# Using centrality measures to improve the classification performance of tweets during natural disasters

## Usando medidas de centralidad para mejorar la clasificación de tweets durante desastres naturales

Rodrigo Vásquez[1]
Fabián Riquelme[2] *
Pablo González-Cantergiani[3]
Cristobal Vásquez[1]

1 CITIAPS, Universidad de Santiago de Chile, Avda. Ecuador 3519, Santiago, Chile. E-mail: {rodrigo.vasquez, cristobal.vasquez}@usach.cl
2 Universidad de Valparaíso, Escuela de Ingeniería Civil Informática, General Cruz #222, 3rd floor, Valparaíso, Chile. E-mail: fabian.riquelme@uv.cl
3 E-mail: gonzalezcantergiani@gmail.com

* Corresponding author: fabian.riquelme@uv.cl

**SUMMARY:**
Las redes sociales como Twitter facilitan la comunicación durante posibles desastres naturales. Un problema recurrente es lograr distinguir en tiempo real los *tweets* más contingentes de un desastre, del flujo masivo de mensajes recibidos. Para tratar este problema, el aprendizaje de máquina permite clasificar *tweets* respecto a su relevancia o credibilidad. En este artículo, se propone el uso de medidas de centralidad para mejorar conjuntos de datos de entrenamiento para el uso de clasificadores de aprendizaje activo. Como caso de estudio, se analizan *tweets* recolectados durante las inundaciones de Santiago de Chile en el año 2016. Este enfoque permite mejorar la consistencia y pertinencia en el proceso de etiquetado, así como la calidad de los clasificadores.
**Palabras clave**: Aprendizaje activo, Twitter, Medida de centralidad, Respuesta a desastres, Usuario Influyente

**ABSTRACT:**
Online social networks like Twitter facilitate instant communication during natural disasters. A key problem is to distinguish in real-time the most assertive and contingent tweets related to the current disaster from the whole streaming. To address this problem, machine learning allows to classify tweets according to their relevance or credibility. In this article, it is proposed to use centrality measures to improve the training data sample of active learning classifiers. As a case study, tweets collected during the massive floods in Santiago of Chile at 2016 are considered. This approach improves the consistency and pertinence of the labeling process, as well as the classifiers' performance.
**Keywords**: Active learning, Twitter, Centrality measure, Disaster response, User influence

Received day of the month of 20 ## (the Editorial Production will complete these fields if the article is approved)
Accepted day of the month of 20 ##

# INTRODUCTION

In time of crisis, microblogging services are used to communicate tactical and actionable information, that helps to understand mass emergency events [1]. These services can help to geolocate and visualize infrastructure damages, storage facilities, needs of the population, shortages, among other needs[1]. Nowadays, Twitter is one of the main online social networking services worldwide. Although it has fewer users than other social networks such as Facebook or Instagram, it has been designed to spread information faster, publicly and persistently. Due to these features, Twitter is a useful tool for information spread, coordination and decision-making during crisis events such as natural disasters [2, 3].

During a crisis, Twitter users tend to spread diverse and scattered information, both in content and level of detail. Aside from the numerous messages unrelated to the disaster, there may also be many informative tweets about damages reports, requests and offers of help, searches for missing, encouragement messages, among others. As a first approach to find these informative tweets, a bag of words related to the current disaster can be defined, and then searching for those tweets that contain some of the terms of this bag of words. In order to classify informative tweets, there exist crowdsourcing platforms on which some volunteers can label different tweets under different criteria. However, labeling time is costly, and during a disaster, reaction must be as quickly as possible. Furthermore, the amount of data could be too large, so labeling a high percentage of tweets can be unfeasible [3]. To avoid this problem, it is common to use *active learning*, a supervised machine learning method that only requires to select a small subset of tweets collected during the disaster event. This smaller collection of tweets can be labeled by volunteers in a crowdsourcing platform and then be used as training data for a supervised learning classifier, i.e., a machine learning application that continues classifying non-labeled tweets automatically. After that, the tweets closer to the decision area (which are the most difficult to classify) are sent to be labeled again by the volunteers, so that the classifier is re-trained with the new labeled instances [4, 5].

It is known that the precision of the classifiers can be increased with a larger number of labeled tweets [5]. However, in this paper it is shown that the quality of the sample data can be improved without increasing the manual labeling effort. Moreover, it could even be considered fewer tweets than those taken in a more standard methodology.

The main contributions of this work are the following:

- A strategy to improve the training sample for supervised classifiers of tweets related to natural disasters is proposed. These classifiers are trained with an active

[1] See, for instance http://aidr.qcri.org

learning approach. As usual, it is required a collection of tweets obtained via streaming API, whose content is restricted by an initial bag of words related to the event (earthquake, hurricane, fire, flood, etc.), its location, etc. However, instead of considering a random selection of the tweets collected over time, it is proposed to use centrality measures to go getting incrementally more relevant (i.e., assertive and contingent) than irrelevant tweets from the collection. These measures must be computed in (almost) real-time. In this paper, two centrality measures are applied. One measure allows to detect tweets with a rich content and a high dispersion capacity through the network. The other measure allows to detect influential users who are acting as opinion leaders, i.e., users that rapidly spread what they said through the network; then, from the most influential users, a random sample of tweets is selected. Since these tweets have been filtered by the initial bag of words, presumably they will have to do with the case study. These measures will be detailed in the Extraction section.

- The previous strategy is applied on tweets collected during a real case study, namely the massive floods in Santiago of Chile at 2016. Through a manual labeling process, it is shown that the training sample significantly improves in pertinence, compared to that provided by a traditional collection of random tweets. A higher *pertinence* means a collection of tweets considered more valuable by the volunteers. Furthermore, for this case study the training sample also improves in *consistency*, i.e., there is a greater agreement among the volunteers to classify the tweets, according to a Fleiss' kappa coefficient of 2/3. Note that, unlike pertinence, the consistency does not depend on the quality of the data set. In fact, the above does not prevent that during another natural disaster, the volunteers could agree that most of the tweets obtained from a random sample are not informative at all. Therefore, in this strategy, the pertinence property is more interesting than consistency.

- The performance of the active learning approach in the real study case is evaluated, proving that the new strategy improves the precision, recall, and F-score.

As far as we know, the strategy of using Twitter centrality measures to improve the performance of classifiers using an active learning approach is new. The remaining of the paper is organized as follows: Related work section presents a brief review about classifiers and centrality measures focused on the Twitter network. In the Methodology section it is described the methodology to improve the training sample, as well as defined the centrality measures used to accomplish this aim. In the Study-case section is described the study case and presented the experiments and main results of the paper. Finally, Conclusions section is devoted to the conclusions, discussion, and future work.

# RELATED WORK

## Classifiers

Social media such as Twitter contain a significant amount of noise, that needs to be filtered in order to capture the most relevant and contingent information [5]. Machine learning algorithms are heavily used in disaster situations to classify, rank and cluster different kinds of data, such as documents, images, messages, among others. In an active learning approach, these algorithms require a preliminary training data set, which is generated manually by humans. In the crisis specific domain, it is especially important to validate the experiments with real and current data, because using pre-existing data sets can impair the accuracy of the classifiers [5]. The latter also makes it very difficult to compare methodologies by using independent data sets.

There is abundant research about classifiers for crisis management. Imran et al. summarize six popular dimensions that researchers use to classify information: subjective or emotional content, information provided, information sources, credibility, time or stages of an event, and location [6]. Avvenuti et al. presents a mapping system that uses natural language processing (NLP) and word embeddings to analyze social messages and classify them into four levels of damage [7]. Kejriwal and Gu propose a pipeline using active learning and fastText as main word embeddings package. Although these are preliminary results, they show relevant corpus without an extensive labeling process [8]. Zheng et al. propose a novel method called semi-supervised expectation maximization identifier (SSEM) that allows to identify new keywords as the disaster evolve [9]. Karime et al. filter tweets to identify those related to a specific natural disaster, mainly earthquakes, floods, fires and storms. In order to evaluate the classifiers, they use the *accuracy* metric, reaching 86.4% after using 90% of the training data. In that study, however, there is no mention of other relevant evaluation criteria, such as *precision*, *recall*, or *F-score* [10]. Imran et al. use a taxonomy formed by two classification criteria, namely *information* and *information type*, to deal with tweets related with the May 22, 2011 Joplin tornado that struck Joplin, Missouri [11]. For the first criteria, they obtain classifiers with a precision of 79% and a recall of 77%, while for the second criteria the best results were obtained in the caution class, with a precision of 85.9% and a recall of 76.5%. This latter taxonomy is also used in the experiments in the Study Case section. Therefore, precision and recall are also used in this study case. In addition, the F-score is calculated as an indicator of the performance of the classifiers, since it is a well-known measure of a test's accuracy, that integrates the precision and recall.

## Influence measures

In social network analysis, the *centrality* of a user refers to its relative importance within the network to which it belongs. In the context of microblogging services, centrality can be determined in terms of *popularity*, *activity*, or *influence* [12]. By definition, a user can be very active or popular, regardless of the quality of the content broadcast by him or her. This kind of users may post a lot of trivial content or no content at all. Therefore, the

content provided by active or popular users is not especially useful to train classifiers. Instead, influential users can be proposed as a meaningful source of relevant tweets to train classifiers. As a matter of fact, influential users can affect the actions of many other users, since they are the most capable to spread information within the network [13]. Knowing the user influence and being able to predict it is useful for many applications, such as viral marketing [14], information propagation [15], expertise recommendation [16], social customer relationship management [17], percolation theory [13], influence spread models [18], among others.

Only for Twitter, there are more than fifty influence measures (i.e., centrality measures to identify influential users) [12]. Each one of these measures provides a different ranking criterion. From all of them, there are only a few that can be computed in almost real-time. Several others, like the topical-sensitive measures (i.e., those that consider content analysis), are computationally expensive, and sometimes do not provide satisfactory results [19].

It is assumed that the tweets published by influential users are considered as relevant by other users. Furthermore, there also exist relevant tweets that may not have been published by influential users. Indeed, the influence can also be spread within the network through multi-layered peripheral user clusters [20]. That is why there also exist metrics to determine the influence of a tweet, regardless of the author. These metrics include features related to the tweet content, the actions generated around the tweet, and the user accounts interacting with the tweet [19].

## TRAINING METHODOLOGY

The proposed training methodology for supervised classifiers using an active learning approach is illustrated in Figure 1. It is inspired by the AIDR's (Artificial Intelligence for Disaster Response) architecture, an open source platform designed to perform automatic classification of crisis-related microblog communications [5]. This methodology can be divided into the following three steps: collection, extraction, and labeling or tagging. This section ends with an explanation of the classifiers design.

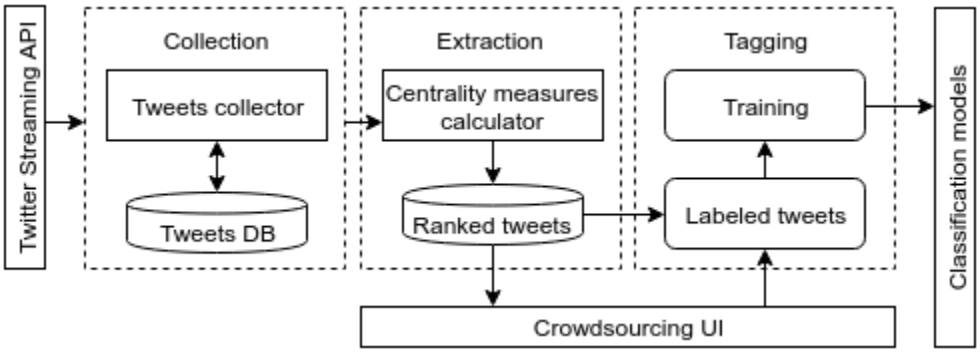


**Figure 1.** A proposal for training methodology to be used on active learning classifiers.

## Collection

When a crisis emerges, the public Twitter Streaming API[2] allows to collect filtered tweets related to the crisis. This API provides two filters for retrieving tweets: defining a polygon via coordinates (also known as geolocation filter) and using keywords (also known as bag of words). Here both are used. The geolocation filters allow to efficiently reduce the search domain, although without certainty of the collected tweets being related with the target disaster [21]. On the other hand, the bags of words have been successfully used for disaster situations [3, 5, 7].

Regardless of how the tweets are collected, it is important to note that the main topics related to a crisis can change over time. For instance, in an earthquake the first tweets spread by users may have to do with the magnitude, epicenter, and infrastructure damage; only after, new tweets about stock outs and specific help requirements emerge [1]. Some authors use techniques such as emerging topics detection or special lexicons to keep improving the bags of words [21]. This aims to collect relevant tweets related to the disaster, and improves the results of the classifier, in terms of precision and recall. Since the approach of this work seeks to improve the classification in the next step of extraction (using centrality measures) these kinds of techniques are not used. Instead, a general bag of words is maintained as a whitelist during the whole collection process.

## Extraction

After each hour of collection, centrality measures allow to extract automatically a list of the most relevant tweets obtained so far. Every time a new list is generated, it is labeled by a group of volunteers (see Tagging section). It is important to reduce this volunteers' effort. Therefore, the size of the lists should be relatively small, and the tweets that continue in the list between one period and the next one must be filtered to be labeled only the first time they appear.

Concretely, two centrality measures are used. They can be computed in almost real-time, using only features that can be obtained by the Twitter Streaming API. Remarkably, these measures do not depend on the language in which the tweets are written. The collected tweets are stored in a document-oriented database. It was chosen MongoDB because it has built-in the map-reduce framework processing[3], allowing to use it as data storage and analysis engine for the metrics calculation.

The first centrality measure detects influential tweets directly, regardless of the author. It considers the most relevant features described by [19] for a tweet, i.e., its number of words, its number of replies and retweets, the highest number of followers of the users who interact with it, and the presence or not of a URL on it. The hypotheses that support these features are the following:

---

[2] https://developer.twitter.com/en/docs

[3] https://docs.mongodb.com/manual/core/map-reduce/

- Tweets with many words can include more content, and thus be more informative.
- Tweets with many replies and retweets are interesting for many people.
- Tweets replied or retweeted by popular users are more able to be spread if they have a URL on them, since they can provide a lot of information beyond the Twitter network.

After each hour of collection, three features lists are generated from the accumulated data set:

(a) tweets sorted by number of words (from highest to lowest),
(b) tweets sorted by number of replies and retweets (from highest to lowest), and
(c) tweets sorted by the highest number of followers of the users who replies or retweets the tweet (from highest to lowest).

Let *m* be the number of tweets in the current collection that has a URL, and let $Trank_f(t)$ be a function that returns the ranking position of the tweet *t* in the feature list *f*. The centrality measure *TRank(t)* for each tweet *t* is defined as:

$$TRank(t) = \begin{cases} \frac{\sum_{f=1}^{3} TRank_f(t)}{3}, & \text{if } t \text{ has a URL} \\ m + 1 + \frac{\sum_{f=1}^{3} TRank_f(t)}{3}, & \text{otherwise} \end{cases} \quad (1)$$

i.e., *TRank(t)* is the ranking of the tweet *t* considering its average ranking in the three features lists mentioned above and taking the presence of a URL as priority.

The second centrality measure considered is the *Social Networking Potential* (*SNP*) [22]. This measure does not allow to detect relevant tweets directly, as TRank. Instead, it is used to detect influential users, according to their capacity to spread their tweets through the network. It is defined as follows:

$$SNP(i) = \frac{Ir(i)+RMr(i)}{2} \quad (2)$$

where the *Interaction Ratio*, *Ir(i)*, and the *Retweet and Mention Ratio*, denoted as *RMr(i)*, are defined as

$$Ir(i) = \frac{\#\text{users who have retweeted } i\text{'s tweets} + \#\text{users mentioning } i}{\#\text{followers of } i} \quad (3)$$

and

$$RMr(i) = \frac{\#\text{tweets of } i \text{ retweeted} + \#\text{tweets of } i \text{ replied}}{\#\text{tweets of } i} \quad (4)$$

For each user, *Ir(i)* measures how many different users interact with user *i*, while *RMr(i)* measures how many tweets of *i* generate a reaction from the audience.

The hypothesis is that during a natural disaster event, users with a high Social Networking Potential have a good chance to be opinion leaders, i.e., they are users whose tweets provide valuable information about the event. Therefore, if random tweets are selected from the ones written by these users during the event, then additional relevant tweets that are not considered by the TRank can be obtained.

## Tagging

In order to train the classifiers, a certain sample with the most relevant tweets obtained in the previous step need to be labeled. Two classification criteria are used, namely the *informativeness* and the *information type*, each one of which contains a different number of labels (see Table 1). These criteria are strongly inspired on the taxonomies of [23][4]. Note that the informativeness criterion has fewer labels than the information type criterion. The labeling process is faster for the first criterion, but it provides less information than the second one.

**Table 1** The different labels for the two considered classification criteria.

| **Informativeness** | |
|---|---|
| **Category** | **Description** |
| Informative and pertinent to disaster | Tweet contains useful information that helps to understand the crisis. |
| not informative, but pertinent to disaster | Tweet refers to the crisis, but it does not contain useful information that helps to understand it. |
| not related to the disaster | Tweet does not provide information related to the crisis. |
| does not apply | Tweet is too short, it cannot be read it, among other problems. |
| **Information Type** | |
| **Category** | **Description** |
| affected individuals | information about people affected, including personal updates about oneself, family, or others. |
| infrastructure and utilities | information on buildings, roads, and services that are damaged, broken, restored or operational. |

[4] Note that there are other possible taxonomies, more adequate to other natural disasters [11].

| | |
|---|---|
| donations and volunteering | information on the needs, requests for help or consultations; offers supplies; volunteer or professional services. |
| caution and advice | information on the warnings issued or lifted, guidance and advice. |
| sympathy and emotional support | thoughts, prayers, expressions of gratitude or sadness, etc. |
| other useful information | information related to the crisis but not covered by the above alternatives. |
| does not apply | the tweet is too short, it cannot be read it, among other problems. |

The labeling process was developed with an open source crowdsourcing framework called *Pybossa*[5], a humanitarian computing platform that has been successfully used in the past in computational systems such as the AIDR platform [5]. Pybossa uses digital collaborators to label tweets collected during disasters, emergencies and crisis events. To minimize bias, it was configured with a triple modular redundancy, so that each tweet was evaluated by at least three voluntaries, in such a way that a classification is considered as correct if at least two-thirds agree. This greater-than-one redundancy allows to use the Fleiss' Kappa coefficient in order to assess the agreement between different volunteers. For the information type criterion, for each tweet the labels to be chosen are those in which more volunteers coincide. Figures 2 and 3 show the labeling process using Pybossa (in Spanish) for both classification criteria. Note that the authors of the tweets are omitted, to avoid privacy problems, as well as bias in the labeling process.

[5]http://pybossa.com/


**Figure 2.** An example of a tweet being labeled with Pybossa by a volunteer, under the Informativeness classification criterion. Here only one alternative is allowed.

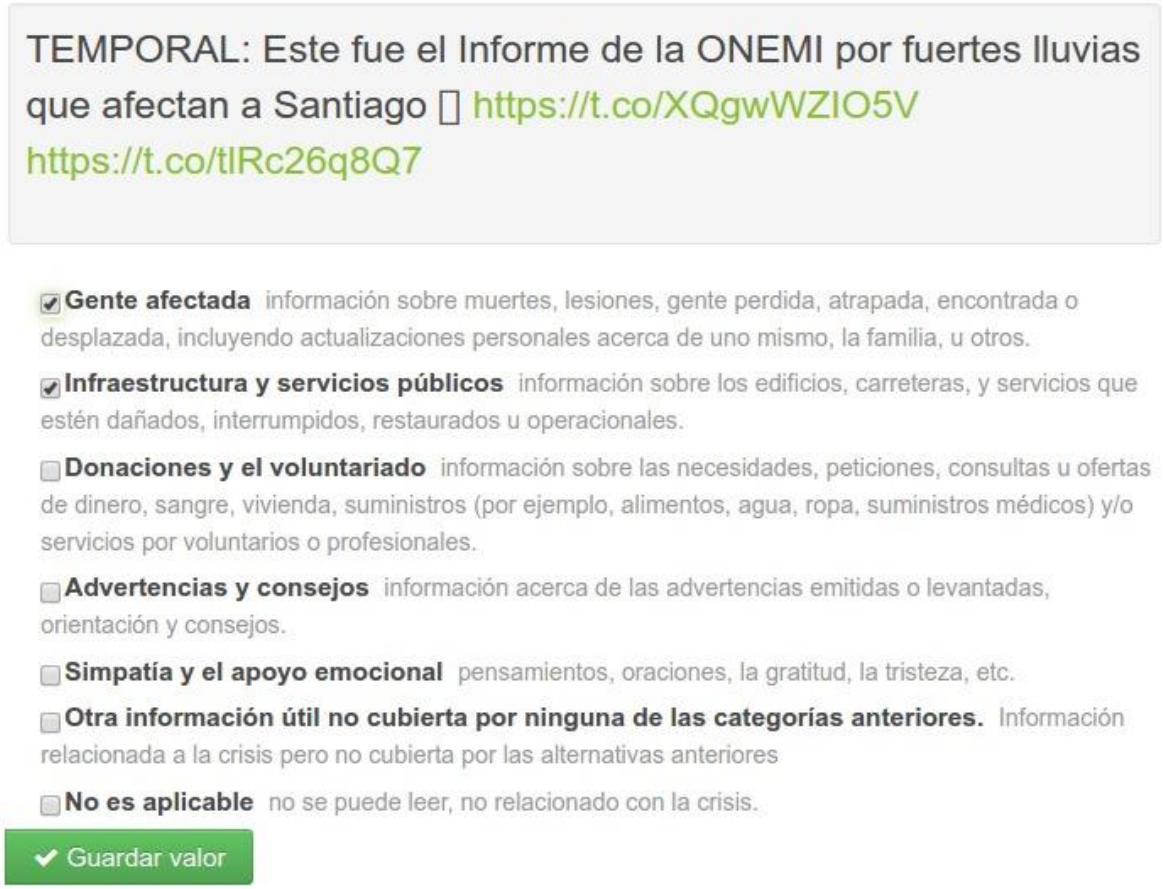


**Figure 3.** An example of a tweet being labeled with Pybossa by a volunteer, under the *Information type* classification criterion. Here more than one alternative is allowed.

Regarding the sample size, it depends on both the required number of volunteers (at least three, in this case) and the estimated time spent by each volunteer to label every tweet. In the context of natural disasters, it is important to remark that time is a scarce resource. Therefore, the number of tweets obtained each hour should be small enough so that the new tweets can be labeled in less than an hour. In total, the whole sample should be labeled in at most a couple of hours. However, the sample size should also be big enough to avoid the bias in the classification. Details of the sample used in this study are shown in the Study Case section.

### Classifier design

The classifiers are trained every hour, adding the last labeled tweets. As stated before, to reduce the volunteer's effort, the tweets that are repeated in the extraction process from one hour to the next one, are labeled only once. Similarly, to diversify the training sample, here repeated tweets are also used only once.

For the feature's extraction process no machine learning techniques were used. Instead, the features used in the literature with the best results were considered [7, 11]. These features are unigrams, bigrams, POS tags [24], hashtags, mentions, number of hashtags and mentions, and length of the tweet.

The classifiers used were decision trees, naive bayes and SVM with the radial-based function, since they are the most used classifiers in the literature [3, 7, 11]. The classification results are obtained from the training set, which changes every hour according to the current tweets collected until that moment.

80% of the data was used for training and 20% for testing. In addition, it was used stratified *k-fold cross-validation* with *k=5*, so that the proportion of the classes for each fold is maintained.

## STUDY CASE: MASSIVE FLOODS IN SANTIAGO OF CHILE AT 2016

On April 17, 2016, due to heavy rains and the negligence of a construction company, Mapocho river overflowed affecting more than 4 million people with cut water service in Chilean Capital [25].

### Collection and extraction

A bag of words was defined at the beginning of the disaster, related to both the place where the accident occurs (Santiago of Chile) and the kind of disaster (massive floods). It is a small bag of words, that emulates a fast response that any non-expert person could have at the beginning of a disaster. The bag of words remains unchanged throughout the whole period of retrieved data. The full list of words is depicted in Table 2.

**Tabla 2:** Bag of words for the massive floods in Santiago of Chile at 2016

| | |
|---|---|
| Keywords | lluvia, santiago, chile, mapocho, inundación, sanhattan, agua, cordillera, río, precipitaciones, santa marta, desborde |

With this approach, 515,763 tweets were successfully retrieved during the first 24 hours of the crisis, from April 17, 2016, 15:09 UTC until April 18, 14:08 UTC. This collection involves 281,132 different user accounts. In addition to the bag of words, the tweets were restricted to those geolocated in Santiago of Chile. This reduced the ambiguity of the sample, and avoided many irrelevant tweets, e.g., those tweets written in Portuguese and related to *Rio* de Janeiro. Finally, a sample set of 134,106 tweets was obtained, with 39,314 users involved. Figure 4 shows the total number of filtered tweets collected each hour, as well as the cumulative amount[6].

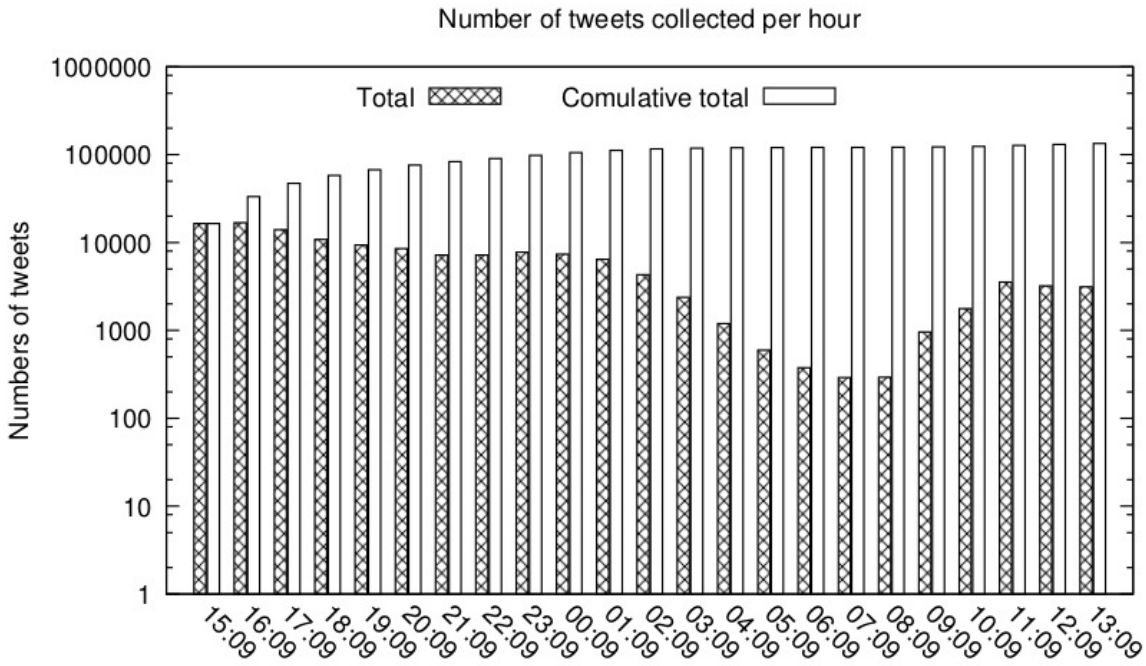


**Figure 4.** Number of tweets collected each hour, and its cumulative amount. The quantities are in logarithmic scale.

From the 134,106 tweets, three different sample sets were obtained:

- Random-sample: 30 tweets obtained randomly per hour.
- IU-sample: 30 tweets obtained randomly each hour from the whole tweets posted by the 50 most influential users at that moment, according to the SNP measure.

[6] This data set is available from this link:
http://citiaps.cl/articles/massive_floods_santiago_2016.zip

- IT-sample: 30 most influential tweets obtained each hour, according to the TRank measure.

At last, leaving out the tweets repeated over time, for each sample were obtained 720, 544, and 404 tweets to be labeled by the volunteers.

## Tagging tasks results

The volunteers used in the labeling process were undergraduate students of Computer Science and Psychology, from the University of Santiago, Chile.

For the labeling process it was considered a Fleiss' kappa coefficient of 2/3, so that every tweet was labeled by three different volunteers, in such a way that a classification was considered as correct when at least two volunteers agree. These agreement results are shown in Table 3. It was also considered the Altman's kappa coefficient assessment shown in Table 4.

For both classification criteria, the IT-sample was the most *consistent*, while for the random-sample, the volunteers disagree more than for the IU-sample. Indeed, only the IT-sample had a good agreement level (0.65) for the first classification criterion and a moderate agreement level (0.42) for the second criterion. The agreement levels for the *information type* criteria are lower because this classification considers more alternatives to be labeled.

**Table 3:** Fleiss' kappa coefficients for both classification criteria, and for each data sample

| | **Random-sample** | **IU-sample** | **IT-sample** |
|---|---|---|---|
| **Informativeness** | 0.49 | 0.54 | 0.65 |
| **Information type** | 0.19 | 0.24 | 0.42 |

**Table 4:** Altman's kappa coefficient assessment [26].

| **Kappa value** | **Strength of agreement** |
|---|---|
| < 0.20 | Poor |
| 0.21 − 0.40 | Fair |
| 0.41 − 0.60 | Moderate |
| 0.61 − 0.80 | Good |
| 0.81 − 1.00 | Very good |

Regarding the pertinence of the samples, Table 5 shows how the volunteers labeled the tweets. For the random-sample, almost one-half of the tweets were considered as not informative, and around 32% did not apply. Note that this could be considered a usual data sample to be labeled in a traditional training approach. However, from both centrality measures, samples with around 70% informative and pertinent tweets are obtained.

**Table 5.** Tweets labeled with both classification criteria

| **Informativeness classification criterion** | | | |
|---|---|---|---|
| **Labels** | **Random-sample** | **IU-sample** | **IT-sample** |
| informative and pertinent | 33.9% | 69.5% | 70.6% |
| not informative, but pertinent | 46.3% | 22.6% | 7.8% |
| not related | 16.3% | 5.9% | 21.0% |
| does not apply | 3.6% | 2.0% | 0.6% |
| **Information type classification criterion** | | | |
| **Labels** | **Random-sample** | **IU-sample** | **IT-sample** |
| affected individuals | 6.7% | 15.4% | 9.8% |
| infrastructure/utilities | 15.1% | 35.4% | 27.6% |
| donations/volunteering | 3.1% | 6.0% | 10.5% |
| caution/advise | 5.8% | 7.9% | 4.5% |
| emotional support | 11.2% | 4.2% | 0.7% |
| other information | 31.9% | 20.2% | 23.5% |
| does not apply | 26.1% | 10.9% | 23.5% |

## Classification results

Regardless of the sample used, there are only considered those tweets that were equally labeled by at least two of the three volunteers, in such a way that the tweets selected provide more consistent information to the training. Thus, for the informativeness

criterion, the random-sample, IU-sample and IT-sample sizes were reduced to 576, 284 and 329 tweets, respectively, while for the informative type criterion, to 469, 234 and 291 tweets, respectively.

It was executed one training per hour (24 hours in total) for each sample (random, IU, IT), and for each type of classifier (SVM, naive bayes, decision trees). In total, $24 \times 3 \times 3 = 216$ executions were done (including within each execution the cross validation of 5 folds). Each execution took just a few minutes of computation.

The average performance results obtained for the different classifiers are shown in Table 6. As stated in the Classifiers section, the precision and recall were used, because they are the metrics used for similar taxonomies [11]. Furthermore, the F-score was included as a test's accuracy measure.

**Table 6.** Average performance of the classifiers.

| **Informativeness classification criterion** | | | | |
|---|---|---|---|---|
| **classifier** | **data sample** | **precision** | **recall** | **F-score** |
| naive bayes | random | 0.70 | 0.73 | 0.67 |
| | IU-sample | 0.82 | 0.77 | 0.78 |
| | IT-sample | 0.82 | 0.81 | 0.81 |
| decision trees | random | 0.55 | 0.56 | 0.56 |
| | IU-sample | 0.74 | 0.76 | 0.75 |
| | IT-sample | 0.77 | 0.81 | 0.78 |
| SVM with radial-based function | random | 0.71 | 0.71 | 0.70 |
| | IU-sample | 0.83 | 0.80 | 0.80 |
| | IT-sample | **0.86** | **0.85** | **0.85** |
| **Information type classification criterion** | | | | |
| **classifier** | **data sample** | **precision** | **recall** | **F-score** |

| | | | | |
|---|---|---|---|---|
| naive bayes | random | 0.41 | 0.39 | 0.36 |
| | IU-sample | 0.39 | 0.34 | 0.34 |
| | IT-sample | **0.54** | 0.55 | 0.51 |
| decision trees | random | 0.31 | 0.32 | 0.30 |
| | IU-sample | 0.31 | 0.32 | 0.30 |
| | IT-sample | 0.51 | 0.58 | 0.51 |
| SVM with radial-based function | random | 0.38 | 0.43 | 0.36 |
| | IU-sample | 0.42 | 0.41 | 0.39 |
| | IT-sample | 0.53 | **0.65** | **0.52** |

As can be seen in Table 6, regardless of the use or not of centrality measures, the results obtained for *Informativeness* are considerably better than those obtained for *Information type*. One reason for these differences could be the greater complexity of the second taxonomy, that is, it has a higher number of labels, and users could select multiple labels for each tweet. Another reason is that some of the categories were poorly labeled by the volunteers, either because of the kind of disaster considered, or because Twitter was not used to disperse that type of information. As a result of the few instances used in some categories during the training, the average precision, recall, and F-score of the classifiers decreased. Examples of this are the donations/volunteering, caution/advise, and emotional support categories, which together represent just around the 20% of the labels for each sample.

Regarding the centrality measures, the TRank (IT-sample) tends to improve in average all the performance metrics. The SNP (IU-sample), on the other hand, improves all the performance metrics for the Informativeness classification criterion, but it fails in the Information type criterion. The best results are obtained for the Informativeness classification criterion using SVM with radial-based function, where the classifiers reach a precision, recall and F-score of around 85%.

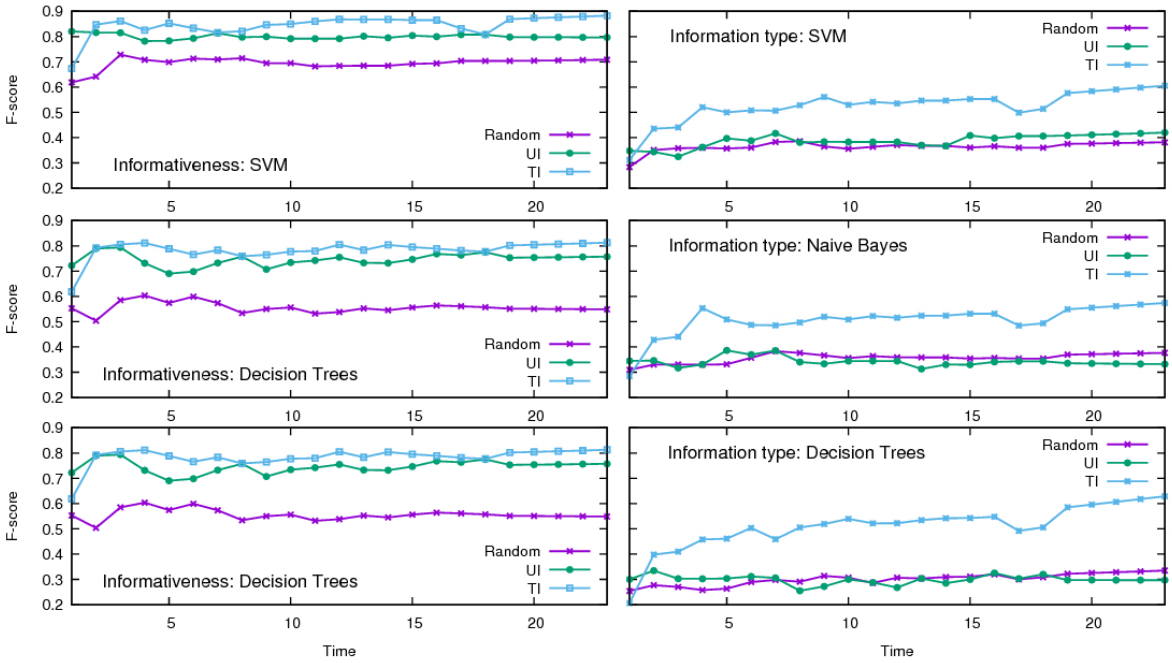


**Figure 5.** F-score obtained for the classifiers through the time, for both classification criteria.

Finally, the F-score obtained for the classifiers over time is illustrated in Figure 5. As mentioned above, the SVM with radial-based function presents the best results for both taxonomies. In general, good quality models are obtained in the first hours of the disaster, especially with the IT-sample. For the Informativeness taxonomy, a good quality model is achieved in the second hour. Then, with the IT-sample, the quality of the classifiers begins to stabilize around an F-score greater than 0.8. Regarding the Information type taxonomy, whereas for the IT-sample the quality of the models improves over time, for the random and IU samples the performance does not seem to evolve. The above indicates that the TRank measure gets tweets that improve the quality of the models over time.

## CONCLUSIONS AND FUTURE WORK

In this paper has been proposed the use of centrality measures to improve the training sample for supervised classifiers of tweets related to natural disasters, using an active learning approach. In this applied context, it is crucial to obtain valuable information quickly, usually starting from a small training data set.

For the considered study case, the new strategy produced a higher consistency in the labeling process, i.e., a greater agreement among the volunteers to classify the tweets (see Table 3). Note that this is not related to the quality of data collection, nor does it imply that it will be fulfilled for other cases of study. Instead, it seems clear that centrality measures produce a higher pertinence of the tweets within the data sample, i.e., these

measures help to provide a collection of tweets considered more valuable by the volunteers. In fact, the centrality measures allowed increasing to more than double the number of informative and relevant tweets (see Table 5).

For the experiments were used two centrality measures: the TRank, used to identify relevant tweets, and the SNP measure, used to identify influential users. While both measures generally improve the results under the Informativeness classification criterion, compared to using a random data sample, the measure that produced substantially better results for both classification criteria was the TRank. This validates the hypotheses defined in the Extraction section for the TRank but relativizes the hypothesis for the SNP measure. Remarkably, to obtain these results, the sizes of the data sets produced by the centrality measures are smaller than the standard sample size. Thus, although the size of the IT-sample is about 60% of the Random-sample, the collected tweets bring more discriminative power to the model. This reduction in sample size translates into a reduction in training time, but more importantly, in human labeling time.

Regarding the performance of the classifiers, the new proposed strategy improves in general the precision, recall, and F-score of the classifiers. The average results are illustrated in Table 6, and the results over time are shown in Figure 5. The best results are obtained for the SVM with radial-based function.

It is important to note the differences of the results obtained for both taxonomies. It seems that the simplest taxonomies allow to train classifiers with better performance, because during the labeling process the volunteers only had to choose among a few categories, so that better training samples are obtained. On the other hand, for more complex taxonomies like Information type, some categories are not sufficiently labeled to obtain a more satisfactory training sample. Hence, while simpler taxonomies can fit into different scenarios, more complex taxonomies may not be appropriate for any natural disaster. In this sense, the development of suitable taxonomies for various types of disasters can be a useful and interesting line of work.

As future work, it would be interesting to apply this methodology to other natural disasters events. Furthermore, although the SNP measure did not produce results as satisfactory as TRank, it could be other influence measures that could give better results. This work leaves open the possibility of looking for other centrality measures more suitable for supervised learning under the active learning approach.

## REFERENCES

[1] S. Vieweg. “Twitter communications in mass emergency: contributions to situational awareness.” Poltrock SE, Simone C, Grudin J, Mark G, Riedl J (eds) CSCW ’12 Computer Supported Cooperative Work, Seattle, WA, USA, February 11-15, 2012.

[2] C. Castillo. “Big Crisis Data: Social Media in Disasters and Time-Critical Situations.” Cambridge University Press, 1 Edition. New York, NY, USA. ISBN: 1107135761 2016.

[3] F. Ofli, P. Meier, M. Imran, C. Castillo, D. Tuia, N. Rey, J. Briant, P. Millet, F. Reinhard, M. Parkan, S. Joost “Combining human computing and machine learning to make sense of big (aerial) data for disaster response”. Big Data 4(1):47–59 2016. DOI: 10.1089/big.2014.0064

[4] M. Imran, C. Castillo, J. Lucas, P. Meier, J. Rogstadius “Coordinating human and machine intelligence to classify microblog communications in crises.” Hiltz SR, Plotnick L, Pfaf M, Shih PC (eds) 11th Proceedings of the International Conference on Information Systems for Crisis Response and Management, University Park, Pennsylvania, USA, May 18-21, 2014.

[5] M. Imran, C. Castillo, J. Lucas, P. Meier, S. Vieweg “AIDR: artificial intelligence for disaster response.” Chung C, Broder AZ, Shim K, Suel T (eds) 23rd International World Wide Web Conference, WWW ’14, Seoul, Republic of Korea, April 7-11, 2014.

[6] M. Imran, C. Castillo, F. Diaz, S. Vieweg “Processing social media messages in mass emergency: A survey”. ACM Comput Surv 47(4):67:1–67:38, 2015. DOI: 10.1145/2771588

[7] Avvenuti, M., Cresci, S., Del Vigna, F., Fagni, T., & Tesconi, M. “CrisMap: a big data crisis mapping system based on damage detection and geoparsing”. Information Systems Frontiers, Springer, 2018.

[8] Kejriwal, Mayank, and Yao Gu. “A pipeline for post-crisis Twitter data acquisition.” arXiv preprint arXiv:1801.05881, 2018.

[9] Zheng, Xin, Aixin Sun, Sibo Wang, and Jialong Han. “Semi-supervised event-related tweet identification with dynamic keyword generation.” Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 1619-1628. ACM, 2017.

[10] S. Karimi, J. Yin, C. Paris “Classifying microblogs for disasters.” Culpepper JS, Zuccon G, Sitbon L (eds) The Australasian Document Computing Symposium, ADCS '13, Brisbane, QLD, Australia, December 5-6, 2013.

[11] M. Imran, S. Elbassuoni, C. Castillo, F. Diaz, P. Meier “Extracting information nuggets from disaster-related messages in social media.” 10th Proceedings of the International Conference on Information Systems for Crisis Response and Management, Baden-Baden, Germany, May 12-15, 2013.

[12] F. Riquelme, P. Gonzalez-Cantergiani “Measuring user influence on Twitter: A survey.” Inf Process Manage 52(5):949–975, 2016. DOI: 10.1016/j.ipm.2016.04.003

[13] F. Morone, HA. Makse H “Influence maximization in complex networks through optimal percolation.” Nature, 2015. DOI: 10.1038/nature14604

[14] D. Kempe, JM. Kleinberg, E. Tardos “Maximizing the spread of influence through a social network-” Getoor L, Senator TE, Domingos PM, Faloutsos C (eds) Proceedings of the

Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 24 - 27, 2003.

[15] J. Golbeck, JA. Hendler “Inferring binary trust relationships in web-based social networks.” ACM Trans Internet Techn 6(4):497–529, 2016. DOI: 10.1145/1183463.1183470

[16] X. Song, BL. Tseng, C. Lin, M. Sun “Personalized recommendation driven by information flow.” Efthimiadis EN, Dumais ST, Hawking D, J¨arvelin K (eds) SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, August 6-11, 2006.

[17] J. Li, W. Peng, T. Li, T. Sun, Q. Li, J. Xu “Social network user influence sense-making and dynamics prediction”. Expert Syst Appl 41(11): 5115–5124, 2014. DOI: 10.1016/j.eswa.2014.02.038

[18] F. Riquelme, P. Gonzalez-Cantergiani, X. Molinero, M. Serna “Centrality measure in social networks based on linear threshold model.” Knowledge Based Systems 140:92–102, 2018. DOI: 10.1016/j.knosys.2017.10.029.

[19] Y. Duan, L. Jiang, T. Qin, M. Zhou, H. Shum “An empirical study on learning to rank of tweets”. Huang C, Jurafsky D (eds) COLING 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China, Tsinghua University Press, 2010.

[20] C. Francalanci, A. Hussain “Discovering social influencers with network visualization: evidence from the tourism domain.” J of IT & Tourism 16(1):103–125, 2016. DOI: 10.1007/s40558-015-0030-3

[21] A. Olteanu, C. Castillo, F. Diaz, S. Vieweg “CrisisLex: A lexicon for collecting and filtering microblogged communications in crises.” Adar E, Resnick P, Choudhury MD, Hogan B, Oh AH (eds) Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014.

[22] I. Anger, C. Kittl “Measuring influence on Twitter.” Lindstaedt SN, Granitzer M (eds) I-KNOW 2011, 11th International Conference on Knowledge Management and Knowledge Technologies, Graz, Austria, September 7-9, 2011.

[23] A. Olteanu, S. Vieweg, C. Castillo “What to expect when the unexpected happens: Social media communications across crises.” Cosley D, Forte A, Ciolfi L, McDonald D (eds) Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW 2015.

[24] CD. Manning, M. Surdeanu, J. Bauer, J. Finkel, SJ. Bethard, D. McClosky “The Stanford CoreNLP natural language processing toolkit.” Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, 2014.

[25] RT “Massive floods hit Chile: Power cuts, mine shut, 4mn people left with no fresh water.” April 19, 2016. March 01, 2019 URL: https://www.rt.com/news/340141-chile-floods-millions-no-water/

[26] D. Altman. “Practical statistics for medical research”. Chapman and Hall/CRC 1 Edition, p. 404. London, UK. ISBN: 9780412276309 1991