Discovering Credible Events in Near Real Time from Social Media Streams

Cody Buntain
Supervised By Dr. Jen Golbeck
Dept. of Computer Science
University of Maryland
College Park, Maryland 20742
cbuntain@cs.umd.edu

ABSTRACT

My proposed research addresses fundamental deficiencies in social media-based event detection by *discovering* high-impact moments and evaluating their credibility rapidly. Results from my preliminary work demonstrate one can discover compelling moments by leveraging machine learning to characterize and detect bursts in keyword usage. Though this early work focused primarily on language-agnostic discovery in sporting events, it also showed promising results in adapting this work to earthquake detection. My dissertation will extend this research by adapting models to other types of high-impact events, exploring events with different temporal granularities, and finding methods to connect contextually related events into timelines. To ensure applicability of this research, I will also port these event discovery algorithms to stream processing platforms and evaluate their performance in the real-time context. To address issues of trust, my dissertation will also include developing algorithms that integrate the vast array of social media features to evaluate information credibility in near real time. Such features include structural signatures of information dissemination, the location from which a social media message was posted relative to the location of the event it describes, and metadata from related multimedia (e.g., pictures and video) shared about the event. My preliminary work also suggests methods that could be applied to social networks for stimulating trustworthy behavior and enhancing information quality. Contributions from my dissertation will primarily be practical algorithms for discovering events from various social media streams and algorithms for evaluating and enhancing the credibility of these events in near real time.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining;
H.3.3 [**Information Search and Retrieval**]: Information Filtering

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2741752.

Keywords

event detection; twitter; social networks; credibility analysis

1. PROBLEM

Social media's ubiquity has had a profound effect on the ways in which we share experiences and information. While much of this shared content might be inane pictures of today's outfit or tonight's dinner, social media has proven an important tool during times of crisis and mass unrest. Such a powerful communication tool is not without its hazards, however, in its ability to spread misinformation and unease. This dichotomy is made clear in the events surrounding the Boston Marathon bombings in April of 2013.

Demonstrating social media's benefits, a Harvard report lauded the Boston Police Department's (BPD) effective use of Twitter "to keep the public informed about the status of the investigation, to calm nerves and request assistance, to correct mistaken information reported by the press" [3]. Additional research showed a significant number of Twitter messages, or tweets, posted immediately after the explosions contained valuable information on severity and location that could help guide first responders [1]. As hinted at in Harvard's report though, the shear volume of *incorrect* information being published in both social and traditional media warranted the BPD's direct intervention. This misinformation ranged from the number of bombs found to whether suspects had been arrested [17]. Similar issues of quality appeared during the 2010 Chilean earthquake, in which the volume of misinformation and rumor being spread on social media became a contributing factor to the sense of chaos and unease surrounding the events, as described by Mendoza et al. [10].

Clearly, it is dangerous to assume all posts on social media are true and accurate; yet, many organizations make just this assumption. This mistake is understandable given social media's volume during such crises, especially when traditional media organizations lack adequate on-the-ground journalistic resources (as is often the case in unanticipated, high-impact events). Unfortunately, in these tense moments, traditional media also lack sufficient capabilities to evaluate credibility in this (mis)information. In these instances, one has to balance *timely* reporting with *accurate* reporting, and when the need to get information out rapidly is great, accuracy often suffers.

While researchers have sought to address these credibility issues in social media [8, 2, 6, 14, 4], assessing credibility in real time still presents a significant gap in the literature. Coupled with this lacuna is an inability to discover the unanticipated but truly important events taking place around such crises. Preliminary work in applying machine learning and trust analysis to social media suggests detecting credible events in or near real time is feasible. Therefore, my dissertation explores the following:

> **By integrating machine learning and high-volume streams across social media networks, one can detect high-impact events, identify specific occurrences within those events, and evaluate credibility of those occurrences in near real time.**

2. STATE OF THE ART

A significant body of research supports the research questions and approaches detailed herein, as others have attempted to address pieces of our objective. This body is well separated into two distinct areas: event detection and credibility analysis.

2.1 Event Detection in Social Media

Detecting events by leveraging digital media has fascinated researchers for over twenty years, with new methods, breakthroughs, and technologies emerging every few years. This subfield has evolved to integrate the latest available techniques and data sources, starting from early digital newsprint in the mid 1990s to blogs in the mid 2000s and now social media. Many techniques in this research fall into one of two categories: does the technique target online or retrospective event detection? Retrospective detection is valuable for understanding the patterns of an event or series of events after the fact (e.g., investigating how a mass protest was organized, or how news of an event spread). The research detailed herein focuses on online/real-time detection as a means to support journalists, first-responders, and other decision makers who are in need of rapid information during times of crisis.

One of the most well-known works in detecting high-impact events online is Sakaki, Okazaki, and Matsuo's 2010 paper on detecting earthquakes in Japan using Twitter [16]. Sakaki et al. show that not only can one detect earthquakes on Twitter but also that it can be done simply by tracking frequencies of earthquake-related tokens. Surprisingly, this approach can outperform geological earthquake detection tools since digital data propagates faster than tremor waves in the Earth's crust. Though this research is limited in that it requires pre-specified tokens and is highly domain- and location-specific (Japan has a high density of Twitter users, so earthquake detection may perform less well in areas with fewer Twitter users), it demonstrates a significant use case and the potential of such applications.

Petrović and his colleagues' research on clustering in Twitter avoids the need for seeding keywords by instead focusing on the practical considerations of clustering large streams of data quickly using Locality Sensitive Hashing (LSH). While typical clustering algorithms require distance calculations for all pairwise messages, LSH facilitates rapid clustering at the scale necessary to support event detection in Twitter streams by restricting the number of comparisons between tweets to only those within some threshold of similarity. Once these clusters are generated, Petrović was able to track their growth over time to determine impact for a given event. This research was originally unique in that it was one of the early methods that did not require pre-specified seed tokens for detecting events and has been very influential in the field, resulting in a number of additional publications to demonstrate its utility in breaking news and for high-impact crisis events [11, 12, 15]. That being said, Petrović's work relies on semantic similarity between tweets, which limits its ability to operate in mixed-language environments.

Finally, the most recent work relevant here is the 2013 paper by Xie et al. on TopicSketch [18]. Like my investigations, TopicSketch's authors seek to perform real-time event detection from Twitter streams "without pre-defined topical keywords" by maintaining acceleration features across three levels of granularity: individual token, bigram, and total stream. As with Petrović's use of LSH, Xie et al. leverage "sketches" and dimensionality reduction to facilitate the event detection task and also relies on language-specific similarities, but Xie et al. also focus only on tweets from Singapore rather than the worldwide stream. In contrast, our approach is differentiated primarily in its language-agnosticism and its use of the unfiltered stream from Twitter's global network.

2.2 Credibility in Social Media

Understanding credibility online has grown into a rich area of research in parallel with the growth of the Internet and has naturally focused on social media sites as they have increased in popularity, use, and scope. Of particular interest is understanding user credibility during times of crisis as mentioned earlier with the work of Mendoza et al. on the 2010 Chilean earthquake [10]. Mendoza et al. followed up this work in 2011 with automated methods for applying their lessons learned [2]. This study leveraged information propagation (via retweet analysis), user certainty, use of external sources (i.e., URL linking), and user characteristics like follower count as features for a supervised machine learning system capable of classifying high- and low-quality information. The authors relied on Amazon's Mechanical Turk to generate a labeled set of tweets with varying degrees of plausibility/credibility and used decision trees to learn this classification task with 86% accuracy. From these features and their classifiers, they concluded the most credible data on Twitter would generally start with one to a few users and exhibit deep re-tweet networks. This result, however, is somewhat contradicted by more recent events like the various instances of compromise of otherwise credible accounts.

At the same time, Qazvinian et al. created a data set of tweets and Twitter users replete with labels of which tweets were rumors versus non-rumors and which users were rumor believers/propagators versus rumor disbelievers [13]. From this data set, the authors built a framework capable of classifying tweets and users accordingly. Their classifiers leveraged features of tweet content and constituent parts of speech, network structure with respect to original tweets versus retweets, and embedded entities like hashtags and external references. Qazvinian et al. then demonstrated feasibility in using Bayesian and log-linear models built around these features to discriminate between rumors and non-rumors successfully. This work is particularly interesting in the context of the research presented in this prospectus because it focuses more on the quality of informational tweets rather than particular users.

It is worth noting here that, in addition to work on credibility in social media, an extensive corpus of research covers issues of trust in users. A number of algorithms attempt to gauge user trustworthiness, potential trust between users, or recommend items based on trust between users [7, 9, 5]. While these resources are certainly important, the research presented in this prospectus focuses more on aggregate analysis of credible information from many users rather than evaluating trust between two specific users.

3. APPROACH

My research covers two major areas: real-time event discovery and rapid credibility analysis. First, I have already developed a burst-oriented technique for identifying high-impact moments from social media streams, called LABurst. The LABurst algorithm is novel in that it does not require domain-specific information (e.g., pre-specified keywords) or language models (often needed for stop word filtering or normalization) (see Section 5). I plan to follow up that work by integrating stream-centric processing frameworks like Apache Storm (as used by Petrović et al.) or Apache Spark Streaming to support real-time event discovery. These two platforms are freely available and provide significant real-time processing potential for research systems. As such, I plan to implement LABurst on at least one, but preferably both, of these systems to evaluate potential and the feasibility of language-agnostic real-time event discovery.

The true novelty of my research, however, comes from my concentration on rapid credibility assessment in social media. Current technology allows traditional media outlets to source information from social media sources, but the credibility of this information is often unavailable and at worst very poor. Fortunately, social media has a number of potential features one can mine in evaluating information credibility. Characteristics of social network structure, information source location, and multimedia availability/metadata all might prove useful indicators in information quality. Machine learning algorithms offer a suite of tools for integrating these features into credibility rankings and metrics that could augment social media analysis.

An initial path I plan to explore is integrating short-term interaction network analysis. Intuition suggests significant differences may exist in how different types of information spread through social networks. Endogenous phenomena like memes may spread primarily through retweets whereas exogenous phenomena instigated by real-world events might spread through loosely coupled networks of users with limited social overlap. Analyzing short-term interaction networks of mentions and retweets might illuminate these structural differences without reliance on knowledge of the entire social graph (which is likely too big to obtain or process in a reasonable amount of time).

It is possible, however, that the highly restricted view of Twitter's interaction graphs imposed by the sparsity of the 1% public Twitter stream might limit the feasibility of this approach. To address this possibility, we also plan to integrate primary-source multimedia evidence. For instance, Twitter and Facebook both provide mechanisms for sharing photos easily, and networks like Instagram and Pinterest focus almost exclusively on this multimedia content. Given this data, it may be possible to leverage event-related keywords and text messages that include multimedia to identify photos related to specific events. The very existence of this additional content may support discriminating between trustworthy real-world events and memes, but this content also can have extensive meta-data attached to it (e.g., location, type of camera used, orientation, and date/time). By integrating these additional sources, we may be able to enhance credibility analysis *and* provide enhanced understanding of events by augmenting textual summaries.

Finally, only approximately 1% of Twitter data is tagged with geolocation information, which leads to a poverty of stimulus for algorithms that rely on location for analysis. As a result, a number of approaches exist to infer the location of a message from features like locations of similar textual content or locations of users mentioned or with whom the author interacts. Though this problem is well studied, many approaches' applicability are limited here given our near real-time constraints; for instance, inferring location from retweets would be problematic given the limited number of retweets occurring very soon after a high-impact event. Therefore, we plan to augment this inference task with additional geolocation data extracted from multimedia metadata. By integrating these location-inference algorithms into our work, existing research shows it may be feasible to infer both user and event locations from social media. This information could be very useful in assessing trustworthiness as, intuitively, an event-related message sourced from the same location as the event should be more trustworthy than a message related to that same event but posted from hundreds of miles away.

4. EVALUATION

Given my two main tasks of event discovery and credibility analysis, my planned experiments revolve around analyzing prior high-impact events that already have significant coverage by traditional media outlets. In particular, I plan to extract events from several important events, like the Boston Marathon Bombing, Westgate Mall Attack, the Euromaidan Revolution in Ukraine, and the mass demonstrations in Ferguson, Missouri, and compare these discovered instances with the reported timelines presented by traditional media outlets.

For event discovery, the primary method for evaluation is the amount of time between when a target event occurs and my implementations actually *discover* that event. Since the majority of my testing data was obtained using Twitter's Public Sample Stream, I can simulate Twitter's stream around the testing events I used previously and calculate this lag time for each event. This evaluation also serves as an ideal comparison since both the Storm and Spark Streaming implementations should produce similar timestamped events discovered in the social media streams. Therefore, I can construct a single test harness to evaluate accuracy and lag time of both implementations and draw comparisons on which is faster versus which discovers more events.

Similarly for event credibility, since I will be examining social media streams and event data after the fact, traditional media should have well-established timelines of the actual occurrences within these testing events. For example, with the Boston Marathon Bombing, we now have a clear understanding of the number of bombs, official response, and mistakes from news organizations. We can use this information to separate true events from rumors and then use these labeled events to evaluate our techniques' ability to discriminate between credible and in-credible events automatically.

5. PRELIMINARY RESULTS

At this early stage in my dissertation work, I have had some success in demonstrating the feasibility of a language-agnostic, domain-independent algorithm for discovering unanticipated and high-impact events from Twitter's unfiltered 1% public sample stream. To accomplish this task, I leveraged machine learning to model temporal patterns around bursts in the Twitter stream and build a classifier, LABurst, to identify tokens experiencing these bursts. I showed this LABurst technique performs competitively with existing burst detection techniques while simultaneously providing insight into and detection of unanticipated moments. To demonstrate LABurst's potential, I compared two baseline event-detection algorithms with our language-agnostic algorithm to detect key moments across three major sporting competitions (2013 World Series, 2014 Super Bowl, and 2014 World Cup). These two existing burst detection methods were a volume-centric burst detection technique (RawBurst), and a similar technique with a pre-determined set of sports-related keywords (TokenBurst). Results showed the technique outperformed one baseline and was competitive with the second baseline even though my technique operates without any domain knowledge. I then went further by transferring these sports-based models to the task of identifying earthquakes in Japan and showed my method detects large spikes in earthquake-related tokens within two minutes of the actual event.

5.1 Event Discovery Results

My first research question was whether LABurst could perform as well as existing methods in detecting events, with a focus on sporting competitions, specifically the final two games of the 2013 MLB World Series, the 2014 NFL Super Bowl, and the final two matches of the 2014 FIFA World Cup. To this end, I first constructed Receiver Operating Characteristic (ROC) curves and used the area under the curve (AUC) to assess performance for each sporting competition. To compare comprehensive performance, we look to Figure 1, which shows ROC curves for all three methods across all three testing events. From this figure, we see LABurst (AUC=0.71) outperforms RawBurst (AUC=0.65) and performs nearly identically to TokenBurst (AUC=0.72). Therefore, it seems the answer to **RQ1** is yes, LABurst can be competitive with existing methods.

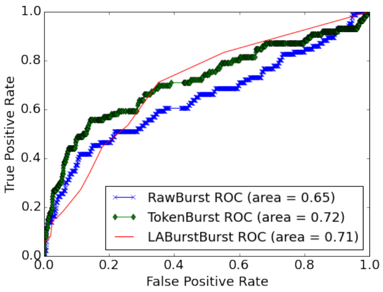


Figure 1: Composite ROC Curves

5.2 Earthquake Detection

My second research question was whether transferring LABurst's models, as trained on the previous sporting events, can compete with existing techniques in a different domain. Figures 2a and 2b show the detection curves for LABurst and the TokenBurst baseline for two earthquakes in Japan, one in 2013 and one in 2014; the red dots indicate the earthquake times as reported by the United States Geological Survey (USGS). The left vertical axis for each figure reports the frequency of the "earthquake" token (as in Sakaki et al. [16]), and the right axis shows the number of tokens classified as bursty by LABurst. From the TokenBurst curve, one can see the token "earthquake" sees a significant increase in usage when the earthquake occurs, and LABurst experiences a similar increase at the same moment for both events. The peak occurring about 50 minutes after the earthquake on 25 October 2013 potentially represents an aftershock event[1]. Given the minimal lag between LABurst and TokenBurst's detection, I showed LABurst is effective in cross-domain event discovery.

6. CONCLUSIONS AND FUTURE WORK

Preliminary results show that by leveraging temporal characteristics to identify bursty tokens and using frequency of these bursty tokens, one can detect significant events across a collection of disparate sporting competitions with a level of performance nearly equivalent to an existing, domain-specific baseline. This approach offers notable flexibility in identifying bursting tokens without normalization and across language boundaries. Finally, these advantages culminate in powerful tool for event *discovery* in that it can unanticipated instances of high interest that one did not expect, regardless of the source language, which makes the LABurst technique particularly useful for journalists and newswire sources who have a need to know about events on the ground, as they happen but cannot know a priori what the event may be about in all cases. Much work remains, however, in developing methods for evaluating truthfulness and credibility in this information and extending this work to a true real-time context.

7. REFERENCES

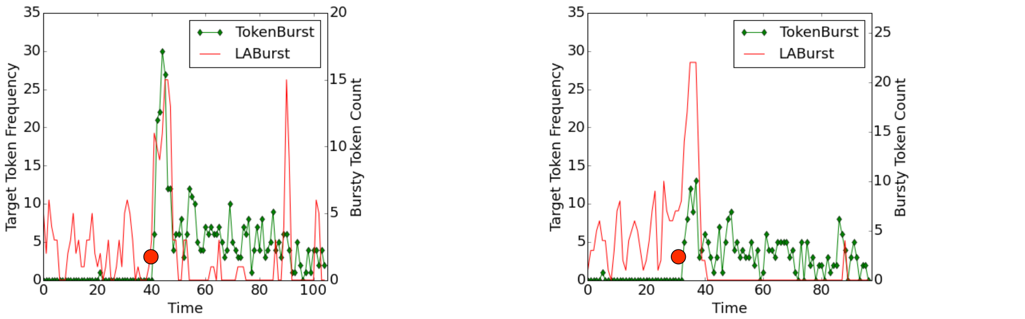
[1] C. A. Cassa, R. Chunara, K. Mandl, and J. S. Brownstein. Twitter as a Sentinel in Emergency Situations : Lessons from the Boston Marathon Explosions. *PLOS Currents Disasters*, pages 1–10, 2013.

[2] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 675–684, New York, NY, USA, 2011. ACM.

[3] E. F. Davis Iii, A. A. Alves, and D. A. Sklansky. Social Media and Police Leadership: Lessons From Boston. In *New Perspectives in Policing Bulletin.* Washington, DC: U.S. Department of Justice, National Institute of Justice, NCJ 244760., 2014.

[4] N. Diakopoulos, M. De Choudhury, and M. Naaman. Finding and assessing social media information sources in the context of journalism. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 2451–2460, New York, NY, USA, 2012. ACM.

[1] http://ds.iris.edu/spud/aftershock/9761021



(a) Honshu, Japan Earthquake - 25 October 2013

(b) Iwaki, Japan Earthquake - 11 July 2014

Figure 2: Japanese Earthquake Detection

[5] J. Golbeck. Combining provenance with trust in social networks for semantic web content filtering. In *Provenance and Annotation of Data*, pages 101–108. Springer, 2006.

[6] A. Gupta and P. Kumaraguru. Credibility ranking of tweets during high impact events. In *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media*, PSOSM '12, pages 2:2—-2:8, New York, NY, USA, 2012. ACM.

[7] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina. The Eigentrust Algorithm for Reputation Management in P2P Networks. In *Proceedings of the 12th International Conference on World Wide Web*, WWW '03, pages 640–651, New York, NY, USA, 2003. ACM.

[8] B. Kang, J. O'Donovan, and T. Höllerer. Modeling topic specific credibility on twitter. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, IUI '12, pages 179–188, New York, NY, USA, 2012. ACM.

[9] X. Liu, A. Datta, K. Rzadca, and E.-P. Lim. StereoTrust: A Group Based Personalized Trust Model. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 7–16, New York, NY, USA, 2009. ACM.

[10] M. Mendoza, B. Poblete, and C. Castillo. Twitter Under Crisis: Can We Trust What We RT? In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 71–79, New York, NY, USA, 2010. ACM.

[11] M. Osborne, S. Moran, R. McCreadie, A. Von Lunen, M. Sykora, E. Cano, N. Ireson, C. Macdonald, I. Ounis, Y. He, and Others. Real-Time Detection, Tracking, and Monitoring of Automatically Discovered Events in Social Media. *Association for Computational Linguistics*, 2014.

[12] S. Petrovic, M. Osborne, R. McCreadie, C. Macdonald, I. Ounis, and L. Shrimpton. Can Twitter replace Newswire for breaking news? In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, volume 2011, 2013.

[13] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei. Rumor Has It: Identifying Misinformation in Microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1589–1599, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[14] S. Ravikumar, R. Balakrishnan, and S. Kambhampati. Ranking tweets considering trust and relevance. In *Proceedings of the Ninth International Workshop on Information Integration on the Web*, IIWeb '12, pages 4:1—-4:4, New York, NY, USA, 2012. ACM.

[15] J. Rogstadius, M. Vukovic, C. A. Teixeira, V. Kostakos, E. Karapanos, and J. A. Laredo. CrisisTracker: Crowdsourced social media curation for disaster awareness. *IBM Journal of Research and Development*, 57(5):4:1–4:13, Sept. 2013.

[16] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM.

[17] S. Siddiqui. Boston Bombings Reveal Media Full Of Mistakes, False Reports, Apr. 2013.

[18] W. Xie, F. Zhu, J. Jiang, E.-p. Lim, and K. Wang. TopicSketch: Real-time Bursty Topic Detection from Twitter. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 837–846. IEEE, 2013.