J Ambient Intell Human Comput (2017) 8:37–45
DOI 10.1007/s12652-016-0373-4

ORIGINAL RESEARCH

Earthquake management: a decision support system based on natural language processing

E. Fersini[1] · E. Messina[1] · F. A. Pozzi[1]

Received: 3 July 2015 / Accepted: 9 April 2016 / Published online: 26 April 2016

Abstract When an earthquake occurs, a huge amount of data is generated by social media users. Social networks play therefore a fundamental role in the development of decision support systems that could help both government and citizens. From user-generated contents, the information about an occurring emergency could be acquired and exploited to understand the critical event and its evolution over time. On the other side, the social interactions among users can be exploited as a dissemination gate to make people informed. In this paper, we present a decision support system for earthquake management based on machine learning and natural language processing to effectively extract and organize knowledge from online social media data. The proposed system, on a real Twitter dataset, has shown significant results for identifying messages related to (real) earthquakes and critical tremors, highlighting those posts provided by spontaneous users and containing any actionable knowledge about damages, magnitude, location and time references.

Keywords Decision support system · Disaster management · Earthquakes

1 Introduction and motivation

During a critical event such as an earthquake, time-sensitive decisions must be taken in order to help people, locating available resources, delivery assistance and disseminate relevant information (Yin et al. 2012; Jennex 2012). The timely acquisition of relevant geospatial data is crucial to plan and coordinate recovery actions in critical situations, especially when a disaster develops rapidly. The contents generated by the user and disseminated through social networks emerge as an alternative source of data that could be integrated in decision support systems in order to help both government and citizens for managing critical situations. From user-generated contents, the information about an occurring emergency could be acquired and exploited to promptly understand the critical event and its evolution over time. Not only emergency professionals but also common citizens may act as human sensors that can observe and monitor the disaster process. The advantage of these sensors is that they result to be often densely distributed around the site where the emergency arises. On the other side, the social interactions among users can be exploited as a dissemination gate to make people informed.

To effectively extract and organize information from the huge amount of data available in online social media, we need to design and implement several efficient computational methods able to deal with the unstructured nature of the user-generated contents. The big challenge is therefore to identify and filter only relevant information embedded in the data deluge of social media, understanding the human sensors through their online posts. Text containing potentially critical information needs to be urgently collected from the human sensors to subsequently extract and filter only actionable knowledge to be provided to emergency services and government authorities with the ultimate goal of speeding up the decision making processes. A prompt understanding of the user-generated contents would therefore enable the accomplishment of complex tasks such as emergency responding (e.g. concentrate rescue teams) and recovering (e.g. post-emergency activities based on damage assessment).

✉ E. Fersini
fersini@disco.unimib.it

[1] University of Milano-Bicocca, Milano, Italy

In this paper, our goal is to infer relevant information from the massive amounts of incoming social media data (Twitter), to finally identify high-value messages related to natural disasters (earthquakes). Although the social networks potentially offer many advantages, they also pose new challenges. The text in a user-generated message could be noisy and potentially containing inaccurate and misleading information. Moreover, the identification of disaster related messages by simply using keyword-based search leads to the retrieval of a large proportion of false alarms. We report in the following two messages containing some keywords related to the target event "earthquake".

Example 1: My daughter is a *natural calamity*: when she is going downstairs, there is an *earthquake*...the lights *shake* and our dog runs away!

Example 2: *Earthquake!* Everything is *shaking!!!!* A hole in Venice Street! Milan, now!!

While the first example is clearly a message unrelated to a critical event, the second one must be further considered to extract in short time some useful knowledge such as geo-localization and information about damages. Machine learning and natural language precessing methods can be effectively exploited in order to solve many challenging issues arising when dealing with user-generated contents. Some initial tentatives to deal with Twitter messages for earthquake management has been recently proposed in the state of the art. In (Sakaki et al. 2010), Sakaki at al. consider Twitter users as *virtual sensors* that can contribute to monitor what happen if an earthquake occurs. The authors investigated the integration of a supervised classification model [Support Vector Machines (SVM)] for earthquake detection and a Bayesian Filtering approach (Kalman and Particle Filtering) for tracking the event over time and location. A more recent study is related to a decision support system based on burst detection (Avvenuti et al. 2014b). The authors investigated a novel approach, to detect unusual occurrences of a phenomenon within a short time window. Although the above mentioned investigations represent a fundamental step towards the design of effective decision support systems for earthquake managements, they suffer of two limitations that the proposed paper intends to overcome:

- In Sakaki et al. (2010), geo-location and temporal information are assumed to be available from Twitter. However, most of these information are missing or cannot be reliable to characterize an occurring earthquake.
- In Avvenuti et al. (2014a, b), the detection of a relevant earthquake event is based a single classification model that can be biased by the natural language uncertainty that characterizes social networks.

2 Natural language processing: capturing situational awareness

2.1 The proposed framework

In order to define a decision support system able to deal with critical earthquake events, several issues need to be modeled:

1. *WHAT*: Twitter is a distributed virtual sensor system (Crooks et al. 2013) that needs to be analyzed to understand if an actual earthquake is occurring or an earthquake mention is not effectively related to a critical event;
2. *WHO*: There are more than 40 million users as "Twitter sensors" (Sakaki et al. 2010) that can contribute to the situational awareness related to an earthquake. It's necessary to distinguish between the "human-being sensors on site" and "news media providers". The identification of human-being sensors is fundamental to capture help requests, establish a contact point on the site and to identify specific location for damage assessment.
3. *WHERE*: Most of the social networks provide a functionality to the users to make a registration on their location. Unfortunately, Twitter users have been slow to adopt geospatial features: 1 % of tweets are geo-localized (Mahmud et al. 2012) (in a random sample of over 1 million Twitter users, only 26 % have listed a city name (e.g., Los Angeles, CA), while the rest are overly general (e.g., California), missing or nonsensical (e.g., Wonderland) (Cheng et al. 2010)).
4. *WHEN*: A tweet can be associated with a time stamp. However, a tweet message can report an opinion or a fact related to the past, biasing therefore the time stamp information. It's therefore important to complement the time reference provided by the social network platform with the temporal information eventually provided by the user in the tweet itself.

In order to effectively reduce the false positive of earthquake warnings due to the noise sensitivity related to language ambiguity, and therefore to provide a more accurate prediction of real target events, a framework based on natural language processing has been developed. The proposed decision support system addresses the above mention issues by formulating several questions to be replied, to identify messages that contribute to a situational awareness able better characterize critical earthquake events:

1. *WHAT*:
   - Is the tweet about a real earthquake event?
   - Does the tweet report any detail about magnitude or damages?

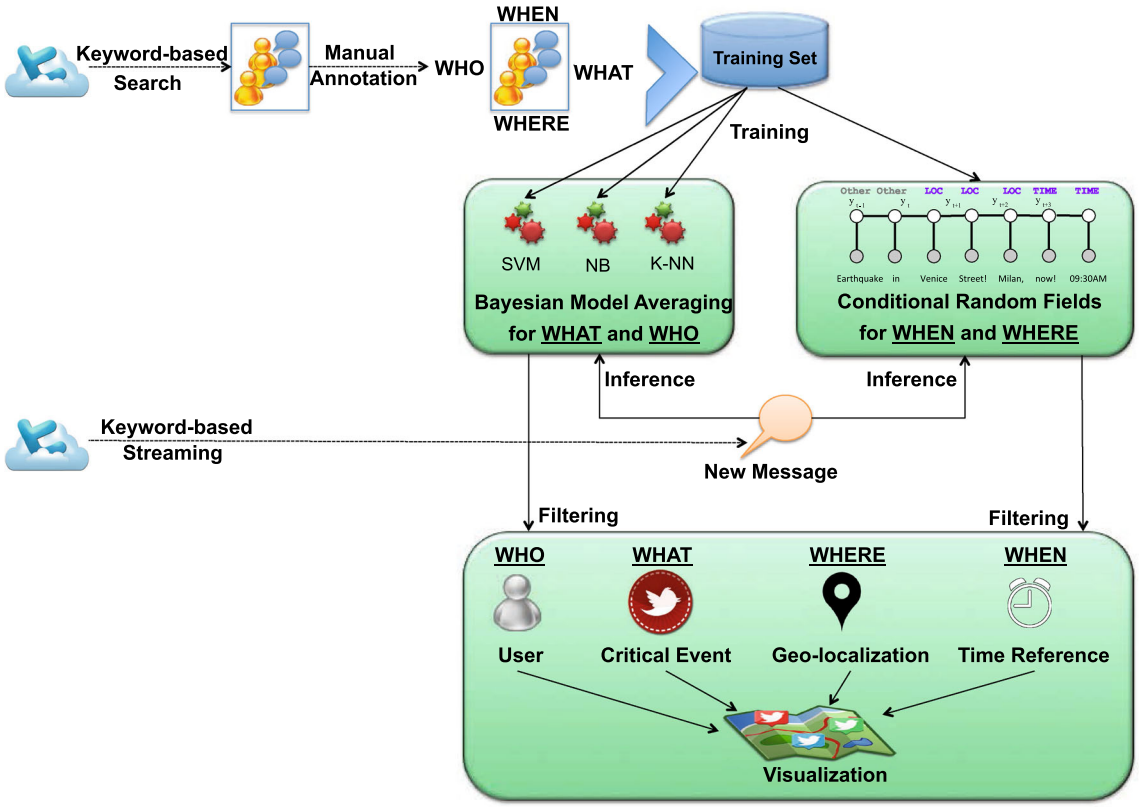


Fig. 1 High level architectural overview of the proposed decision support system

- Is the tweet about a critical earthquake event?

2. *WHO*:
   - Is the author a spontaneous user or media news?
3. *WHERE*:
   - Is there any mention about the earthquake location?
4. *WHEN*:
   - Is there any temporal mention about an occurring earthquake?

A high level architecture of the proposed decision support system is reported in Fig. 1. The system exploits the Twitter API[1] to collect posts related to earthquakes. Through the Twitter Search API we collected a set of messages that have been subsequently manually annotated (with respect to the above mentioned Who, What, Where and When questions) to create the corresponding training set. In particular, the following terms have been used to collect earthquake related messages: *earthquake*, *quake*, *tremor*, *seism*, *seismic swarm*, *aftershock* and *magnitude*. Once the training sets have been created, two components based on natural language processing are instantiated. The first module, designed to address the WHO and WHAT issues, is based on an ensemble classification approach called Bayesian Model Averaging. The second component, conceived for tackling the WHEN and WHERE issues, is built upon a popular probabilistic model for structured prediction known as Conditional Random Fields. The models enclosed in both components are trained according to the annotated data, to be subsequently used to process new incoming messages. Through the Twitter Streaming API, new messages containing the above mentioned keywords are captured, to be subsequently submitted to the WHO, WHAT, WHEN and WHERE models previously trained. Once the prediction of the models are provided, they are filtered to identify high-value messages containing situational awareness information useful for decision making purposes. The filtered information are finally visualized on a map by using several kinds of perspectives. These characteristics make the proposed system suitable to be adopted as early warning system during an earthquake as well as post-event support for damage assessment and recovery.

2.2 WHO and WHAT: Bayesian model averaging

The response to the *WHO* and *WHAT* questions can be viewed as a classification problem, which could be addressed

[1] https://dev.twitter.com.

by any traditional machine learning algorithm. In particular, given a set of user-generated contents collected according to specific keywords, it is possible to annotate them to create a training set and finally induce a classifier. However, there is no agreement on which methodology is better than others: one classifier could perform better than others with respect to a given application domain, while a further approach could outperform the others when dealing with a given language or linguistic register. The uncertainty left by the natural language expression can introduce a bias in the prediction models, therefore reducing the generalization capabilities of the classifiers. In order to overcome this limitation, an ensemble of different classifiers could lead to more robust and accurate classification. To this purpose, in this paper we investigated a Bayesian Model Averaging (BMA) approach (Pozzi et al. 2013; Fersini et al. 2014). In particular, all the possible classifiers in the hypothesis space are combined in a voting mechanism that could exploit their marginal prediction capabilities and their reliabilities. Given a message s and a set C of independent classifiers, the probability of label $l(s)$—related to the what and who questions—can be estimated by the following paradigm:

$$P(l(s) \mid C, \mathcal{D}) = \sum_{i \in C} P(l(s) \mid i, \mathcal{D}) P(i \mid \mathcal{D}) \tag{1}$$

where $P(l(s) \mid i, \mathcal{D})$ is the marginal distribution of the label predicted by classifier i and $P(i \mid \mathcal{D})$ denotes the posterior probability of model i. The posterior $P(i \mid \mathcal{D})$ can be computed as:

$$P(i \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid i) P(i)}{\sum_{j \in C} P(\mathcal{D} \mid j) P(j)} \tag{2}$$

where $P(i)$ is the prior probability of i and $P(\mathcal{D} \mid \cdot)$ is the model likelihood. In Eq. (2), $P(i)$ and $\sum_{j \in C} P(\mathcal{D} \mid j) P(j)$ are assumed to be a constant and therefore can be omitted. Therefore, BMA assigns the optimal label $l^*(s)$ to s according to the following decision rule:

$$\begin{aligned} l^*(s) = \arg\max_{l(s)} P(l(s) \mid C, \mathcal{D}) &= \sum_{i \in C} P(l(s) \mid i, \mathcal{D}) P(i \mid \mathcal{D}) \\ &= \sum_{i \in C} P(l(s) \mid i, \mathcal{D}) P(\mathcal{D} \mid i) P(i) \\ &= \sum_{i \in C} P(l(s) \mid i, \mathcal{D}) P(\mathcal{D} \mid i) \end{aligned} \tag{3}$$

The implicit measure $P(\mathcal{D} \mid i)$ can be easily replaced by an explicit estimate, known as F_1-measure, obtained during a preliminary evaluation of the classifiers i. In particular, by performing a cross validation each classifier can produce an averaged measure stating how well a learning machine generalizes to unseen data. Considering ϕ-folds for cross validating a classifier i, the measure $P(\mathcal{D} \mid i)$ can be approximated as

$$P(\mathcal{D} \mid i) \approx \frac{1}{\iota} \sum_{\iota=1}^{\phi} \frac{2 \times P_{i\iota}(\mathcal{D}) \times R_{i\iota}(\mathcal{D})}{P_{i\iota}(\mathcal{D}) + R_{i\iota}(\mathcal{D})} \tag{4}$$

where $P_{i\iota}(\mathcal{D})$ and $R_{i\iota}(\mathcal{D})$ denotes precision and recall obtained by classifier i at fold ι. According to Eq. 3, we take into account the vote of each classifier by exploiting the prediction marginal instead of a 0/1 vote and we tune this *probabilistic claim* according to the ability of the classifier to fit the training data[2] . This approach allows the uncertainty of each classifier to be taken into account, avoiding over-confident inferences.

2.3 WHERE and WHEN: conditional random fields

The response to the *WHEN* and *WHERE* questions can be viewed as a segmentation and labelling problem on a text, which could be addressed by following a sequential learning paradigm. In our case, once each token of the Twitter messages has been manually annotated with the tags *Location*, *Time* and *Other*, Conditional Random Fields (Lafferty et al. 2001) have been exploited to train the underlying probabilistic model and automatically label new incoming messages.

A conditional random field is an indirected graphical model that defines the joint distribution $P(y|x)$ of the predicted labels (hidden states) $y = y_1, \ldots, y_N$ given the corresponding tokens (observations) $x = x_1, \ldots, x_N$. Now, consider X as the random variable over a words sequence (tweet) to be labeled, and Y is the random variable over corresponding label sequences over a finite label alphabet $\mathcal{Y}$. The joint distribution $P(X, Y)$ is represented by a conditional model $P(Y|X)$ from paired observation and label sequences, and the marginal probability $p(X)$ is not explicitly model. The formal definition of CRF is given below:

Definition 1 *(Conditional random fields)* Let G = (V, E) be a graph such that $Y = (Y_v)_{v \in V}$, so that Y is indexed by the vertices of G. Then (X, Y) is a Conditional Random Field, when conditioned on X, the random variables Y_v obey the Markov property with respect to the graph: $p(Y_v|X, Y_w, w \neq v) = p(Y_v|X, Y_w, w \sim v)$, where $w \sim v$ means that w and v are neighbors in G.

In our case study, hidden states y correspond to date/-time and locations (and *other* for all the rest), while each observation x denotes a social media post (tweet) with the corresponding tokens. Concerning the extraction of dates and times, the model is able to identify not only simple references like "09:30 AM", but also indications like "5

[2] Concerning the learning feature space, all the models enclosed in BMA are trained using a traditional vector space model representation (Salton et al. 1975) with a Boolean weighting schema.

mins ago". In the latter case, the system computes the exact time reference (e.g. 10-11-2015:08.46.32 PM) using the time stamp related to the post publication. An instance of the CRF model for the proposed framework is reported in Fig. 2.

3 Experimental investigation

In order to evaluate the proposed decision support system, an experimental investigation has been conducted. A dataset has been created by collecting 1500 Italian tweets mentioning the keywords *earthquake*, *quake*, *tremor*, *seism*, *seismic swarm*, *aftershock* and *magnitude*. The obtained tweets have been manually annotated as follows:

1. **WHAT**:
   - (Q1) Real earthquake event? YES or NO
   - (Q2)Any information about magnitude or damages? YES or NO
   - (Q3)Critical earthquake event? YES or NO
2. **WHO**:
   - (Q4)Spontaneous user ? YES or NO
3. **WHERE**:
   - (Q5)Any mention about the earthquake location? LOCATION and NO INFO
4. **WHEN**:
   - (Q6)Any temporal mention about the earthquake? TIME and NO INFO

The inter-agreement between annotators has been computed according to the Fleiss' kappa statistics (Fleiss 1971), which measures the reliability agreement of labeling over that which would be expected by chance (when multiple annotators are involved).

In our case, the inter-agreement statistics $\kappa = 0.70$ indicates a substantial agreement among annotators. The statistics about the collected dataset are reported in Fig. 3.

In order to evaluate the proposed approach, Accuracy, Precision, Recall and F-Measure have been computed using a tenfolds cross validation. Concerning the baseline classifiers to enclose in BMA, Decision Trees (DT) (Quinlan 2014), SVM (Cortes and Vapnik 1995), Naive Bayes (NB) (McCallum and Nigam 1998), Logistic Regression (LR) and K-Nearest Neighbors (KNN) (Aha et al. 1991) have been considered. All these models have been also compared to a traditional ensemble approach known as Voting. Voting has been evaluate according to the following decision rules:

- Majority Voting (MV): the final label is selected by a democratic voting
- Maximum posterior (Max): the final label is selected according to the maximum a posteriori probability among the classifiers
- Product of posteriors (Prod): the decision is determined by the product of the posterior probabilities
- Average of posteriors (Mean): the decision is determined according to the mean of a posteriori probabilities

By analyzing Fig. 4, some considerations can be drawn for the WHAT and WHO questions. First of all, it emerges that although the datasets annotated with respect to the four questions are quite unbalanced, in most of the cases all the approaches are able to guarantee remarkable accuracies.

The only exception is related to the discrimination of spontaneous users, i.e. human beings on the site, with respect to news media providers. The lower performance related to the speaker identification is mainly due to the similarity of the language between the two subjects. We report in Fig. 5 an example showing the impersonal writing style related to an earthquake. Although the messages are really similar from a linguistic point of view, the first one reported in Fig. 5a has been posted by a news media provider, while the second one in Fig. 5b has been posted by a spontaneous user.

Focusing on the overall recognition abilities of the investigated models, it emerges that the ensemble classification based on BMA outperforms both the other voting mechanisms and the baseline classifiers in every configuration of the studied datasets[3]. These results confirm our initial hypothesis that the BMA ensemble is able to deal with the ambiguity and uncertainty of the natural language better than other approaches.

If we consider Precision, Recall and F-Measure, we can grasp more peculiar behaviors of the considered models. We report in Tables 1, 2, 3 and 4 the above mentioned measures distinguished in positive (YES denoted by "+") and negative (NO represented by "−") messages, for the considered baseline methods as well as for the ensembles. The best performance are reported in bold. Considering the results obtained on the four questions, it is easy to note that the traditional methods mostly obtain high precision and low recall on a given target class, and low precision and high recall on the other class. Consider for instance the KNN classifier in Table 1. KNN achieves 0.9026 and 0.4879 on precision and recall for the positive class, while for the negative one the performance are 0.8093 and 0.9764 respectively. A similar behavior can be observed on the

[3] T-Test rejects $H_0 : \mu BMA - \mu other = 0$, where the critical region is $T > 2.92$ and $T = 3.08$ with $\alpha = 0.05$. Then the test does not reject $H_1 : \mu BMA - \mu other > 0$.

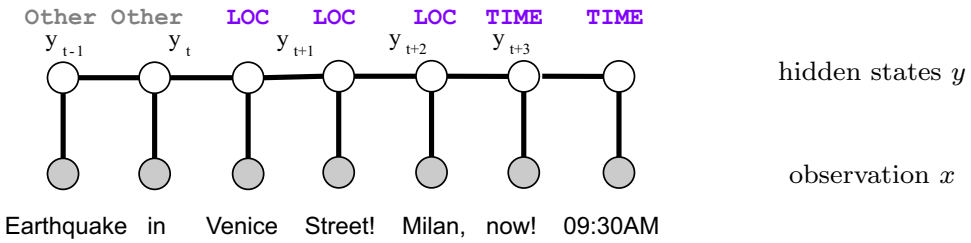


Fig. 2 Graphical representation of a linear-chain CRF: *white circles* represent hidden states y (the output sequence of labels) and the *grey ones* denote the observation x (the input sequence of tokens). The grounded CRF has been used for extracting locations (LOC) and time references (TIME) in earthquake related messages

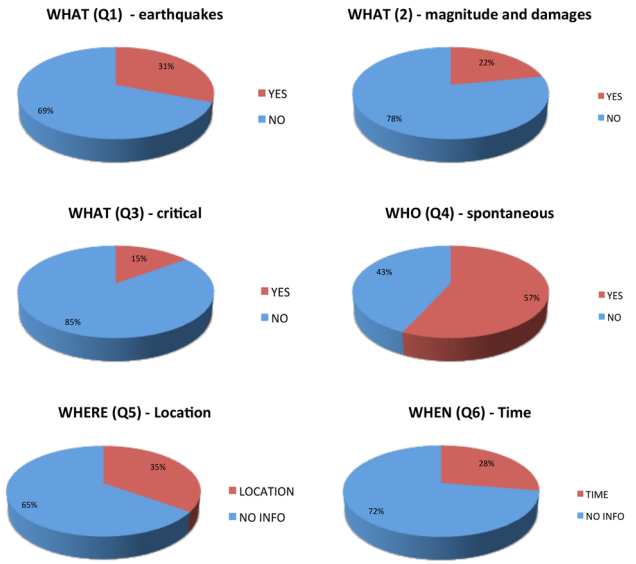


Fig. 3 Dataset distribution

other approaches. The only exception is BMA, which guarantees a good trade-off of precision and recall for both classes. In fact, the F+ and F− measures confirm that BMA has a good ability to capture earthquake related posts as well as correctly identify messages with high informative value. Analogous considerations can be drawn for the other questions, where BMA emerges as the most stable and robust method for the classification task.

Concerning the WHEN and WHERE questions, a summary of the performance are reported in Table 5. Precision, Recall and F-Measure have been detailed for the three types of entities, i.e. Location, Time and Other, to be detected in a user-generated contents. It is easy to note that the proposed system shows different performance for capturing information about the location of an earthquake and the time reference of the target event. While for the time reference CRFs are able to achieve a good compromise between precision and recall, for the identification of locations the task becomes more difficult. In particular, the proposed system is able to correctly classify the tokens related to a geo-spatial information, but it shows a low recall due to the variety of expressions for identifying the same location. For instance, different messages referring to the same location *L'Aquila* can mention the site in several ways such as *lAquila*, *l'Aquil*, *AquilaAbruzzo* and *#AQ*. In this case the misspelling of a location has a great impact in the inference phase, dramatically reducing the recall performance.

Considering the promising results on precision, but the reduced in recall, it emerges that the proposed system

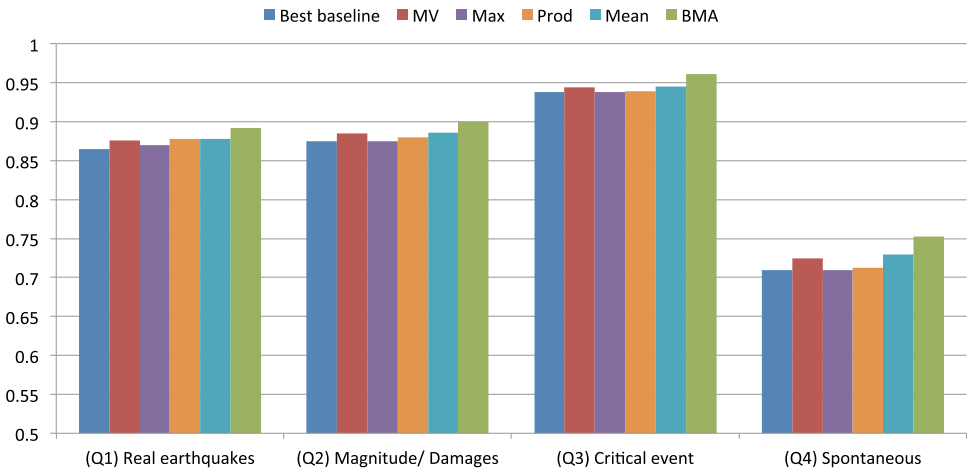
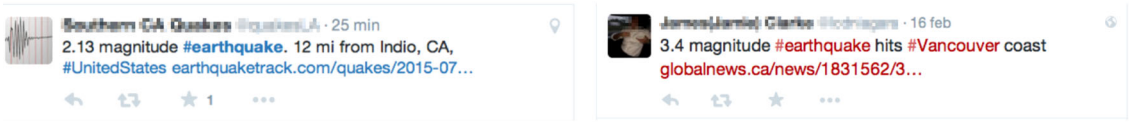


Fig. 4 Accuracy comparison for the WHAT and WHO questions



(a) News media provider **(b)** Spontaneous user

Fig. 5 Example of news media provider and spontaneous tweets

Table 1 Precision, recall and F-measure per class related to the question WHAT-(Q1)

	WHAT (Q1)—Is the tweet about a real earthquake event?						
	BMA	MV	DT	SVM	NB	REG	KNN
P+	0.8424	0.8519	0.7719	0.8350	0.7681	0.8165	**0.9026**
R+	0.7247	0.6984	0.6781	0.6862	**0.7510**	0.7024	0.4879
F+	**0.7791**	0.7675	0.7220	0.7533	0.7595	0.7552	0.6334
P−	0.8837	0.8747	0.8629	0.8695	**0.8893**	0.8743	0.8093
R−	0.9391	0.9455	0.9100	0.9391	0.8982	0.9291	**0.9764**
F−	**0.9105**	0.9087	0.8858	0.9030	0.8937	0.9008	0.8850

Bold numbers denote the best performing approach

Table 2 Precision, recall and F-measure per class related to the question WHAT-(Q2)

	WHAT (Q2)—Any information about magnitude or damages?						
	BMA	MV	DT	SVM	NB	REG	KNN
P+	0.7744	0.7719	0.7411	0.7455	0.6104	0.7406	**0.9077**
R+	0.6571	0.6286	0.5971	0.5943	**0.8275**	0.6200	0.3371
F+	**0.7110**	0.6929	0.6614	0.6614	0.7025	0.6750	0.4917
P−	0.9073	0.9005	0.8924	0.8919	**0.9571**	0.8976	0.8413
R−	0.9461	0.9477	0.9412	0.9428	0.8792	0.9388	**0.9903**
F−	**0.9263**	0.9235	0.9161	0.9166	0.9165	0.9177	0.9098

Bold numbers denote the best performing approach

Table 3 Precision, recall and F-measure per class related to the question WHAT-(Q3)

	WHAT (Q3)—Is the tweet about a critical earthquake event?						
	BMA	MV	DT	SVM	NB	REG	KNN
P+	0.8288	0.8287	0.8211	0.7990	0.5972	0.7991	**0.8387**
R+	0.7731	0.7521	0.7521	0.6849	**0.9034**	0.7185	0.5462
F+	**0.8000**	0.7885	0.7851	0.7376	0.7191	0.7566	0.6616
P−	0.9605	0.9570	0.9569	0.9458	**0.9813**	0.9512	0.9246
R−	0.9719	0.9726	0.9711	0.9696	0.8926	0.9681	**0.9815**
F−	**0.9661**	0.9647	0.9640	0.9576	0.9348	0.9596	0.9522

Bold numbers denote the best performing approach

should be take into account the possibility of text normalization (Fauffman and Kugal 2010) before proceeding with the segmentation and classification task with CRFs.

Table 4 Precision, recall and F-measure per class related to the question WHO-(Q4)

	WHO (Q4)—Is the author a spontaneous user or media news?						
	BMA	MV	DT	SVM	NB	REG	KNN
P+	0.7484	0.7435	0.6868	**0.7526**	0.7512	0.7182	0.7009
R+	**0.7607**	0.7541	0.7155	0.7310	0.6858	0.7585	0.6615
F+	**0.7545**	0.7488	0.7009	0.7416	0.7170	0.7378	0.6807
P−	**0.6766**	0.6686	0.6012	0.6568	0.6270	0.6546	0.5829
R−	0.6618	0.6560	0.5679	0.6818	**0.6993**	0.6058	0.6263
F−	**0.6691**	0.6623	0.5841	0.6691	0.6611	0.6293	0.6038

Bold numbers denote the best performing approach

Table 5 Performance of CRF on the WHEN and WHERE questions

Label	Precison	Recall	F-measure
Other	0.9476	0.9827	0.9647
Location	0.8489	0.3850	0.5297
Time	0.7434	0.7789	0.7607
Avg. performance	0.8466	0.7155	0.7517

Orthographic normalization and syntactic disambiguation, would improve not only the recall, but also precision and therefore the overall ability of detecting location and time reference of a real earthquake event.

A final consideration about the proposed decision support system relates to the time efficient of the system itself. A time sensitive decision strictly depends on the computational complexity related to the models inference phase. We can therefore distinguish between the two macro-models, i.e. Bayesian Model Averaging for dealing with WHAT and WHO and Conditional Random Fields for addressing WHEN and WHERE. Concerning BMA, let assume the inference phase of a given classifier be proportional to $O(1)$. Since BMA is composed of N distinct classifiers, it results to be linear in time complexity with respect to the number of models enclosed into the ensemble. Regarding CRFs, its computational complexity is mainly related to the Viterbi algorithm used for labeling the textual tokens belonging to a tweet. The complexity of this algorithm is $O(T \cdot S^2)$, where T denotes the number of tokens enclosed in a post and S represents the state space that in our case is composed of three hidden states

corresponding to the labels *Other*, *Location* and *Time*. From a practical point of view, we have estimated the time performance of the proposed system on a Desktop PC with Windows 7 64-bit Operating System, Pentium Quad Core i7 3.10GHz Processor and 8GB RAM. On this configuration, the average time required by BMA is 0.0007 *s/tweet*, while for CRF the average complexity is 0.008 *s/tweet*. These average time performance provide an evidence about the applicability of the proposed system to a real scenario of emergency management.

4 Discussion

The promising results described in the previous sections represent a first step towards the definition of a general-purpose disaster management system. In fact, although the experimental investigations have revealed the potential of natural language processing as a key element for the management of earthquake events, some issues need to be discussed:

- **Language variability.** Dealing with keyword-based searches (both offline and streaming) is a key issue that needs to be properly addressed to guarantee an adaptive system able to deal with the language variability of user-generated contents in social networks. People talk about an earthquake—or any other emergency—in a wide variety of ways, using emerging words or keywords (e.g. hashtags) that can not be a priori predicted. The proposed system should be therefore extended to capture the semantics of the messages in order to deal with the language variability and dynamism that characterize a real social networking environment. A possible solution to this issue is represented by a deep learning neural network architecture aimed at learning how to represent different words that are semantically related. Through a word-embedding strategy [(e.g. Skyp-Gram model (Mikolov et al. 2013)] it will be possible to project words onto a multi-dimension vector space such that the proximity between two vectors will indicate the semantic similarity between their associated words. According to this projection, the pre-defined keywords can be used as seeds to obtain the words with the highest similarity via the learned word vector representations.
- **Applicability to other emergencies.** The machine learning methods enclosed in the proposed framework work well under the assumption that the training and test data are drawn from the same feature space (words) and the same distribution. When a different type of emergency needs to be addressed, the feature space and the distribution change implying a new training phase for deriving the corresponding statistical models. In a real world scenario, it is expensive or impossible to recollect the needed training data and rebuild the models. In order to overcome this limitation, we need to adapt the learned models to new emergencies. To this purpose, a transfer learning strategy (Pan and Yang 2010) could be adopted for exploiting the models learned on a different disaster, reducing therefore the need and effort to recollect the training data and derive new models for new emergencies.
- **Adaptability to other social networks.** The proposed framework is mainly based on Twitter monitoring, thanks to the publicly available data provided by the users. In order to converge to a more efficient and effective management system, the proposed architecture should be extended to collect the user sensing from other social networks like Facebook, Google+ and Flickr. However, considering that the contents generated on some social networks are not public by default, the proposed architecture should be enriched to deal with the privacy issues. To this purpose, the system should integrate several connectors, adapters and privacy-preserving applications for collecting specific contents on specific social networking platforms. For instance, an integration with Facebook would require to develop a specific APP, to adopt "Graph API" with the corresponding application registration (access_token) and finally collect Facebook posts by the available API. Through a verified APP and Graph API, it will be possible to search in the Facebook graph different types of objects like posts, groups, places and pictures that correspond to specific emergency keywords.

5 Conclusion

In this paper, a novel decision support system based on natural language processing has been proposed to address earthquake early warning signals from social networks. The system has been experimented on a real dataset, showing significant results for identifying user-generated contents related to (real) earthquakes and critical tremors, highlighting those posts provided by spontaneous users and containing any knowledge about damages, magnitude, location and time references. Although we are conscious that the proposed system is a first step towards an effective decision support system, we believe that it represents a good starting point for addressing security issues. Concerning the future work, the improvement of recognition of spontaneous users is a first issue. In order to increase the discrimination between human being on the site and news media providers, several linguistic characteristics, such as

punctuation, uppercase and emoticon, will be included in the dataset representation for a better training of the classification models. A final goal, the proposed decision support system will be applied in other contexts such as river floorings, storms, wildfire as well as civil disorders.

References

Aha DW, Kibler D, Albert MK (1991) Instance-based learning algorithms. Mach Learn 6(1):37–66

Avvenuti M, Cresci S, La Polla MN, Marchetti A, Tesconi M (2014a) Earthquake emergency management by social sensing. In: Proceedings of the international conference on pervasive computing and communications workshops, pp 587–592

Avvenuti M, Cresci S, Marchetti A, Meletti C, Tesconi M (2014b) EARS (earthquake alert and report system): a real time decision support system for earthquake crisis management. In: Proceedings of the international conference on knowledge discovery and data mining, pp 1749–1758

Cheng Z, Caverlee J, Lee K (2010) You are where you tweet: a content-based approach to geo-locating twitter users. In: Proceedings of the 19th ACM international conference on information and knowledge management, pp 759–768

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297

Crooks A, Croitoru A, Stefanidis A, Radzikowski J (2013) Earthquake: twitter as a distributed sensor system. Trans GIS 17(1):124–147

Fauffman M, Kugal K (2010) Syntactic normalization of twitter messages. In: Proceedings of the international conference on natural language processing

Fersini E, Messina E, Pozzi FA (2014) Sentiment analysis: Bayesian ensemble learning. Decis Support Syst 68:26–38

Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychol Bull 76(5):378

Jennex M (2012) Managing crises and disasters with emerging technologies: advancements. IGI Global, Hershey, PA

Lafferty JD, McCallum A, Pereira FCN (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th international conference on machine learning, pp 282–289

Mahmud J, Nichols J, Drews C (2012) Where is this tweet from? inferring home locations of twitter users. In: Proceedings of the international conference on weblogs and social media, pp 511–514

McCallum A, Nigam K (1998) A comparison of event models for naive bayes text classification. In: Workshop on learning for text categorization-15th national conference on artificial intelligence, pp 41–48

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems, pp 3111–3119

Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Trans Knowl Data Eng 22(10):1345–1359

Pozzi FA, Fersini E, Messina E (2013) Bayesian model averaging and model selection for polarity classification. In: Natural language processing and information systems, pp 189–200. Springer, Berlin

Quinlan JR (2014) C4. 5: programs for machine learning. Elsevier, London

Sakaki T, Okazaki M, 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, pp 851–860

Salton G, Wong A, Yang C-S (1975) A vector space model for automatic indexing. Commun ACM 18(11):613–620

Yin J, Lampert A, Cameron M, Robinson B, Power R (2012) Using social media to enhance emergency situation awareness. IEEE Intel Syst 6:52–59