SUPPLEMENTARY FILE - CONTRASTIVE ANALYSIS FOR SCATTERPLOT-BASED REPRESENTATIONS OF DIMENSIONALITY REDUCTION

August 26, 2021

1 Tweets classification

The *tweets* dataset was created after downloading *tweets* mentioning any COVID-19 common symptom (fever, high fever, cough, dry cough, difficulty breathing, shortness of breath) inside the territory of São Paulo state (Brazil) from March 2020 to August 2020. In the end, the dataset contains 72425 *tweets*. To construct the dataset analyzed in the paper, we manually classified 9989 *tweets* as relevant or not. The non-relevant *tweets* are the ones talking about jokes, news articles, or informative *tweets*. Then, using these 9989, we tested various classification algorithms to construct a bigger dataset, as described in the paper.

We trained two different classifiers in order to select the best one at classifying the *tweets* based on relevant or non-relevant. For these two classifiers, *Logistic Regression* [1] and *CatBoost* [2], five approaches were used to create the matrix representation of the *tweets* collection, as shown in Table 1. Firstly, *FastText* [3] and *Word2Vec* [4] words embeddings were trained for the *tweet* collection, where each *tweet* was represented by the mean of the word embedding with 600 dimensions. Then, we also grid-searched for the best parameters (min. document frequency and number of n-grams) of *Bag of Words* representation for each classifier, resulting in different parameters for each one. Finally, we concatenate the matrices representation of *FastText* and *Bag of Words* (in this case with fixed parameters) to generate a unique representation with 11167 parameters. In order to classify, we grid-searched for the best parameters of *Logistic Regression* and fixed the parameters for *CatBoost* (due to excessive time computation). The parameters for each strategy are shown in Tables 2 and 3.

Strategy	Dimensionality
FastText	600
Word2Vec	600
Bag of Words	L.R.: 193120, C.B.: 12370
FastText & Bag of Words	11167
Word2Vec & Bag of Words	11167

Table 1: Dimensionality of the dataset representations.

Strategy	**Logistic Regression**	**CatBoost**
FastText	C:0.1, penalty:'l2', class_weight: None	depth:5, l2_leaf_reg:0.1, learning_rate:0.03
Word2Vec	C:0.1, penalty:'l2', class_weight: None	depth:5, l2_leaf_reg:0.1, learning_rate:0.03
FastText & BOW	C:0.1, penalty:'l2', class_weight: None	depth:5, l2_leaf_reg:0.1, learning_rate:0.03
Word2Vec & BOW	C:0.1, penalty:'l2', class_weight: None	depth:5, l2_leaf_reg:0.1, learning_rate:0.03
Bag of Words	C:0.1, penalty:'l2', class_weight:'balanced'	depth:5, l2_leaf_reg:0.1, learning_rate:0.03

Table 2: Tuned hyper-parameters after grid-search.

Strategy	min_df	ngram_range	stop_words
FastText	3	(1,3)	portuguese
Word2Vec	3	(1,3)	portuguese
Logistic Regression	0	(1,3)	None
CatBoost	3	(1,2)	None

Table 3: Parameters used for bag-of-words construction.

We performed 10-fold cross validation for each classifier to visualize the accuracy distribution according to each strategy, as shown in Figure 1. Note that in both situations the *Word2Vec & Bag of Words* representation represented the best distributions of accuracy.

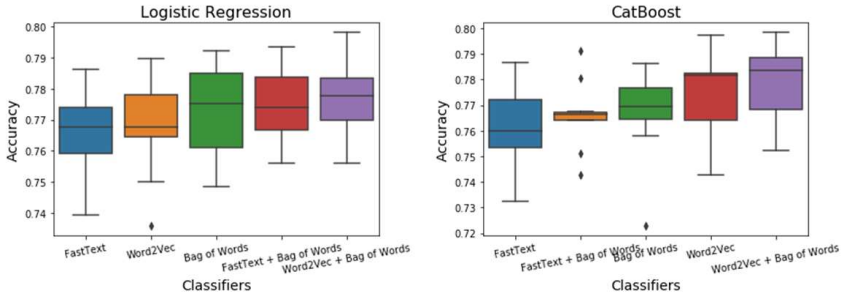


Figure 1: Performance comparison between Logistic Regression and CatBoost.

Finally, we compared the two classifiers with BERT [5] language model. Figure 2 shows that BERT was superior at classifying the *tweets* so that we picked it as the classifier to construct the dataset. That is, the BERT model was used to retrieve only the relevant *tweets* among the 72425 ones.

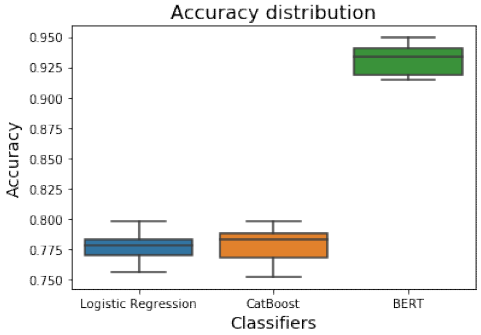
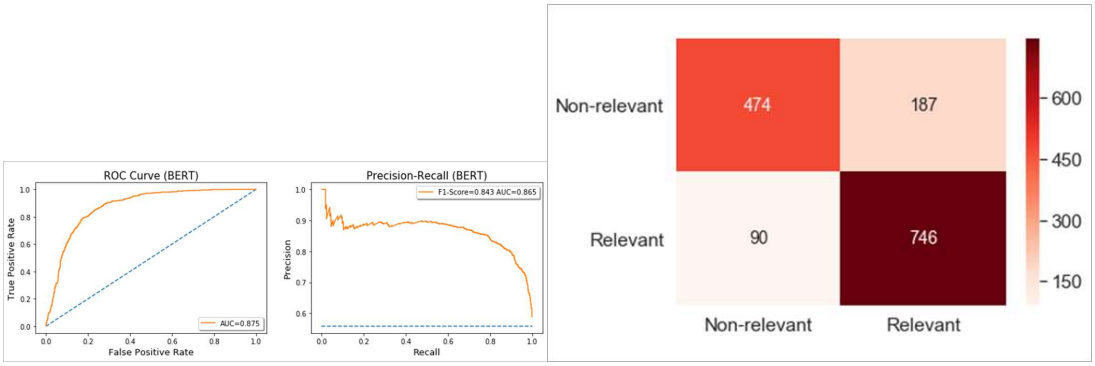


Figure 2: Performance comparison among the classifiers.

To further validate the BERT classifier, we tested it using our hold-out data (a test set with 1497 *tweets*). Figure 3 shows the Precision-Recall and ROC curves, as well as the confusion matrix.



(a) ROC and Precision-Recall curves. (b) Confusion matrix.

Figure 3: Evaluation metrics.