Inferencia causal en series de tiempo de Twitter y encuestas políticas

Albanese, Federico[1,2,†], Baldonado, Juan Manuel[3], and Feuerstein, Esteban[2,3]

[1]Instituto de Cálculo (IC), Universidad de Buenos Aires (UBA)
[2] Instituto de Ciencias de la Computación (ICC), Universidad de Buenos Aires (UBA)
[3]Departamento de Computación - FCEyN - UBA
[†]falbanese@dc.uba.ar

Agosto 2022

Palabras Claves: Inferencia causal; series de tiempo; Twitter; redes sociales; técnicas de procesamiento del lenguaje natural.

Resumen

Las redes sociales han sido utilizadas como medios para la discusión política de los ciudadanos [1]. En este trabajo nos propusimos analizar la influencia que ejerce el discurso en las redes sociales sobre la opinión pública de los candidatos políticos en contextos electorales. Para ello conformamos un dataset con 4.4 millones de tweets políticos durante las elecciones presidenciales estadounidenses entre Trump y Biden del 2020 y analizamos 229 encuestas presidenciales realizadas por 29 encuestadores. Luego, armamos series temporales con los resultados de las encuestas y con la cantidad, tópico del que hablan y sentimiento (positividad / negatividad) de los tweets que mencionan a cada candidato, usando técnicas de procesamiento del lenguaje natural de forma similar a trabajos previos [2]. Aplicando herramientas de inferencia causal en series de tiempo [3] [4] como la causalidad de Granger [5], Información Mutua Condicional [6] y Base de Funciones Radiales [7], encontramos resultados estadísticamente significativos de una relación causal. En particular, el sentimiento con el que se habla de los candidato y ciertos tópicos particulares que se debaten en Twitter impactan sobre la intención de voto que finalmente recibe cada candidato presidencial.

Referencias

[1] Albanese, F., Lombardi, L., Feuerstein, E., & Balenzuela, P. (2020). Breaking the Communities: Characterizing community changing users using text mining and graph machine learning on Twitter. arXiv e-prints, arXiv-2008.

[2] Albanese, F., Pinto, S., Semeshenko, V., & Balenzuela, P. (2020). Analyzing mass media influence using natural language processing and time series analysis. Journal of Physics: Complexity, 1(2), 025005.

[3] Nath, R., Gupta, N. K., Gupta, N., Tiwari, P., Kishore, J., & Ish, P. (2022). Effect of COVID-19 pandemic on tuberculosis notification. The Indian Journal of Tuberculosis, 69(3), 364.

[4] Tabari, N., Biswas, P., Praneeth, B., Seyeditabari, A., Hadzikadic, M., & Zadrozny, W. (2018, July). Causality analysis of Twitter sentiments and stock market returns. In Proceedings of the first workshop on economics and natural language processing (pp. 11-19).

[5] Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. Econometrica: journal of the Econometric Society, 424-438.

[6] Wyner, A. D. (1978). A definition of conditional mutual information for arbitrary ensembles. Information and Control, 38(1), 51-59.

[7] Ancona, N., Marinazzo, D., & Stramaglia, S. (2004). Radial basis function approach to nonlinear Granger causality of time series. Physical Review E, 70(5), 056221.

Versión en Ingles

Title: Time series causal inference between the political discourse on Twitter and opinion polls

Key words: Causal inference; Time series; Twitter; social media; Natural language processing.

Abstract

In recent years, social networks have been the place where citizens exchange their political opinions [1]. In this work, we analyzed the causal influence between the discourse in social networks on the opinion of political candidates in electoral contexts. Therefore, we created a dataset with 4.4 million political tweets during the 2020 US presidential election between Trump and Biden and analyzed 229 presidential polls conducted by 29 different pollsters. Then, we create time series with the poll's results and with the quantity, topic and sentiment (positive / negative) of the tweets that mentioned the candidates, using natural language processing techniques similarly to previous works [2]. Using time series causal inference tools [3] [4] such as Granger causality [5], conditional mutual information [6] and radial basis function [7], we found significant results of causal relationships. In particular, the positivity / negativity of the tweets

and some topics discussed on Twitter had a causal impact on each presidential candidate's polling.