Universidade de São Paulo

Faculdade de Filosofia, Letras e Ciências Humanas

Departamento de Ciência Política

FLS 6497 -- Ciência de Dados e Aprendizado de Máquina em Ciência Política

Final Assignment

Twitter and Vaccinations: Natural Language Processing using BERT

Isabel Seelaender Costa Rosa ;

Pedro Henrique de Santana Schmalz ;

Luiz Guilherme Roth Cantarelli .

São Paulo

2022

Acknowledgements

We thank Lorena Barberia and her team for their collaborative efforts in the construction of the dataset employed in this analysis. This work would not have been possible without the great efforts of Rebeca de Jesus Carvalho, Gustavo Fernandes de Paula and André Garibe.

I. Introduction

The Covid-19 vaccination has been a relevant topic of debate amongst political elites and the general public since the pandemic's onset still in 2020[1]. Even though immunizers for the infection caused by the SARS-CoV-2 virus became available in Brazil only in 2021, the theme was broadly discussed still during the first pandemic year. In this context, candidates and officeholders used social media to express and publicize their positioning regarding the issues of vaccines and vaccination.

This paper presents a study of the debate, amongst brazilian political elites, surrounding the theme of Covid-19 vaccination in the country. For this purpose, we collect and analyze Twitter posts about vaccination for SARS-CoV-2, from mayoral candidates, elected mayors and elected governors from 2020 and 2021. We then use the deep learning Bidirectional Encoder Representations from Transformers (BERT) algorithm to categorize tweets according to subject and positions expressed by subjects regarding vaccines and vaccination for Covid-19.

The BERT algorithm is a pre-trained, open-sourced, Natural Language Processing (NPL) tool, developed in 2018, by researchers at Google AI Language. Its functionalities include a varied list of language/NLP tasks, ranging from sentiment analysis to text prediction and generation. Here, we employ BERT for sentiment analysis, in an effort to identify and categorize tweets by Brazilian political elites according to their positioning regarding Covid-19 vaccines and vaccination. Hence, we attempt to better understand and differentiate positive, neutral and negative positionings from subjects.

As argued by Barberá et al. (2020), text analysis became increasingly relevant for a number of research questions in social sciences. In their paper "*Automated Text Classification of News Articles: A Practical Guide*", authors emphasize significant decreases in time and cost of text analysis since the advent of automated text classification methods. Combined with the widespread reach of digital text archives these aspects have led to an "(...) explosion in the extent and scope of textual analysis.". (BARBERÁ et al., 2020).

In the following sections we present an in-depth description of our efforts to analyze positionings regarding Covid-19 vaccines and vaccination, from Brazilian political elites,

[1] The World Health Organization (WHO) declared the infection caused by the SARS-CoV-2 virus a pandemic on March 11th, 2020.

through the BERT algorithm. Section II is devoted to explaining and detailing the collection methods employed in the construction of our dataset containing Twitter posts from 2020 and 2021 for different players. The following section presents our analysis of tweets using the BERT algorithm and presents some partial results.

II. Selecting the Corpus and Manual Analysis: Twitter and Vaccination Dataset

The *corpus* selected for this paper includes all Twitter posts surrounding the themes of Covid-19 vaccines and vaccination from Brazilian mayoral candidates, elected mayors and governors during the period between 2020 and 2021. Firstly, based on data registered in the TSE (Superior Electoral Court), a list including all candidates running for mayoral positions in the 2020 elections that took place along all cities in Brazil, was used to identify active Twitter accounts used by these candidates within the researched period. From an initial sample including 147 candidates, a total of 114 candidates were identified – active profiles for the remaining 33 were not found[2]. From the remaining 114 candidates from the 12 Twitter accounts identified did not publish in the analyzed year on topics of interest to the research.

Afterwards 18 municipalities were selected in all five regions of Brazil in an effort to ensure regional representativeness. The selected cities were: Barretos, Belém, Belo Horizonte, Campinas, Campo Grande, Cuiabá, Curitiba, Fortaleza, Manaus, Niterói, Pelotas, Porto Alegre, Ribeirão Preto, Rio de Janeiro, Salvador, São Caetano do Sul, São José do Rio Preto e São Paulo. Table 1 presents the selected municipalities across the five Brazilian regions and summarizes the number of tweets for all candidates.

Table 1. Number of Vaccine and Vaccination Tweets by Municipality

Municipality	State	Region	# of Tweets	Tweets (%)
Barretos	São Paulo	Southeast	1	0,05%
Belém	Pará	North	31	1,43%
Belo Horizonte	Minas Gerais	Southeast	154	7,11%
Campinas	São Paulo	Southeast	39	1,80%
Campo Grande	Mato Grosso do Sul	Center-West	17	0,78%

[2] The determination of Twitter accounts' statuses was performed manually by coders and considered criteria such as the existence of a personal or professional candidate profiles, as well as, the existence of any activity from these subjects.

Cuiabá	Mato Grosso	Center-West	13	0,60%
Curitiba	Paraná	South	384	17,73%
Fortaleza	Ceará	Northeast	116	5,36%
Manaus	Amazonas	North	53	2,45%
Niterói	Rio de Janeiro	Southeast	120	5,54%
Pelotas	Rio Grande do Sul	South	13	0,60%
Porto Alegre	Rio Grande do Sul	South	366	16,90%
Ribeirão Preto	São Paulo	Southeast	26	1,20%
Rio de Janeiro	Rio de Janeiro	Southeast	208	9,60%
Salvador	Bahia	Northeast	201	9,28%
São Caetano do Sul	São Paulo	Southeast	2	0,09%
São José do Rio Preto	São Paulo	Southeast	5	0,23%
São Paulo	São Paulo	Southeast	417	19,25%
TOTAL	-	-	2166	100,00%

From all active Twitter profiles, an API (Application Programming Interface) was used to collect 151,512 posts. After the first collection round, TSE data was used to expand the number of subjects in the sample and the timeframe, by selecting Twitter profiles of mayors elected in 2020 and governors in office for both 2020 and 2021. The resulting sample consisted of 175 profiles from 147 mayors and 28 governors in the same 18 pre-selected municipalities, for the period encompassing the years 2020 and 2021.

II.I Keywords

For this analysis we adopt a keyword search as our main strategy for selecting a *corpus*. According to Barberá et al. (2020), this method is preferable in comparison to other approaches such as employing subject categories, mainly due to the fact that it allows researchers to maintain control over the breadth of the search. Additionally, authors emphasize other relevant advantages in the keyword search method, such as easy replicability and transportation across alternative or additional universes of documents. The latter is especially important in the case of our analysis, as we expand our analysis to include elected mayors and governors.

In spite of the multiple advantages of this approach, there are also some challenges specific to keyword search. In particular, the selection of keywords might be biased or inappropriate, and relevant terms might be susceptible to changes over time. Moreover, there are also practical problems involving the selection of keywords such as the existence of synonyms, spelling mistakes and linguistic variations of similar terms. As it will become clearer throughout this section, our selection was based on a number of different criteria, considering these particular constraints.

The resulting set of keywords is a product of four test-trials, and has been built based on human observation of spelling alternatives, term frequency and employment. A list containing the keywords was used to select Tweets by candidates that referred - directly or indirectly - to SARS-CoV-2 vaccination. Table 2 summarizes the key terms employed in the filtering process that was performed using R software. The filter considered both upper and lower cases. Orthographic and spelling issues were included after a preliminary analysis regarding common variations adopted by Twitter users, preventing the loss of information due to simple issues. Some tweets that were selected were discarded once it was determined that the content while citing vaccines or a specific vaccine platform did not, in fact, refer to the topic of COVID-19 vaccination.

Table 2. Terms Employed to Filter Candidate Twitter Posts

SEARCH TERMS	KEY-WORDS
Vaccines and vaccination – general references	"vacin", "vassina", "vasina", "imunizacao", "imunisacao", "vachina", "vaxina", símbolos de vacina
COVID-19 Vaccines and **Laboratories**	"coronavac", "comunavac","sinovac", "astrazenica", "astrazeneca","oxford", "oxfort", “Oxfor”,"vaxzevria","pfizer", "pfeizer", "pfaizer", "faizer","feizer", "biontech", "biontec", "comyrnaty", "comirnaty", "biontech manufacturing gmbh", "jansen", "janssen", "j&j", "johnson & johnson","jhonson & jhonson","jonson", "j & j","johnson", "johnsons", "jhonson", "jancen", "Ad26.COV2.S", "vacina moderna","vacina da moderna", "spikevax","moderna biontech", "mRNA-1273", "CX-024414","sputnik","sputinik","sputink", "sputinic", "sputinikV", "gamaleya", "gamaleia", "covaxin", "covachin", "bharat biotech", "novavax", "covavax", "Nuvaxovid", "NVX-CoV2373", "TAK-019", "SARS-CoV-2 rS with Matrix-M1 adjuvant","Serum Institute of India", "Novavax

	Formulation","sinopharm", "BIBP", "Sinofarm", "butantan", "butanta","fiocruz", "@fiocruz", "fiocrus", "fiocruz"
Preventive Treatments – pharmacological treatments other than COVID-19 vaccines	"tratamento precoce", "tratamento precosse","tratamento preventivo", "cloroquina", "hidroxicloroquina", "hidroxocloroquina", "hidrosicloroquina", "hidrocicloroquina", "idroxicloroquina", "cloroquinha", "spray nasal", "spray nazal", "ivermectina", "invermectina"
Localities	"virus chines", "vacina da china", "vacina chinesa", "virus chines", "vacina britanica", "vacina cubana", "vacina da rússia", "vacina russa"
Others	“doriavac", "comunavirus"

II.II Producing a Training Dataset.

Following data collection and filtering processes, we devoted ourselves to a fundamental task for supervised machine learning processes: the development of a training dataset, in which we coded vaccine tweets manually. Criteria to be followed by coders was defined collectively after a first round of exploration. For this stage we had 5 individual coders, working simultaneously to code and revise each other's work.

As postulated by Barberá et al. (2020), the manual coding of tweets is devoted to the estimation of a “(...) model of sentiment [Y] as labeled by humans as a function of the text objects, the features of which compose the independent variables.”. (BARBERÁ et al., 2020). Our unit of analysis corresponds to an individual tweet[3].

The choice for coding segments rather than sentences was appropriate, especially considering the content format, that is, the tweet as a coherent *corpus* of text, dedicated to communicating ideas through a set of phrases. In this sense, exploring stages similarly suggested that understanding was improved by considering all the text incorporated in posts, rather than analyzing separate sentences. The resulting dataset was subsequently employed to train the BERT algorithm. Next sections are devoted to the description of our analysis and the main results obtained.

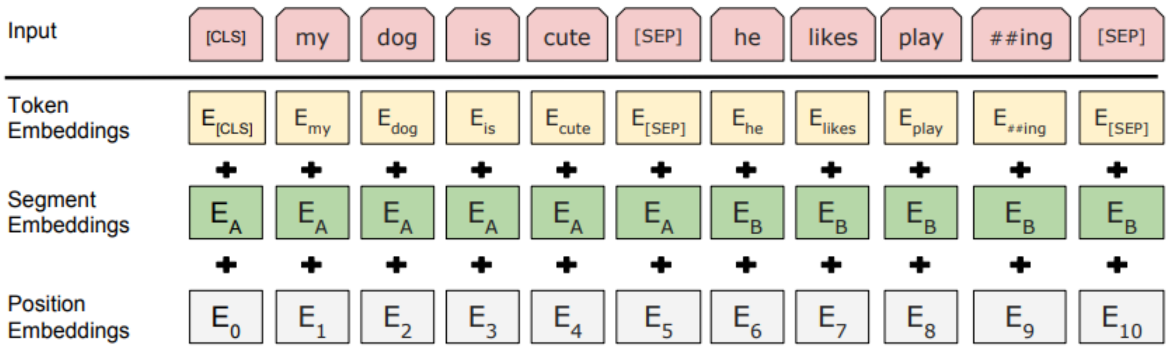
[3] Importantly, video content and pictures/images were not considered during coding due to the increased difficulty this type of content analysis imposes for any SML processes.

III. Using the BERT Algorithm to Predict Sentiment: Analysis and Main Results

BERT (Bidirectional Encoder Representations from Transformers) is a deep learning model developed by Devlin et al. (2019), and trained - on a large dataset of text - to identify the meaning and context of words and sentences. It can be used for a variety of natural language processing tasks. Substantially, the idea behind BERT is to train a model to be able to understand the context and meaning of words in a sentence, by considering preceding and succeeding words, a process which characterizes BERT as a ‘bidirectional’ model.

Specifically, BERT uses a type of neural network denominated *a transformer*, which is able to process input sequences efficiently and concomitantly. The transformer employs self-attention mechanisms to weight the importance of different parts of the input sequence, allowing it to focus on the most relevant parts of the input while processing it. It is important to notice that the model does not process text in its natural form, requiring text conversion into tokens. Figure 1 shows how each phrase is translated into an input for BERT to process it.

Figure 1 - BERT input representation.



Source: Devlin et al. (2019).

Moreover, BERT uses a particular type of tokenization called WordPiece tokenization. This method consists of dividing words into subwords based on frequency, so that the most common subwords are given their own tokens and the less common subwords are combined into a single token. For example, the word "university" will be tokenized as follows:

["univ", "##ers", "##ity"],

with "##" indicating that the following characters are part of a subword. This allows the model to recognize common word components and avoid having to treat each word as a completely separate entity.

In addition to WordPiece tokenization (Wu et al., 2016), BERT also adds special tokens to the input to indicate the beginning and the end of a sentence, and to separate different sentences within a document. These special tokens are important for helping the model to understand the structure and context of the input text. Overall, the tokenization process is an important step for preparing text imputed to a BERT model, as it allows the model to effectively process the input and to recognize text meaning. Important to notice, in this analysis, a pre-trained BERT model for Brazilian Portuguese, *BERTimbau*, was used (Souza, Nogueira and Lotufo, 2020), for better performance considering our database. To ensure results were not random, a K-fold Cross-Validation was ensued, with 10 folds.

Finally, a stratified sampler was put in place to ensure each fold/epoch would have sufficient observations for each group, as analysis's classes presented meaningful balance issues: the vast majority of tweets in the dataset are favourable tweets (4007), followed by neutral ones (1621) and with a minority of 263 unfavourable (negative) tweets. Figure 2 illustrates the distribution of the categories.

Figure 2. Distribution of the Classes

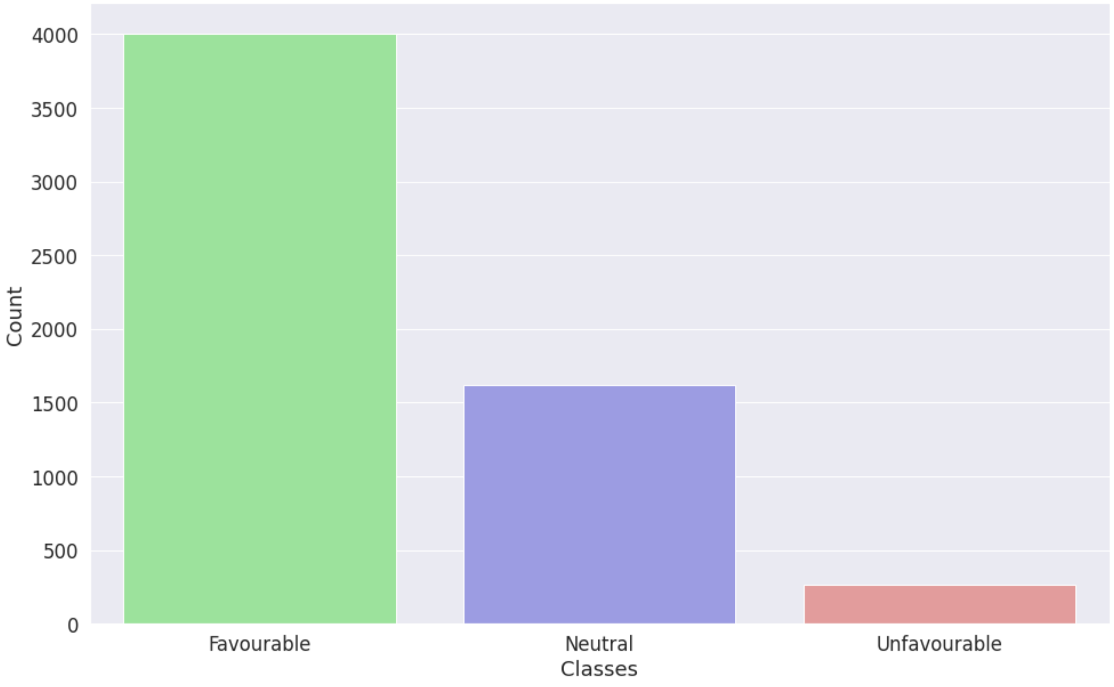


Table 3 shows the results of the classification by class. The overall results were as follows: 1) Training Loss - 0.14; 2) Validation Loss - 0.83; 3) Validation Accuracy - 0.77; 4) F1 Score - 0.77.

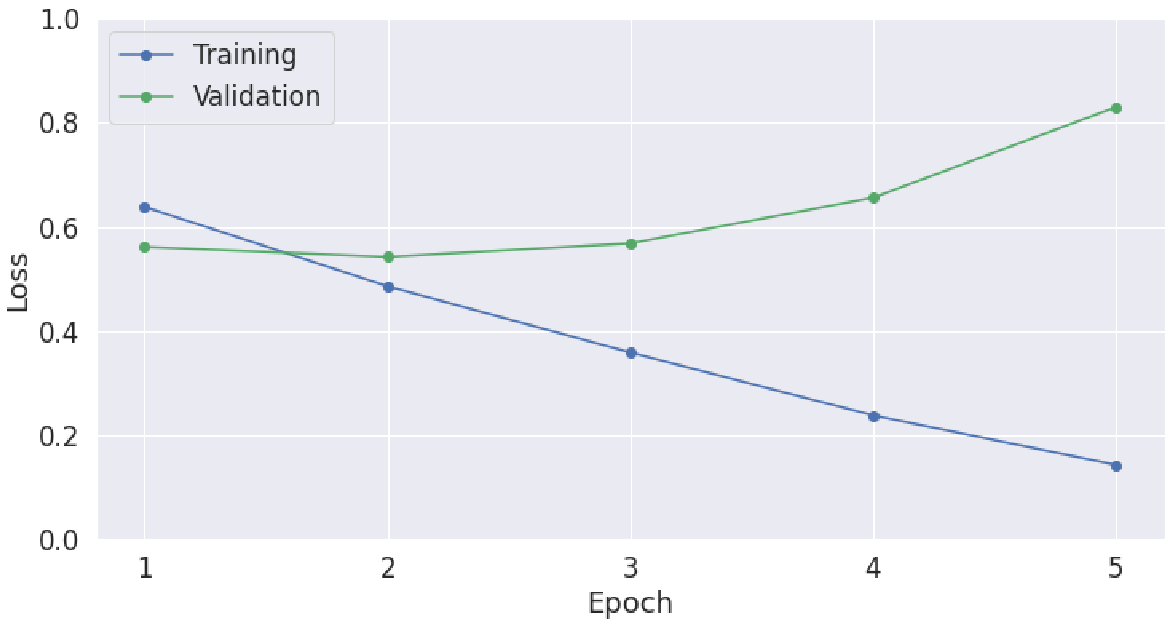
Table 3. Model Results by class (10 fold average).

Class	**Precision**	**Recall**	**F1-Score**
Favourable	0.9	0.77	0.83
Neutral	0.57	0.73	0.62
Unfavourable	0.6	0.8	0.65

Results suggest that the model performs better when predicting favorable tweets, in comparison to the other two classes. This is expected, given that this category amounts to the majority of the dataset. However, the model yields similar results in the classification of neutral and unfavorable tweets, even though there are far more neutral tweets.

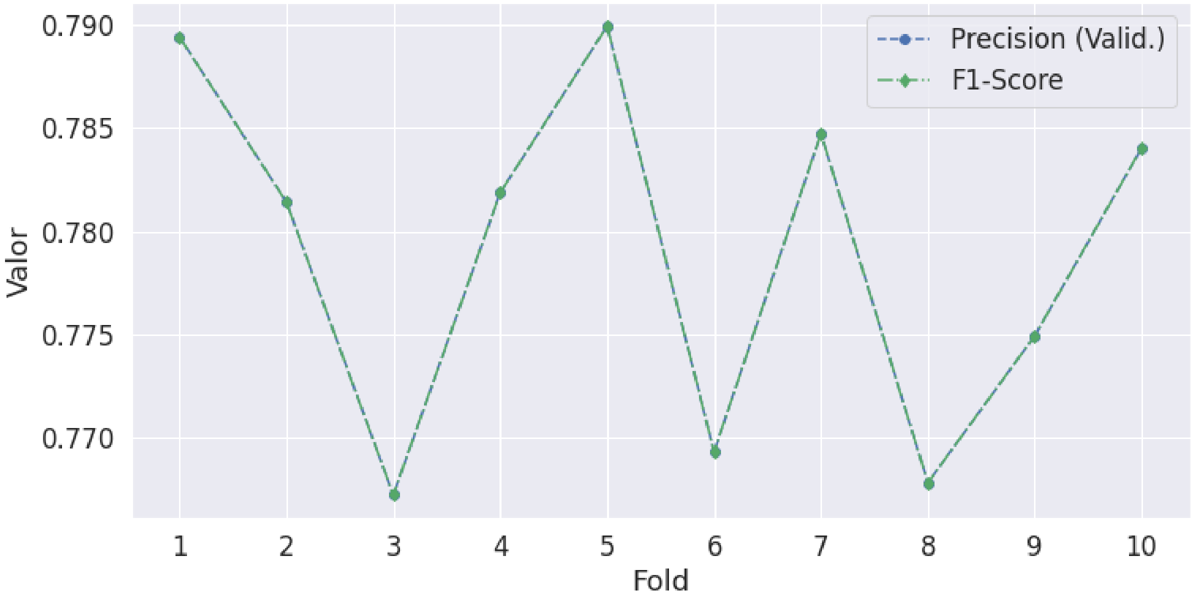
One of the possible reasons for that, as shown in figure 3, is that there is an overfitting within the model. With 5 epochs, the validation loss gradually increases as the training loss decreases. This movement is clearer after the second epoch. Potentially, retraining the model with only 2 epochs would suffice and render better results.

Figure 3. Training and Validation Loss (10 fold avg.)



Finally, figure 4 presents the precision and F1-Score metrics by fold. As shown in the figure, the results vary over the values of 0.75 and 0.79 for each fold, showing that the class imbalance affected the classification, yielding poorer results.

Figure 4. Precision and F1-Score by fold.



Conclusion and Next Steps

In summary, the model performed fairly well for predicting positive vaccination tweets. BERT is especially suited for shorter text, performing well with tweets. Nevertheless, for the other minor classes (neutral and negative tweets), we observed performance issues, with similar results in both cases. These issues are likely to result from the limited number of observations available for negative and neutral posts, in contrast with the vast majority of positive tweets.

Next steps will encompass changes in the number of epochs (maximum of two epochs), as we identify overfitting issues starting from the third epoch. Moreover, by employing oversampling and undersampling strategies, we expect to address the lack of class balance issues to improve the model's capacity for the minor classes.

References

BARBERÁ, P.; BOYDSTUN, A.; LINN, S.; MCMAHON, R.; & NAGLER, J. Automated Text Classification of News Articles: A Practical Guide. Political Analysis, 29(1), 19-42. 2020. doi:10.1017/pan.2020.

DEVLIN, J. et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. [S.l.]: arXiv, 2018. DOI: 10.48550/ARXIV.1810.04805. Available at: https://arxiv.org/abs/1810.04805>.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: pretrained BERT models for Brazilian Portuguese. In: 9TH Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear). [S.l.: s.n.], 2020.

WU, Y. et al. Google’s neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144. 2016.