LANDSLIDE INFORMATION SERVICE BY INTEGRATING MULTIPLE INFORMATION SOURCES

A Thesis Presented to
The Academic Faculty

by

Qixuan Hou

In Partial Fulfillment of the Requirements

for the Research Option in the College of Computing

Georgia Institute of Technology May 2018

LANDSLIDE INFORMATION SERVICE BY INTEGRATING MULTIPLE INFORMATION SOURCES

Approved by

Professor Dr. Calton Pu, Advisor

School of Computer Science

at the College of Computing

Georgia Institute of Technology

Professor Dr. Ling Liu

School of Computer Science

at the College of Computing

Georgia Institute of Technology

TABLE OF CONTENTS

TABLE OF CONTENTS ..... **- 3 -**

LIST OF TABLES ..... **- 4 -**

LIST OF FIGURES ..... **- 5 -**

INTRODUCTION ..... **- 6 -**

LITMUS: A LANDSLIDE DETECTION SERVICE ..... **- 8 -**

INTEGRATING NEWS TO LITMUS ..... **- 12 -**

INTEGRATING MULTILINGUAL INFORMATION TO LITMUS ..... **- 16 -**

CONCLUSION & FUTURE WORK ..... **- 20 -**

REFERENCE ..... **- 21 -**

LIST OF TABLES

Table 3.1. Summary of news data

Table 3.2. Summary of Twitter data

Table 3.3. Matched events in news

Table 3.4. Events summary

Table 3.5. Events summary

Table 4.1. Summary collected Chinese tweets

Table 4.2. Summary collected English tweets

Table 4.3. Summary of events reported by Chinese tweets

Table 4.4. Summary of events reported by English tweets

LIST OF FIGURES

Figure.2.1 A Japanese tweet related to a landslide event

Figure.2.2 Tweets irrelevant to natural disaster with “landslide” keyword

Figure.2.3 English infrastructure pipeline for LITMUS

Figure.2.4. LITMUS live demonstration of landslide events

Figure 3.1. LITMUS pipeline

Figure 3.2. Sample tweet about an event not validated by news

Figure 4.1. LITMUS multilingual pipeline

Figure 4.2. LITMUS pipeline

CHAPTER I

INTRODUCTION

Detection of natural disasters mostly depends on physical sensors, but few physical sensors are available for the detection of multi-hazards, such as landslides (Musaev, Wang & Pu, 2015). As a popular platform for real time information with users all around the world, social media augments the traditional way of natural disaster detection by providing supplemental timely data. LITMUS (Landslide Detection by Integrating Multiple Sources) is the application we built to provide landslide detection service that combines data from both physical and social information services by filtering and then joining the information flow from those services based on their spatiotemporal features (Musaev, Wang & Pu, 2015). In this thesis, we have explored additional information sources (news source and multilingual information from Twitter), which, if integrated to LIMTUS, can potentially increase credibility and coverage of the application.

Credibility is a perceived quality and credibility perceptions which result from evaluating trustworthiness and expertise. "Trustworthiness is defined by the terms well-intentioned, truthful, unbiased, and so on. Expertise is defined by terms such as knowledgeable, experienced, competent, and so on" (Fogg & Tseng, 1999). Without losing credibility, we would like also to improve the coverage of LITMUS. Improving coverage means increasing the number of events reported by LITMUS and the events cover diverse regions throughout the world.

News source is a traditional way of the public to collect timely information. There are many researchers from different areas studying the trustworthiness of the news (Salwen, Garrison, & Driscoll, 2005) (El-Nawawy, 2006) (Kohring & Mattes, 2007), and most of them focus questioning the credibility of politics news. News publishers generally focus on events from certain region. For example, Xinhua, the biggest news agency in China, dedicates to news within China and WSB-TV focuses on the events around Atlanta. In order to achieve better coverage, diverse news sources should be included.

With the trend of globalization, social media platforms, including Twitter and Facebook, are also popular in non-English-speaking countries, such as Japan, Indonesia, and European countries. A tremendous amount of data on social media is written in different languages. Although English

has been regarded as a standard international language, it is common for people from each country to exchange information in their own language(s). Therefore, discussions on the events in non-English spoken countries are more likely to have postings in the local languages in addition to English (Hou, Musaev, Yang, & Pu, 2017). We would like to collect tweets in other languages and to analyze if we can improve LITMUS coverage and credibility.

The rest of the thesis is organized as follows. Chapter II outlines the main components of LITMUS. Chapter III discusses integrating news into LITMUS and evaluating the results. Chapter IV shows the evaluation of integrating multilingual tweets into LITMUS. Chapter V concludes the thesis.

LITMUS: A LANDSLIDE DETECTION SERVICE

Landslides often escape physical sensor detection, and their discovery depends critically on human reportings, in modern days through social media. LITMUS integrates information from physical sensors and social media to provide information on landslides worldwide, including social networks, Twitter, Facebook, and YouTube. Like many applications that process and filter social media information, LITMUS supported only English initially due the availability of natural language processing (NLP) tools. Since landslides occur in many countries around the world, we soon realized that LITMUS is missing significant information on landslides in non-English speaking countries, such as Japan, China, and Brazil, where landslides occur frequently. For instance, Twitter has approximately 328 million active users monthly and 79% of the users are outside the U.S. (Twitter, 2017). Figure 2.1 shows an example of a Japanese tweet on a landslide in Japan.

Figure.2.1 A Japanese tweet related to a landslide event

Figure.2.2 Tweets irrelevant to natural disaster with “landslide” keyword

The basic architecture of LITMUS (designed for English social media information filtering) has been described in our previous paper (Musaev, Wang & Pu, 2015). In this section, we briefly outline the five stages of LITMUS filtering pipeline in this section to make the paper self-contained.

Readers already familiar with LITMUS may safely skip to the next section. The experiments focus on the LITMUS processing of Twitter data in five stages, as shown in Figure 2.3.

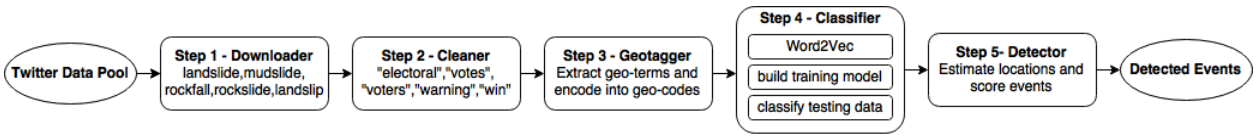

Figure.2.3 English infrastructure pipeline for LITMUS

Stage 1 – Downloader. "Landslide" and "mudslide" are identified as English search terms to retrieve data about landslides from Twitter. LITMUS implements Twitter Streaming APIs to access Twitter's global stream of Tweet data with identified search terms.

Stage 2 – Cleaner. Social media platforms contain noises. For example, instead of natural disasters, "Landslide" is a song name performed by the music group, Fleetwood Mac. See Figure 2.2 for a tweet with "landslide" keyword as a song name. "Landslide" is also used to describe any election in which the victor wins by an overwhelming margin, as the tweet shown in Figure 2.2 demonstrates. LITMUS filters out these two situations using a set of stop words, including "Fleetwood Mac", "song", "election", "won", and "votes".

Stage 3 – Geotagger. Without annotated locations, detected landslide events are not useful. However, less than 1% of Twitter data contain geo-coordinates even if Twitter provides service to include user locations (Jurgens, 2013). In order to retrieve geo-coordinates from Twitter data, LITMUS, firstly, extracts geographical terms (geo-terms) by Stanford NER, a Java implementation of a Named Entity Recognizer (Finkel, Grenager, & Manning, 2005), and then geocodes with Google Maps APIs to convert geo-terms into geographic coordinates (geo-code) (Google, 2017). The surface of the Earth is regarded as a grid of cells, and geo-codes are eventually grouped into cells.

Stage 4 – Classifier. Classifier, a complement to cleaner stage, is another filtering stage in LITMUS to remove noises in social media data. Classifier is based on Word2Vec and Support Vector Machine (SVM) (Google, 2017). A part of tweet data is labeled manually as relevant or irrelevant to landslide disasters for training purposes. We firstly run word-segmentation on tweets, and then convert the words to vectors using the built Word2Vec model. SVM algorithm is

implemented by Weka, a Java open source library, to build training model based on the manually labeled tweets (Hall et al., 2009). With the built training model, Classifier automatically labels each item as either relevant or irrelevant to landslide disasters.

Stage 5 – Detector. Detector groups these tweets into cells to determine where landslides may have happened. After previous stages, LITMUS has tweets which are labeled as either relevant or irrelevant by Classifier and have cell ID assigned by Geotagger. Detector collects both relevant and irrelevant tweets from each month and groups them based on their cell IDs. Using a ranking strategy, LITMUS computes a landslide probability score to determine whether there is a landslide event in this cell location in that month. Since Detector is language independent, the same Detector is used for different languages.

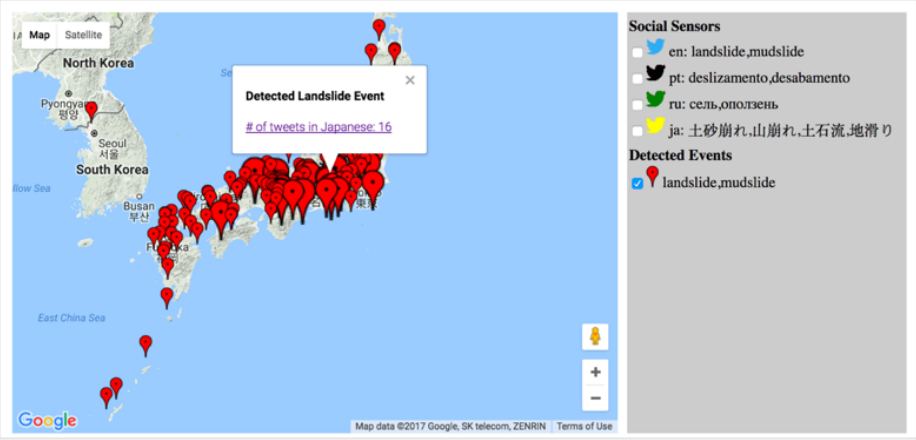

Figure.2.4. LITMUS live demonstration of landslide events

LITMUS integrates information from physical sensors and social media to provide worldwide landslide information. Figure 2.4 is a live demonstration of LITMUS, which can be accessed by the link in (GRAIT-DM, 2017). Disaster management systems always tend to adopt multiple information sources to achieve comprehensive event coverage. However, the more information sources they include, the more noises they need to face. LITMUS presents a generic method of compositing multiple information sources and filtering out unreliable and useless information, and landslide is chosen as a demonstration of the method (Musaev et al., 2014) (Musaev et al., 2014) (Musaev et al., 2015) (Musaev et al., 2015) (Lai et al., 2015) (Musaev, Wang, & Pu, 2015). The method can be easily implemented to support other information system. Tien et al. built a system to detect

damage and failure events of critical public infrastructure based on LITMUS prototype (Tien et al., 2016). In order to further improve LITMUS event credibility, coverage and timeliness, multilingual information and news are considered as potential sources. The thesis evaluated the improvement by integrating potential sources.

CHAPTER III

INTEGRATING NEWS TO LITMUS

Introduction

Newspaper is a traditional way for the public to gain information about the events around certain areas. With the development of online societies, people tend to browse news published online via phones and laptops. Credibility of online news and news from social media is studied by many researchers from different areas (Salwen, Garrison, & Driscoll, 2005) (El-Nawawy, 2006) (Kohring & Mattes, 2007) but most of them question the trustworthiness of news about politics. Additionally, since news articles are published by professional journalists, we regard events from news source as the authorities and we would like to study the accuracy of LITMUS system by comparing the events detected by Twitter and news.

Evaluation

News source is expensive to access. There are only limited news sources which provide developer friendly APIs to download historical news. We decided to use Google news API to access historical news because of its good coverage (Google, 2018). Google declared to crawl and index news articles from over 30,000 news sources and blogs and from 54 different countries.

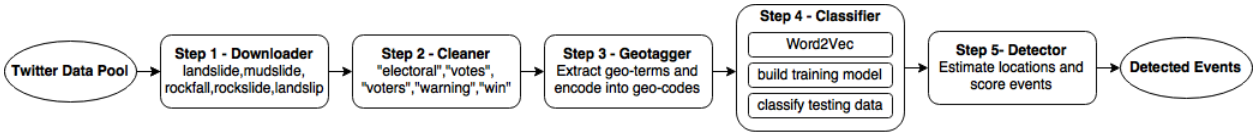

Figure 3.1. LITMUS pipeline

We collected news for three months from January 2018 to March 2018. The same pipeline was applied to process news data as shown in Figure 3.1. The number of news downloaded, geotagged, classified and the number of events detected are shown in Table 3.1.

News	201801	201802	201803	total
Downloader	2057	3634	3202	8893
Cleaner	1852	3383	3137	8372
Geotagger	1828	3334	3083	8245
Classifier	715	855	904	2474
%useful news	0.3475936	0.2352779	0.2823235	0.2781963
Events	245	241	295	781

Table 3.1. Summary of news data

We collected tweets in the same time period and processed tweets as the pipeline shown in Figure 3.1 as well. The number of tweets downloaded, geotagged, classified and the number of events detected are shown in Table 3.2.

tweets	201801	201802	201803	total
Downloader	6584	13984	10606	31174
Cleaner	6158	13196	8386	27740
Geotagger	5832	12942	8096	26870
Classifier	1017	3342	853	5212
%useful tw	0.1544654	0.2389874	0.0804262	0.1671906
Events	135	203	111	449

Table 3.2. Summary of Twitter data

With the additional news source, we would like to evaluate if the events can be validated (credibility) and if more events can be added to the system (coverage) by different sources.

Credibility. News is regarded more trustworthy than tweets because news is published by authorities and is designed to report events to the public. We regard news as the authority and validate if LITMUS Twitter-detected events can be validated by collected news. As shown in Table 3.3, there are about 35% Twitter-detected events which can be validated by collected news. After reviewing the Twitter-detected events which cannot be validated by news, we find some of the events are true events and these events are reporting local landslides or flooding. One example is an event reported on February 26, 2018 by QLDTrafficNCSC, shown in Figure 3.2. The event mentioned that Schmidt road in Eagleby was closed due to flooding. This type of local event is not often reported by major news source, but Google News API tends to only include major news source from each country.

events	201801	201802	201803	total
# Events detected by tweets	135	203	111	449
# Events matched in news	49	56	50	155

Table 3.3. Matched events in news

Figure 3.2. Sample tweet about an event not validated by news

We should not conclude the 65% Twitter-detected events which cannot be validated by news are not true events. In order to ensure the quality of the events reported by LITMUS, we previously decided to consider events with more than three tweets as true events and delete the events with less than three tweets from the final reports. The events with less than three tweets are regarded as unconfident events. The threshold is selected arbitrarily in order to ensure the quality of the LITMUS, but it also decreases the coverage. As shown in Table 3.4, there are about 68% events deleted from final reports.

events	201801	201802	201803	total
# Events detected by tweets	135	203	111	449
# Events > 3 tweets	35	80	28	143
# Events <= 3 tweets	100	123	83	306
# Events <= 3 tweets and validated by news	26	24	33	83
# total events reported by LITMUS	61	104	61	226

Table 3.4. Events summary

We would like to apply news source to validate unconfident events with less than three tweets. If the events can be validated by news, we can add it back to the report to increase the system coverage without losing quality assurance. There are about 27% of unconfident events validated by news.

Therefore, the final approach is, firstly, to include confident events with more than three tweets and, then, to include events which can be matched by news source. With the final approach, the system can report about 75 events each month.

Coverage. By comparing data from Table 3.1. and Table 3.2, events reported by news are much more than the events reported by tweets. In order to increase the coverage of LITMUS, we add news as a new information source and include events with more than three news items. As shown

in Table 3.5, there are 78 new events added to the system and we can reach 304 total reported events with LITMUS.

events	201801	201802	201803	total
# events reported by tweets	61	104	61	226
# events reported by news	32	17	29	78
# total events reported by LITMUS	93	121	90	304

Table 3.5. Events summary

Discussion & Future Works

The quality and the coverage of the LITMUS can be both improved by integrating news source. Two different information sources have different characteristics and specialties. Events reported by news tend to gain more authorities but these events often cover large-scale events in different countries because the news sources collected by Google News API are major news agents. If we can include local news source, the event coverage of news source will be increased. However, it is hard to find a developer's friendly way to collect local news.

Tweets outperform news on local events. Tweets are posted by the public from different areas and local events are more likely captured by tweets. Without the authorities, the quality of events reported by tweets keeps being questioned. We apply threshold and news validation to improve the quality. For the further study, we would like to explore official Twitter users. Official users on Twitter are those who typically are public figures in music, acting, fashion, government, politics, religion, journalism, media, sports, business, and other key interest areas (Twitter, 2018). The badge appears next to the name on account's profile. Especially the official users who dedicate to report local weather or nature disasters might have better authorities than the general Twitter users. We will analyze their impacts in the future studies.

CHAPTER VI

INTEGRATING MULTILINGUAL INFORMATION TO LITMUS

Introduction

As a global phenomenon, social media are open to all languages and all people. Although English has been adopted by many as a standard international language, it is common for people from each country to exchange information in their own language(s). Consequently, discussions on the events that happened in a country are more likely to have postings in the local language in addition to English. Disaster management systems always tend to adopt multiple information sources to achieve comprehensive event coverage. In order to further improve LITMUS event coverage, multilingual information is considered as a potential source. We integrate Chinese tweets into the system as an example to study the LITMUS improvements in credibility and coverage.

Evaluation

Chinese is selected as an example to study the potential improvements with integrating multilingual information from Twitter. Localizing the application in each language of interest is nontrivial. Two approaches of integration of additional languages into LITMUS are studied previously. The studies demonstrated that using a machine translation tool to automatically translate social media data from another language into English, and then using our English filters produces comparable or slightly better results than manually developing native filters (Hou et al., 2017). The pipeline for integrating new languages to LITMUS is shown in Figure 4.1.

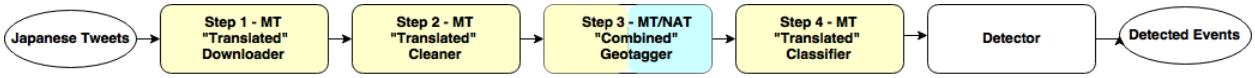

Figure 4.1. LITMUS multilingual pipeline

We collected Chinese tweets from 2014 to 2017. The pipeline above was applied to process the collected tweets. The number of tweets downloaded, geotagged, classified is shown in Table 4.1.

Chinese	2014	2015	2016	2017 total	
Downloader	9255	13036	10138	7380	39809
Cleaner	9189	12982	10102	7336	39609
Geotagger	7046	10531	7662	4453	29692
Classifier	5483	5270	4077	2218	17048
%useful tw	0.5924365	0.4042651	0.4021503	0.300542	0.4282449

Table 4.1. Summary collected Chinese tweets

We collected English tweets in the same time period and processed tweets as the pipeline shown in Figure 4.2. The number of tweets downloaded, geotagged, classified is shown in Table 4.2.

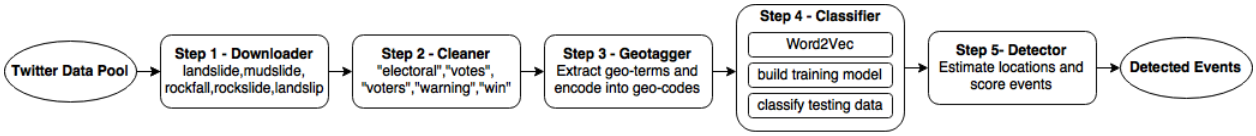

Figure 4.2. LITMUS pipeline

English	2014	2015	2016	2017 total	
Downloader	969361	548420	649873	403188	2570842
Cleaner	918927	522926	590250	349810	2381913
Geotagger	758973	433306	519292	305743	2017314
Classifier	261054	152209	138493	64693	616449
%useful tw	0.2693052	0.2775409	0.2131078	0.1604537	0.2397849

Table 4.2. Summary collected English tweets

With the additional language tweets, we would like to evaluate if the events can be validated (credibility) and if more events can be added to the system (coverage).

Credibility. As mentioned in Chapter Three, in order to ensure the quality of the system, we regard events with less than three tweets as unconfident events and news is used to validate the unconfident events to prevent deleting true events from the final report. We apply the same idea for multilingual tweets.

Firstly, we would like to use English tweets to validate the unconfident events reported by Chinese tweets. Then, reversely, we can use Chinese tweets to validate the unconfident events reported by English tweets. The results are shown in Table 4.3. and 4.4. There are about fifty unconfident events can be validated by another language each month. The approach will increase the coverage of LITMUS without losing information quality.

	2014	2015	2016	2017	
# events reported by Chinese	757	625	541	450	2373
# unconfident events reported by Chinese	431	373	322	288	1414
# events validated by English	69	48	57	39	213
# total events reported by Chinese	395	300	276	201	1172

Table 4.3. Summary of events reported by Chinese tweets

	2014	2015	2016	2017	
# events reported by English	2389	1985	2210	1553	8137
# unconfident events reported by English	1012	833	1004	709	3558
# events validated by Chinese	51	40	41	34	166
# total events reported by English	1428	1192	1247	878	4745

Table 4.4. Summary of events reported by English tweets

Coverage. As discussed in the introduction section, even though English is considered as standard language, people tend to report local events in their local languages. Since Chinese is used as an additional language, it is reasonable to expect events which can be reported by Chinese tweets but might be missed by English tweets. Therefore, we would like to investigate the number of events which are only reported by Chinese. Contrarily, there are also events reported by English only. If we combine the confidents events reported by both languages, the coverage of LITMUS will be improved.

As shown in Table 4.3. and Table 4.4., there are 1172 events reported by Chinese and 4745 events reported by English. 213 events are reported by both languages. If we combine the confident events, we can collect 5704 events from both languages. With Chinese as an additional language source, the coverage of LITMUS can be increased by about 20%.

Discussion & Future Work

As shown in the tables above, the quality and the coverage of the LITMUS can be both improved by integrating news source. Different languages, especially languages other than English, target different regions of the worlds. Even English tends be spoken by people throughout the world, majorities from the country where English is not the native language prefer to exchange information in their native language. With adding additional languages, the coverage of the events can be improved dramatically. Additionally, with the fast development of transportations, people

can easily travel to the other end of the world within days. One event might be reported by people from different countries with different languages. Integrating additional languages provides a way to validate the detected events.

However, it is not trivial to immigrate LITMUS pipeline to other languages. There are different approaches proposed previously to integrate new language to LITMUS system. We would like to further study the general way of localizing the applications to other language, especially the applications which use multiple information source, including social media information.

CHAPTER V

CONCLUSION & FUTURE WORK

Integrating additional information sources into LITMUS improves the trustworthiness and event coverage. It is risky to depend on information from one source. Multiple sources serve as validation agents for each other. LITMUS uses additional sources to validate the unconfident events reported by the system. Different sources provide LITMUS distinct and unique information, which helps to achieve the completeness.

In the further studies, we would like also to investigate if integrating additional information sources can improve the timeliness of LITMUS. There might be some information source which can report certain type of the events earlier than others.

The incorporation of additional information support is considered a laborious and expensive process. Especially, integrating multilingual support in application processing social media data has been regarded as expensive localization process. We studied different approaches previously and concluded that using automated translation to translate other languages to English and then processing data with existing English pipeline achieves comparable or better results (in terms of false positives, false negatives, and F1-score) as manually developing new pipelines for additional languages (Hou et al., 2017). Built on top of this conclusion, we are able to explore the advantages of integrating information in other languages.

Our future work includes exploring other potential sources to increase credibility, coverage and timeliness of LITMUS. We are also interested in how to reduce the costs of integrating new sources into an existing system.

REFERENCE

Musaev, A., Wang, D., & Pu, C. (2015). *Multi-hazard Detection by Integrating Social Media and Physical Sensors.* Social Media for Government Services, 395-409. doi:10.1007/978-3-319-27237-5_17

Fogg, B. J., Tseng, S. (1999). *The elements of computer credibility*. Proceedings of Computer Human Interface SIG conference, 80–87.

Qixuan, H., Musaev, A., Yang, Y., & Pu, C. (2017). *A Comparative Study of Increasing Automation in the Integration of Multilingual Social Media Information.* 2017 IEEE 3rd International Conference on Collaboration and Internet Computing (CIC), San Jose, CA, US. 319-327. doi:10.1109/CIC.2017.00049

Twitter Inc., https://about.twitter.com/company, accessed on 07/10/2017.

Jurgens, D. *That's What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships*. in Seventh International AAAI Conference on Weblogs and Social Media, 2013.

Finkel, J., Grenager, T., & Manning, C. *Incorporating Non-Local Information into Information Extraction Systems by Gibbs Sampling.* Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370.

Google Inc. *The Google Geocoding API.* https://developers.google.com/maps/documentation/geocoding/, accessed on 07/20/2017.

Google Inc. *Word2Vec*. https://code.google.com/archive/p/word2vec/, accessed on 04/10/2016.

Google Inc. *Google News API*. https://newsapi.org/s/google-news-api, accessed on 04/18/2018.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. *The WEKA data mining software*. ACM SIGKDD Explorations Newsletter, vol. 11, no. 1, 2009 .

RCN/SAVI: GRAIT-DM. *LITMUS Live Demonstration*. https://grait-dm.gatech.edu /demo-multi-source-integration/, accessed on 07/20/2017.

Musaev, A., Wang, D., Cho, C., Pu, C. *Landslide Detection Service Based on Composition of Physical and Social Information Services.* ICWS 2014.

Musaev, A., Wang, C., Pu, C. *LITMUS: Landslide Detection by Integrating Multiple Sources*. ISCRAM 2014

Musaev, A., Wang, C., Shridhar, S., Pu, C. *Toward a Real-time Service for Landslide Detection: Augmented Explicit Semantic Analysis and Clustering Composition Approaches*. ICWS 2015.

Musaev, A., Wang, C., Shridhar, S., Pu, C. *Fast Text Classification Using Randomized Explicit Semantic Analysis*. IEEE IRI 2015.

Lai, C., Donahue, J., Musaev, A., Pu, C. *Nimbus: Tuning Filters Service on Tweet Streams.* IEEE BigData Congress 2015.

Musaev, A., Wang, C., Pu, C. *Multi-Hazard Detection by Integrating Social Media and Physical Sensors*. Social Media for Government Services, Springer 2015, pp 395-409.

Tien, I., Musaev, A., Denas, D., Ghadi, A., Goodman, S., Pu, C. *Detection of Damage and Failure Events of Critical Public Infrastructure Using Social Sensor Big Data*. IoTBD 2016.

Salwen, M. B., Garrison, B., & Driscoll, P. D. (2005). Online news and the public. Mahwah NJ: Lawrence Erlbau.

Thorson, K., Vraga, E., Ekdale, B. (2010). *Credibility in context: How uncivil online commentary affects news credibility.* Mass Communication and Society, Taylor & Francis.

El-Nawawy, M. (2006). *US public diplomacy in the Arab world: The news credibility of Radio Sawa and Television Alhurra in five countries.* Global Media and Communication.

Kohring, M., Matthes, J. (2007). *Trust in news media: Development and validation of a multidimensional scale.* journals.sagepub.com.

Twitter (2018). *Official Twitter Account.* https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts. Accessed 04/12/2018.

Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. Journal of the American Society for Information Science and Technology,60(11), 2169-2188. doi:10.1002/asi.21149