Semantic Textual Similarity for Abridging Clinical Notes in Brazilian Electronic Health Records

Lucas T. Bandeira[1], **Bernardo S. Consoli**[1], **Renata Vieira**[2], **Rafael H. Bordini**[1]

[1]School of Technology – Pontifical Catholic University of Rio Grande do Sul
Porto Alegre – RS – Brazil.

[2]School of Technology – University of Évora
Évora – Portugal.

l.treviso@edu.pucrs.br, bernardo.consoli@acad.pucrs.br,

renatav@uevora.pt, rafael.bordini@pucrs.br

Abstract. *With the growing importance of the use of information from electronic patient records in the development of machine learning models, there is also a need for a holistic understanding of those records, in particular abridging the clinical notes so that important information is used in the training process without the repetition that is commonly found in such notes. This paper presents the pre-processing of clinical notes from the BRATECA Dataset, a Brazilian tertiary care data collection, aiming at removing repeated information resulting from the interaction between healthcare providers and patients, considering assigned values of semantic similarity between sentences in clinical notes.*

1. Introduction

In the Artificial Intelligence field, there is significant interest of researchers in developing systems capable of supporting decision making in the healthcare domain [Shamout F 2021]. However, some of the data categories included in the electronic patient records have several characteristics that make their use difficult, which highlights the need to maintain better databases. Although, the most used databases for these purposes are formed by data extracted from hospital records in English-speaking countries, which does not represent the clinical reality in Brazil.

The BRATECA Dataset is one of the few national resources for the development of research projects in computational medicine. In this study, we contribute in that direction by abridging the clinical notes, in particular to remove repeated information, since it is customary for healthcare providers in Brazil to repeat known conditions and ongoing treatments when writing clinical notes. Nevertheless, this repetition obstructs the training of accurate machine learning models for clinical tasks, introducing biased inaccuracies rooted in duplicated content. Our paper outlines the experiments conducted to identify and eliminate potentially repetitive sentences from the clinical notes within BRATECA.

The remainder of this work is organized as follows: Section 2 describes previous work on BRATECA, semantic similarity and word embeddings; Section 3 describes the neural network used; Section 4 describes the data resources used; Section 5 describes the process of abridging clinical notes; Section 6 provides final considerations on the research developed.

2. Related Work

The method identified to generate meaningful data for an AI model considering the heterogeneity of information is through word embeddings, which are vector representations of words in a multidimensional space. As the semantic value of a word is also mapped, it can be inferred that it is possible to perform calculations to discover how similar two sentences are. Thus, studies involving the application of word embeddings were considered, in addition to databases with clinical and general domain resources in Portuguese.

[Consoli et al. 2022] proposed a new collection of Brazilian clinical data containing more than 70,000 admissions, representing a total of more than 2.5 million clinical notes in free text, aiming to create a dataset with Brazilian clinical information.

[Schneider et al. 2020] presents a BERT model trained on clinical texts from electronic medical records of Brazilian hospitals and texts from the biomedical literature. BioBERTpt is the result of transferring knowledge encoded in a multilingual BERT model to a corpus of clinical and biomedical data.

The research conducted by [Mutinda et al. 2021] is noteworthy for creating a dataset of Japanese clinical records through an approach that evaluates the semantic similarity between clinical notes using BERT. The raw text of the documents was first divided into sentences, and a new data collection was created by combining all possible sentence pairs.

Finally, the work carried out by [Real 2021] should be highlighted, aiming to offer a new benchmark for computational semantic tasks in Portuguese, by providing a dataset composed of pairs of sentences annotated with semantic similarity indexes.

3. Neural Network

The developed model consists of fine-tuning BioBERTpt to generate a value for a regression task. In brief, we used the base model in the embedding layer of the selected architecture to create vector representations of the clinical text inputted into the model. To consider both directions of the sequence of values generated by BERT, we added a bidirectional layer during training of the developed model. Next, a Max Pooling layer is applied to filter the numerical vector used as word embeddings. Three Dense Layers are then added, with a Dropout layer between them to randomly turn off nodes and prevent overfitting. The last Dense Layer is constructed with only one node, which generates a single value as the output of the regression task. In this case, the value represents a similarity index between a pair of sentences, ranging from 1 (for sentences containing extremely different information) to 5 (for sentences with practically equal information).

The input text is prepared for the model using bioBERTpt's pre-trained Tokenizer, which employs a WordPiece approach to convert sentences into words and subwords represented by ids. Special tokens like [SEP] are incorporated to signify sentence endings, while the [PAD] token is used to standardize input sizes. The Tokenizer processes two sentences as input, generating a two-dimensional vector containing input ids and attention masks. It's important to highlight that the model creation and data processing script were developed using Python, utilizing libraries including TensorFlow, Keras and spaCy. The model's performance evaluation in the task was executed using Pearson's correlation coefficient, which yielded a score of 0.73.

4. Resources

This work required the use of linguistic resources that cover both the health domain and the general domain. How these resources were used is explained in more details below.

4.1. Work with ASSIN 2

ASSIN 2 is a shared task in the field of natural language processing, which focused on identifying the semantic similarity between pairs of sentences written in Portuguese. Therefore, this dataset was chosen to train the different tested models during the project, as it is one of the few resources in Portuguese with annotated semantic similarity values. As the developed model needed to be trained for a semantic similarity task, we searched for data with similarity values assigned to it. This way, it would be possible to later apply the trained model on clinical data.

4.2. Work with BRATECA

BRATECA presents its information in a free text format, making it highly unstructured for natural language processing tasks. Thus, to eliminate repeated information and compare sentences, it was necessary to process the data by transforming them into sentences that are understandable to a model.

The project's focus is to create patient representations from heterogeneous data, which requires structuring. Due to the free-text format of the records, a pre-trained pipeline in Portuguese from the spaCy library was used to preprocess the records, as well as regular expressions to remove special characters and clean the patient records. After splitting the clinical records into sentences, semantic similarity values were computed by using the previously trained model based on BioBERTpt and ASSIN 2 Dataset. To create patient representations, the clinical records were structured by removing special characters, splitting them into sentences, and eliminating duplicate information. The abridging process was performed by defining the Cartesian product between the sentences and assigning a similarity index to remove highly similar sentences. By processing BRATECA in this way, a new dataset was created that represents a patient's clinical records in a format more suitable for use with AI models.

5. Results

Patients were selected from 4 ranges of clinical records, namely 10, 100, 500, and 1000 records, to gather information on the percentage of repeated information in different hospitalization scenarios. This was done to determine how much information is repeated for patients who are hospitalized for a few days and patients who are hospitalized for a few months.

To create patient representations, all clinical records for a given patient were first selected. The first clinical note was split into sentences and all sentences were added to a new dataset because they contained new information. For subsequent records, the text was also split into sentences and each sentence was compared to the information already in the dataset. If the similarity index between them was greater than or equal to 4, the new sentence was ignored because it duplicated information that had already been obtained.

Thus, in addition to structuring the clinical notes into less heterogeneous sentences, it was possible to eliminate the repeated information present in the clinical records

and consequently significantly reduce the number of clinical notes needed to represent the hospitalization of a patient. Table 1 shows the reduction in the number of clinical notes after eliminating the repeated information using the trained model, varying from 55.45% to 71.45% across the four groups of patients. Furthermore, we checked how many of those sentences were exactly the same for each patient. To achieve this, we compared the Unicode values and calculated the percentage of sentences that were identical. Table 2 presents the results of this comparison, which revealed that a substantial proportion of sentences in clinical notes were duplicates to previous sentences.

Table 1. Clinical notes with similarity less than 4

Patient Id	Clinical Records			
	Clinical Notes	*Sentences Before*	*Sentences After*	*Reduction*
17	10	110	49	55.45%
293	97	661	264	60.06%
76	469	1836	678	63.07%
668	929	4463	1274	71.45%

Table 2. Clinical notes with equal sentences

Patient Id	Clinical Records			
	Clinical Notes	*Sentences Before*	*Sentences After*	*Equal Sent.*
17	10	110	72	34.55%
293	97	661	383	42.06%
76	469	1836	1210	34.10%
668	929	4463	3011	32.54%

5.1. Discussions

The study emphasizes the importance of eliminating redundant information from patient records. The findings suggest that this can significantly decrease the number of clinical notes needed to document a patient's hospitalization. By analyzing the semantic similarity of each sentence in a patient's hospitalization record, it was possible to achieve a 70% reduction in the number of sentences while preserving the same informational value as the original record.

It's important to highlight that the performance of the developed model was hindered in some cases due to grammatical and structural errors in the clinical notes. These errors made it challenging to understand drug names and updates on treatments, particularly when records included abbreviations and lacked white spaces, hindering the model's ability to extract meaningful information from some clinical notes.

6. Conclusion

This study tested the application of an architecture based on Word Embeddings from BioBERTpt that was fine-tuned for a regression task. The goal was to generate a semantic similarity index between sentences of clinical notes in Portuguese to reduce the heterogeneity of the existing information in the records of the BRATECA. We found that the information presented in the clinical records is often repeated countless times during the patient's hospital stay, regardless of its length. Thus, it is worth noting that we developed a method to reduce the number of sentences to be processed by an artificial intelligence model while preserving all the information that existed before creating the patient representations. This approach would reduce the computational cost while using BRATECA data.

References

Consoli, B., dos Santos, H. D. P., Ulbrich, A. H. D. P. S., Vieira, R., and Bordini, R. H. (2022). BRATECA (Brazilian tertiary care dataset): a clinical information dataset for the Portuguese language. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5609–5616, Marseille, France. European Language Resources Association.

Mutinda, F., Yada, S., Wakamiya, S., and Aramaki, E. (2021). Semantic textual similarity in japanese clinical domain texts using bert.

Real, L., F. E. G. O. H. (2021). The assin 2 shared task: A quick overview. Methods Inf Med.

Schneider, E., Souza, J., Knafou, J., Copara, J., Oliveira, L., Gumiel, Y., Ferro Antunes de Oliveira, L., Teodoro, D., Paraiso, E., and Moro, C. (2020). Biobertpt – a portuguese neural language model for clinical named entity recognition. pages 65–72. Association for Computational Linguistics.

Shamout F, Zhu T, C. D. (2021). Machine learning for clinical outcome prediction. volume 14, pages 116–126. Institute of Electrical and Electronics Engineers Inc.