Mining Social Networks to Learn about Rumors, Hate Speech, Bias and Polarization

Barbara Poblete

@bpoblete
www.barbara.cl

Department of CS, University of Chile
Millennium Institute for Fundamental Research on Data

About Me

I'm a **Data Mining** person and I've also worked in **Information Retrieval** (search engines) almost two decades

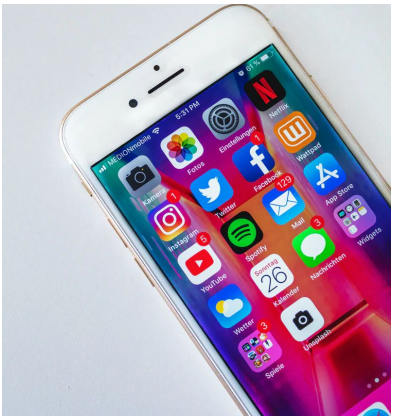
I specialize in **Web data mining** and **online social media** analysis

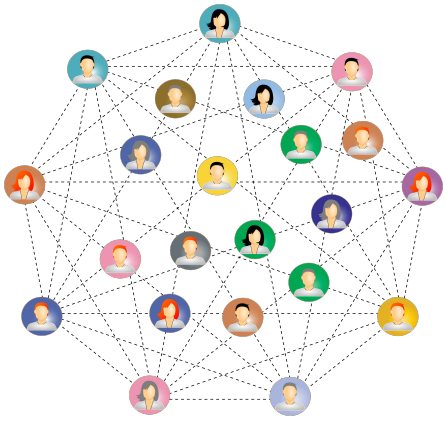
I use ML as a tool, my focus is on **evaluation** and using available tools in the best possible way to **discover knowledge**.

I value **gaining understanding from data,** being able to **reproduce,** and to **generalize** results more than gaining marginal improvements

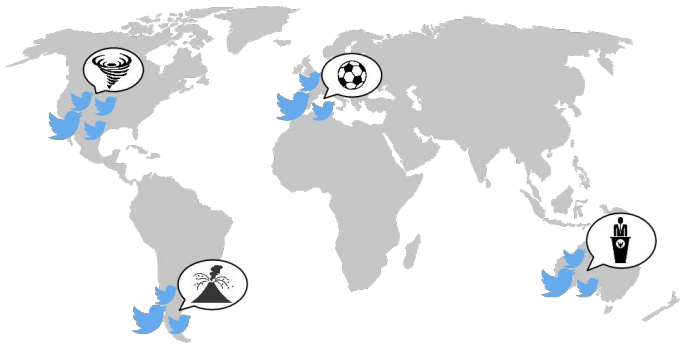
The Social Web

shift from passive browsing of content to massive producers





New information media

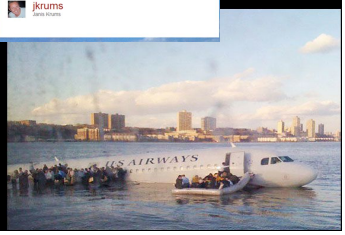


Real-time information from the place of events

Image: courtesy of Jazmine Maldonado

Users become creators and publishers of their own content

Short messages



Social media has many advantages but it has also made us **vulnerable** to how online communication is shaped

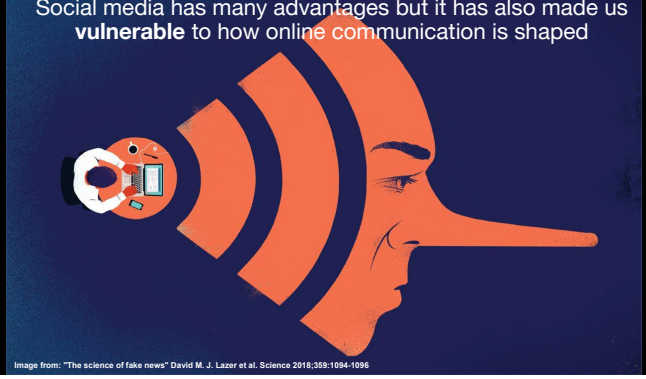


Image from: "The science of fake news" David M. J. Lazer et al. Science 2018;359:1094-1096

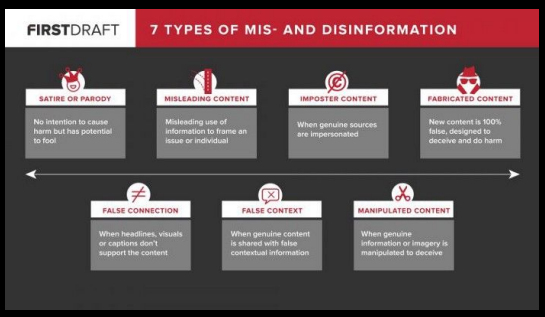
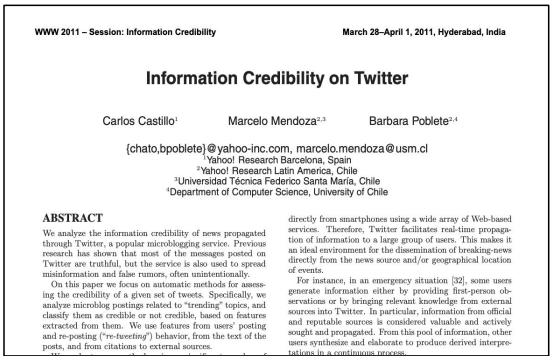


Image from Claire Wardle: "Fake news. It's complicated". (Thanks to R. Baeza-Yates for sending me this)

10 years ago...



WWW 2011 – Session: Information Credibility March 28–April 1, 2011, Hyderabad, India

Information Credibility on Twitter

Carlos Castillo[1] Marcelo Mendoza[2,3] Barbara Poblete[2,4]

{chato,bpoblete}@yahoo-inc.com, marcelo.mendoza@usm.cl
[1]Yahoo! Research Barcelona, Spain
[2]Yahoo! Research Latin America, Chile
[3]Universidad Técnica Federico Santa María, Chile
[4]Department of Computer Science, University of Chile

ABSTRACT

We analyze the information credibility of news propagated through Twitter, a popular microblogging service. Previous research has shown that most of the messages posted on Twitter are truthful, but the service is also used to spread misinformation and false rumors, often unintentionally.

On this paper we focus on automatic methods for assessing the credibility of a given set of tweets. Specifically, we analyze microblog postings related to "trending" topics, and classify them as credible or not credible, based on features extracted from them. We use features from users' posting and re-posting ("re-tweeting") behavior, from the text of the posts, and from citations to external sources.

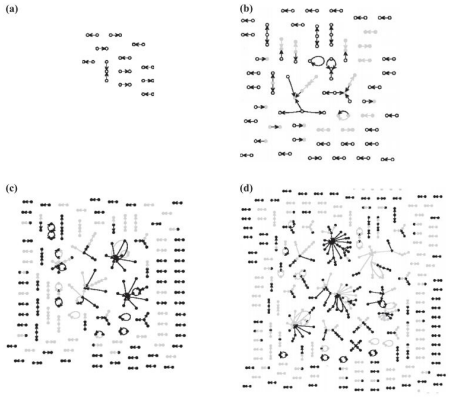
directly from smartphones using a wide array of Web-based services. Therefore, Twitter facilitates real-time propagation of information to a large group of users. This makes it an ideal environment for the dissemination of breaking-news directly from the news source and/or geographical location of events.

For instance, in an emergency situation [32], some users generate information either by providing first-person observations or by bringing relevant knowledge from external sources into Twitter. In particular, information from official and reputable sources is considered valuable and actively sought and propagated. From this pool of information, other users synthesize and elaborate to produce derived interpretations in a continuous process.

Real-time misinformation

- Chilean Earthquake
  - Sat Feb 27, 2010. 03:34 local time
  - 8th largest recorded in history
  - 8.8 Mw, 90 s, 500-600 casualties
- Communications
  - Almost impossible for 2-3 hours
  - First video images 6-7 hours after quake

#terremotochile



(a) 27 Feb (b) 28 Feb

(c) 01 Mar (d) 02 Mar

- **Large majority of tweets were helpful**
- Some tweets were not
  - Celebrities and public figures "killed"
  - False tsunami warnings
  - False reports of looting
  - ...

Manual classification results

Case	# of unique tweets	% of re-tweets	# of unique "affirms"	# of unique "denies"	# of unique "questions"
Confirmed truths					
The international airport of Santiago is closed	301	81	291	0	7
The *Viña del Mar International Song Festival* is canceled	261	57	256	0	3
Fire in the Chemistry Faculty at the University of Concepción	42	49	38	0	4
Navy acknowledges mistake informing about tsunami warning	135	30	124	4	6
Small aircraft with six people crashes near Concepción	129	82	125	0	4
Looting of supermarket in Concepción	160	44	149	0	2
Tsunami in Iloca and Duao towns	153	32	140	0	4
TOTAL	1181		1123	4	30
AVERAGE	168,71		160,43	0,57	4,29
False rumors					
Death of artist Ricardo Arjona	50	37	24	12	8
Tsunami warning in Valparaiso	700	4	45	605	27
Large water tower broken in Rancagua	126	43	62	38	20
Cousin of football player Gary Medel is a victim	94	4	44	34	2
Looting in some districts in Santiago	250	37	218	2	20
"Huascar" vessel missing in Talcahuano	234	36	54	66	63
Villarrica volcano has become active	228	21	55	79	76
TOTAL	1682		502	836	216
AVERAGE	240,29		71,71	119,43	30,86

Can we automatically detect false tweets as events are unfolding?

Not really
and not without external sources, but...

Information credibility

- Perceived quality
- Made of multiple dimensions
- Examples:
  - Same article written by male/female author
  - Same tweet by male/female author
  - Same headline in news/twitter
  - ...

Automatic credibility classification

- **Newsworthiness** **92% precision at 92% recall**
- **Credibility** **87% precision at 83% recall**
- We used *text*, *network*, *propagation* and *top-tweet element* features
- We even showed that our model could be used to classify events **online** and in **other languages** (mostly lang agnostic)

Some findings

- We found that the **less you rely on external sources** the more likely you are to **believe false information**
- Credible information was propagated by prolific and well connected users
- Tweets that had URLs were found to be more credible (specially if it was from a popular domain)
- Information looked more credible if it didn't have exclamation marks or if it expressed a negative sentiment

Now back to 2021...

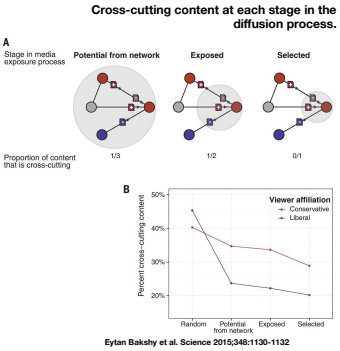
Since then, there have been thousands of academic works about social media credibility

How can we be worse than ever?

Filter Bubbles

Exposure to politically opposed content is greatly reduced by:

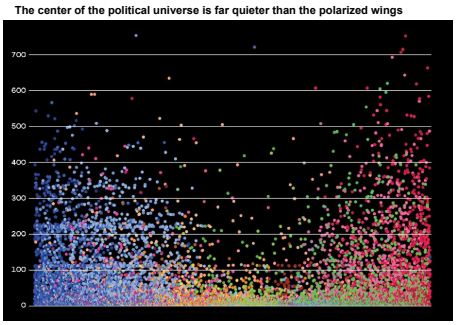
- Our selection of **"friends"** and
- The **recommendation algorithm**
- Social networks show us a homogeneous view that **reduces tolerance to alternative views and amplifies polarization**.
- This increases the probability of accepting news compatible with our ideology



Eytan Bakshy et al. Science 2015;348:1130-1132

Echo chambers

- Use of **bots** (automated or paid accounts) in polarized communities
- **Amplify the effect** of false news
- 15% of accounts on Twitter
- 60M accounts in FB



J. Kelly and C. François. This is what filter bubbles actually look like. MIT Technology Review. 2018.

Facebook leading misinformation campaigns

70 countries where **disinformation campaigns** have been detected

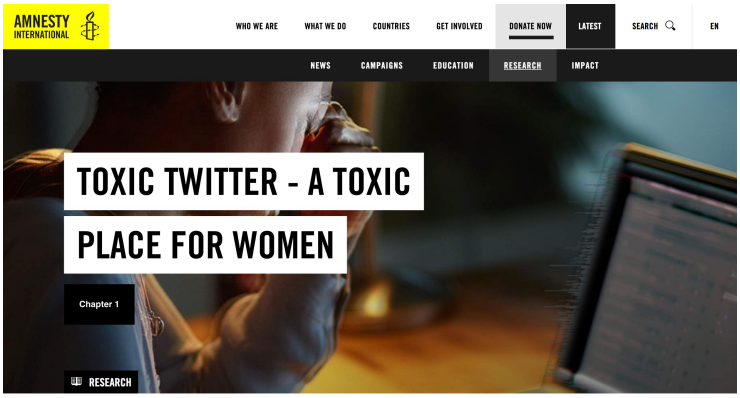
Cambridge Analytica collected private information from **87 million** Facebook users

This allowed them to create **highly detailed profiles of people's preferences** and design campaigns (intervention in US and Brexit elections)

We have **stopped trusting traditional news media** and now, more than ever, we rely **only on social media**

Algorithms that optimize user engagement are **biased towards fake information**

There are NO incentives for platforms to change this...



Automatic hate speech detection

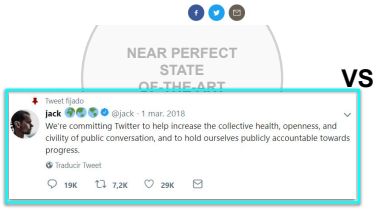
NEAR PERFECT STATE OF-THE-ART RESULTS

VS

THE REALITY OF HATE SPEECH IN SOCIAL MEDIA

U.S. • TWITTER

Twitter Apologizes for Mishandling Reported Threat From Mail-Bomb Suspect



NEAR PERFECT STATE OF-THE-ART

VS

THE REALITY OF HATE SPEECH IN SOCIAL MEDIA

We show that hate speech detection is a multidimensional problem and that existing approaches **fail due to faulty implementation** of machine learning and **bias in the data** (performance drop from F1 ~93% to 51% in SOTA)

Hate Speech Detection is Not as Easy as You May Think: A Closer Look at Model Validation

Aymé Arango
aarango@dcc.uchile.cl
Department of Computer Science
University of Chile
IMFD, Chile

Jorge Pérez
jperez@dcc.uchile.cl
Department of Computer Science
University of Chile
IMFD, Chile

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

ABSTRACT

Hate speech is an important problem that is seriously affecting the dynamics and usefulness of online social communities. Large scale social platforms are currently investing important resources into automatically detecting and classifying hateful content, without much success. On the other hand, the results reported by state-of-the-art systems indicate that supervised approaches achieve almost perfect

1 INTRODUCTION

Automatic detection of hate speech has become an increasingly relevant research topic in the past few years [11, 26, 27]. The worldwide adoption of online social networks has created an explosion in the volume of text-based social exchanges. Social media communications can strongly influence public opinion and some social platforms are said to have enough social capital to influence the

Some of the main problems we found:

- Use of the **complete dataset** in the **training phase** (without separation of the test set), e.g., to generate word embeddings before training
- **Oversampling** hate speech class **before train/test split**
- **Data bias**:
  - One of the most popular datasets only had tweets without user information. We mapped tweets to users to find that 90% of HS messages were generated by 3 users! (the classifier was learning to identify users)

How did we find these problems?

- By having **multilingual needs**!
- We attempted to transfer results of HS detection models from English to Spanish, but were not able to achieve comparable performance
- We tried to reproduce results in English and found multiple errors and problems in the data
- We ended up publishing a paper at SIGIR 2019 about why results were not reproducible

Overestimation of the state of the art
is more common than we think,
specially since classification models have
become more **obscure** to the user and
if no attention is paid to the **data**

We always have to look at the data!

You **can't really say that you've solved a problem**,
until you successfully apply your solution/method
in a **completely different scenario**

What are we (currently) doing?

- We are working on **hate-specific multilingual word embeddings**
- We have shown that **we need much less data** to **outperform sophisticated methods** trained on massive amounts of data
- These findings indicate us that there are **cross-cutting patterns of hate speech** that are **language independent**

A **multidisciplinary** approach is **needed**
to target fake-news and hate speech

This is **not a problem that technologists can solve**
and we **should not frame** it in that way

Although Spanish is one of the top-5 languages it is extremely difficult to find open NLP resources (models, train data, benchmarks)


COMMUNICATIONS of the ACM

HOME | CURRENT ISSUE | NEWS | BLOGS | OPINION | RESEARCH | PRACTICE

Home / Magazine Archive / November 2020 (Vol. 63, No. 11) / Minding the AI Gap in LATAM / Full Text

LATIN AMERICA REGIONAL SPECIAL SECTION: HOT TOPICS

Minding the AI Gap in LATAM

By Barbara Poblete, Jorge Pérez

Communications of the ACM, November 2020, Vol. 63 No. 11, Pages 61-63

10.1145/3416969

Comments

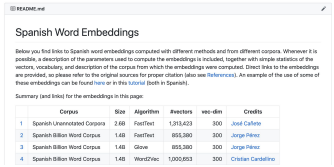
VIEW AS: SHARE:

Credit: WordClouds.com

Societies and industries are rapidly changing due to the adoption of artificial intelligence (AI) and will face deep transformations in upcoming years. In this scenario, it becomes critical for under-represented communities in technology, in particular developing countries like Latin America, to foster initiatives that are committed to developing tools for the local adoption of AI. Latin America, as well as many non-English speaking regions, face several problems for the adoption of AI technology, including the lack of diverse and representative resources for automated learning tasks. A highly problematic area in this regard is natural language processing (NLP), which is strongly dependent on labeled datasets for learning. However, most state-of-the-art NLP resources are allocated to English. Therefore, creating efficient NLP tools for diverse languages requires an important investment of time and financial resources. To

B. Poblete, J. Pérez. Communications of the ACM, November 2020, Vol. 63 No. 11, Pages 61-63 10.1145/3416969

Pre-trained neural word representations

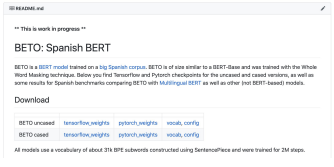

Spanish Word Embeddings

Below you find links to Spanish word embeddings computed with different methods and from different corpora. Wherever it is possible, a description of the parameters used to compute the embeddings is included, together with simple statistics of the vectors, vocabulary, and description of the corpus from which the embeddings were computed. Direct links to the embeddings are provided, so please refer to the original sources for proper citation (also see References). An example of the use of some of these embeddings can be found here or in this tutorial (both in Spanish).

Summary (and links) for the embeddings in this page:

	Corpus	Size	Algorithm	#vectors	vec-dim	Credits
1	Spanish Unannotated Corpora	2.6B	FastText	1,313,423	300	José Cañete
2	Spanish Billion Word Corpus	1.4B	FastText	855,380	300	Jorge Pérez
3	Spanish Billion Word Corpus	1.4B	Glove	855,380	300	Jorge Pérez
4	Spanish Billion Word Corpus	1.4B	Word2Vec	1,000,653	300	Cristian Cardellino

Neural-attention NLP model


** This is work in progress **

BETO: Spanish BERT

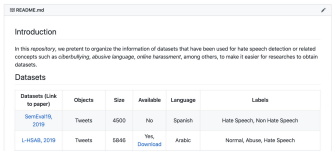
BETO is a BERT model trained on a big Spanish corpus. BETO is of size similar to a BERT-Base and was trained with the Whole Word Masking technique. Below you find Tensorflow and Pytorch checkpoints for the uncased and cased versions, as well as some results for Spanish benchmarks comparing BETO with Multilingual BERT as well as other (not BERT-based) models.

Download

BETO uncased	tensorflow_weights	pytorch_weights	vocab, config
BETO cased	tensorflow_weights	pytorch_weights	vocab, config

All models use a vocabulary of about 31k BPE subwords constructed using SentencePiece and were trained for 2M steps.

Hate-speech evaluation datasets

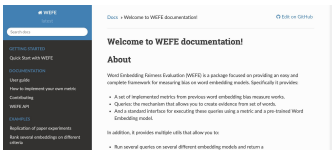

Introduction

In this repository, we pretent to organize the information of datasets that have been used for hate speech detection or related concepts such as cyberbullying, abusive language, online harassment, among others, to make it easier for researches to obtain datasets.

Datasets

Datasets (Link to paper)	Objects	Size	Available	Language	Labels
SemEval19, 2019	Tweets	4500	No	Spanish	Hate Speech, Non-Hate Speech
L-HSAB, 2019	Tweets	5846	Yes, Download	Arabic	Normal, Abuse, Hate Speech

Fairness evaluation for word representations


WEFE

Welcome to WEFE documentation!

About

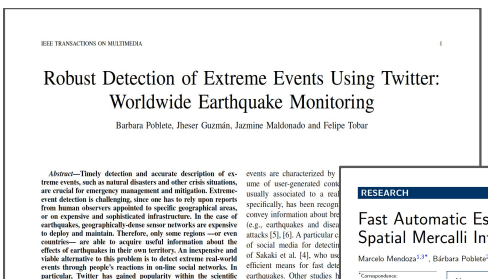
Word Embedding Fairness Evaluation (WEFE) is a package focused on providing an easy and complete framework for measuring bias on word embedding models. Specifically it provides:

- A set of implemented metrics from previous word embedding bias measure works.
- Queries: the mechanism that allows you to create evidence from set of words.
- And a standard interface for executing these queries using a metric and a pre-trained Word Embedding model.

In addition, it provides multiple utils that allow you to:

- Run several queries on several different embedding models and return a

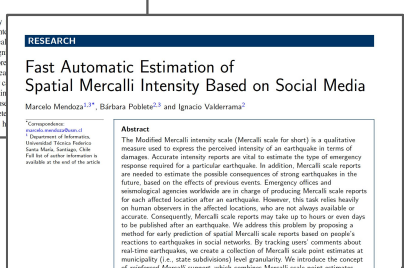
There is **value in the collective** of online social networks, we need to provide tools to **make good quality (aggregated) content available**


IEEE TRANSACTIONS ON MULTIMEDIA

Robust Detection of Extreme Events Using Twitter: Worldwide Earthquake Monitoring

Barbara Poblete, Jheser Guzmán, Jazmine Maldonado and Felipe Tobar

Abstract—Timely detection and accurate description of extreme events, such as natural disasters and other crisis situations, are crucial for emergency management and mitigation. Extreme-event detection is challenging, since one has to rely upon reports from human observers appointed to specific geographical areas, or on expensive and sophisticated infrastructure. In the case of earthquakes, geographically-dense sensor networks are expensive to deploy and maintain. Therefore, only some regions —or even countries— are able to acquire useful information about the effects of earthquakes in their own territory. An inexpensive and viable alternative to this problem is to detect extreme real-world events through people's reactions in on-line social networks. In particular, Twitter has gained popularity within the scientific community for providing access to real-time "citizen sensor"


RESEARCH

Fast Automatic Estimation of Spatial Mercalli Intensity Based on Social Media

Marcelo Mendoza[1,3*], Bárbara Poblete[2,3] and Ignacio Valderrama[2]

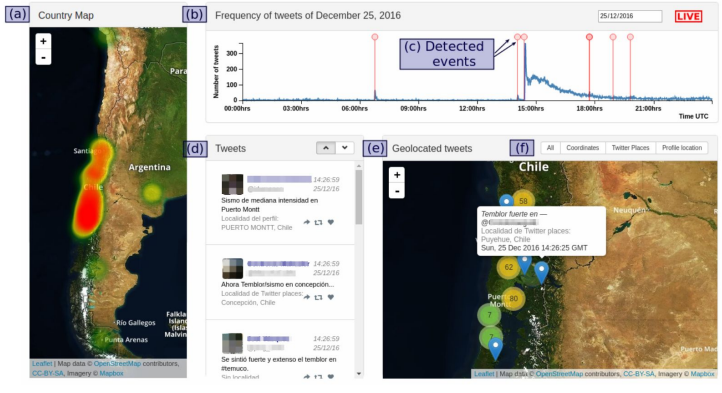
*Correspondence:

[1] Department of Informatics, Universidad Técnica Federico Santa María, Santiago, Chile

Full list of author information is available at the end of the article

Abstract

The Modified Mercalli intensity scale (Mercalli scale for short) is a qualitative measure used to express the perceived intensity of an earthquake in terms of damages. Accurate intensity reports are vital to estimate the type of emergency response required for a particular earthquake. In addition, Mercalli scale reports are needed to estimate the possible consequences of strong earthquakes in the future, based on the effects of previous events. Emergency offices and seismological agencies worldwide are in charge of producing Mercalli scale reports for each affected location after an earthquake. However, this task relies heavily on human observers in the affected locations, who are not always available or accurate. Consequently, Mercalli scale reports may take up to hours or even days to be published after an earthquake. We address this problem by proposing a method for early prediction of spatial Mercalli scale reports based on people's reactions to earthquakes in social networks. By tracking users' comments about real-time earthquakes, we create a collection of Mercalli scale point estimates at municipality (i.e., state subdivisions) level granularity. We introduce the concept


(a) Country Map

(b) Frequency of tweets of December 25, 2016

25/12/2016 LIVE

(c) Detected events

Number of tweets

00:00hrs 03:00hrs 06:00hrs 09:00hrs 12:00hrs 15:00hrs 18:00hrs 21:00hrs

Time UTC

(d) Tweets

Sismo de mediana intensidad en Puerto Montt

Localidad del perfil: PUERTO MONTT, Chile

Ahora Temblor/sismo en concepción...

Localidad de Twitter places: Concepción, Chile

Se sintió fuerte y extenso el temblor en #temuco.

(e) Geolocated tweets

(f) All Coordinates Twitter Places Profile location

Chile

Temblor fuerte en —

Localidad de Twitter places: Puyehue, Chile

Sun, 25 Dec 2016 14:26:25 GMT

Argentina

Río Gallegos

Punta Arenas

Falkland Islands (Islas Malvinas)

Neuquén

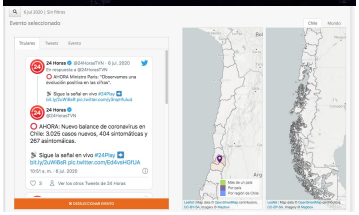
Puerto Montt

Puerto Madryn

Leaflet | Map data © OpenStreetMap contributors, CC-BY-SA, Imagery © Mapbox

http://www.twicalli.cl

http://galean.cl



Mining Social Networks to Learn about Rumors, Hate Speech, Bias and Polarization

Barbara Poblete

@bpoblete
www.barbara.cl

Department of CS, University of Chile
Millennium Institute for Fundamental Research on Data