Proposal: Automatic topic modelling of official communication from Spanish far right deputies during the COVID crisis

Olivier R. Philippe

Introduction

The present proposal focuses on the discussion by the Spanish far right elected representative during the COVID-19 crisis. It explores the topics around which the Twitter discourse from a selected VOX members' accounts revolves around. The topics detection model will be perform using different unsupervised topic modelling methods and manual labelling of the resulting topics distribution for further analysis.

Context

Twitter is often used as an alternative media for health information [12] In case of epidemic and pandemic the uncertainty and fast news can lead to an increase of false information as the verification is not often done by the users and lies and false information spread faster than reliable ones [15].

This phenomenon happened during the H1N1 outbreak in 2009 the Ebola in 2013 and the Zika epidemic in 2015 [10], [16], [4]. Naturally the COVID-19 pandemic raises the same issues [18],[7].

There is a complete descriptive publication on the discussion on Twitter about the COVID-19 and the relation with misinformation [13]. However, the COVID-19 is more general than a health issues, considering the novelty, extend and impact of the confinement and health misinformation is only one aspect. Populism and far right in particular are also analysed and their own reaction to the COVID-19 and the spreading of conspiracy theories [9], [1] and the increase of hate messages during this pandemic [14].

In this context of rapid change and fertile context for misinformation, it is important to unveil the ensemble of topics that far right talk about, and not only about the risk posed by health misinformation. This is why topic modelling is an important tool to tackle this issue.

Data and methods

Topic modelling has been used in the context of COVID-19 crisis. Wicke and Bolognesi used the latent Dirichlet allocation LDA method an their existing framework to model discourse on COVID-19 [17], Lwin et al. did a sentiment analysis on the Twitter trends [8] and Haman et al. on the communication by state leaders [6].

A contrario to existing studies using large set of available tweets about the COVID-19 [2], or following keywords of pre-identified issues and hashtags, our data target is a finite set of users, the entire elected deputies elected in 2019 and present on Twitter.

The data collection covers the entire period of 1st January 2020 until the 15th of July 2020, accessed using the REST API from Twitter [11]. The method used is a combination of a fine-tuned language model [19], [5] on the dataset and using the created word embeddings with LDA [3] to model the topic distribution. From there, it is possible to compare between different discourses made by elected officials in Spain, and unveil the difference between parties and see if the far right differs from the others parties.

References

[1] Ahmed, W. et al. 2020. COVID-19 and the 5G Conspiracy Theory: Social Network Analysis of Twitter Data. *Journal of Medical Internet Research.* 22, 5 (May 2020), e19458. DOI:https://doi.org/10.2196/19458.

[2] Banda, J.M. et al. 2020. A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration. (May 2020). DOI:https://doi.org/10.5281/zenodo.3723939.

[3] Blei, D.M. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research.* 3, (2003), 993–1022.

[4] Bode, L. and Vraga, E.K. 2018. See Something, Say Something: Correction of Global Health Misinformation on Social Media. *Health Communication.* 33, 9 (Sep. 2018), 1131–1140. DOI:https://doi.org/10.1080/10410236.2017.1331312.

[5] Devlin, J. et al. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs].* (May 2019).

[6] Haman, M. 2020. The use of Twitter by state leaders and its impact on the public during the COVID-19 pandemic. (2020), 16.

[7] Huang, B. and Carley, K.M. 2020. Disinformation and Misinformation on Twitter during the Novel Coronavirus Outbreak. *arXiv:2006.04278 [cs].* (Jun. 2020).

[8] Lwin, M.O. et al. 2020. Global Sentiments Surrounding the COVID-19 Pandemic on Twitter: Analysis of Twitter Trends. *JMIR Public Health and Surveillance.* 6, 2 (May 2020), e19447. DOI:https://doi.org/10.2196/19447.

[9] Mason, A. 2020. Europe's Future: The Impact of Covid-19 on Populism. (2020), 10.

[10] Miller, M. et al. 2017. What Are People Tweeting About Zika? An Exploratory Study Concerning Its Symptoms, Treatment, Transmission, and Prevention. *JMIR Public Health and Surveillance.* 3, 2 (Jun. 2017), e38. DOI:https://doi.org/10.2196/publichealth.7157.

[11] POST statuses/filter: *https://developer.twitter.com/en/docs/tweets/filter-realtime/api-reference/post-statuses-filter.*

[12] Radzikowski, J. et al. 2016. The Measles Vaccination Narrative in Twitter: A Quantitative Analysis. *JMIR Public Health and Surveillance.* 2, 1 (Jan. 2016), e1. DOI:https://doi.org/10.2196/publichealth.5059.

[13] Singh, L. et al. 2020. A first look at COVID-19 information and misinformation sharing on Twitter. *arXiv:2003.13907 [cs].* (Mar. 2020).

[14] Velásquez, N. et al. 2020. Hate multiverse spreads malicious COVID-19 content online beyond individual platform control. (2020).

[15] Vosoughi, S. et al. 2018. The spread of true and false news online. *Science.* 359, 6380 (Mar. 2018), 1146–1151. DOI:https://doi.org/10.1126/science.aap9559.

[16] Wang, Y. et al. 2019. Systematic Literature Review on the Spread of Health-related Misinformation on Social Media. *Social Science & Medicine.* 240, (Nov. 2019), 112552. DOI:https://doi.org/10.1016/j.socscimed.2019.112552.

[17] Wicke, P. and Bolognesi, M.M. Framing COVID-19: How we conceptualize and discuss the pandemic on Twitter. 41.

[18] Xaudiera, S. and Cardenal, A.S. 2020. Ibuprofen Narratives in Five European Countries During the COVID-19 Pandemic. *Harvard Kennedy School Misinformation Review.* (Jul. 2020). DOI:https://doi.org/10.37016/mr-2020-029.

[19] Xie, Q. et al. 2020. Monolingual and multilingual topic analysis using LDA and BERT embeddings. *Journal of Informetrics.* 14, 3 (Aug. 2020), 101055. DOI:https://doi.org/10.1016/j.joi.2020.101055.