Semantic Textual Similarity: In Defense of Wordnet-Based Methods

Eduardo Corrêa Gonçalves[1]

[1]Escola Nacional de Ciências Estatísticas (ENCE/IBGE)
Rio de Janeiro – RJ – Brazil

eduardo.correa@ibge.gov.br

Abstract. *Wordnets have long been used as a tool for evaluating the semantic similarity between short texts. In addition to being simpler than recent deep learning approaches, methods based on wordnets offer an important advantage: they deliver results that are easy to interpret as their decisions are usually taken by considering the proximity between graph nodes. In this work, we explore a lightweight approach based on a Portuguese wordnet to solve the ASSIN 2 Semantic Textual Similarity (STS) shared task. In this task, each object of a dataset consists of a pair of Portuguese sentences annotated with its semantic score and the goal is to learn an STS model to estimate the similarity value of new, previously unseen, sentence pairs. Experiments show that our results are competitive with state-of-the-art methods in terms of mean squared error.*

1. Introduction

Semantic Textual Similarity (STS) is the task of assessing the degree of semantic equivalence between two short pieces of text [Chandrasekaran and Mago 2021, Agirre et al. 2012]. A popular application of STS can be found in Question Answering (QA) Systems [Soares and Parreiras 2020], i.e., systems that automatically answer questions posed by humans in natural language. Consider, for instance, two distinct users of a medical QA System who are interested in obtaining information about the symptoms of diabetes. The first user could ask "what are the signs of diabetes?" whereas the second might pose the question as "how can I check if I have diabetes?". Once the questions are equivalent, the system should be capable of providing the same answer for both users.

In addition to QA Systems, there are several other important applications of STS, varying from plagiarism detection [Ferrero et al. 2017] to the comparison of product descriptions [de Lima and Gonçalves 2022]. As a result, a number of competitions (challenges and shared tasks) to promote research in STS have been run over the last few years, such as n2c2/OHNLP Clinical STS Track [Wang et al. 2020], SemEval Task on STS [Agirre et al. 2012, Cer et. al 2017], and ASSIN, *Avaliação de Similaridade Semântica e Inferência Textual* (Evaluating Semantic Similarity and Textual Entailment) [Fonseca et al. 2016, Real et. al. 2019]. The second and last edition of ASSIN – which we will refer to as ASSIN 2 in the remainder of this paper – is the focus of the present work.

The organizers of ASSIN 2 shared task made available a dataset composed of about 10,000 pairs of sentences in Brazilian Portuguese (6,500 for training, 500 for validation, and 2,448 for testing). Each pair is assigned a semantic similarity score between 1.0 (the sentences are completely unrelated) and 5.0 (the sentences are

equivalent). For instance, the pair "*Um homem está tocando uma flauta*" ("A man is playing a flute") and "*Um homem está tocando um instrumento*" ("A man is playing an instrument") is scored with 4.5. A total of nine teams participated in the challenge whose aim was to produce the best model in terms of Pearson correlation (ρ), with Mean Squared Error (MSE) being considered as a secondary evaluation metric.

At the end of the challenge, state-of-the-art BERT-based models have shown remarkable performance, obtaining the overall best results [Fonseca and Alvarenga 2019; Rodrigues et al. 2019a; Rodrigues et al. 2019b]. However, despite their effectiveness, it is necessary to observe that BERT and its variants suffer from a drawback: they were designed to maximize predictive performance, but do not consider the comprehensibility (interpretability) of the model. This is a relevant issue for application domains in which it is necessary to determine how a model came to its conclusions, either for legal/transparency reasons or to follow ethical guidelines. For example, this is often the case of natural language processing (NLP) applications in the context of public administration, as stressed in [Anthopoulos and Wood 2021; Darrazão et al. 2023; de Lima and Gonçalves 2022].

In this paper, we explore a lightweight approach based on a Portuguese wordnet, namely Onto.PT [Gonçalo Oliveira and Gomes 2014], to solve the ASSIN 2 STS task. More specifically, the proposed strategy is based on the combination of a few traditional lexical and distributional features with semantic features computed with the utilization of Onto.PT. The choice of a wordnet-based solution was mainly motivated by the fact that this kind of structure provides the user with means for understanding how the model is generating predictions, since model's decisions are usually taken by considering the proximity between nodes in the wordnet graph. Experiments with ASSIN 2 collection showed that our proposal achieved competitive MSE results compared with most of the state-of-the-art deep learning methods.

The remainder of the paper is divided as follows. Section 2 reviews work related to our proposal and gives a short overview on wordnets. Section 3 outlines the advantages of wordnet-based STS models and presents our proposed approach. Section 4 reports the experimental results. Concluding remarks and future directions are given in Section 5.

2. Related Work and Background

This section revises the work related to our proposal and gives an overview of wordnet concepts relevant to this paper.

2.1. Best-Performing Methods from ASSIN 2

In this subsection, we briefly describe six of the best performing algorithms developed for ASSIN 2. The model of Rodrigues et al. (2019a) ranked first place in the competition – with a Pearson ρ of 0.826 and an MSE of 0.52 – using a solution based on a pre-trained multilingual version of BERT-Base. To improve the effectiveness, they added one untrained layer of neurons, and then trained the new model using the ASSIN 2 training set along with the Brazilian Portuguese training set of the first ASSIN task (ASSIN 1).

The Stilingue team [Fonseca and Alvarenga 2019] attained the best MSE performance (0.47) and the second-best Pearson correlation score (0.817). Their proposal consists of a wide / deep learn model (based on multilingual BERT-Base and Universal Sentence Encoder-Large multilingual) combined with 18 features that describe lexical,

syntactic and semantic information from the sentences in the dataset (e.g.: jaccard similarity, negation agreement, difference in the amount of tokens between the two sentences, among others).

The method proposed in [Rodrigues et al. 2019b] consists of a stacked ensemble approach that combines the predictions generated by two models: a multilingual BERT model fine-tuned over ASSIN 1 and ASSIN 2 datasets; and a RoBERTa model fine-tuned over the automatic translation of the datasets into English. The method obtained a Pearson ρ of 0.785 (third best) and MSE of 0.59 (fourth best).

Following a different approach, more related to the present work, Santos et al. (2019) developed a traditional machine learning framework based on the evaluation of a collection of lexical, distributional, syntactic, and semantic attributes. Feature selection strategies were employed as preprocessing step to discover the subset of input features most relevant for the STS task. As a result, the original set of 71 attributes was reduced to 12, comprising only lexical and distributional features (surprisingly, none of the syntactic and semantic features were considered as relevant). The technique achieved competitive results, with Pearson ρ and MSE of 0.740 and 0.60, respectively.

Other good-performing methods were proposed by Cabezudo et al. (2019) and de Souza et al. (2019). The first fine-tuned multilingual BERT on ASSIN 2 corpus without any extra feature whilst the later trained a Siamese neural network model using various distinct features, including lexical-based, word2vec embeddings, and also incorporating similarity metrics obtained from a multilingual wordnet.

2.2. Wordnets

A wordnet is a lexical database of a given language [de Paiva et al. 2016, Fellbaum 1998]. In this kind of structure, words are organized into groups of synonymous lexical items, known as *synsets*, which are linked to each other according to their conceptual-semantic relations. Examples of such relations, among others, include "is-a", which links more specific synsets (called hyponyms) to more general ones (called hypernyms) and "antonymy", which indicates semantic opposition between two synsets. Princeton's WordNet (PWN) [Fellbaum 1998], the first wordnet released, was manually created by a multidisciplinary team in the early 1990s having English as its target language. Since then, wordnets in dozens of other languages including Portuguese have been developed and successfully established [de Paiva et al. 2016].

A wordnet can be constructed as a graph where nodes are synsets and edges represent their semantic relations. Figure 1 shows an excerpt from Onto.PT [Gonçalo Oliveira and Gomes 2014] – the Portuguese wordnet used in this study – showing five different synsets and the "is-a" (hyponym-hypernymy) relations between them. In this example, the topmost node represents the general concept {*cardume*, *peixe*, *peixes*} ({shoal, fish, fishes}). Three hyponym (more specific) synsets are linked to this synset, denoting three distinct kinds of fishes: {*truta*, *truite*} (trout fish), {*linguado*} (halibut), and {*bocaréu*, *biquerão*, *manjuba*, *enchova*, *anchova*} (five distinct Portuguese names for anchovy fish). In the same way, {*truta-salmonada*} (salmon trout, i.e., a specific kind of trout) is linked to its hypernym {*truta*, *truite*}. Aside from relations, each synset in the wordnet hold two properties: its part of speech and a dictionary-style definition called gloss (see Figure 2).

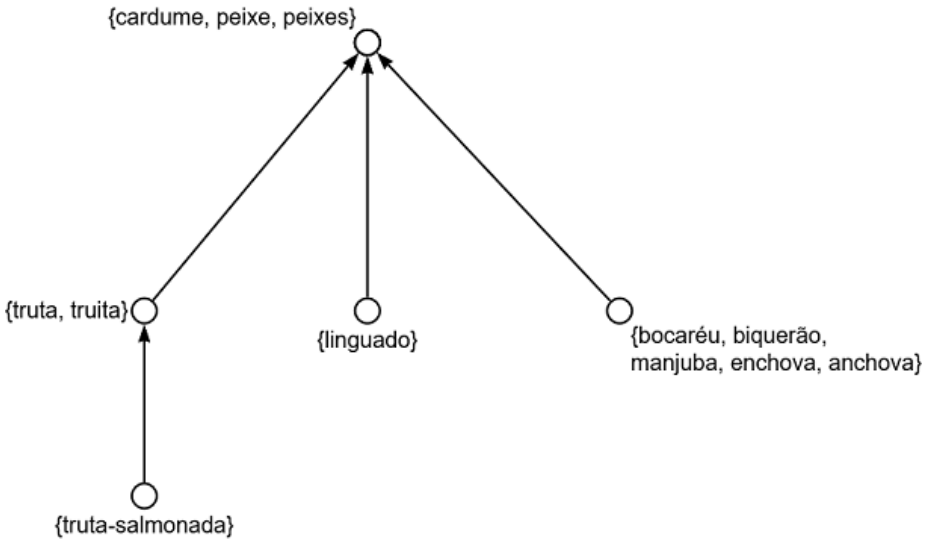

Figure 1. An excerpt from Onto.PT

{**truta-salmonada**} (substantivo): *truta com pintas rosas e carne mais avermelhada*[1].

Figure 2. A synset and its properties

Given a wordnet, there are a few different approaches for measuring the semantic similarity between two words w_1 and w_2 [Pilehvar and Navigli 2015]. Initially, it should be verified if both words belong to the same synset. If this occurs, then there is a sense of w_1 and w_2 in which they are synonymous. Otherwise, the most adopted approach is the one known as *edge counting*. According to this strategy, the similarity between w_1 and w_2 is computed by finding the shortest path between a synset containing w_1 and another synset containing w_2 in the wordnet graph. The less the number of edges in the path, the more semantically similar the words are.

The above approach can straightforwardly be extended to account for the STS task by computing the similarity between two sentences as the average of the similarity between the closest (most similar) word pairs in both sentences. This idea has been applied with good results in [Li et al. 2006, Croft et al. 2013, de Lima and Gonçalves 2022].

3. In Defense of Wordnet-based STS

When it comes to means of incorporating semantic knowledge into NLP algorithms, there are two main approaches: wordnets and word embeddings [Gonçalo Oliveira 2018]. Recently, BERT contextual embeddings [Delvin et al. 2019] and its variants have been preferred over wordnets due to their remarkable performance in a number of distinct NLP tasks – including the STS task [Fonseca and Alvarenga 2019; Rodrigues et al. 2019a; Rodrigues et al. 2019b, Wang et al. 2020].

[1] {**salmon trout**} (noun): species of trout that has pink spots and a redder meat

Although wordnets' disadvantages compared to BERT and other embedding technologies are widely acknowledged, their advantages are rarely mentioned. Nonetheless, wordnets have appealing properties for STS applications. First, as shown in Section 2, wordnets are theoretically simple and intuitive. Aside from this, wordnets like PWN are more formalized since they have been created and maintained by experts who are responsible for grouping synsets and defining relations amongst them [Gonçalo Oliveira et al. 2021].

Second and more importantly, methods based on wordnets can deliver results that are easy to interpret, as their decisions are usually taken by considering the graph topology (i.e., semantically similar concepts are located in nodes that are close to each other). On the other hand, BERT architecture was designed to maximize predictive performance but do not consider the comprehensibility (interpretability) of the model. This is a disadvantage because in certain domains, like public administration [Anthopoulos and Wood 2021; Darrazão et al. 2023; de Lima and Gonçalves 2022], the ability of users to understand relevant aspects of the modeling process is also important or even required due to legal reasons, transparency issues or to follow ethical guidelines [Freitas 2014].

Motivated by these issues, in this paper we explore a lightweight wordnet-based approach to solve the ASSIN 2 STS task. Our approach is based on a small set of five features, where each takes account of either lexical / distributional similarity or semantic similarity between Portuguese sentences. The semantic features are computed with the utilization of Onto.PT [Gonçalo Oliveira and Gomes 2014], the largest Portuguese wordnet. In the next subsection we describe our approach. First, we present the preprocessing steps that were performed before generating the features. Next, the features themselves are described.

3.1. The Proposed Approach

3.1.1 Preprocessing

Two preprocessing steps were carried out in the ASSIN 2 datasets: stop word removal and stemming. In the stop word removal step, pronouns, articles, prepositions, conjunctions, and linking verbs were removed from the sentences. However, adverbs were kept as they can modify the meaning of an entire sentence. The process was done with the use of the NLTK standard Portuguese stoplist [Bird et al. 2009].

Next, we assessed the coverage of Onto.PT with respect to the set of tokens (individual words or unigrams) present in the ASSIN 2 training set. We identified that from a total of 2,253 distinct tokens in the training set, only 63.43% (1,429) could be found as lexical items in Onto.PT. To mitigate this problem, we decided to submit both, the ASSIN 2 collection and Onto.PT to a stemming process employing the RSLP algorithm [Orengo and Huyck 2001]. In the stemming process, suffixes common in the Portuguese language (due to plurals or tenses) are trimmed to reduce any word to its stem. As a result, the number of distinct tokens in the ASSIN 2 training dataset was reduced from 2,253 words to 1,466 stems, where 1,364 (93.04%) could be found in Onto.PT. Table 1 summarizes the results of the stemming process. From now on, the terms stem and token will be used interchangeably in this paper, since in our approach tokens are represented by their stems.

Table 1. Summary of the stemming process

ASSIN 2 training dataset	Number of distinct tokens	Number and percentage of tokens found in Onto.PT
Before stemming	2,253	1,429 (63.43%)
After stemming	1,466	1,364 (93.04%)

3.1.2 Features

In this subsection, we present the set of five features used to build our STS model. In the definitions throughout the text, we adopted the following notation:

- *t* and *h*: the two sentences whose similarity score is to be computed with their words transformed into stems.
- Tok_t and Tok_h: the set of tokens obtained from *t* and *h*, respectively. As aforementioned, tokens are represented by the stems of the words in *t* and *h*.

Semantic Features

Two semantic features were used in our model. Both were computed according to the wordnet-based semantic similarity function presented in Equation (1).

$$F_{wordnet}(t,h) = \frac{|ExtTok_t \cap Tok_s|}{max(|Tok_t|,|Tok_s|)} \quad (1)$$

In Equation (1), $ExtTok_t$ corresponds to Tok_t augmented with additional tokens that are somehow related to each token in Tok_t according to a wordnet. Greater values indicate higher similarity between the sentences. We employed two different approaches to determine $ExtTok_t$, which consequently led to the generation of two different semantic features, named $F_{wordnet\_synonyms}$ and $F_{wordnet\_hypernyms}$:

- $F_{wordnet\_synonyms}$: to compute this feature, $ExtTok_t$ was generated by augmenting Tok_t with the synonyms found in Onto.PT for each of its tokens.
- $F_{wordnet\_hypernyms}$: in this case, $ExtTok_t$ was generated by augmenting Tok_t with the hypernyms found in Onto.PT for each of its tokens.

Lexical and Distributional Features

Following the approach of some of the teams that participated in ASSIN 2 [Fonseca and Alvarenga 2019, Santos et al. 2019, de Souza et al. 2019], we combined the above semantic features with three additional features that explore either lexical or distributional information contained in the sentences. They are described below:

- $F_{TokensRatio}$: corresponds to the ratio of the amount of tokens (stems) in *t* to the amount of tokens in *h*. The rationale is that semantically similar sentences are expected not to have a large difference in their corresponding number of tokens.
- $F_{n\text{-}grams}$: corresponds to the cosine between the character *n*-grams vectors of *t* and *h*. In this work, we chose *n*=5 since our preliminary experiments found character *5*-grams to be more effective than *2*-grams, *3*-grams, and *4*-grams.
- $F_{tf\text{-}idf}$: the cosine of the TF-IDF vectors of the sentences. This distributional feature reflects the importance of each word stem in a sentence.

4. Results

We performed two distinct experiments. The first was carried out on the validation dataset in order to compare the performance of the features when used alone and when combined. I.e.: we created six different regression models, the first five trained with a single feature and the last one trained with the complete set of features ($F_{wordnet\_synonyms}$, $F_{wordnet\_hypernyms}$, $F_{TokensRatio}$, $F_{n\text{-}grams}$, and $F_{tf\text{-}idf}$). The models were trained using the multi-layer perceptron regressor implementation available at scikit-learn [Pedregosa et al. 2011], with default parameters, except for the maximum number of iterations (*max_iter*), which was set to 1,000. Results are shown in Table 2. It is possible to observe that, amongst the models trained with a single feature, those that performed better in the validation dataset were the model trained with $F_{tf\text{-}idf}$ (Pearson ρ of 0.694 and MSE of 0.51) and the one trained with $F_{wordnet\_synonyms}$ (Pearson ρ and MSE of 0.688 and 0.51, respectively). Nonetheless, the regression model trained with the complete set of features achieved superior performance in the validation dataset, with Pearson correlation of 0.730 and MSE of 0.46. It is worth mentioning that we also evaluated combinations of two, three, and four features, but they did not perform as effectively as the model trained with the full set of features.

Next, we conducted a second experiment on the test collection. We compared the performance of our proposed wordnet-based approach against the Pearson ρ and MSE results originally obtained by the nine teams that participated in ASSIN 2 (these results had been previously published in [Real et al. 2019]). The comparison includes the six methods presented in Section 2: ASAPPy [Santos et al. 2019], Deep Learning Brasil [Rodrigues et al. 2019b], IPR [Rodrigues et al. 2019a], NILC [Cabezudo et al. 2019], PUCPR [Souza et al. 2019], and Stilingue [Fonseca and Alvarenga 2019]. Results are presented in Table 3. In this table, the rank obtained by each method in each performance metric (Pearson and MSE) is presented in parenthesis.

Table 2. Preliminary results on the validation set

Feature(s) used to build the model	Pearson	MSE
Tokens Ratio ($F_{TokensRatio}$)	0.354 (5)	0.85 (5)
Character 5-gram ($F_{5\text{-}grams}$)	0.568 (4)	0.66 (4)
TF-IDF ($F_{tf\text{-}idf}$)	0.694 (2)	0.51 (2)
Semantic Feature – Synonyms ($F_{wordnet\_synonyms}$)	0.688 (3)	0.51 (2)
Semantic Feature – Hypernyms ($F_{wordnet\_hypernyms}$)	0.117 (6)	0.96 (6)
Full wordnet-based model – all features combined ($F_{TokensRatio} + F_{5\text{-}gram} + F_{tf\text{-}idf} + F_{wordnet\_synonyms} + F_{wordnet\_hypernyms}$)	**0.730 (1)**	**0.46 (1)**

Table 3. Final results on the test dataset: our method versus ASSIN 2 participants

Method	Pearson	MSE
Wordnet-based model ($F_{TokensRatio} + F_{5\text{-}gram} + F_{tf\text{-}idf} + F_{wordnet\_synonyms} + F_{wordnet\_hypernyms}$)	0.735 (6)	0.52 (2)
ASAPPj	0.652 (9)	0.61 (7)
ASAPPy	0.740 (5)	0.60 (6)
Deep Learning Brasil	0.785 (3)	0.59 (5)
IPR	**0.826** (1)	0.52 (2)
L2F/L2F INESC	0.778 (4)	0.52 (2)
LIACC	0.493 (10)	1.08 (10)
NILC	0.729 (7)	0.64 (8)
PUCPR	0.678 (8)	0.85 (9)
Stilingue	0.817 (2)	**0.47** (1)

Table 3 shows that our method obtained competitive results with state-of-the-art BERT methods in terms of MSE (second best result) even though it was built using basic semantic features (proportion of synonyms and hypernyms) combined with a small set of lexical and distributional features. These results encourage us to continue investigating other, more sophisticated approaches based on wordnets to solve STS tasks.

In what follows, some drawbacks related to the use Onto.PT in STS problems will be discussed. These drawbacks might have been responsible for negative impacts on the effectiveness of the proposed method. First, it is important to state that differently from PWN and several other wordnets, Onto.PT was not handcrafted by experts. Instead, it was built through an automated process of extracting, clustering and connecting terms present in Portuguese dictionaries, thesauri, and wordnets. Consequently, it has limitations and errors. For instance, Gonçalo Oliveira and Gomes (2014) reports that an evaluation by two judges on a random sample of 774 distinct Onto.PT synsets showed that only 73.9 % of those were considered correct by both judges.

Aside from this, other relevant disadvantage of Onto.PT for STS is the fact that 56.82% of its synsets are directly connected to the root node of the graph (i.e., they do not have a hypernym). This characteristic of Onto.PT topology has hindered us from evaluating edge counting algorithms, which is the category of similarity algorithms most commonly adopted by wordnet-based STS systems [Li et al 2006, Croft et al. 2013, de Lima and Gonçalves 2022]. Another important disadvantage is that only 39.48% of synsets in Onto.PT have an associated gloss. Thus, it is not possible to employ gloss-based approaches often used for disambiguating short texts and thus improving the performance of PLN systems [Pilehvar and Navigli 2015].

5. Conclusions and Future Work

This work addressed the Portuguese STS task. We explored a lightweight wordnet-based approach that is suitable for use in domains where not only the effectiveness, but also the interpretability of the model is important. We evaluated our proposed approach on ASSIN 2 collection and achieved an MSE of 0.52 and Pearson Correlation of 0.735.

As future work, we first plan to evaluate other Portuguese and multilingual wordnets [Gonçalo Oliveira 2018; de Paiva et al. 2016]. We also intend to follow an approach similar to Rodrigues et al. (2019b) – one of the best-performing methods from ASSIN 2 – by evaluating the use of PWN over the automatic translation of the ASSIN 2 datasets into English. Since PWN is not prone to the same limitations as Onto.PT and is more complete the other Portuguese wordnets, the translation will allow the evaluation of several distinct similarity measures [Pilehvar and Navigli 2015], including those based on edge counting, along with the incorporation of gloss-based techniques for disambiguation.

Other two topics for future research that seem to deserve special attention are the following: (i) taking into consideration not only individual words (unigrams) as tokens during the STS process, but compound words as well (bigrams and trigrams); (ii) evaluating the performance of different regression algorithms instead of MLP, prioritizing transparent and easily auditable techniques.

References

Agirre, E., Cer, D., Diab, M., and Gonzalez-Agirre, A. (2012). "SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity". In: Proc. of the 6th Intl' Wksp on Semantic Evaluation (SemEval-2012), ACL, p. 385–393.

Anthopoulos, T. and Wood, M. (2021) "Automated coding of Standard Industrial and Occupational Classifications (SIC/SOC)", https://statswiki.unece.org/display/ML/Machine+Learning+Group+2021, June.

Bird, S., Loper, E., and Klein, E. (2009). Natural language processing with python, O'Reilly Media Inc.

Cabezudo, M. A. S., Inácio, M., Rodrigues, A. C., Casanova, E., and de Sousa, R. F. (2019). "NILC at ASSIN 2: Exploring Multilingual Approaches". In: Proc. of the ASSIN2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese co-located with XII Symp. in Inf. and Human Language Technology (STIL), CEUR, p. 49–58.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., and Specia, L. (2017). "SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation". In: Proc. of the 11th Intl' Wksp on Semantic Evaluation (SemEval-2017), ACL, p. 1–14.

Chandrasekaran, D. and Mago, V. (2021). Evolution of semantic similarity: A survey. In *ACM Comput. Surv.*, 54(2), pages 41:1–41:37. ACM.

Croft, D, Coupland, S., Shell, J., Brown, S. (2013) "A Fast and Efficient Semantic Short Text Measure", In: Proc. of the 13rd UK Workshop on Computational Intelligence (UKCI), IEEE, p. 221–227.

Darrazão, E., Amorim, V., Oliveira, K., Gomes-Jr, L. (2023). "Engenharia e Avaliação de Features para Extração de Informação em Notas Fiscais". In: Anais da XVIII Escola Regional de Banco de Dados (ERBD), SBC, p. 80–89.

Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2019). "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding". In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers), ACL, p. 4171–4186.

Fellbaum, C. (1998). WordNet: an electronic lexical database, MIT Press, Cambridge.

Ferrero, J., Besacier, L., Schwab, D., and Agnès, F. (2017). "CompiLIG at SemEval-2017 Task 1: Cross-Language Plagiarism Detection Methods for Semantic Textual Similarity". In: Proc. of the 11th Intl' Wksp on Semantic Evaluation (SemEval-2017), ACL, p. 109–114.

Fonseca, E., and Alvarenga, J. P. R. (2019). "Wide And Deep Transformers Applied to Semantic Relatedness and Textual Entailment". In: Proc. of the ASSIN2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese co-located with XII Symp. in Inf. and Human Language Technology (STIL), CEUR, p. 68–76.

Fonseca, E. R., Borges dos Santos, L., Criscuolo, M., and Aluísio, S. M. (2016). Visão geral da avaliação de similaridade semântica e inferência textual. In *Linguamática*, 8(2), pages 3–13. UMinho / UVigo.

Freitas, A. A. (2014). Comprehensible classification models – A position paper. In *SIGKDD Explorations*, 15(1), pages 1–10. ACM.

Gonçalo Oliveira, H. (2018). Distributional and knowledge-based approaches for computing Portuguese word similarity. In *Information*, 9(35), pages 1–21. MDPI.

Gonçalo Oliveira, H. and Gomes, P. (2014). ECO and Onto.PT: A flexible approach for creating a Portuguese wordnet automatically. In *Language Resources and Evaluation*, 48(2), pages 373–393. Springer.

Gonçalo Oliveira, H., Aguiar, F. S. S., and Rademaker, A. (2021). "On the Utility of Word Embeddings for Enriching OpenWordNet-PT", In: Proc. of the 3rd Conf. on Language, Data and Knowledge (LDK 2021), OASIcs, p. 21:1–21:13.

Li, Y., McLean, D., Bandar, Z. A., O'Shea, J. D., Crockett, K. (2006). Sentence similarity based on semantic nets and corpus statistics. In *IEEE Transactions on Knowledge and Data Engineering*, 18(8), pages 1138–1150. IEEE.

de Lima, L. S. G. and Gonçalves, E. C. (2022). "Similaridade Semântica de Nomes de Produtos Alimentícios Utilizando Wordnets do Português". In: Proc. of the XV Seminar on Ontology Research in Brazil (ONTOBRAS 2022) and VI Doctoral and Masters Consortium on Ontologies (WTDO 2022), CEUR, p. 23–31.

Orengo, V. M. and Huyck, C. (2001). "A Stemming Algorithm for the Portuguese Language". In: Proc. of the 8th Symposium on String Processing and Information Retrieval, IEEE, p. 186–193.

de Paiva, V., Real, L., Gonçalo Oliveira, H., Rademaker, A., Freitas, C., Simões, A. (2016) "An overview of Portuguese WordNets", In: Proc. of the 8th Global WordNet Conference (GWC 2016), ACL, p. 74–81.

Pedregosa et al. (2011). Scikit-learn: Machine learning in python. In *JMLR* 12, pages 2825–2830.

Pilehvar, M. T. and Navigli, R. (2015). From senses to texts: An all-in-one graph-based approach for measuring semantic similarity. In *Artificial Intelligence*, 228, pages 95–128. Elsevier.

Real, L., Fonseca, E., and Gonçalo Oliveira, H. (2019). "Organizing the ASSIN 2 Shared Task". In: Proc. of the ASSIN2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese co-located with XII Symp. in Inf. and Human Language Technology (STIL), CEUR, p. 1–13.

Rodrigues, R., Couto, P., and Rodrigues, I. (2019a). "IPR: The Semantic Textual Similarity and Recognizing Textual Entailment Systems". In: Proc. of the ASSIN2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese co-located with XII Symp. in Inf. and Human Language Technology (STIL), CEUR, p. 39–47.

Rodrigues, R. C., da Silva, J. R., de Castro, P. V. Q., da Silva, N. F. F., Soares, A. S. (2019b). "Multilingual Transformer Ensembles for Portuguese Natural Language Tasks". In: Proc. of the ASSIN2 Shared Task: Evaluating Semantic Textual Similarity

and Textual Entailment in Portuguese co-located with XII Symp. in Inf. and Human Language Technology (STIL), CEUR, p. 27–38.

Santos, J., Alves, A. and Gonçalo Oliveira, H. (2019). “ASAPPpy: a Python Framework for Portuguese STS”. In: Proc. of the ASSIN2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese co-located with XII Symp. in Inf. and Human Language Technology (STIL), CEUR, p. 14–26.

Soares, M. A. C. and Parreiras, F. S. (2020). A literature review on question answering techniques, paradigms and systems. In *Journal of King Saud University - Computer and Information Sciences*, 32(6), pages 635–646. Elsevier.

de Souza, J. V. A., Oliveira, L. E. S., Gumiel, Y. B., Carvalho, D. R., Moro, C. M. C. (2019). “Incorporating Multiple Feature Groups to a Siamese Neural Network for Semantic Textual Similarity Task in Portuguese Texts”. In: Proc. of the ASSIN2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese co-located with XII Symp. in Inf. and Human Language Technology (STIL), CEUR, p. 59–68.

Wang, Y., Fu, S., Shen, F., Henry, S., Uzuner, O., and Liu, H. (2020). Overview of the 2019 n2c2/OHNLP Track on Clinical Semantic Textual Similarity. In *JMIR Med Inform.*, 8(11):e23375. JMIR.