

Mapping prescribed burns and wildfires from Twitter with natural language processing and information retrieval techniques

K Arthur Endsley^{A,B} and Jessica L McCarty^A

^AMichigan Tech Research Institute (MTRI), 3600 Green Court, Suite 100, Ann Arbor, MI, jmccarty@mtu.edu

^BCorresponding author e-mail: kaendsle@mtu.edu

Abstract: New media are increasingly used to capture ambient geographic information in multiple contexts, from mapping the evolution of the Tahrir Square protests in Egypt to predicting influenza outbreaks. 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, Twitter is used to communicate prescribed agricultural or other burning and the emergence, response to, and containment of wildfires. A prototype for operational prescribed and wildland fire detections from social media, specifically Twitter, was created using natural language processing and information retrieval techniques. The intent is to identify and locate prescribed burns and wildland fires in the contiguous United States often missed by satellite detections with the hope of providing relevant, spatio-temporal fire data for emission estimates, inventories and burned area mapping efforts. The authors present their methodology and an evaluation of its performance in collecting relevant tweets, extracting metrics such as area burned and geolocating the fire events using the GeoNames geographic gazette. Compared to two operational satellite fire products, this data mining effort mapped fires potentially unknown to the satellite record.

Additional keywords: social media, text mining, geocoding, prescribed fire, wildland fire, geospatial techniques

Introduction

The automated extraction of volunteered and ambient geospatial information from social media has proven to be useful in a variety of contexts, from mapping the evolution of the 2011 Tahrir Square protests in Egypt (Stefanidis *et al.* 2013) to predicting influenza outbreaks (Lamos *et al.* 2010). There is already a significant body of work showcasing the use of Twitter for disaster alerting, mitigation, and response (e.g. MacEachren *et al.* 2011; USGS 2012) where information extracted from the social media platform often performs equal to or better than conventional techniques (Petrovic *et al.* 2013). In fact, according to Sakaki *et al.* (2010), the operational earthquake alerting system in Japan provides alerts faster than the national Meteorological Agency. In the domain of prescribed burns and wildland fires, however, Twitter has previously

been used only for qualitative assessments and case studies (e.g. Vieweg *et al.* 2010), such as a single fire in France (Longueville *et al.* 2009).

Twitter's popularity for such applications can likely 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 (e.g. Australia and New Zealand, as seen during the 2013 New South Wales fires) to report prescribed burns in advance (sometimes as part of a reporting obligation) or to communicate emergence, response to, and containment of wildfires. These reports, like all Twitter messages (or 'tweets'), are limited to 140 characters of unstructured text.

A prototype framework was developed to collect messages from Twitter potentially describing wildfire or prescribed burn activity data, to extract from them information such as the type of fire and acreage burned, and to map the location of the fires in near-real time. The goal of this research was to assess the predictive power of information extracted from new media, specifically Twitter, for gleaning useful occurrence information for environmental and natural hazard applications. Specifically, the aim was to identify and locate known wildland and prescribed fire events in the contiguous United States (CONUS) often missed by satellite detections, thus providing relevant spatio-temporal fire data for emission estimates, ongoing burned area mapping efforts, and natural hazard management.

Materials and methods

Natural language processing

The general approach employed in this study requires the use of natural language processing (NLP), information retrieval (IR), and data mining techniques. The specific steps involved are, in this order: (1) To collect potential tweets of interest from the Twitter network; (2) Detect and remove duplicate tweets (or 'retweets'); (3) Determine the area burned, if possible; (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 words), *removing stopwords* (removing common and largely meaningless words like 'if, and, or, but') and parts-of-speech *tagging* are applied at almost every step of the process to facilitate computation (Russell 2011).

Tweets concerning potential fire activity were collected using the Twitter Search application programming interface (API), which allows access to millions of tweets going back approximately one week in time (Twitter 2012). While no estimate of average daily fire-related tweet volume is currently available, it is likely to be a small fraction of the approximately 5,700 tweets generated every second. In defining 'fire-related' we distinguish between (1) relevant, well-formed tweets that provide actionable information about prescribed burn(s) or wildland fire(s) and (2) everything else. Specifically, this research distinguishes between tweets such as, 'Enjoying a camp fire tonight; I love the smell of wood smoke!' (not well-formed or relevant) and, 'Prescribed fire today: 50 acres at Sebastian Tree Farm, near #Micco Road and US1' (both well-formed and relevant). Also of interest to this research is the avoidance of tweets such as, 'I see smoke outside! I wonder what's #burning?' which is relevant, but not well formed.

Search terms submitted to the Twitter Search API were chosen by the authors such that the bulk of tweets returned were almost exclusively well-formed and relevant (after Vieweg *et al.* 2010). The terms are case-insensitive and may be partially contained by the words of the tweet (e.g. a search for ‘fire’ also returns ‘FIRE’ and ‘wildfire’). The chosen query also contained words that are known to be associated with those fire-related tweets that are most likely to convey information about the location and size of the fire, such as state abbreviations and the following terms with their common abbreviations: *road, highway, county, acre.*

Given that, in this study, the tweets returned typically contain the search terms ‘acre’ or ‘acres,’ the acreage burned is extracted from the tweet text using basic regular expressions. The date and time of the fire are taken from the tweet's metadata, assuming that the date and time of authorship is sufficiently proximate to the date and time of the fire incident. This assumption is borne out by the authors’ experience reading fire-related tweets.

Fire type classification is assumed to be relatively straightforward. The classification rests solely on a tweet's inclusion of certain tokens (i.e. words) thought to be singularly representative of prescribed burns. This approach proceeds from the assumption that the majority of wildfire communication on Twitter is about wildfires and that less common communications about prescribed burns include one or more of a finite set of tokens that are exclusively associated with prescribed burns (e.g. ‘prescribed’ and ‘#RxFire’)

Retweet detection

For the purposes of this study, only unique reports describing wildfires or prescribed burns are of interest. The intent is to avoid documenting the same fire twice or more in order to produce accurate estimates of bulk acreage burned and to reduce the overall number of tweets that need to be processed. Therefore, once a tweet sufficiently describing a fire is captured, the aim is to filter out all later tweets describing the same fire. These might even be exact or near duplicates of an earlier tweet. We have observed that these so-called ‘retweets’ have no reliably definitive signifiers in the text or their metadata to distinguish them from original compositions. A common convention of using the letters ‘RT’ to signify a retweet is not universally adopted. The approach used in this study is to determine the similarity, or distance, between two tweets as if they were arbitrary points in some feature space. If the tweets fall within a certain distance of one another, then one or the other must be a retweet.

An efficient means of comparing the text of any two tweets is provided by an implementation of locality-sensitive hashing (LSH) called minhashing (Moulton 2012; Jaffer 2013), which is also used by Twitter Inc. themselves for search engine optimization (Twitter 2011). In a minhashing procedure, we compare sets of hashes made from the inclusive n-tuples (or n-grams) of the ordered words in the tweet text—sets like ‘50, acres, burned’ (a 3-gram or trigram) are mapped to a new text string or hash using a special, one-way mathematical relationship allowing for quick comparisons. If the tweets were authored by two different Twitter users, the younger (more recent) tweet is marked as a retweet.

Geocoding

This research aims to locate prescribed burns and wildland fires associated with relevant and well-formed tweets collected through the Twitter Search API. Some tweets are explicitly geotagged, meaning they are given geographic coordinates by their authors, usually from embedded global positioning system (GPS) measurements appended to the metadata by a mobile device. However, in practice, these account for less than 1% of all tweets (Cheng *et al.* 2010, Lee *et al.* 2011). Therefore, the majority of tweets must be geocoded. Geocoding is generally defined as an attempt to determine the real-world coordinates of an entity (e.g. a building, a topographic feature, a person) from unstructured or loosely-structured data. The classic example of geocoding is transforming postal addresses into latitude and longitude coordinates. Similarly, geocoding in this application is concerned with assigning geographic coordinates 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 is a synthesis of named-entity recognition (NER) and gazetteering with clustering used to resolve ambiguous cases. NER is an NLP technique that chunks adjacent words into meanings (such as ‘location’ or ‘person’) based on their parts-of-speech. Here, 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 searched for in the geographic gazette. The geographic gazette used in this study is the GeoNames collection (GeoNames.org), 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 clustering is used to pick a centroid of geographic locations based on all of the retrieved locations.

Two types of clustering were tested in this study, k-means clustering (with $k=2$) and a proprietary clustering scheme termed ‘pair-wise clustering’ which seeks to minimize the distance first between gazette entries that have the same toponym (e.g. the city of Riverside, California and Riverside County in California) and then to minimize the distance between each toponym centroid, arriving at a final centroid which is assigned as the tweet’s coordinates. 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.

Results

Twitter data mining prototype

A total of 13,241 tweets were collected starting July 4, 2012 and continuing until June 11, 2013 () when the version of the Twitter Search API being used at the time was deprecated (Twitter 2013b) and tweets could no longer be collected. The collection thus spans 342 days with an

average rate of 38 tweets per day and represents the contributions of 6,351 unique users. A performance evaluation of the methods described in the preceding section was performed over this dataset 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 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%.

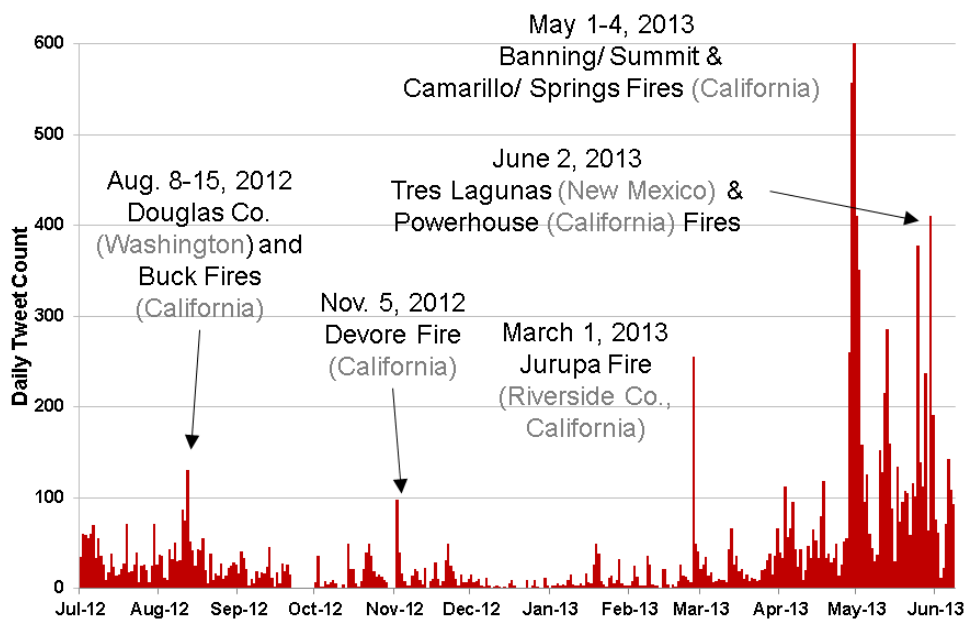


Fig. 1: Many notable fire events, as well as data holidays due to active development, can be seen in the time series histogram above which displays the total number of tweets each day over the period of study. Tweet volumes are higher in the early summer months.

One consequence of the query submitted to the Search API is that recovered tweets were tailored for the extraction of acreage burned. Thus, it is unsurprising that the overall accuracy of the method in correctly extracting acreage burned exceeded 99% in this study. We found that fewer than 1% of the tweets were classified as prescribed burns by our methodology. There were too few prescribed burns in the entire Twitter collection analyzed in this study, as verified by human evaluators, to provide a representative sample. Of the wildland fires described by tweets in this study, over 96% were correctly classified as wildfire.

To implement the detection of prescribed burns and wildfires from Twitter in near-real time, almost all of the delay between a fire's occurrence and its accurate mapping should be confined to the time required to author tweet describing it. To that end, we benchmarked the system's performance in processing 100 tweets at a stretch, which were randomly shuffled before each iteration, to measure the speed of our processing framework and to allow us to estimate the rate at which we can process tweets. On an Intel Core i7 CPU at 3.0 GHz, 100 tweets could be all but geocoded in 2.9 seconds. With geocoding, the time required jumps to 8.1 minutes for 100 tweets, which should hardly be surprising, as this step requires multiple database lookups (in the geographic gazette) per tweet. However, as the highest throughput of fire-related tweets to date seen in this study has been 600 tweets per day (less than 1 tweet every 2 minutes, on average), this performance meets and exceeds current requirements for near-real time mining of fire-related Twitter messages.

Geocoding performance

Assessment of the geocoding algorithm performance again involved a human evaluator who, in this case, rather than making a discrete classification, 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 was located. This includes sources of information not used by the algorithm, such as web pages linked in the tweet. The intended effect is to compare the 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 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

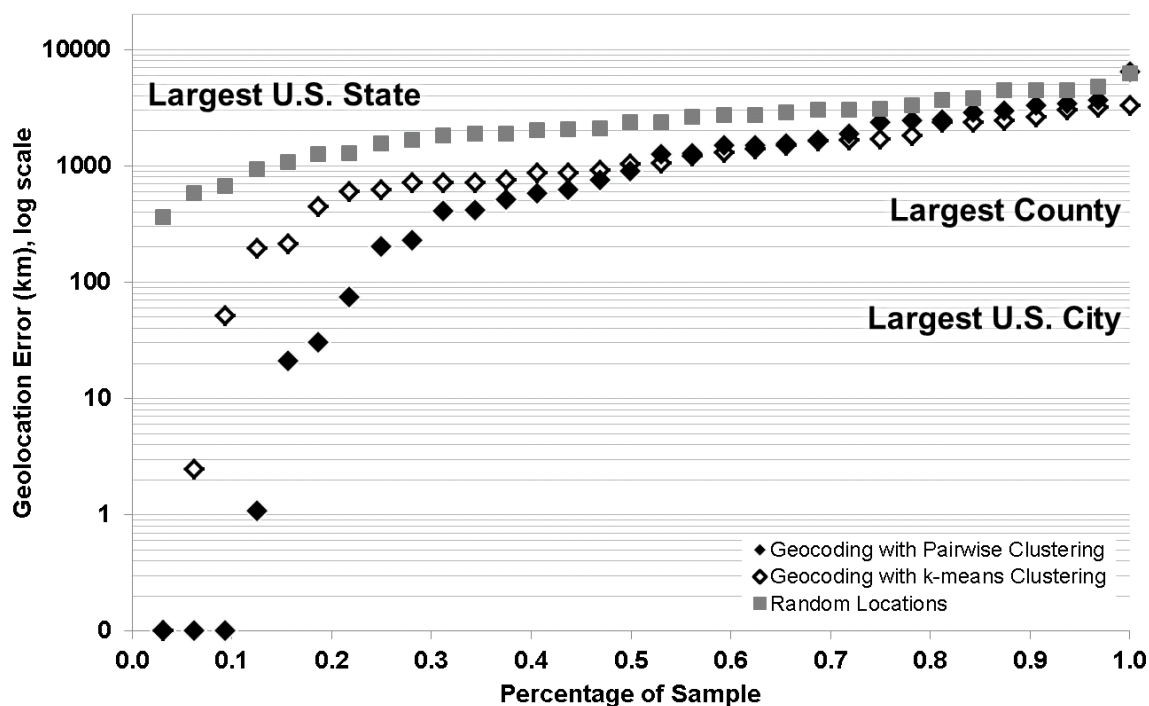


Fig. 2: The geocoding performance of the algorithm developed in this study is compared to that of a random geocoding for two different clustering schemes. Though both clustering schemes perform better than random, the proprietary clustering scheme, pairwise clustering, performs significantly better than k-means.

While the geocoder obviously performs better than chance with either clustering scheme, there is 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 is a significant threshold as it represents 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 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.

The geocoded results were further compared to the satellite record: the positional accuracy of our geocoding against the MODIS Active Fire Product (Giglio *et al.* 2003) and the accuracy of our 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 burned area estimates. This implies that the remaining collection, equal to 1,697 tweets from July 2012 to June 2013, may contain references to fires potentially unknown and currently missed in the satellite record.

Comparison with official fire activity and burned area

As estimates of burned area were also extracted from the Tweet text describing a fire, this study included a comparison with official burned area statistics from the National Interagency Fire Center (NIFC), which provides metrics on both wildfires and prescribed burns. According to the NIFC, approximately 4.09 million acres of wildfire burned between January 1 and October 15, 2013 (NIFC 2013). Over the same period, we estimate approximately 17 million wildfire acres burned; a 315% percent surplus in burned area, but nonetheless on the order of magnitude expected. As only a limited record of this study is available so far and the NIFC reporting is aggregated by year, direct comparisons over arbitrary ranges of time are not possible. This study did collect data during the bulk of the fire season in North America for 2012, from July 4 to December 31 of that year. The NIFC estimate for wildfire acres burned is 9.3 million, compared to our estimate of 7.6 million. The NIFC does provide ‘to date’ estimates, which allow for partial-year comparison with the continuous record currently produced by this study.

Prescribed burn acreage estimates were also compared to both the NIFC estimates and to the National Association of State Forester’s and the Coalition of Prescribed Fire Council’s 2012 National Prescribed Fire Use Survey Report (Melvin 2012). According to the latter report, 20.2 million acres of prescribed burning from forestry and agriculture took place in 2012 while the NIFC estimates burning from federal agencies that same year accounted for 1.9 million acres. The approach described in this study, which yielded only 114 tweets thought to describe prescribed burns, can account for only about 15,000 acres from July 4 to December 31, 2012. It is difficult to determine whether prescribed fires are not often being discussed on Twitter or if the collection of tweets in this study were not sufficiently representative of prescribed fire communication on Twitter.

Conclusions and discussion

This study describes a prototype framework for mapping prescribed burn and wildfire events based on Twitter messages in near-real time. The results can be explored in their spatio-temporal context online through an interactive web map () developed by the authors (<http://spatial.mtri.org/fireminer/>). The software architecture is based entirely on open-source software in the Python programming language and on established information retrieval and natural language processing techniques including named entity-recognition and gazetteering.

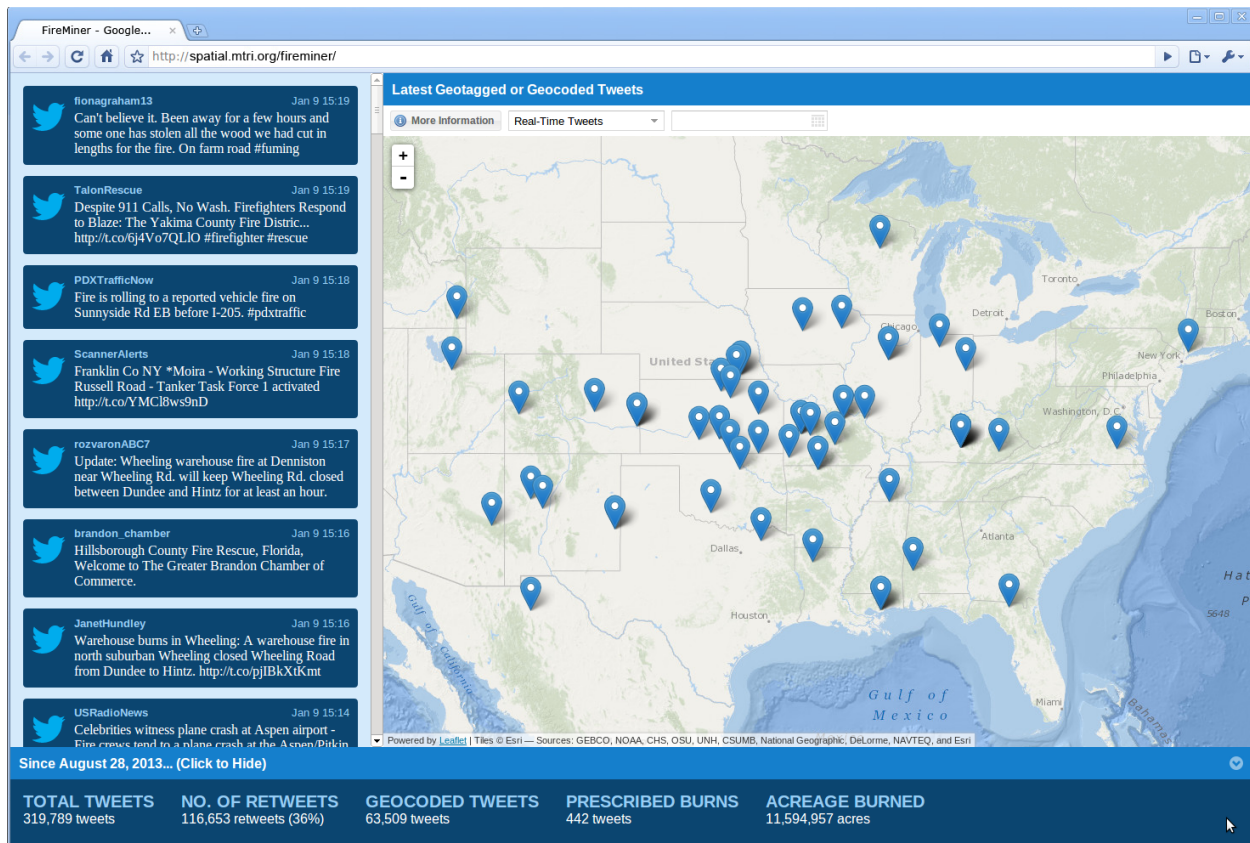


Fig. 3: The screenshot above displays the website where the tweets collected and geocoded during this study can be found, in addition to an ongoing, near-real time stream of tweets that are continuously analyzed.

The retrieval of fire-related tweets demonstrated high precision, as did the differentiation between retweets and non-retweets, such that very few irrelevant, misleading or redundant fire incident reports were admitted. The query submitted to the Search API was sufficiently and arbitrarily constrained such that virtually all of the tweets collected are well-formed and relevant to wildfires or prescribed burns, though this has not yet been quantified. The more interesting question that needs to be addressed, rather than measuring and incrementally boosting the performance of the methodology employed here, is: How can we distinguish between fire-related and non-fire related tweets in near-real time from a large, random sample of tweets?

Fig. 4 shows the word cloud produced by analyzing the most common non-trivial words in the collection of tweets from this study. Future improvements to the Twitter Search API query might 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 (forestry vs. agriculture as outlined in the National Prescribed Fire Survey) would benefit from word cloud visualizations to determine best key terms and/or terms causing false detections. 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 with the appropriate key terms to identify location, burned area, and fire type. This type of outreach could further enhance the value of social media data mining for fire managers seeking to monitor wildfire outbreaks, conditions, and post-fire regrowth from crowd-sourced information.



Fig. 4: A word cloud based on the tweets collected provides a view of the common terms used in fire-related tweets and can be helpful for designing new or improved queries to the Twitter Search API.

While the geocoding performance, in particular, needs to be significantly improved before implementation in operational fire event detection and reporting, 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 work might benefit from filtering out entities that do not meet certain criteria (e.g. population, feature type, urban versus rural index). Alternatively, ambiguous results could be ranked by these criteria, which allows for either fuzzy or discrete matching. In addition, this study did not consider other contextual information available such as the location and biography from the author's Twitter profile. 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 fires, it could be used to disambiguate results from full-text retrievals based on the tweet text and/or biography text.

Future applied research related to this project should focus on improving disambiguation in the geocoding and on the selection and evaluation of machine learning algorithms for discriminating between fire-related and non-fire related tweets. Significant gains in these areas would allow for an operational wildfire and prescribed burn detection system that operates in near-real time on large volumes of tweets, such as those delivered by the real-time Twitter Streaming APIs. Such a system would complement the satellite remote sensing record, particularly concerning small fires (< 200 acres) and prescribed burns that are not captured by existing emission inventories. Current attempts at more comprehensive emission inventories, like SmartFire 2 (Raffuse *et al.* 2011), can be labor and time intensive to incorporate ground reports from state and federal agencies (i.e., ICS-209 reports). This study illustrates that social media can be used to inform and possibly expand current fire emission inventories to create a comprehensive emissions inventory that includes fires from the satellite records, state and federal reports, and verified data from social media.

Acknowledgments

This research was made possible through internal research and development funding at the Michigan Tech Research Institute (MTRI). The authors would like thank research interns David Banach, Scott Hawkins, and James Whitehead for their assistance as human evaluators.

References

- Cheng Z, Caverlee J, Lee K (2010) You are where you tweet: A content-based approach to geolocating Twitter users. *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. Toronto, Canada.
- 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]
- Lamos 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]
- Melvin M (2012) 2012 National Prescribed Fire Use Survey Report. Available at http://www.stateforesters.org/sites/default/files/publication-documents/2012_National_Prescribed_Fire_Survey.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://moultono.wordpress.com/article/simple-simhashing-3kbzhsxyg4467-6/> [Verified 9 January 2014]
- NIFC (2013) National Interagency Fire Center: Statistics. Available at: http://www.nifc.gov/fireInfo/fireInfo_statistics.html [Verified 9 January 2014]
- Petrovic S, Osborne M, McCreddie 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]
- Raffuse SM, Larkin NK, Pryden DA, TM Dedecko (2011) New Methods for Modeling and Monitoring Wildfires Using Multiple Data Sources: SmartFire v2. *AGU Fall Meeting Abstracts - 2011*. Available at <http://adsabs.harvard.edu/abs/2011AGUFMNH53A1717R> [Verified 10 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.
- Yang W, Di L (2004) An accurate and automated approach to georectification of HDF-EOS swath data. *Photogrammetric Engineering & Remote Sensing* **70**, 397-404.