RELATING BIG DATA TO LOCAL NATURAL HAZARDS: LESSONS LEARNED FROM DATA MINING THE TWITTER API FOR VOLUNTEERED GEOGRAPHIC INFORMATION ON EARTHQUAKES, WILDFIRES, AND PRESCRIBED FIRES IN THE CONTIGUOUS UNITED STATES

J. L. McCarty[a], E. Levin[b], K. A. Endsley[a,c], S. T. Aden[b], and J. Bialas[b]

[a] Michigan Tech Research Institute (MTRI), 3600 Green Court, Suite 100, Ann Arbor, MI –
jmccarty@mtu.edu

[b] School of Technology, Michigan Technological University,1400 Townsend Drive, Houghton, MI
– elevin@mtu.edu

[c] School of Natural Resources and Environment, University of Michigan, Dana Building, 440 Church Street, Ann Arbor, MI 48109

Commission IV, WG IV-2

KEYWORDS: Social Media, Text Mining, Geocoding, Prescribed Fire, Wildland Fire, Earthquake, Geospatial Techniques

ABSTRACT

New media are increasingly used to capture ambient and volunteered geographic information in multiple contexts, from mapping the evolution of the social movements to tracking infectious disease. The social media platform Twitter is popular for these applications; it boasts over 500 million messages ('tweets') generated every day from as many total users at an average rate of 5,700 messages per second. In the United States, Japan, and Chile to name a few, Twitter is officially and unofficially used as an emergency notification and response system in the event of earthquakes, wildfires, and prescribed fires. A prototype for operational emergency detections from social media, specifically Twitter, was created using natural language processing and information retrieval techniques. The intent is to identify and locate emergency situations in the contiguous United States, namely prescribed fires, wildfires, and earthquakes, that are often missed by satellite detections. The authors present their methodologies and an evaluation of performance in collecting relevant tweets, extracting metrics such as area affected and geo-locating the events. Lessons learned from data mining Twitter for spatiotemporally-explicit information are included to inform future data mining research and applications.

1. INTRODUCTION

The automated extraction of volunteered and ambient geospatial information from social media has proven to be useful in a variety of contexts (Lampos et al. 2010; Stefanidis et al. 2013). A significant body of work currently exists that showcases the use of Twitter for disaster alerting, mitigation, and response (MacEachren et al. 2011; USGS 2012; Wendel 2015) where information derived from Twitter often performs equal to or better than conventional techniques (Petrovic et al. 2013; Wendel 2015). Sakaki et al. (2010) found that the operational earthquake alerting system in Japan provides alerts faster than the national Meteorological Agency. A young boy in Chile has over 405,000 users for his twitter-based earthquake information system (https://twitter.com/AlarmaSismos). In the domain of prescribed burns and wildland fires, however, Twitter has previously been used only for

qualitative assessments and case studies (Longueville et al. 2009; Vieweg et al. 2010).

Twitter's popularity for natural hazard detection and alert applications can be attributed to its volume of users and frequency of use: 200 million active users, 60% of those on mobile devices, sending over 500 million tweets every day (Moore 2013; Twitter 2013a). Twitter is used by a number of federal, state and local officials in the United States as well as by motivated individuals in a number of countries world-wide to report prescribed burn plans in advance (sometimes as part of a reporting obligation) or to communicate detection, response to, and containment of wildfires. These publicly announced fire reports, like all Twitter messages (or 'tweets'), are limited to 140 characters of text, making it easily accessible for mobile phone users and the SMS 160 character limit.

This paper describes the experience of using Twitter to detect earthquake activity in southern California as well as prescribed and wildland fires in the contiguous United States (Endsley and McCarty, Accepted). We report lessons learned from using two separate approaches: Natural Language Processing (NLP) with the streaming Twitter API and a structured text search with the Twitter archive data. Both of these data mining methods of social media aimed to assess the predictive power of information extracted from Twitter for useful event occurrence information for environmental and natural hazard applications.

2. MATERIALS AND METHODS
2.1 Natural language processing of tweets for prescribed and wildland fire reporting

A prototype framework was developed to collect messages from the streaming Twitter API that potentially described wildfire or prescribed burn activity data for the contiguous United States, to extract from information from the tweets such as the type of fire and acreage burned, and to map the location of the tweet-derived fires in near-real time. The general approach employed NLP, information retrieval (IR), and data mining techniques. The specific steps involved were, in order: (1) Collect potential wildland and prescribed fire tweets; (2) Detect and remove duplicate tweets (or 'retweets'); (3) Determine burned area of the prescribed and wildland fires; (4) Classify the fire incident as either a prescribed burn or wildfire; and (5) Locate the fire described in the tweet with real world coordinates. Basic NLP techniques such as tokenizing (breaking text up into individual signifiers, usually single words), removing 'stopwords' (common words like 'if, and, or, but') and parts-of-speech (POS) tagging are applied at almost every step of the process to facilitate computation (Russell 2011).

A total of 13,241 tweets were collected starting July 4, 2012 and continuing until June 11, 2013 (Fig. 1) when the version of the Twitter Search API being used at the time was deprecated (Twitter 2013b) and tweets could no longer be collected. This version of the Twitter Search API allowed access to millions of tweets going back approximately one week in time (Twitter 2012). The collection spanned 342 days with an average rate of 38 tweets per day and represents the contributions of 6,351 unique users. In defining 'fire-related' tweets, we distinguish between (1) relevant, well-formed tweets that provide actionable information about prescribed burn(s) or wildland fire(s) and (2) everything else.

Search terms submitted to the Twitter Search API were chosen such that the bulk of tweets returned were almost exclusively well-formed and relevant, following the method of Vieweg et al. (2010). The terms were case-insensitive and could be partially contained by other words (e.g. a search for 'fire' also returns 'FIRE' and 'wildfire'). The chosen query also contained words that were most likely to convey information about the location and size of the fire, such as U.S. state abbreviations and the following terms with associated abbreviations: road, highway, county, and acre.

Acreage burned was extracted from the tweet text using basic regular expressions. The date and time of the fire were taken from the tweet's metadata, assuming that the date and time the tweet was written was proximate to the date and time of the fire incident. This approach assumed that the majority of fire tweetsn on Twitter is about wildfires with one or more of a finite set of tokens that are exclusively associated with prescribed burns (e.g. 'prescribed' and '#RxFire').

2.1.1 Retweet detection to determine single fire events

Only unique reports describing wildfires or prescribed burns were of interest, to avoid documenting the same fire twice in order to produce accurate estimates of burned area and to reduce the overall number of tweets that needed to be processed. Once a tweet was determined to sufficiently describe a fire, we aimed to filter out all later tweets describing the same fire - including exact or near duplicates of an earlier tweet. These so-called 'retweets' had no reliably definitive textual signifiers or the metadata to distinguish them from original compositions. The common convention of using 'RT 'within a tweet to signify a retweet was not universally adopted across the Twitterverse. Our approach determined the similarity, or distance, between two tweets as if they were arbitrary points in some feature space so that tweets within a certain distance of one another indicated the one or the other must be a retweet.

We employed locality-sensitive hashing (LSH) called minhashing (Moulton 2012; Jaffer 2013) to compare the text of any two tweets, which is also used by Twitter Inc. for search engine optimization (Twitter 2011). If the tweets were authored by two different Twitter users, the younger (more recent) tweet was marked as a retweet.

2.1.2 Geocoding of fire tweets

Some tweets were explicitly geotagged, with geographic coordinates derived from global positioning system (GPS) chips in a mobile device. This geocoding was appended to the metadata of each tweet. However, in practice, geocoded tweets account for less than 1% of all tweets in the Twitterverse (Cheng et al. 2010, Lee et al. 2011). Therefore, the majority of tweets for any data mining application must be geocoded. We defined geocoding for this project as an attempt to determine the real-world coordinates of an entity from unstructured or loosely-structured data. Geographic coordinates were assigned to a tweet based on its textual content, specifically, the tweet text and some metadata from the user profile (after Leetaru et al. 2013). The approach used in this study was a synthesis of named-entity recognition (NER) and gazetteering with clustering to resolve ambiguous cases. NER is an NLP technique that chunks adjacent words into meanings (such as 'location' or 'person') based on POS. NER is used to identify words or groups of words with potential geographic significance—so-called toponyms, or words that might be successfully mapped to a place through gazetteering.

Gazetteering is the process of using a geographical lookup table (gazette) to associate place names with their coordinates. Each toponym is then searched for in the geographic gazette. We used the GeoNames collection (GeoNames.org) as the geographic gazette, which contains over 2.1 million geographic entities in the United States and U.S. territories. If more than one match is found per tweet (i.e., if more than one term in the tweet is found in the gazette) and if those matches are not duplicates of the same gazette entry, then the tweet has ambiguous geographic coordinates. Only one set of coordinates can be assigned to the tweet, so k-means clustering was used to pick a centroid of geographic locations based on all of the retrieved locations. The developed algorithm searches for two clusters in k-means clustering (k=2) as it assumes that it is retrieving two types of entries from the gazette: those that describe the true location of the fire and those that describe faraway places with similar names (e.g. Riverside, California versus Riverside, Iowa). In k-means clustering, the centroid of the 'tightest' cluster is assigned as the geographic coordinates of the tweet where 'tightest' is determined as the minimum total distance between cluster members.

2.2 Extracting earthquake information from structured text

While similar NLP methods can be used to evaluate the extent of an area affected by an earthquake, certain key details regarding an earthquake event, namely the location and magnitude of the quake at the epicenter, are more standardizable than in the case of fires. Volunteer reporting on social media regarding earthquakes has been practiced and improved by various individuals and organizations (USGS 2012; Wendel 2015). While this diminishes the capacity for social media to perform the actual detection of an emergency, social media can still provide information regarding the extent of the

affected area. The earthquake data mining framework developed here seeks to integrate this work that has already been done by public services and scientists.

The USGS provides up to date information on earthquakes that documents its detection of a seismic event, including features such as depth, locational uncertainty, and distances from nearby populated geographical features (USGS 2012). All of this information is included as text or via a link generated by a USGS Twitter account (@USGSted). All of this information builds a much more precise query that can be made at the Twitter API, with a keyword search that is delimited in both time and space.

In this project, the Twitter Search API was queried for any and all tweets relating to earthquakes located in and around southern California. This query was conducted between January 2014 and December 2014. Initial earthquake locations were determined from the epicenter metadata provided by the @USGSted Twitter account. Similarly, earthquake-impacted areas for each individual earthquake were estimated from the distances and/or nearby cities metadata. Twitter queries were targeted within these estimated earthquake-impacted areas. Tweets from the earthquake-impacted area were recorded, counted, and then mapped on Google Maps and ESRI ArcGIS to classify the total number of tweets coming from populated towns or cities within these areas. Fig. 2 shows an example from an earthquake detected in November 2014 nearby Anchorage, Alaska.
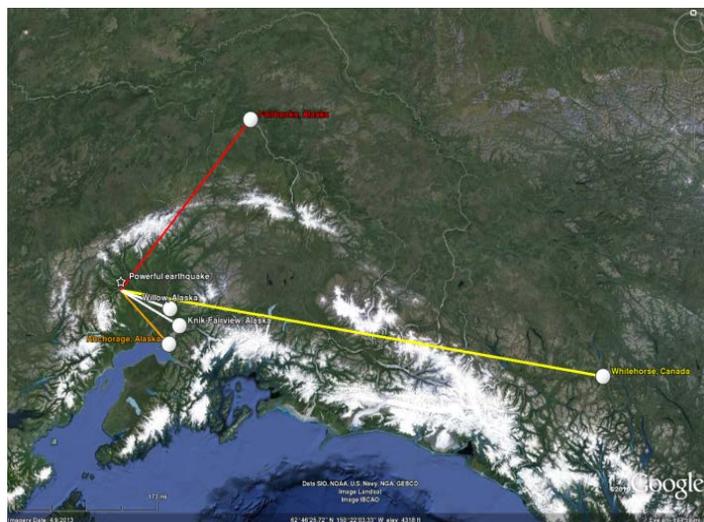
Figure 1: The results of the earthquake Twitter approach for a November 2014 earthquake near Anchorage, Alaska.

2.2.1 The expanding geographic gazetteer: challenges of inherited structure data

One of the challenges of this method was the ever-expanding geographic gazetteer. While initially this Twitter query approach was targeting southern California, earthquake data from the USGS included global locations. During the study period,

there were no significant seismic activities within southern California, so the approach to include all English-languages from global locations of earthquake activity. Like the wildland and prescribed fire study, this algorithm would need to develop a custom gazette for individual regions of interest, implemented through geolocation metadata analysis (e.g., human evaluator) or machine learning.

While building on available USGS data is a more direct way of obtaining well-formed and relevant earthquake-related tweets, relying on the classification schema provided by the Twitter API would cause relevant tweets that lack the metadata for full classification of earthquake event to be missed. For example, if a Twitter user reporting on earthquake damages does so from an account that does not have geotagging enabled, then the API query would not return this tweet as potential relevant variable within the established earthquake-impacted areas. As was discovered in the fire data mining analysis, the overall lack of geocoded tweets further complicated geospatial queries and filtered out potentially relevant data for mapping outputs (i.e., no further NLP to discover textual location).

3. RESULTS

3.1 Lesson Learned: Custom Geographic Gazetteer

Assessment of the geocoding algorithm performance for the wildland and prescribed fire data mining algorithm involved a human evaluator who manually attempted to determine the location of 60 tweets randomly sampled from those located successfully (if not accurately) by the algorithm, excluding those explicitly geotagged. Of the original 60, the human evaluator was able to determine the unambiguous location of the fires in 32 tweets. The evaluator employed any information in the tweet and on the web to learn where the fire referred to within the text of the tweet was located. This includes sources of information not used by the algorithm, such as web pages linked in the tweet or official fie webpages from federal, state, and local agencies. The intended effect was to compare the performance of an algorithm to the best geocoder available: human intuition applied to the largest library of spatial information available (the internet).

For each of the 32 tweets in the wildland and prescribed fire tweets sample, the Vincenty distance (Vincenty 1975) between the actual and geocoded locations were calculated and compared to one of 32 random locations within exactly the same geographic extent allowed by the geocoder. Two clustering schemes were also compared, resulting in three assessments: a random geocoding, geocoding using the defined algorithm with k-means clustering (k=2), and geocoding using the defined algorithm with pair-wise clustering. The results are displayed as a flipped cumulative frequency plot in Fig. 2.
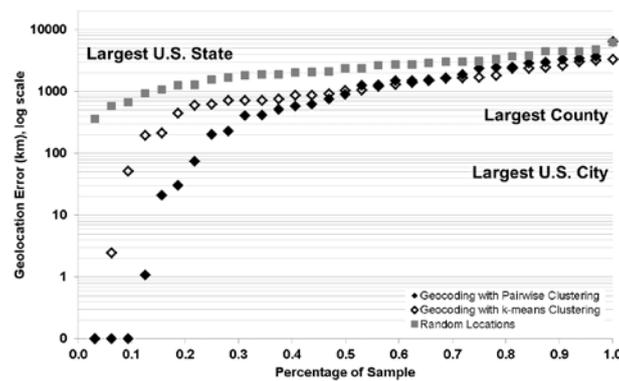
Figure 2: The geocoding performance of the algorithm developed in this study is compared to that of a random geocoding for two different clustering schemes. Though both clustering schemes perform better than random, the proprietary clustering scheme, pairwise clustering, performs significantly better than k-means.

In the case of the fire data mining application, the automated geocoder of tweets performed better than chance with either clustering scheme. However, there was significant room for improvement. With pair-wise clustering, the best clustering in terms of geocoding performance, less than 10% of the geocoded tweets in the sample were geocoded to within 8 km of the actual location of the fire. This distance was a significant threshold as it represented the approximate instantaneous field-of-view (IFOV) of the Moderate Resolution Imaging Spectroradiometer (MODIS) sensor at swath's edge (in the 1 km resolution data), which ranges from approximately 4.83 km to 9.71 km (Yang and Di 2004). The 1 km MODIS Active Fire Product is still used by the U.S. Forest Service for early warning fire mapping (http://activefiremaps.fs.fed.us/) and in the NASA Earth Observing System Data and Information System (EOSDIS; https://earthdata.nasa.gov/data/near-real-time-data/firms/active-fire-data). Roughly 15% of the samples were geocoded to within 22 km of the true location and 22% to within 100 km.

3.1.1 The 'gold standard': Human evaluators still needed

A performance evaluation of the fire data mining methods was performed using human evaluators, which are justifiably the 'gold standard' in tweet text interpretation. Evaluators used in this project were all research professionals familiar with both the geospatial and fire sciences. In each evaluation, the results of the algorithm were compared to that of one human evaluator (in some cases, out of a group) over a random sample of tweets that had been identified as non-retweets—that is, each performance evaluation is an assessment in light of the retweet detection algorithm having been applied.

Assessment of the retweet detection algorithm for the fire data mining application involved a manual classification of 120 tweets from a simple random sample of all tweets, all of which were classified by the algorithm. The evaluator considered as a retweet any tweet for which there could be found an earlier tweet of sufficient

similarity. The 'sufficient similarity' criterion is, of course, based on the human evaluator's own intuitive assessment. The human classifier's results were compared to the retweet detection algorithm. The algorithm achieved an overall accuracy of 72% with a recall, or true positive rate, of 82%. The majority of the error is commission error. Almost half of the truly independent tweets were mistakenly classified as retweets by the algorithm. The naïve approach to the problem of retweet detection is predicated solely upon the presence of the 'RT' token in the tweet text. A comparison of this approach was made to the naïve approach. While the naïve approach's omission error indicates that some retweets lack the 'RT' token, the overall accuracy of the naïve approach exceeded 83%.

This research found that region-specific earthquake detection data mining applications still need human evaluators to extract contextual location information from streaming tweets. As most earthquake epicentre data is provided automatically from a global system of seismological sensors and volunteer scientists, determining earthquake-impacted areas would require more than the 1% tweets to be geodocoded and/or a sophisticated interpretation of location from the text of the tweets themselves. This research has shown that sophisticated machine learning and/or trained human evaluators necessary for extracting geospatial information from tweets.

3.3 Lesson Learned: More rigorous integration of metadata

Tweets are generally considered unstructured text in data mining applications. However, public service announcements on social media often contain information in a consistently structured format (USGS 2012). While the computational price may become prohibitive with regard to machine learning and making common sense inferences from unstructured text, maintaining compatibility with reliable data sources and structures could provide valuable context to the information parsed from the unstructured data. This would entail that both the Twitter API and governmental agencies using Twitter to facilitate the transfer of data (i.e., @USGSted) must use consistent metadata structure and labeling as well as provide details on deprecation of the metadata schema. Future applied research related gleaning natural hazard-related information from Twitter should focus on the selection and evaluation of machine learning algorithms and techniques for discriminating relevant tweets and their approximate geolocation.

3.4 Lesson Learned: Data mining Twitter can produce geospatial data of natural hazard occurrences unknown to the satellite record

The geocoded results of the prescribed and wildfire data mining application were further compared to the satellite record. The positional accuracy of the tweet geocoding against the MODIS Active Fire Product (Giglio et al. 2003) and the accuracy of tweet-derived burned area estimates against the MODIS Burned Area (MCD45A1) Collection 5.1 product (Roy et al. 2008). In terms of positional

accuracy, only about 4% of the geocoded tweets were located within 8 km of the corresponding fire's location in the MODIS Active Fire record. Approximately 33% of the geocoded tweets in our collection correspond with the coincident MODIS-based MCD45A1 burned area estimates. This implies that the targeted fire tweet collection, equal to 1,697 tweets from July 2012 to June 2013, likely contained references to fires potentially unknown and currently missed in the satellite record.

3.5 Lesson Learned: Systematically re-evaluate search terms and NLP corpus

Fig. 3 shows the word cloud produced by analyzing the most common non-trivial words in the collection of wildland and prescribed fire tweets. Future improvements to the Twitter Search API query would better integrate key terms based on their prominence in a word cloud visualization. Additionally, future data mining activities aimed at differentiating types of wildland and prescribed fire (e.g. forestry vs. agriculture vs. rangeland) would benefit from word cloud visualizations to determine best key terms and/or terms causing false detections.

Figure 3: A word cloud based on the tweets collected provides a view of the common terms used in fire-related tweets within the contiguous United States and can be helpful for designing new or improved queries to the Twitter Search API.

Further improvements could be made to any social media-based data mining prototype through outreach efforts to educate current fire and land managers, scientific community, citizen scientists, and the general public to safely share their observations of prescribed and wildland fires on social media. This type of outreach would need to provide appropriate and standardized key terms to identify location, burned area, and fire type. However, this type of outreach would require systematic evaluations of search terms or text within the NLP corpus to ensure the right keywords and/or terms are being communicated. Moreover, this type of outreach – established through analysis of continuously evolving search terms - could further enhance the value of social media data mining for natural hazards managers seeking to monitor earthquakes, earthquake damage zones, wildfire outbreaks, fire conditions, and post-fire regrowth from crowd-sourced information.

4. CONCLUSION AND DISCUSSION

This paper has presented data mining methodologies used to detect and map fire and earthquake-related information from Twitter. Any geocoding approach needs to be rigorously tested before implementation in an operational natural hazard event detection system. In general, the geocoding error can be attributed almost entirely to failures in disambiguation between common toponyms (Leetaru et al. 2013). Multiple solutions to this problem are now known but perhaps the most appropriate and effective for this application would be to construct a custom geographic gazette. Starting with the GeoNames.org database, for instance, future data mining work would benefit from filtering out entities that do not meet certain criteria (e.g. population, feature type, urban versus rural index, landscape and/or locational suitability indices for targeted natural hazards). Alternatively, ambiguous results could be ranked by these criteria, which allows for either fuzzy or discrete matching of geolocation for targeted tweets. Including the contextual information of the location and biography from the author's Twitter profile could also be used to further refine custom geographic gazetteers. Leetaru et al. (2013) describe both custom gazette generation and these disambiguation approaches, comparing the effect on accuracy in choosing between geocoding on tweet text, the profile biography, and the author's stated location. They found that the majority of accurate gazette matches came from the stated location in the author's Twitter profile. While this location would be too coarse for mapping wildland and prescribed fires, it could be used to disambiguate results from full-text retrievals based on the tweet text and/or biography text and would likely be adequate for earthquake and earth impact detection.

The wildland and prescribed fire prototype described in this paper treats incoming text as unstructured data and only classifies specific information regarding the event and its location. It does not take into account how information may already be structured within the data. In addition to helping to curb geocoding error, providing a more distributed classification schema would offer more options for discerning relevant tweets and even integrating further information. Future applied research related gleaning natural hazard-related information from Twitter should focus on the selection and evaluation of machine learning algorithms for discriminating relevant tweets.

REFERENCES

Cheng Z, Caverlee J, Lee K (2010) You are where you tweet: A content-based approach to geo-locating Twitter users. *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. Toronto, Canada.

Endsley KA, McCarty J L (Accepted) Mapping prescribed burns and wildfires from Twitter with natural language processing and information retrieval techniques. *Proceedings of the International Smoke Symposium*. Hyattsville, Maryland, USA.

Giglio L, Descloitres J, Justice CO, Kaufman Y (2003) An enhanced contextual fire detection algorithm for MODIS. *Remote Sensing of Environment* **87**, 273-282. doi: 10.1016/S0034-4257(03)00184-6.

Jaffer, S (2013) Finding similar items using minhashing. Available at http://www.toao.com/posts/finding-similar-items-key-store-minhashing.html October 19, 2013. [Verified 9 January 2014]

Lampos V, Bie T, Cristianini N (2010) Flu detector - Tracking epidemics on Twitter. *Machine Learning and Knowledge Discovery in Databases* **6323**, 599–602. doi:10.1007/978-3-642-15939-8_42

Lee CH, Yang HC, Chien TF, Wen WS (2011) A novel approach for event detection by mining spatio-temporal information on microblogs. *2011 International Conference on Advances in Social Networks Analysis and Mining,* 254–259. doi:10.1109/ASONAM.2011.74

Leetaru K, Wang S, Cao G, Padmanabhan A, Shook E (2013) Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday* **18**. doi:10.5210/fm.v18i5.4366

Longueville BD, Smith RS, Luraschi G (2009) 'OMG, from here I can see the flames!': A use case of mining Location Based Social Networks to acquire spatio-temporal data on forest fires. *Proceedings of the 2009 International Workshop on Location Based Social Networks*. Seattle, Washington, U.S.A.: Association for Computing Machinery.

MacEachren AM, Robinson AC, Jaiswal A, Pezanowski S, Savelyev A, Blanford J, Mitra P (2011) Geo-twitter analytics: Applications in crisis management. *Proceedings, 25th International Cartographic Conference.* Available at http://nzdis.org/projects/projects/berlin/repository/revisions/61/entry/trunk/MastersDocs/Papers/Filer_Papers/MacEachren_ICC_2011.pdf [Verified 9 January 2014]

Moore H (2013, September 13) Twitter heads for stock market debut by filing for IPO. *The Guardian*. New York, NY, U.S.A. Available at http://www.theguardian.com/technology/2013/sep/12/twitter-ipo-stock-market-launch [Verified 9 January 2014]

Moulton R (2012) Simple simhashing: Clustering in linear time. *Ryan Moulton's Articles*. Available at http://moultano.wordpress.com/article/simple-simhashing-3kbzhsxyg4467-6/ [Verified 9 January 2014]

Petrovic S, Osborne M, McCreadie R, Macdonald C, Ounis I, Shrimpton L (2013) Can Twitter replace Newswire for breaking news? *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM). Association for the Advancement of Artificial Intelligence.* Available at http://homepages.inf.ed.ac.uk/miles/papers/short-breaking.pdf [Verified 9 January 2014]

Roy DP, Boschetti L, Justice CO, J Ju (2008) The Collection 5 MODIS Burned Area Product - Global Evaluation by Comparison with the MODIS Active Fire Product. *Remote Sensing of Environment* **112**, 3690-3707. doi:10.1016/j.rse.2008.05.013.

Russell MA (2011) Chapter 8: Blogs et al..: Natural Language Processing (and Beyond). In Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites. (O'Reilly Media, Inc.)

Stefanidis A, Crooks A, Radzikowski J (2013) Harvesting ambient geospatial information from social media feeds. GeoJournal, **78**, 319–338. Available at http://link.springer.com/article/10.1007/s10708-011-9438-2 [Verified 9 January 2014]

Twitter (2011) The engineering behind Twitter's new search experience. (I. Twitter, Ed.) *Engineering Blog*. Twitter. Available at https://blog.twitter.com/2011/engineering-behind-twitter's-new-search-experience [Verified 9 January 2014]

Twitter (2012) Using the Twitter Search API. (I. Twitter, Ed.) *Developers*. Available at http://dev.twitter.com/docs/using-search [Verified 9 January 2014]

Twitter (2013a) New Tweets per second record, and how! (I. Twitter, Ed.)*Twitter Engineering Blog.* Available at https://blog.twitter.com/2013/new-tweets-per-second-record-and-how [Verified 9 January 2014]

Twitter (2013b) Calendar of API changes. (I. Twitter, Ed.) *Developers*. Retrieved from https://dev.twitter.com/calendar [Verified 9 January 2014]

USGS (2012) U.S. Geological Survey: Twitter Earthquake Detector (TED). Available at http://recovery.doi.gov/press/us-geological-survey-twitter-earthquake-detector-ted/ [Verified 9 January 2014]

Vieweg S, Hughes A, Starbird K, Palen L (2010) Microblogging during two natural hazards events: what twitter may contribute to situational awareness. *ACM (Ed.), CHI 2010: Crisis Informatics* pp. 1079–1088. Atlanta, GA, U.S.A.: ACM. Available at http://dl.acm.org/citation.cfm?id=1753486 [Verified 9 January 2014]

Vincenty T (1975) Direct and Inverse Solutions of Geodesics on the ellipsoid with Application of Nested Equations. *Survey Review* **22**, 88-93.

Wendel, J (2015), Internet users act as earthquake trackers, Eos, 96, doi:10.1029/2015EO025457. Published on 2 March 2015.

Yang W, Di L (2004) An accurate and automated approach to georectification of HDF-EOS swath data. *Photogrammetric Engineering & Remote Sensing* **70**, 397-404.