Capturing and analyzing social representations. A first application of Natural Language Processing techniques to reader's comments in COVID-19 news. Argentina, 2020

Germán Rosati[1,2,3 +], Laia Domenech[1 +], Adriana Chazarreta[1,2 +], Tomás Maguire[1 +]

[1] Instituto de Altos Estudios Sociales, Universidad Nacional de San Martín, ARG
[2] Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET, ARG
[2] Programa de Investigaciones Sobre el Movimiento de la Sociedad Argentina (PIMSA), ARG
[+] Equal contributor

german.rosati@gmail.com

Abstract. We present a first approximation to the quantification of social representations about the COVID-19, using news comments. A web crawler was developed to construct the dataset of reader's comments. We detect relevant topics in the dataset using Latent Dirichlet Allocation, and analyze its evolution during time. Finally, we show a first prototype to the prediction of the majority topics, using FastText.

Keywords: NLP, News Comments, COVID-19, Social Representations.

1 Introduction

Generally speaking, social or collective representations are sociopsychological constructs that perform a symbolic role, representing something –an object– to a person or group [8]. They consist in systems of values, ideas and practices [8], and are a largely discussed theme in social science. From early approaches linked to the marxist notion of ideology [4] [7] to more modern definitions of the term [8]. Historical studies have also addressed this issue: for example the "inherent ideas" in different forms of pre-industrial riots [9]. These collective representations should not be seen as logical and consistently thought structures: they can be formed by incoherent fragments of ideas.

They have been classically studied using a variety of sources [11]. The most relevant ones are sampling surveys [1] [8], manual text analysis [9] and in-depth interviews [11]. More modern approaches have worked with social networks [3], especially analyzing political representations and its diffusion. However, other spontaneous discourse sources, like news reader's comments seem not to wake the same interest as a source for analyzing these problems. There were, however, some attempts: Schuth et al [10] tried to extract the discussion structure of reader's comments in several dutch newspapers.

Is it possible to construct relevant information about social representations regarding a specific theme using news reader's comments? Which are the main topics discussed in news comments? How do these topics evolve over time? The aim of this paper is to present some preliminary results of an ongoing research addressing these questions using some novel data sources.

We present the whole workflow (from data gathering, preprocessing, modeling and predicting) of the reader's comments on COVID-19 news in five argentinian papers.

2 Data and methods[1]

2.1 Scraping news and comments

We study the comments of online news articles available in five newspapers (with national circulation): Página 12, La Nación, Clarín, Infobae and Ámbito Financiero. In order to get the reader's comments it was necessary to scrap the news themselves first. To do so, we made a series of GDELT[2] requests every one or two week intervals. This provided an input of all the article links by argentinian newspapers labeled as "COVID-19".

We developed a crawler, which we applied once we had each set of articles. This crawler ran through every link of our input database and scraped all the comments of each article. The final database contains all the 385.255 comments produced between 13/03/2020 and 01/06/2020.

[1] The web scraping, preprocess, modeling and prediction stages were developed in Python using standard libraries (BeautifulSoup, Selenium, scikit-learn, etc.). The visualizations were produced in R using ggplot library.

[2] GDELT (Global Database of Events, Language and Tone) -https://www.gdeltproject.org/- is a huge free access database about human society considered as the "bigger, most complete and with higher resolution" ever created. It is growed day by day monitoring online news in several countries with more than 100 languages, saving published news and identifying topics, locations, themes and emotions present in each article. It is updated every 15 minutes.

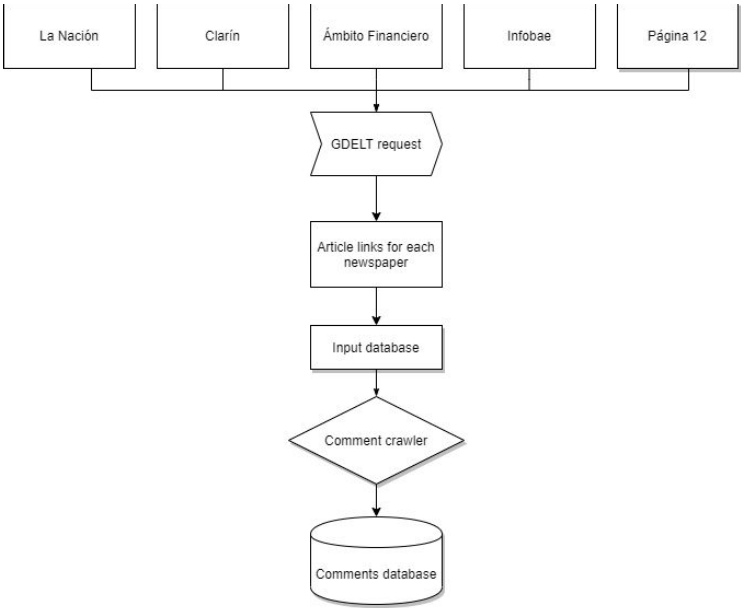

Fig. 1. Flowchart of data acquisition process.

2.2 Topic modeling with Latent Dirichlet Allocation

The second step was to train a topic model using LDA [2] to create an instrument which could be useful for an automatic tag of each comment.

LDA is a method for analyzing large quantities of unstructured data, in which each topic is characterized by a distribution of words. Each distribution of words provides a 'thematic summary' of the topic. In other words, we can read each topic as an answer to what are the social representations surrounding our input documents.

3 Experiments and results

We applied several standard preprocessing operations to text data: removing stopwords, punctuation, digits, web links and usernames and transforming to lowercase. We also constructed the Term-Frequency matrix using l1 normalization and TF-IDF weighting.

After iterating we found ten topics which were conceptually relevant.

Table 1. Topic identification and labeling

Order	Words distribution	Label
Topic 01	[alberto peronismo impuesto chile acuerdo razon excelente paso tipo ministro brasil ignorante test gato casos]	Government, cases and testing at the regional level
Topic 02	[vos sos gamurra tenes porota cuba troll venezuela foto matanza mano barbijo decis paga quiero]	Insults between readers
Topic 03	[gente pandemia china argentina salir cuarentena argentinos muertos gusta gobierno paises virus digo infectados hablar]	Testing, cases and deaths at the local level
Topic 04	[globo pobre vieja jajajaja arroyo grande fideos mujer cosa sueldo albertitere miedo patria queda dictadura]	Miscelaneous
Topic 05	[larreta peronista inutil seguro deja viejo cabeza verguenza kk ojo peronistas veo basura pasa idea]	Macrism, antiperonism
Topic 06	[che cuarentena ja anos muertos fernandez titere pena argentina vas mira favor hijo economia delincuentes]	Government and the economic policy
Topic 07	[gobierno alverso dolar virus cientificos coronavirus presidente inutiles covid leer default bolsonaro meses inflacion pobres]	Insults to the government, economic crisis
Topic 08	[macri jajaja nota comentario ladrones chorros cree voto diario clarin jaja sale alguien cerebro llama]	Macrism, antiperonism
Topic 09	[cara medicos votaron cubanos falta politicos deuda aca provincia mundo comer odio conurbano espero perdon]	Insults to the government, economical crisis
Topic 10	[millones gente pagar cuarentena impuestos casa gobierno personas presos trabajar anda anos ojala trabajo plan]	Public administration

Topics 07 and 09 were merged as “Insults to the government, economical crisis” as well as topics 05 and 08 which were labeled as “Anti-peronist macrism”. We finally got 8 topics which can be plotted during time:

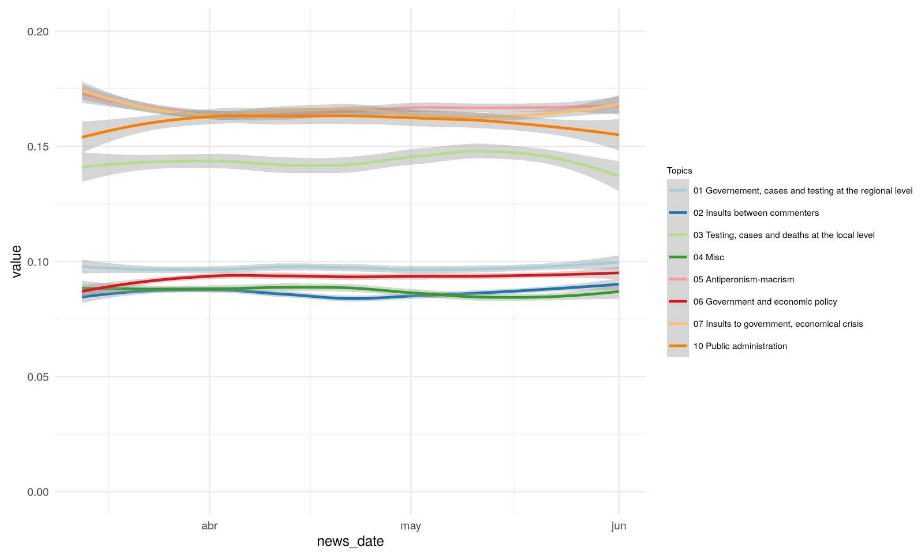

Fig. 2. Time evolution of topic composition (mean of each topic) in reader's comments on COVID-19 news

The antiperonist-macrism debate along with the insults towards the government and economical crisis are the two main topics in the distribution. The debate regarding public administration is a close third main topic: for a while between april and may, it reaches the means of the first two topics as well, but it begins a decreasing tendency afterwards. It seems like an argued debate gained strength in the first weeks of the quarantine and until the midst of it, but as we are reaching more recent months people began to drop it. Other relevant topic is the one that talks about testing, cases and deaths at the local level.

In a preliminary model we made, we kept in the input dataframe the duplicated comments. The main topic turned out to be the critics towards the government and there was a topic which represented the trolls, bots and spammers. The fact that the removal of these comments makes those topics disappear can indicate that this is a preliminary way of controlling trolls (which generally copy and paste the same message multiple times) and analyzing comments of people.

Table 2. Illustrative comments (random sample between comments with high prevalence of topics 05 and 08)

Topic order	Text of sampled comment
Topic 05 - Macrism, antiperonism	- LA CUARENTENA SIGUE HASTA EL 2023… PARA ALBERTITO, EL CORONAVIRUS ES SU TABLA DE SALVACIÓN...
	- EL LAVADO DE K ULO, ES PARA LA FASE 3 !! OK?
	- El verdadero virus que destruyó Argentina es el peronismo. Cristina, Alverso, el General, Menem,Duhalde, Eva. Todos lo mismo. El día que logremos una vacuna efectiva contra el peronismo, ese dia va a arrancar el pais
	- Como es posible que la oposicion se prestara a este bochorno¡ lamentable R Larreta!!!
	- los gobierno no..... los peronistas.... no generalices....
Topic 08 - Macrism, antiperonism	- SI ESTUVIERA MAURICIO....LAS INSTITUCIONES FUNCIONARIAN !!!!!
	- Mauricio,¿Qién es ese individuo?
	- Naaaa ya estabamos quebrados, Macri lo hizo
	- NO SOLO ORKO CEREBRO DE AMEBA,..SINO QUE UN REVENRENDO PE-LO-TU-2,.ESTAMOS INFECTADOS DE ESTE VIRUS KAKA
	- Son K el ADN de chorros no se los quita nadie

As can be seen in the previous table, both topics seem to capture the so-called “grieta” in argentina. They contain positive and (mostly) negative positions and images in relation to current administration, peronism in general, and the previous national government.

As a general appreciation, the evolution of the complete topic composition seems to be rather stable during the period.

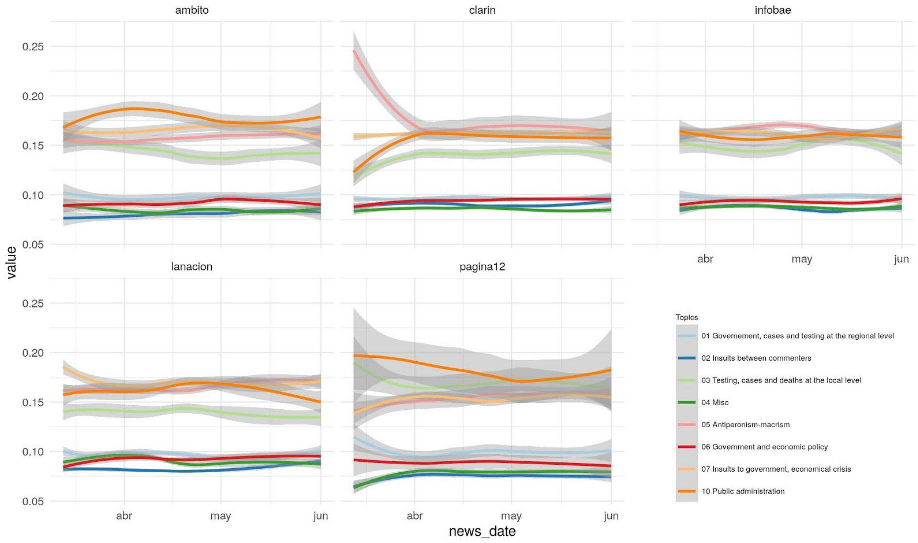

Fig. 3. Time evolution of topic composition (mean of each topic) in reader's comments on COVID-19 news by newspaper

When we plot the same evolution but conditioning over newspaper, we see that there aren't many differences amongst what the readers of each site talk about. In most of the cases, the topic regarding public administration is the highest, except in Clarin where the antiperonist-macrism is. In general, we see that there aren't substantial differences within the topic distribution of each media. This opens the question to whether the commenters (which may have a distinct profile from readers) of newspapers with different political ideologies have different profiles or not. In terms of what they talk about when reading news, we can offer a preliminary answer: no, they don't.

3.1 Predicting majority topic with FastText

Since the topic detection seemed to be rather unstable (we tried it in different periods of time and the topics changed) we tried to convert the topic detection into a supervised problem. The idea is to predict the main topic of a determined comment. We present in this section the first approximation to this problem.

After estimating the LDA model, we detected the most relevant topic in each comment and we used that topic in order to tag it. We only tagged the comments if the majority topic were to have a prevalence higher than 0.3, this is to say, we tagged only those comments which presented a clearly distinctive category.

We applied a FastText classification to predict the tag. FastText [5] [6] [12] makes the vectorial representation of each word in a vocabulary, taking into account the morphology by considering subword units and representing words by a sum of its character n-grams.

Our training set was constructed using a random sample of comments per day from our main database, and the test set was a result of the remaining comments: for each day in the analyzed period we use a 30% as a test set and the 70% remaining, as a training set. Our final results from the test set were validated by precision and recall metrics.

Since the class distribution of the dataset was not severely imbalanced, we tried a classifier with no over or undersampling. After performing an auto tuning process of the classifier, the final embedding produced has 100 dimensions.

Table 3. Performance metrics of fasttext algorithm in majority topic prediction

Labels	Precision	Recall	F1	n
Government, cases and testing at the regional level	0.90	0.89	0.89	5733
Insults between readers	0.89	0.87	0.88	3450
Testing, cases and deaths at the local level	0.92	0.93	0.93	16537
Miscelaneous	0.90	0.87	0.88	3792
Macrism, antiperonism	0.82	0.85	0.84	2795
Government and the economic policy	0.89	0.89	0.89	4991
Insults to the government, economic crisis	0.90	0.88	0.89	3649
Macrism, antiperonism	0.90	0.81	0.85	3488
Insults to the government, economical crisis	0.89	0.84	0.87	2725
Public administration	0.93	0.95	0.94	22316
Total (weighted)	0.91	0.91	0.91	69476

The performance of the classifier seems promising. It will also provide the word embedding which could be used to perform more in-depth analysis about the semantic structure of these comments.

4 Conclusions and future work

We have presented the first results of an ongoing investigation about the possibilities that news reader’s comments and Natural Language Processing techniques have as a tool for studying social representations. We use the COVID-19 theme as a test case, but one of the main goals of the research is to develop a tool and a workflow useful for other themes or situations.

Using labelled links to digital news articles, provided by GDELT, we were able to construct a dataset of comments framed in the COVID-19 discussion. LDA topic modeling allowed us to detect some relevant themes of discussion in this frame. Topics regarding critics, negative views, insults to the government and the economic crisis were particularly relevant. Discussion of public administration and the problem of testing cases were also a relevant theme in news comments. But one of the main findings was the existence of two topics which capture the so-called “grieta” in argentinian political discussion: the peronism-anti peronism axis.

We have also trained a preliminary fasttext model to predict the majority topic of each comment with good first results. It would be possible to use word-embedding representation to capture other relevant semantic dimensions of the corpus analyzed. At the methodological level some of the future steps are

- restate the problem as a multilabel classification problem: train the model to predict the three more relevant topics
- perform a deeper parameter search
- test most advanced NLP frameworks such as BERT in order to see if it is possible to achieve a higher accuracy in the topic prediction

At a theoretical level, several research questions arise: how stable is the existence of the peronism-anti peronism topics? Is it possible to observe it in comments framed in another discussion? These questions are to be addressed in the near future.

References

1. Adorno, T., Frenkel-Brunswik, E., Levinson, D., Sanford, N.: Studies in authoritarian personality. 1st edn. New York: Harper & Row (1950).
2. Blei, D., Ng A.Y., Jordan, M.: Latent Dirichlet Allocation. The Journal of Machine Learning (3), 993-1022 (2003).
3. Calvo, E., Aruguete, N.: #Tarifazo. Medios tradicionales y fusión de agenda en redes sociales. Inmediaciones de la Comunicación (13), 189-213 , (2018).
4. Gramsci, A.: Notas sobre Maquiavelo, sobre la política y sobre el estado moderno. 1st edn. Buenos Aires: Nueva Visión, Buenos Aires (1972).
5. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of Tricks for Efficient Text Classification. arXiv:1607.01759 (2016).
6. Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., Mikolov, T.: FastText.zip: Compressing text classification models. arXiv:1612.03651 (2016).
7. Marx, K., Engels. F.: The German Ideology. 1st edn. International Publishers, New York (2004).
8. Moscovici, S.: El psicoanálisis, su imagen y su público. 1st edn. Huemul, Buenos Aires (1979).
9. Rudé, G.: Ideology and Popular Protest. 1st edn. Lawrence & Wishart, Great Britain (1980).
10. Schuth, A., Marx, M., Rijke, M.: Extracting the Discussion Structure in Comments on News-Articles. 9th ACM International Workshop on Web Information and Data Management, pp. 97-104. Lisbon, Portugal (2007).
11. Wagner, W., Duveen, G., Farr, R., Jovchelovitch, S., Lorenzi-Cioldi, F., Markova, I., Rose, I.: Theory and method of social representations. Asian Journal of Social Psychology (2), 95–125 (1999).
12. Xu, J., Du, Q.: A Deep Investigation into fastText. IEEE 21st International Conference on High Performance Computing and Communications, pp. 1714-1719. Zhangjiajie, China (2019).