Real-time Geosocial Media Event Detection and Prediction

Assignment for Research Methods in Computer Science course at Ryerson University

Richard Wen

Department of Civil Engineering, Ryerson University, Toronto, ON

December 3, 2017

Contents

1 Introduction

The wide availability of mobile devices have enabled millions of people to share online content, such as text, images, sound, and videos, from any location with wireless Internet connection. Social media platforms, such as Facebook (Facebook, 2017) and Twitter (Twitter Inc, 2017), are commonly used to share large amounts of online content in near real-time. This online content produces valuable sources of real-time locational data, known as geosocial media data, that may provide information on current real-world events such as traffic jams, natural disasters, disease spread, and terrorist attacks. Geosocial media data can be used to detect and predict real-world events given particular locations and times. However, human errors, inconsistencies, noise, high volumes, and constant changes make it difficult to extract useful information from geosocial media data. These issues cause a divide in the methods and approaches for geosocial media event detection and prediction, where standards, comparisons, and integration between different data sources and use cases are rare. This proposal documents a plan to develop a generalized framework and open source software for detecting and predicting real-world events using geosocial media data.

The objective of this proposal is to develop a framework and accompanying software to detect and predict real-world events in real-time with geosocial media data. A literature review was done to provide background knowledge on current research on event detection and prediction methods and applications. An approach, built on the knowledge from the literature review, was developed to satisfy the objectives. Recent progress was detailed to provide preliminary results and relevant past work related to the objectives. A discussion of the impacts was provided to address the importance and effect of the proposed research work.

The remaining sections are organized as follows:

- **Section 2** details the objectives of the proposed research
- **Section 3** provides a literature review of current research
- **Section 4** details the proposed approach to satisfy the objectives and recent progress
- **Section 5** discusses the impact of the proposed research and provides concluding summaries and remarks

2 Objectives

This section provides details objectives of this proposal. The main objective is to develop the following for detecting and predicting real-world events using geosocial media data:

1. Framework that can be applied to a wide variety of applications and data
2. Open Source Software based on (1)

2.1 Framework

The framework objective requires that the following components be identified and developed:

- **Data Sources**: Popular geosocial media platforms and data sources
- **Data Structures**: Geosocial media data structures
- **Event Detection Methods**: Common event detection methods and patterns
- **Event Prediction Methods**: Common event prediction methods and patterns
- **Output**: Resulting human-readable output information
- **Use Cases**: Common applications of geosocial media event detection and prediction

2.2 Software

The software objective requires that the following open source components be identified and developed:

- **Databases**: Popular databases used for geosocial media data
- **Event Detection and Prediction Software**: Libraries or packages for event detection and prediction algorithms and models
- **Information Software**: Libraries or packages for displaying and extracting information from model outputs
- **Online Platform**: Online websites to host and distribute software
- **Testing Software**: Libraries or packages to conduct standard unit tests
- **Documentation Software**: Libraries or packages to document software for a wide audience

3 Literature Review

This section provides a literature review to provide background knowledge on current research related to the topic of *"real-time geosocial media event detection and prediction"*. Papers were selected from the Association for Computing Machinery (ACM) and Institute of Electrical and Electronics Engineers (IEEE) Xplore digital libraries. Figure 1 shows the distribution of selected papers for review by year. Appendix A provides details of the methods used for paper selection and review.

3.1 Event Detection and Prediction

Event detection and prediction methods using geosocial media data are focused on statistical, machine learning, clustering, and network based approaches. Alvanaki et al. (2012) use sliding time windows to compute statistics about social media tags in order to detect unusual correlation shifts that are probable emergent topic events. Khabiri et al. (2012) used closeness measures, term selection, and dynamic sliding windows to annotate social media messages with informative and more

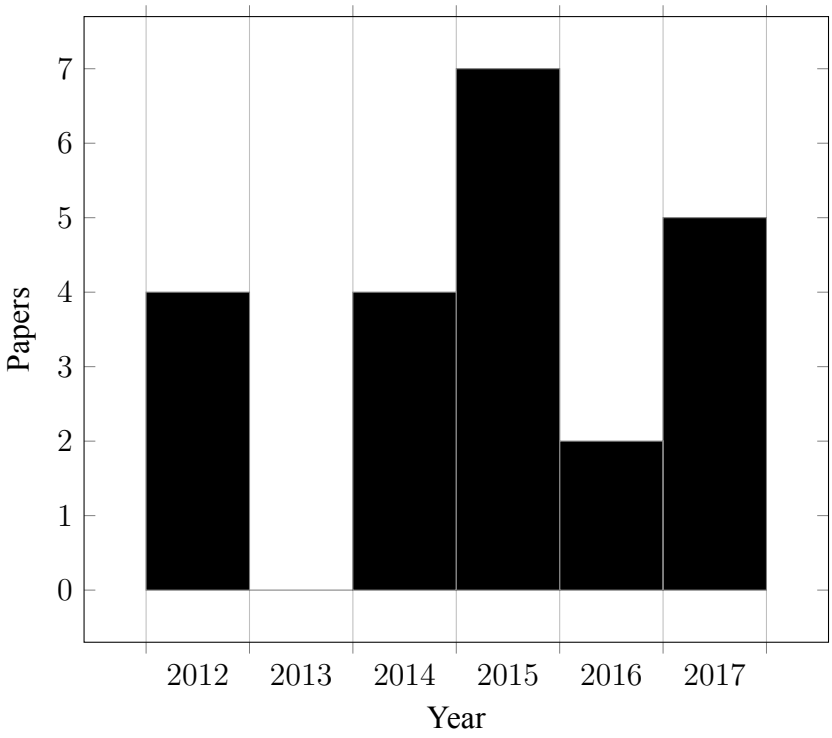

Figure 1: **ACM and IEEE Published Papers Found from 2012 to December 2, 2017.** A total of 22 papers were reviewed. Black bars represent the selected papers filtered from the potential papers using the literature review methods described in Appendix A.

abstract terms, however the techniques used were infeasible for real-time annotation when smoothing was implemented in the sliding windows. Sofean and Smith (2012) used machine learning methods to classify medical-related tweets on Twitter based on world topics relative to diseases and symptoms. Middleton et al. (2014) suggested the use re-tweets from Twitter for credibility measures, and context filters to narrow natural disaster events to specific classes such as positive, negative, or urgent reports with natural language text. Hazra et al. (2015) built a web application with event sockets that extracted news-worthy Twitter events, which can be personalized to particular interests, but machine learning may be needed improve the user recommendation system, and historical events were not considered. Enoki et al. (2015) used in-memory stores for real-time topic querying given social diffusion of popular Twitter messages. Abrol et al. (2015) used cloud computing to profile and geo-locate millions of users for social behavioural patterns, changes, and events in real-time, but required improvements to handle simultaneously emerging new classifications. Buntain (2015) used social media features such as location and metadata to evaluate the credibility of detected events. **?** used a technique which identified rapid increases in social media posts to detect events and improve the performance of other real-time tracking systems. Lee et al. (2017) used machine learning methods to extract important word features from social media data to predict flu activity 2 to 3 weeks ahead in real-time. Li et al. (2017) used semantic and similarity clustering to identify old and new geosocial media events with the same meaning in real-time. Tsikerdekis (2017) compared several optimization techniques to reduce the volume or velocity of social media data for identity deception detection, and concluded that data velocity is a major fac-

tor in considering detection methods. A majority of the methods presented in the literature focused on geosocial media event detection methods, but few mentioned or considered event prediction in real-time.

3.2 Visualization

After detecting or predicting geosocial media events, the model outputs must be visualized into human-understandable information to appropriately act and respond to each event. Zubiaga et al. (2012) used summarization techniques to reduce the large volume of Twitter data to provide a higher level textual abstraction of social media events and topics. Calderon et al. (2014) designed streaming graphs to visualize real-time Twitter sentiments for emergency management, where a study on 21 randomy selected participants concluded that visualization of real-time social media data required interactive interfaces, geo-location context, and human cognition and reasoning theory. Middleton et al. (2014) mapped real-time social media data using textual geo-parsing, spatial clustering, and a combination of various data sources, such as Volunteered Geographic Information (VGI), online mapping services, and gazetteers, to visually display areas of potential events. Xia et al. (2014) developed a web-based visualization, consisting of a web map with complementary photos and graphs, for multiple social media sources such as Instagram, Twitter, and Foursquare to provide real-time events and news trends for a city. Tsirakis et al. (2015) created a web-based dashboard application that displays the current trending events, influential users, and popular topics in real-time, but had considerations with the scalability of algorithms, multiple data sources, and language integrations. **?** used background knowledge from Wikipedia to summarize social media events relative to photos and videos in real-time. Kumar and Sinha (2016) used nodes to represent social media users for creating visual networks of Twitter users to analyze and detect the strength of social relationships among users. **?** used natural language processing techniques to detect traffic incidents for a real-time traffic alerts and warnings system. The majority of the visualization approaches proposed in the literature involved interactive, web-based, interfaces that transformed lower level details of model results into higher level abstract visuals to produce human-understandable information.

3.3 Applications

Event detection and prediction methods are widely used for real-world applications related to the categories of disaster, disease, travel, and the environment. Figure 2 shows the number of papers in each category from the literature. Sofean and Smith (2012) created a real-time surveillance architecture to detect disease-related social media postings on Twitter. Riga and Karatzas (2014) used machine learning models to cluster Twitter tweets for urban air quality monitoring and soft sensing. Middleton et al. (2014) used geosocial media event detection to obtain real-time crisis maps and reports for hurricanes and earthquakes in disaster response. Semwal et al. (2015) used machine learning techniques to detect and predict daily co-occurring traffic event issues with social media data. Bodnar et al. (2017) used social media data together with electrical consumption data to discover relevant energy-related topics and to gain real-time insight on energy consumption of users in urban areas. Lee et al. (2017) used geosocial media data with historical diseases datasets to predict future influenza events in real-time. The majority of the papers in the literature conducted experiments and tests in disaster applications, followed by disease, travel, and environment in order.

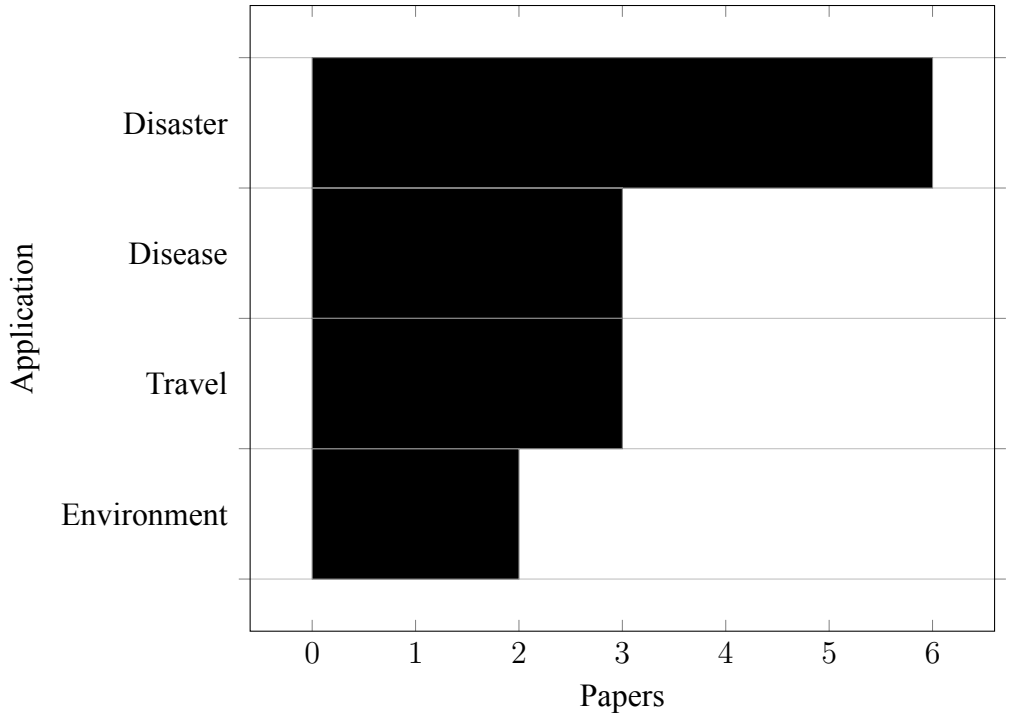

Figure 2: **Applications in ACM and IEEE Published Papers from 2012 to December 2, 2017.** Disease refers to human health-related disease spread events. Disaster refers to natural disasters and human-related emergencies. Environment refers to environmental-related applications such as air pollution and urban energy utility. Travel refers to vehicular traffic and transportation related applications. Black bars represent the number of papers in each application category, where papers were selected using the literature review methods in Appendix A.

4 Approach

This section provides details on the methods for the proposed approach and milestones to achieve the research objectives. Recent progress toward the objectives is also mentioned in this section.

4.1 Methods

x

4.2 Recent Progress

Recent progress involved the partial identification of several framework components and development of a small software package. The identified framework and software components are provided in Table 1 and 2 respectively. A small software package was developed for Node.js (Node.js Foundation, 2017) named *"twitter2pg"* (Wen, 2017) to conveniently extract real-time Twitter data into a relational PostgreSQL database (The PostgreSQL Global Development Group, 2017). The package has been downloaded 259 times as of December 2, 2017 after approximately a month of

release, and consists of documentation, unit tests, and automatic Linux builds for continuous tests every month.

Table 1: **Identified Framework Objective Components.**

Component	**Identified**
Data Sources	Twitter Streaming API, Programmable Web
Data Structures	Unstructured (JSON), Location Points, Time Stamp
Event Detection Methods	Frequency, Sliding Window, Normalization, Clustering, Sampling, Graphs, Machine Learning
Output	Textual Summary, Webmap, Wordcloud
Use Cases	Influenza, Earthquake, Psychosocial, Energy, Traffic, Air Quality

Table 2: **Identified Software Objective Components.**

Component	**Identified**
Databases	PostgreSQL, MongoDB, MySQL, Hbase, Cassandra, Accumulo, GeoMesa
Event Detection and Prediction Software	Massive Online Analysis (MOA), scikit-learn, Apache Spark, Apache Kafka
Information Software	Leaflet, Carto, D3.js
Online Platform	Github, PyPi, npm
Testing Software	travis Continuous Integration (CI), Docker
Documentation Software	HTML, Markdown

5 Conclusion and Impact

This proposal presented a potential framework and accompanying software for geosocial media event detection and prediction based on current research. The framework was proposed to provide a more consistent approach to working with geosocial media data, and to more easily allow non-experts to have a standard solution for a wide variety of applications such as traffic management, disease control, and disaster response. The impacts of a framework and software for geosocial media event detection and prediction are related to approach consistency improvements, standardized solutions, and promotion of transparency in social media research.

Appendices

Appendix A Literature Review Methods

The paper selection process involved identifying reputable digital libraries using the Journal Impact Factor (JIF) measure (Garfield, 2006b), followed by using automatic search queries to produce an initial list of potential papers. The potential papers were then further filtered by manual selection criteria to produce a list of selected papers for reviewing. The literature review process is seen in Figure 3.

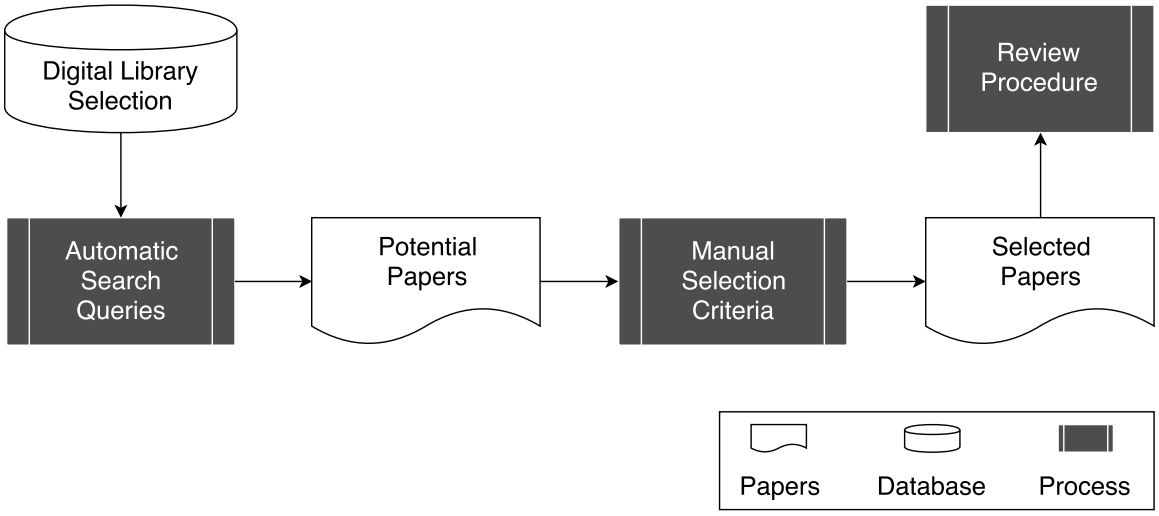

Figure 3: **Literature Review Methods.**

A.1 Digital Library Selection

The papers for the literature review were found with the search engines available in the Association for Computing Machinery (ACM) (Association for Computing Machinery, 2017) and Institute of Electrical and Electronics Engineers (IEEE) Xplore (Institute of Electrical and Electronics Engineers, 2017) digital libraries. A search for the top journals in computer science by journal impact factor (Garfield, 2006b) was done using the InCites journal citation reports web tool (Clarivate Analytics, 2017a). A majority of ACM and IEEE journals were found to be in the first quartile of journal impact factor values for the computer science category. A visualization of the top 25 journals in computer science by journal impact factor in 2016 is shown in Figure 4.

The search for the top 25 computer science journals was based on the Journal Impact Factor (JIF) (Garfield, 2006b) measure, and was done using the InCites Journal Citation Reports (JCR) web tool (Clarivate Analytics, 2017a). The search used the following options available on InCites:

- **Categories**:
  - COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE
  - COMPUTER SCIENCE, CYBERNETICS

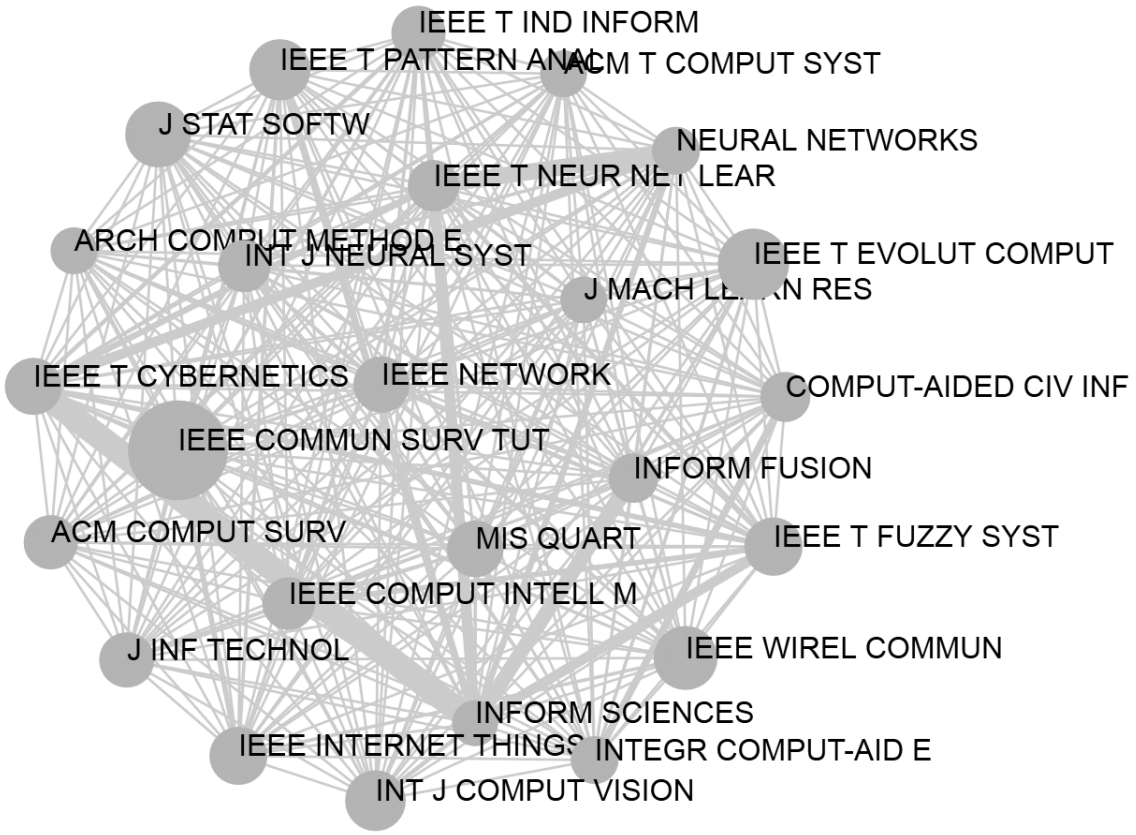

Figure 4: **Top 25 Computer Science Journals by Journal Impact Factor from InCites Journal Citation Report in 2016.** Gray circles represent the Journal Impact Factor, where higher Journal Impact Factor values are represented by larger sizes. Connected lines represent the citation relationships between each journal, where thicker lines mean stronger relationships.

  - COMPUTER SCIENCE, HARDWARE & ARCHITECTURE
  - COMPUTER SCIENCE, INFORMATION SYSTEMS
  - COMPUTER SCIENCE, INTERDISCIPLINARY APPLICATIONS
  - COMPUTER SCIENCE, SOFTWARE ENGINEERING
  - COMPUTER SCIENCE, THEORY & METHODS

- **JCR Year**: 2016
- **Edition**: Science Citation Index Expanded (SCIE) (Garfield, 2006a) and Social Sciences Citation Index (SSCI) (Klein et al., 2004)
- **Category Schema**: Web of Science (Clarivate Analytics, 2017b)
- **JIF Quartile**: Quarter 1 (Q1)

A.2 Automatic Search Queries

Potential papers were found using search engine queries in the ACM (Association for Computing Machinery, 2017) and IEEE Xplore (Institute of Electrical and Electronics Engineers, 2017) digital libraries identified in Appendix A.1. Search queries were modified from the defaults and sorted by relevance. Each search query was defined to filter for potential papers with the following requirements:

(a) **Publication**: Published in ACM or IEEE

(b) **Year**: Published from 2012 to December 2, 2017

(c) **Keywords**: Contains the keywords *"real time"* and *"social media"* in the paper title, and *"prediction"*, *"predict"*, *"detection"*, or *"detect"* anywhere in the text

The query syntax in the ACM digital library was accessed through the advanced search page by clicking *"show query syntax"*. The *"+"* symbol includes each keyword in the title. *"gte"* and *"lte"* represent *"greater than or equal to"* and *"less than or equal to"* respectively. The publication date query syntax must be manually generated using the web interface. The full advanced query syntax used for the ACM digital library to return potential papers is shown below:

```
"query": { acmdlTitle:(+real +time +social +media) AND (prediction predict detection detect) }

"filter": {"publicationYear":{ "gte":2012, "lte":2017 }},
{owners.owner=HOSTED}
```

The command search in the IEEE Xplore digital library was accessed through the advanced search page by clicking *"command search"*. Refinements were manually applied using the web interface to filter command search results for the years 2012 to 2017 and to search in *"Full Text & Metadata"*. The command search used for the IEEE Xplore digital library to return potential papers is shown below:

```
"Document Title":"real time" AND "Document Title":"social media" AND ("prediction" OR "predict"
    OR "detection" OR "detect")
```

A.3 Manual Selection Criteria

The potential papers from Appendix A.2 were further filtered with the abstracts and paper length. The abstracts were inspected for relevancy to the topic: *"real-time geosocial media event detection and prediction"*. This included mentions of methods that deal with detecting or predicting real-world events in real-time using geosocial media data. After inspections of the abstract, each paper was further evaluated for practicality by searching for mentions of event prediction or detection applications, benchmarks, and experiments in the results sections. The manual selection criteria sought to find papers with the following characteristics:

(a) **Detailed**: Paper contained sufficient details and explanations to obtain a general understanding of the methods and results

(b) **Relevant**: Paper had mentions of real-time geosocial media event detection or prediction

(c) **Practical**: Paper had conducted experiments, benchmarks, or applications using described event detection or prediction methods

A.4 Review Procedure

A literature review of the papers selected using the methods in Appendix A.3 was done with the following procedure:

1. **Identify** methods used for real-time geosocial media event detection or prediction
2. **Summarize** methods in (1)
3. **Summarize** applications and results for the methods in (1)
4. **Discuss** limitations, possible improvements, and future directions relative to the summaries from (2) and (3)

References

Abrol, S., Rajasekar, G., Khan, L., Khadilkar, V., Nagarajan, S., McDaniel, N., Ganesh, G., and Thuraisingham, B. (2015). Real-time stream data analytics for multi-purpose social media applications. In 2015 IEEE International Conference on Information Reuse and Integration, pages 25–30.

Alvanaki, F., Michel, S., Ramamritham, K., and Weikum, G. (2012). See what's enblogue: Real-time emergent topic identification in social media. In Proceedings of the 15th International Conference on Extending Database Technology, EDBT '12, pages 336–347, New York, NY, USA. ACM.

Association for Computing Machinery (2017). Acm digital library. Retrieved December 2, 2017 from `https://dl.acm.org/`.

Bodnar, T., Dering, M. L., Tucker, C., and Hopkinson, K. M. (2017). Using large-scale social media networks as a scalable sensing system for modeling real-time energy utilization patterns. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 47(10):2627–2640.

Buntain, C. (2015). Discovering credible events in near real time from social media streams. In Proceedings of the 24th International Conference on World Wide Web, WWW '15 Companion, pages 481–485, New York, NY, USA. ACM.

Calderon, N. A., Arias-Hernandez, R., and Fisher, B. (2014). Studying animation for real-time visual analytics: A design study of social media analytics in emergency management. In 2014 47th Hawaii International Conference on System Sciences, pages 1364–1373.

Clarivate Analytics (2017a). Incites journal citation reports. Retrieved December 2, 2017 from `https://jcr.incites.thomsonreuters.com/`.

Clarivate Analytics (2017b). Web of science. Retrieved December 2, 2017 from `https://webofknowledge.com/`.

Enoki, M., Yoshida, I., and Oguchi, M. (2015). Performance of system for analyzing diffusion of social media messages in real time. In 2015 IEEE International Conference on Systems, Man, and Cybernetics, pages 801–806.

Facebook (2017). Facebook. Retrieved December 2, 2017 from `https://www.facebook.com`.

Garfield, E. (2006a). Citation indexes for science. a new dimension in documentation through association of ideas. International journal of epidemiology, 35(5):1123–1127.

Garfield, E. (2006b). The history and meaning of the journal impact factor. Jama, 295(1):90–93.

Hazra, T. K., Ghosh, A., Sengupta, A., and Mukherjee, N. (2015). Mitigating the adversities of social media through real time tweet extraction system. In 2015 International Conference and Workshop on Computing and Communication (IEMCON), pages 1–8.

Institute of Electrical and Electronics Engineers (2017). Ieee xplore digital library. Retrieved December 2, 2017 from `http://ieeexplore.ieee.org/Xplore/home.jsp`.

Khabiri, E., Caverlee, J., and Kamath, K. Y. (2012). Predicting semantic annotations on the real-time web. In Proceedings of the 23rd ACM Conference on Hypertext and Social Media, HT '12, pages 219–228, New York, NY, USA. ACM.

Klein, D. B., Chiang, E., et al. (2004). The social science citation index: A black box with an ideological bias? Econ Journal Watch, 1(1):134–165.

Kumar, P. and Sinha, A. (2016). Real-time analysis and visualization of online social media dynamics. In 2016 2nd International Conference on Next Generation Computing Technologies (NGCT), pages 362–367.

Lee, K., Agrawal, A., and Choudhary, A. (2017). Forecasting influenza levels using real-time social media streams. In 2017 IEEE International Conference on Healthcare Informatics (ICHI), pages 409–414.

Li, Q., Nourbakhsh, A., Shah, S., and Liu, X. (2017). Real-time novel event detection from social media. In 2017 IEEE 33rd International Conference on Data Engineering (ICDE), pages 1129–1139.

Middleton, S. E., Middleton, L., and Modafferi, S. (2014). Real-time crisis mapping of natural disasters using social media. IEEE Intelligent Systems, 29(2):9–17.

Node.js Foundation (2017). Node.js. Retrieved December 2, 2017 from `https://nodejs.org`.

Riga, M. and Karatzas, K. (2014). Investigating the relationship between social media content and real-time observations for urban air quality and public health. In Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14), WIMS '14, pages 59:1–59:7, New York, NY, USA. ACM.

Semwal, D., Patil, S., Galhotra, S., Arora, A., and Unny, N. (2015). Star: Real-time spatio-temporal analysis and prediction of traffic insights using social media. In Proceedings of the 2Nd IKDD Conference on Data Sciences, CODS-IKDD '15, pages 7:1–7:4, New York, NY, USA. ACM.

Sofean, M. and Smith, M. (2012). A real-time architecture for detection of diseases using social networks: Design, implementation and evaluation. In Proceedings of the 23rd ACM Conference on Hypertext and Social Media, HT '12, pages 309–310, New York, NY, USA. ACM.

The PostgreSQL Global Development Group (2017). Postgresql. Retrieved December 2, 2017 from `https://www.postgresql.org/`.

Tsikerdekis, M. (2017). Real-time identity deception detection techniques for social media: Optimizations and challenges. IEEE Internet Computing, PP(99):1–15.

Tsirakis, N., Poulopoulos, V., Tsantilas, P., and Varlamis, I. (2015). A platform for real-time opinion mining from social media and news streams. In 2015 IEEE Trustcom/BigDataSE/ISPA, volume 2, pages 223–228.

Twitter Inc (2017). Twitter. Retrieved December 2, 2017 from `https://twitter.com`.

Wen, R. (2017). twitter2pg. Retrieved December 2, 2017 from https://www.npmjs.com/package/twitter2pg.

Xia, C., Schwartz, R., Xie, K., Krebs, A., Langdon, A., Ting, J., and Naaman, M. (2014). Citybeat: Real-time social media visualization of hyper-local city data. In Proceedings of the 23rd International Conference on World Wide Web, WWW '14 Companion, pages 167–170, New York, NY, USA. ACM.

Zubiaga, A., Spina, D., Amigó, E., and Gonzalo, J. (2012). Towards real-time summarization of scheduled events from twitter streams. In Proceedings of the 23rd ACM Conference on Hypertext and Social Media, HT '12, pages 319–320, New York, NY, USA. ACM.