Toward a Real-time Service for Landslide Detection: Augmented Explicit Semantic Analysis and Clustering Composition Approaches

Aibek Musaev, De Wang, Saajan Shridhar, Chien-An Lai, Calton Pu
Georgia Institute of Technology, Atlanta, Georgia
{*aibek.musaev, wang6, saajan, calai, calton.pu*}*@gatech.edu*

***Abstract*—The use of Social Media for event detection, such as detection of natural disasters, has gained a booming interest from research community as Social Media has become an immensely important source of real-time information. However, it poses a number of challenges with respect to high volume, noisy information and lack of geo-tagged data. Extraction of high quality information (e.g., accurate locations of events) while maintaining good performance (e.g., low latency) are the major problems. In this paper, we propose two approaches for tackling these issues: an augmented Explicit Semantic Analysis approach for rapid classification and a composition of clustering algorithms for location estimation. Our experiments demonstrate over 98% in precision, recall and F-measure when classifying Social Media data while producing a 20% improvement in location estimation due to clustering composition approach. We implement these approaches as part of the landslide detection service LITMUS, which is live and openly accessible for continued evaluation and use.**

***Keywords*-landslide detection service; semantic relatedness; clustering composition; social media; event detection**

I. Introduction

LITMUS uses the following set of keywords to extract the data from Social Media related to landslides: *landslide*, *mudslide*, *rockslide*, *rockfall*, and *landslip*. Here is an example of a relevant data item returned by the Twitter Streaming API[1] that contains information regarding a recent event in Indonesia:

- "Indonesia rescuers use earth-movers in landslide rescue as toll rises to 24: JAKARTA (Reuters) - Indonesian... http://t.co/2Vk2k0gObu"

However, most of the data returned by social information services are irrelevant to landslide as a natural disaster. The following are frequent examples of irrelevant data items from Social Media that we consider as noise:

- *landslide* as an adjective describing an overwhelming majority of votes or victory: "The Bills held P Manning to 57 rating today in Denver - worst this year by a landslide - only game without a TD pass this year."
- *landslide* as the Fleetwood Mac song "Landslide" from the 1975 album *Fleetwood Mac*: "But time makes you bolder•Even children get older•And I'm getting older too. -Fleetwood Mac #landslide #dreams #gypsy #fleetwoodmac #onwiththeshow #thegrouch #eligh #music #family #longhairdontcare #theforum #california"
- *mudslide* as a popular cocktail: "The best dessert I found at Brightspot yesterday, not too sweet! @creamy-comfort #baileys #dessert #mudslide #brightspot brightspot"

One of the trivial approaches for finding irrelevant items is based on the presence of specific words in the items' texts, including *election*, *vote*, *fleetwoodmac* or specific phrases from song lyrics. However, even after applying this labeling technique, many unlabeled items remain that require a more sophisticated labeling approach as demonstrated in this tweet:

- "A serious breakdown of the numbers shows the better player is Michael Jordan. In a landslide. http://53eig.ht/1GvjDT1"

[1] suggested to employ a machine learning technique called text classification to automatically label each tweet as either relevant or irrelevant to a disaster event, such as earthquake. A research in natural language processing has found Explicit Semantic Analysis (ESA) to be successful for text classification [2]. However, the ESA algorithm relies heavily on leveraging existing knowledge of Wikipedia, which is very time-consuming and parts of it may be irrelevant for our purpose. Hence, the first contribution of this paper is rapid classification by augmenting ESA, such that instead of using all concepts from Wikipedia articles, we determine a subset of concepts based on a training set that allows us to rapidly classify Social Media texts while leveraging the capabilities of ESA as a superior text classifier. Our augmented ESA approach allows a user to rapidly classify unstructured texts, such that a preprocessing step is more than 7 times faster and throughput is an order of magnitude faster compared to the original ESA approach.

Not only the data from Social Media contain a lot of noise, but most of the data do not have geo-location either [3]. A common approach for geo-tagging such data is to look for mentions of places in Social Media texts using a gazetteer [4] or a named entity recognition (NER) approach, which generates fewer irrelevant locations [5]. However, even the NER based approach may extract incorrect locations. Con-

[1] https://dev.twitter.com/streaming/reference/post/statuses/filter

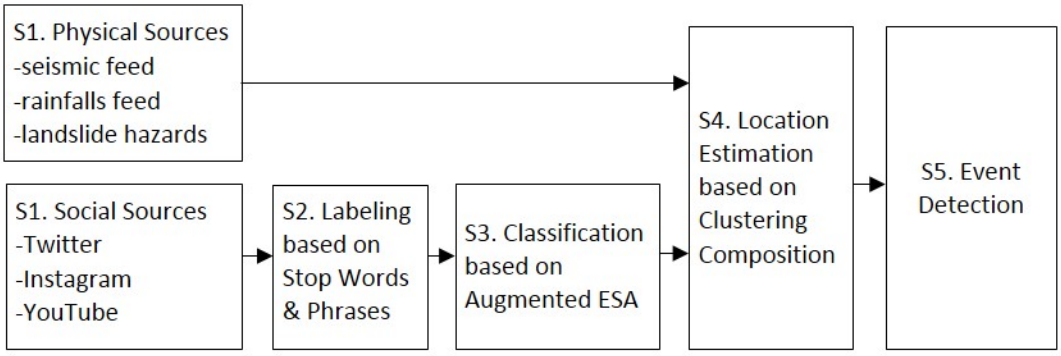

Figure 1. System Pipeline

sider the following tweet that was posted in December 2014:

- "On the Front Page of Personal Thailand Search for survivors begins after Indonesia landslide kills 18, leaves 90... http://t.co/bcwUzWNqmb"

The NER library incorrectly extracts *Thailand* as the location entity for this tweet, which is an outlier as the location for majority of tweets regarding the disaster event in Indonesia is determined correctly. That is why we propose to cluster Social Media texts based on semantic clustering and to find location outliers for each such cluster.

A further challenge in identifying locations of the detected events is that a single event may comprise multiple locations, which is important to address in order to avoid reporting the same event multiple times. Consider the following tweets mentioning locations affected by mudslide:

- "#LosAngeles News Amid Mudslide Concerns, Glendora Residents Prepare for More Rain: ... http://t.co/VhwIlQ6nCC"
- "Mudslide covers yard of an evacuating resident in Azusa, CA. Taken by @smasunaga: "This is a regulation hoop" http://t.co/xuhVVrHLbx"

Glendora[2] and Azusa[3] are neighboring cities in California that were affected by the same mudslide event, which is why we propose that outlier removal using semantic clustering should be followed by Euclidean clustering, such that locations that are in close proximity to one another are grouped into one cluster. Thus, the second contribution of this paper is that a composition of clustering algorithms is needed for accurate estimation of locations of the detected events. Based on our knowledge, this is the first work that employs a composition of clustering algorithms to accurately estimate geographic locations based on unstructured texts.

The rest of the paper is organized as follows. We describe the details of system components in Section II followed by implementation notes in Section III. In Section IV we present an evaluation of system components using real data and compare detection results generated by LITMUS with an authoritative source. We summarize related work in Section V and conclude the paper in Section VI.

[2]http://cityofglendora.org/about-glendora

[3]http://www.ci.azusa.ca.us/index.aspx?nid=569

II. System Description

LITMUS performs a series of processing steps before generating a list of detected landslides - see Figure 1 for an overview of the system pipeline.

LITMUS starts by collecting the data from multiple social and physical information services. The data collection component currently supports the seismic feed from USGS, the rainfall reports from the TRMM project, and the global landslide hazards map by NGI as its physical sources; as well as Twitter, Instagram and YouTube as its social sources.

The data from Social Media requires additional processing as it is usually not geo-tagged and contains a lot of noise. Hence, LITMUS attempts to determine the relevance of the social items to landslide as a disaster and labels the items accordingly. This is performed using labeling based on the presence of stop words and phrases in the items' texts. It is followed by classification based on augmented ESA approach on the remaining unlabeled items. Next LITMUS applies the geo-tagging component based on NER and estimates event locations using a composition of clustering algorithms.

The final component considers each cluster as a potential event and computes its landslide probability using a Bayesian model integration strategy.

The following subsections provide implementation details of the system components. Classification based on augmented ESA as well as location estimation based on clustering composition are the paper's main contributions, which is why they are described in separate sections for clarity.

S1. Data Collection

LITMUS collects data from both physical and social information services. Physical services alone are not sufficient as there are no physical sensors that would detect landslides directly. However, they can detect potential causes of landslides, including earthquakes and rainfalls.

The seismic feed is provided by the United States Geological Survey (USGS) agency [6]. USGS supports multiple feeds of earthquakes with various magnitudes. The data is provided in a convenient GeoJSON format[4], 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 that gets updated every minute.

The rainfalls data is available due to the Tropical Rainfall Measuring Mission (TRMM) [7]. 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 collected by LITMUS.

In addition to the described physical information services, LITMUS also collects data from social information services, including Twitter, Instagram and YouTube. There is a separate data collection process based on the capabilities provided by each information service.

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. This is implemented by connecting to a public stream provided by Twitter whereby tweets matching one or more keywords are returned in real-time. The connection is long-lived and held by Twitter servers indefinitely barring server-side error, excessive client-side lag or duplicate logins among other reasons.

Both YouTube and Instagram provide a pull type of API that LITMUS uses to periodically download items containing landslide keywords. This approach requires developers to implement a mechanism that avoids data duplication in the system. LITMUS uses item IDs to make sure there are no duplicates.

Finally, LITMUS incorporates another physical information source, which is a static map of areas on the planet that are likely to have landslides [8]. It is a 2.5-minute[5] grid of global landslide and snow avalanche hazards based upon the work of the Norwegian Geotechnical Institute (NGI). This dataset is based on 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.

[4]http://earthquake.usgs.gov/earthquakes/feed/v1.0/geojson.php

[5]http://education.usgs.gov/lessons/coordinatesystems.pdf

S2. Labeling based on Stop Words & Phrases

As we mentioned earlier, there are several common irrelevant topics discussed in Social Media that are easy to detect due to the use of specific words, including *election*, *vote* and *fleetwood*, or the use of the lyrics from popular rock songs to describe a user's mood at the moment - see [5], [9] for examples from Social Media. Stop words and phrases are easy to understand and fast to execute. Hence, LITMUS attempts to label items from Social Media using stop words and phrases before applying classification algorithm described next. The reason why we label items instead of removing them is because we want to penalize candidate landslide locations whose majority label is negative. For detailed description of the penalized classification approach we refer the reader to [5].

S3. Classification based on Augmented ESA

In [2] Gabrilovich, et al. described Explicit Semantic Analysis (ESA) approach and used it for text categorization as well as for computing the degree of semantic relatedness between fragments of natural language text. ESA represents the meaning of any text in terms of Wikipedia-based concepts. Concepts are the titles of Wikipedia articles characterized by the bodies of those articles. In ESA a word is represented as a column vector in the TF-IDF table (table T) of Wikipedia concepts and a document is represented using its interpretation vector, which is a centroid of the column vectors representing its words. An entry $T[i, j]$ in the table of size $N \times M$ corresponds to the TF-IDF value of term t_i in document d_j, where M is the number of Wikipedia documents (articles) and N is the number of terms in those documents.

The ESA approach proved to successfully measure semantic relatedness [2], but it requires a substantial amount of computation in order to build the semantic interpreter. The ESA authors reported that it took 7 hours to parse a complete XML dump of Wikipedia back in 2009 [2], whereas the number of articles in Wikipedia only increased since then. We propose to augment the ESA approach to rapidly classify the texts of Social Media items as either relevant or irrelevant. Rapid classification is achieved by reducing the number of Wikipedia concepts to consider as follows.

Given a training dataset we propose to group similar items using a clustering algorithm, such as K-means. For each such cluster the top N terms are selected next based on their TF-IDF values. Using the selected terms as Wikipedia concepts, the ESA method can then be applied to build table T also referred to as an *inverted index*. Inverted index is used to generate values of the interpretation vectors for each text in the training dataset. These rows of vector values are used by a classifier to build a model, which can be utilized to predict relevance of social items. For each social item's text an interpretation vector is computed using the inverted index

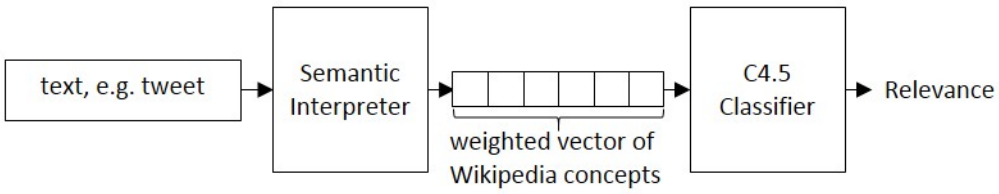

Figure 2. Classification based on augmented ESA

and its relevance label is predicted using the classifier's model - see Figure 2. The dataset collection is described in Subsection IV-A.

In addition to using a subset of Wikipedia concepts for rapid computation, we also described how to use interpretation vector values for classification purposes. The original ESA approach uses a *bag of words* (BOW) approach in conjunction with the top 10 concepts of all the interpretation vectors. Classification based on interpretation vector values represents a *bag of concepts* (BOC) approach, because the dimensions in those vectors are Wikipedia concepts. This method fully utilizes the strengths of the ESA approach, because all of the selected concepts are used for classification purposes instead of their subset [2].

S4. Location Estimation based on Clustering Composition

A. Location Estimation Using Semantic Clustering

Majority of items from Social Media do not have geo-location, although each of the supported social sources, namely Twitter, Instagram and YouTube, allow users to disclose their location when they send a tweet, post an image or upload a video. For example, only 0.8% of tweets have geo-location in our evaluation dataset - see Table I. That is why LITMUS contains a geo-tagging component that attempts to determine the locations of the discussed events by looking for mentions of places in the textual description of the social items. Then it assigns geographic coordinates based on the found geo terms.

In order to find mentions of places in the texts, LITMUS employs an NLP technique called named entity recognition (NER). This technique attempts to recognize various entities in a text, including organizations, persons, dates and locations. We are interested in the location entity for geo-tagging purposes. Once location entities are determined, we can use Google Geocoding API [10] to obtain corresponding geographic coordinates.

LITMUS utilizes Stanford CoreNLP library, which is a Java suite of NLP tools [11], to identify all location entities mentioned in Social Media texts. However, the CoreNLP library occasionally extracts incorrect entities. Consider the following tweet that was posted in December 2014:

- "DTN Mongolia: At least 24 dead in Java landslide: A landslide destroyed a remote village in Java, Indonesia, k... http://t.co/mQUGKYSxWZ"

The NER library incorrectly extracts *Mongolia* as the location entity for this tweet. This is an outlier as for most tweets regarding the disaster event in Indonesia, the library extracts correct geo-terms. That is why we propose to cluster social items based on semantic distance and for each cluster to find such outliers, such that if an overwhelming geo-term exists in a cluster then the location for all social items in the cluster is set to that geo-term. In this particular example, the overwhelming geo-term in the cluster to which these tweets belong to is *Indonesia*, that is why the location for this tweet is reset by LITMUS accordingly.

B. Location Estimation Using Euclidean Clustering

In order to estimate locations of landslide events based on data from multiple information services, originally we employed a cell-based approach [9]. The surface of the Earth was represented as a grid of cells and each geo-tagged item was mapped to a cell in this grid based on the item's geographic coordinates.

Obviously, the size of these cells is important. 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

Social Media	Raw Data	Data geo-tagged by user	Data geo-tagged by LITMUS
Twitter	149798	1242 (0.8%)	55054 (36.8%)
YouTube	6533	416 (6.4%)	2749 (42%)
Instagram	4929	788 (16%)	1139 (23.1%)

Table I
Overview of evaluation dataset

impossible to distinguish one event from another. The size we used was a 2.5-minute grid both in latitude and longitude, which corresponds to the resolution of the Global Landslide Hazard Distribution described earlier. That was the maximum resolution of an event supported by the system.

The formulas to compute a cell's row and column based on its latitude (N) and longitude (E) coordinates are as follows:

$$row = (90^\circ + N)/(2.5'/60') = (90^\circ + N) * 24 \quad (1)$$

$$column = (180^\circ + E)/(2.5'/60') = (180^\circ + E) * 24 \quad (2)$$

For example, Banjarnegara whose geographic coordinates are N = -7.3794368, E = 109.6163185 will be mapped to cell (1983, 6951).

However, a problem with the integration of multiple sources based on cell-based approach is that locations belonging to the same event may be mapped to different cells. This leads to incorrect conclusion that there are multiple events instead of a single one. Consider the following tweets that were posted in December 2014:

- "One village in central Java Banjarnegara Buried landslide - Bubblews http://t.co/iCLRVNNcpG via @GoBubblews"
- "#UPDATE: 12 dead,100 others missing in Jemblung, Indonesia after a landslide was triggered by torrential downpours http://t.co/Npweb5VveG"

The NER library extracts location entity *Banjarnegara* for the first tweet, which is mapped to cell (1983, 6951), and location entity *Jemblung* for the second tweet, which is mapped to cell (1985, 6953). Although the cells are different, but the described event is the same[6]. Jemblung is a village in Banjarnegara regency of Central Java province in Indonesia. These two places are geographically located inside one another even though they are mapped to different cells based on their geographical coordinates. That is why we propose to cluster social items based on Euclidean distance instead of solely relying on the cell-based approach to make sure we do not report the same event multiple times. This approach will map tweets that are in close proximity to one another to the same cluster. However, a large number of items from social and physical information services will slow down the execution of a clustering algorithm. For example, our evaluation dataset in December 2014 contains 42k geo-tagged social items. That is why instead of clustering individual items based on their geographic coordinates, we propose to cluster their cells. The total number of candidate cells during the evaluation period is 539, which is significantly less than the number of geo-tagged items. Cells are defined by (row, column) positions that we treat as (X, Y) coordinates for the clustering algorithm based on Euclidean distance.

[6]http://news.xinhuanet.com/english/world/2014-12/13/c_133851351.htm

S5. Event Detection

After all potential event locations are estimated using the clustering composition approach described earlier, we compute the probability of landslide occurrence in those locations based on a Bayesian model integration strategy. For detailed description of the event detection component and how the data from both physical and social sources are fed into LITMUS we refer the reader to [5], [9].

III. Implementation Details

The texts of Social Media items in the evaluation dataset of December 2014 were grouped into 500 clusters using K-means. For each cluster the top 10 terms based on their TF-IDF values have been selected. The top terms from all clusters have been added to a set and the number of distinct terms in this set is equal to 714. These terms are treated as Wikipedia concepts from this point on. Clustering of the texts in the evaluation dataset is not time-consuming as it contains only 161,260 items.

Unlike the original ESA approach, there is no step of parsing Wikipedia XML dump as the subset of Wikipedia concepts to be used is predetermined using the K-means clustering approach described above. Thus, the contents of only 714 Wikipedia articles are used to build the semantic interpreter. Building the semantic interpreter is a one-time operation that takes less than an hour, which is much faster than the 7 hours reported for the original ESA approach. After the semantic interpreter is built, the generation of interpretation vectors for textual input is several thousand words per second, which is an order of magnitude faster than the original ESA approach [2].

For evaluation of classification performance we used Weka [12], which is an open source suite of machine learning software written in Java. Weka's implementation of C4.5 algirthm is called J48.

Computations of semantic distance as well as Euclidean distance based clustering were performed using the implementation of agglomerative clustering in SciPy [13]. It clusters observation data using a given metric in the $N \times M$ data matrix, where N is a number of observations and M is a number of dimensions. The observation data for semantic clustering are interpretation vector values whereas the observation data for Euclidean clustering are (row, column) positions of the cells.

Both clustering and classification processes are fast to execute because we only consider geo-tagged items as opposed to a complete set of data. In addition, the classification process uses a pre-built model based on training data to classify incoming items and the number of features is only 714.

IV. Evaluation Using Real Data

In this section we analyze LITMUS performance using real-world data during the evaluation period. In particular,

Classifier	Precision	Recall	F-Measure	Class
Naïve Bayes	0.902	0.847	0.874	relevant
	0.281	0.395	0.328	irrelevant
	0.821	0.787	0.802	(weighted avg.)
C4.5	0.989	0.992	0.991	relevant
	0.949	0.930	0.939	irrelevant
	0.984	0.984	0.984	(weighted avg.)

Table II
OVERVIEW OF CLASSIFICATION RESULTS

	Locations based on NER	Locations based on cell-based approach	Locations based on semantic clustering	Locations based on Euclidean clustering
Locations	684	539	509	493

Table III
EVALUATION OF LOCATION ESTIMATION

we provide three sets of experiments designed to evaluate the system components described in the paper. We start with the results of classification of individual items from Social Media based on semantic relatedness to Wikipedia concepts using Naïve Bayes and C4.5 classifiers. Next we describe the preliminary results of landslide detection by LITMUS and compare them with an authoritative source. Finally, we provide location estimation results using clustering composition approach and demonstrate the improvements made in estimating the actual number of detected events unreported by the authoritative source.

The experiment on evaluation of the actual thresholds used by the clustering algorithms is omitted due to lack of space. Also, we did not include comparison of classification results based on augmented ESA approach versus original ESA, because we were unable to compute a semantic interpreter using the latest Wikipedia XML dump within a reasonable amount of time.

A. Dataset Description

We select the month of December 2014 as the evaluation period. Here is an overview of the data collected by LITMUS during this period - see Table I. Majority of items in each social source do not contain geo-location, which is why we apply the geo-tagging component.

In order to collect the ground truth dataset for the evaluation month, 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 the evaluation month represent a set of candidate events. 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 text contains a URL, then we look at the URL to confirm the candidate item's relevance to disasters. 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. A cell is thus relevant if at least one social item mapped to that cell is relevant, whereas the cell is irrelevant if all of its social items are irrelevant. It should be noted that we consider all events reported by USGS as ground truth.

B. Evaluation of Classification based on Augmented ESA

For evaluation of classification performance we used two algorithms: Naïve Bayes and C4.5.

Naïve Bayes was chosen as it is a commonly used baseline classifier algorithm. It remains popular despite its strong (naive) independence assumptions between the features. One of the main advantages of this algorithm is its high scalability and it has shown good performance in various complex real-world situations.

C4.5 is a decision tree based algorithm. We chose it as an alternative classifier algorithm, because we wanted an algorithm to reflect the process of building the ground truth dataset described earlier. In particular, during the process of manually labeling items from Social Media we noticed that we could almost instantly tell whether a given social item was relevant to landslide as a natural disaster or not. There were several common relevant and irrelevant topics discussed in Social Media that were easy to spot due to the use of specific words. Each time a particular word was used we could predict with high probability the label of the whole text. Hence, our hypothesis was that a decision tree based algorithm could predict accurate labels based on

the thresholds of the relevance of terms to the concepts represented as features.

The following table contains classifications results of the evaluation dataset using a 10-fold cross validation approach - see Table II.

C. Preliminary Landslide Detection Results

In addition to the seismic feed described above, USGS provides a variety of other data, including a continually updated list of landslide events reported by other reputable sources[7]. This list contains links to articles describing landslide events as well as the dates when they were posted. In December 2014 USGS listed 72 such links. LITMUS detected 71 out of 72 events. There was only one event reported by USGS that LITMUS did not detect, namely: "Landslides Impede the Movement of Traffic in Two Directions in Bulgaria's Smolyan Region" posted on December 4th[8]. It is a rather minor local event, which explains why it did not receive much attention in Twitter, Instagram or YouTube.

During the same evaluation period LITMUS also detected 238 landslide locations unreported by USGS. The next section evaluates location estimation based on clustering composition and computes the actual number of detected, but unreported events, which is smaller than 238.

D. Evaluation of Location Estimation based on Clustering Composition

The next table contains the results of location estimation - see Table III. The CoreNLP library detected 684 distinct locations based on Social Media texts from the evaluation dataset. Cell-based approach mapped these locations to 539 cells. Semantic clustering removed 5.5% of outlier locations and Euclidean clustering reduced the total number of locations to 493.

Based on the final set of locations generated by the clustering composition approach the actual number of the detected events that were unreported by the authoritative source is equal to 190 instead of 238. This represents a 20% improvement in location estimation due to our clustering composition approach.

V. Related Work

Recent research in Natural Language Processing has found ESA to be successful for text classification. [14] used ESA to find the semantic relatedness between German Words using German-Language Wikipedia and found it to be superior for judging semantic relatedness of words compared to a system based on the German version of WordNet (GermaNet). [15] found approaches based on ESA to perform better than those that are solely based on hyperlinks. However, the stock algorithm for ESA relies heavily on leveraging existing knowledge of Wikipedia, which is very time-consuming and parts of it may be irrelevant for our purpose. Our implementation of ESA is augmented such that instead of using all Wikipedia concepts, we use top concepts extracted implicitly from our dataset. This allows us to classify rapidly without necessarily having to make a large external repository of knowledge tractable first, while leveraging the capabilities of ESA as a superior text classifier.

Cell based approach to identify clusters in tweets and other social media items originating from a geo-location has been used in research in the past. In the context of Twitter, less than 0.42% of tweets are geo referenced [3]. Thus, [16] attempted to assign geo-coordinates to non-geo tagged tweets to increase the chance of finding localized events. Then, they searched for geo-terms and counted the number of key terms that co-occur in different tweets within a short amount of time to detect a theme. [17] used cell-based approach to cluster tweets into geographical boundaries to detect unusual geo-social events. Other works like [18] extended this idea to detecting events in real-time from a Twitter stream and to track the evolution of such events over time. However, their work does not explore the application of cell-based approach of integrating multiple sources to detect natural disasters. Also, they do not perform very well in situations when there are multiple geo-terms within the same text. Our approach for geo-tagging applies semantic clustering to remove location outliers followed by Euclidean clustering to group related incidents to make sure we do not report the same event multiple times.

Many researchers have explored the use of social media to detect events, such as natural disasters. Guy, et al. [19] introduced TED (Twitter Earthquake Detector) that examines data from social networks and delivers hazard information to the public based on the amount of interest in a particular earthquake. Sakaki, et al. [1] proposed an algorithm to monitor tweets and detect earthquake events by considering each Twitter user as a sensor. Our system LITMUS is based on a multi-service composition approach that combines data from both physical and social information services - see [9], [5] for more information. Furthermore, this work is a refinement of LITMUS that focuses on improving classification and location estimation results.

VI. Conclusion

Real-time disaster detection based on Social Media faces multiple critical challenges including filtering out noise that is present in majority of items from Social Media as well as estimating locations of the events based on the filtered items. In this paper we describe two novel techniques that we implemented to improve the quality of landslide detection service called LITMUS. To rapidly classify items from Social Media as either relevant or irrelevant to landslide as a natural disaster, we augment ESA classification by

[7]http://landslides.usgs.gov/recent/

[8]http://www.focus-fen.net/news/2014/12/04/356314/

extracting a subset of Wikipedia concepts to be used as classification features using a clustering algorithm, such as K-means. Then we estimate locations of the events described by the classified items using a composition of clustering algorithms, namely semantic clustering to remove location outliers followed by Euclidian clustering to avoid reporting separate instances of the same event multiple times. Our experiments demonstrate that this approach not only helps to remove noise rapidly, but also improves the quality of location estimation of the detected events.

During our work on this project we noticed that Social Media users often discuss past events, especially if the damage or affected areas were substantial. For example, many months after Typhoon Haiyan tore through the Philippines, users still discussed it on Social Media as it still resonated with them. As we are developing an automated notification system that people and organizations can subscribe to in order to receive real-time information on detected landslides, we want to make sure that LITMUS can distinguish previous events from the new ones. Finally, we hope that comprehensive and accurate real-time information about disaster events can be useful to various communities, including government agencies and general public, which is why LITMUS is live and openly accessible for continued evaluation and improvement of the system[9].

ACKNOWLEDGEMENTS

This research has been partially funded by National Science Foundation by CNS/SAVI (1250260, 1402266), IUCRC/FRP (1127904), CISE/CNS (1138666, 1421561) programs, and gifts, grants, or contracts from Fujitsu, HP, Intel, 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.

REFERENCES

[1] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *WWW*, 2010.

[2] E. Gabrilovich and S. Markovitch, "Wikipedia-based semantic interpretation for natural language processing," *Journal of Artificial Intelligence Research*, vol. 34, no. 2, pp. 443–498, 2009.

[3] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: a content-based approach to geo-locating twitter users," in *CIKM*, 2010.

[4] E. A. Sultanik and C. Fink, "Rapid geotagging and disambiguation of social media text via an indexed gazetteer," in *ISCRAM*, 2012.

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

[5] A. Musaev, D. Wang, and C. Pu, "LITMUS: a Multi-Service Composition System for Landslide Detection," *IEEE Transactions on Services Computing*, vol. PP, no. 99, 2014.

[6] USGS, "United States Geological Survey agency: Earthquake activity feed from the United States Geological Survey agency," http://earthquake.usgs.gov/earthquakes/, accessed on 2/1/2015.

[7] TRMM, "Tropical Rainfall Measuring Mission: Satellite monitoring of the intensity of rainfalls in the tropical and subtropical regions," http://trmm.gsfc.nasa.gov/, accessed on 2/1/2015.

[8] CHRR, et al., "Global Landslide Hazard Distribution," http://sedac.ciesin.columbia.edu/data/set/ndh-landslide-hazard-distribution/, accessed on 2/1/2015.

[9] A. Musaev, D. Wang, and C. Pu, "LITMUS: Landslide Detection by Integrating Multiple Sources," in *ISCRAM*, 2014.

[10] Google Inc., "The Google Geocoding API," https://developers.google.com/maps/documentation/geocoding/, accessed on 2/1/2015.

[11] The Stanford Natural Language Processing Group, "Stanford CoreNLP," http://nlp.stanford.edu/software/corenlp.shtml, accessed on 2/1/2015.

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

[13] E. Jones, T. Oliphant, and P. Peterson, "SciPy: Open source scientific tools for Python," 2001.

[14] I. Gurevych, C. Mller, and T. Zesch, "What to be? - electronic career guidance based on semantic relatedness," *Annual Meeting-Association for Computational Linguistics*, vol. 45, no. 1, 2007.

[15] I. H. Witten and D. Milne, "An effective, low-cost measure of semantic relatedness obtained from wikipedia links," in *AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy*, 2008.

[16] K. Watanabe, M. Ochi, M. Okabe, and R. Onai, "Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs," in *20th ACM international conference on Information and knowledge management*, 2011.

[17] R. Lee and K. Sumiya, "Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection," in *2nd ACM SIGSPATIAL international workshop on location based social networks*, 2010.

[18] H. Abdelhaq, C. Sengstock, and M. Gertz, "Eventweet: Online localized event detection from twitter," *Proc. VLDB Endow.*, vol. 6, no. 12, pp. 1326–1329, Aug. 2013. [Online]. Available: http://dx.doi.org/10.14778/2536274.2536307

[19] M. Guy, P. Earle, C. Ostrum, K. Gruchalla, and S. Horvath, "Integration and dissemination of citizen reported and seismically derived earthquake information via social network technologies," vol. 6065, 2010.