A Lightweight and Real-Time Worldwide Earthquake Detection and Monitoring System Based on Citizen Sensors[*]

Jazmine Maldonado[†]
INRIA Chile
Department of Computer Science
University of Chile
Santiago, Chile
jazmine.maldonado@inria.cl

Jheser Guzmán
Department of Computer Science
University of Chile
Santiago, Chile
jguzman@dcc.uchile.cl

Barbara Poblete
Department of Computer Science
University of Chile
Santiago, Chile
bpoblete@dcc.uchile.cl

Abstract

We propose an algorithm and system that detects earthquakes worldwide in real time based on reports of social media users, or "citizen-sensors". Earthquake detections are based on user postings in any language and from any region. This approach is unsupervised, adapting automatically to changes in the input data stream, and only requires a general list of keywords for each language. Our method is noise tolerant and simple, providing good results both in terms of precision and recall. This complements prior work that mostly consists of supervised approaches that focus on performing detections in a specific geographical area and are difficult to generalize to a global scope. We demonstrate the effectiveness of this approach by using it within a real-time on-line system, which is publicly available and currently in use at National Seismology Center in Chile and Oceanographic and Hydrological Office of the Chilean Army. The quantitative evaluation of our system, performed during a 9-month period, shows that our solution is competitive to the best state-of-the-art methods. Overall, our findings indicate that our approach is an effective low-cost alternative for earthquake monitoring at a global scale.

Introduction

In highly seismic countries, the study of earthquakes is a critical task. Countries such as Japan, China, the U.S. and Chile, just to name a few, devote significant resources to deploy dense earthquake sensor networks (i.e., seismographic networks) to detect and describe these events. Seismographic networks are used to estimate the way in which an event is experienced on different parts of the earth's surface. Nevertheless, the quality of this estimation can vary depending on several factors, including the depth at which the earthquake occurred and local terrain conditions. Seismographs have high-sensitivity that allows them to detect even microearthquakes (i.e., very low-magnitude earthquakes), which are not commonly felt by people. However, seismographs are expensive to deploy and maintain and not all geographical regions are well covered by a network of these sensors. Due to this, in many cases, human reports and crowdsourcing initiatives are used to determine if an earthquake was a *"felt earthquake"* (i.e., an earthquake perceived by the population) (Atkinson and Wald 2007). This allows to make better estimations of the actual strength with which the earthquake was perceived, the damage that it caused, and the geographical areas where it hit. These reports allow seismologists to estimate accurately the size of an earthquake, which in turn allows establish its overall strength.

The description and study of all earthquakes, *large and small*, allows experts to gain insights of current seismic activity in certain regions and produce complete earthquake catalogs (Stein and Wysession 2009). In addition, emergency response agencies worldwide also strive to determine people's perception of the shaking produced by a seismic event in all of the geographical areas where the event was felt. Even if the earthquake was small, knowing how it was perceived in an area allows governments to design disaster response policies and estimate the damage that a future high magnitude earthquake might produce. However, certain conditions such as lack of seismographic network coverage or sparce population, can difficult obtaining complete reports of earthquake impact (USGS 2017d; SHOA 2017; ONEMI 2017; Wyss and Zibzibadze 2009).

In their influential work, Sakaki et al. [2010] showed that social network users, or so-called "citizen sensors" (Sheth 2009), can be used as a low-cost means for quick earthquake detection and epicenter location. Since then, several studies have addressed the problem of detecting and describing earthquakes using microblog data (Avvenuti et al. 2014b; Cameron et al. 2012; Earle et al. 2010; Earle, Bowden, and Guy 2012). In particular, Young et al. [2013] have made a case for the need lo leverage citizen sensor data into seismological research, and how it can provide inexpensive high-quality information for areas which are not covered well by existing data collection methods. This research has focused on Twitter[1], a microblogging and social networking site, which is characterized by short-text 140-character messages (called *tweets*), and is currently used worldwide by

[*] This research has been partially funded by: the Nucleus Center for Semantic Web Research NC120004, CONICYT 2010 Doctoral Scholarship (J.G.) and 2016 Masters Scholarship (J.M.), CSN UChile grants 9927/2016 and 11482/2016 (J.M., J.G.), Inria Chile (J.M.) and FONDEF ID16I10222 (B.P.).

[†] This work was done while J.M. was a graduate student at University of Chile.

[1] `https://www.twitter.com`

over 300 million users (Twitter 2016). About 80% of Twitter users access the service using mobile devices, which contributes to the immediacy of its diffusion of information, becoming a preferred news source for journalists and general users, particularly during crisis situations and natural disasters (Castillo 2016; Mendoza, Poblete, and Castillo 2010).

Despite the usefulness of social media data for earthquake detection and description, our state-of-the-art review shows that the problem is not closed in terms of precision, recall and geographical coverage (Young et al. 2013). We observe that existing approaches have high-precision for some large high-magnitude occurrences, which are perceived by many people, but low recall when considering the complete range of "felt earthquakes". This occurs in some cases because of the noise found in Twitter data (i.e., irrelevant messages), which requires systems to trigger alarms only for detections that are performed with very high-confidence in order to avoid false positives (Earle et al. 2010; Earle, Bowden, and Guy 2012; Avvenuti et al. 2014b). Certain systems manage to improve the recall of high-magnitude events, but at the expense of narrowing their scope to country level using supervised approaches and strict ad hoc filters on the input data (Avvenuti et al. 2014b; Sakaki, Okazaki, and Matsuo 2013). However, these approaches are extremely difficult to scale and replicate to other countries, due to the high-cost of tuning and labeling that they require.

As a consequence, (to the best of our knowledge) there are no systems that can provide a good trade-off between precision/recall for multiple regions and languages. Nor there no are publicly available systems for real-time worldwide earthquake detection (and description) using citizen sensors.

Our contribution. We created a methodology for real-time earthquake detection based on microblog messages. Our method is noise tolerant, unsupervised, easy to parametrize and robust in terms of providing good precision and recall for high-magnitude and low-magnitude earthquakes that were perceived by citizen sensors. The approach is simple, requiring little computational resources, which allows us to perform multilingual worldwide detection. In addition, we implemented a visual web-based system for our approach, called Twicalli[2], which is available to the public and is currently used as a decision support tool at the National Seismology Center in Chile and the Oceanographic and Hydrological Office of the Chilean Army. Our method bases its analysis on tracking signals created from aggregated messages related to earthquakes, including arrival rates of location mentions in tweets. This allows the user of the system, for example, to disambiguate different events that occur simultaneously or within a short time interval of each other. This type of disambiguation is not straight forward using seismographs (Kennett and Engdahl 1991; Sambridge and Kennett 2001).

In this article we describe our methodology for earthquake detection and description, as well as our web-based system. We present a quantitative evaluation of our detection algorithm over a 9-month period using several ground truth criteria. We discuss the performance of our method in relation to that reported in the literature. This is the most long-term evaluation to date for this type of system. Our results show that our system is very competitive, achieving 0.99 precision and 0.85 recall (F-measure of 0.91) for earthquakes with ≥ 4.0 magnitude that were felt by people. For ≤ 4.0 earthquakes, most of which were not felt by people (and therefore cannot be detected using citizen sensors), we achieve 1.00 precision and 0.15 recall (F-measure of 0.26). This result outperforms significantly most of the results reported by other supervised and unsupervised systems in similar studies. In the only case that our system comes close to, but does not improve the performance of a competing system, our method provides the trade-off that it is detecting earthquakes worldwide in an unsupervised manner (as opposed to being supervised and local to one country).

Overall, we provide a simple and scalable tool for earthquake detection in real-time, which can even be run on a personal computer. The proposed approach solves existing issues that limited the possibility of using Twitter for worldwide detection.

Reproducibility. Our system is publicly available online, as well as our method's source code and ground truth[3].

Related Work

There are 2 main types of approaches for earthquake detection based on social media: *probabilistic temporal models* and *Short Term Average vs Long Term Average Algorithm (STA/LTA)*. All systems are based on retrieving public Twitter messages that are likely to be reporting real-time earthquake occurrences.

Probabilistic Temporal Models

Probabilistic temporal models are used in the work of Sakaki et al. [2013; 2010] (who extend the work of Okazaki et al. [2010]), and in the work of CSIRO Australia researchers (Yin et al. 2012; Robinson, Power, and Cameron 2013).

Sakaki et al. [2013] introduce a detailed temporal model based on an exponential distribution to identify real-time earthquake occurrences and a geo-spatial model to detect the epicenter of an event. Their system, which is specific to Japan, filters messages which are likely to refer to an earthquake, which will be then classified using an SVN classifier into relevant or non-relevant. They validate their system quantitatively using a set of official earthquake reports from Japan's Meteorological Agency (JMA). Their proposed system has a strong trade-off between precision and recall, given by the threshold used as the number of relevant messages needed for detection.

Researchers at CSIRO Australia (Yin et al. 2012) propose a temporal model, based on a binomial distribution of message arrival rates, to detect disasters in Australia and New Zealand. ESA (Robinson, Power, and Cameron 2013), their proof of concept system available online (ESA 2012), is designed as an emergency situation support system and shows

[2] term coined by M. Strohmaier (Strohmaier 2010). System available at `http://twicalli.cl`

[3] Available at: `https://github.com/dicotips/BurstDetector`

geographical information of messages that contain disaster information.

Short Term Average vs Long Term Average Algorithm (STA/LTA)

The STA/LTA algorithm is commonly used in seismology to detect and time seismic phases (Stein and Wysession 2009). This algorithm is used by Earle et al. [2010; 2012] to track arrival rates of messages that contain keywords associated with earthquake occurrences in real-time. This approach is targeted towards worldwide earthquake detection for which they use a list of keywords that refer to the term "earthquake" and its variations in different Western languages. To reduce false alarms induced by noisy messages, that contain the specified keywords but within a different context, they assume a compromise between the number of detections and detecting only high-scale events.

Avvenuti et al. [2014a; 2014b] also use this approach for their earthquake detection system specifically designed for Italy. Their system, EARS, is meant to detect earthquakes and improve crisis response in that country. EARS collects messages extracting geographical information based on the corpus' text to provide information related to affected locations. Similarly to the work of Sakaki et al. [2013], an important part of the pipeline of EARS is that of applying strict filters to keep only messages that are very likely to refer to a real-time earthquake to minimize noise and prevent false positives. This filtering is based on several features and classifiers, which are trained to detect relevant messages in Italian. This strict filter prevents the use of words commonly used to report earthquakes, such as "tremando" (shaking), because they are also frequently used in contexts other than earthquake reports. This approach has a trade-off between precision and recall, because events that are not perceived as a large earthquake will likely not meet the threshold to trigger an alert.

Scope of the State-of-the-Art Solutions

In Table 1 we summarize relevant characteristics of prior work and of our own. Existing approaches have an important supervised component and are customized for a particular geographical region, except for the work of Earle et al.

Systems based on probabilistic models require an initial training period to estimate the probability distribution of the input data, which is done offline. These systems must be retrained periodically in order to adapt them to the dynamic changes that occur in Twitter's input stream (e.g., volume variations, seasonal trends, etc.)

Prior work relies on strict ad hoc filters on the input data to reduce the amount of noisy messages. The authors of these works indicate clearly that the data cleaning process and fine selection high-confidence keywords is fundamental to the performance of their methods. This also removes messages that can potentially be useful for earthquake detection, decreasing the recall of total earthquakes. In addition, supervised approaches provide classifiers that have been customized for a particular region allowing them to determine with high-precision messages that correspond to real-time earthquakes in a particular language. These models are trained on manually labeled data for earthquakes in a particular region/language.

It can be very difficult to scale supervised approaches to other countries due to the high cost involved in the creation of ad hoc manual filters for each region, labeling of messages and fine tuning of classifiers. Existing unsupervised approaches, on the other hand, do not apply such strict filters but require a high activation threshold to detect events (i.e., a significant portion of users reporting a strong earthquake) in order to reduce false detections. Furthermore, the introduction of a larger set of keywords in several languages increases the noise in the input stream for these methods, degrading performance (Avvenuti et al. 2014b; Earle, Bowden, and Guy 2012).

Our approach complements existing work by being unsupervised, with low parametrization cost and more tolerant to noise. It does not require to apply strict filters to the input data stream.

Earthquake Detection Methodology

The goal of our work is to provide a methodology that does not rely on supervised classification of messages, and that adapts automatically to changes in the input data stream.

Our algorithm is inspired in an existing unsupervised approach for *burst detection* in text streams (Guzman and Poblete 2013). Although, there are several emerging event detection methods for streaming microblog data, we select this one due to its linear complexity (on the number of input messages) and because it only requires an simple initial parameter setup. Other techniques for event detection can require periodical fine-tuning of parameters, or periodical training (Mathioudakis and Koudas 2010; Sankaranarayanan et al. 2009), or have higher computational complexity, such as (Zhou, Chen, and He 2015; Zhao et al. 2014), which are polynomial and quadratic respectively.

Our method aggregates into a unique discrete-time signal all messages that match a broad set of keywords related to earthquake reports. This approach is noise tolerant (i.e., tolerant to messages in the input that are not related to earthquakes), allowing us to include keywords that may introduce noisy messages into the input data. This differs from prior work, which requires as input only messages for which there is high-confidence that they are reporting a real-time earthquake. In this sense, prior work restricts the use of keywords that may be used in other contexts, in order to avoid false detections.

Our method relies on the collective organization phenomenon that arises in "citizen sensors", when exposed to a stimulus produced by a real-time earthquake, which causes the complete input signal to increase in frequency significantly. This increase will be much larger than that normally observed due to noise in the signal.

Relative Arrival Rate Monitoring

The core of our method detects changes in the *relative arrival rate* of relevant messages. We now define this quantity. Consider a data stream of time-indexed messages $\mathcal{S} =$

Work	Techniques	Coverage	Keywords	Active
Ours	Unsupervised	World-Wide	earthquake, sismo, quake, temblor, temblando, gempa, lindol, tremblement, erdbeben, deprem, seisms, sisme, zelzele, terremoto, scossa and translations of earthquake in Japanese, Chinese, Greek and Persian	Yes
EARS	Supervised	Italy	scossa, terremoto	Yes
Sakaki et. al.	Supervised	Japan	earthquake, shaking	Yes
Earle et al.	Unsupervised	World-Wide	earthquake, gempa, sismo, temblor, terremoto	No
ESA	Supervised	Australia & New Zealand	earthquak, #eznq	Yes

Table 1: Summary of the scope of the state-of-the-art and our solution.

$\langle S_1, S_2, \ldots, S_n \rangle$ where $r : \mathcal{S} \rightarrow \mathbb{R}^+$ describes the arrival time of messages, and $S_i \subseteq T$ corresponds to the attributes of message S_i, with T corresponding to a set of possible attributes in the messages, such as words, locations, hashtags, inferred sentiments, and/or other elements. Some of the possible attributes in T constitute *elements of interest*, which are application-dependent attributes (in our case, attributes of messages that might be related to earthquakes), we denote these attributes as $K (\subseteq T)$.

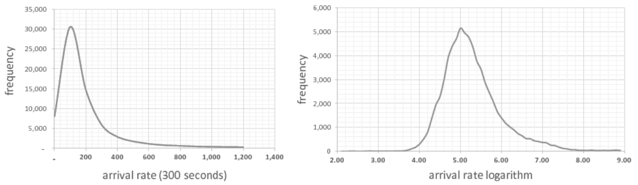
Let $w_i = \langle w_i^s, w_i^e \rangle$ denote a time window spanning from time w_i^s to time w_i^e. Let $M_i = \{s \in \mathcal{S} : w_i^s \leq r(s) \leq w_i^e\}$ denote the messages of $\mathcal{S}$ inside that time window. Within these messages, we count how many of them contain elements of interest, denoting this quantity by $\mathrm{freq}(K, M_i) = |\{S \in M_i : K \cap S \neq \emptyset\}|$.

The data stream is processed in batches, splitting arriving messages in consecutive time-windows of fixed length t. For a given time window w_i containing messages M_i we define the *relative arrival rate* (λ) of elements of interest in this window as:

$$\lambda(K, M_i) = \frac{\mathrm{freq}(K, M_i)}{t|M_i|} \tag{1}$$

We track a generic input discrete-time signal ($\mathcal{S}$) over time, in order to determine when a positive variation of the signal's relative arrival rate (λ) is significantly larger than variations due to noise observed in the past. The signal's "burstiness" is estimated based on the magnitude of the change experienced by its relative arrival rate (λ) with respect to the previous time-window. The next step is to automatically determine if a significant positive change in the signal has occurred within the current time-window.

A common approach used in the literature to determine if the relative arrival rate (λ) experiments a significant variation during time-window w_i is to track the standard deviation of λ for all time-windows up until w_{i-1} (Kleinberg 2003; Mathioudakis and Koudas 2010; Nguyen et al. 2013). This idea assumes (and requires) that the arrival rate (λ) has an exponential distribution. However, our empirical analysis, shown in Figure 1(a), of the dataset described in Section "Experimental Analysis", indicates that this is not the case of earthquake-related message arrival-rate distribution.



(a) Distribution of the frequency of messages.

(b) Distribution of the logarithm of the frequency of messages.

Figure 1: (a) Original (log-normal distribution) and (b) transformed (normal distribution) data.

The data, shown in Figure 1(a), actually resembles a log-normal distribution. Therefore, we apply a log transformation after which the data resembles a normal distribution $\log(\lambda) \sim \mathcal{N}(\mu, \sigma^2)$, shown in Figure 1(b). Therefore, instead of tracking changes in the relative arrival rate (λ), we model the distribution as if it were a log-normal and track the *logarithm of this function* as a normal distribution, defined as $\tilde{\lambda}$:

$$\tilde{\lambda}(K, M_i) = \frac{\ln(\mathrm{freq}(K, M_i))}{t|M_i|} \tag{2}$$

Using the transformed function $\tilde{\lambda}(K, M_i)$ we compute its z-score to track variations in the input signal's relative arrival rate. We compute the z-score at time-window w_i as

$$z\text{-score}(K, M_i) = \frac{\tilde{\lambda}(K, M_i) - \mu_i}{\sigma_i} \tag{3}$$

where μ_i and σ_i are, respectively, the mean and standard deviation of the observed values of $\tilde{\lambda}$ during time windows $w_1, w_2, \ldots, w_{i-1}$.

The proposed method triggers an alert warning that an earthquake has been detected when $z\text{-score}(K, M_i) \geq \theta$, where θ is an experimentally defined threshold. In the following Section we describe our system and the methodology used to determine the initial parameters and threshold setup.

Experimental Setup

We use our detection algorithm to create a real-time earthquake detection system with a visual web-based interface. In this section we describe the experimental setup used to evaluate our system including how to tune its input parameters.

Twitter Dataset

Our input data stream $\mathcal{S}$, consists of tweets obtained from the public Twitter endpoint from January 25th to October 25th, 2016 (9 months). Technically, since we only have access the sample of the complete data stream provided publicly by Twitter, we retrieve the messages already filtered by the keywords, or attributes of interest, specified in K (instead of retrieving the full stream and filtering it afterwards). Therefore, we use the Stream API to retrieve directly tweets that match a logical "OR" of keyword-based predicates. Tweets are then preprocessed in a standard fashion, normalizing text and removing duplicated messages from the same user. Text processing is done for multiple languages, including non-western languages, such as Chinese, Japanese and Arabic. In total, our dataset consists of $53,557,475$ tweets.

Tweets containing geographical coordinates from mobile devices are represent only 8% of our dataset. We enhance geo-location information by adding inferred location information, similar to that used by (Robinson, Power, and Cameron 2013). Specifically, we use a simple gazetteer for to extract locations from message text and user profiles. Using this procedure we are able to geo-locate 53.6% of the collection ($28,723,948$ tweets).

Ground Truth Datasets

We use two publicly available earthquake catalogs as ground truths, one corresponding to a global sensor network, and another to a local sensor network. We use earthquake detections for the same 9-month period as our Twitter data (January 25th to October 25th, 2016). Both catalogs constitute official reports and describe the estimated location of the earthquake's epicenter and magnitude (USGS 2017b).

Global catalog (USGS): is obtained from the United States Geological Survey (USGS 2017a). This catalog is very complete for earthquakes worldwide over 4.5 magnitude (Earle et al. 2010). However, for lower magnitudes, they focus mostly on certain regions in the U.S. For the time period of our study the USGS reports $9,470$ earthquakes with magnitude ≥ 4.0 from all over the world. We use this dataset to evaluate and compare performance with systems that have worldwide coverage.

Local catalog (GUC): is focalized in one country and is obtained from the National Seismology Agency in Chile (GUC 2017). This catalog is based on GUC's dense local sensor network and is very complete for earthquakes in Chile. In addition to epicenter and magnitude, this catalog also contains earthquake *intensity*, indicating whether an earthquake was felt by people and how much damage it produced (in the modified Mercalli intensity scale (USGS 2017c)). For the time period of our study the GUC reported 662 earthquake events of magnitude $\geq$ 4.0, 476 of them classified as "felt earthquakes". We use this dataset to evaluate and compare performance with systems that have local scope.

z-score (θ)	Precision	Recall	F-Measure
0.5	48.1%	79.9%	60.1%
1.0	62.6%	65.0%	63.8%
1.5	**88.3%**	**54.2%**	**67.1%**
2.0	92.3%	29.0%	44.2%

Table 2: Algorithm performance using different values for the event alarm threshold θ.

We note that during period of our study, the GUC reported $1,373$ earthquakes (≥ 3.6), but the USGS reported only 436 earthquakes in Chile with the same magnitude (only 32% coverage). Supporting the need to have complementary information sources for complete coverage of earthquakes in areas that are not well covered by global or local seismographic networks.

System Parameter Setup

The proposed methodology requires three input parameters, described in Section "Earthquake Detection Methodology": (1) the *threshold for triggering earthquake alarms* θ, (2) the list of *earthquake-related attributes* (K), such as keywords, and (3) the *time-window size* (t), described next:

(1) Earthquake detection threshold (θ): In order to determine if there has been a significant variation in signal $\mathcal{S}$ we empirically search for the optimal z-score value threshold, θ Using a 2-month data sample we test different values of θ empirically against actual earthquake detections and select the value that optimizes the F-measure. Table 2 shows the different values obtained, from which we select $\theta = 1.5$.

(2) Keywords related to earthquakes (K): In our system the so-called elements of interest, defined in the previous Section, correspond to a list of keywords related to earthquake occurrences. We extend the initial set of keywords provided by researchers at USGS, used in Earle et al. [2012] and Avvenuti et al. [2014a; 2014b]. However, unlike prior systems, our list of keywords is meant to include as many words as possible that are related to earthquake occurrences in every language, even if this includes ambiguous terms (i.e., that might not refer always to real-time earthquakes). Noisy messages that can be retrieved using ambiguous terms do not affect the performance of our method, which is noise tolerant. For example, a keyword such as "quake" can be used to refer to a video game or comic character of the same name; and the word "shaking" (temblando) is commonly used in Spanish and Italian within different context such as: shaking because one is cold, or excited. The keyword list consists of the terms in figure 2.

If new terms need to be added, the list of elements K can be updated dynamically.

earthquake, sismo, quake, temblor, temblando,
gempa, lindol, tremblement, erdbeben, deprem,
σεισμός, seismós, séisme, زلزله, زمین‌لرزه, zelzele,
terremoto, scossa, 地震, 海啸, 津波, 地震

Figure 2: Selected keywords used to retrieve tweets related with earthquakes in the world.

(3) Time-window size (t): The window size corresponds to the time length (in seconds) of the batches in which our algorithm will process the input data (i.e., the time-step used to calculate arrival rates). This parameter is determined by analyzing the input signal and selecting the smallest value of t that minimizes the signal's relative standard deviation (i.e., where its frequency becomes more stable). Intuitively, if the time-window is too small, then messages containing elements in K will be distributed more randomly in each window. This is, we will have windows with no occurrences and others with or more. The larger the window the more likely that we can estimate the frequency in the following window. To determine t we use the following procedure, described in (Guzman and Poblete 2013):

For the given input signal we compute the relative arrival rate for successive time windows w_i of size t. We repeat this process for different sizes of t, ranging from 0 to 3 600 seconds (one hour) for a 2-month period and select the t with the smallest relative standard deviation (in this case $\geq$ 300 seconds), shown in Figure 3. We note that in practice, our system runs two identical parallel processes 150-seconds apart, in order to decrease detection time in half.

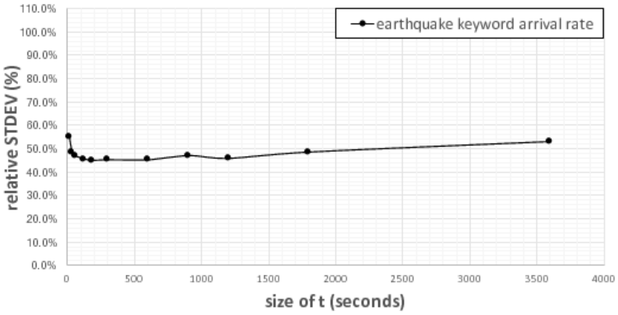


Figure 3: Relative standard deviation of the earthquake-keyword signal in relation to different window-sizes (in seconds).

Evaluation

In this section we present an evaluation of our method for earthquake detection. We evaluate our method using 3 types of variations of the input signal $\mathcal{S}$ created from messages that contain the earthquake related keywords K. These 3 types of signals are:

1. Earthquake keywords: This signal represents the relative arrival rate of messages that contain any element in K, where K corresponds to a list of earthquake-related keywords.
2. Geolocation from text: Tweets selected for the *earthquake keywords* signal are each labeled with the country names mentioned in the corpus of the tweet. These tweets are then separated into several streams, one stream for each country. Then we monitor each country as a separate signal.
3. Geologation from user: Created in a similar manner to the signal *geolocation from text* signals, with the difference that location labels are extracted from the user profile.

Evaluation Methodology

We create an evaluation that attempts to meet the different criteria used in the prior work to which we compare our system (Sakaki, Okazaki, and Matsuo 2010; Avvenuti et al. 2014b; Robinson, Power, and Cameron 2013; Earle, Bowden, and Guy 2012). We replicate their experiments using our previously described Twitter dataset and ground truths, which are equivalent to those used previous evaluations. However, we were not able to reproduce other systems in order to compare them on the exact same dataset. This is because we do not have access to the source codes, models and implementation details of those systems. As well as due to the high costs associated to supervised approaches (e.g., data labeling and parametrization), described in the Section "Related Work". Therefore, we report our results and then compare them in the closest possible manner to the results reported by the other systems.

However, there are certain considerations regarding the use of catalogs ground truth data for social-based systems:

Not all earthquakes in a catalog are felt by people. Therefore, any system relying on so-called "citizen sensors" *can only* potentially detect the subset of earthquakes that were actually perceived by humans. Ideally, a perfect ground truth for social-based systems should only consider earthquakes that were perceived by people.

Not all earthquakes felt by people that are in the catalog are officially labeled as "felt earthquakes". This discrepancy occurs when there is human oversight from the people in charge of officially reporting earthquake intensity (mostly for lower intensity earthquakes). Ideally, a perfect ground truth would have all perceived earthquakes labeled as "felt earthquakes".

Earthquake detections performed by seismological agencies, under 4.0 *magnitude, are not always complete (they do not have perfect recall).* Earthquake reports are based on detections from multiple seismographs in the network. However, if an earthquake occurs in an area without seismographs or with low density of these sensors, then it will not be reported. This refers to catalog completeness and is also discussed in (Earle et al. 2010)

Magnitude	Strict (GUC)	Strict (USGS)	Super Strict	Moderate
≥4.0	662	9470	201	419
<4.0	3834	5761	66	131

Table 3: Number of earthquakes belonging to each set used in the evaluation.

In order to deal with these limitations, and in an attempt to provide comparability to prior work, we propose three different evaluation criteria (summarized in Table 3):

Strict This evaluation considers as a ground truth earthquakes in a global (or local) earthquake catalog. Independently on whether the earthquake was labeled as a "felt earthquake" or not (similar criteria has been used in (Earle et al. 2010; Avvenuti et al. 2014a; 2014b)).

Super-strict This evaluation considers as a ground truth only the earthquakes in a global (or local) earthquake catalog that have been reported as "felt earthquakes" (similar criteria has been used in (Sakaki, Okazaki, and Matsuo 2013; Avvenuti et al. 2014a; 2014b)).

Moderate This evaluation considers as a ground truth "felt earthquakes" (same as the super-strict criteria) and in addition adds to this set any earthquake detected by the social-based system if, and only if it matches exactly a non-felt earthquake in the catalog. In this criteria it is more flexible; we assume that if users on Twitter talk about an earthquake, and at the same time seismographs also detected an earthquake, then it can be counted as an actual earthquake.

Results

	Signal Type (USGS)	P	R	F-M
≥ 4.0	Earthquake keywords	**0.95**	**0.75**	**0.84**
	Geolocation from text	0.94	0.79	0.86
	Geolocation from user	0.95	0.81	0.88
< 4.0	Earthquake keywords	0.58	0.18	0.27
	Geolocation from text	0.53	0.75	0.62
	Geolocation from User	0.54	0.79	0.64

Table 4: Global strict evaluation with the USGC catalog.

	Signal Type (GUC)	P	R	F-M
≥ 4.0	Earthquake keywords	**1.00**	**0.59**	**0.74**
	Geolocation from text	0.96	0.41	0.57
< 4.0	Earthquake keywords	0.95	0.02	0.04
	Geolocation from text	0.84	0.12	0.21

Table 5: Local strict evaluation with the GUC catalog.

	Signal Type (GUC)	P	R	F-M
≥ 4.0	Earthquake keywords	**0.99**	**0.85**	**0.91**
	Geolocation from text	0.98	0.64	0.78
< 4.0	Earthquake keywords	1.00	0.15	0.26
	Geolocation from text	0.94	0.23	0.37

Table 6: Super-strict evaluation using GUC "felt earthquakes".

Signal Type (GUC)	P	R	F-M
Earthquake keywords	**1.00**	**0.93**	**0.96**
Geolocation from text	0.96	0.58	0.72

Table 7: Moderate evaluation using GUC "felt earthquakes" + earthquakes in the catalog that match a detection by the system.

Tables 4, 5, 6 and 7 show our detailed results using the *strict*, *super-strict* and *moderate* ground truth selection criteria, respectively. We divide our results by scope, *local* or *global*, and by magnitude as done in other valuations, ≥ 4.0 and < 4.0. We note that the *super-strict* and *moderate* evaluations are only available for the GUC catalog, because "felt earthquakes" labels are only available for that catalog.

Overall, the best performance of our system is achieved by the *earthquake keywords* signal. According to the *moderate* evaluation, which we consider the most accurate in terms of ground truth, our system has $p = 1.00$ (precision), $r = 0.93$ (recall), and *f-m*$= 0.96$, reported in Table 7. In addition, we can observe that the signals that only consider location mentions in tweets (*geolocation from text*) and in user profiles (*geolocation from user*) also provide information for detecting events, shown in Tables 4, 5, 6 and 7. In practice these 3 signals are used together for detecting and locating an event. First detection is performed using the *earthquake keywords* signal and then the location is identified from the most "bursty" country in the *geolocation from text* signals.

Global scope comparison. First, we compare our method to the **unsupervised global scope** approach of Earle et al. [2010; 2012]. In their evaluation Earle et al. uses a *global strict* criteria using the USGS catalog, with $p = 0.94$, $r = 0.01$ and *f-m*$= 0.02$. Using the same criteria on the USGS catalog, our system improves that performance with $p = 0.95$, $r = 0.75$ and *f-m*$= 0.84$, for magnitude ≥ 4.0. For magnitude < 4.0 our system has $p = 0.54$, r=0.79 and *f-m*$= 0.64$. These results are shown in Table 4.

Local scope comparison. Secondly, we compare our method to the **supervised local scope** approaches EARS (Avvenuti et al. 2014a; 2014b), Sakaki et al. [2013; 2010] and ESA, by CSIRO Australia researchers (Yin et al. 2012; Robinson, Power, and Cameron 2013). Those results are summarized in Table 8. These methods have supervised components and ad hoc filters that are highly customized for a particular region, while our approach does not. In the evaluations reported for these methods all of them validate their

Approach	Precision	Recall	F-Measure	Catalog	Earthquakes
Ours (≥ 4.0)	0.99	0.85	0.91	GUC	201
Ours (< 4.0)	1.00	0.15	0.26	GUC	66
EARS (≥ 4.0)	1.00	1.00	1.00	INGV	7
EARS (< 4.0)	0.28	0.09	0.14	INGV	397
Sakaki et al.	0.75	0.80	0.77	JMA	1,136
Earle et al.	0.94	0.01	0.02	USGS	5175
ESA[4]	0.85	≈ 0.2	≈ 0.32	GeoNet, GA	≈ 98

Table 8: Comparison of other works related with seismic detection systems. The earthquakes columns are the number of events used in each evaluation.

work against a local-scope catalogs based on dense local seismographic networks (INGV, JMA and GeoNet GA, respectively), which are likely to report more earthquakes than the global USGS catalog. This is equivalent to our evaluation with the GUC catalog. The evaluation for which EARS reports its best results (for $\geq$ 4.0, $p = 1.00$, $r = 1.00$, and for $<$ 4.0, $p = 0.28$, $r = 0.09$ *f-m*= 0.14), considers only earthquakes in the local catalog that produced at least one tweet in their dataset. This criteria lies somewhere in between *moderate* and *super-strict*. We note that their evaluation for magnitude ≥ 4.0 was performed with only 7 earthquakes. Using those same criteria our system reports a similar performance for ≥ 4.0 ($p = [1.00, 0.99]$, $r = [0.93, 0.85]$ and *f-m*= $[0.96, 0.91]$) considering that our system was evaluated with 201 earthquakes. For earthquakes < 4.0 our system improves performance ($p = 0.94$, $r = 0.23$, *f-m* $= 0.37$). The evaluation of Sakaki et al. reports results for felt earthquakes (in JMA intensity scale ≥ 3) and are therefore comparable to our *moderate* criteria for earthquakes ≥ 4.0. In this case our system outperforms considerably their results, as well as those of the ESA system.

Additional results for comparison. The aforementioned results for local scope systems were obtained using different catalogs and attempted to simulate similar conditions of prior experiments. Therefore, as an additional effort to reproduce prior experiments, we also report the performance of our system according to the USGS catalog for Italy ($p = 1.00$, $r = 0.90$, *f-m*= 0.95), Japan ($p = 0.98$, $r = 0.77$, *f-m*= 0.86), Australia ($p = 0.91$, $r = 0.77$, *f-m*= 0.83) and New Zealand ($p = 0.96$, $r = 0.80$, *f-m*= 0.87).

False positives. These cases were uncommon, one example was a concert of a Korean pop band, in which people tweeted that they were ”shaking with excitement” (*temblando* de emoción), and another, a massive earthquake drill. Both cases created collective bursts of the target keywords from a specific location.

Visualization

Our system, called Twicalli, is designed to provide information about “felt earthquakes” to seismologists, emergency offices, and general audiences[5]. The Web application displays information in real time. It shows the frequency of messages containing earthquake-related keywords per minute during the last 24 hours. This view also presents geographical information for the messages posted during the last five minutes. In real-time mode, this system is useful for observing where the earthquake begins, how the seismic waves propagate, and the manner in which notifications about the earthquake are disseminated by social media users.

The Web application can also be used to explore past data. In exploration mode, users can observe in detail messages published during past earthquakes detected by the system for a selected time period.

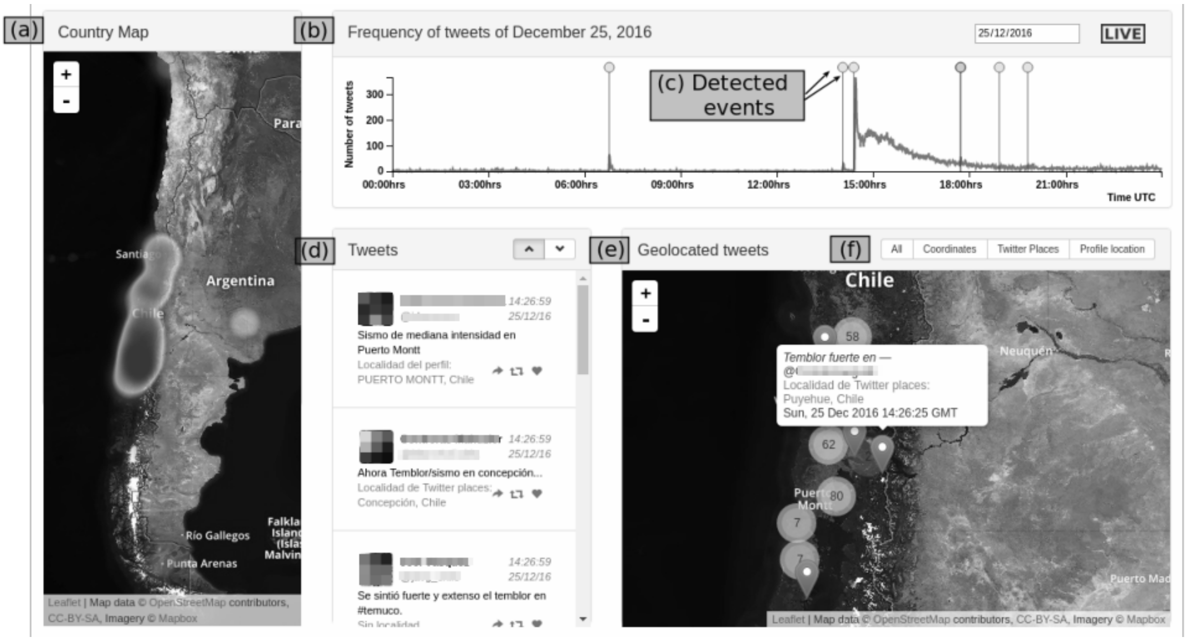
Figure 4 shows a portion of the visual interface of the web-based system, this view is updated in real-time. Clicking on any of of the visual elements displays information related to each event. Users can select a particular time period by dragging the mouse over the frequency time-series displayed in this panel, which will automatically update all of the other views in order to show only the information corresponding to the selected period. In addition, buttons (f) allow the interface users to filter location markers according to a location's source, this is useful because some location sources may be more accurate than others. When an event occurs the heatmap (d) will automatically zoom-in the area from which most of the initial messages emerged.

Figure 4, shows a major 7.6-magnitude earthquake occurred in Chile. The time-series displayed, shows a large burst in earthquake-related message frequency, which corresponds to the moment in which that event occurs. In Figure 4 we select the first two minutes after the 7.6-magnitude earthquake strikes. Maps (d) and (e) show message concentration during the first two minutes of the event.

Conclusions

We have presented a simple and efficient approach for earthquake detection based on citizen sensors. Our approach differs from existing work in that it is unsupervised and tolerant to noisy messages, allowing to monitor events worldwide

[5] The online version of Twicalli displays a specialized view of Chile, however, the backend allows for worldwide detection. We expect to have an interface with a world view in the near future.



[redacted] 14:26:59 25/12/16
Sismo de mediana intensidad en Puerto Montt
Localidad del perfil:
PUERTO MONTT, Chile

[redacted] 14:26:59 25/12/16
Ahora Temblor/sismo en concepción...
Localidad de Twitter places:
Concepción, Chile

[redacted] 14:26:59 25/12/16
Se sintió fuerte y extenso el temblor en #temuco.
Sin localidad

Temblor fuerte en —
@[redacted]
Localidad de Twitter places:
Puyehue, Chile
Sun, 25 Dec 2016 14:26:25 GMT

Figure 4: (Converted to grayscale for publication purposes) Interface showing an earthquake on December 25th, 2016. (a) Heat map of the country. (b) Signal showing the number of tweets each 60 seconds. (c) Marker indicating an event, on click, information of the event is displayed. (d) Latest tweets with buttons to reorder. (e) World map with clustered markers, showing where an event is located. (f) Buttons that filter the markers considering the source of the location information, so users can filter messages according to their location.

and in any language. The initial parametrization cost is very low and the algorithm adapts automatically over time. Our experimental results show that our algorithm is competitive in relation to the state-of-the-art improving both precision and recall. By being tolerant to noise, our method allows us to retain more information for earthquake description and detect more seismic events. Having more messages allows us to have information for event description to, for example, differentiate between consecutive events that occur in different locations. This contributes to create more complete earthquake catalogs. In addition, we roughly improve in a 100% the detection of "felt earthquakes" above 4.0 magnitude, according to our *moderate* evaluation criteria.

Our approach is limited by the geographical coverage of Twitter users, and therefore cannot detect quickly earthquakes that occur in areas without users. For future work we expect to extend this system to detect and describe other types of unexpected events, such as natural disasters.

Acknowledgment

We thank Carlos Castillo (Eurecat) and Sergio Barrientos (National Seismology Center in Chile, CSN/GUC) for their valuable comments and feedback, which has contributed to improve our work. We also thank Michelle Guy and Paul Earle (USGS) for providing the keywords and initial insight into this problem.

References

Atkinson, G. M., and Wald, D. J. 2007. "did you feel it?" intensity data: A surprisingly good measure of earthquake ground motion. *Seismological Research Letters* 78(3):362–368.

Avvenuti, M.; Cresci, S.; La Polla, M. N.; Marchetti, A.; and Tesconi, M. 2014a. Earthquake emergency management by social sensing. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on*, 587–592. IEEE.

Avvenuti, M.; Cresci, S.; Marchetti, A.; Meletti, C.; and Tesconi, M. 2014b. Ears (earthquake alert and report system): a real time decision support system for earthquake crisis management. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1749–1758. ACM.

Cameron, M. A.; Power, R.; Robinson, B.; and Yin, J. 2012. Emergency situation awareness from twitter for crisis management. In *Proceedings of the 21st international conference companion on World Wide Web*, 695–698. ACM.

Castillo, C. 2016. *Big Crisis Data: Social Media in Disasters and Time-Critical Situations*. Cambridge University Press.

Earle, P.; Guy, M.; Buckmaster, R.; Ostrum, C.; Horvath, S.; and Vaughan, A. 2010. Omg earthquake! can twitter improve earthquake response? *Seismological Research Letters* 81(2):246–251.

Earle, P. S.; Bowden, D. C.; and Guy, M. 2012. Twitter earthquake detection: earthquake monitoring in a social world. *Annals of Geophysics* 54(6).

ESA. 2012. Csiro australia – home page. https://esa.csiro.au/ausnz/index.html. [Online; accessed 3-May-2017].

GUC. 2017. National seismology center. http://www.sismologia.cl/. [Online; accessed 3-May-2017].

Guzman, J., and Poblete, B. 2013. On-line relevant anomaly detection in the twitter stream: an efficient bursty keyword detection model. In *Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description*, 31–39. ACM.

Kennett, B., and Engdahl, E. 1991. Traveltimes for global earthquake location and phase identification. *Geophysical Journal International* 105(2):429–465.

Kleinberg, J. 2003. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery* 7(4):373–397.

Mathioudakis, M., and Koudas, N. 2010. Twittermonitor: Trend detection over the Twitter stream. In *Proceedings of the 2010 international conference on Management of data*, 1155–1158. ACM.

Mendoza, M.; Poblete, B.; and Castillo, C. 2010. Twitter under crisis: Can we trust what we rt? In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, 71–79. New York, NY, USA: ACM.

Nguyen, T.; Phung, D.; Adams, B.; and Venkatesh, S. 2013. Event extraction using behaviors of sentiment signals and burst structure in social media. *Knowledge and information systems* 37(2):279–304.

Okazaki, M., and Matsuo, Y. 2010. Semantic twitter: analyzing tweets for real-time event notification. In *Recent Trends and Developments in Social Software*. Springer. 63–74.

ONEMI. 2017. Basic integral security guide for visitors and foreign residents in chile. http://www.santiago.diplo.de/contentblob/422340/Daten/14470/PDF_Handbuch_ONEMI_englisch.pdf. [Online; accessed 3-May-2017].

Robinson, B.; Power, R.; and Cameron, M. 2013. A sensitive twitter earthquake detector. In *Proceedings of the 22nd international conference on World Wide Web companion*, 999–1002. International World Wide Web Conferences Steering Committee.

Sakaki, T.; Okazaki, M.; and Matsuo, Y. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, 851–860. ACM.

Sakaki, T.; Okazaki, M.; and Matsuo, Y. 2013. Tweet analysis for real-time event detection and earthquake reporting system development. *Knowledge and Data Engineering, IEEE Transactions on* 25(4):919–931.

Sambridge, M., and Kennett, B. 2001. Seismic event location: nonlinear inversion using a neighbourhood algorithm. *Pure & Applied Geophysics* 158(1-2):241.

Sankaranarayanan, J.; Samet, H.; Teitler, B. E.; Lieberman, M. D.; and Sperling, J. 2009. Twitterstand: news in tweets. In *Proceedings of the 17th acm sigspatial international conference on advances in geographic information systems*, 42–51. ACM.

Sheth, A. 2009. Citizen sensing, social signals, and enriching human experience. *IEEE Internet Computing* 13(4):87.

SHOA. 2017. Home page. http://www.shoa.cl/en/. [Online; accessed 3-May-2017].

Stein, S., and Wysession, M. 2009. *An introduction to seismology, earthquakes, and earth structure*. John Wiley & Sons.

Strohmaier, M. 2010. Measuring earthquakes on twitter: The twicalli scale. https://mstrohm.wordpress.com/2010/01/15/measuring-earthquakes-on-twitter-the/-twicalli-scale/. [Online; accessed 25-July-2017].

Twitter. 2016. Twitter company — about. https://about.twitter.com/. [Online; accessed 25-July-2017].

USGS. 2017a. Data and products. https://earthquake.usgs.gov/data/. [Online; accessed 3-May-2017].

USGS. 2017b. Earthquake glossary. https://earthquake.usgs.gov/learn/glossary/?term=magnitude. [Online; accessed 3-May-2017].

USGS. 2017c. The modified mercalli intensity scale. https://earthquake.usgs.gov/learn/topics/mercalli.php. [Online; accessed 3-May-2017].

USGS. 2017d. Usgs earthquake hazards program. https://earthquake.usgs.gov/. [Online; accessed 3-May-2017].

Wyss, M., and Zibzibadze, M. 2009. Delay times of worldwide global earthquake alerts. *Natural hazards* 50(2):379–387.

Yin, J.; Lampert, A.; Cameron, M.; Robinson, B.; and Power, R. 2012. Using social media to enhance emergency situation awareness. *IEEE Intelligent Systems* 27(6):52–59.

Young, J. C.; Wald, D. J.; Earle, P. S.; and Shanley, L. A. 2013. Transforming earthquake detection and science through citizen seismology.

Zhao, L.; Chen, F.; Dai, J.; Hua, T.; Lu, C.-T.; and Ramakrishnan, N. 2014. Unsupervised spatial event detection in targeted domains with applications to civil unrest modeling. *PloS one* 9(10):e110206.

Zhou, D.; Chen, L.; and He, Y. 2015. An unsupervised framework of exploring events on twitter: Filtering, extraction and categorization. In *AAAI*, 2468–2475.