PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
SCHOOL OF ENGINEERING

A TWITTER-BASED CITIZEN CHANNEL FOR NATURAL DISASTER SITUATIONS

ALFREDO COBO OBERPAUR

Thesis submitted to the Office of Research and Graduate Studies
in partial fulfillment of the requirements for the degree of
Master of Science in Engineering

Advisor:
JAIME NAVÓN C.

Santiago de Chile, January 2015

PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
SCHOOL OF ENGINEERING

A TWITTER-BASED CITIZEN CHANNEL FOR NATURAL DISASTER SITUATIONS

ALFREDO COBO OBERPAUR

Members of the Committee:
JAIME NAVÓN C.
DENIS PARRA S.
CARLOS ORREGO U.
RODRIGO CIENFUEGOS C.

Thesis submitted to the Office of Research and Graduate Studies
in partial fulfillment of the requirements for the degree of
Master of Science in Engineering

Santiago de Chile, January 2015

To my family and everyone who helped me on this long journey.

ACKNOWLEDGEMENTS

First I would like to thank my advisor Jaime Navón. He constantly encouraged me to keep going and gave me guidance and support in so many aspects. I definitely wouldn't have entered a post graduate program without his advice.

My second advisor Denis Parra for his patience and time. His dedication was fundamental get this work done in a reasonable time.

My office colleagues and lifetime friends for all the great conversations and encouragement during our stay in the department. It's an honor to be a part of the family, even if some are continents away.

To all the other friends who leaved a mark in the department. Especially the people living in the near "fishbowls", I had very enjoyable times with you.

My parents, brothers, and by extension all of my wide family, for their love, patience and support.

Finally I thank God for giving me the strength to surpass lots of obstacles during all these years in the university.

This work has been partially supported by the National Research Center for Integrated Natural Disaster Management CONICYT/FONDAP/15110017.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

During the first 24 hours after an earthquake strikes, there is a huge need for information from the affected citizens. Most of this information comes from traditional media, either directly from official sources or mediated through journalists of radio and TV. During recent years the online social networks (in particular Twitter) have become an important alternative information channel, but the amount and diversity of messages poses the challenge of information overload to end users. Under this scenario, the goal of our research is to develop an automatic classifier of tweets to feed a mobile application that reduces the difficulties that citizens face to get relevant information in an earthquake situation. In this paper, we contribute by presenting a prototype of our application and the details of our classification model. Although previous works have presented prediction models to find relevant and credible messages, they provide few details about the effect of class imbalance and number of latent dimensions. By using a dataset from the Chilean earthquake of 2010, we first build and validate a ground truth, and then we contribute by presenting in detail the effect class imbalance and dimensionality reduction over 5 classifiers, showing that some of them are barely affected and others, though showing good performance under specific conditions, are extremely susceptible to variations on the parameters. Overall, we found that random forest is the most stable model when we tested its behavior over different conditions.

Keywords: Social networks, Twitter, Natural disasters, Context awareness, Machine learning

RESUMEN

Durante las primeras 24 horas en que un terremoto ocurre, se genera una gran necesidad de informarse por parte de la ciudadanía. La mayor parte de esta información proviene de los medios más comunes, desde fuentes oficiales o mediada a través de periodistas de radio y televisión. Durante los últimos años las redes sociales (en particular Twitter) se han posicionado como un importante canal alternativo de información. Chile es un país con alta actividad sísmica, por lo que este tipo de desastre natural ocurre bastante frecuentemente. Por otra parte existe en este país una alta penetración de smartphones y es también uno de los líderes en el uso de redes sociales. Esto explica el gran aumento de ciudadanos que se conectan a Twitter y no a la radio o la televisión cuando algo sucede. Pero aun cuando este medio es rápido y bidireccional, es bastante ruidoso. En este trabajo describimos el desarrollo de un clasificador automático que utiliza algoritmos de aprendizaje de máquinas para filtrar el flujo de mensajes, seleccionando aquellos que son considerados ”relevantes” o ”relacionados” con el evento de desastre natural. En este trabajo se muestran detalles acerca del rendimiento de este clasificador. El modelo fue entrenado con un set de datos extruidos desde Twitter que fueron enviados durante y después del terremoto de magnitud 8.8 ocurrido en el 2010. Los mensajes seleccionados se usan para alimentar a una aplicación web móvil que los ciudadanos pueden acceder desde sus teléfonos.

Palabras Claves: Redes sociales, Twitter, Desastres naturales, Concienciación, Aprendizaje de máquinas

1. INTRODUCTION

Chile is well known for being one of the most seismic countries of the world. Having natural disasters related to this activity is quite common, due to its geographic location over the Nazca plate and thus being part of the ring of fire. This condition is shared with Japan, New Zealand and other countries around the Pacific Ocean. One of the memorable earthquakes that stroke recently was the one of the 27 February of 2010, which was located near the coast of Pelluhue in the Maule region. This movement was quite catastrophic, because it brought also a tsunami that hit a wide portion of the coast. Considering the direct consequences more than 700 people lost their lives (Fritz et al., 2011). These events were a difficult test on public agencies, putting pressure to the authorities and creating a sense of need to respond better during these hazards.

Information channels are relevant for Chilean society. The citizens inform themselves everyday mainly through radio, newspapers and television. Hence when a meaningful event arises people tend to tune in to traditional sources to know what is happening. These channels transmit relevant and trustworthy information from a spectator point of view. In fact this is one of their strengths, having a connection with the public. However the interaction that they have is mainly one sided, due to natural restrictions of the channel. When a disaster event strikes people tend to listen to the radio, because traditionally it was the fastest medium to obtain first hand information.

Besides traditional media we have online social networks, that are having a great impact on Chilean society. These facilitate bidirectional communication among their users. This new information channel in which any person can contribute at any time is known now as social media. The most relevant microblogging network to get a sense of what's happening at a given time is Twitter, which is also one of the most used in the country.

When a disaster strikes we usually see that traditional media has not much information to communicate in the first hours (Puente, Pellegrini, & Grassau, 2013). They just transmit

anecdotal information about what's happening to their reporters or near the radio station. However at the same time people are messaging constantly with their smartphones over their online networks, and a great part of this information is really not used in a broad sense.

In this document we attempt to construct a new citizen channel based on Twitter to contribute to situational awareness in the context of an earthquake. The scope is to build classification algorithms that will provide the messages to this channel. Therefore our goals are the following.

- Create several classifiers that will automatically filter relevant messages to the citizen channel.
- Test the performance and behaviour of the model with different train conditions.
- Select the best classifier among them, being this the that will be inserted in the channel architecture.

This channel will be accessed through special applications installed in the citizens smartphones. Furthermore, the same applications will allow the citizens to make relevant contributions to this channel.

Before getting into the channel construction itself it is convenient to have some understanding of Twitter, how messages are collected, and the fundamentals of machine learning techniques that were used to classify the messages. We also explain that this research is part of a bigger research project carried out by the Natural Research Center for Integrated Natural Disaster Management (CIGIDEN), that addresses natural disaster from many different perspectives.

1.1. CIGIDEN

CIGIDEN (Natural research center for integrated natural disaster management) is a research center created with the purpose of developing the knowledge and to transfer the

technology needed to address the complexity of the challenges involved in situation of natural disasters. These challenges include technical, social and policy that go beyond reducing direct casualties, brought on by the interactions between the natural, human and built environment. Preparedness, response, recovery and mitigation are the phases of the cycle. The Center has an integrated and interdisciplinary research approach oriented towards reducing the risk of natural disasters, in collaboration with state organisms responsible for the protection of infrastructure and the wellbeing of society.

Our work is related with specific objectives of one of the research project called "Human behavior and communication strategies during the disaster response phase". The general objectives are the following: To study the human response to external messages during the emergency response phase of a natural disaster. To generate suitable recommendations and protocols for authorities, media, and the citizens regarding information during the emergency response phase of a disaster.

The specific objectives are:

- To describe how psychosocial factors influence individual and collective behavior during the emergency response phase of a natural disaster.
- Specification of communication protocols, warnings and signs standards for the dissemination of response information by authorities.
- To build classifiers, based on machine learning algorithms, to put in the hands of the citizens a relevant flow of disaster-related information posted in the social networks by the citizens themselves.

Our work is fundamental for completing the last specific objective. To design and build a classifier that could be used at the kernel of a citizen channel is our main goal.

1.2. Twitter

1.2.1. Characteristics and mechanics

Twitter is an online social network created in 2006 and one of the most successful ones in Chile. The network has more than 280 million active users per month who send more than 500 million messages per day.

The basic communication element of this network is the tweet or short message, which can have a maximum length of 140 characters. Within this message the user can write anything they want, mention other users and include relevant hashtags. Hashtag is a mechanic that consist on putting the # symbol before each keyword, which are used to add relevance to the message. In some way the hashtag is an index that people use to search and interact with a wider audience. Twitter has also a basic mechanic to make connections, to follow. A user can follow multiple accounts to read the messages that are broadcasted by them. In this way a personalized stream of messages is build, which can be read whenever the builder wants.

When a tweet is read, the reader has three options, to do nothing about it, favorite it or retweet it. To favorite a message allows accessing it later in an easier way, it behaves like a bookmark. To retweet a message is the main feature of this network. It allows the user to rebroadcast the tweet to their followers. This mechanic allows a message to reach an exponential amount of users quickly (Kwak, Lee, Park, & Moon, 2010).

To better introduce the network let us propose an analogy. This service can be thought as a network of radio broadcasters, in which every listener can hear the channels they want and produce the content they want. They construct a personal stream of radio programs and messages on their own, and when something appears to be interesting enough they can retransmit it again through the channel to all their listeners.

Another advantage of Twitter as a news channel is that it is bidirectional. Citizens not only can get the important news about an event but they can become reporters themselves by making contributions to the Twitter stream.

However it is known that the influence of every account is not the same. There are notable differences in the efforts needed to spread a message between influential and common users (Morales, Borondo, Losada, & Benito, 2014). This difference notes an unfair advantage from messages emitted from influential accounts. So in case of a natural disaster a normal user would probably not know about the existence of many relevant messages, just given the topology of the network.

All of these characteristics are enhanced by the fact that the network is accessible through mobile devices, so when something happens the people tend to look as a first information source their Twitter stream. In the same way users tend to post the information first on this channel through their devices, including photos or links to other pages to add more context.

1.2.2. Public Access

Twitter provides an interface to their network data through an API, allowing developers to gather information about streams, users and messages. In fact lots of studies have been held using this source of data.

At first there was just one big public interface to gather historical data. This meant that the developer could only know what had been tweeted for the past couple of minutes. In order to gather the data a repetitive program had to be set up, because it had a lot of restrictions including requests per minute and number of messages gathered on each request. First studies based on Twitter activity used this type of techniques for the data collection (Vieweg, Hughes, Starbird, & Palen, 2010).

Recently, a new interface was offered known as streaming API. This method allowed developers to listen to Twitter constantly. Basically the developer connects a permanent listener to the service, and whenever the message matches the stream query criteria it appears on the channel. Afterwards the tweet can be processed and saved into a new database. The query used by the developer can include words, hashtags, places or specific users that should be listened to. Obviously these connections have limitations, including

not having more than one stream per application and a hard limit of 1% of the total stream, known as the 'firehose'. Luckily the subset is known to be fairly random (Morstatter, Pfeffer, Liu, & Carley, 2013).

Because of the "retweet" mechanism, Twitter is known to be extremely fast. A posted message can reach all over the world in just a few seconds. Most people get information about important news (for instance the death of Michael Jackson) through their Twitter timelines well before they listen to the event by radio or TV. This is important in situations of natural disaster when there is a need to get the information to the citizens as fast as possible. When a stream is set to listen to Twitter messages the influence of the accounts over the messages is flattened, and the retweet factor becomes less relevant (Taxidou & Fischer, 2014). This does not take away the speed of the channel.

The main problem of Twitter however is that when a general stream is set up one can quickly determine that this is a very noisy channel. If we consider just Spanish words and listen to messages sent from Chile, the tweets received are about just anything, because the stream does not discriminate inherent characteristics or context. It's similar to listen to a conversation of a full stadium all at once, making it very difficult to identify relevant messages without paying great attention.

One of the challenges of social media for emergency management is the information overload (Hiltz & Plotnick, 2013). Given this premise, the value of our contribution appears within the context of natural disasters. Our model will reduce the information overload of the Twitter stream, providing a cleaner and simpler channel for the users consumption.

1.3. Machine learning algorithms

Beeing a consequence of the noisy nature of the Twitter streams, information overload was considered as a serious problem (Hiltz & Plotnick, 2013). Therefore machine learning techniques were evaluated, because they could allow us to filter the stream and to

address the information overload issue. This is especially critical for the people who are not familiar with the mechanics of the social service.

Machine learning algorithms in a broad sense are basically processes that can effectively learn from data. This means that they are models which can be trained in order to make prediction or take decisions over new data. Normally to solve general problems a developer explicitly programs an analytical solution, but there are situations where this approach is nearly impossible to apply. These algorithms are used when this type of problems appear, managing to reduce the complexity of a classification task (Flach, 2012).

Normally to train these models a ground truth is needed. From this base set the main characteristics are first extracted as features. Once the features are represented, the ground truth is divided in two parts, a training and a test set. With the training set the model is adjusted to predict over it and generalize to do it over new data. With the test set the model is scored, using it as an input and comparing the results afterwards.

In a broad sense there are two approaches of these algorithms, supervised and unsupervised learning. Supervised learning models are used over labeled data. This means that the dataset needs to have the pretended outcome beforehand as a value. Unsupervised learning is used when there are unlabeled data and hidden structures are wanted to be discovered, so the model tries to identify patterns from the given dataset (Flach, 2012).

A supervised learning approach is more suitable to our needs. We want a specific group of messages from Twitter and discard the rest of them. We can only know about the relevance of a message in a specific scenario by asking real people what their opinion is. This information can be added to a dataset to utilize this type of modeling.

The main techniques in the supervised learning sector are decision trees, linear regressions, naive Bayes variations, neural networks, support vectors machines (SVM), or ensembles like random forests. From all these techniques it is very difficult to know beforehand what the best one will be, or if any of them will perform above a baseline (Saad,

2014). For doing this is critical to know which criteria will be used, which scores are more important and to evaluate the risks and problems of each of them.

1.4. Risks and limitations

The main risk when a machine learning algorithm is trained is the overfitting. Overfitting means that the model will not generalize very well over the labels given in the dataset, and it will consider instead very specific cases. So when new data are given to the model, the performance will be very poor. To avoid this problem robust algorithms are needed and to use a balanced dataset.

Other problem is the dimensionality of the dataset. In our approach we need to use text processing, so it's logical that the number of dimensions will be very high. The quantity will be determined by the vocabulary used over the whole training set. Some algorithms simply cannot operate on high dimensional spaces or tend to converge very slowly on them, resulting in very impractical methods. This sparse matrix used for text processing is called a tf-idf or term frequency–inverse document frequency, which takes every word in a vocabulary and obtains a value based on frequency that represents how much information is gained with it for each of the documents.

This risk of high dimensionality can be reduced using dimensionality reduction algorithms. One of them is performing a latent semantic analysis (LSA), which use a singular value decomposition (SVD) strategy to extract a representative number of rows or documents from the tf-idf matrix. After that new queries can be computed over the reduced space (Landauer, Foltz, & Laham, 1998).

The main problem to this approach is that when a new document is being fit to this dimension reduction model, if the words contained did not appear in the tf-idf matrix the weights assigned are zero. Therefore the new document hasn't got a chance to being classified as a relevant message.

On the other hand we have latent dirichlet allocation (LDA), which reduces the dimensionality and allows unseen documents, adding a small random weight when the terms are not being seen, thus giving it a chance to be classified. Additionally this type of models can be upgraded with new documents and are known for being adequate for working with streams (Hoffman, Bach, & Blei, 2010).

For training the model it was very important to consider the noisy nature of Twitter. It was critical that the classifier performs well in those spaces. Therefore custom noise was added to the training stage, so the negative labels could have more variance, thus reducing the overfitting for this class and allowing to discard messages in a wider range.

1.4.1. Search and scoring

There are some basic scores that can be considered to evaluate the performance of a machine learning model. Precision, recall, accuracy, f1 score and ROC AUC were the ones used in our research. Each one of them evaluates a metric assuming a contingency table is build from the results of each model. Thus it is usable to know in advance what does each one mean.

- Precision is the positive predictive value, or the proportion of retrieved documents that are relevant. It tells us how good the model is in predicting a relevant message.
- Recall is the true positive rate, or the proportion of relevant documents that are successfully retrieved. It tells us how many relevant messages could be obtained from all the initial relevant messages.
- Accuracy is the proportion of all the correct results among all the examined messages. If all the messages were correctly determined the value is 1.
- F1 score is the harmonic mean of precision and recall.
- From the receiver operating characteristic (ROC) plot the area under the curve (AUC) can be measured. It basically tells the probability that a classifier will

rank a randomly chosen relevant message higher than a randomly chosen negative one.

To determine the best possible model for the citizen channel a grid search was needed. This method is to simply list several combinations of parameters and to search in all the possible combinations of them, considering some criteria of evaluation. Also whenever a model is trained a k-fold cross-validation has to be performed, dividing the training set into a number of folds and training systematically, leaving one of the folds out for testing each time. Finally the best performing model on average is selected.

1.5. Dataset construction

As the most plausible approach for the filtered stream was a supervised learning algorithm, a labeled dataset had to be constructed and validated. To do this we needed to select an odd number of individuals to manually classify our data. In this way a democratic agreement could be used to determine the ground truth of each message. In our case the number of different classes was two, so with three classifiers a simple majority criteria could be used.

Even though an agreement is possible, the labeled set had to be validated using statistical analysis methods. Generally agreement analysis is performed to know if the people are classifying with a similar criteria. When this is not the case one can assume that the rules are not clear or that the messages do not provide enough information to make a clear statement. Therefore is critical to have a clear set of rules and a preprocessed dataset, so the risks of having a low agreement are reduced.

To do this agreement analysis we used a set of tests. Raw agreement was first analyzed, to observe in general whether the opinions of the classifiers were aligned. Afterwards statistical significance was calculated using a bootstrap process, to test marginal homogeneity (Efron & Tibshirani, 1994). Finally intraclass correlation and Fleiss kappa

metric were computed. If marginal homogeneity was found, these statistical test would be valid too.

Intraclass correlation gives a score of how much homogeneity, or consensus, there is in the labels given by manual classifiers (McGraw & Wong, 1996). The interpretation of this score is quite general, but commonly over a 0.7 is known to be strong agreement indicator. Fleiss kappa indicates the reliability of agreement between raters. A score over 0.6 is considered as a substantial agreement and above 0.8 is considered nearly perfect (Landis & Koch, 1977).

2. A TWITTER-BASED CITIZEN CHANNEL FOR NATURAL DISASTER SITUATIONS

The following chapter is a paper, submitted for publication in the Journal Disasters.

2.1. Introduction

Chile is a country that is frequently punished by natural disasters. It suffered a major earthquake (8.8 Richter) not far ago in 2010 and new one (8.2 Richter) in 2014. Sometimes these events are followed by Tsunamis that strike villages situated along the very long coastal line of the country. In the minutes immediately after the event, affected people experiment an urgent need to get information of different kinds. First about the event itself, how big it was, where was the epicenter, and so on. People also needed to know about their relatives and friends. Sometimes the person needs specific information about where he can get help, when the basic services will be restored, etc. Historically, the main source of information in the minutes after a quake was the radio. There is however a lapse of time where the radio has no much information to communicate and they just broadcast anecdotal information about what is happening near the station or where some of the reporters happen to be at the moment (Puente et al., 2013). But a new channel of information has emerged in the last few years. People are turning to online social networks and in particular to Twitter to learn what is going on and organize themselves. This is especially true among youngsters who carry their smartphones at all times (Valenzuela, Arriagada, & Scherman, 2012). Twitter has two big advantages as a news channel over the radio: first a very fast propagation speed (Kwak et al., 2010) and second it is bidirectional, that is, everyone can contribute with his own contents to the message stream. In Chile there has been a very fast penetration of mobile devices and a large segment of the population owns a smartphone (Ureta, 2008). Furthermore, Chilean citizens seems to like online social networks a lot (one of the fastest growing rates in the world).– Chilean authorities have taken notes of both the rising popularity of Twitter and the ubiquity of smartphones

and they have open Twitter accounts to inform the citizens. For instance, ONEMI (National Office for Emergencies), SHOA (Army Hydrographic Service) and others tweet every time an important event occurs. One problem to adopt Twitter as a main source of emergency news is that for an important segment of the population it is complicated. For a senior citizen, to create an account and then to follow the relevant sources, not to mention the possibility to write his own messages can be near impossible. To face this problem we built a friendly web application that lower the technology barriers. But the main problem is that Twitter is a very noisy channel, that can produce information overload (Hiltz & Plotnick, 2013). Together with that message from ONEMI the user will be getting many non-relevant messages in his timeline that could hide the important ones. Using machine learning algorithms, we designed and build a piece of software that is able to classify a message as ”relevant” or ”non-relevant” where relevant are the ones that contain some information relative to an earthquake event. To train the classifier we used a stream of Twitter messages that was captured the minutes after the major chilean earthquake of 2010. To this end, the training set of messages were classified as relevant or non relevant by human classifiers so this can be used as a ”ground truth”. Our classifier is the most important piece of the citizen channel solution architecture that affected people can access through their mobile devices, to get relevant information and also to post new disaster related information that can be used by others. Figure 2.1 shows the architecture of the system and 2.2 provides a few snapshots from the mobile web application in action (Molina, 2015).

The rest of the paper is organized as follows. In the next section we provide a review of the most relevant literature and related work. In Chapter 2.3 we explain the methodology we used to build the automatic classifier. Chapter 2.4 presents the results we get when we put our classifier to close in Chapter 3 with conclusions and future work.

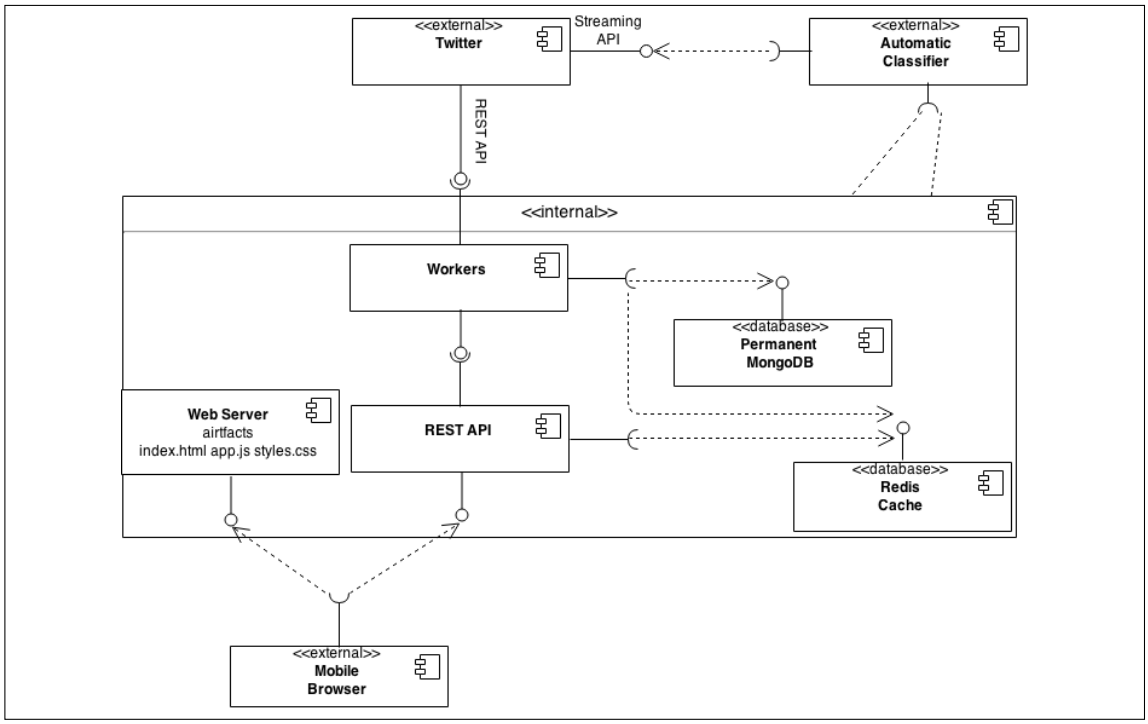

FIGURE 2.1. Role of the Classifier in the general system architecture.

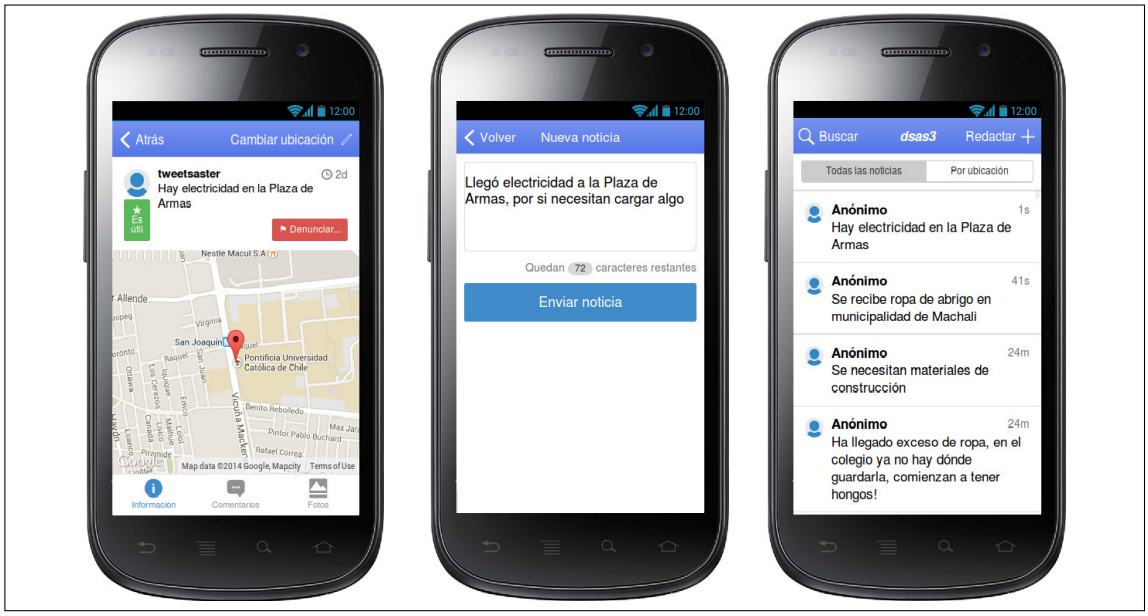

FIGURE 2.2. The mobile web application using the citizen channel.

2.2. Related work

In order to present the related work we divided them into several groups. First manual classification research and post processing are presented. Other feature approaches and analysis are shown. Then several tools for disaster management are reviewed, finishing with our particular goals.

Manual classification. There have been many attempts to capture and process the twitter messages generated in situations of natural disasters. The first attempts were simple manual classification. Vieweg et al. (Vieweg et al., 2010) manually classified situational messages about Oklahoma Grassfires of April 2009 and the Red River Floods that occurred in March and April 2009. Imran et al. (Imran, Elbassuoni, Castillo, Diaz, & Meier, 2013) did the same process for the Joplin tornado of 2011 but he used crowdsourcing services afterwards to perform automatic classification using machine learning techniques. Nevertheless we are not aware of any real time Spanish language automatic classification attempt needed to feed a citizen information channel for natural disaster events.

Post processing. Regarding post processing of the messages there is also relevant work. Castillo et al. (Castillo, Mendoza, & Poblete, 2011) assessed the credibility of the messages, while Mendoza et al. (Mendoza, Poblete, & Castillo, 2010) classified dissemination of false rumors and confirmed news of the Chilean 2010 earthquake. We addressed relevance of messages according to a certain criteria, using post processing similar to these works.

Feature generation approaches. There have also been attempts to improve the performance of the algorithms by generating new features. For example Gimplel et al. (Gimpel et al., 2011) used part of speech recognition in English while Kouloumpis et al. (Kouloumpis, Wilson, & Moore, 2011) and Liu et al. (Liu, Li, & Guo, 2012) used several tools such as sentiment analysis to add features to the training set. We also experimented with a variety of features, focused on content. We included also a proposed one. The time gap between the instant of the event and the moment when each message was posted.

Network features. Another type of analysis can be made over the a generated network graph. Wu et al.(Wu, Hofman, Mason, & Watts, 2011) examined the information generated and consumed by twitter users, resulting in distinguishable groups and high concentration. Lee et al. (Lee, Mahmud, Chen, Zhou, & Nichols, 2014) studied the likelihood of a user to make a retweet to spread information. This type of research was not made over the data set, considering the structure of the network important, but not essential to our analysis.

Tools for disaster management. There have been several attempts in constructing frameworks to deal with the information overload produced by twitter messages (Hiltz & Plotnick, 2013). Most of these frameworks provide ways to filter relevant messages in order to add situational awareness. Caragea et al. (Caragea et al., 2011) made a framework to aid NGOs and first responders to record and classify and aggregate data from the Haiti 2010 earthquake. Power et al. (Power, Robinson, & Wise, 2013) characterized tweets as a fast source of information for situation awareness. Abel et al. (Abel, Hauff, Houben, Stronkman, & Tao, 2012) made a tool to explore information from Twitter and other web streams. Middleton et al. (Middleton, Zielinski, Necmioğlu, & Hammitzsch, 2014) developed a decision support system to give awareness in earthquake and tsunami events. Morstatter et al. (Morstatter, Kumar, Liu, & Maciejewski, 2013) created a system to gain knowledge and visualize events. More specific frameworks have been used to detect earthquakes. Robinson et al. (Robinson, Power, & Cameron, 2013) used word frequencies to detect bursts in Australia, while Walther et al. (Walther & Kaisser, 2013) detected real world events and geolocalized them. Sakaki et al. (Sakaki, Okazaki, & Matsuo, 2010) predicted earthquake locations and typhoon trajectories in order to alert the population about incoming situations. Recently research has been done in crowdsourced tagging, so the algorithms can be repeatedly trained over time. Imran et al. (Imran, Castillo, Lucas, Meier, & Vieweg, 2014) proposed a framework to actively tag messages during an event. These frameworks are generally designed to help the official agencies and tend to forget

that the citizen are also in need of situational awareness. In our research the final product would be targeted to them, affection our assumptions and decisions.

Previous research efforts served as a guideline for our work. The main difference was the context, a Spanish channel for earthquake situations guided to Chileans. The focus of our efforts was made towards the citizen and not necessarily to the official agencies situational awareness, gathering relevant messages and delivering them in real time. The challenge was to know if a classifier of these characteristics could be constructed based on content features.

2.3. Methodology

2.3.1. Building the ground truth

We wanted to train a model that could be able to predict ”relevant” messages, taking a supervised learning approach. The tweets used for this training were obtained from a known earthquake dataset (Mendoza et al., 2010), which were sent before and after the critical event (2010/02/27 03:34:08). They started at midnight of the 27th February and ended at midnight of the 2 of March. These data were not labeled beforehand, and the relevance of messages wasn't explicit. Therefore we built a ground truth performing manual labeling so it could be used to train supervised learning classifiers.

Due to limited resources and time constraints we gathered a subset of the whole dataset, so we could have a fine control over each message. It was important also to have control over the people that were going to classify, because of the Chilean context and terms that appeared in the data.

A subset of 5000 tweets was obtained using systematic sampling, to have a more homogeneous set over time. After this we removed the ones which were not written in Spanish. First with language processing tools like textcat and tm packages , followed by a manual inspection of every message. The processing was done due to the lack of

http://CRAN.R-project.org/package=tm
http://CRAN.R-project.org/package=textcat

language information in the set. Subsequently similar phrases were removed using 10% Lavenshtein distance as minimum tolerance. Afterwards a manual review was done, to ensure low redundancy. All of these resulted in a set of 2187 messages: 524 tweets for day one, 529 for day two, 618 for day three and 516 for day four.

Once the base set was defined, we provided a known criteria to label each message (Imran et al., 2013). If the tweet belonged to one of the following categories it corresponded a "true" label, "false" otherwise. Meaning for each one that it was or wasn't relevant to the earthquake situation.

- **Caution and advice.** The message conveys/reports information about some warning or a piece of advice about a possible hazard of an incident.
- **Casualties and damage.** The message reports the information about casualties or damage done by an incident.
- **People missing, found, or seen.** The message reports about the missing or found person affected by an incident or seen a celebrity visit on ground zero.
- **Information source.** The message conveys/contains some information sources like photo, footage, video, or mentions other sources like TV, radio related to an incident.

Using these categories 6 people classified the dataset, dividing them in groups to produce 3 labels for each message.

2.3.2. Validation of ground truth

In a democratic way we wanted to have a unique label for each tweet, selecting with a simple majority criteria. But before doing this we needed to be sure that the dataset was reliable enough. To prove that the labels were mostly correct we needed to analyze the people agreement.

The raw agreement, calculated as the proportion of agreements divided by all the possible cases of agreement, was 74.2%. This meant for us at first sight that there was a reasonable agreement between all raters.

Afterwards intraclass correlation was calculated, which asses rating reliability by comparing the variability of each subject to all the variations of all subjects and ratings. Before calculating it was necessary to know if the data could be used in an ANOVA analysis, meaning to test for identically and independently distributed variables. We performed a simple non parametric bootstrap of 10000 repetitions for consensus and non consensus.

TABLE 2.1. Validation tests and their correspondent p-value.

Test	Score	Value	s.e.	p-value
Sum consensus	1631	1631	20.58	¡.001
Sum non consensus	568	568	20.52	¡.001
ANOVA (F-test)	0.649	6.54	3.92	¡.001
Fleiss Kappa	0.645	0.645	52.4	¡.001

Having the validated basis a fully-crossed, 2-way ANOVA and the Fleiss Kappa metric were obtained, shown in Table 2.1. The interpretation of the ANOVA test is a moderate agreement (McGraw & Wong, 1996) and the Fleiss Kappa gives a substantial strength of agreement (Landis & Koch, 1977).

2.3.3. Classifier

After building and validating the ground truth the next step would be to construct the classifier. For that matter we explain the feature selection, dimensionality reduction and class imbalance problems.

Feature selection. The dataset had mainly two groups of features, user based and content based. From the user we extracted number of followers and friends, which are directly usable as is. From the text we preformed text preprocessing, including tokenization and Spanish snowball stemming. From the corpora we used hashtags, words and user mentions, removing everything else. The number of resulting features were 4766 using

a tfidf vectorizer for the filtered content, considering a minimum frequency value of one word.

Additional considerations. We wanted also to test the models under different conditions. First we wanted to reduce the number of features, so a wider variety of models could be used due to memory restrictions. Latent dirichlet allocation (LDA) was chosen over latent semantic indexing (LSI) in this regard, because it can handle unseen documents giving a prediction when the words are not previously known (Bíró, Szabó, & Benczúr, 2008) and for LSI there is no natural way to do it (Wei & Croft, 2006). The gensim implementation was used to perform the LDA reduction.

An interesting topic was the noisy nature of Twitter, therefore we considered noise addition as a relevant issue. When additional not relevant messages were added to the original set a class imbalance problem appeared, that could affect the models performances (Wang & Yao, 2012). To address this issue we used the boundary SMOTE algorithm (Han, Wang, & Mao, 2005) to over sample the relevant messages. This was done before each round of training.

In order to add the required noise we gathered another set of tweets from Twitter streaming API, connecting a geographic localized query to the service for about 5 months, from 16/05/2014 to 27/10/2014. This query was a rectangle over Chile, so every tweet in this dataset was from or nearby this country. Afterwards tweets that were not recognized as Spanish by Twitter were removed. Additionally seismic activity related tweets were filtered, starting at magnitude 4. The messages that were 20 minutes before until 2 hours after an event were also removed*. Systematic sampling was used to extract the messages from this set in order to add them as noise to the ground truth before each training phase. The proportion of not relevant messages were added as a 20, 40, 60 and 80 percent of the ground truth length.

---

http://radimrehurek.com/gensim/

TABLE 2.2. The best scores for each classifier. (For every score the best is marked in bold)

Model	Precision	Recall	F1 score	Accuracy	AUC	Dimensions	Noise prop.
Baseline	0.625	0.545	0.53	0.5	0.568	-	0
Bernoulli NB	**0.831**	0.226	0.355	0.594	0.605	2000	0.0
Logistic Regression	0.827	0.641	0.722	0.756	0.834	1000	0.6
Linear SVM	0.687	**0.677**	0.682	0.687	0.719	1000	0.6
Random Forest	0.807	0.673	**0.734**	**0.758**	**0.844**	1000	0.8

In the training phase cross validation of 5 folds was used in each iteration of a grid search for every algorithm. We wanted to compare the performance under all the mentioned conditions. So we put logistic regression, random forest, support vector machines and naive Bayes variants under test and precision, recall, accuracy, AUC and f-score were measured to compare each one.

2.4. Results

We explored several algorithms and our results are shown in Table 2.2. The best exponent of each classifier was compared by accuracy. This method was used in every cross-validation and grid search.It is important to note that for comparison and sufficiency purposes we defined a simple criteria to set a baseline. The word earthquake (terremoto in Spanish) was used to classify each message as relevant or not relevant. Thus whenever this word appeared, the document was marked as relevant. This criteria performed better than random guessing.

In selecting the appropriate classifier it was important to remember that it would work at the heart of our citizen channel, to deliver relevant messages when an earthquake event occurred. In this sense it would work as a noise reduction tool, so the citizen looking for information would not need to read through the whole twitter messaging stream. Having this in mind it is easy to see that false positives were not really a big concern because we were not aiming at eliminating the noise completely. We really wanted however not to miss relevant messages (false negatives) and therefore recall was our main goal.

The results shown at Table 2.2 indicate that the best recall score was given to the linear SVM model. However random forest did better at the scores that evaluate precision and recall. We preferred this model considering that its recall was slightly below the best one, but outperformed it clearly in precision and as a consequence in other scores. Furthermore the random forest performed the best in the highest noise context, showing the best adaptation of the whole set of models.

2.4.1. Dimensionality reduction

While representing the content on a space with reduced dimensions some questions arose, such as how the models would be affected by them. The latent dimension reduction effect on the behaviour of each model is shown in Figure 2.3. This figure shows how the performances varied over noise proportion addition. Considering that with more noise, more oversampling was needed to address the class imbalance problem.

When the latent dimensions were few, the best algorithm was the logistic regression. However as the dimensions increase, random forest tended to perform better. Even with the highest considered number of latent dimension it performed above all the others. In general the scores got better with more dimensions, but worse with a very high amount.

The behaviour of the models on the recall scores variated more as shown in Figure 2.4. As the number of latent dimensions increased the algorithms performed around the baseline, however Bernoulli naive Bayes got often sufficient recall. Logistic regression got better as the dimensions grew, worsening when the number was very high. The best recall scores were located at the 1000 mark, all of the models tended to get better with those parameters.

2.4.2. Effects of class imbalance

Considering AUC when noise was added the stability of each model was clear. As noise grew, the scores got better or at least maintained its behaviour. Balancing the

classes proved to be helpful to random forest and logistic regression AUC, and neutral for Bernoulli naive Bayes.

The recall scores were very unstable when the dimensions were low. The noise addition helped consistently the logistic regression and random forest, but with the others that was not the case. Linear SVM was the most unstable overall. When we considered the baseline, the majority of the models did not surpass it in terms of recall. Again a local section between 40% and 80% with 1000 dimensions was the best score region for almost every model. Overall the trained algorithms took advantages from the noise addition and class imbalance measures.

Another important consideration had to do with computational performance. We did not want that the classifier became a bottleneck in the architecture of the whole system, affecting its effective throughput. The chosen algorithm according to the best performance as a classifier, corresponded to a random forest, which because it was relatively simple, had also very good performance from a computational point of view. This reduced the risk of the classifier becoming a bottleneck in situations of very high traffic of messages.

2.5. Conclusions

As the new generations take over and technology makes possible for anyone to own a sophisticated mobile device, Twitter and social networks will be used more and more to get fast information about special events. A natural disaster event is no exception and recent experience in Chile demonstrated the important role of social networks. Our goal was to leverage the main advantages of Twitter to produce a citizen to citizen information channel whose architecture we described before. A key component of this architecture was an automated classifier that can filter the huge and noisy flow of twitter messages, discarding all messages that were not related to the event. This channel was used to feed a mobile web application that the citizen could use at the time of the event. The building of the classifier involved many challenges including the definition of a reliable and validated ground truth and the selection of an appropriate algorithm in the context that the classifier

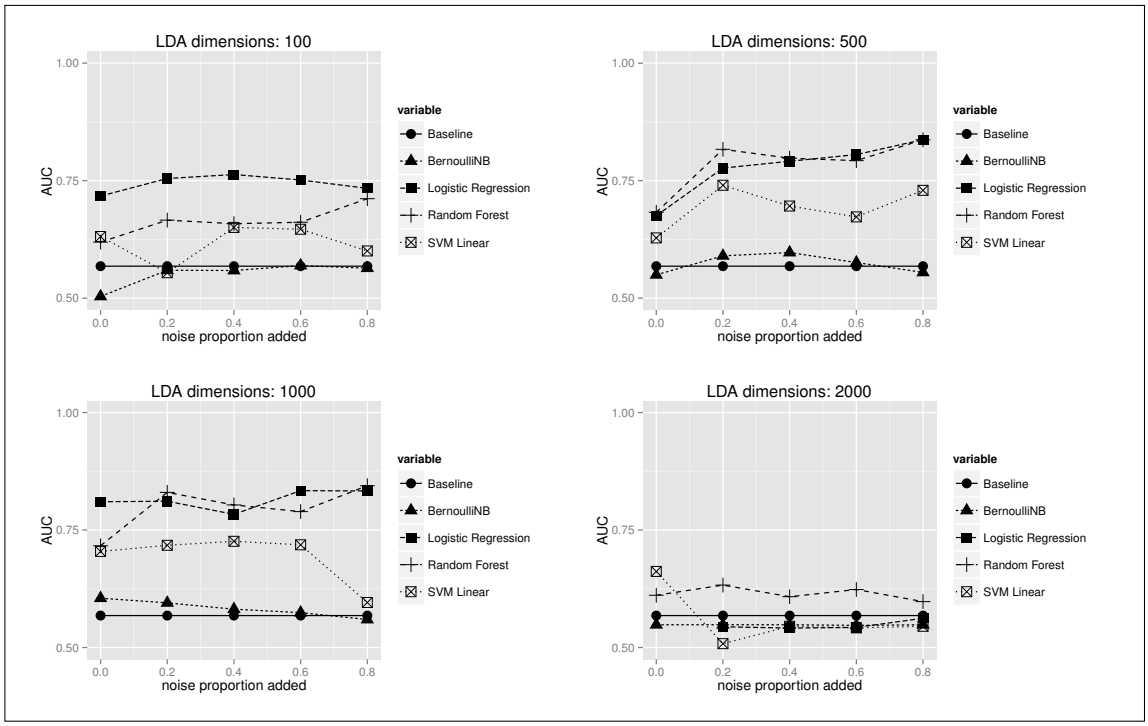

FIGURE 2.3. Variation of AUC scores when latent dimensions are set to different values.

was going to be used. After analyzing and comparing several classifiers we finally could get one that performed remarkably well for the purposes of our citizen channel. The selected model was a random forest that had 0.807 precision, 0.673 recall and 0.844 ROC AUC, outperforming our baseline and all other classifiers evaluated. Having this satisfying and concrete result allowed us to make a big step toward the implementation of our system. Final evaluation tests for this model would be when the architecture is in full operation and many citizens use the application during a forthcoming seismic event.

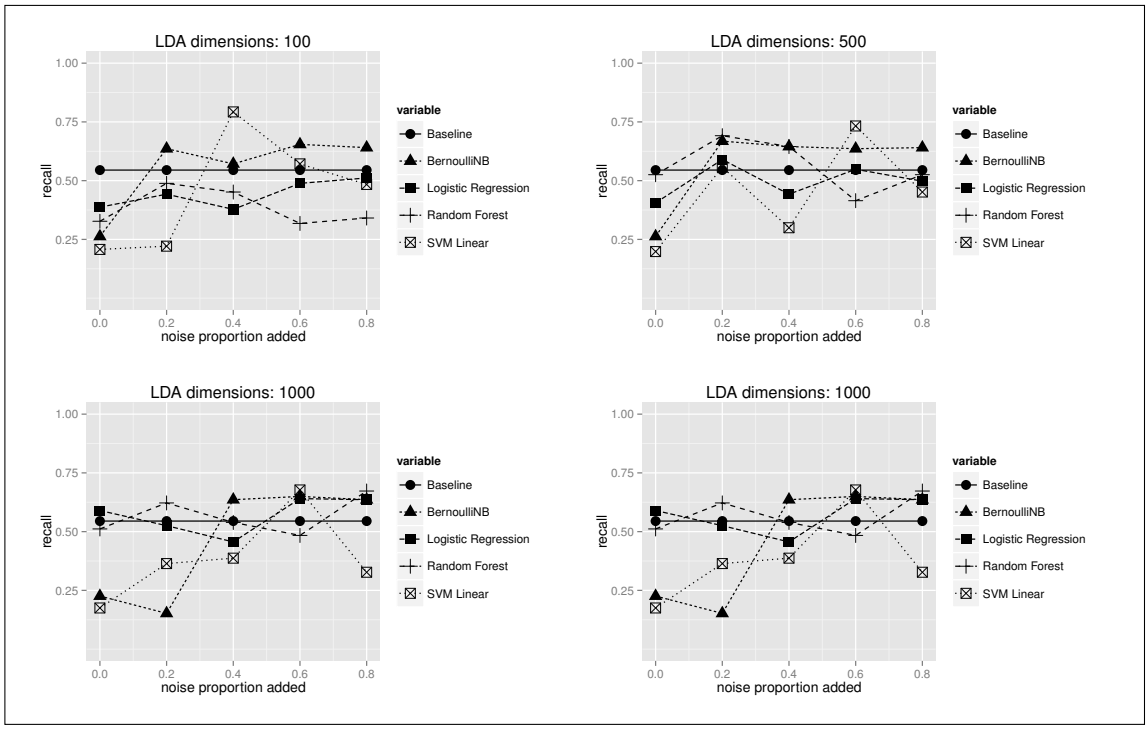

FIGURE 2.4. Variation of recall scores when latent dimensions are set to different values.

3. CONCLUSIONS

We have proposed a supervised trained model that can predict labels of twitter messages in an chilean earthquake context. Relevant messages can be filtered through it in order to construct a citizen channel. The criteria used to determine which model was good enough for this research was the one that had better scores overall and having special attention to the recall metric. In this way the channel could obtain a high quantity of relevant messages, considering that some noise was acceptable, since it was natural to the raw input stream.

The particularity of our approach is the context in which the model would be working. This is in an earthquake context, in a specific language and for chilean people. These held an important challenge for us, because most of the reviewed techniques were aimed at the english language and a general context. Furthermore, additional features which are generally considered we could not include.

Initially, when the dataset was being prepared we faced some problems. Because in our first approach we did not evaluated if the people did experienced the 2010 earthquake, we didn't know how much it would impact the base set. Not having defined this criteria led us to poor agreement scores, forcing us to do the manual classification again. The second time we fixed the requirements to chileans, so the context in which the messages were sent wouldn't be lost. The scores got much better and allowed us to have a good agreement between raters.

We learnt that It is really critical to know the context in which the people live, because it gives additional hidden information about the relevance of the messages. For this reason, in case a crowdsourced labeled dataset is used for this kind of application, the participants context should be a main concern for the researchers. We recommend a personal review of the participants, or alternatively having a massive classification with the grand majority of them being chilean and having experienced the natural disaster.

After the dataset was confirmed, the main problem we faced was the time it took for each model to run. If we considered a simple model we needed to wait at least a couple of days to see the results. So an important restriction was the limitations in the available hardware used to do the training and this, in fact, limited the number of algorithms we could test. Some algorithms complete only one run, others struggle to converge due to the high number of features in the dataset.

When we considered the performance of the algorithms in general, we could see that some were not fitted to work well in our domain since they achieved a low score, quite similar to random guessing. The fact that we picked a random forest does not mean that this would be the best one in other domains or even a similar domain but different situation or different context. In fact, a very careful analysis of all the scores was needed to get to the best model for us. Furthermore, this best model did not won by a wide margin and we had to favor the simplest model in case of similar scores.

This work made us realize how difficult is to filter through noisy channels. Obviously our intention was to obtain a good model in a very specific context, but we aimed to make it as general as we can. It is very likely that part of the noise gathered has some information that is relevant for other kind of events. Having this in mind it would be really interesting to explore more types of natural disasters, for example consider tsunamis, volcanic eruptions or wildfires as separate classifiers and join them in a single citizen channel.

Although we are quite satisfied with the results of the filtering model, it leaves a lot of room for improvement. One direction is to consider a much wider time range for the data. Another improvement would be to use a more recent event to train the classifier. We assumed that the habits of Twitter users would not change much as the years pass and this could not be so true. Using a more recent event would bring also better features to work with, considering that every day more people use geolocalized devices and that the Twitter API is evolving constantly.

Regarding the Twitter API we found that it was not difficult to access present data, because connecting a new stream is quite easy. The problems appeared when we needed data from a particular place and time. Searching streams back in time was simply impossible given the API restrictions, so an old proven dataset of an actual event had to be used. The only plausible way to gather enough information was to listen constantly to Twitter and hope that a disaster took place within the time of the research.

Another challenge was to save large amounts of messages that Twitter threw at us. For this we needed a simple database that could handle the format in which tweets came. We used a non SQL database to handle this, inserting the whole document at once. Even though the architecture could handle saving the entire stream, the space it took began to grow. In the course of one weeks a filtered stream can easily occupy more than 40 GB in a hard drive. So it was really important to have enough space in the servers.

In a wider context of the Cigiden subproject we believe that this is a first important step but a lot of work still needs to be done. If we assume that our classifier perform well in terms of separating the relevant messages from the noise, we still need to test the performance of the whole system which includes also many other pieces of software and yet even more important, the citizens themselves. Will be the classifier capable of filtering at the rate needed in a situation of a real earthquake?. How will all the rest of the technology will work in the context of partial or null connectivity? Are we going to be able to convince the citizens first to install the applications in their smartphones and then to use them during a real disaster ? Many of these issues are not directly related to this research but they need to be addressed if we want to provide real help.

Other architectures are being explored to add context awareness to any natural disaster event. One promising idea is to let actual people classify in real time the tweets. After that the framework learns automatically from the stream. A model that improves constantly could be a better approach to do the citizen channel. As a small step toward making the users to participate more directly in the classification could be a rating system, to get real time feedback of the performance of the working model and retrain it using those scores.

We have already incorporated in our prototype mobile application the functionality to classify and rate the messages.

References

Abel, F., Hauff, C., Houben, G.-J., Stronkman, R., & Tao, K. (2012). Semantics+ filtering+ search= twitcident. exploring information in social web streams. In *Proceedings of the 23rd acm conference on hypertext and social media* (pp. 285–294).

Bíró, I., Szabó, J., & Benczúr, A. A. (2008). Latent dirichlet allocation in web spam filtering. In *Proceedings of the 4th international workshop on adversarial information retrieval on the web* (pp. 29–32).

Caragea, C., McNeese, N., Jaiswal, A., Traylor, G., Kim, H.-W., Mitra, P., ... others (2011). Classifying text messages for the haiti earthquake. In *Proceedings of the 8th international conference on information systems for crisis response and management (iscram).*

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. In *Proceedings of the 20th international conference on world wide web* (pp. 675–684).

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap* (Vol. 57). CRC press.

Flach, P. (2012). *Machine learning: the art and science of algorithms that make sense of data*. Cambridge University Press.

Fritz, H., Petroff, C., Catalán, P., Cienfuegos, R., Winckler, P., Kalligeris, N., ... Synolakis, C. (2011). Field survey of the 27 february 2010 chile tsunami. *Pure and Applied Geophysics*, *168*(11), 1989-2010.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., ... Smith, N. A. (2011). Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies: short papers-volume 2* (pp. 42–47).

Han, H., Wang, W.-Y., & Mao, B.-H. (2005). Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *Advances in intelligent computing* (pp. 878–887). Springer.

Hiltz, S. R., & Plotnick, L. (2013). Dealing with information overload when using social media for emergency management: emerging solutions. In *Proceedings of the 10th international iscram conference* (pp. 823–827).

Hoffman, M., Bach, F. R., & Blei, D. M. (2010). Online learning for latent dirichlet allocation. In *advances in neural information processing systems* (pp. 856–864).

Imran, M., Castillo, C., Lucas, J., Meier, P., & Vieweg, S. (2014). Aidr: Artificial intelligence for disaster response. In *Proceedings of the companion publication of the 23rd international conference on world wide web companion* (pp. 159–162).

Imran, M., Elbassuoni, S. M., Castillo, C., Diaz, F., & Meier, P. (2013). Extracting information nuggets from disaster-related messages in social media. *ISCRAM, Baden-Baden, Germany*, *11*.

Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter sentiment analysis: The good the bad and the omg! *Proceedings of the fifth annual conference on weblogs and social media ICWSM*, *11*, 538–541.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on world wide web* (pp. 591–600).

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, *25*(2-3), 259–284.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159–174.

Lee, K., Mahmud, J., Chen, J., Zhou, M., & Nichols, J. (2014). Who will retweet this?: Automatically identifying and engaging strangers on twitter to spread information. In *Proceedings of the 19th international conference on intelligent user interfaces* (pp. 247–256). New York, NY, USA: ACM.

Liu, K.-L., Li, W.-J., & Guo, M. (2012). Emoticon smoothed language models for twitter sentiment analysis. In *Proceedings of aaai.*

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological methods*, *1*, 30.

Mendoza, M., Poblete, B., & Castillo, C. (2010). Twitter under crisis: Can we trust what we rt? In *Proceedings of the first workshop on social media analytics* (pp. 71–79).

Middleton, S. E., Zielinski, A., Necmioğlu, Ö., & Hammitzsch, M. (2014). Spatio-temporal decision support system for natural crisis management with tweetcomp1. In *Decision support systems iii-impact of decision support systems for global environments* (pp. 11–21). Springer.

Molina, C. (2015). *Diseño e implementación de aplicación movil informativa para desastres naturales* (Unpublished master's thesis). Pontificia Universidad Católica de Chile, Santiago, Chile.

Morales, A., Borondo, J., Losada, J., & Benito, R. (2014). Efficiency of human activity on information spreading on twitter. *Social Networks*, *39*, 1–11.

Morstatter, F., Kumar, S., Liu, H., & Maciejewski, R. (2013). Understanding twitter data with tweetxplorer. In *Proceedings of the 19th acm sigkdd international conference on knowledge discovery and data mining* (pp. 1482–1485).

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In *Proceedings the seventh international aaai conference on weblogs and social media.*

Power, R., Robinson, B., & Wise, C. (2013). Comparing web feeds and tweets for emergency management. In *Proceedings of the 22nd international conference on world wide web companion* (pp. 1007–1010). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee.

Puente, S., Pellegrini, S., & Grassau, D. (2013). How to measure professional journalistic standards in television news coverage of disasters? 27-f earthquake in chile. *INTERNATIONAL JOURNAL OF COMMUNICATION*, *7*, 1896–1911.

Robinson, B., Power, R., & Cameron, M. (2013). A sensitive twitter earthquake detector. In *Proceedings of the 22nd international conference on world wide web companion* (pp. 999–1002).

Saad, F. (2014). Baseline evaluation: an empirical study of the performance of machine learning algorithms in short snippet sentiment analysis. In *Proceedings of the 14th international conference on knowledge technologies and data-driven business* (p. 6).

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

Taxidou, I., & Fischer, P. M. (2014). Online analysis of information diffusion in twitter. In *Proceedings of the companion publication of the 23rd international conference on world wide web companion* (pp. 1313–1318). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee.

Ureta, S. (2008). Mobilising poverty?: Mobile phone use and everyday spatial mobility among low-income families in santiago, chile. *The Information Society*, *24*(2), 83–92.

Valenzuela, S., Arriagada, A., & Scherman, A. (2012). The social media basis of youth protest behavior: The case of chile. *Journal of Communication*, *62*(2), 299–314.

Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010). Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1079–1088).

Walther, M., & Kaisser, M. (2013). Geo-spatial event detection in the twitter stream. In *Advances in information retrieval* (pp. 356–367). Springer.

Wang, S., & Yao, X. (2012). Multiclass imbalance problems: Analysis and potential solutions. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, *42*(4), 1119–1130.

Wei, X., & Croft, W. B. (2006). Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international acm sigir conference on research and development in information retrieval* (pp. 178–185).

Wu, S., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Who says what to whom on twitter. In *Proceedings of the 20th international conference on world wide web* (pp. 705–714).