How did COVID-19 change the words? A study with social media and complex networks

Vanesa Copa[1] Sebastián Pinto[2,4 *], Laura Kaczer[1,3 *]

[1] Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales, Argentina
[2] Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales, Departamento de Física, Argentina.
[3] CONICET - Universidad de Buenos Aires, Instituto de Fisiología, Biología Molecular y Neurociencias (IFIByNE), Argentina
[4] CONICET - Universidad de Buenos Aires, Instituto de Física Interdisciplinaria y Aplicada (INFINA), Argentina
*Equal contribution
vanesajc2@gmail.com

Abstract. Understanding the change in word meaning across different contexts and time periods is crucial for revealing the role of language in social and cultural evolution. In this study, we examine the diachronic effect of the Covid-19 pandemic on the use and meaning of words through the lens of complex networks theory. Specifically, we analyze how Twitter users express themselves and how their language use was affected by the pandemic. Using network analysis, we construct networks corresponding to three different years (2019, 2020, 2021) to represent the relationship between the words used by Twitter users, allowing us to track semantic changes quantitatively. Our study focuses on a set of Twitter users that were compared through time. Overall, our research sheds light on the methodological possibilities offered and how complex networks theory can help us understand this impact.

Keywords: Diachronic semantic shifts, Covid-19, Complex networks, Twitter.

1 Introduction

Understanding the change in the meaning of words in different contexts and time periods is crucial for revealing the role of language in social and cultural evolution. The COVID-19 pandemic, one of the most influential global events in modern human history, offers a unique opportunity to quantify the drift of word concepts [3] that could eventually lead to semantic change. Therefore, in this study, we aimed to explore collective changes in the mental lexicon as a consequence of the COVID-19 pandemic. We are interested in addressing the plasticity of lexical-semantic representations in the context of the COVID-19 pandemic using a large volume of data from the analysis of language from the social network Twitter. A previous study addressed the plasticity and malleability of words caused by the COVID-19 pandemic by using a large word association database [4]. As a motivation for this work, we study whether it is possible to reproduce these controlled trials using data obtained from social networks such as Twitter.

Our hypothesis is that changes in the use and meaning of different terms will be reflected in the variation in topological properties of the semantic networks constructed in different stages of the pandemic (2019, 2020 and 2021). To this end, we studied the semantic networks associated with different terms (related or unrelated to COVID-19).

2 Dataset

The dataset consists of original public geo-tagged tweets whose publication date belongs to the period before 2019 and different stages of the COVID-19 pandemic years 2020 and 2021. The tweets were limited to those written in Spanish, which corresponds to Argentina.

We selected 200 study users to perform a follow-up somewhat analogous to that produced by participants in the word association experiment to follow their tweets between January 1, 2019, and December 31, 2021. The criteria for selecting users were based on their having some activity in the three periods and having used the term "vacuna" (vaccine) at some point during the pandemic. All the tweets produced by the selected users were incorporated into the dataset. The tweets and accompanying metadata were downloaded from the Twitter API v2, which allows access to all historical content on the platform for academic purposes (https://developer.twitter.com/en/products/twitter-api/academic-research).

The resulting dataset consisted of 3,477,423 tweets published by 200 authors, of which 1,116,160 belonged to 2019, 1,237,920 to 2020, and 1,123,343 to 2021. The three datasets are balanced.

The sizes of vocabulary were 264022 , 272021, and 251482 words for, respectively, 2019, 2020 and 2021.

3 Construction of the Semantic Network

3.1 Preprocessing

Preprocessing prepares texts to translate the textual information into the numerical information necessary for the construction of semantic networks. Tokenization was performed by discarding URLs and mentions, emoji symbols, and special characters. Each word was then transformed to lowercase, and empty words were removed using the list of empty words in the Python NLTK package. In addition, the tokens were normalized through lemmatization using a library from the NLP department of Stanford University, Stanza [5].

3.2 Weighted Semantic Networks

After pre-processing the tweets, we generated complex semantic networks. A semantic network or network representation scheme represents linguistic knowledge, in which concepts and their interrelationships are represented by a graph.

One way to construct these networks is by measuring term co-occurrences using tweets. A link between the two terms is established if they are mentioned in the same tweet and its weight represents the number of co-occurrences. Because of the power law distribution of word frequency in natural language, some words can often be observed in the vicinity of other words simply because they are very common, and not because of the co-occurrence that they are really indicative of some meaning. In this study, we used normalized pointwise mutual information (NPMI) [2]. Pointwise Mutual Information (PMI) is a measure of how much the actual probability of a particular co-occurrence of events p(w1,w2) differs from what we would expect based on the probabilities of the individual events and the assumption of independence p(w1)p(w2). In our case, co-occurrence is defined as binary features within each tweet. NPMI is the normalization of the PMI, where NPMI values between -1 and 1 are obtained.

3.3 Graph Processing, Size Reduction

To identify meaningful clusters, the graphs were aggregated and then reduced in size. The graphs were aggregated by averaging the weights obtained for the different pairs (w1,w2) of all users. To reduce the size of the obtained network, the Weighted Network Reduction algorithm is applied, which extracts the truly relevant connections that form the backbone and preserves the edges that represent statistically significant deviations with respect to a null model for local assignment of edge weights [6]. A p-value of 0.05 is used. Finally, the giant component was selected.

4 Preliminary results

By averaging the networks of 200 users and reducing the size of the networks after applying the Weighted Network Reduction algorithm for 2019, 2020, and 2021, we obtained 20702, 21640, and 20255 nodes, respectively, for each main component.

4.1 Community detection

Community detection is important not only for characterizing the graph but also for providing information about the network's formation and functionality. The Louvain community detection algorithm [1] was used for the giant component of each period. Subsequently, for each period, ten communities containing the largest number of nodes were selected. For example, in 2019, these communities contained

97% of the total nodes. On the other hand, the centrality measure eigenvector (eigenvector centrality) was analyzed. This indicates that a node is important if it is linked to another important node. Therefore, the first 30 nodes of each community were filtered according to the obtained eigenvector value. Based on these nodes, each community was labeled as a possible topic.

Tables 1 and 2 show the main words related to what we called "food and nutrition", and "Covid-19 pandemic".

Table 1. Main words related to "food and nutrition"

Año	Main words, sorted by Eigenvector
2019	eat, buy, milk, cheese, bread, meat, water, egg, rich, sweet, price, coffee, fruit, wine, rice, barbecued, potatoes, oil, chicken, flour
2020	put, go out, day, stay, good, house, buy, fine, come, water, eat, after, take, carry, use, milk, food, egg, rich, white
2021	eat, buy, meat, cheese, rich, milk, bread, stay, water, food, egg, barbecued, fruit, coffee, wine, chicken, oil, oven, noodle, sweet

Table 2. Main words related to "Covid-19 Pandemic"

Año	Main words, sorted by Eigenvector
2020	coronaviru, case, covid, health, person, arrive, quarantine, virus, doctor, first, vaccine, contagion, hospital, chinese, patient, via, disease, infect, control, travel
2021	vaccine, argentinian, covid, country, vaccinate, arrive, dose, first, health, case, vaccination, new, vaccinated, use, major, doctor, sputnik, risk, virus, apply, coronaviru

As can be observed, each community contained a set of words that were closely related to the defined meaning.

4.2 Inspection of the semantic networks

To determine the changes that have occurred in the different periods of time (2019, 2020, and 2021) in the semantic networks associated with the incorporation of new meanings, a list of control words and a list of pandemic words were defined. Based on the word association experiment, the selected pandemic words were "protocol", "alcohol", "bubble", "corona", and "virtual", and the control words were "agreement", "clear", "thing", "word", and "history".

For instance, Figure 1 shows a subgraph containing the word "protocol" in the three periods. It can be observed that the qualitative neighborhood of the study node changes in the three periods. In the 2019 period, it can be observed that the word is closer to abortion law issues, while in 2020 it became semantically closer to health-related issues and finally in 2021 the term began to be related to education. This example shows that the obtained semantic networks allow the detection of lexical semantic changes that occurred in the studied periods.

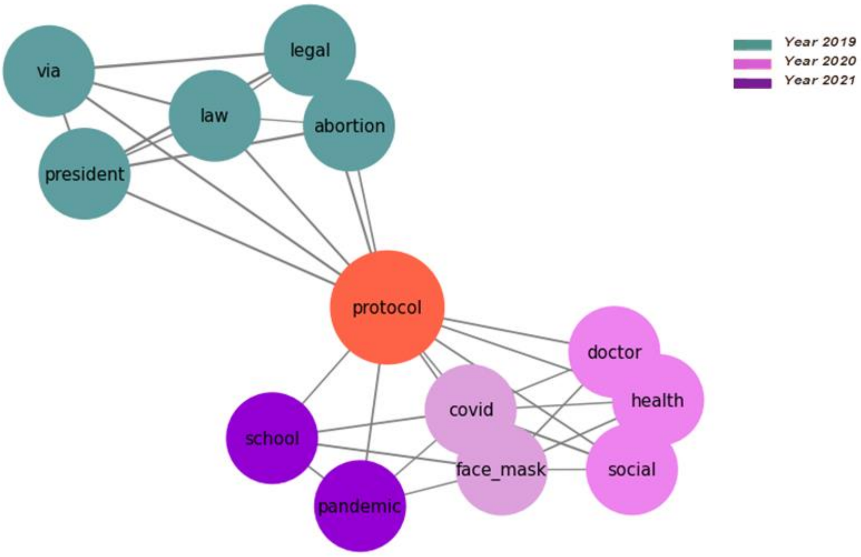

Fig. 1. Subgraph of word "protocol" from the three periods. Each period is pointed out in the figure. It can be observed the change in the neighborhood of the word "protocol" associated with a change in its semantic use.

5 Conclusions

In this study, we conducted a comparative semantic analysis during the pre-pandemic and Covid-19 pandemic periods by taking information from social media. To some extent, we aim to reproduce a large-scale word association task [4] using information provided by users from Twitter. We propose that the use of complex networks, built based on the language used on the Twitter social network, can be seen as a new methodology that could be helpful in tracking semantic changes produced in different periods of time [3]. Preliminary results show that semantic changes can be detected both in the emergence of new communities in the semantic network and in the change of the neighborhood of targeted words.

References

1. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment 2008(10), P10008 (2008)

2. Bouma, G.: Normalized (pointwise) mutual information in collocation extraction. Proceedings of GSCL 30, 31–40 (2009)

3. Carrillo, F., Cecchi, G.A., Sigman, M., Slezak, D.F.: Fast distributed dynamics of semantic networks via social media. Computational intelligence and neuroscience 2015, 50–50 (2015)

4. Laurino, J., De Deyne, S., Cabana, Aa. Kaczer, L.: The pandemic in words: tracking fast semantic changes via a large-scale word association task. Open mind, 1-19 (2023)

5. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D.: Stanza: A python natural language processing toolkit for many human languages. arXiv preprint arXiv:2003.07082 (2020)

6. Serrano, M.A., Bogun ́a, M., Vespignani, A.: Extracting the multiscale backbone of complex weighted networks. Proceedings of the national academy of sciences 106(16), 6483–6488 (2009)