Brazilian public agreements proposals clustering model based on BERT and kMeans

Douglas Farias Cordeiro[1], Leandro Rodrigues da Silva Souza[2], Núbia Rosa da Silva[3]
[1] Federal University of Goiás (UFG)
[2] Goiano Federal Institute (IFGoiano)
[3] Federal University of Catalão (UFCat)

Abbreviated abstract: Public agreements are one of the main strategies for transferring public financial resources. The identification and classification regarding the object and objectives is commonly performed manually. We use a pre-trained BERT model to generate embeddings from agreements of the Ministry of Agriculture, Livestock and Supply of Brazil. An unsupervised clustering model based on the kMeans method is presented. The results show the potential of using BERT to support typical demands of public administration.

4th Conference on
Statistics and
Data Science
Salvador, Brazil (online)
December 1-3, 2022

Problem, Data, Previous Works

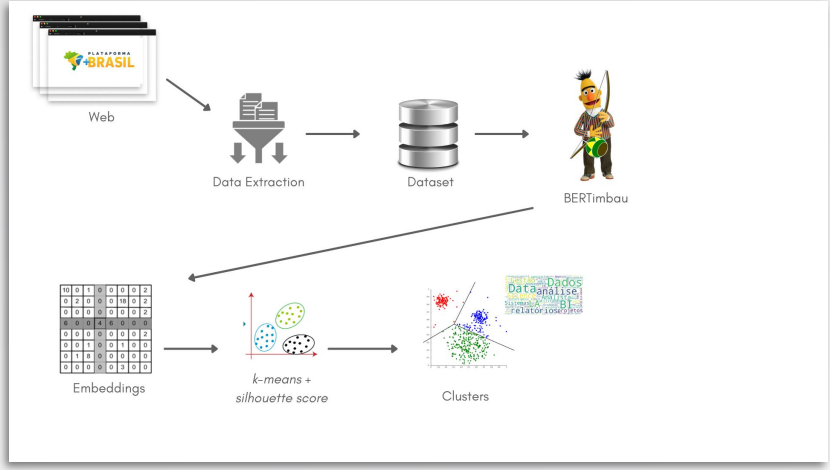
- A large part of the application of public resources by the Union is carried out through **public agreements**
- One of the major problems is the absence of an attribute that allows the categorization of proposed agreements
    - Our proposal is to use a solution based on BERT (SOUZA *et al.*, 2020) and kMeans to build an unsupervised grouping and classification model
- Dataset: Data from the Platform +Brasil[1]
    - We selected data from proposals for public agreements and applied a filter for the year 2021 and the state of Bahia, Brazil;
    - 1371 proposals were extracted. Attribute of interest: **Object of the Proposal**

It is the product of the agreement. It may involve carrying out a project, activity, service, acquisition of goods or an event of reciprocal interest. Examples: construction of local roads.

4th Conference on Statistics and Data Science
Salvador, Brazil (online)
December 1-3, 2022

Methods

- We use the BERT pre-trained model to generate the embedding matrix (SOUZA *et al.*, 2020);
- For the grouping and classification model we used a kMeans algorithm, with definition of the value of k through the silhouette method;
- The textual data of the groups were summarized using a cloud of words → helping to understand the content of each group (there is no prior labeling of the data).



SOUZA et *al*. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In: Proceedings of Brazilian Conference on Intelligent Systems, 2020. p. 403-417.

Results and Conclusions

- Two main thematic groups were identified: 1) acquisition of equipment; 2) Construction/recovery of roads and public spaces.
- The results make it possible to quantify and qualify the proposals under different perspectives according to specific business needs.