**Supplementary Material 1 - Automatized categorization tests**

We performed the verification of the success of categorization through machine learning analyses. The Twitter database, due to its volume, only had a random sample categorized (1,000 posts).

We outline commands to select and evaluate prediction models in an automated manner. In addition to machine learning algorithms, it was also necessary to adopt text representation techniques to get final models. The procedures were done using two libraries (packages) in Python (Gensim and Scikit-learn). From the provided database, the text is transformed from text format to vector format, utilizing embeddings (matrix representation of the text). Embedding generation is an unsupervised process, so it was not necessary to delete posts. Embeddings present differences related to the transformation of text into matrices that influence the final performance of the classification algorithms.

With this, we were able to train and evaluate the machine learning algorithms, adopting the division of the database between classified and unclassified. We trained classifiers to predict the response variable (categories), so we could obtain the prediction accuracy for each algorithm. Accuracy was obtained through cross-validation with samples gathered randomly from the data set. Cross-validation was done using the Scikit-learn package. The algorithms for transformation and classification used for the analysis were:

**Transformation Algorithms (Embeddings)**

TI-IDF (Term frequency – inverse document frequency): Representation of the term-document matrix with calculated weights, where the use of parameters is not necessary (Jurafsky, 2000).

LSA (Latent Semantic Analysis): Execution using the term-matrix document (with weights calculated by TLDF). In this test, we used a dimensional space of 100 and 300. The reconstruction of the term-document matrix was used as an embedding (Deerwester et al., 1990).

TopicLSA: Alteration of the previous algorithm, in which, instead of using matrix reconstruction as embedding, we use the V'S' matrix. The name is used because the generated embedding is not represented concerning the vocabulary, but rather, to the dimensions resulting from the method (called topics by Gensim).To select the number of dimensions, the same LSA values were tested (Deerwester et al., 1990).

**Classifiers**

Naïve Bayes: algorithm where it was necessary to test whether the conditional distribution of values is Gaussian, without the need for other parameters (Zhang, 2004).

Logistic regression: in the case of logistic regression, the main parameter used is the regularization factor, responsible for penalizing model parameter values. We use the value 10 with a low penalty applied (Böhning, 1992).

SVM (Support vector machine): in the case of SVM, what is most important is the penalty for wrong classifications and the Kernel function used. In this case, we adopted the logistic regression values, with the difference that the penalty logic in this case is inverse, thus, smaller values are more penalized. We adopted the linear function for the Kernel (Boswell, 2002).

Random Forest: for the algorithm, it is necessary to define the number of decision trees. We chose 500 decision trees. In addition, the criterion for evaluating tree breakage in a minimum number of samples per leaf was selected as a parameter, the value adopted was 5 (Breiman, 2001).

Finally, we obtained different classification results by the algorithm for categories. For the jaguar’s database, the best accuracy in the classification with validation from random sampling occurred with the SVM embedding LSA Topic algorithm (0.71) and the best accuracy using validation without random sampling occurred with the SVM algorithm with embedding LSA (0.62). For the puma database, the best accuracy in classification with validation from random sampling occurred with the Naive Bayes embedding LSA algorithm (0.64) and the best accuracy using validation without random sampling occurred with the Naive Bayes algorithms embedding TI-IDF and Random Forest embedding LSA (0.58, 0.58).

**References**

Böhning, D. (1992). Multinomial logistic regression algorithm. *Annals of the institute of Statistical Mathematics* 44, 197–200.

Boswell, D. (2002). Introduction to support vector machines. *Departement of Computer Science and Engineering University of California San Diego* 11.

Breiman, L. (2001). Random forests. *Machine learning* 45, 5–32.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science* 41, 391–407.

Jurafsky, D. (2000). *Speech & language processing*. Pearson Education India.

Zhang, H. (2004). The optimality of naive Bayes. *Aa* 1, 3.