Multi-hazard Detection by Integrating Social Media and Physical Sensors

Aibek Musaev, De Wang and Calton Pu

Abstract Disaster Management is one of the most important functions of the government. FEMA and CDC are two examples of government agencies directly charged with handling disasters, whereas USGS is a scientific agency oriented towards disaster research. But regardless of the type or purpose, each of the mentioned agencies utilizes Social Media as part of its activities. One of the uses of Social Media is in detection of disasters, such as earthquakes. But disasters may lead to other kinds of disasters, forming multi-hazards such as landslides. Effective detection and management of multi-hazards cannot rely only on one information source. In this chapter, we describe and evaluate a prototype implementation of a landslide detection system LITMUS, which combines multiple physical sensors and Social Media to handle the inherent varied origins and composition of multi-hazards. Our results demonstrate that LITMUS detects more landslides than the ones reported by an authoritative source.

Keywords LITMUS · Social media · Physical sensor · Disaster management · Landslide detection

1 Introduction

Government through its agencies plays a critical role in disaster management. There are multiple government agencies dealing with various aspects of disasters, including FEMA and CDC. The Federal Emergency Agency (FEMA) is a

A. Musaev (✉) · D. Wang · C. Pu
College of Computing, Georgia Institute of Technology, 266 Ferst Drive,
30332 Atlanta, Georgia
e-mail: aibek.musaev@gatech.edu

D. Wang
e-mail: wang6@gatech.edu

C. Pu
e-mail: calton.pu@cc.gatech.edu

S. Nepal et al. (eds.), *Social Media for Government Services*,
DOI 10.1007/978-3-319-27237-5_17

federal agency under the Department of Homeland Security, which is responsible for coordinating the response to a disaster. The Centers for Disease Control and Prevention (CDC) is a federal agency under the Department of Health and Human Services. It is responsible for emergency preparedness and response. Unlike these two major agencies that are directly charged with handling disasters, the United States Geological Survey (USGS) is a scientific agency. It studies the landscape of the United States, its natural resources and the natural hazards that threaten it. But regardless of the type or purpose, all of these agencies utilize Social Media as part of their activities.

The agencies maintain a number of Social Media accounts as part of their mission to disseminate information to the public and even offer digital toolkits to integrate such information into third party tools.[1] USGS uses Social Media channels to inform the public about various natural hazards, including earthquakes, landslides and volcanoes.[2] However, Social Media itself can be used as a source of data for disaster management instead of solely relying on physical sensors. A good example of exploring the data from Social Media is Twitter data streams functioning as social sensors [1]. Also, many existing disaster management systems adopt multiple information sources, including news channels. However, they all face the challenge of integrating multiple information sources in the way that preserves the useful information while limiting the amount of noise. We cannot depend on a single information source to make decisions, since each information source has its advantages and disadvantages. For instance, Social Media sources can provide real-time streaming information, but not all of such information is related to disasters that we are interested in. In fact, there is a high amount of noise in Social Media, which has been elaborated in our previous research study on denial of information [2–4]. Also, one interesting example of the noise about "landslide" is the 70s rock song "Landslide" by Fleetwood Mac. Twitter filter for the word "landslide" gets more tweets on this song than landslide disasters that involve soil movement. News channels provide reliable and mostly verified information sources. Unfortunately, they normally have high latency that may be up to several days after the occurrence of a disaster.

Besides, disasters like multi-hazards present more significant challenges, since there are no effective physical sensors that would detect multi-hazards directly. Landslide, which can be caused by earthquakes, rainfalls and human activity among other reasons, is an illustrative example of a multi-hazard. After investigating existing approaches using physical and social sensors, we proposed a new landslide detection service—LITMUS [5–7] and also implemented a prototype system in practice, which is based on a multi-service composition approach to the detection of landslides. More concretely, LITMUS has the following benefits compared with traditional or existing approaches for natural disaster detection:

[1] http://www.cdc.gov/socialmedia/tools/guidelines/socialmediatoolkit.html.

[2] https://twitter.com/usgsnewshazards.

- It composes information from a variety of sensor networks including both physical sensors (e.g., seismometers for earthquakes and weather satellites for rainfalls) and social sensors (e.g., Twitter and YouTube). Besides providing wider coverage than a system relying on a single source, it improves detection accuracy and reduces the overall latency.
- It applies state-of-art filters for each social sensor and then adopts geo-tagging to integrate the reported events from all physical and social sensors that refer to the same geo-location. Such integration achieves better landslide detection when compared to an authoritative source. Meanwhile, the geo-location information not only provides the base for the integration, but also enables us to do real-time notification in the future.
- It provides a generic approach to the composition of multiple heterogeneous information services and uses landslide detection as an illustrative example, i.e. it is not tied to disaster detection and can be applied to other application areas involving service composition. Traditional approach to the composition of web services makes strong assumptions about services, which it then uses to select services when composing a new service, such as quality of service [8] or service license compatibility [9]. In practice, the real world services do not satisfy such assumptions. The claim we make is that more information services should provide a more solid result and we demonstrate that it is the case with LITMUS.

The rest of the chapter is organized as follows. Section 2 provides an overview of the LITMUS system. We introduce the supported physical and social sources, and describe implementation details of each system component. In Sect. 3, we present an evaluation of landslide detection using real data and compare the results generated by LITMUS with an authoritative source. We summarize related work in Sect. 4 and conclude the chapter in Sect. 5.

2 System Overview

There are several stages in the LITMUS prototype that are implemented by the corresponding software components—see Fig. 1 for an overview of the system pipeline.

The data collection component downloads the data from multiple social and physical sources using provided API. The data from Social Media requires additional processing as it is usually not geo-tagged and contains a lot of noise. That is why the data from Social Media is geo-tagged followed by the filtering out of irrelevant items using stop words/phrases and classification algorithms. The integration component integrates the data from social and physical sources by performing grid-based location estimation of potential landslide locations followed by the computation of landslide probability to generate a report on detected landslides. This report includes all of the data related to detected landslides, i.e. the physical sensor readings as well as all tweets, images, and videos that were used to detect them.

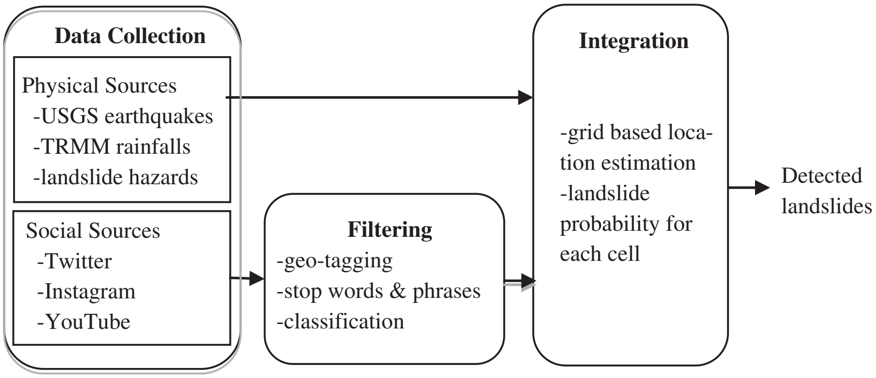

Fig. 1 Overview of system pipeline

2.1 Data Collection Component

Social Media feeds. There is a separate data collection process based on the capabilities provided by each data source. Among the currently supported data sources, Twitter has the most advanced API for accessing its data. In particular, it provides a Streaming API, which returns tweets in real-time containing the given keywords. Instead of storing the incoming tweets directly into a data store, LITMUS writes the tweets to a set of intermediate files first. The intermediate layer was introduced for two reasons. On the one hand we wanted to increase overall robustness, such that even if the data store failed we would still have the original files that we could restore the data from. On the other hand it allows us to easily switch to another data store if needed. The file structure of the intermediate layer is as follows:

```
<source_type>_< event_type>_<year>/<month>/<day>/<hour>/<min>.json
```

Note that when there are multiple incoming items per minute, then they get appended to the same file. The item IDs are used to make sure there are no duplicate records. The rate of incoming items containing landslide keywords is moderate, but we plan to add support for other types of events that would have a much higher rate of incoming items, such as “ebola” for instance. So, a file structure as this makes sure that the data is broken into manageable chunks.

The next step is to upload the incoming items to a data store. We use Redis, because it is an in-memory data store that is widely used and it is open source [10]. We keep the latest 30 days worth of data in the data store to maintain a fixed memory footprint. The new data is periodically uploaded into Redis and obsolete items are removed. The rest of the system works with Redis directly instead of files.

Both YouTube and Instagram provide a pull type of API that LITMUS uses to periodically download items containing landslide keywords. Again, the items from these Social Media get stored into the described file structure and the new items are periodically uploaded into Redis.

The rainfalls data is available due to the Tropical Rainfall Measuring Mission (TRMM) [11]. TRMM is a joint space project between NASA and the Japan Aerospace Exploration Agency (JAXA). The mission uses a satellite to collect data about tropical rainfalls. TRMM generates various reports based on its data, including a list of potential landslide areas due to extreme or prolonged rainfall. In particular, it generates reports of potential landslide areas after 1, 3, and 7 days of rainfall. The data is provided in HTML format, which LITMUS periodically downloads, parses and saves extracted content into data storage for further analysis. TRMM project has been operating since December 1997. However, on July 8, 2014 pressure readings from the fuel tank indicated that the TRMM satellite is near the end of its fuel. The satellite is estimated to be shutdown in February 2016, but JAXA may stop distribution of the radar data prior to that date. As of January 1, 2015 the data is still available.

The seismic feed is provided by the United States Geological Survey (USGS) agency [12]. USGS supports multiple feeds of earthquakes with various magnitudes. The data is provided in a convenient GeoJSON format, which is a format for encoding a variety of geographic data structures. LITMUS uses a real-time feed of earthquakes with 2.5 magnitude or higher, which gets updated every minute. USGS includes event id, which is used to avoid duplicate records in the system.

Global Landslide Hazards Distribution is another physical source that LITMUS supports [13]. It provides a 2.5 min grid of global landslide and snow avalanche hazards based upon the work of the Norwegian Geotechnical Institute (NGI). This source incorporates a range of data including slope, soil, precipitation and temperature among others. The hazard values in this source are ranked from 6 to 10, while the values below are ignored. The reason why this particular source is supported is because the landslides detected by LITMUS to occur in the landslide hazardous areas are more likely to be determined correctly as opposed to the landslides detected to occur in other areas.

2.2 *Filtering Component*

Geo-tagging. All Social Media supported by LITMUS allow users to disclose their location when they send a tweet, post an image or upload a video. However, based on the evaluation dataset collected in November 2014 very few users actually use this functionality. In particular, less than 0.77 % of all tweets are geotagged in our dataset. That is why we analyze the textual descriptions of the items from Social Media to see if they mention geographic terms in them.

A common approach implementing this idea is based on the use of a gazetteer. A gazetteer is a dictionary that maps geographic terms to geographic coordinates. An exact match of a sequence of words is performed against the gazetteer. Since we do not know in advance which particular word or sequence of words is a geographic term, all possible sequences are considered. This approach requires the presence of a local and relatively small gazetteer, since requests to remote or large gazetteers will significantly slow down the system, as the number of sequences of words in a text is very high.

Another weakness of this approach is that gazetteers often have geo terms that are common nouns, so they are used in texts a lot. For example, “Goes” is a city in Netherlands and “Enterprise” is a city in the United States. Most likely both words will be useless geo terms for the purposes of landslide detection and would have to be excluded from consideration by the system. Also, many news sources contain geographic terms in them, such as “Boston Globe” or “Jamaica Observer”. A geo-tagging algorithm would have to have a list of news sources in order to ignore such geographic terms automatically.

This is only a small fraction of issues that would have to be addressed in a geo-tagging algorithm based on the use of a gazetteer. Which is why LITMUS implements an alternative approach that employs a natural language processing technique called named entity recognition (NER).

NER implementations locate and classify elements in a text into pre-defined categories, including names of persons, organizations, dates and locations. For geo-tagging purposes LITMUS extracts sequences of words recognized as locations from text. Then it checks the found geo-terms against a local gazetteer. There is an open source project called GeoNames that provides a free gazetteer dump with more than 10 million places.[3] If the geo term is not found there, LITMUS makes a remote call to the Google Geocoding API[4] to obtain corresponding geographic coordinates, i.e. latitude and longitude values.

See Experimental Evaluation section for the results of the geo-tagging analysis performed by LITMUS during the evaluation period.

Stop words and phrases. During the process of building the ground truth dataset described below, we noticed that we could almost instantly tell whether a given social item was irrelevant to landslide as a natural disaster or not. There were several common irrelevant topics discussed in Social Media that were easy to spot due to the use of specific words, including “election”, “vote”, “parliament” and “Fleetwoodmac”, e.g.:

> What does the Republican election landslide mean?: VIRGINIA (WAVY) — What does the Republican landslide in the… http://t.co/2Alrs48SwK

> Landslide… and every woman in the Tacoma Dome wept with the beautiful @StevieNicks @fleetwoodmac #fleetwoodmacworldtour

Another common irrelevant topic is the use of the lyrics from a popular rock song from the 70’s to describe a user’s mood at the moment, e.g.:

> Well I’ve been afraid of changing cause I built my life around you #LandSlide

In this case instead of a particular stop word, we use excerpts from the lyrics of a popular song as a stop phrase instead.

[3] http://www.geonames.org/.

[4] https://developers.google.com/maps/documentation/geocoding/.

Stop words and phrases are easy to understand and fast to execute. So, LITMUS attempts to filter out items using stop words and phrases first before applying classification algorithm described next on the remaining items.

Classification algorithm. To decide whether an item from Social Media is relevant or irrelevant to landslide as a natural disaster, we propose the following approach. The textual description of each item is compared against the texts of relevant Wikipedia articles and the texts of irrelevant articles. Then we use the relevance of the article that is most similar to the given item as our decision.

For a list of relevant articles, we use the landslide keywords as Wikipedia concepts, namely landslide, landslip, mudslide, rockfall, and rockslide. These articles are downloaded, parsed and all HTML markup is removed, so that only their content is used for analysis. In addition to these articles, we also use a set of articles describing actual occurrences of landslides, mudslides, and rockslides, including 2014 Pune landslide, 2014 Oso mudslide, and Frank Slide. For a list of irrelevant articles, we use the landslide stop words to download the corresponding Wikipedia articles, namely Landslide victory, Blowout (sports), Election, Landslide (song), and Politics. Similarly, these articles are downloaded, parsed and all HTML markup is removed, so that only their texts are used for analysis.

To compute the distance between social items and these Wikipedia articles we use a formula named after Swiss Professor Paul Jaccard. He compared how similar different regions were based on the following formula:

$$\frac{Number\ of\ species\ common\ to\ the\ two\ regions}{Total\ number\ of\ species\ in\ the\ two\ regions}$$

This formula gives 0 if the sets have no common elements and 1 if they are the same. This is the opposite of what we need as a similarity measure, so we use the following formula instead:

$$Jaccard\ distance = 1 - \frac{Intersection(A, B)}{Union(A, B)},$$

where A and B are the sets that we want to compare.

Each article is converted to a bag of words representation or more precisely a set of words. Each incoming item from Social Media is also converted to a set of words representation. Now these sets can be used to compute the Jaccard distance between them.

Using this approach we were able to successfully classify items in November 2014. Table 1 lists the examples of items from Social Media together with the smallest Jaccard distance values and corresponding Wikipedia concepts. See the Experimental Evaluation section for more details on the experiment.

2.3 Integration Component

Previously the items from social sources have been geo-tagged and classified as either relevant or irrelevant to landslide as a natural disaster. The items from physical

Table 1 Examples of classification of items

Text	Jaccard distance	Wikipedia concept	Decision
Bad weather hampers rescue operations at Sri Lanka's landslide http://t.co/vYYgwRL1S6 #ANN	0.9916317991631799	2014 Pune landslide	1
Bertam Valley still deadly: After a mudslide claimed four lives and left 100 homeless, the danger is far from… http://t.co/ZiauH2YVvJ	0.9913366336633663	2014 Oso mudslide	1
#bjpdrama World's knowledge in 1 hand site: BJP got landslide Will India become a 1 party state like China Russia http://t.co/jGhp1j84az	0.9847715736040609	Landslide victory, wave election	0

sources are already geo-tagged and there is no need to classify them, as they are all considered relevant to landslide as a natural disaster. Now that we have the items' geographic coordinates, namely their latitude and longitude values, we want to integrate the data based on those values. One possible way of doing it is to divide the surface of the planet into cells of a grid. Items from each source are mapped to the cells in this grid based on their latitude/longitude values. Obviously, the size of these cells is important, because it can range from the smallest possible size to the one covering the whole planet. The smaller the cells, the less the chance that related items will be mapped to the same cell. But the bigger the cells, the more events are mapped to the same cell making it virtually impossible to distinguish one event from another.

Currently we use a 2.5-min grid both in latitude and longitude, which corresponds to the resolution of the Global Landslide Hazard Distribution described above. This is the maximum resolution of an event supported by the system at the moment.

The total number of cells in our grid is huge as cells are 2.5 min in both latitude and longitude, there are 60 min per degree, latitude values range from $-90°$ to $+90°$ and longitude values range from $-180°$ to $+180°$. But the actual number of cells under consideration is much smaller, because LITMUS only analyzes non-empty cells. For example, there are only 1192 candidate cells during the evaluation month of November 2014 as you can see in the Experimental Evaluation section below.

Next we consider each non-empty cell to decide whether there was a landslide event there. To calculate the probability of a landslide event w in cell x, we use the following weighted sum formula as the strategy to integrate data from multiple sources:

$$P(w|x) = \sum_i R_i \frac{\sum_j POS_{ij}^x - \sum_j NEG_{ij}^x - \sum_j STOP_{ij}^x}{\sum_i N_i^x}$$

Here, R_i denotes ith sensor's weight or confidence; POS_{ij}^{x} denotes positively classified items from sensor i in cell x, NEG_{ij}^{x} denotes negatively classified items from sensor i in cell x, $STOP_{ij}^{x}$ denotes the items from sensor i in cell x that have been labeled as irrelevant based on stop words and stop phrases, and N_i^x denotes the total number of items from sensor i in cell x.

In our prototype, we use prior F-measure R as the confidence for each sensor, since F-*measure* provides a balance between precision and recall, namely $F\text{-}measure = 2 * \frac{precision*recall}{precision+recall}$. To generate results in the range from 0 to 1, we normalize the values of F-*measure* into a scale between 0 and 1.

Finally, it should be noted that the given formula generates a score between 0 and 1 that can be used to rank all location cells based on the probability of a landslide occurrence there.

3 Experimental Evaluation

In this section, we perform an evaluation of LITMUS using real-world data. In particular, we design an experiment to compare the performance of landslide detection by LITMUS versus an authoritative source. We show that LITMUS manages to detect 41 out of 45 events reported by the authoritative source during evaluation period as well as 165 additional locations. We also describe the collection of the ground truth dataset and provide the details of the dataset collected by LITMUS during this period.

3.1 Evaluation Dataset

We select the month of November 2014 as the evaluation period. Here is an overview of the data collected by LITMUS during this period—see Table 2.

For each geo-tagged item, LITMUS also computes its cell based on its latitude and longitude. The total number of cells during the evaluation period is equal to 1192. Hence, there are 1192 candidate locations that LITMUS has to mark as either relevant or irrelevant to landslide as a natural disaster.

Table 2 Overview of evaluation dataset

Social media	Raw data	Geo-tagged data
Twitter	83,909	13,335
Instagram	2026	460
YouTube	7186	2312

3.2 Ground Truth Dataset

In order to collect the ground truth dataset for the month of November, we consider all items that are successfully geo-tagged during this month. For each such geo-tagged item, we compute its cell based on its latitude and longitude values. All cells during November represent a set of candidate events, which is 1192 as shown above. Next we group all geo-tagged items from Social Media by their cell values. For each cell we look at each item to see whether it is relevant to landslide as a natural disaster or not. If the item's textual description contains URL, then we look at the URL to confirm the candidate item's relevance to landslides. If the item does not contain a URL, then we try to find confirmation of the described event on the Internet using the textual description as our search query. If another trustworthy source confirms the landslide occurrence in that area then we mark the corresponding cell as relevant. Otherwise we mark it as irrelevant. It should be noted that we consider all events reported by USGS as ground truth as well.

Overall, there are 212 cells that we marked as relevant. The following are a few examples of social activity related to the events in those cells:

> Landslide on route to Genting Highlands: PETALING JAYA: A landslide occurred at 4.2KM heading towards Genting… http://t.co/AYfCKy6H2n

> Major back up on HWY 403 Toronto bound in Hamilton due to mudslide. ALL lanes closed at 403 between Main & York. http://t.co/QcRJdjydR1

> Trains cancelled between Par and Newquay due to landslip http://t.co/IcGsdS3y5r

3.3 Comparison of Landslide Detection Versus Authoritative Source

In November 2014 USGS posted links to 45 articles related to landslides.[5] LITMUS detects events described in 41 of them, i.e. over 90 % of events reported by the authoritative source were detected by our system. In addition to 41 locations described in these articles, LITMUS managed to detect 165 locations unreported by USGS during this period.

Hence, there are only 4 events reported by USGS that were missed by LITMUS during this period. Next we provide explanation why LITMUS did not detect the events described in these articles.

Out of these 4 articles, 2 did not report recent natural disasters. In particular, one article suggests that Bilayat grass, also called trap grass, can be used to prevent landslides in the hills of Uttarakhand.[6] The other article describes the reopening of

[5] http://landslides.usgs.gov/recent/index.php?year=2014&month=Nov.

[6] http://timesofindia.indiatimes.com/city/dehradun/Now-a-grass-that-could-prevent-landslides/articleshow/45196678.cms.

the Haast Pass in New Zealand.[7] It was closed nightly since a major slip last year and it will stay open due to a three-net system that protects the pass against rock fall.

The third article describes a minor event that did not receive much attention in Twitter, Instagram or YouTube. In particular, this article is a link to an image in Wikipedia of a minor rock fall on Angeles Crest Highway in California.[8]

Finally, the fourth article is about a route in Costa Rica that remains closed due to recent landslides in that area.[9] There were many tweets on this subject in Spanish, but not much activity in English. LITMUS currently supports English language only, which is why it missed this event. We are already working on adding support for other languages, including Spanish. See Conclusion and Future Work section for more details.

As we mentioned earlier, LITMUS detected 165 locations unreported by the authoritative source during this period. The reasons why LITMUS manages to detect more landslide events than the authoritative source are twofold. On the one hand we claim that our approach is comprehensive as it is fully automated, so it processes all items from each supported data source as opposed to a manual approach where an expert may miss an event due to a human error or human limits. On the other hand LITMUS integrates multiple sources in its analysis, both physical and social, and we plan to add more sources over time. See Conclusion and Future Work section for more details.

Overall, LITMUS detected 41 locations reported by USGS and 165 locations more, which is 206 locations out of 212 total ground truth locations, i.e. a landslide detection rate of over 97 % during this period.

4 Related Work

Event analysis using Social Media received a lot of attention from the research community recently. Guy et al. [14] introduced Twitter Earthquake Dispatcher (TED) that gauges public's interest in a particular earthquake using bursts in social activity on Twitter. Sakaki et al. [1] applied machine learning techniques to detect earthquakes by considering each Twitter user as a sensor. Cameron et al. [15] developed platform and client tools to identify relevant Twitter messages that can be used to inform the situation awareness of an emergency incident as it unfolds. Musaev et al. [5–7] introduced a landslide detection system LITMUS based on integration of multiple social and physical sources. We provide an overview of LITMUS implementation in this work, demonstrate its advantages using a recent evaluation period and describe enhancements made.

[7] http://www.radionz.co.nz/news/regional/258610/pass-reopens-with-rock-fall-protection.

[8] http://en.wikipedia.org/wiki/File:Minor_rockfall_on_Angeles_Crest_Highway_2014-11-05.jpg.

[9] http://thecostaricanews.com/route-27-remains-closed-due-to-landslides.

Document classification or document categorization is one of the most studied areas in computer science due to its importance. The problem is to assign a document to one or more classes or categories from a predefined set. Sakaki et al. [1] described a real-time earthquake detection system where they classified tweets into relevant and irrelevant categories using a support vector machine based on features such as keywords in a tweet, the number of words, and their context. Musaev et al. [6] improved the overall accuracy of supervised classification of tweets by converting the filtering problem of each item to the filtering problem of the aggregation of items assigned to each event location. Gabrilovich and Markovitch [16, 17] proposed to enhance text categorization with encyclopedia knowledge, such as Wikipedia. Each Wikipedia article represents a concept, and documents are represented in the feature space of words and relevant Wikipedia concepts. Their Explicit Semantic Analysis (ESA) method explicitly represents the meaning of any text as a weighted vector of Wikipedia-based concepts and identifies the most relevant encyclopedia articles across a diverse collection of datasets. In our work we identify two classes of Wikipedia articles that contain either relevant or irrelevant to landslides articles. Then we use Jaccard distance instead of a weighted vector to find the most similar article to a given social item. Finally we use the article's class as a decision for the social item's relevance to landslides.

Accurate identification of disaster event locations is an important aspect for disaster detection systems. The challenge for Social Media based analysis is that users do not disclose their location when reporting disaster events or that they may use alias or location names in different granularities in messages resulting in inaccurate location information. Cheng et al. [18] proposed and evaluated a probabilistic framework for estimating a Twitter user's city-level location based on the content of tweets, even in the absence of any other geospatial cues. Hecht et al. [19] showed that 34 % of users did not provide real location information, and they also demonstrated that a classifier could be used to make predictions about users' locations. Sultanik and Fink [20] used an indexed gazetteer for rapid geo-tagging and disambiguation of Social Media texts. Musaev et al. [7] evaluated three geotagging algorithms based on the use of gazetteer and named entity recognition approaches. In our work we employ the named entity recognition approach to identify all location entities mentioned in Social Media first. Then we use a public gazetteer to retrieve geographic coordinates for the found locations. If there is no match in the gazetteer, then LITMUS uses the Google Geocoding API to convert locations into geographic coordinates.

5 Conclusion and Future Work

In this chapter, we described and evaluated a prototype implementation of a landslide detection system called LITMUS, which combines multiple physical sensors and Social Media to handle the inherent varied origins and composition of multihazards. LITMUS integrates near real-time data from USGS seismic network, NASA TRMM rainfall network, Twitter, YouTube, Instagram as well as a global landslide hazards map. The landslide detection process consists of several stages of Social Media filtering and integration with physical sensor data, with a final ranking of relevance by integrated signal strength. Our results demonstrate that with such approach LITMUS detects 41 out of 45 reported events as well as 165 events that were unreported by the authoritative source during the evaluation period.

As we showed in the Experimental Evaluation section, LITMUS missed four events reported by USGS in November 2014. One of the events did not have much activity in English, but it did receive more attention in Spanish as it occurred in Costa Rica. That is why we are already working on adding support to LITMUS for event detection in other languages, including Spanish and Chinese. The data from Social Media in different languages can be considered as additional data sources, which will increase the coverage of event detection by LITMUS. It should also be noted that different languages have varying amounts of noise depending on the used keywords. For example, a "mudslide" in Russian is "оползень". We were surprised to find that the overwhelming majority of items in Social Media containing this word are relevant to mudslide as a natural hazard, which is an interesting fact that we plan to explore.

One of our objectives in this project is to analyze the possibility of predicting landslides in LITMUS. We have been collecting data in LITMUS since August 2013. Our plan is to eventually be able to predict landslide events based on the data from multiple sources, both physical and social. Landslides are an illustrative example of a multi-hazard disaster and we plan to study the possibility of predicting landslides in LITMUS using not only real-time data feeds from multiple sources, but also historical data that we collected.

We also believe that comprehensive and real-time information about landslide events can be useful not only to government agencies, but also research and journalism communities. That is why we are developing an automated notification system that people and organizations can subscribe to in order to receive real-time information on major landslides. This service will provide all relevant information collected by LITMUS, including tweets, images and videos related to each detected event.

Finally, the prototype landslide detection system LITMUS is live and openly accessible,[10] collecting data and displaying detection results in real-time for continued evaluation and improvement of the system.

Acknowledgements This research has been partially funded by National Science Foundation by CNS/SAVI (1250260, 1402266), IUCRC/FRP (1127904), CISE/CNS (1138666, 1421561), NetSE (0905493) programs, and gifts, grants, or contracts from Fujitsu, Singapore Government, and Georgia Tech Foundation through the John P. Imlay, Jr. Chair endowment. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or other funding agencies and companies mentioned above.

[10] https://grait-dm.gatech.edu/demo-multi-source-integration/.

References

1. Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes twitter users: Real-time event detection by social sensors. In *19th International Conference on World Wide Web (WWW)*. Raleigh, North Carolina.
2. Wang, D., Irani, D., & Pu, C. (2011). A social-spam detection framework. In *8th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference*. Perth, Australia.
3. Wang, D., Irani, D., & Pu, C. (2013). A study on evolution of email spam over fifteen years. In *9th International Conference Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom)*. Austin, Texas.
4. Wang, D. (2014). Analysis and detection of low quality information in social networks. In *IEEE 30th International Conference on Data Engineering Workshops (ICDEW)*. Chicago, Illinois.
5. Musaev, A., Wang, D., & Pu, C. (2014). LITMUS: Landslide detection by integrating multiple sources. In *11th International Conference Information Systems for Crisis Response and Management (ISCRAM)*. Pennsylvania: University Park.
6. Musaev, A., Wang, D., Cho, C.-A., & Pu, C. (2014). Landslide detection service based on composition of physical and social information services. In *21st IEEE International Conference on Web Services (ICWS)*. Anchorage, Alaska.
7. Musaev, A., Wang, D., & Pu, C. (2014). LITMUS: A multi-service composition system for landslide detection. In *IEEE Transactions on Services Computing* (No. 99).
8. Ran, S. (2003). A model for web services discovery with QoS. In *ACM SIGecom Exchanges* (Vol. 4, no. 1).
9. Gangadharan, G., Weiss, M., DAndrea, V., & Iannella, R. (2007). Service license composition and compatibility analysis. In *5th International Conference on Service Oriented Computing (ICSOC)*. Vienna, Austria.
10. Redis: An open-source advanced key-value store. http://redis.io. Accessed January 1, 2015.
11. Tropical Rainfall Measuring Mission (TRMM). http://trmm.gsfc.nasa.gov. Accessed January 1, 2015.
12. Earthquakes Hazards Program, United States Geological Survey. http://earthquake.usgs.gov. Accessed January 1, 2015.
13. Center for Hazards and Risk Research—CHRR—Columbia University, Center for International Earth Science Information Network—CIESIN—Columbia University, and Norwegian Geotechnical Institute—NGI. (2005). *Global Landslide Hazard Distribution*. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). http://dx.doi.org/10.7927/H4P848VZ. Accessed January 1, 2015.
14. Guy, M., Earle, P., Ostrum, C., Gruchalla, K., & Horvath, S. (2010). Integration and dissemination of citizen reported and seismically derived earthquake information via social network technologies. In *Intelligent Data Analysis IX*. Tucson, Arizona.
15. Cameron, M. A., Power, R., Robinson, B., & Yin, J. (2012). Emergency situation awareness from twitter for crisis management. In *1st Workshop on Social Web for Disaster Management (SWDM)*. Lyon, France.
16. Gabrilovich, E., & Markovitch, S. (2006). Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. In *National Conference on Artificial Intelligence (AAAI)*. Boston, Massachusetts.
17. Gabrilovich, E., & Markovich, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *20th International Joint Conference on Artificial Intelligence (IJCAI)*. Hyderabad, India.
18. Cheng, Z., Caverlee, J., & Lee, K. (2010). You are where you tweet: A content-based approach to geo-locating twitter users. In *19th ACM international conference on Information and Knowledge Management (CIKM)*. Toronto, Ontario, Canada.

19. Hecht, B., Hong, L., Suh, B., & Chi, E. H. (2011). Tweets from Justin Bieber's heart: The dynamics of the "location" field in user profiles. In *Conference on Human Factors in Computing Systems (CHI)*. Vancouver, Canada.
20. Sultanik, E. A., & Fink, C. (2012). Rapid geotagging and disambiguation of social media text via an indexed gazetteer. In *9th International Conference Information Systems for Crisis Response and Management (ISCRAM)*. Vancouver, Canada.