Consiglio Nazionale delle Ricerche

Towards a Timely Prediction of Earthquake Intensity with Social Media

S. Cresci, M. La Polla, A. Marchetti, C. Meletti, M. Tesconi

IIT TR-12/2014

Technical report

Settembre 2014

Istituto di Informatica e Telematica

Towards a Timely Prediction of Earthquake Intensity with Social Media

Stefano Cresci
Institute for Informatics and Telematics (IIT)
National Research Council (CNR), Pisa, Italy
stefano.cresci@iit.cnr.it

Mariantonietta N. La Polla
Institute for Informatics and Telematics (IIT)
National Research Council (CNR), Pisa, Italy
mariantonietta.lapolla@iit.cnr.it

Andrea Marchetti
Institute for Informatics and Telematics (IIT)
National Research Council (CNR), Pisa, Italy
andrea.marchetti@iit.cnr.it

Carlo Meletti
National Institute of Geophysics and Volcanology
Pisa, Italy
carlo.meletti@pi.ingv.it

Maurizio Tesconi
Institute for Informatics and Telematics (IIT)
National Research Council (CNR), Pisa, Italy
maurizio.tesconi@iit.cnr.it

ABSTRACT

A growing number of people is turning to Social Media in the aftermath of emergencies to search and publish critical and up to date information. Retrieval and exploitation of such information may prove crucial to decision makers in order to minimize the impact of disasters on the population and the infrastructures. Yet, to date, the task of the automatic assessment of the consequences of disasters has received little to no attention. Our work aims to bridge this gap, merging the theory behind statistical learning and predictive models with the data behind social media. Here we investigate the exploitation of Twitter data for the improvement of earthquake emergency management. We adopt a set of predictive linear models and evaluate their ability to map the intensity of worldwide earthquakes. The models build on a dataset of almost 5 million tweets and more than 7,000 globally distributed earthquakes. We run and discuss diagnostic tests and simulations on generated models to assess their significance and avoid overfitting. Finally we deal with the interpretation of the relations uncovered by the linear models and we conclude by illustrating how findings reported in this work can be leveraged by existing emergency management systems. Overall results show the effectiveness of the proposed techniques and allow to obtain an estimation of the earthquake intensity far earlier than conventional methods do. The employment of the proposed solutions can help understand scenarios where damage actually occurred in order to define where to concentrate the rescue teams and organize a prompt emergency response.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.2.8 [**Database Management**]: Database Applications—*Data mining*; I.5.1 [**Pattern Recognition**]: Models—*Statistics*

General Terms

Algorithm, Experimentation, Measurement

Keywords

Predictive modeling, feature selection, information retrieval, social sensing, web mining, emergency management

1. INTRODUCTION

Social Media (SM) is the most effective, sophisticated and powerful way to gather preferences, tastes and activities of groups of users in the context of Web 2.0 [24]. Therefore, SM users could be regarded as social sensors, namely as a source of information about situations and facts related to the users (e.g., their preferences or experiences) and their social environment [33], as asserted by the Social Sensing paradigm. Among the most widespread SMs are online social networks (OSNs) such as Facebook, Twitter and Weibo. These platforms have grown bigger in the last few years, becoming a primary hub for public expression and interaction. The advantage of exploiting OSNs compared to traditional methods of investigation lies in the spontaneous participation of the users, in that their contribution is made without pressure or influence from others. Twitter, in particular, counting a total of 645 million active users and 58 million messages shared every day[1], introduced a policy and a message format which encourages users to make their messages, commonly known as tweets, by default publicly available. Furthermore, the 140 characters limitation imposed to Tweets length causes Twitter users to share more topic specific content. The global spread of the Twitter phenomenon thus enabled a new wave of experimentation and research

[1] http://www.statisticbrain.com/twitter-statistics/

on web stream mining. Now, it has been shown that this vast amount of data encapsulates useful signals driven by our everyday life [20]. Data mining and statistical learning methods can be applied to retrieve and extract useful knowledge from the *digital shadows* cast everyday on the web.

Given this picture, it's not surprising that SM data, and in particular Twitter data, has been deeply studied by academia for a broad variety of purposes. Emergency Management is a promising field of application for Social Sensing since it is possible to retrieve and exploit the content shared on these web platforms to gather up to date information on emerging situations of potential danger [4]. Emergency management is one of the research fields which have attracted the most attention in the last few years. Many studies have focused on the understanding of the on-the-ground community knowledge during disasters [16]. Such knowledge, reflected on the social web by the multitude of emergency related messages, can greatly contribute to improve situational awareness during all kinds of crises. Yet, despite many compelling findings along this line of research, the task of the automatic assessment of the consequences of disasters has received little to no attention. Our work aims to bridge this gap, merging the theory behind statistical learning and predictive models with the data behind social media.

We apply our methodology to the field of earthquake emergency management. Other previous works employed web mining techniques in the same field [25] [3], however almost no effort has been made towards the assessment of earthquake consequences. The severity of an earthquake is described by both *magnitude* and *intensity.* Magnitude characterizes earthquakes measuring the energy released and is accurately and timely measured by seismographs. By contrast, intensity indicates the local effects and potential for damage produced by an earthquake. It is verified after the earthquake has occurred with direct surveys on the field. Intensity can also be estimated from instrumental measurements by using empirical relationships.
We address one central question. Can social media data predict earthquake intensity?
Models capable of accurately and timely defining observed earthquake intensity may have a huge impact on the mitigation of the damages. Currently adopted estimations can only infer shaking level distribution in the epicentral area in terms of ground-motion parameters and of instrumentally derived intensities [28]. This scenario does not take into account information coming from the earthquake-stricken area and can be greatly improved with the analysis of data automatically collected in near-real-time site by site. The importance of social data, such as eyewitness reports, towards the estimation of earthquake intensity have long been asserted. Data collected from online surveys is already employed to obtain a more accurate characterization of the consequences of earthquakes [10]. However, such a solution obviously lacks responsiveness and the output of this system is often subject to updates even weeks after the earthquake[2]. Here our objective is to leverage data collected from on-the-ground social sensors in a responsive predictive system.

The models we propose build on a dataset of almost 5 million tweets and more than 7,000 globally distributed earthquakes. Each analyzed earthquake is described by 45 distinct numeric features and by their interactions. On average, the proposed models predict earthquake intensity with an error as low as 0.5 on a continuos 1 → 10 scale. These results can help understand scenarios where damage actually occurred in order to define where to concentrate the rescue teams and organize a prompt emergency response.

The remainder of this paper is organized as follows: Section 2 describes the related work. Section 3 discloses the details of our dataset, Section 4 describes the adopted methodology. Section 5 details the results of our study and Section 6 draws the conclusions and describes future work.

2. RELATED WORK

Several initiatives, both in scientific and in application environments, have been developed with the aim of exploiting information available on Social Media (SM). To the best of our knowledge predictive models with SM have never been applied to emergency management. In this section we survey works in the fields of both social media emergency management and prediction from social media data. We aim to leverage previous experiences in both research fields and apply them to the prediction of earthquake intensity.

2.1 Social media emergency management

Works such as [4] highlighted how disasters and emergencies cause changes in human behavior and dynamics which are clearly distinguishable in SM. These studies showed the potential for the exploitation of such information and paved the way for the current research in the fields of social media emergency management. The vast majority of existing works have focused on the event detection and information dissemination tasks, with only a few exceptions dealing with the automatic damage assessment, thoroughly discussed in the remainder of this section.

To date the only work which adopted a predictive modeling approach to earthquake damage assessment with SM is described in [6]. The goal is the estimation of the earthquake intensity via statistical models. Authors benefited from a Twitter data grant[3] and performed a combined analysis on both tweets and seismographic data[4]. Achieved results are overall interesting with some of the proposed models able to map seismographic and Twitter data into an estimation of the earthquake intensity. However, the study proposed in [6] still leaves many open issues. Among all proposed models, those trained only with Twitter data achieved the worst performances. Researchers experimented only with a few basic tweets features, which might have negatively influenced the results. Moreover the proposed models are not thoroughly discussed and are only evaluated by means of mean squared error (MSE). Therefore it is difficult to assess the statistical significance of the proposed solutions and it remains unclear what Twitter may contribute to earthquake intensity estimation. Our results, discussed in Section 5, shed light on this issue and show how our detailed analysis leads to differ-

[2] We provide a detailed discussion of currently adopted intensity estimation techniques in Section 3.1

[3] https://engineering.twitter.com/research/data-grants

[4] https://blog.twitter.com/2014/using-twitter-to-measure-earthquake-impact-in-almost-real-time

ent conclusions. In addition, while researchers in [6] focused on earthquakes in Japan, we collected and analyzed data for earthquakes from all over the world.

Other works have previously focused on the analysis of SM data for the detection of earthquakes. Researchers in [25] and [26] had the goal of creating an early warning system (EWS) for the real-time detection of earthquakes and tornadoes in Japan. The proposed system is based on bayesian statistics and was able to timely detect 67.9% (53 of 78) of the earthquakes which occurred over two months. The system described in [25] and [26] only focuses on the event detection task.
USGS efforts towards the development of an earthquake detection system based solely on Twitter data are described in [12]. The proposed solution is evaluated with different settings according to the sensitivity of the event detection module. However, even with its best configuration the system could only detect 48 globally distributed earthquakes out of the 5,175 earthquakes reported during the same time window. USGS kept on working on the project and recently announced the official employment of a Twitter earthquake detection system named TED (Tweet Earthquake Dispatch). As explained by USGS, such detection system proved more responsive than those based on seismographs in regions where the number of seismographic stations is low[5].

Situational awareness during emergencies is the goal of the work described in [7] and [31]. The system performs event detection and authors propose some first solutions towards the task of the automatic damage assessment of disasters. Standard burst detection algorithms are employed for the event detection task. After the detection of an event the system mines the content of new tweets and outputs word-clouds to help users understand the nature of the detected emergency.

Other works related to the emergency management have studied communication patterns and information diffusion in OSNs in the aftermath of disasters. The study described in [9] shows how OSNs can be used as a reliable source of spatio-temporal information. Researchers investigate Twitter activity during a major forest fire in the south of France in July 2009. Other similar studies have been carried out in [16], [21], [11] showing the importance of OSNs in the communications after a disaster. These studies encourage the exploitation of this information and motivate further research in this field such as the one that we are proposing.

In [3] we designed and developed a system for the detection and the assessment of the consequences of earthquakes. The proposed solution employs data mining and natural language processing techniques on SM data to enhance situational awareness after seismic events. The system proved highly effective over a seven months long testing period, however it still lacks some important features. The work we are proposing in this paper complements well with novel earthquake emergency management systems like the one detailed in [3]. It can enhance actual damage assessment procedures and deliver critical information to decision makers in the aftermath of strong earthquakes.

As highlighted by the works surveyed in this section, social media emergency management is a relatively young field of study. We believe this is one of the reasons why the automatic assessment of the consequences of disasters is a largely unexplored line of research. In addition, preliminary studies such as the ones proposed in [25], [7], [3] are prerequisites for a deeper knowledge of the role of social media in emergency situations, which will eventually lead to the employment of automatic damage assessment techniques in novel emergency management systems.

2.2 Prediction with social media

Predictive models have long been trained on social media data to explain both real world and virtual world dynamics. Fields of application for such techniques range from syndromic surveillance, social network analysis and event forecast to sales prediction, product recommendation and many more.

The work discussed in [15] dates back to 2005 and was among the first studies to focus on the impact of web discussions and chatter towards the prediction of real world trends. Researchers exploited postings in blogs, media, and web pages to successfully predict book sales. Other subsequent works related to prediction of purchase behaviors are described in [5] and [32]. In particular, the latter study focused on the correlations between Facebook account data and eBay purchases. In [29], authors study the dynamics of time patterns related to content published on the web. Part of the work also focused on the prediction of new time patterns exploiting a small number of new observations. Works presented in [14] and [18] apply predictive modeling to social network analysis. The former thoroughly discusses the design and evaluation of an ordinary least squares (OLS) regression model for the estimation of tie strength on Facebook. The latter work employs a multilevel logistic regression approach and focuses on the impact of network structure in the breaking of ties.

Prediction of web content change is studied in [23] where authors analyze the impact of novel page features to the predictive power of content change models. The work discussed in [17] proposes a model based on dwell time to predict search satisfaction. Authors first employ a regression model to infer dwell time characteristics, then apply a binary classification for the prediction of search satisfaction. An application of LASSO regression for the prediction of real world phenomena is discussed in [20]. Authors perform two case studies on the prediction of (i) rainfall levels and (ii) influenza-like illness in the UK. Researchers in [1] focus on the forecast of future events by training autoregressive models over a corpus of news articles from the New York Times.

Works described in this section represent a small survey of the application of predictive models on social media data. Although the application context differs from ours, the encouraging results of these studies demonstrate the power of predictive models and social media and pave the way for the exploitation of such techniques in other research fields as well.

[5] http://www.livescience.com/45385-earthquake-alerts-from-twitter.html

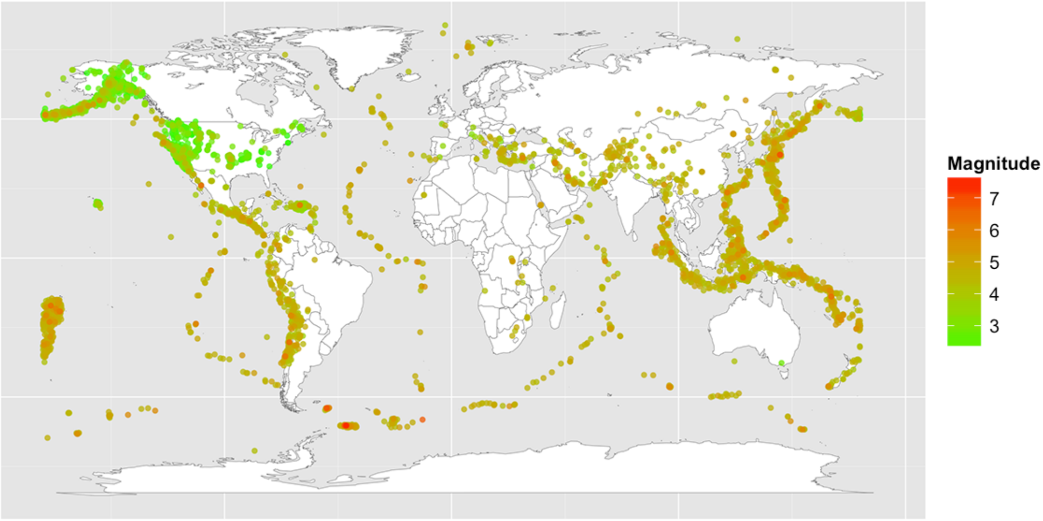

Figure 1: Geographical view of the 7,283 earthquakes covered by our study

3. THE DATASET

Our dataset is composed of (i) earthquake data, acquired with a semi-automatic procedure from USGS, and (ii) Twitter data. USGS earthquake data serves as our ground truth, while messages collected from Twitter are exploited to compute our earthquake features.

We started collecting earthquake-related tweets for a period of 90 days, spanning from October 18th 2013 to January 15th 2014. We then queried USGS for data about worldwide earthquakes occurred in the same time window. The exploitation of automatic crawling strategies for data collection gives us an advantage in terms of amount of data to use in our experiments compared to works such as [14] where data acquisition is performed manually.

3.1 USGS Earthquakes

We started collecting earthquake data directly from USGS's search web page[6] which we queried specifying time boundaries and a minimum magnitude of 2. As a result we obtained a list of 7,283 globally distributed earthquakes. For each of these earthquakes we further queried USGS for additional data.

Specifically, USGS performs many different analyses for every earthquake giving access to detailed information ranging from magnitude, epicenter, depth to intensity estimations, cities near the epicenter, etc. However, not every earthquake presents all these information. This is because USGS collects data about earthquakes striking anywhere in the world, but some calculations are computed only for seismic events which occur in areas covered by its seismic network. For instance, data about earthquakes having a magnitude lower then 4 is available almost only for earthquakes hitting US territory. As further explained in Section 5 we set up several different experimental settings to account for the diversity in earthquake data.

Intensity estimations are computed both at an aggregate level and at a local level. In the aggregate estimation a single intensity value is given for each earthquake. Instead, local estimations produce more values per earthquake which may differ between the various stricken locations. In this work we focus on the prediction of aggregate intensity estimations, as opposed to [6] where the focus is on local values. Intensity estimations, both at the aggregate and local level, are computed by the ShakeMap[7], Did You Feel It?[8] (DYFI?) and PAGER[9] systems. The ShakeMap system outputs intensity values based on empirical relationships that convert recorded or estimated accelerations in intensities [28]. The intensity estimation computed by ShakeMap outputs values according to the MMI (Modified Mercalli Intensity) 1 → 12 continuos scale and the aggregate MMI value is defined as follows.

Definition 1. The **MMI** value is the maximum estimated instrumental intensity for the event.

The DYFI? system outputs intensity values based on online survey reports [10]. The intensity estimation computed by DYFI? outputs values according to the CDI (Community Decimal Intensity) 1 → 10 continuos scale and the aggregate CDI value is defined as follows.

Definition 2. The **CDI** value is the maximum reported intensity for the event.

As further explained in Section 4 we exploit CDI and MMI values for earthquakes as the dependent variables in our OLS regression models. We do not exploit PAGER estimations for our predictive models since the PAGER system builds on ShakeMap estimations.

Figure 1 displays the epicenter and magnitude of each of the 7,283 earthquakes contained in our dataset. As shown in figure, a limited number of earthquakes occurred in sea regions. Such seismic events obviously didn't cause any damage and therefore we removed them from the training-set employed to build our statistical models. We discarded those earthquakes occurred more than 50km away from the coast and

[6] http://earthquake.usgs.gov/earthquakes/search/

[7] http://earthquake.usgs.gov/research/shakemap/

[8] http://earthquake.usgs.gov/research/dyfi/

[9] http://earthquake.usgs.gov/research/pager/

Group	*Countries*	*Language* (ϵ^{Lang})
CNA: Central and North America	United States, British and U.S. Virgin Islands, Canada, etc.	English
CSA: Central and South America	Puerto Rico, Mexico, Chile, Peru, Argentina, etc.	Spanish
ROW: Rest of the World	Indonesia, Japan, Philippines, New Zealand, Taiwan, etc.	English, Spanish

Table 2: Earthquake to language association

those more than 50km deep from the surface, with the exception of seismic events having a particularly high magnitude.

3.2 Twitter Data

We exploited the Twitter Streaming API[10] for Twitter data acquisition. The Streaming API gives access to a global stream of messages, optionally filtered by search keywords. We set search keywords so as to maximize the trade-off between completeness and specificity in earthquake-related collected messages. We based the keyword selection process on our previous studies in this field [2, 3] and on other related works such as [12, 25, 26].

For our study we analyze tweets in the two most widespread languages: English and Spanish. This allows us to thoroughly evaluate the impact of earthquakes hitting the American region while still giving us the chance to experiment with earthquakes from the rest of the world and specifically the Asian region. Although expecting decreased performances for models working on earthquakes outside the American region, we believe it is interesting to evaluate the possibility to predict worldwide seismic intensity with only messages in English and Spanish languages.

Table 1 summarizes data collected during our 90 days long data acquisition phase.

	English	*Spanish*	*Total*
Tweets	3,362,873	1,623,248	4,986,121
Tweet Entities	3,785,805	1,920,007	5,705,812
Users	2,000,491	897,551	2,898,042

Table 1: Twitter dataset statistics

4. METHOD

Building on the work discussed so far, we want to address the following research issues:

- Can social media data predict earthquake intensity?
- What are the limitations of such predictive models based solely on social media data?

We exploit data collected from Twitter to design features associated with earthquakes. These features, together with their mutual interactions, are then employed as predictive variables in the proposed statistical models. To compute features we must first link each earthquake ϵ_j with a corresponding set of tweets T^{ϵ_j}. We based this link on both temporal and geographical information about the earthquakes. As summarized in Table 2, geographical information is exploited to create three distinct groups of earthquakes. Only tweets τ with a given language τ^{Lang} are used to compute features for earthquakes of a specific group.

$$T_{lang}^{\epsilon_j} = \{\tau : \tau^{Lang} = \epsilon_j^{Lang}\} \tag{1}$$

We further exploited temporal information by only considering tweets published during a given time window after the occurrence time of the earthquake. Leveraging previous experiences [2] we modeled the time window length $\epsilon^{\Delta t}$ according to the magnitude ϵ^{Mag} of the earthquake, so that stronger earthquakes received a wider time window than weaker ones.

$$\epsilon^{\Delta t} \propto \epsilon^{Mag} \tag{2}$$

$$T_{time}^{\epsilon_j} = \{\tau : \epsilon_j^{Time} < \tau^{Time} \leq \epsilon_j^{Time} + \epsilon_j^{\Delta t}\} \tag{3}$$

This is because stronger earthquakes are likely to generate more messages and data which is critical towards the damage assessment is likely to be shared over a longer time. As thoroughly explained in the remainder of this section, our features are then normalized over the length of the time window to avoid introducing a bias.

The final set of tweets T^{ϵ_j} associated to an earthquake ϵ_j is computed as in the following:

$$T^{\epsilon_j} = T_{lang}^{\epsilon_j} \cap T_{time}^{\epsilon_j} \tag{4}$$

Together with the set of tweets, for our features we also exploit a set of user accounts A^{ϵ_j} associated to an earthquake ϵ_j which is computed selecting all the accounts α that posted at least one tweet among the ones belonging to the set T^{ϵ_j}.

$$A^{\epsilon_j} = \{\alpha : \alpha \leftarrow \tau \in T^{\epsilon_j}\} \tag{5}$$

4.1 Feature Extraction

In order to characterize earthquakes we designed a large set of features to act as potential predictors of earthquake intensity. Our 45 distinct features fall into four different categories according to the nature of the information they aim to capture. This categorization is also exploited to assess the contribution and predictive power of the single classes of features as well as their combined effects. During the design process we also picked features which fit well with Twitter's dynamics and that could still be carried over to other social medias.

Features that are influenced by the length of the observation time are indicated as F_i^* and are normalized to account for the different time windows associated to every earthquake. Given the time window length $\epsilon_j^{\Delta t}$ of the j$^{\text{th}}$ earthquake ϵ, normalized features are defined as:

$$F_{i,j} = \frac{F_{i,j}^*}{\epsilon_j^{\Delta t}} \tag{6}$$

Table 3 gives formal definitions for a subset of all the extracted features.

[10] https://dev.twitter.com/docs/api/streaming

#	*Label*	*Definition*
Profile Features		
F_1^*	distinct_acc	$\lvert A \rvert$
F_2^*	distinct_acc_in_same_country	$\lvert A_{same} \rvert$, where $A_{same} = \{\alpha : \alpha \in A \wedge \alpha^{Country} = \epsilon^{Country}\}$
F_5	avg_acc_dist_from_epicenter	$\overline{D}$, where $D = haversine(\alpha^{Loc}, \epsilon^{Loc}) \; \forall \alpha \in A$
Tweet Features		
F_8^*	productivity	$\lvert T \rvert$
F_{10}^*	earthquake_hashtag_count	$\lvert T_{ear} \rvert$, where $T_{ear} = \{\tau : \tau \in T \wedge \tau$ contains "$\#earthquake$"$\}$
F_{12}^*	location_hashtag_count	$\lvert T_{loc} \rvert$, where $T_{loc} = \{\tau : \tau \in T \wedge \tau$ contains "$\#\epsilon^{Loc}$"$\}$
F_{15}	avg_character_count	$\overline{\lvert K \rvert}$, where $K = \{\zeta : \zeta \leftarrow \tau \in T\}$
Time Features		
F_{22}	avg_time_between_msgs	$\overline{\tau_{i+1}^{Time} - \tau_i^{Time}}$, for $i = 1, \ldots, \lvert T \rvert - 1$
Linguistic Features		
$F_{28}^*, \ldots, F_{42}^*$	proto_word_count	$\overline{\lvert W \rvert}$, where $W = \{\omega : \omega \leftarrow \tau \in T \wedge \omega = \eta_k\}$

Table 3: Formal definition of a sample of features

Profile features. Most OSNs allow users to provide basic information about themselves such as: user name, location, short bio, etc. Statistics reported in previous works [8], [22] showed that profile information is often unreliable and does not contain enough high-quality data to be employed in complex data mining tasks. For the sake of experimentation we aimed at verifying these claims and we implemented a few basic profile features, mainly based on the account location α^{Loc}: number of distinct accounts (F_1^*); number of distinct accounts that tweeted from the same country of the earthquake (F_2^*); number of distinct accounts that tweeted from a neighbour country (F_3^*); number of distinct countries derived from accounts locations (F_4^*); average (F_5), minimum (F_6) and variance (F_7) of accounts distances from the epicenter. These features can help understand whether a relation exists between earthquake intensity and the geographic distribution of reports around the epicenter.
We exploited the Geonames[11] database for the conversion from the user location string to the geographic coordinates and geographic distances are computed by means of the Haversine formula.

Tweet features. To compute tweet features we exploited tweet entities, tweet metadata and the structure of the messages. Tweet entities comprehend urls, mentions, hashtags as well as photos attached to a tweet. Among these information we found hashtags to be particularly useful for our goal. Eyewitness reports of earthquakes usually carry the #earthquake or #quake hashtag and the same also applies to Spanish language messages. Other hashtags related to the location of the epicenter or the country hit by the earthquake are sometimes used as well. Among tweet metadata we exploited the geographic information of geolocated tweets. Regarding the structure of tweets we tried to measure the complexity of the messages. As highlighted in [2], the emotional state of users after an emergency tends to be reflected by the length, use of punctuation and number of capital letters in the messages shared. Thus we defined the following set of features that takes into account the results of the previous analyses: total number of tweets (F_8^*); ratio between the total number of tweets and the average number of tweets shared during the same time of the day for all other days (F_9^*); number of #earthquake (or similar) hashtags (F_{10}^*); number of hashtags with the name of the country hit by the earthquake (F_{11}^*); number of hashtags with the name of the location hit by the earthquake (F_{12}^*); average (F_{13}) and variance (F_{14}) of the total number of words ω among messages; average (F_{15}) and variance (F_{16}) of the total number of characters ζ among messages; average ratio between number of capital letters and total number of characters in messages (F_{17}); average ratio between number of punctuation characters and total number of characters in messages (F_{18}); average (F_{19}), minimum (F_{20}) and variance (F_{21}) of tweets distances from the epicenter.

Time features. Time features aim at grasping the bursty nature of emergency reports. Other previous studies have exploited bursty characteristics of message streams for the tasks of topic or event detection [19], [7, 31], [3]. Here we want to evaluate whether the quantification of such characteristics contributes to the prediction of the intensity of an emergency. Therefore we designed the following set of features, based on the publication time τ^{Time} of messages: average time delay between one message and the next one (F_{22}); average (F_{23}), minimum (F_{24}), maximum (F_{25}) and variance (F_{26}) in the number of messages per minute; longest streak of messages having a maximum delay of 5 seconds between one another (F_{27}).

Linguistic features. Many recent works such as [22] and [14] demonstrated the power of linguistic features towards classification and prediction tasks. In literature, bag-of-words techniques are opposed to approaches based on automatically bootstrapped prototypical words (or proto-words). Proto-words represent typical expressions that characterize a specific class of users. Here we adapt the proto-words algorithm originally proposed in [22] and employed in a user classification task. Our goal is to extract prototypical ex-

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

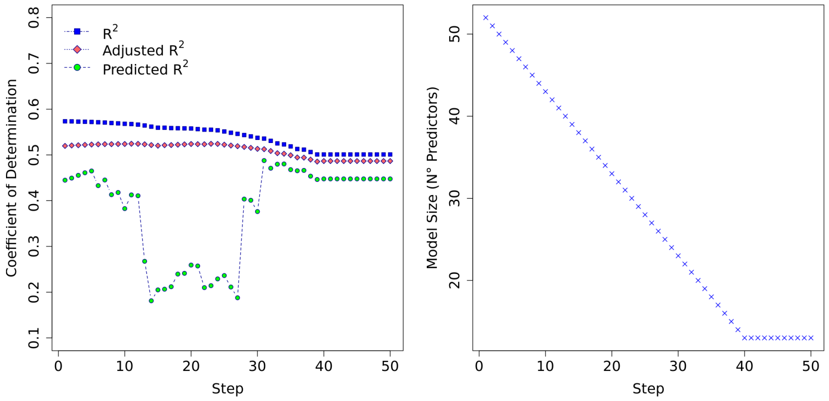

Figure 2: Simulation results for CDI ∈ [2, 10] and CNA earthquakes over 50 steps

pressions used in reports related to severe earthquakes. We are confident that employing such proto-words as predictors in our models will greatly increase the accuracy of earthquake intensity estimations. Given $|C|$ classes of events, each class $c_i \in C$ is represented by a set of seed events.

$$S^i = \{\epsilon_1^i, \ \dots \ , \epsilon_n^i\} \tag{7}$$

We compute frequencies for every unigram η_k found in a message τ associated to one of the seed events ϵ^i.

$$T^i = T^{\epsilon_1^i} \cup \ \cdots \ \cup T^{\epsilon_n^i} \tag{8}$$

$$T_{\eta_k}^i = \{\tau : \tau \in T^i \wedge \tau \ contains \text{ ``}\eta_k\text{''}\} \tag{9}$$

$$\text{seedfreq}(\eta_k, c_i) = \frac{\left|T_{\eta_k}^i\right|}{|T^i|} \tag{10}$$

The term seedfreq(η_k, c_i) represents the normalized frequency of the unigram η_k for the class c_i. The normalization term $\left|T^i\right|$ is necessary to account for the different cardinalities among sets of messages T^i of the different classes. The score of the unigram η_k for the class c_i is then computed as in the following:

$$\text{score}(\eta_k, c_i) = \frac{\text{seedfreq}(\eta_k, c_i)}{\sum_j \text{seedfreq}(\eta_k, c_j)} \tag{11}$$

This score indicates how much an unigram η_k is representative of the class c_i. For our experiments we chose 3 classes of events: strong earthquakes that caused damages and casualties (STR), moderate earthquakes widely felt by the population but without severe consequences (MOD) and light earthquakes felt only by a small number of social sensors (LIG). We picked the top 10 unigrams from the STR class as predictors, together with the top 5 unigrams from the MOD class. Unigrams of the LIG class are not directly exploited to compute features but instead serve as contrast terms to highlight typical expressions of the other classes. This resulted in 10 features computed as the number of times an unigram of the STR class was used in the earthquake reports ($F_{28}^*, \ \dots \ , F_{37}^*$), plus 5 features computed the same way for unigrams of the MOD class ($F_{38}^*, \ \dots \ , F_{42}^*$). We also added aggregate features as the total number of unigrams of the STR class used in reports (F_{43}^*); the total number of unigrams of the MOD class used in reports (F_{44}^*); and the total number of unigrams from both the STR and MOD classes (F_{45}^*).

4.2 Earthquake Intensity Model

We modeled earthquake intensity as a linear combination of our predictive variables, plus terms for pairwise interactions:

$$y_i = \beta_0 + \sum_{j=1}^{n} \beta_j F_{j,i} + \gamma I_i + \varepsilon_i \tag{12}$$

We adopted a linear model over more complex ones since we are not only interested in predictive power, but we also focus on the relations between our predictive variables and earthquake intensity. Although slightly increasing model complexity, pairwise interactions are commonly included in predictive models and proved to significantly boost predictive power without impairing model interpretability [14] [13]. In the definition of our model y_i represents the intensity of the i$^{\text{th}}$ earthquake; β_0 is the intercept term of our linear model; β_j are the coefficients of our predictors $F_{j,i}$; γ is the coefficient vector of the interaction terms; I_i is the vector of the interactions and ε_i represents the error term.

Because of the large number of regressors together with their pairwise interactions, we employed feature selection techniques to include in our model only the most influential predictors. We started from a model trained exploiting all the 45 predictive variables presented in Section 4.1 and we ran a stepwise model selection algorithm to chose which features to add or remove from the models. During the forward steps the starting model is expanded with the addition of the most influential interactions. In the backward steps, least relevant variables are removed from the model. Two widespread criteria for model selection are the Akaike's information criterion (AIC) and Schwarz's Bayesian information criterion (BIC). We experimented with both criteria and decided to employ BIC in our procedure for the sake of model parsimony. In fact BIC penalizes the number of parameters more strongly than AIC does [30]. In addition, we ran a series of simulations to overcome the trade-off between predictive power and model complexity. Finally, being aware of the criticism raised with regards to stepwise model selection, we evaluated candidate models exploiting the domain specific knowledge of a seismologist.

In our simulations we evaluated the different model structures by means of the predicted residual sum of squares (PRESS) statistic [27]. The PRESS statistic is a form of leave-one-out cross-validation (LOOCV) used in regression

Model	R^2	*Adjusted* R^2	*Predicted* R^2	*MAE*	*MSE*	*n*	*p*	*p-value*
dep. var. $CDI \in [1, 10]$								
$CNA_{mag>2}$	0.4820	0.4637	0.4125	0.66	0.67	734	25	$< 2.2 \times 10^{-16}$
$CSA_{mag>4}$	0.7769	0.7195	0.5980	0.38	0.24	182	37	$< 2.2 \times 10^{-16}$
$ROW_{mag>4}$	0.5917	0.5551	0.5358	0.47	0.52	147	12	$< 2.2 \times 10^{-16}$
dep. var. $CDI \in [2, 10]$								
$CNA_{mag>2}$	0.5357	0.5125	0.4877	0.46	0.33	465	22	$< 2.2 \times 10^{-16}$
dep. var. $MMI \in [1, 12]$								
$ALL_{mag>2}$	0.6074	0.5202	0.2286	0.45	0.42	89	16	2.69×10^{-9}

Table 4: Earthquake intensity prediction results

analysis to measure the accuracy of a model versus a sample of observations that were not themselves used to train the model. Therefore for each model structure generated by the stepwise algorithm we computed the PRESS statistic, the PRESS residuals and the Predicted R^2 values:

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - \hat{y}_{i,-i})^2 \tag{13}$$

$$\text{PRESS}_{resid} = \{y_1 - \hat{y}_{1,-1},\ \dots\ , y_n - \hat{y}_{n,-n}\} \tag{14}$$

$$\text{Predicted } R^2 = \text{corr}(Y, \hat{Y}_{-i})^2 \tag{15}$$

Models that are over-parameterised (overfitted) would tend to give small residuals for observations of the training-set but large residuals for unseen observations. It is possible to avoid overfitting and assess a model's ability to generalize by analyzing R^2 and Adjusted R^2 values versus Predicted R^2 values. In contrast to R^2 and Adjusted R^2, Predicted R^2 values drop as an overfitted model looses its ability to generalize.

Our simulation procedure builds on these considerations and also plots intermediate results which are useful to evaluate models performances during the iterative selection process. Figure 2 shows evaluation plots related to the backward model selection part of the algorithm and resulting from a simulation targeting the estimation of CDI intensity values $\in [2, 10]$ over the earthquakes of the Central and North American group (CNA). Prior to step 31 Predicted R^2 values show massive fluctuations and a significative difference in comparison to the stable trends of R^2 and Adjusted R^2. This should raise concerns towards overfitting and the model's ability to generalize. As model complexity is reduced Predicted R^2 values also manifest a stable trend, comparable to those of R^2 and Adjusted R^2. This simulation seems to suggest the model generated at step 31 as a good candidate for the estimation of CDI values over CNA earthquakes. All the experimental settings proposed in Section 5 have been evaluated with this procedure. In addition, candidate models resulting from simulations have been further diagnosed to check models assumptions and assess their statistical significance.

Although widely used in many research fields such as biostatistics, to the best of our knowledge the PRESS statistic and the Predicted R^2 are almost never employed for predictive analyses on social media data.

5. RESULTS

We set up 5 different experiments building on the geographical grouping of earthquakes as reported in Table 2 and on USGS earthquake intensity estimations which we described in Section 3.1. Intensity estimations based on online surveys (CDI) are the most frequent among the collected earthquakes. We are interested in evaluating the relation between CDI values and intensity estimations with social media data. We started with 3 experiments aimed at mapping the whole scale of 1 → 10 CDI values. Each experiment is based on earthquakes from one of the three geographical regions and resulted in the top 3 models described in Table 4. Comparing results from these 3 models can help understand to what extent predictive power is affected by the differences in earthquake magnitude and by our language assumptions. Specifically, as anticipated in Section 3.1, 98.5% of the earthquakes of the CNA group have a magnitude value between 2 and 4, as underlined by the $CNA_{mag>2}$ label. Earthquakes in the CSA and ROW groups instead have magnitude values almost always higher than 4, hence the labels $CSA_{mag>4}$ and $ROW_{mag>4}$. In other words, earthquakes of the CNA group are almost completely non-overlapping with the ones in the CSA and ROW groups, with regards to magnitude. All the models proposed in Table 4 show p-values $\ll 0.001$ assessing their statistical significance. Among all the proposed models, the CSA one achieves the best performances with a Predicted R^2 close to 0.6 , MSE as low as 0.24 and MAE = 0.38 on a continuos 1 → 10 scale.

The substantially lower performance of the model trained on CNA data is probably to be imputed to the much lower severity of the earthquakes in the 2 → 4 magnitude range. These results seem to suggest that the prediction of earthquake intensity is more difficult for low magnitude earthquakes compared to high magnitude ones. This can be intuitively explained by considering that earthquakes having a magnitude < 3 are felt by a very limited number of persons thus resulting in a reduced number of shared messages.

As anticipated in Section 3.2, constraining the analysis of earthquakes outside the American region to messages only in English and Spanish language resulted in a slightly worse intensity estimation. It is worth noticing however that despite a 0.18 reduction in R^2, the ROW model exhibits a MAE value of 0.47 which still reflects accurate predictions. Since a considerable number of earthquakes in the ROW group occurred in Japan and Taiwan, we would expect a better fit by extending the analyses to also include messages in Japanese and Chinese languages.

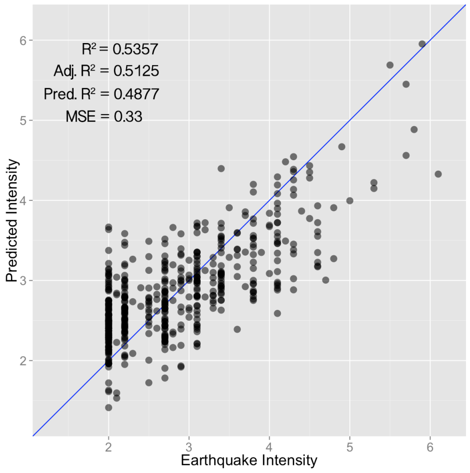

Figure 3: Regression plot for the CNA model trained over earthquakes with CDI $\in [2, 10]$

In addition to these experiments, we were also interested in evaluating the prediction of CDI values over a reduced $2 \rightarrow 10$ scale. The technical motivation for this experiment is based on an analysis of the distribution of CDI estimations in that the 1 values form a distinct cluster of points, separated from the remaining values. Specifically, CDI values are uniformly distributed over the $[2, 10]$ range, while there are no observations in the $(1, 2)$ interval. This is because a CDI value of 1 means that the earthquake was not felt by anyone and actual intensity estimations start from a CDI value of 2 [10]. From a theoretical point of view the CDI value of 1 encodes a different kind of information than the remaining values. Furthermore, while online surveys also store information about earthquakes not felt by the population, Twitter users almost never share messages about earthquakes they did not feel. The low number of USGS estimations for earthquakes of the CSA and ROW regions having a CDI ≥ 2 (49 and 26 respectively) did not allow to adequately experiment in these areas. Instead, we trained a new model with the CNA earthquakes reducing the interval of the CDI dependent variable to $[2, 10]$. Noticeably the new model produces overall better predictions, with MSE $=$ 0.33 and MAE $=$ 0.46 in contrast to MSE $=$ 0.67 and MAE $=$ 0.66 for the model trained on the whole CDI $1 \rightarrow 10$ interval. Figure 3 shows the regression plot for this model: despite the good fit the model exhibits a slight tendency to underestimate.

Among all the earthquakes of our dataset, USGS computed MMI intensity estimations only for a small portion of them. Therefore we did not have enough observations to propose different models based on geographical areas. Instead, we set up a single experiment comprising all 89 earthquakes carrying an MMI estimation, disregarding the location of the epicenter. Results for this experiment are labeled $ALL_{mag>2}$ in Table 4. Although exhibiting encouraging R^2, Adjusted R^2, MAE and MSE values, the Predicted R^2 is substantially lower which seems to indicate an overfitted model. The inconclusive results with MMI estimations ask for subsequent experimentations on more data.

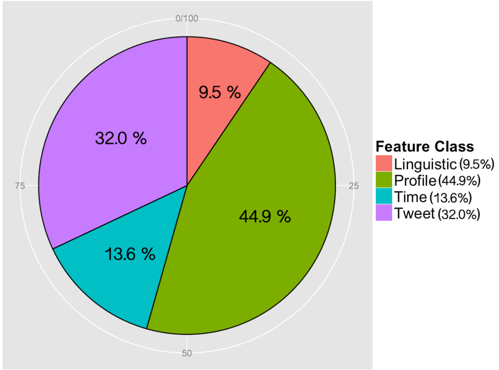

Figure 4: Features classes contribution towards predictive power

Figure 4 represents the contribution of the different classes of features towards the predictive power of the CNA model trained on the limited CDI $2 \rightarrow 10$ scale. The weight of each feature class is computed by summing the absolute values of the β coefficients of its features. As shown, there is no dominant class in the model and all 4 classes give a significative contribution to the prediction. The profile and tweet classes alone provide 76.9% of the predictive power. This somewhat contrasts with results presented in [8] and [22] claiming the unreliability of profile features for data mining tasks. Linguistic features exhibit the smaller contribution, anyway they appear in the majority of the interaction terms. Despite their relatively small direct impact on predictive power, in our experiments they play an important modulating role for the other parameters. Among the 4 classes, time features seem to be the most marginal ones because of their relatively light direct contribution and limited presence in interaction terms.

6. CONCLUSIONS AND FUTURE WORK

In this study we shed light on the possibility to exploit social media data towards the prediction of earthquake intensity. We leveraged experiences in previous works and demonstrated the impact of Twitter earthquake reports on intensity estimations. The proposed models build on a large dataset and exploit 45 distinct features belonging to 4 different classes.

Results discussed in Section 5 are overall encouraging and show significative correlations between the messages shared in social media and the consequences of worldwide earthquakes. This correlation is particularly strong for intensity estimations based on online survey data (CDI). Instead, the prediction of intensity estimations based on empirical relationships (MMI) is still an open issue which requires further investigation.

We achieved a Predicted R^2 value close to 0.6 with an MSE of 0.24 over a $1 \rightarrow 10$ scale for the best performing model. This is a remarkable performance in comparison to the only work available in literature addressing the same issue [6] and to other comparable works such as [14]. Discussed results seem to favor the employment of predictive techniques in novel earthquake emergency management systems.

It is worth noticing that the proposed results are based on unfiltered social media data, which is known to be particu-

larly noisy. We believe that employing data filtering techniques could further improve prediction accuracy by removing most of the noise. Therefore, such techniques should always be adopted when applying predictive models in deployed emergency management systems.
We are confident that the proposed study addresses fundamental challenges towards the understanding and the exploitation of social media data for the enhancement of modern emergency management procedures.

In the future we will apply our predictive models to novel earthquake emergency management systems, such as the one described in [3]. We also plan to experiment with more complex regression models and we look forward to extending our analyses to the local earthquake intensity estimations.

7. REFERENCES

[1] G. Amodeo, R. Blanco, and U. Brefeld. Hybrid models for future event prediction. In *CIKM'11*. ACM.

[2] M. Avvenuti, S. Cresci, M. La Polla, A. Marchetti, and M. Tesconi. Earthquake emergency management by social sensing. In *PERCOM Workshops*. IEEE, 2014.

[3] M. Avvenuti, S. Cresci, A. Marchetti, C. Meletti, and M. Tesconi. Ears (earthquake alert and report system): a real time decision support system for earthquake crisis management. In *KDD'14*. ACM.

[4] J. P. Bagrow, D. Wang, and A.-L. Barabasi. Collective response of human populations to large-scale emergencies. *PloS one*, 2011.

[5] R. Bhatt, V. Chaoji, and R. Parekh. Predicting product adoption in large-scale social networks. In *CIKM'10*. ACM.

[6] L. Burks, M. Miller, and R. Zadeh. Rapid estimate of ground shaking intensity by combining simple earthquake characteristics with tweets. In *Tenth US National Conference on Earthquake Engineering*, 2014.

[7] M. A. Cameron, R. Power, B. Robinson, and J. Yin. Emergency situation awareness from twitter for crisis management. In *WWW'12*. ACM.

[8] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *CIKM'10*. ACM.

[9] B. De Longueville, R. S. Smith, and G. Luraschi. Omg, from here, i can see the flames!: a use case of mining location based social networks to acquire spatio-temporal data on forest fires. In *International Workshop on Location Based Social Networks*. ACM, 2009.

[10] L. Dengler and J. Dewey. An intensity survey of households affected by the northridge, california, earthquake of 17 january 1994. *Bulletin of the Seismological Society of America*, 1998.

[11] P. Earle, M. Guy, R. Buckmaster, C. Ostrum, S. Horvath, and A. Vaughan. Omg earthquake! can twitter improve earthquake response? *Seismological Research Letters*, 2010.

[12] P. S. Earle, D. C. Bowden, and M. Guy. Twitter earthquake detection: earthquake monitoring in a social world. *Annals of Geophysics*, 2012.

[13] D. Gergle, R. E. Kraut, and S. R. Fussell. The impact of delayed visual feedback on collaborative performance. In *CHI'06*. ACM.

[14] E. Gilbert and K. Karahalios. Predicting tie strength with social media. In *CHI'09*. ACM.

[15] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins. The predictive power of online chatter. In *KDD'05*. ACM.

[16] A. L. Hughes and L. Palen. Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, 2009.

[17] Y. Kim, A. Hassan, R. W. White, and I. Zitouni. Modeling dwell time to predict click-level satisfaction. In *WSDM'14*. ACM.

[18] F. Kivran-Swaine, P. Govindan, and M. Naaman. The impact of network structure on breaking ties in online social networks: unfollowing on twitter. In *CHI'11*. ACM.

[19] J. Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 2003.

[20] V. Lampos and N. Cristianini. Nowcasting events from the social web with statistical learning. *Transactions on Intelligent Systems and Technology (TIST)*, 2012.

[21] A. Murakami and T. Nasukawa. Tweeting about the tsunami?: mining twitter for information on the tohoku earthquake and tsunami. In *WWW'12*. ACM.

[22] M. Pennacchiotti and A.-M. Popescu. Democrats, republicans and starbucks afficionados: user classification in twitter. In *KDD'11*. ACM.

[23] K. Radinsky and P. N. Bennett. Predicting content change on the web. In *WSDM'13*. ACM.

[24] A. Rosi, M. Mamei, F. Zambonelli, S. Dobson, G. Stevenson, and J. Ye. Social sensors and pervasive services: Approaches and perspectives. In *PERCOM Workshops*. IEEE, 2011.

[25] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *WWW'10*. ACM.

[26] T. Sakaki, M. Okazaki, and Y. Matsuo. Tweet analysis for real-time event detection and earthquake reporting system development. *Transactions on Knowledge and Data Engineering (TKDE)*, 2013.

[27] T. Tarpey. A note on the prediction sum of squares statistic for restricted least squares. *The American Statistician*, 2000.

[28] C. Wills, M. Petersen, W. Bryant, M. Reichle, G. Saucedo, S. Tan, G. Taylor, and J. Treiman. A site-conditions map for california based on geology and shear-wave velocity. *Bulletin of the Seismological Society of America*, 2000.

[29] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *WSDM'11*. ACM.

[30] Y. Yang. Can the strengths of aic and bic be shared? a conflict between model identification and regression estimation. *Biometrika*, 2005.

[31] J. Yin, A. Lampert, M. Cameron, B. Robinson, and R. Power. Using social media to enhance emergency situation awareness. *IEEE Intelligent Systems*, 2012.

[32] Y. Zhang and M. Pennacchiotti. Predicting purchase behaviors from social media. In *WWW'13*. ACM.

[33] A. Zhou, W. Qian, and H. Ma. Social media data analysis for revealing collective behaviors. In *KDD'12*. ACM.