Automatic Classification of Portuguese Proverbs

Jorge Baptista
University of Algarve, Faro, Portugal
INESC-ID Lisbon, Portugal

Sónia Reis
University Algarve, Faro, Portugal
INESC-ID Lisbon, Portugal

Abstract

In this paper, natural language processing (NLP) and machine learning methods and tools are applied to the task of topic (thematic or semantic) classification of Portuguese proverbs. This is a difficult task since proverbs are usually very short sentences. Such classification should allow an easier selection of the most relevant proverbs for a given situation, considering their context in discourse or within a text. For that, we used, on the one hand, a collection of +32,000 proverbial expressions organized "thematically" into a large set of previously attributed topics (+2,200) and, on the other hand, the ORANGE data mining toolkit, along with the NLP and machine learning tools it provides. Since the classification provided in the collection of proverbs is, for the most part, based only on a keyword in the body of the proverbs, 2 experiments were set up, to determine the feasibility of the task with a modicum of effort and the most promising configurations applicable. Different sample sizes, 100 and 50 proverbs randomly selected per topic, corresponding to Scenario 1 and 2, respectively, were contrasted; several preprocessing strategies were explored, and different data representation methods tested against several learning algorithms. Results show that Neural Networks is the best performing model, achieving the best classification accuracy of 70% and 61%, in the two different experimental scenarios, Scenario 1 and 2, respectively. Some of the inaccurate classification cases seem to indicate that the machine learning approach can sometimes do a better job than a human classifier, especially considering the manual attribution of the topics by the collection's author, the sheer number of topics involved, and the very unbalanced distribution of proverbs per topic. Based on the results achieved, the paper presents some proposals for future work to cope with such difficulties.

2012 ACM Subject Classification Human-centered computing

Keywords and phrases Portuguese Proverbs, Automatic Topic Classification, Machine Learning

Digital Object Identifier 10.4230/OASIcs.SLATE.2022.2

Supplementary Material *Dataset*: `https://doi.org/10.13140/RG.2.2.22354.02242`

Funding *Jorge Baptista*: This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT), under project ref. UIDB/50021/2020.

1 Introduction

This paper aims at the automatic topic/thematic classification of (European) Portuguese proverbs. The semantic (or thematic) classification of proverbs is not, *a priori*, a trivial task, not only for the very nature of proverbs as short sentences, relatively insulated from the surrounding text; but especially because of the idiomatic (figurative) character of many of these expressions; this difficulty is also due to the fact that, for a given theme, the meaning of the words forming the proverb is not necessarily related to the expression by which we might designate that same theme. For example, in [5], we find *Água mole em pedra dura, tanto dá até que fura* lit.: "water soft on stone hard, so much it hits until it bores the stone" "Water dropping day by day wears the hardest rock away" [1] under the topic PERSEVERANÇA

[1] Equivalent suggested by: `https://www.sk.com.br/sk-proverbios-portugues-ingles.html`

"perseverance". Arguably, even if no clue is given in its wording, it may be easier to asign this topic to this very well known proverb, than to the following one: *A pouco e pouco fia a velha a touca* "Little by little the old woman spins the cap" which the author also classes under the same topic, but, as it is much less known and used, where the lexical clues may hint at several other potential meanings (persistence, carefulness/wisdom, progressive, wealth accumulation, etc.). Compare how much easier it is to classify under that same topic a proverb that explicitly has that word as its subject, and whose meaning is clearly denoted by its elements: *A perseverança sempre/tudo alcança/vence* "Perseverance always pays off/wins out everything" (Lexical or punctuation variants of the same *paremiological unit* are represented in a compact way using "/". On the definition of this concept, see [10].) Added to these difficulties is the fact that, on the one hand, some proverbs can be used in different situations, with subtle variations in meaning.

Identifying the theme of a proverb can be useful in several situations. According to [8], "the task for the automatic classification of proverbs which can assist users in selecting a suitable proverb according to a given context prove[s] to be important and interesting but challenging." Thus, when serving as a *motto* to a text, either in the title or in the *lead* of a newspaper article, or even in the conclusion of the text, the proverb anticipates (or ends the text with) a judgment of value that the author avoids making in the first person, taking refuge in the so-called "popular wisdom". This may be viewed as the underlying *rationale* behind the work of [6], who assessed different methods of associating proverbs to news headlines, exploring different semantic similarity metrics.

In this paper we address the practical problem of automatic topic classification of Portuguese proverbs and try to provide answers to the following research questions:

1. How are proverbs classified by topic/theme in the relevant literature? How good is the adequacy, the consistency or even the usefulness of extant classifications in the available collections of Portuguese proverbs?
2. To what extent can the available classification of Portuguese proverbs be captured by NLP and ML tools and techniques, using only the words in the proverbs? Is the size of each class relevant for this task, considering that proverb classifications in the literature are very fragmentary (a large number of topics, with a small number of proverbs each), in spite of the large collections of proverbs available?

This paper is organized as follows: Next, Section 2 presents the most similar works found in the literature on machine learning strategies to proverb classification. Section 3 presents a critical assessment of the collection of proverbs used here as corpus, its structure and composition. Then, Section 4, explains the sampling method used to select the expressions here used and outlines the two experiments conducted using the ORANGE [2] data mining toolkit [2], along with the results obtained. In Section 5, a short discussion ensues, Finally, Section 6 concludes the paper and presents prospects for future work.

2 Related work

There seems to be very little related work dealing with the automatic, machine-learning based, classification of proverbs. These are presented in this Section.

Noah and Ismail (2008) [8] is the first work, to the best of our knowledge, to address this challenge in a similar way, aiming at helping an end-user to better select the most appropriate proverb for a given context. The authors comment on the non-trivial nature of the task,

[2] `https://orangedatamining.com/`

since proverbs are very concise structures, thus limiting the type and number of features that can be extracted from them and the classification techniques that can be applied. The paper draws on a relatively small corpus of 1,000 Malayan proverbs, split into two partitions 50% for training and 50% for testing. The data was divided into equal-sized 5 classes, according to the respective topics (FAMILY, LIFE, DESTINY, SOCIAL and KNOWLEDGE). These topics are generic in nature, and encompass proverbs that pertain to many situations and contexts. Three classification scenarios were considered: (i) the proverbs, alone; (ii) the proverbs along with an explanation of their meaning; and (iii) the proverbs, their explanation and an example of a sentence containing those proverbs. The data was preprocessed, removing the stopwords from the proverbs, to achieve a reduction of the number of extracted features, as well as obtaining better performing ones. No other preprocessing step was indicated. Two Naïve Bayes models (multinomial and Bernouilli's multivariate) were tested. In the training stage, a classification accuracy (CA) of 99% is reported. In the testing step, the best reported result was the third scenario (proverb+explanation+example), with a CA=72.2% for the multinomial model and 68.2% for the multivariate. The authors also report an increase in the performance depending on the size of the vocabulary, from a CA of 36% for a minimal vocabulary of 632 words, to the reported value of 72.2% with a 3,203 words' vocabulary.

For this paper, only the first scenario can be reproduced. Firstly, no collection of Portuguese proverbs was found with their respective explanation. Secondly, while some experiments have been made to determine the distribution of proverbs from the Portuguese *paremiological minimum* [10] in 3 large corpora [11], in most cases, the proverb forms by itself an isolated sentence, with very little formal relation with the surrounding text, working as a textual "island", as it has been already remarked in the literature [3, p.37; 221-222]. This offers little to no help in the classification of proverbs by topic.

Next, Mendes & Oliveira (2020) [6] assessed different similarity metrics in order to associate proverbs with news headlines. This is basically an automatic recommendation task, similar to the one outlined by [8]. The corpus of news headlines was obtained using a news API from online Portuguese newspapers, involving some keywords (CLIMA "climate", AMBIENTE "environment" and AQUECIMENTO GLOBAL "global warming"), retrieved during the 3 months prior to February 2020 (the precise number of news headlines was not provided). The corpus of proverbs consisted of a list of approximately 1,600 expressions, drawn from the project NATURA (U. Minho) [1] [3] (The selection criteria of these proverbs, from the source database, which contains 2,293 proverbs, was not given). Both corpora were processed (tokenized, part-of-speech tagged and lemmatized, and stopwords removed). To estimate semantic similarity between the news headlines and the proverbs, several basic techniques were used (Jaccard coefficient, word count and TF-IDF vectorization), as well as static word embeddings (Glove and FastText) and BERT (references within the paper). To assess the results, the authors used a questionnaire, which revealed that most of the time people could establish a relation between the automatically selected proverb and the news headline. This was more obvious when there were common lexical elements between them, thus recommending the use of simpler computational methods, like the Jaccard coefficient. Deeper semantic representations, such as word embeddings (BERT), produced poorer results, which is explained by the authors by the figurative nature typical of these proverbial expressions.

[3] `https://natura.di.uminho.pt/~jj/pln/proverbio.dic`

Though this paper does not frame the proverbs' classification within an extrinsical evaluation, that is, in an applicational context, as a recommendation task; it does compare bag-of-words with word embeddings data representation methods, as well as different types of learning models for the classification. A previously classified list of proverbs were used, selecting the most frequently occurring topics within a very large collection (+32,000 expressions), and using same-sized classes.

3 Corpus

For this work, we used a relatively large collection [5] of over 32,000 proverbs, organized by what the author called an "interpretative classification essay" (p.12). This is a task that has raised doubts for the author himself, who even mentions that some proverbs included in the collection "do not contain any "far-reaching principle" nor contain a moral maxim confirmed by the course of generations", while some cases can be deemed as just "non-sense". The author also added a list of 167 proverbs "without classification": "[...] some dozens of sayings, of difficult interpretation because its meaning is multiple or diffuse, or eventually so personalized that only a deep investigation could, in some cases, determine it (p.14, our translation)". The extensive list of proverbs from this collection had already been digitized and integrated into the database that was the main source of [9]'s work.

Going through the collection, it turns out that the "organization" of the proverbs collected there is essentially based on the fact that most proverbs under a given "topic" present that same word or cognate forms of the word designating the "topic" itself. While this may very well be adequate for the topic *brigas* "fights" and proverbs like *Brigam dois se um quer* "Two fight if one will", it is much less obvious when the proverb's interpretation is predominantly idiomatic, like the inclusion under the topic *brilho* "shine" of the proverb *O brilho abre o trilho* (lit.:"The shine/glow opens the trail/way") "Someone's success/fame makes it easier for something to happen or for someone to do something". In other cases, cognate words are split into distinct topics: PERDA (noun) "loss", PERDER (verb) "loose" and PERDER-SE (pronominal verb) "loose oneself/get lost", without a consistent distribution of the proverbs within each group, namely, proverbs with the verb but noun in the group headed by the noun, and vice-versa. It is unlikely that joining some topics designated by cognate words from different parts-of-speech would reduce this fragmentary classification, since this seldom happens in this collection.

This also leads to a fragmentary nature of the "classes" construed by the author. There are cases where a given "theme/topic" presents only one, two or very few proverbs, or even different variants of the same proverb, sometimes even with the same keyword, e.g. ALQUIMIA "alchemy": *Alquimia e/está provada :/, ter renda e não pagar/gastar nada* "Alchemy is proven :/, to have income and don't spend/pay anything". In other cases, a synonym of the keyword is used instead: (ALMOFARIZ "mortar": *Nem corte sem chocarreiro, nem* gral *sem malhadeiro* (*gral* = *almofariz* "mortar") "Neither a court without a court jester, nor a mortar without a pestle".

In the other extreme situation, some "topics" are so broad, e.g. AGRICULTURA "agriculture", that the proverbs there included address the most disparate situations: from the quality of the soil to produce: *Terra negra dá bom pão* "Black soil makes good bread"; to the relationship between weather and crops: *Ano de rosas, ano de pão* "Year of roses, year of bread"; or the need to care for the crops: *Vinha sem guarda, vindima feita* "Unguarded vine, harvest done". In other cases (e.g. ALIMENTO "food"), different food products (e.g. *carne* "meat" and *peixe* "fish"; *fruta* "fruit" and *legumes* "vegetables"; *leite* "milk", *pão* "bread" and *mel* "honey") give title to the different "subclasses".

Finally, several proverbs occur more than once in the collection, but they are not classified consistently, e.g. the proverb *Se queres mel, suporta a abelha* "If you want honey, support the bee" appears both under the "topic" MEL "honey" and as a subtype of ALIMENTO "food" (but not under ABELHA "bee"), thus confirming the difficulty in identifying (and naming) the proper topic of a given proverbial expression.

Despite all the reservations that this so-called "classification" may rise, there are few works, especially recent ones, that present this kind of topic/thematic organization (we know of [12], for instance). Other collections present only alphabetically ordered listings of the words appearing in the proverbs [4], eventually accompanied by a remissive index of the vocabulary [7]. For lack of a better resource, this collection [5] was used to this paper, not least because its size, but also because it had already been digitized into the database that was used for [9], and it only required that topics be associated to the proverbs.

4 Methods

In total, [5] collects 32,067 proverbs, distributed over 2,227 different "topics" (an average of approximately 14,4 proverbs per topic). There is a long tail of topics with only one proverb (482), 112 topics with 50 or more proverbs each (14,078 proverbs), and only 43 topics with 100 (9,537 proverbs) or more.

A subcorpus with the largest classes, that is, the topics with the largest set of proverbs, were selected. Two classifications scenarios were considered in these experiments:

- **Scenario 1**: only the 43 most frequent topics, with +100 proverbs each, were selected;
- **Scenario 2**: the 112 topics with 50+ proverbs were considered.

A random sample of 100 and 50 proverbs per topic and for each scenario, respectively, was retrieved from the subcorpus [4]. Each topic has exactly the same number of proverbs in both scenarios. The proverbs in Scenario 2 are included in Scenario 1. Using the ORANGE [2] machine-learning toolkit, the workflow presented in Figure 1 was set up.

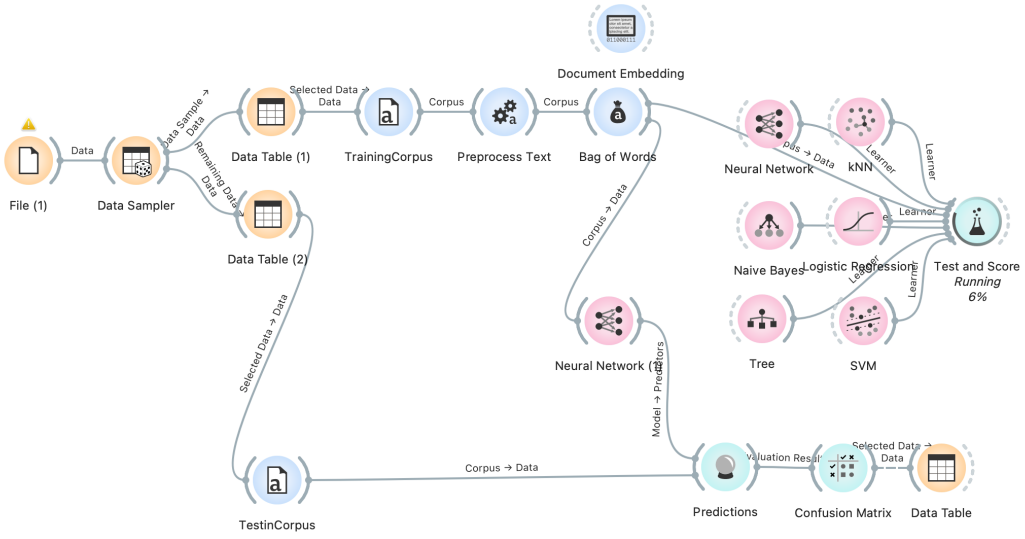

Figure 1 ORANGE workflow configuration.

[4] Available at: `http://dx.doi.org/10.13140/RG.2.2.22354.02242`

Each scenario was tested consecutively. The datasets for the two scenarios consist of samples of 4,300 and 5,600 proverbs each. As shown in Fig. 1, a DataSampler widget was used to split the data into fixed-sized, training/testing partitions, with replicable, stratified sampling options selected, 4,000/300 in Scenario 1 and 5,000/600 in Scenario 2, respectively. Preprocessing consisted of text transformation (conversion to lowercase), tokenization (keeping words and punctuation) and part-of-speech tagging (averaged perceptron tagger), using the default configurations of the system. Two data representation options were compared: Bag-of-Words (BoW) and DocumentEmbeddings. After preliminary experiments, it was clear that the Document Embeddings data representation method consistently produced very poor results (Classification Accuracy (CA): 0.030) against the bag-of-words (CA: 0.415), so it was discarded. Several models were evaluated, using the Test&Score widget in a 10-fold cross-validation setting: Neural Networks (NN), Nearest Neighbour (kNN), Naive Bayes (NB), Logistic Regression (LR), a forward pruning decision tree algorithm (Tree), and Support Vector Machines (SVM). Table 1 shows the performance of each model, estimated by the Test&Score tool in Scenario 1, and Table 2 in Scenario 2. Evaluation of the best performing model (Neural Networks) with the testing corpus yielded the results shown in Table 3.

Table 1 Test&Score: Evaluation of different models in Scenario 1: 43 most frequent topics and 100 randomly selected proverbs per topic. Metrics: AUC: area under the ROC curve, CA: classification accuracy, F1: harmonic mean (of Precision and Recall).

Model	AUC	CA	F1	Precision	Recall
SVM	0.69522	0.01375	0.01066	0.05234	0.01375
kNN	0.78336	0.36150	0.36166	0.43864	0.36150
Tree	0.76424	0.44850	0.45341	0.47162	0.44850
Naive Bayes	0.93931	0.59500	0.58550	0.60109	0.59500
Logistic Regression	0.95132	0.66150	0.66299	0.66799	0.66150
Neural Network	0.95068	**0.68825**	0.68531	0.68640	0.68825

Table 2 Test&Score: Evaluation of different models in Scenario 2: 112 most frequent topics and 50 randomly selected proverbs per topic.

Model	AUC	CA	F1	Precision	Recall
SVM	0.67129	0.00220	0.00310	0.01805	0.00220
Tree	0.60671	0.12900	0.13154	0.17824	0.12900
kNN	0.74069	0.24480	0.25509	0.37730	0.24480
Naive Bayes	0.91620	0.48580	0.48219	0.52645	0.48580
Logistic Regression	0.92985	0.57740	0.57985	0.59201	0.57740
Neural Network	0.93662	**0.60480**	0.60425	0.61241	0.60480

Table 3 Results from best performing model, Neural Networks (NN), in both Scenario 1 (train: 4,000 proverbs / test: 300 proverbs) and Scenario 2 (train: 5,000/test: 600).

Scenario	Model	AUC	CA	F1	Precision	Recall
Scenario 1	NN	0.95900	**0.70300**	0.69900	0.71800	0.70300
Scenario 2	NN	0.94300	0.61200	0.61000	0.64600	0.61200

5 Discussion

From the Test&Score widget (Tables 1 and 2), the best performing model in both scenarios is the Neural Networks (CA: 0.688 and 0.605, respectively), with Logistic Regression following close behind (CA: 0.662 and 0.577, respectively). The relative performance of the tested models is similar and consistent across the two scenarios, with only kNN and Tree changing ranks. Overall, the reduction of the number of proverbs by topic degraded the models' predicted performance. The testing step (Table 3), however, produced slightly better results, while keeping the same difference. The best performance was achieved in Scenario 1 (CA: 0.703).

The experimental designs presented in related work (Section 2) do not allow for a straightforward comparison of results. Even so, we notice that the best performing models in the experiments on Malayan proverbs, while including in the training dataset not only the proverb itself, but also both a definition and an example of the proverb in its context of use, only achieved a 72.2% accuracy.

Looking deeper into the classification errors in Scenario 2, we find that in 77 out 233 errors the proverb contains the predicted keyword, and in most cases, instead of the target keyword, a cognate word, from a different part-of-speech appears: Proverb: *Prometer e não dar é dever e não pagar* "To promise and not give is to owe and not pay"; Actual: *promessas* "promises" (noun-pl.); Predicted: *dever* "to owe" (verb)/"duty" noun.

Some of these results are very interesting, for even if the topic classification is inaccurate, some clear relation can be established between the predicted tag and the respective proverb. In some cases, one can even say that the machine learning model produced a better result than the human annotator: Proverb: *Onde entra o ar e o sol, não entra o doutor* "Where the air and the sun enter, the doctor does not enter"; Actual: *sol* "sun"; Predicted: *saúde* "health"; Proverb: *Mal vai ao passarinho na mão do menino* "Badly goes the little bird in the little boy's hand"; Actual: *aves* "birds"; Predicted: *criança* "child".

6 Conclusion and future work

This study aimed at an automatic topic (semantic) classification of Portuguese proverbs. The Orange data mining toolkit was used to produce models that could appropriately assign a topic/theme to Portuguese proverbs. Different models were trained and tested in 2 experimental Scenarios, varying the number of proverbs (100 and 50) per topic. These were drawn from a large collection [5] of Portuguese proverbs (over 32,000 expressions), already classified according to a large set of topics (over 2,200). The data sampling, the preprocessing configuration and the data representation methods were presented and discussed. In these experiments, the best performing model was Neural Networks, achieving a best classification accuracy of 70% in Scenario 1 (100 proverbs per topic) and 61% in Scenario 2 (50 proverbs per topic). Results are encouraging and in line with previous related work [8]. However, there is still much room for improvement. Going through the cases of inaccurate automatic classification, it seems that sometimes the model does a better job in predicting the appropriate class than the human classification.

Some of the problems observed derive from the topic classification itself as it was presented in the proverb collection, since this seems to have been less than rigorous and often inconsistent, perhaps as a consequence of the large set of proverbs here collected and the (apparently) manual procedure in the classification (no explicit criteria nor method of classification was provided). Furthermore, the high number of classes and the non-uniform distribution of proverbs across these classes do not allow for a straightforward application of the machine-learning methods here used to the entire collection.

In the future, we would like to be able to automatically classify the entire database of proverbs (+114,000) used by [9], which contains those proverbs of the collection here used [5], but also encompasses 3 other collections. For this, there is still a long way to go. First, we believe that the set of topics must be substantially reduced, as the machine-learning methods available in the ORANGE toolkit degrades when over 100 classes are used. Secondly, the classification, in our view, should be more semantic in nature, and not so much based on the presence of a given keyword, as it seems to be the case in [5]. It is likely that, in spite of the *caveat* from [6], different word embeddings (other than from those distributed with ORANGE), purposefully built for Portuguese (e.g. BERTimbau, [14]; LX-DSemVectors [13]) could improve the results. Eventually, vocabulary filtering or other machine-learning techniques such as clustering algorithms could be used to this end, and help determine the most adequate set of topics. Finally, semantic similarity metrics could be used to rank/associate same-topic proverbs and then to confront those data-driven classifications with human assessment, or, alternatively, with instances of proverbs in context [11].

References

1 José João Almeida. Dicionário aberto de calão e expressões idiomáticas [online]. Available at. `http://natura.di.uminho.pt/~jj/pln/calao/dicionario.pdf`, 2014.

2 Janez Demšar, Tomaž Curk, Aleš Erjavec, Črt Gorup, Tomaž Hočevar, Mitar Milutinovič, Martin Možina, Matija Polajnar, Marko Toplak, Anže Starič, Miha Štajdohar, Lan Umek, Lan Žagar, Jure Žbontar, Marinka Žitnik, and Blaž Zupan. *Orange*: Data Mining Toolbox in Python. *Journal of Machine Learning Research*, 14:2349–2353, 2013.

3 Ana Lopes. *Texto Proverbial Português: elementos para uma análise semântica e pragmática*. PhD thesis, Universidade de Coimbra, Coimbra, Portugal, 1992.

4 José Pedro Machado. *O Grande Livro dos Provérbios*. Editorial Notícias, Lisboa, 1996.

5 José Ricardo Marques da Costa. *O Livro dos Provérbios Portugueses*. Editorial Presença, Lisboa, 1999.

6 Rui Mendes and Hugo Gonçalo Oliveira. Comparing different methods for assigning Portuguese proverbs to news headlines. In *11th International Conference on Computational Creativity ((ICCC'20)*, pages 153–160, 2020.

7 António Moreira. *Provérbios Portugueses*. Editorial Notícias, 1996.

8 S.A. Noah and F. Ismail. Automatic classifications of Malay proverbs using naïve Bayesian algorithm. *Information Technology Journal*, 7:1016–1022, 2008.

9 Sónia Reis. *Expressões proverbiais do português: Usos, variação formal e identificação automática*. PhD thesis, Universidade do Algarve, Faro, Portugal, 2020.

10 Sónia Reis and Jorge Baptista. Determinação de um mínimo paremiológico do português europeu. *Acta Scientiarum. Language and Culture*, 2(42):e52114, 2020.

11 Sónia Reis, Jorge Baptista, and Nuno Mamede. Provérbios portugueses usuais: distribuição em corpora. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 325–334. Sociedade Brasileira de Computação, 2021.

12 Fernando Ribeiro de Mello. *Nova Recolha e Provérbios Portugueses e outros lugares-comuns*. Edições Afrodite, 2 edition, 1986.

13 João Rodrigues, António Branco, Steven Neale, and João Silva. LX-DSemVectors: Distributional Semantics Models for Portuguese. In *International Conference on Computational Processing of the Portuguese Language (PROPOR 2016)*, pages 259–270. Springer, 2016.

14 Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In Ricardo Cerri and Ronaldo C. Prati, editors, *Intelligent Systems. BRACIS 2020*, pages 403–417. Springer, 2020.