

## Article

# Sensing Earthquake Disaster Information: A Named Entity Recognition Approach using Twitter Collaborative Data

Aldo Hernandez-Suarez<sup>1</sup> , Gabriel Sanchez-Perez<sup>1</sup>, Karina Toscano-Medina<sup>1</sup>, Hector Perez-Meana<sup>1,\*</sup> , Jose Portillo-Portillo<sup>1</sup> and Victor Sanchez<sup>2</sup>

<sup>1</sup> Instituto Politecnico Nacional, ESIME Culhuacan, Mexico City 04440, Mexico; ahernandezs1325@alumno.ipn.mx (A.H.-S.); (G.S.-P.); (K.T.-M.); (H.P.-M.); (J.P.-P.)

<sup>2</sup> Department of Computer Science, University of Warwick, Coventry CV4 7AL, UK; v.f.Sanchez-Silva@warwick.ac.uk

\* Correspondence: hmperezm@ipn.mx; Tel.: +52-55-5624-2000

**Abstract:** In recent years, online social networks have received important consideration in spatial modelling fields given the critical information that can be extracted from them for events in real time; one of the most latent issues is that regarding various natural disasters such as earthquakes. Although it is possible to retrieve data from these social networks with embedded geographic information provided by GPS, in many cases this is not possible. An alternative solution is to reconstruct specific locations using probabilistic language models, more specifically those based on Name Entity Recognition (NER), which extracts names from a user's description about an event occurring in a specific place (e.g., a collapsed building on a specific avenue). In this work, we present a methodology to use twitter as a social sensor system for disasters. The methodology scores NER locations with a kernel density estimation function for different subtopics originating from a natural disaster and that maps them into a geographic space is proposed. The proposed methodology is evaluated with tweets related to the 2017 earthquake in Mexico.

**Keywords:** statistical analysis; data mining

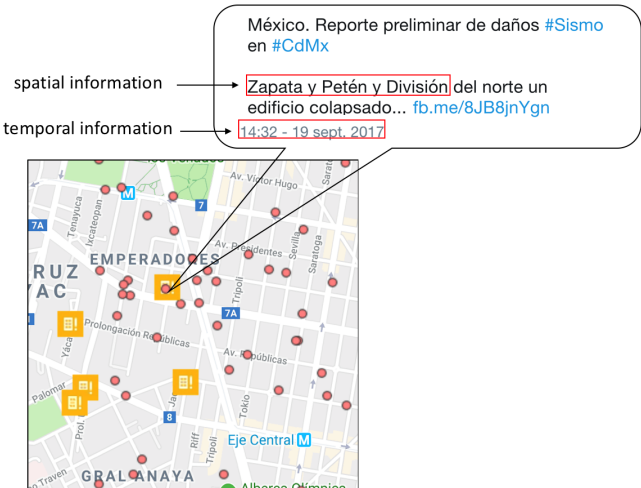
## 1. Introduction

Although there are sensors that can detect in advance various natural disasters such as earthquakes (Mexico City's Alarm System senses quakes coming from southern states) [1], the devastating consequences in urban areas are still severe and uncountable. The relief efforts after such disasters can inspire the participation of civil society, which together with rescue teams and civil protection and security institutions, can help to inform, rescue, and provide restoration as quickly as possible. The active participation of civil society in the aftermath can not only strengthen the society itself but also improve the trust in the information obtained from non-traditional sources [2,3]. For example, thanks to the widespread of wireless communication networks and mobile technologies, the dissemination of information, such as news and reports, has served as a vital way to contact aid services and make appropriate decisions in a timely and more flexible manner [4]. Similarly, other data from personal mobile devices have played an important role in disaster relief efforts. As evidence, in the 2010 earthquake in Haiti, the use of instant messages sent by civilians in different locations facilitated the report of trapped people and the provision of emergency medical assistance and basic needs such as food, water, and shelter [5]. Personal mobile phones can also be used by survivors to send messages about their current status to relatives and the community. These messages can then be forwarded to rescue teams. Figure 1 illustrates an example of a survivor of an earthquake using their mobile phone to communicate with their relatives.



**Figure 1.** Through a messaging system, an earthquake survivor describes their situation inside a collapsed building. The messages translated to English are: *My love. The roof fell. We are trapped. My love I love you. I love you so much. We are on the 4th floor. Near the emergency stair. There are 4 of us. My love are you ok?*

Thanks to the ubiquitous availability of mobile Internet [7], personal mobile devices can be linked to online social networks, providing a synchronisation between applications (e.g. between Twitter and Facebook), which allows to communicate behaviours and activities in real-time across various platforms [6]. Since online social networks are applications based on the Internet protocol and enable the creation and exchange of user-generated content [8], information extracted from these networks can include temporal and spatial data related to different events [9]. The information extracted from these networks can then be represented as geo-referenced patterns that establish relationships between the publication and the geographical and temporal characteristics of the publishing entity. As an example, a post (tweet) on Twitter with temporal and spatial data is shown in Figure 2.



**Figure 2.** A tweet providing the location (spatial information) of a collapsed building along with a time stamp (temporal information) one day after the 2017 earthquake in Mexico City.

The behaviour of users on social networks and their posted information can be considered as a social sensor system since the data generated on a large scale closely resembles those acquired by traditional sensor systems [10,11]. Below are some characteristics that reinforce the notion of social networks as being sensor systems [12,13]:

- **Sensor operation:** Sensors acquire data related to various events as a result of real-time observations. Mobile phones are equipped with various sensors, such as cameras, that can acquire also acuaire data related to various events.[11,14,15].
- **Processing of sensed data:** When information is processed, metadata such as geographic information can be included if navigation systems, e.g. GPS, are available. If the information is then published on a social network, users have the ability to comment and retransmit it along with the geographic information [16].

Twitter, one of the most popular social networks, can be used as an alternative communication channel for exchanging information related to disasters such as fires, floods, hurricanes, and earthquakes. In this work, we then propose a methodology to use Twitter as a social sensor system for disasters, by exploiting the data resulting from observations and experiences posted by the users. Within this context, the collected data is expected to be in textual form (sentences of maximum 140 characters, referred to as tweets) and with well-defined geographic information (spatial attributes). However, it has been shown that only a very small percentage of users use navigation systems such as GPS to reference their information [17]. Some studies propose estimating the event location by exploting some features available in Twitter , such as searching for updates related to the event within a known geographical region [18]. Other studies propose to detect and group textual patterns to approximate the location of the event in a given time period [19], or to use geographic coders to approximate the geographic coordinates by accounting for the frequency of occurrence of the event in a well-defined spatial region [20]. In this work, our proposed methodology is based on the fact that the location of the event can be implicitly described in a tweet. A probabilistic language model known as Named Entity Recognition (NER) is then computed [21,22] to extract entities from the text that can be used to calculate spatial features, such as latitude and longitude coordinates. To detect these entities, the non-grammatical nature of tweets in Spanish as well as informal abbreviations and lexicons (for example, describing a locality using a hashtag) are accounted for. Temporal features are computed by clustering similar tweets related to an event [23]. The proposed methodology employs a Kernel Density Estimation (KDE) algorithm, a well-known statistical technique, to detect and monitor the hotspot dynamics of detected events [24].

## 2. Related Work

The detection of events related to natural disasters is the subject of recent research in several fields of sensors, natural language processing, and automatic and statistical learning. The main aim is to detect, monitor, and disseminate information in a timely manner with some degree of trust. This section describes some works related to this growing field of research. As described in [25], Twitter has been recently used as a platform for dissemination of diverse information related to various natural disasters such as wildfires [26], floods [25], hurricanes [27], and earthquakes [28], achieving situational awareness. In Table 1, works that employ data extracted from Twitter and other online social networks to sense events related to natural disasters are listed.

**Table 1.** Related work that contributes to natural disasters sensing using data extracted from Twitter and other online social networks.

Title	Description
Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors-2010	To detect a target event, this work classifies tweets based on features such as the keywords, and the number of words and their context. It then estimates a probabilistic spatio-temporal model to find the centre and the trajectory of the target event. To this end, each Twitter user is assumed to be a sensor and Kalman filtering and particle filtering is applied for location estimation with ubiquitous/pervasive computing. The authors claim that a 96% probability of correctly detecting an earthquake can be achieved by monitoring tweets [29]
Public health implications of social media use during natural disasters, environmental disasters, and other environmental concerns	This work analyses how social media can be used to disseminate information, predict data and provide early warnings within the context of environmental awareness and health promotion. The works also analyses how social media can be used as an indicator of public participation during environmental concerns. The authors have found evidence supporting social media as a useful surveillance tool during natural disasters, environmental disasters, and other environmental concerns. Public health officials can use social media to gain insight into public opinions and perceptions. Social media allows public health workers and emergency responders to act more quickly and efficiently during crises [30]
Real-Time Crisis Mapping of Natural Disasters Using Social Media	In this work, the authors propose a social media crisis mapping platform for natural disasters that uses statistical analysis with geoparsed real-time tweet data streams matched to locations from gazetteers, street maps, and volunteered geographic information (VGI). Geoparsing results are benchmarked against existing published work and evaluated across multilingual datasets. Two case studies compare five-day tweet crisis maps to official post-event impact assessment from the US National Geospatial Agency (NGA), compiled from verified satellite and aerial imagery sources [31]
Tweedr: Mining twitter to inform disaster response	In this paper, the authors introduce Tweedr, a Twitter-mining tool that extracts actionable information for disaster relief workers during natural disasters. The Tweedr pipeline consists of three main parts: classification, clustering and extraction. In the classification phase, they use classification methods (sLDA, SVM, and logistic regression) to identify tweets reporting damage or casualties. In the clustering phase, they use filters to merge tweets that are similar; and finally, in the extraction phase, they extract tokens and phrases that report specific information about different classes of infrastructure damage, types of damage, and casualties [32]
A linguistically-driven approach to cross-event damage assessment of natural disasters from social media messages	In this work, the authors focus on the analysis of Italian social media messages for disaster management. Their aim is to detect those messages conveying critical information for the damage assessment task. A main novelty of this study is the focus on out-of-domain and cross-event damage detection, and the investigation of the most relevant tweet-derived features for these tasks. They conduct different experiments by resorting to a wide set of linguistic features qualifying the lexical and grammatical structure of a text, as well as ad-hoc features specifically extracted for this task [33]
Combining machine learning topic models and spatio-temporal analysis of social media data for disaster footprint and damage assessment	The authors propose a crisis mapping system by analysing the textual content of disaster reports from a twofold perspective. A damage detection component employs an SVM classifier to detect mentions of damage among emergency reports. A novel geoparsing technique is proposed and used to perform message geolocation. They report on a case study to show how the information extracted through damage detection and message geolocation can be combined to produce accurate crisis maps. The crisis maps detect both highly and lightly damaged areas, thus opening up the possibility to prioritise rescue efforts where they are most needed [34]

Table 1. Cont.

Title	Description
From social sensor data to collective human behaviour patterns: Analysing and visualising spatio-temporal dynamics in urban environments.	This paper presents an approach to analyse social media posts to assess the footprint of and the damage caused by natural disasters by combining machine-learning techniques (Latent Dirichlet Allocation) for semantic information extraction with spatial and temporal analysis (local spatial autocorrelation) for hot spot detection. The results demonstrate that earthquake footprints can be reliably and accurately identified,. The results also show that a number of relevant semantic topics can be automatically identified without a priori knowledge, revealing clearly differing temporal and spatial signatures. Furthermore, a damage map that indicates where significant losses have occurred is also presented [35]

3. Proposed Methodology

This section describes the proposed methodology to use Twitter as a social sensor system for natural disasters, as depicted in Figure 3.

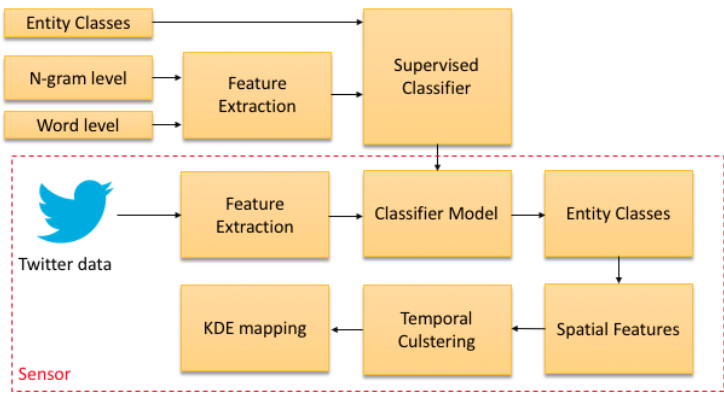


Figure 3. Proposed methodology to use Twitter as a social sensor system for disasters

First, data are gathered with entity classes related to locations, people, organisations and other denominations (that do not belong to any rigid designator) [36]. In the training stage, the entity classes are analysed at a word level and then expanded to a sentence level using an n-gram model. Features are then extracted to train a classifier using supervised learning. In the sensing stage, tweets are scraped from Twitter API, and grouped in one of three topics  $T \in \{\text{disaster areas, missing individuals, shelters}\}$ . After extracting Features are extracted from the tweets to predict entity classes. A request to Google Maps [37] is sent to obtain the spatial features (latitude and longitude coordinates) from the predicted entities. The set of spatial features is then scored by the occurrences of entities in the same spatial region. Temporal features correspond to the number of mentions of the entity in the same spatial region during a time period. These features are used by KDE to estimate the entity class location. Estimated locations are finally plotted on a Basemap surface.

4. Data Gathering

Data gathering is achieved by querying well-identified terms in Spanish related to natural disasters. Here, we scrap tweets using Twitter API Search and Stream. The purpose of this step is to create a training dataset,  $X_t$ , comprising named entities and classes,  $y$  (which are described in the next section), and to create clusters of tweets collected over a time period. The querying terms used are well known urban spaces of a city [38], as well as descriptive terms related to natural disasters.

Twitter characteristics, such as retweets and mentions, can contribute to the propagation strength of each tweet and thus, as a temporal feature [39]. A sample query,  $q$ , sent to the Twitter API Search to collect tweets to construct  $X_t$  is shown next:

$$q = [\text{\#sismo}, \text{ayuda}, \text{avenida tlalpan}]. \tag{1}$$

$q$  contains the following words in English = #earthquake,help,tlalpan avenue

5. Word Representations: Word Level Analysis and the n-gram Model

Word level analysis is a widely used language model [40,41] to describe words given a certain context. Each word is mapped to a new representation based on its co-occurrences and semantics. In this work, a word level classifier for Spanish language, known as Polyglot, is adapted to perform the entity extraction for each tweet. Table 2 tabulates the set of classes,  $y$ , used in this work.

Table 2. Entity classes for extraction and classification.

Class	Description
LOC	Location representing a street, avenue, region, county, or any administrative division
ORG	Reference to an organisation, institution, or establishment
PER	Reference to a person or a group of people
O	Any other criteria

Based on the word-level classification proposed in [42], let us define a tweet by  $\mathcal{T}_i^n = \{w_{i-n}, \dots, w_i, \dots, w_{i+n}\}$  as a phrase centered around the word  $w_i$  with a window size  $2n + 1$ , and a target function  $F : \mathcal{T} \mapsto y$ . First the tweet  $\mathcal{T}_i^n$  is mapped to its embedding representation  $\phi_i^n$ . Next, a discriminative model  $\psi_y$  trains a set  $X_t$  of word embeddings to score a tag  $y$  using a neural network with one hidden layer of size  $h$ . A one-vs-all classifier, the function  $F$  is constructed and penalised by Equation (2) :

$$F = \frac{1}{m} \sum_{\forall i} \max(0, 1 - \psi_{t_i}(\phi_i^n) + \max(\psi_y(\phi_i^n))), \forall y \neq t_i \tag{2}$$

where  $t_i$  is the correct class of the word  $w_i$  and  $m$  is the size of the training set. According to [43], an n-gram model is a suitable implementation to capture patterns of local co-occurrence within sentences comprising contiguous words [44]. Unlike the word-level analysis, an n-gram model allows us to construct a word vocabulary  $w \in \mathcal{T}$  in which  $w^n - 1$  independent words with probability of the form  $P(w_n | w_i^{n-1})$  coexist, where  $w_n$  is the word occurrence given  $w_i^{n-1}$  word particles in a tweet. Table 3 exemplifies the n-gram model with the corresponding entity classes.

Table 3. N-gram model examples with their corresponding entity classes.

n-Gram	Tweet in Spanish	English Translation
1-gram	ayuda gente atrapada en $\overbrace{\text{edificio}}^{\text{LOC}}$ $\overbrace{\text{Alvaro}}^{\text{LOC}}$ $\overbrace{\text{Obregon}}^{\text{LOC}}$ $\text{1-gram}$ $\text{1-gram}$ $\text{1-gram}$	help people trapped in an Alvaro Obregon building
2-gram	derrumbe en el multifamiliar de $\overbrace{\text{av. tlalpan}}^{\text{LOC}}$ $\text{2-gram}$	a collapse in tlalpan avenue multifamily building
3-gram	se cayó el $\overbrace{\text{Soriana de Taxqueña}}^{\text{LOC}}$ $\text{3-gram}$	Taxqueña’s Soriana has fallen down



A training set  $X_t$  of size  $m$  is designed with a pre-classification for both, word-level and sentence-level (n-gram) representations. Each element of  $X_t$  is denoted as a document  $d_i \in X_t$ , with terms  $v$  defined by a single occurrence of a term such that  $v_k \in d_i$ , mapped to a vocabulary  $\mathcal{V} = v_1, \dots, v_k$  of size  $k$

### 5.1. Feature Extraction

The representation of  $X_t$  can be done by vector space modelling, so the importance of each term  $v_j$  contained in a document  $d_i$  can be approximated. The algorithm used in this work is based on Term Frequency-Inverse Document Frequency, which normalises the frequency of the set of documents  $X_t$  to assign weights with respect to their importance and relevance. Based on the vocabulary  $\mathcal{V}$  defined with  $k$  terms and the training set with  $m$  documents, the algorithm estimates the number of occurrences  $\forall d_i \in X_t$ , producing a set  $\mathcal{D} \in \mathbb{R}^{m \times k}$ , in which each element is the frequency of the vocabulary of terms in the documents  $d_i$  such that  $\mathcal{D} = f(d_{1,1}), \dots, f(d_{m,k})$ . The term frequency,  $tf_{i,k}$ , is defined by a vocabulary term  $v_k$  in a document  $d_i$  as described in Equation (4):

$$tf_{i,k} = \frac{f(d_{1,k})}{\max\{f(d_{1,i}), \dots, f(d_{m,k})\}}. \quad (3)$$

For each  $v_i \in \mathcal{V}$ , the number of documents in which each term occurs at least once is computed. This is known as the document frequency,  $df_i$ , of the document  $d_i$ . To retrieve the inverse document frequency for  $m$  documents, Equation (4) is used:

$$idf_i = \log \frac{m}{df_i}. \quad (4)$$

From the equation above, each pair of document-term is then represented by its weight  $W_{i,k} = tf_{i,k} \times idf_i$ . The resulting set is a weighted feature array  $\mathcal{W} = \{W_{1,1}, \dots, W_{m,k}\}$ .

### 5.2. Supervised Learning

Supervised learning has been extensively used in extensive research works for NER [45–48] as an alternative to classic adaptations of probabilistic language models, such as conditional random fields and hidden Markov models. In this section, we present an algorithm based on random forests to create a classification model to address the sensor design. Random forests is an algorithm assembles decision trees and combines weak learning to build a more robust classification model [49,50]. This is how the generalisation of an error reduces the avoidance of inherent adjustment to different classification models. In this work, the algorithm for the classification of entities of the set of weighted documents,  $X_{t_{\mathcal{W}}}$ , is described in Algorithm 1.

---

#### Algorithm 1: Training Samples

---

```

function RANDOMFORESTTRAINING( $X_{t_{\mathcal{W}}}$ );
  for  $i = 1 : k$  do
     $B \leftarrow$  a bootstrap sample of  $X_{t_{\mathcal{W}}}$  with size  $n$ ;
     $C \leftarrow$  a grown decision tree from the bootstrapped sample  $B$ ;
    foreach  $node$  in  $C$  do
      Select  $f(d)$  features from  $C$ ;
      Split the node by maximising the information gain of  $f(d)$ ;
    Aggregate the prediction by each tree to assign the class label by majority voting
  end function

```

---

Algorithm 1 uses a statistical technique called bootstrapping, which aims to reduce the variance of random forest learning given  $k$  independent weighted documents in  $X_{t_W}$ , each with variance  $\sigma_{t_W}^2$ . Bootstrapping considers the variance of the mean  $\bar{X}_{t_W}$  of each document  $d_i$ , which is given by  $\frac{\sigma_{t_W}^2}{n}$ , where  $n$  is the number of weights. To reduce such variance and increase the prediction accuracy of the entity classes,  $X_{t_W}$  is split into several bootstrapping training sets  $B \in X_{t_W}$ . A decision tree is trained in the  $b$ th bootstrapped training set in order to attain a predictive function  $\hat{f}^{b*}(d_i)$ . Resulting predictions are taken from the average, as specified by Equation (5):

$$\hat{f}(d_i) = \frac{1}{B} \sum_{\forall b,i} \hat{f}^{b*}(d_i). \quad (5)$$

Finally, to predict an entity class, a majority voting classifier combines predicted classes of each individual tree,  $C$ , and selects the entity class  $\hat{y}$  that receives the most votes, as follows:

$$\hat{y} = \text{mode}\{C_1(d_i), C_2(d_i), \dots, C_k(d_i)\}. \quad (6)$$

### 5.3. Spatial and Temporal Information

To be able to find the spatial information, the entities,  $\hat{y}$ , classified by the random forest model are used to request Google API locations, obtaining as a result an address described in terms of a string of characters and geographic coordinates. The entities are first compared by querying the space model representation,  $\hat{Y}$ , of entities using the cosine similarity measure [51] to find similar entities (given some threshold  $\theta \in \{0, \dots, 1\}$ ). Therefore, if an entity is similar to an already found entity, it is assigned the same spatial information. Figure 4 shows the proposed spatial information extraction.

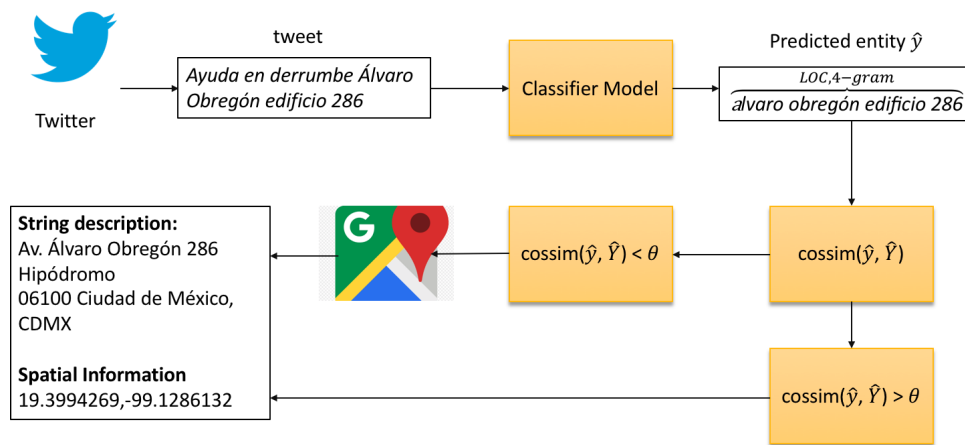


Figure 4. Spatial information extraction.

For the temporal features, a window of time is taken from the first round of tweets collected. When a tweet is denoted as initial, the corresponding spatial information is extracted along with the corresponding time-stamp (date of creation),  $ts$ . All the subsequent retweets from the initial tweet are considered to have a threshold  $\theta = 1$  (for total similarity) and are assigned a creation time-stamp equal to the difference between the time-stamp of the parent tweet and their actual time-stamp, i.e.,  $(ts_{parent} - ts_{children})$ ,  $\{\forall ts_{parent} \in \hat{y}_{parent}, \forall ts_{parent} \in \hat{y}_{children}\}$ . When a tweet is similar to others according to the threshold  $\theta$ , the date of creation is then calculated based on that of the similar tweets. Tweets are then re-ordered by date of creation, from the oldest to the most recent one, i.e.,  $sort(\hat{y}_1 \mapsto ts_1, \dots, \hat{y}_n \mapsto ts_n)$ .



#### 5.4. Kernel Density Estimation

Several works have employed [52–55] KDE [56], a statistical method to visualise hotspots from estimated spatial points distributed on a two-dimensional spatial probability density function. In this work, this method is applied to spatial information from predicted entity classes. Based on Basemap [57], we can for example identify geographic areas prone to be dangerous, detect areas with the most missing individuals, and locate shelters. In order to quantify the incoming entities of a certain class from the random forest model at a spatial point  $g$ , Equation (7) is computed as suggested in [55]:

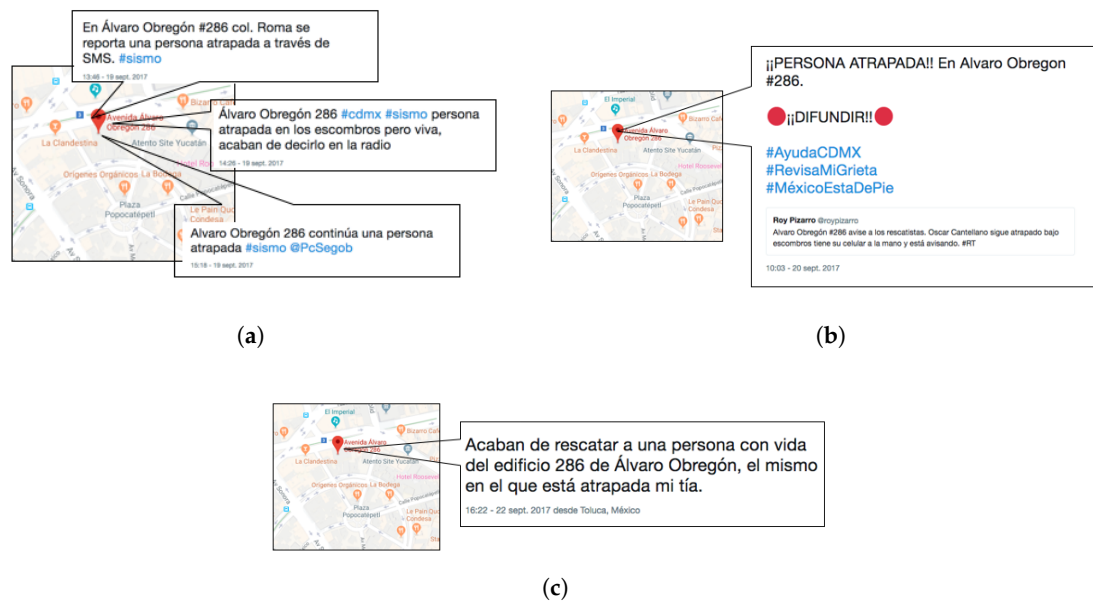
$$f(g) = \gamma(g, h) = \frac{1}{Ph} \sum_{w_i} K \left( \frac{\|g - g_{w_i}\|_{\ell_2}}{h} \right) \quad (7)$$

where  $h$  is the bandwidth,  $P$  is the number of spatial features of the topic  $T \in \{\text{disaster areas, missing individuals, shelters}\}$  within the window of time,  $w_i$  maps is the entity of spatial information from the temporal window,  $K$  is a density function,  $\ell_2$  is the vector norm, and  $g_i$  is a location related to the disaster.

#### 6. Sensing Information: A Case study of the 2017 Mexico Earthquake

On 19 September 19 2017, at 1:14 p.m., an earthquake of 7.1 magnitude in the Richter scale with an epicenter in Axochiapan, Morelos, an adjacent state of Mexico City, impacted the urban infrastructure of the city and surrounding areas. Although the alarm system is efficient when epicentres occur in the Pacific Ocean coast, in the case of this natural disaster, the sensors were so close to the city that the evacuations took place 11 seconds after the earthquake started. It was not to be expected that Twitter users reported events related to the disaster zones in real time. In addition to army and navy personnel, a large number of individuals took to the streets to offer humanitarian aid to people who were in risk areas. Days later, a number of official and collaborative shelters were set up in churches, parks, schools, and other places to offer help to the victims. Figures 5a–c show sample tweets collected over a three-day observation window by the proposed methodology.

We compare our proposed classifier based on random forest learning with the work presented in [58]. For our classifier, we use a 3-day observation window. In [58], the extraction of localities embedded in social network messages contain the following metrics: lexical, morpho-syntactic, and lexical expansion, which are associated with three classes—damage, non-damage, and non-relevant. For the events that occurred during the earthquake of 20 May 2012 in Northern Italy, that work trains messages using a Support Vector Machine (SVM) for word embeddings (EMB) Table 4 tabulates the performance metrics of the class *damage* from [58] and the entity classes LOC, O, ORG, PER of this work (Note that the entity class LOC is the most comparable metric with the class *damage* from [58]).



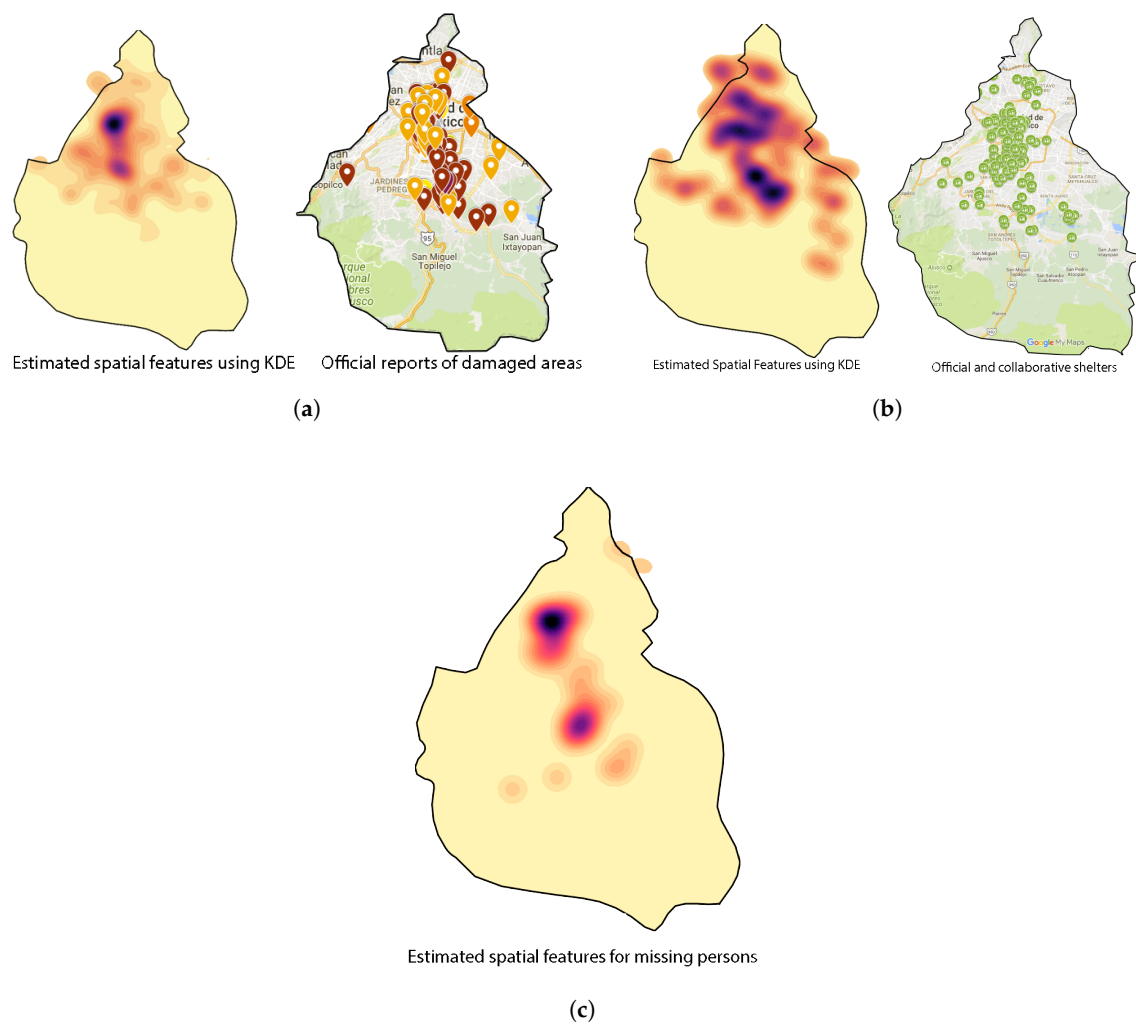
**Figure 5.** The event starts at 1:46 pm, almost half an hour after the earthquake. The localised entity corresponds to the street of Av. Álvaro Obregón, number 286, with geographic coordinates 19.4162205, –99.1705947. The other classified entities have the same similarity and are ordered temporarily until the last report at 4:22 p.m. of the third observation day. (a) The event starts, users start reporting that a person is trapped in a collapsed building; (b) a day later users continue reporting that a person is in the rubble, and information is already disseminated in a retweet; (c) on the third day, the event ends with the victim reported as rescued.

**Table 4.** Performance metrics for entity classification.

Dataset	Classifier	Entity Class	Precision	Recall	F-1 Score
19 September 2017 Mexico Earthquake	Random Forest	LOC	0.97	0.89	0.91
19 September 2017 Mexico Earthquake	Random Forest	O	0.98	0.96	0.97
19 September 2017 Mexico Earthquake	Random Forest	ORG	0.79	0.76	0.77
19 September 2017 Mexico Earthquake	Random Forest	PER	0.98	0.90	0.97
20 May 2012 Italy Earthquake (L'Aquila)	EMB-SVM	Damage	0.85	0.82	0.72
20 May 2012 Italy Earthquake (Emilia)	EMB-SVM	Damage	0.85	0.87	0.52
20 May 2012 Italy Earthquake (Sradegna)	EMB-SVM	Damage	0.82	0.43	0.33

Figures 6a–c present the hotspots obtained from the KDE of spatial features collected over a span of three days as described in 5.4. These results are compared from those obtained using official data from the September 19 earthquake, which are publicly available as *Mapeo Verificado19s* [59]. The following categories are shown in the Figures using official data (note that there are no official reports of missing people, so it was not possible to compare this category):

- **Official Damages:** includes collapsed buildings, major risks, minor risks, and wall collapses;
- **Official Shelters :** official government assistance and aids;
- **Collaborative Shelters:** non-official collaborative assistance and aids;
- **Mexico City's Monitor System:** which includes major risks, collapsed buildings and gas hazards;
- **Harvard MIT:** collaborative data from community computed by Harvard University and the MIT.



**Figure 6.** Estimations using KDE using data collected over a 3-day window. (a) The hotspots of estimated spatial features related to damages and collapses are shown and compared with official reports. (b) The hotspots of estimated spatial features related to official and collaborative shelters are shown and compared with official reports. (c) Although there are no official reports of missing persons, the hotspots of estimated spatial features related to this category are plotted.

## 7. Conclusions

In this work, a methodology to use Twitter as a social sensor system was proposed. The methodology extracts features from tweets by analysing entities at the word and sentence levels (n-gram model). It then uses a random forest classifier, which combines different small decision trees and uses majority voting to classify entities extracted from newly seen tweets. Classified entities are then used to extract spatial features using the Google Maps API, which allows reconstructing the entity's location by assigning it geographic coordinates. Using KDE, the compendium of spatial features collected over an temporal window is also to visualise on two-dimensional map the hotspots associated with important information related to the disaster, such as highly affected areas or areas with the most missing individuals. The methodology is evaluated using tweets posted during the earthquake of 19 September 2017 in Mexico. A 3-day observation window is used to collect spatial and temporal features. Classification results are compared with those of a similar work which uses word embeddings to classify damaged areas after an earthquake using a Support Vector Machine and Word Embeddings. Results showed that the random forest classification accurately classified the entities over the 3-day observation window.

**Author Contributions:** A.H.-S. and V.S. developed the proposed algorithm and carried out the analysis of the final results. K.T.-M. and H.P.-P. developed the computer program used to evaluate the performance of the proposed algorithm for Named Entity Recognition. J.P.-P. and G.S.-P. developed the computer programs for Kernel Density Estimation, whose results are presented in the evaluation results' sections. All authors participated in the write-up and review.

**Acknowledgments:** The authors thank the National Science and Technology Council of Mexico (CONACyT), and the Instituto Politécnico Nacional for the financial support for this research.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

- Ekström, G.; Dziewonski, A.M.; Steim, J.M. Single station CMT; Application to the Michoacan, Mexico, earthquake of September 19, 1985. *Geophys. Res. Lett.* **1986**, *13*, 173–176.
- Gao, H.; Barbier, G.; Goolsby, R.; Zeng, D. *Harnessing the Crowdsourcing Power of Social Media for Disaster Relief*; Arizona State Univ Tempe: Tempe, AZ, USA, 2011.
- Teets, J.C. Post-earthquake relief and reconstruction efforts: The emergence of civil society in China? *China Q.* **2009**, *198*, 330–347.
- Smith, P.C.; Simpson, D.M. Technology and communications in an urban crisis: The role of mobile communications systems in disasters. *J. Urban Technol.* **2009**, *16*, 133–149.
- Heinzelman, J.; Waters, C. *Crowdsourcing Crisis Information in Disaster-Affected Haiti*; US Institute of Peace: Washington, DC, USA, 2010.
- Hayashi, N.J.; Ott IV, E.S.; Tsang, A.Y.; Fukuda, M.; Wascovich, D.; Quoc, M. Multimedia Sharing in Social Networks for Mobile Devices. U.S. Patent No. 8,046,411, 25 October 2011.
- Dhillon, H.S.; Huang, H.; Viswanathan, H. Wide-area wireless communication challenges for the Internet of Things. *IEEE Commun. Mag.* **2017**, *55*, 168–174.
- Kaplan, A.M.; Haenlein, M. Users of the world, unite! The challenges and opportunities of Social Media. *Bus. Horiz.* **2010**, *53*, 59–68.
- García-Palomares, J.C.; Salas-Olmedo, M.H.; Moya-Gómez, B.; Condeco-Melhorado, A.; Gutierrez, J. City dynamics through Twitter: Relationships between land use and spatiotemporal demographics. *Cities* **2018**, *72*, 310–319.
- Sagl, G.; Resch, B.; Hawelka, B.; Beinatz, E. From social sensor data to collective human behaviour patterns: Analysing and visualising spatio-temporal dynamics in urban environments. In *Proceedings of the GI-Forum*; Herbert Wichmann Verlag: Berlin, Germany, 2012; pp. 54–63.
- Aggarwal, C.C.; Abdelzaher, T. Social sensing. In *Managing and Mining Sensor Data*; Springer: Boston, MA, USA, 2013; pp. 237–297.
- Aggarwal, C.C. (Ed.) *Managing and Mining Sensor Data*; Springer Science & Business Media: 2013.
- Abdelzaher, T.; Anokwa, Y.; Boda, P.; Burke, J.; Estrin, D.; Guibas, L.; Kansal, A.; Madden, S.; Reich, J. Mobiscopes for human spaces. *IEEE Pervasive Comput.* **2007**, *6*, 20–29.
- Xu, Z.; Mei, L.; Choo, K.K.R.; Lv, Z.; Hu, C.; Luo, X.; Liu, Y. Mobile crowd sensing of human-like intelligence using social sensors: a survey. *Neurocomputing* **2018**, *279*, 3–10.
- Hernandez-Suarez, A.; Sanchez-Perez, G.; Toscano-Medina, K.; Martinez-Hernandez, V.; Perez-Meana, H.; Olivares-Mercado, J.; Sanchez, V. Social Sentiment Sensor in Twitter for Predicting Cyber-Attacks Using  $\ell_1$  Regularization. *Sensors* **2018**, *18*, 1380.
- Wang, R.Q.; Mao, H.; Wang, Y.; Rae, C.; Shaw, W. Hyper-resolution monitoring of urban flooding with social media and crowdsourcing data. *Comput. Geosci.* **2018**, *111*, 139–147.
- Lee, K.; Ganti, R.; Srivatsa, M.; Mohapatra, P. Spatio-temporal provenance: Identifying location information from unstructured text. In *Proceedings of the 2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, San Diego, CA, USA, 18–22 March 2013; pp. 499–504.
- Li, R.; Lei, K.H.; Khadiwala, R.; Chang, K.C.C. Tedas: A twitter-based event detection and analysis system. In *Proceedings of the 2012 IEEE 28th International Conference on Data Engineering (ICDE)*, Washington, DC, USA, 1–5 April 2012; pp. 1273–1276.
- Feng, W.; Zhang, C.; Zhang, W.; Han, J.; Wang, J.; Aggarwal, C.; Huang, J. STREAMCUBE: hierarchical spatio-temporal hashtag clustering for event exploration over the twitter stream. In *Proceedings of the*

- 2015 IEEE 31st International Conference on Data Engineering (ICDE), Seoul, Korea, 13–17 April 2015; pp. 1561–1572.
20. Sisco, M.R.; Bosetti, V.; Weber, E.U. When do extreme weather events generate attention to climate change? *Clim. Chang.* **2017**, *143*, 227–241.
  21. Nadeau, D.; Sekine, S. A survey of named entity recognition and classification. *Linguisticae Investig.* **2007**, *30*, 3–26.
  22. Bontcheva, K.; Derczynski, L.; Roberts, I. Crowdsourcing named entity recognition and entity linking corpora. In *Handbook of Linguistic Annotation*; Springer: Dordrecht, The Netherlands, 2017; pp. 875–892.
  23. Jeon, Y.; Cho, C.; Seo, J.; Kwon, K.; Park, H.; Chung, I.J. Rule-Based Topic Trend Analysis by Using Data Mining Techniques. In *Advanced Multimedia and Ubiquitous Engineering*; Springer: Singapore, 2017; pp. 466–473.
  24. Silverman, B.W. *Density Estimation for Statistics and Data Analysis*; Routledge: 2018.
  25. Kongthon, A.; Haruechaiyasak, C.; Pailai, J.; Kongyoung, S. The role of Twitter during a natural disaster: Case study of 2011 Thai Flood. In Proceedings of the PICMET'12 Technology Management for Emerging Technologies (PICMET), Vancouver, BC, Canada, 29 July–2 August 2012; pp. 2227–2232.
  26. Sachdeva, S.; McCaffrey, S. Using Social Media to Predict Air Pollution during California Wildfires. In Proceedings of the ACM 9th International Conference on Social Media and Society, Copenhagen, Denmark, 18–20 July 2018; pp. 365–369.
  27. Hughes, A.L.; Palen, L. Twitter adoption and use in mass convergence and emergency events. *International J. Emerg. Manag.* **2009**, *6*, 248–260.
  28. Earle, P.S.; Bowden, D.C.; Guy, M. Twitter earthquake detection: Earthquake monitoring in a social world. *Ann. Geophys.* **2012**, *54*, doi:10.4401/ag-5364.
  29. Sakaki, T.; Okazaki, M.; Matsuo, Y. Earthquake shakes Twitter users: real-time event detection by social sensors. In Proceedings of the ACM 19th International Conference on World Wide Web, Raleigh, NC, USA, 26–30 April 2010; pp. 851–860.
  30. Finch, K.C.; Snook, K.R.; Duke, C.H.; Fu, K.W.; Tse, Z.T.H.; Adhikari, A.; Fung, I.C.H. Public health implications of social media use during natural disasters, environmental disasters, and other environmental concerns. *Nat. Hazards* **2016**, *83*, 729–760.
  31. Middleton, S.E.; Middleton, L.; Modafferi, S. Real-time crisis mapping of natural disasters using social media. *IEEE Intell. Syst.* **2014**, *29*, 9–17.
  32. Ashktorab, Z.; Brown, C.; Nandi, M.; Culotta, A. Tweedr: Mining twitter to inform disaster response. In *ISCRAM*; May 2014.
  33. Cresci, S.; Tesconi, M.; Cimino, A.; Dell'Orletta, F. A linguistically-driven approach to cross-event damage assessment of natural disasters from social media messages. In Proceedings of the ACM 24th International Conference on World Wide Web, Florence, Italy, 18–22 May 2015; pp. 1195–1200.
  34. Cresci, S.; Cimino, A.; Dell'Orletta, F.; Tesconi, M. Crisis mapping during natural disasters via text analysis of social media messages. In *International Conference on Web Information Systems Engineering*; Springer: Cham, Switzerland, 2015; pp. 250–258.
  35. Resch, B.; Usländer, F.; Havas, C. Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment. *Cartogr. Geogr. Inf. Sci.* **2018**, *45*, 362–376.
  36. Kripke, S.A. Naming and necessity. In *Semantics of Natural Language*; Springer: Dordrecht, The Netherlands, 1972; pp. 253–355.
  37. Erol, M.H.; Bulut, F. Real-time application of travelling salesman problem using Google Maps API. In Proceedings of the IEEE Electric Electronics, Computer Science, Biomedical Engineering's Meeting (EBBT), Istanbul, Turkey, 20–21 April 2017; pp. 1–5.
  38. Krier, R.; Rowe, C. *Urban Space*; Academy Editions: London, UK, 1979.
  39. Spiro, E.; Irvine, C.; DuBois, C.; Butts, C. Waiting for a retweet: Modeling waiting times in information propagation. In Proceedings of the 2012 NIPS Workshop of Social Networks and Social Media Conference, December 2012; Volume 12. Available online: <http://snap.stanford.edu/social2012/papers/spiro-dubois-butts.pdf> accessed on June 18, 2018.
  40. Turian, J.; Ratinov, L.; Bengio, Y. Word representations: A simple and general method for semi-supervised learning. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010; Association for Computational Linguistics: pp. 384–394.



41. Liu, X.; Zhang, S.; Wei, F.; Zhou, M. Recognizing named entities in tweets. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, Portland, Oregon, 19–20 June 2011; Association for Computational Linguistics: pp. 359–367.
42. Al-Rfou, R.; Kulkarni, V.; Perozzi, B.; Skiena, S. Polyglot-NER: Massive multilingual named entity recognition. In Proceedings of the 2015 SIAM International Conference on Data Mining, June 2015; Society for Industrial and Applied Mathematics: pp. 586–594.
43. Brown, P.F.; Desouza, P.V.; Mercer, R.L.; Pietra, V.J.D.; Lai, J.C. Class-based n-gram models of natural language. *Comput. Linguistics* **1992**, *18*, 467–479.
44. Habash, N. The use of a structural n-gram language model in generation-heavy hybrid machine translation. In *Natural Language Generation*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 61–69.
45. Uszkoreit, H.; Staack, O.; Budde, K.; Halleck, F.; Schmidt, D. Detecting Named Entities and Relations in German Clinical Reports. In Proceedings of the Language Technologies for the Challenges of the Digital Age: 27th International Conference, Berlin, Germany, 13–14 September 2017; Springer; Volume 10713, p. 146.
46. Zhang, Q.; Fu, J.; Liu, X.; Huang, X. Adaptive Co-attention Network for Named Entity Recognition in Tweets. In *AAAI*; 2018.
47. Goyal, A.; Gupta, V.; Kumar, M. Recent Named Entity Recognition and Classification techniques: A systematic review. *Comput. Sci. Rev.* **2018**, *29*, 21–43.
48. Zhu, Q.; Li, X.; Conesa, A.; Pereira, C. GRAM-CNN: A deep learning approach with local context for named entity recognition in biomedical text. *Bioinformatics* **2017**, *34*, 1547–1554.
49. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.
50. Speck, R.; Ngomo, A.C.N. Ensemble learning for named entity recognition. In *International Semantic Web Conference*; Springer: Cham, Switzerland, 2014; pp. 519–534.
51. Steinbach, M.; Karypis, G.; Kumar, V. A comparison of document clustering techniques. In *KDD Workshop on Text Mining*; August 2000; Volume 400, pp. 525–526.
52. Li, L.; Goodchild, M.F.; Xu, B. Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartogr. Geogr. Inf. Sci.* **2013**, *40*, 61–77.
53. Sims, K.M.; Weber, E.M.; Bhaduri, B.L.; Thakur, G.S.; Resseguie, D.R. Application of social media data to high-resolution mapping of a special event population. In *Advances in Geocomputation*; Springer: Cham, Switzerland, 2017; pp. 67–74.
54. Huang, H.; Dong, Y.; Tang, J.; Yang, H.; Chawla, N.V.; Fu, X. Will Triadic Closure Strengthen Ties in Social Networks? *ACM Trans. Knowl. Discov. Data (TKDD)* **2018**, *12*, 30.
55. Gerber, M.S. Predicting crime using Twitter and kernel density estimation. *Decis. Support Syst.* **2014**, *61*, 115–125.
56. Sheather, S.J.; Jones, M.C. A reliable data-based bandwidth selection method for kernel density estimation. *J. R. Stat. Soc. Ser. B Methodol.* **1991**, 683–690.
57. Hunter, J.D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95.
58. Avvenuti, M.; Cresci, S.; Del Vigna, F.; Fagni, T.; Tesconi, M. CrisMap: A Big Data Crisis Mapping System Based on Damage Detection and Geoparsing. *Inf. Syst. Front.* **2018**, 1–19, doi:10.1007/s10796-018-9833-z.
59. Sismo Verificado 19s. Available online: <http://google.org/crisismap/a/gmail.com/v19s> (accessed on 5 July 2018).