Evaluating the cohesion of municipalities' discourse during the COVID-19 pandemic

Victor Antonio Menuzzo[1], **André Santanchè**[1], **Luiz Gomes-Jr**[2]

[1] Instituto de Computação – UNICAMP – Campinas – SP – Brazil

[2]Departamento de Informática – UTFPR – Curitiba – PR – Brazil

v234996@dac.unicamp.br, santanch@ic.unicamp.br, lcjunior@utfpr.edu.br

Abstract. *Social media has been used as a method to alert and raise awareness among the population to help fight the COVID-19 pandemic. We argue that the discourse of municipalities and their respective mayors may have an influence on the behavior of the population and thus directly impact COVID-19 outcomes. This paper analyzes the diversity and cohesion of these discourses through posts published on Facebook, evaluating (i) diversity of topics discussed, (ii) topic evolution, and (iii) deviation from a central discourse. We also combine this information with epidemiological data to assess impact in the outcomes. In particular, we present two different Latent Dirichlet allocation (LDA) models to analyze how topics are being discussed by municipalities/mayors and compare how cohesion is related to the evolution of the pandemic. Our initial analysis suggests that municipalities tend to employ a unified discourse as a response to the worsening of epidemic outcomes. The results of our study could help to inform governments of better communication strategies in this and future health crisis.*

1. Introduction

Different types of communication are used nowadays, but with the growth of social networks, written communication over these platforms posts has become increasingly popular. Therefore, it is important to assess the patterns in these online tools, especially regarding government communication. Linguists have been studying text analysis for years, seeking to understand the relationships between language and the context of human societies [Gill 2000]. From these studies, a sub-area has gained considerable prominence: discourse analysis [Reed 1997, Gill 2000, Widdowson 2007].

The term discourse refers to the meaning the writer wants to express with that text and what meaning the text will have for the reader [Widdowson 2007]. It focuses on the ways writers use language to achieve certain goals, for example, handling conflicts and/or building trust [Fairclough 2003].

In this paper, we analyze the discourse of municipalities/mayors in relation to the pandemic, adopting the topic modeling approach for discourse analysis. We developed two *LDA* models to assess the evolution of the topics discussed (Section 4.1). Different metrics were used to understand the diversity of topics over time (Section 4.2). A metric based on the distance from a central discourse was also developed to identify cities whose discourse deviate from the general trend. Among our main results, a negative correlation was observed between this distance and number of deaths by COVID-19, suggesting an interplay between communication and epidemic outcomes.

2. Related Work

One approach to assess the meaning of different types of discourse is based on the analysis of the topics in the text. Topic modeling was initially approached by [Deerwester et al. 1990], where a term-document matrix decomposition strategy was proposed with the goal of discovering the different subjects covered in the document through the co-occurrence of words [O'callaghan et al. 2015].

LDA (Latent Dirichlet allocation) is one of the most used techniques to perform topic modeling. It is widely applied in the fields of medicine [Liu et al. 2020, Goyal and Gomeni 2013, Kandula et al. 2011] and social network analysis [Li et al. 2015, Shahbazi and Byun 2020]. LDA is a generative probabilistic model, in which documents are represented as random mixtures of latent topics, considering that each topic is characterized by a distribution of words [Blei et al. 2003].

Applying topic modeling to analyze discourse cohesion has been done in several works. Rashed et al. [Rashed et al. 2019] used the strategy to understand the behavior of extremist communities in the darknet. The paper identified that cohesion between members grows over time. They also used LDA to analyze the topics with greater cohesion through a complex network model. The analyses identified the most central nodes of the extremist communities, thus making it possible to determine the network influencers.

To analyze different positions of political speeches, Zirn and Stuckenschmidt [Zirn and Stuckenschmidt 2014] used LDA to generate results from the transcribed speeches. With the applied method, it was possible to determine topics related to policy areas within different discourses. As their main contribution, the authors proposed the analysis of documents that contain more than one topic in a fully automatic way.

Another line of research focuses on how cohesion impacts in pandemic cases such as COVID-19. Jewett et al. [Jewett et al. 2021] analyzed that the more cohesive and resilient a society is, the faster its recovery after a disaster. Furthermore, during these periods, social inequalities are more revealed, highlighting the importance of prioritizing investments for marginalized groups [Fløttum 2010].

3. Data sources and preprocessing

For developing topic modeling and obtaining the analyzes described above we used Facebook posts from municipalities' official pages and those of their respective mayors. Socioeconomic and data related to COVID-19 were also used to perform analyzes on the discourses obtained. We extracted data from: (i) **Facebook posts** - The 26 Brazilian capitals were selected for the study; data were collected from January 2020 to April 2021 using the CrowdTrangle[1], in which only verified accounts were included; (ii) **COVID-19 reports**: data on number of COVID-19 news cases, new deaths, and vaccines were obtained from Brasil.io/datasus[2]; (iii) **Socioeconomic indicators**: data on the Brazilian population were obtained through IBGE[3].

Two different corpora were developed: one using only posts related to COVID-19 and the other with all posts published by the selected pages. To filter whether a post

[1]https://www.crowdtangle.com

[2]https://ourworldindata.org

[3]https://www.ibge.gov.br/acesso-informacao/dados-abertos.html

is related to COVID-19 or not, we employed the approach from Melo and Figueiredo [de Melo and Figueiredo 2020], where the words commonly used in this topic are described. A total of 58,709 posts were obtained, 22,763 of them being related to COVID-19 and 35,946 not related. A preprocessing of the text for each post was performed, with all links, city names and stop-words removed. We also applied spacy lemmatization to reduce each word to its base form.

4. Data analysis

We applied diverse techniques to assess the cohesion and diversity of topics during the pandemic. Our hypothesis is that municipalities/mayors, when deviating from a standard discourse, could worsen the outcomes of the pandemic.

4.1. Topic evolution analysis

To analyze the topics discussed throughout the pandemic, we developed two LDA models. The first used only posts related to COVID-19; the second uses all posts. The first point to consider when creating the models was to find the optimum number of topics. To define this we measure the coherence score [Stoica et al. 2005], assessing the quality of the topics learned. Five topics were detected for the model that uses only posts related to COVID-19 and 10 for the model that uses all posts.

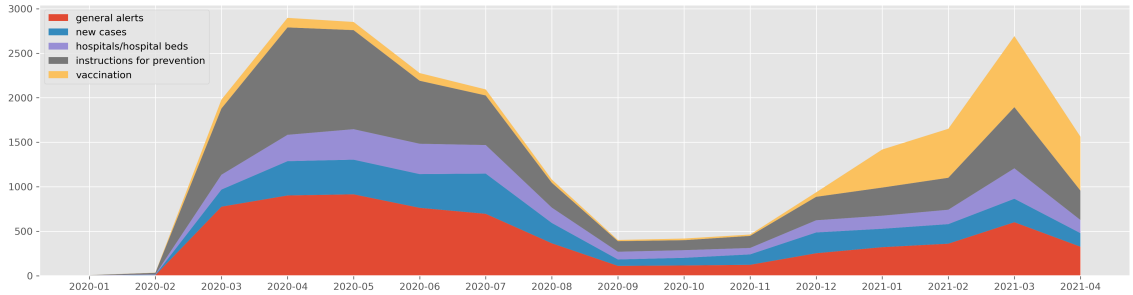


Figure 1. Topics - Model only COVID-19 posts (Number of posts by year/month)

The COVID-19 posts model fitted the five topics, which we then manually classified as: general alerts (warnings from municipalities about measures taken to combat the pandemic); new cases (information related to new cases in the city); hospitals/hospital beds (information about the status of hospitals); instructions for prevention (instructions how to wear a mask/stay at home); and vaccination (posts about vaccination in the city). These topics are shown in Figure 1.

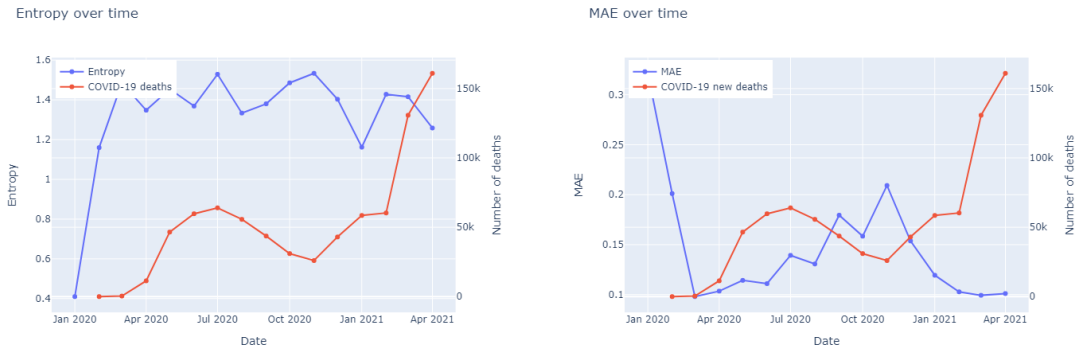
Figure 1 shows the evolution of the number of posts referring to each topic during the analyzed period. We can see that during the initial phase of the pandemic, the discourse was focused on alerting and giving instructions to the population. During the pandemic, a decrease in the publication of posts related to COVID-19 can also be seen. Finally, in the second peak, where there is large emphasis on vaccination.

Analyzing the model with all posts (not detailed here for brevity), the influence of topics not related to COVID-19 is evident. The distribution of topics remains constant over time in almost all topics, but the most discussed topics are not related to COVID-19 for most of the period. Posts related to construction in the city, education, municipality services, etc. have a larger influence in the corpus.

4.2. Discourse diversity and cohesion

To quantify the evolution of the discourse throughout the pandemic (using the COVID-19 posts model), we employed two main approaches. First, to represent the diversity of the topics in the posts, we calculated the entropy of the monthly topic distribution. Figure 2a shows the evolution of the monthly entropy. The general downward trend indicates that the distribution of topics tends to become less uncertain/diverse over time. This can be an indication that the discourse tends to converge, probably as a result of governments taking the issues more seriously and also reflecting the scientific consensus as the disease is better understood.

To assess the cohesion of the discourse, i.e. how much the municipalities deviate from a central discourse, we calculated the mean absolute error of the distribution of topics for each municipality against the distribution of topics for each month. Figure 2b shows the average error over time. It is interesting to note an inverse correlation between the distance and deaths. This suggests that governments tent to unify the discourse only during epidemic peaks, possibly missing the opportunity to better inform the population during less critical times.



(a) Entropy COVID-19 posts model **(b) MAE COVID-19 posts model**

Figure 2. Entropy and MAE COVID-19 posts model

Analyzing which municipalities are carrying out the most cohesive discourse is another central point to understand. To measure how much a municipality deviates from a central discourse, we measure the weighted distance between its monthly topic distribution and the overall topic distribution for that month. We use exponentially decreasing weights to give more emphasis to less important topics – i.e. marginal topics have greater influence over the distance, making municipalities posting about them stand out as outliers. Municipalities with shorter distances are considered more coherent with the central discourse. Table 1 and Table 2 show the average distances for the top and bottom municipalities according to the metric. However, in the model with only COVID-19 posts, we can distinguish both the capitals that are using a cohesive discourse and those that are not.

Table 1. Top 5 distances

City	COVID-19 model distance
Palmas	0.362
Manaus	0.364
Florianópis	0.373
Curitiba	0.376
Salvador	0.379

Table 2. Bottom 5 distances

City	COVID-19 model distance
Vitória	0.589
Macapá	0.602
Fortaleza	0.607
Teresina	0.617
Rio branco	0.786

Finally, to analyze whether the discourses influence epidemic outcomes and/or vice versa, we calculated the correlations between the distances for each municipality and their epidemic outcomes (total cases and total deaths, population normalized). To capture possible temporal influences, we shifted the outcomes (cases/deaths) in +1 and -1 month in relation to distance. When using shift +1 we are assuming that COVID-19 outcomes influence the cohesion of the discourse. When using shift -1 we are assuming that the cohesion of the discourse influences the COVID-19 outcomes. High correlations [Correlation: -0.43, p-value: 0.12] between deaths and distance were obtained using shift +1. This result suggests that as deaths increase, distances tend to decrease, i.e., cities use a more coherent discourse as the pandemic worsens. The small sample (26 cities) makes it hard to obtain more statistically significant correlations, which we plan to address by expanding the number of municipalities covered.

5. Conclusion and future work

This article explores the potential of applying a topic modeling approach to analyze the evolution of discourse, taking as a scenario the municipalities/mayors during the pandemic.

Through this technique, we were able to: (i) find results suggesting that the discourses of the municipalities tend to become less diverse over time, getting closer and closer to a central discourse; (ii) find a significant negative correlation between deaths and discourse distance, indicating that the speech tends to become more cohesive in response to the worsening of the pandemic. We also proposed a distance metric to identify possible cities that are further away from this central discourse.

Our future work focuses on performing regression analysis using the distance variables and COVID-19 reporting. It would also be interesting to increase the number of cities covered to achieve greater statistical significance.

Acknowledgments

- This study was financed in part by CNPq grant number 428459/2018-8.
- Data from CrowdTangle, a public insights tool owned and operated by Facebook.

References

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *JMLR*, 3:993–1022.

de Melo, T. and Figueiredo, C. M. (2020). A first public dataset from brazilian twitter and news on covid-19 in portuguese. *Data in brief*, 32:106179.

Deerwester, S. et al. (1990). Indexing by latent semantic analysis. *JASIST*, 41(6):391–407.

Fairclough, N. (2003). *Analysing discourse: Textual analysis for social research*. Psychology Press.

Fløttum, K. (2010). A linguistic and discursive view on climate change discourse. *ASp. la revue du GERAS*, (58):19–37.

Gill, R. (2000). Discourse analysis. *Qualitative researching with text, image and sound*, 1:172–190.

Goyal, N. and Gomeni, R. (2013). A latent variable approach in simultaneous modeling of longitudinal and dropout data in schizophrenia trials. *ECNP*, 23(11):1570–1576.

Jewett, R. L., Mah, S. M., Howell, N., and Larsen, M. M. (2021). Social cohesion and community resilience during covid-19 and pandemics: A rapid scoping review to inform the united nations research roadmap for covid-19 recovery. *IJHS*, page 0020731421997092.

Kandula, S., Curtis, D., Hill, B., and Zeng-Treitler, Q. (2011). Use of topic modeling for recommending relevant education material to diabetic patients. volume 2011, page 674. AMIA.

Li, A. et al. (2015). Attitudes towards suicide attempts broadcast on social media: an exploratory study of chinese microblogs. *PeerJ*, 3:e1209.

Liu, Q. et al. (2020). Health communication through news media during the early stage of the covid-19 outbreak in china: digital topic modeling approach. *JMIR*, 22(4):e19118.

O'callaghan, D. et al. (2015). An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, 42(13):5645–5657.

Rashed, M., Piorkowski, J., and McCulloh, I. (2019). Evaluation of extremist cohesion in a darknet forum using ergm and lda. ASONAM '19, page 899–902, New York, NY, USA. ACM.

Reed, J. T. (1997). Discourse analysis. *A handbook to the exegesis of the New Testament*, pages 189–218.

Shahbazi, Z. and Byun, Y.-C. (2020). Analysis of domain-independent unsupervised text segmentation using lda topic modeling over social media contents. *Int. J. Adv. Sci. Technol*, 29:5993–6014.

Stoica, P., Moses, R. L., et al. (2005). Spectral analysis of signals.

Widdowson, H. G. (2007). *Discourse analysis*, volume 133. Oxford University Press Oxford.

Zirn, C. and Stuckenschmidt, H. (2014). Multidimensional topic analysis in political texts. *DKE*, 90:38–53.