Spatio and Temporal Characterization of Chilean News Events in Social Media

Jazmine Maldonado
Department of Computer Science
University of Chile
Santiago, Chile
jamaldonadof@ug.uchile.cl

Vanessa Peña-Araya
Department of Computer Science
University of Chile
Santiago, Chile
vpena@dcc.uchile.cl

Barbara Poblete
Department of Computer Science
University of Chile
Santiago, Chile
bpoblete@dcc.uchile.cl

ABSTRACT

Online Social Networks play a leading role in news consumption. As a consequence, most newspapers and other media use these platforms to promote their content. However, the geographic bias in the media, in addition to the demographic bias in Online Social Networks can lead to inaccurate and incomplete view of the news in a country. Being aware of these two kinds of bias in news published in Online Social Networks is useful to understand the context in which events develop. We selected Chile as a case study to observe these problems. Chile is a country with a high degree of participation in Online Social Networks and suffers from both issues: media covers mostly news from Santiago, its capital, and most of Online Social Networks users are located in this city. We built a dataset of Chilean news headlines extracted from Twitter.

We conducted a characterization of news and messages which comment them. We focus on the geographical and temporal features of news. In this paper we present the results of this analysis in addition to the description of the dataset. Our findings show that as expected, news and Twitter users are mainly concentrated in Chile's capital. In addition, users in Chile focus on local news paying little attention to international events. We observed that a considerable number of users discussing Chilean news are located outside of Chile. We conclude that users in Chile are subject to bias in news media coverage of information, which privileges news from the largest cities.

Keywords

Geo-temporal analysis, event coverage in Social Media, case study

1. INTRODUCTION

Online social networks have become an important source of information. Their growth has allowed for regular citizens and not only traditional news media to inform when an important event happens. Indeed, it is not uncommon that the first information about breaking news is published by a regular user instead of a journalist. However, even if the involvement of users in Online Social Networks has helped decrease the bias in information coverage introduced by news media, by no means are these platforms geographically representative source of information. We assume the task of characterizing news extracted from Online Social Network by the geographical distribution of the location they mention, in addition to that of the Online Social Network users that share them. We believe that this analysis will help understand the context in which news events develop, in addition to unveiling biases in information coverage.

Chile is one of the top ten countries using Twitter [1]. As Chile is a very centralized country, the geographical distribution of Twitter users and news visibility is not homogeneous and mainly focused in Santiago, its capital. To observe the characteristics of the geographical behaviour of Twitter users and news over time, we built a dataset by gathering news headlines from Chilean newspapers' Twitter accounts. For each news topic extracted from the headlines we retrieve its keywords and the tweets [2] that comment about it. In this work we present a case study of this dataset.

2. RELATED WORK

This case study is based on an extension of the work of Vanessa Peña-Araya, Mauricio Quezada and Barbara Poblete [2] which presents spatio-temporal event models.

The work of Graells-Garrido and Lalmas [1] covers how the physical centralization of Chilean population affects the participation of Twitter users. Although we also cover centralization, our work is broader as it covers other issues and considers behaviour of the Chilean media.

3. EVENT EXTRACTION AND DATASET CONSTRUCTION

To build the Chilean news dataset, we conduct a 2 steps process. We first collect news events from Chilean Twitter accounts of media entities. The second step is to geolocate where events occur and also the users who comment them.

[1] http://www.forbes.com/sites/victorlipman/2014/05/24/top-twitter-trends-what-countries-are-most-active-whos-most-popular/

[2] tweets: messages posted in Twitter composed of maximum 140 characters

TAIA'15, August 13, 2015, Santiago, Chile.

Event detection starts by periodically collecting tweets from a set of selected twitter news accounts from Chile. The set of accounts was manually curated and complete including accounts from online newspapers in regions of Chile. It also includes radio accounts, TV news accounts, institutions and some journalist accounts. Headlines are processed to retrieve a representative pair of keywords of each one, and then, they are clustered into events. For each event we retrieve the tweets in Spanish. This search aims to obtain messages commenting about the event after and before the event was published by the media.

In the second step of the process users and events are geotagged. To geolocate users, the system takes the user location field in their Twitter profile and resolves using CLAVIN [3]. On the other hand to geolocate an event we extract the locations that are mentioned in an event's headlines and comments. To obtain locations with higher-level of administrative divisions than country or region, we use two sources of geographical locations. First of all, the system searches for text matches in a dedicated list of Chilean cities or country names. If no location is matched or the number of mentions of those that were indeed found are too small, the system uses CLAVIN. Even when both methods are used, some events are not possible to geotag. This happens because there were not locations mentioned in tweets related to those events or because the places mentioned are cities which cannot be identified using CLAVIN.

In the set of events that can not geotagged there are events that do not mention names of cities but mention other information like the names of the local soccer teams, names of some monument known by the population, names of people, etc. Some of these cases were identified and they are considered in the process of geotag an event, but using our approach they must be identified manually so they are exceptions.

As a result of using this process over a period of six months, we constructed a dataset of Chilean interest events which we plan to share with the community for research purposes. The dataset is composed of events, and each one contains: (i) the date when it was detected; (ii) its most representative keywords; (iii) one or more locations related to it (where it happened, countries involved, etc); and (iv) a set of tweets IDs commenting about it. In addition, each tweet ID is geolocated, when possible.

4. CHILEAN NEWS EVENTS

In this section we analyze the Chilean news dataset focused on its geographical and temporal characteristics. We first give an overview of the dataset (subsection 4.1). In subsection 4.2 we analyze the geographical distribution of Chilean news and Chilean Twitter users. We finally inspect the interest that Chilean media and Twitter users have in international countries in subsection 4.3.

4.1 Dataset Overview

The dataset is composed of events collected over a period of six months, starting from November 1, 2014 to April 30, 2015. Table 1 provides an overview of dataset in terms of the number of events and Twitter users gathered. It also gives the number and percentages of events and users that were geotagged (one or more locations were assigned). From the total of 6507 geotagged events, 4740 of them (72.8%) are news concerning Chile, and 1767 (27.2%). are news about other countries. For the purpose of this paper, we call the first kind of news *national events*. On the other hand, news events concerning countries outside Chile are referred to as *international events*.

	Users	Events	Tweets
geotagged	2,179,255(44%)	6,507(60%)	14,121,553(51%)
non-geotagged	2,796,443(56%)	4,415(40%)	13,618,515(49%)
Total	4,975,698	10,922	27,740,068

Table 1: General information and totals of events and users of the dataset

4.2 Geographical distribution of Chilean news and Twitter users

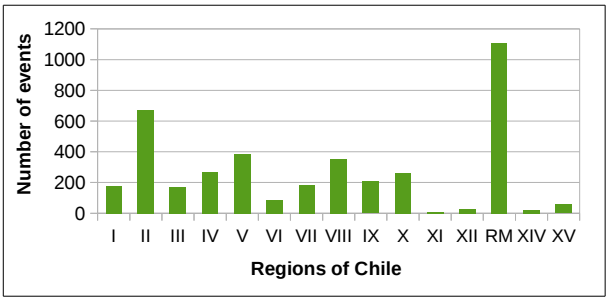

Figure 1: Number of events detected for each region of Chile

To analyze the geographical distribution of Chilean events it is important to know that Chile is divided in to 15 *regions*. Regions are the first-level administrative division of Chile and are named by a Roman numeral. We divide events in two categories: (i) events regarding specific Chilean locations i.e. into the set of tweets associated with the event was possible to find names of cities or regions of Chile a considerable number of times, and (ii) events concerning the whole nation i.e. only was possible to detect the name of Chile in majority but not names of cities or regions. For the first category we found events like natural disasters, regional initiatives, accidents, etc. For events in which the whole country is involved, we found news such as political topics, a new law approvals, TV show topics, national soccer matches, among others. Events in category (ii) occur when at the moment of geotagging the events, they are tagged as Chilean events but without information about region or city. From the total 4740 Chilean events in the dataset, 3450 of them are events associated with specific locations in Chile and 1290 are general Chilean events.

Figure 1 displays the distribution of news events related with each region of Chile. The figure only represents events concerning a particular Chilean region, category (i), and it does not include events belonging to (ii) event category. Nevertheless, although we consider headlines for all regions of Chile, most events refer to the capital of the country Santiago which is located in the Metropolitan Region (RM). Region II also has many events, mainly because on March 25 a large temporary affected the area leaving a large number of

dead and missing persons. Furthermore, the regions XI and XII of Chile, have very few events that mention them.

From this information, one can ask whether the concentration of news events in the capital is because no newsworthy events occurs in other regions of the country, or if instead it is the product of bias in information coverage by the media. From the data we observe that the national media tends to inform only about big disasters in smaller localities, but little about everyday problems of people there. However, even low impact news in the capital, such as malfunction of Santiago's subway trains, are covered by national TV news.

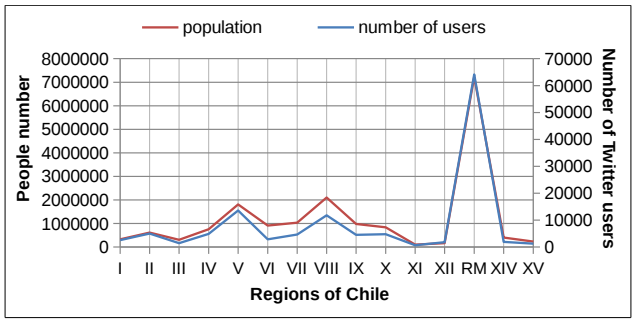

Figure 2: Number of Twitter users and population of each region of Chile

Regarding geolocated Twitter users, from the total of users who comment on news published in Chile (national and international), 8,4% of them are in Chile. The rest consists of users from other countries, among them, there are some Spanish-speaking countries with a considerable amount of users as Spain, Mexico, Argentina, Venezuela and Colombia, which together represent 62% of users in the dataset. The remaining 30% is made up of users located in 245 other countries. The user geotagging process obtains the information from Twitter user location field, so is unknown if the geolocated Twitter users effectively live in the country that they say they live, or if they are Chileans that live in those countries. As for the geographical distribution of Chilean users, 27.8% of them only put "Chile" in their location field, and the remaining 72.2% have "Chile" in addition to a region name or a city name.

The Figure 2 shows the distribution of Chilean users that specified in which part of Chile they are from and the population of each region of Chile. In the same way of events distribution, the concentration of users is mainly in the capital, Santiago, and the difference between users distribution is even more drastic than event distribution. But we also can observe the distribution of Twitter users is very similar to population in each region. The *centralization* is a known problem of Chile in which most of the population and resources are located in Santiago and somehow this condition is reproduced in the number of Twitter users in the dataset we build. About the relation between the number of tweets and users per region, there are in average 18 tweets per user in each region with an standard deviation of 1,8.

4.3 Geopolitical interest of Chilean media and Twitter users

We present the geopolitical interest in international countries by Chilean media and Chilean Twitter users. As we previously mentioned, 27.2% of news events are international events, involving other countries than Chile. This means that, even if the seeds of events are Chilean media entities, a considerable part of headlines are international news, showing the Chilean media interest in external countries.

Of all collected data, the events with more tweets are international events or related with other countries that are not Chile, for example, the Oscar Awards, the Nepal earthquake, Charlie Hebdo Shooting, etc. As expected, people that comment about them are from different Spanish-speaking countries and the most are not from Chile. The international events with more interest of Chilean people are related with nearby countries, have less than 6,000 Chilean tweets and are not in the top 30 of most tweeted by Chileans.

Although Chilean media and users have interest in international events, Chilean users are by far more interested in Chilean news. Table 2 shows the top five events most tweeted by Chilean people.

Description	Location	Chilean Tweets
Villarrica Volcano Eruption	IX - La Araucanía	42,483
Forest Fire in Valparaíso	V - Valparaíso	29,253
Elqui River Overflow	Chile	25,684
Chile's Teleton	Chile	25,050
Resignation of the Health Minister Helia Molina	Metropolitan Region of Santiago	22,723

Table 2: Most commented events by users in Chile

Locations of the events were originally in Spanish, and they were translated to English for the purpose of this paper. Even though the dataset includes several international events tweeted by Chilean users, the top five events most twitted by Chileans occur in Chile. About Chilean engagement, in average, we observe that users in Chile tweet approx. 2 tweets per event, in those with the most activity.

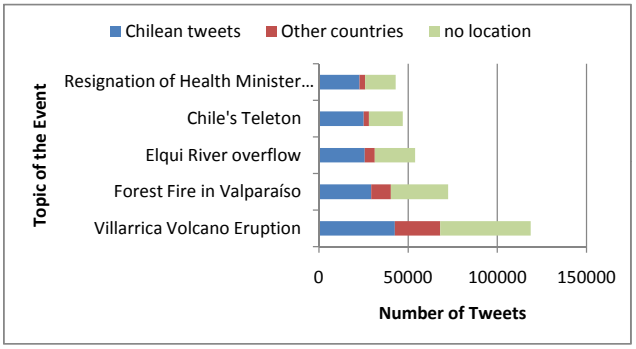

Figure 3: User participation in the events with more tweets by users in Chile

Figure 3 shows the number of tweets of users of other countries in events with highest participation of Chilean people. Although the Chilean participation is higher than in other events, users in other countries also have an important presence in these national events.

To inspect which countries are relevant for Chilean media and Twitter users, Figure 4 shows the top ten coun-

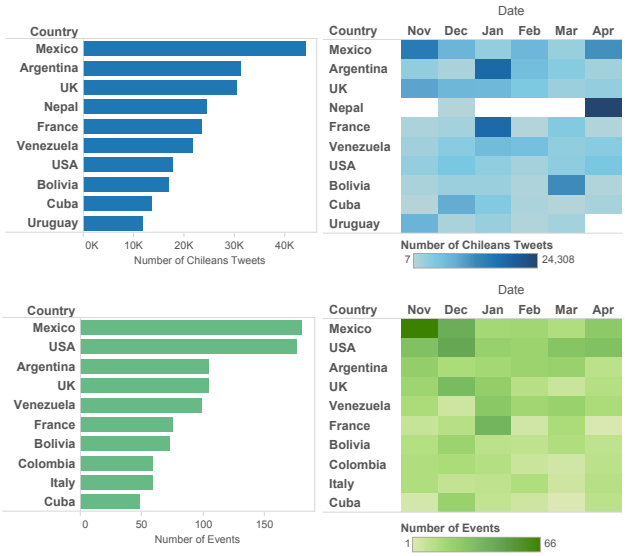

Figure 4: Top Ten countries that concentrated the most interest for Chilean users (in Blue) and for Chilean Media (in Green) during the period Nov/2014 to Apr/2015.

tries found in events. On the left side it shows the amount of events concerning each country, and on the right side it shows the distribution of events per month. As can be observed, the interest of media is similar to the interest of Chilean people. The only exception are Nepal and Uruguay, which appear to be more interesting for Chilean people but were not covered much by media. In the upper-right side of the visualization, we can see that the most tweeted events in Nepal are mainly in April, the month when the Nepal Earthquake happened and after the Everest avalanche. In the case of Uruguay, we can see that events are more evenly distributed over time, so they are probably not the same event but several different events involving the country. A look at event's keywords, showed us that the most tweeted Uruguayan events are related with soccer matches between Chile and Uruguay in November.

Regarding time, we can observe that Chilean media has a more even coverage of countries over time, than the interest of Chilean users. Indeed, there are periods of time when Chileans did not tweet about certain countries, like Nepal, to only pay attention when important news happens. Mexico and USA appear to be covered by Media and Twitters users in a similar way. Analysing the keywords of Mexican events, we observe that the news of the Mexican's students disappearance was extensively covered by Chilean media and also very commented by Chileans users. It is important to notice that, even if the event happened before we started our collection, people and media still continued to talk about it, being as it was a very persistent topic over time. Different is the case of nations like United Kingdom, France and Italy. By inspecting events keywords involving these countries, we observe that the most frequents events are related with soccer. This could be due to the fact that several Chilean soccer players are members of important European teams. Only France has an important spike in January, the month when a series of terrorist attacks occurred.

5. CONCLUSIONS

We have presented our preliminary findings of Chilean news on Twitter. For this we have built a dataset of events from Chilean news media on Twitter and gathered the tweets that discuss them. By analyzing events by their geographical and temporal characteristics, we have three main conclusions. The first one is that even if Chilean media has interest in international events, Chilean Twitter users are more interested in local news. In fact, Chilean users only focus on international news for extremely high-impact events (e.g. terrorist attacks and natural disasters) or for soccer matches. Our second finding is that Twitter users from countries outside Chile are interested in Chilean news: a considerable number of them commented about Chilean events, particularly for natural disasters. As we mention before, a large part of users geotagged with countries outside of Chile and tweet in Chilean events could very well be Chileans living abroad. Finally, we were able to confirm that Chilean media displays from geographical bias in news coverage as most of news were from Santiago, Chile's capital. This is most likely driven by commercial interests moved by audience distribution in the country.

6. ACKNOWLEDGMENTS

We acknowledge Janani Kalyanam from UCSD and Mauricio Quezada for designing and allowing us to use their news event collection procedure.

The authors were partially supported by FONDECYT Grant 11121511 and the Millennium Nucleus Center for Semantic Web Research under Grant NC120004.

References

[1] E. Graells-Garrido and M. Lalmas. Balancing diversity to counter-measure geographical centralization in microblogging platforms. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, HT '14, pages 231–236, New York, NY, USA, 2014. ACM.

[2] M. Quezada, V. Peña Araya, and B. Poblete. Location-aware model for news events in social media. In *38th Annual ACM SIGIR Conference*, Santiago, Chile, August 2015.

[3] B. Technologies. CLAVIN: Cartographic Location And Vicinity INdexer. `http://clavin.bericotechnologies.com/`, 2012–2013.