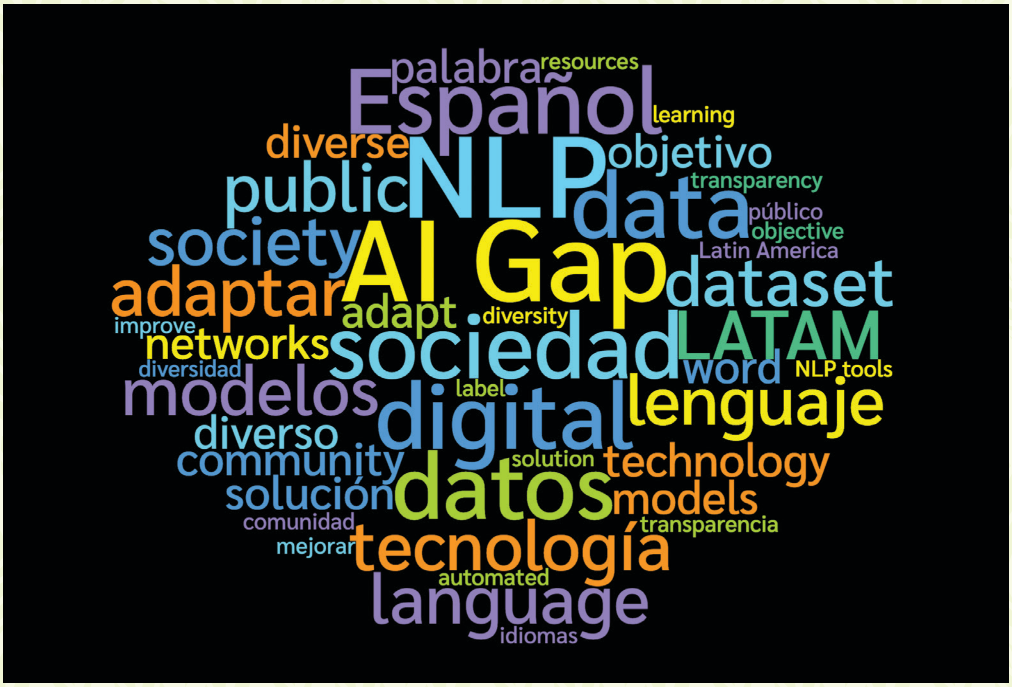
Natural Language Processing | DOI:10.1145/3416969

Minding the AI Gap in LATAM

BY BARBARA POBLETE AND JORGE PÉREZ

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 underrepresented 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 deal with such issues, our group has worked toward creating *language-agnostic approaches* as well as adapting and improving existing NLP techniques to local problems. In addition, we have focused on producing new state-of-the-art NLP publicly available data and models in Spanish. Next, we briefly present some of them.



WORD CLOUD GENERATED BY WORDCLOUDS.COM

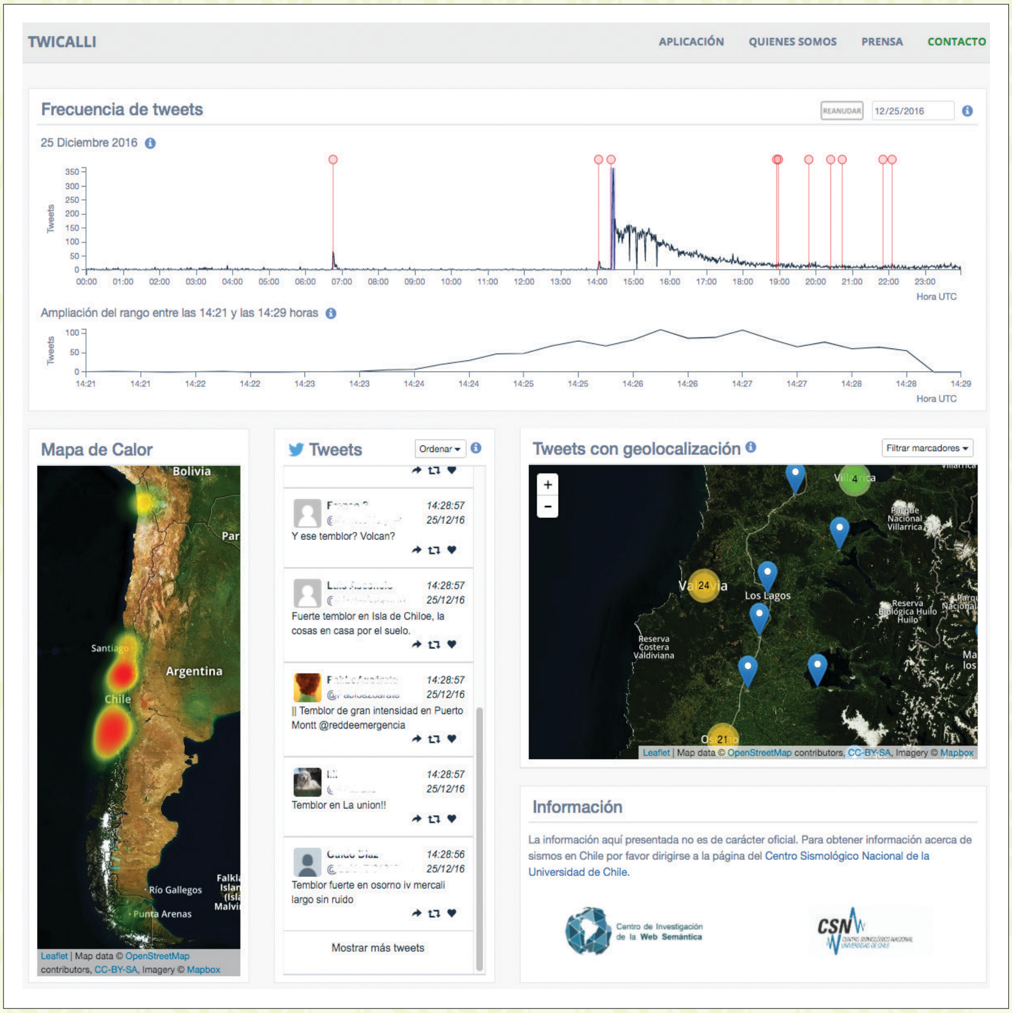
Our group has worked toward creating *language-agnostic approaches* as well as adapting and improving existing NLP techniques to local problems.

Twicalli, a social seismograph, and other crisis management tools. Timely detection and accurate description of natural disasters and other crisis situations are crucial for emergency management. This is challenging and important for our region, since one must rely on human observers appointed to specific geographical areas or on advanced infrastructure. In the case of earthquakes, geographically dense sensor networks are expensive. A viable inexpensive alternative to this problem is to detect events through people's reactions in online social networks, particularly on Twitter.[a]

Nevertheless, the massive number of messages in the data stream, along with the noise they contain, create a number of difficulties for worldwide detection. The common solution used to date has been to learn from labeled data to identify user messages related to a real-time earthquake.[2,12] This approach does not scale well globally across countries and languages. Consequently, our group proposed a simple, yet efficient solution that, generally speaking, identifies the unusual increase in the frequency of multilingual textual features related to earthquakes within the Twitter stream.[10] This method only requires a one-off semi-supervised initialization and can be scaled to track multiple features and thus, multiple

[a] Microblogging and social networking service http://twitter.com

crisis situations. Experimental results validate our approach as a competitive open source alternative to leading solutions, with the advantage of working independently of language and providing worldwide scalability. An instance of this framework is currently available as *Twicalli* (http://twicalli.cl) that provides visual "social seismograph" for the Chilean territory (as shown in Figure 1). Twicalli is used by the National Seismology Center in Chile, among other emergency response agencies nationwide. Furthermore, we are in the process of incorporating a novel machine learning-based model for automatically estimating the Modified Mercalli Intensity Scale[b] of an earthquake,[9] and detecting in real time other types of crisis situations.[13]



TWICALLI

APLICACIÓN QUIENES SOMOS PRENSA CONTACTO

Frecuencia de tweets

REANUDAR 12/25/2016

25 Diciembre 2016

Ampliación del rango entre las 14:21 y las 14:29 horas

Mapa de Calor

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

Tweets

Ordenar

14:28:57 25/12/16
Y ese temblor? Volcan?

14:28:57 25/12/16
Fuerte temblor en Isla de Chiloe, la cosas en casa por el suelo.

14:28:57 25/12/16
|| Temblor de gran intensidad en Puerto Montt @reddeemergencia

14:28:57 25/12/16
Temblor en La union!!

Guido Diaz
14:28:56 25/12/16
Temblor fuerte en osorno iv mercali largo sin ruido

Mostrar más tweets

Tweets con geolocalización

Filtrar marcadores

+
−
Villarrica
4
Parque Nacional Villarrica
Valdivia
24
Los Lagos
Reserva Costera Valdiviana
Reserva Biológica Huilo Huilo
Osorno
21

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

Información

La información aquí presentada no es de carácter oficial. Para obtener información acerca de sismos en Chile por favor dirigirse a la página del Centro Sismológico Nacional de la Universidad de Chile.

Centro de Investigación de la Web Semántica

CSN Centro Sismológico Nacional Universidad de Chile

Figure 1. Twicalli interface showing a large earthquake in Chile on December 25, 2016 and its aftershocks. (Top right) Shows the frequency of earthquake-related messages per minute. (Left) Shows a heat map of geographical message density distribution. (Bottom center) Displays user messages. (Bottom right) Shows fine-grained geographical message navigation. Image courtesy of Jazmine Maldonado.

Political data and the Open Constitution Project. In 2016, Chile went through a collective open process to establish the guidelines of what a new political constitution should consider. From a technological point of view, an interesting aspect of the process was the use of digital platforms to collect the output of the deliberative instances that included discussions produced by 8,000+ small assemblies across the country. This resulted in a dataset of 200,000+ political arguments, which was openly published in a raw and anonymous form. This dataset was manually systematized through months of work. Hence, with the goal of making this process less time consuming, more objective, and scalable, we worked toward creating ways to automate at least parts of the systematization. With the help of an interdisciplinary group of linguists and experts in argumentation, we first determined the most important tasks and then developed machine learning methods to solve them. For example, we addressed the task of classifying raw text arguments into the corresponding "constitutional concept." For instance, the raw text "*the state should provide free education for all*" should be assigned to the concept "*right to education.*" This was challenging, given we had more than 100 different constitutional concepts. Our best method achieved a top-5 accuracy of more than 90%.[6] We created a visualization for exploring the dataset

With the help of an interdisciplinary group of linguists and experts in argumentation, we first determined the most important tasks and then developed machine learning methods to solve them.

[b] https://www.usgs.gov/natural-hazards/earthquake-hazards/science/modified-mercalli-intensity-scale

(http://constitucionabierta.cl/), which was widely used by the public and press, providing a much needed transparency to such an important process (see Figure 2).

Two problems naturally arise in the context of our work analyzing social and political discussions; that of incivility, or hate speech,[1,7] and the now ubiquitous problem of bias against minorities.[3,8] Modern automatic tools for processing human-generated text should consider these issues as essential. However, these tasks are extremely challenging even for the English language. Specifically, our work has addressed how the lack of diversity in training data for hate-speech detection models induces an overestimation of their performance in state-of-the-art approaches, and how this affects transferring this knowledge to other domains,[1] such as Spanish.

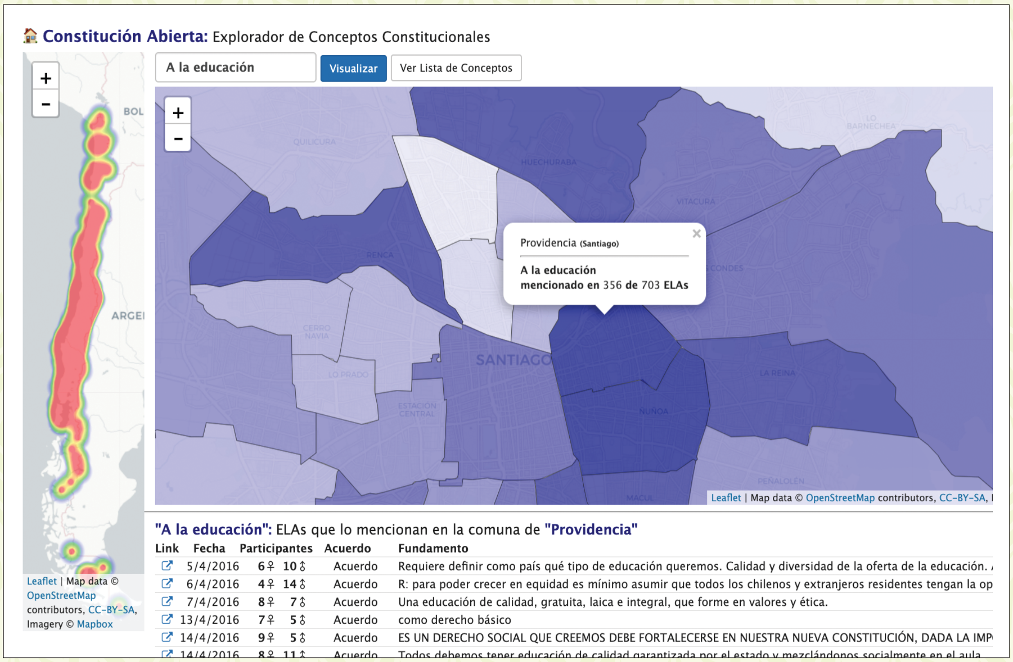
Spanish NLP resources. An issue we have constantly faced, as have other researchers working with Spanish text data, is the lack of high-quality resources for developing, training, and testing models. Some LATAM and Spanish groups have been tackling this problem by systematically producing freely available NLP resources for the whole research community.[4,5,11,14,15] One set of resources of particular importance are Word Embeddings trained from big corpora.[5,14,15] Our group has also developed a Spanish pretrained model based on Neural Self-Attention[4] that has improved the state of the art in many NLP tasks in Spanish. We hope more groups in Latin America can join efforts to produce resources to improve NLP research and applications.

Acknowledgments. BP and JP were partially funded by Fondecyt grants 1191604 and 1200967, respectively. 🅲

An issue we have constantly faced, as have other researchers working with Spanish text data, is the lack of high-quality resources for developing, training, and testing models.

References

1. Arango, A., Pérez, J. and Poblete, B. Hate speech detection is not as easy as you may think: A closer look at model validation. *SIGIR* 2019: 45-54.
2. Avvenuti, M., Cresci, S., Marchetti, A., Meletti, C. and Tesconi, M. (2014, August). "EARS (earthquake alert and report system) a real time decision support system for earthquake crisis management. In *Proceedings of the 20th ACM SIGKDD Intern. Conf. Knowledge Discovery and Data Mining*, Aug. 2014, 1749–1758.
3. Badilla, P., Bravo-Marquez, F. and Pérez, J. WEFE: The word embeddings fairness evaluation framework. To appear in *Proceedings of the 29th Intern. Joint Conf. Artificial Intelligence*, 2020.
4. Cañete, J., Chaperon, G., Fuentes, R., Ho, J-H., Kang, H. and Perez, J. Spanish pre-trained BERT model and evaluation data. In Practical Machine Learning for Developing Countries at ICLR 2020; https://github.com/dccuchile/beto
5. Etcheverry, M. and Wonsever, D. Spanish word vectors from Wikipedia. In *Proceedings of 2016 Intern. Conf. Language Resources and Evaluation.*
6. Fierro, C., Fuentes, C., Pérez, J. and Quezada, M. 200K+ crowdsourced political arguments for a new Chilean constitution. In *Proceedings of the 4th Workshop on Argument Mining*, Sept. 2017, 1-10.
7. Fortuna, P. and Nunes, S. A survey on automatic detection of hate speech in text. *ACM Comput. Surv. 51*, 4 (2018), 85:1–85:30.
8. Garg, N., Schiebinger, L., Jurafsky, D. and Zou, J. Word embeddings quantify 100 years of gender and ethnic stereotypes. In *Proceedings of the National Academy of Sciences*, *115*, 16 (2018), E3635–E3644.
9. Mendoza, M., Poblete, B. and Valderrama, I. Nowcasting earthquake damages with Twitter. *EPJ Data Sci. 8*, 3 (2019). https://doi.org/10.1140/epjds/s13688-019-0181-0
10. Poblete, B., Guzmán, J., Maldonado, J. and Tobar, F. Robust detection of extreme events using Twitter: Worldwide earthquake monitoring. *IEEE Trans. Multimedia 20*, 10, (Oct. 2018), 2551–2561; doi: 10.1109/TMM.2018.2855107.
11. Recursos, Grupo de Procesamiento de Lenguaje Natural, Universidad de La República; https://www.fing.edu.uy/inco/grupos/pln/recursos.html
12. Sakaki, T., Okazaki, M. and Matsuo, Y. Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th Intern. Conf. World Wide Web*, Apr. 2010, 851–860.
13. Sarmiento, H., Poblete, B. and Campos, J. Domain-independent detection of emergency situations based on social activity related to geolocations. In *Proceedings of the 10th ACM Conf. Web Science*, 2018; https://doi.org/10.1145/3201064.3201077
14. Soares, F., Villegas, M., Gonzalez-Agirre, A., Krallinger, M. and Armengol-Estapé, J. Medical word embeddings for Spanish: Development and evaluation. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop at ACL 2019.*
15. Spanish Word Embeddings; https://github.com/dccuchile/spanish-word-embeddings



🏠 **Constitución Abierta:** Explorador de Conceptos Constitucionales

Providencia (Santiago)

A la educación
mencionado en 356 de 703 ELAs

"A la educación": ELAs que lo mencionan en la comuna de **"Providencia"**

Link	Fecha	Participantes	Acuerdo	Fundamento
↗	5/4/2016	6♀ 10♂	Acuerdo	Requiere definir como país qué tipo de educación queremos. Calidad y diversidad de la oferta de la educación.
↗	6/4/2016	4♀ 14♂	Acuerdo	R: para poder crecer en equidad es mínimo asumir que todos los chilenos y extranjeros residentes tengan la op
↗	7/4/2016	8♀ 7♂	Acuerdo	Una educación de calidad, gratuita, laica e integral, que forme en valores y ética.
↗	13/4/2016	7♀ 5♂	Acuerdo	como derecho básico
↗	14/4/2016	9♀ 5♂	Acuerdo	ES UN DERECHO SOCIAL QUE CREEMOS DEBE FORTALECERSE EN NUESTRA NUEVA CONSTITUCIÓN, DADA LA IMP
↗	14/4/2016	8♀ 11♂	Acuerdo	Todos debemos tener educación de calidad garantizada por el estado y mezclándonos socialmente en el aula

Figure 2. Interface of the website http://constitucionabierta.cl showing the distribution of the constitutional concept "right to education" in the 2016 Chilean Constitutional discussions dataset.

Barbara Poblete is an associate professor in the Department of Computer Science at Universidad de Chile and an associate researcher at the Millennium Institute for Foundational Research on Data, Santiago, Chile.

Jorge Pérez is an associate professor in the Department of Computer Science at Universidad de Chile and an associate researcher at the Millennium Institute for Foundational Research on Data, Santiago, Chile.