkey's wedding” is simultaneous sunshine and rain [20].

The fifteen weather types were bad, cloudy, cold, cool, extreme wind, fog, hail, hot, monkey's wedding, nice, rain, snow, sunny, thunderstorms, and wind.

It is important to note that at this point in time, this instantiation of the μ Model was designed to process tweets individually. This instantiation did not consider changing weather trends over one day. It only processed individual tweets.

9. Test Data

Twitter has the facility to extract tweets. Current free facilities only allow tweets for the past day to be extracted. For this reason, data was collected during the period of October through November 2013. Although it is acknowledged, that the weather in Pretoria during that period only covered one season (southern hemisphere spring), the techniques used to instantiate the μ Model can be extended using a year's worth of data.

The dataset consisted of approximately 20 tweets per day over approximately 60 days. Odd numbered days were considered to be research data for the algorithms and even numbered days were used as test data. Both data sets had "ground truth" measurements for each day from the weather bureau. Although it is understood that the weather in Pretoria was in spring during the 60 days of the collection of this data, the algorithms would work with a larger collection of data.

Approximately 600 tweets were used for the creation of the stop word list and the creation of the vocabulary lists for the 15 different weather types. The remaining tweets were then manually inspected to ensure that they did, in fact, cover weather topics. For example, one of the tweets removed from the data set was

- bafana bafana beat spain for real i think its going to snow in pretoria in December well done boys

This tweet was referring to an unexpected soccer outcome which could be a rare as snow falling in Pretoria

10. Results

During testing, the algorithm could not classify just under 15% of the tweets (13.6%). This was due to various vocabulary which was used in the tweets. These tweets expressed opinions about the weather and did not describe the weather. Examples of tweets which could not be classified include:

- awesome cinema time in this weather sterland pta
- but then again pta weather is on another level
- starting to hate pretorias weather

The attempt to classify a monkey's wedding could not be tested since there were no tweets at all about that phenomenon in test data.

Tweets about dramatic weather changes (such as thunderstorms) were much easier to determine. People were eager to tweet descriptions of the thunderstorms, lightening, hail, etc. The algorithm determined thunderstorms with a high degree of accuracy (above 90%) but this was due more to the human factor of people describing transitional events with more emotion and enthusiasm.

- another huge storm in pretoria tonight
- real beatiful thunderstorm and rain all the way from jhb up to km from my home
- in pta
- thunder and rain in pretoria nothing beats a highveld thunderstorm on a monday
- is that thunder i hear pta weather
- heavy lightning over pretoria east

- this lightening in pta is beyond me

Tweets about ongoing weather conditions (such as a week of sunshine) were not as descriptive as tweets about thunderstorms. As such, it was difficult for the algorithm to pick up on key words although with additional work, it could be improved.

The instantiation of the μ Model for could identify thunderstorms with above 90% accuracy. With other weather conditions, the instantiation was not as accurate.

During the course of this research, a massive thunderstorm hit the Pretoria-Johannesburg-Soweto area (in the province of Gauteng) [21]. According to news reports and amateur photos posted on Twitter and Facebook, the hailstones ranged in size up to the size of golf balls and tennis balls. The twitter posts included dramatic descriptions of the storm as can be seen following:

- ja that lightning neh made me sober
- storms and hail assaulted pretoria north akasia and soshanguve few minutes ago zinc roofs and broken glass everywhere
- eish was n huge storm in equestria pretoria my kar net n klomp duike en selfs die huis was nie sterk genoeg nie
- houses have been damaged and roads flooded by a hail storm in mamelodi and hammanskraal outside Pretoria
- even my home is destroyed this rain destroyed lot of homes around pta
- horrific hail storm that took over pretoria hours ago left me scared as hell
- storm damage in mamelodi pretoria houses destroyed on a construction site

The instantiation of the μ Model easily identified these tweets as describing a thunderstorm and tagged the tweets as thunderstorm with 100% accuracy.

Also during the course of this study, an interesting meteorological condition occurred. On November 6, 2013 during the period of testing, “sun halos” or “sun dogs” occurred. These are caused by light from the sun being affected by atmosphere ice crystals in high cirrus clouds creating an arc around the sun [22]. Because the instantiation of the μ Model was completely unprepared for such phenomenon, it could not report on them.

11. Business Benefits

Although many countries have weather bureaus to report on weather situations, some countries do not. In such cases, it would be beneficial to be able to use a crowd sourcing mechanism to obtain weather reports. In addition, even in countries where there are national weather bureaus which broadcast weather situations, in times of extremely fast changing weather (such as tornadoes or heat induced thunderstorms), the traditional weather reports often do not keep up with the changing situation and crowdsourced weather summaries can be helpful.

In cities that experience heat induced thunderstorms including hail, the damage to buildings and vehicles, coupled with injuries to people, can be expensive. While doing research for this paper, Pretoria and Johannesburg experienced a dramatic thunderstorm on November 28, 2013 [21]. According to these news reports, the damage to vehicles and property has amounted to millions of Rands of insurance claims [23, 24]. An automated system that could extract intelligence from crowdsourced Twitter feeds could have easily sent out additional warnings to people who were not connected to social media. These warnings could have been sent out via SMS (Short Message System) so that people who could not afford smartphones and Internet connectivity could have also received warnings about the storm using inexpensive technology.

12. Mitigation of False Information

One of the problems which needs to be handled is the possibility of false information being maliciously posted on Twitter. This problem is not specific to Twitter but is a general problem with social media. In 2007 and in 2011 the Johannesburg/Pretoria area was subject to weather panic due to hoax email warnings of tornadoes and high winds [25, 26].

The project presented in this paper proposes a *linguistic* or *language* approach to determining the current weather by monitoring Twitter. In order to help mitigate against false information, an additional *meteorological* analysis of the results would have to be implemented. For example, assume that the linguistic analysis of the weather tweets consistently summarised the weather as "cold and dry" in Pretoria. If tweets began to be published warning of a "tornado", the meteorological analysis should flag that as false information. Tornadoes, which are rare in any case in Southern Africa, extend downwards from cumulonimbus clouds which are formed by rising hot moist air [27].

This paper only dealt with a linguistic of language analysis of Twitter data. The implementation of a meteorological analysis of Twitter is beyond the scope of this paper.

13. Conclusions

This paper investigated the use of crowdsourced weather reports as a basis for summarised weather reports. This investigation included the automatic extraction of Twitter data related to weather and a programmatic linguistic analysis of that data. The specific crowdsourced weather reports were extracted from Twitter and analysed using an implementation of the μ Model.

This research showed that with exceptional weather conditions (such as thunderstorms), it is possible to report on the weather condition by analysis Twitter feeds. This is consistent with the research done in Japan on using Twitter feeds to report on earthquakes. Both earthquakes and extreme thunderstorms are events about which people will post on Twitter.

This research, however, showed that with unexceptional weather, such as ongoing sunshine, the tweets provided by people are not as descriptive and not as common.

From this, it can be concluded that automatically analysing Twitter feeds can produce accurate summarised reports (above 90%) of weather conditions for dramatic, changing weather conditions.

References

[1] A. Java, X. Song, T. Finin and B. Tseng. Why we twitter: Understanding microblogging usage and communities. *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, August 12-15, 2007, San Jose, California* pp. 56-65. 2007.

[2] H. Kwak, C. Lee, H. Park and S. Moon. What is twitter, a social network or a news media? *Proceedings of the 19th International Conference on World Wide we, April 26-30, 2010, Raleigh, North Carolina, USA* pp. 591-600. 2010.

[3] T. Sakaki, M. Okazaki and Y. Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. Presented at Proceedings of the 19th International Conference on World Wide Web. 2010

[4] J. Alsever. What is crowdsourcing? *BNET.Com, March 7*2007.

[5] J. Howe. The rise of crowdsourcing. *Wired Magazine 14(6),* pp. 1-4. 2006.

[6] *Volunteer Match*. http://www.volunteermatch.org/

[7] *Mobile Volunteering*. http://www.mobilevolunteering.co.u/

[8] A. R. Hevner, S. T. March, J. Park and S. Ram. Design science in information systems research. *Management Information Systems Quarterly 28(1),* pp. 75-106. 2004.

[9] V. K. Vaishnavi and W. Kuechler Jr. *Design Science Research Methods and Patterns: Innovating Information and Communication Technology* 2007.

[10] A. R. Hevner and S. T. March. The information systems research cycle. *Computer 36(11),* pp. 111-113. 2003.

[11] S. T. March and G. F. Smith. Design and natural science research on information technology. *Decision Support Systems 15(4),* pp. 251-266. 1995.

[12] B. J. Oates. *Researching Information Systems and Computing* 2006.
[13] E. Archer, F. Engelbrecht, W. Landman, A. Le Roux, E. Van Huyssteen, C. Fatti, C. Vogel, I. Akoon, R. Maserumule and C. Colvin. *South African Risk and Vulnerability Atlas* 2010.
[14] S. K. Sikdar, B. Kang, J. O'Donovan, T. Hollerer and S. Adal. Cutting through the noise: Defining ground truth in information credibility on twitter. *Human 2(3),* pp. pp. 151-167. 2013.
[15] K. Krippendorff. *Content Analysis: An Introduction to its Methodology* 1980.
[16] J. Ellen. All about microtext-A working definition and a survey of current microtext research within artificial intelligence and natural language processing. Presented at Icaart (1). 2011.
[17] L. Butgereit, "A Model for Automated Topic Spotting in a Mobile Chat Based Mathematics Tutoring Environment," 2012.
[18] W. J. Wilbur and K. Sirotkin. The automatic identification of stop words. *J. Inf. Sci. 18(1),* pp. 45. 1992.
[19] E. Hatcher and O. Gospodnetic. *Lucene in Action* 2004.
[20] P. Silva. South african english: Oppressor or liberator. *The Major Varieties of English* 1997.
[21] *Hailstorms cause damage in Gauteng* [http://www.iol.co.za/news/south-africa/gauteng/hailstorms-cause-damage-in-gauteng-1.1614069#.UpsH_0OJRVI].
[22] L. Kirkpatrick and G. Francis, *Physics: A Conceptual World View.* California: Brooks/Cole, 2010.
[23] *JHB storm aftermath keeps insurance companies bus*y. Available: http://ewn.co.za/2013/11/30/JHB-storm-aftermath-keeps-insurance-companies-busy.
[24] *Standard Bank Implements Catastrophe Plan after Gauteng Stor*m. Available: http://www.risksa.com/standard-bank-implements-catastrophe-plan-after-gauteng-storm/.
[25] *Storm Hoax Puts Gauteng on Edg*e. Available: http://www.iol.co.za/news/south-africa/storm-hoax-puts-gauteng-on-edge-1.374035?ot=inmsa.ArticlePrintPageLayout.ot.
[26] *E-Mail Hoax - Tornadoes and Hurricane*s. Available: http://www.stormchasing.co.za/articles-and-news/articles-and-news/217-tornadoes-email-hoax.
[27] P. D. Tyson and R. A. Preston-Whyte. *The Weather and Climate of Southern Africa* (Second Edition ed.) 2000.