Programa Interdisciplinar de Pós-Graduação em

Computação Aplicada

Mestrado Acadêmico

Allan de Barcelos Silva

O USO DE RECURSOS LINGUÍSTICOS PARA MENSURAR A SEMELHANÇA SEMÂNTICA ENTRE FRASES CURTAS ATRAVÉS DE UMA ABORDAGEM HÍBRIDA

São Leopoldo, 2017

UNIVERSIDADE DO VALE DO RIO DOS SINOS — UNISINOS
UNIDADE ACADÊMICA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA
NÍVEL MESTRADO

ALLAN DE BARCELOS SILVA

O USO DE RECURSOS LINGUÍSTICOS PARA MENSURAR A SEMELHANÇA SEMÂNTICA ENTRE FRASES CURTAS ATRAVÉS DE UMA ABORDAGEM HÍBRIDA

SÃO LEOPOLDO
2017

Allan de Barcelos Silva

O USO DE RECURSOS LINGUÍSTICOS PARA MENSURAR A SEMELHANÇA SEMÂNTICA ENTRE FRASES CURTAS ATRAVÉS DE UMA ABORDAGEM HÍBRIDA

Dissertação apresentada como requisito parcial para a obtenção do título de Mestre pelo Programa de Pós-Graduação em Computação Aplicada da Universidade do Vale do Rio dos Sinos — UNISINOS

Orientador:
Prof. Dr. Sandro J. Rigo

Co-orientador:
Prof[a]. Dr. Isa M. Alves

São Leopoldo
2017

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)

Silva, Allan de Barcelos

O uso de recursos linguísticos para mensurar a semelhança semântica entre frases curtas através de uma abordagem híbrida / Allan de Barcelos Silva — 2017.

87 f.: il.; 30 cm.

Dissertação (mestrado) — Universidade do Vale do Rio dos Sinos, Programa de Pós-Graduação em Computação Aplicada, São Leopoldo, 2017.

"Orientador: Prof. Dr. Sandro J. Rigo, Unidade Acadêmica de Pesquisa e Pós-Graduação".

1. Processamento de Linguagem Natural. 2. Similaridade Semântica Textual. 3. Linguística. 4. Aprendizagem de Máquina. 5. *Support Vector Machines*. 6. *Word embeddings*. 7. *Principal Component Analysis*. I. Título.

CDU 004.4'414

Bibliotecária responsável: Vanessa Borges Nunes — CRB 10/1556

Allan de Barcelos Silva

O USO DE RECURSOS LINGUÍSTICOS PARA MENSURAR A SEMELHANÇA SEMÂNTICA ENTRE FRASES CURTAS ATRAVÉS DE UMA ABORDAGEM HÍBRIDA

Dissertação apresentada à Universidade do Vale do Rio dos Sinos – Unisinos, como requisito parcial para obtenção do título de Mestre em Computação Aplicada.

Aprovado em 14 de dezembro de 2017

BANCA EXAMINADORA

Isa Mara da Rosa Alves – Universidade do Vale do Rio dos Sinos

Thiago Alexandre Salgueiro Pardo – Universidade de São Paulo

José Vicente Canto dos Santos – Universidade do Vale do Rio dos Sinos

Prof. Dr. Sandro José Rigo (Orientador)

Visto e permitida a impressão
São Leopoldo, 14 de fevereiro de 2018.

Prof. Dr. Rodrigo da Rosa Righi
Coordenador PPG em Computação Aplicada

Aos nossos pais e filhos.

Pour examiner la vérité il est besoin, une fois dans sa vie,
de mettre toutes choses en doute autant qu'il se peut.
— RENÉ DESCARTES

AGRADECIMENTOS

Para a realização deste trabalho, algumas pessoas com quem convivi foram essenciais de diferentes maneiras. Sendo assim, presto meus mais sinceros agradecimentos:

Ao professor e amigo Dr. Sandro José Rigo, orientador deste trabalho, por toda dedicação, atenção, oportunidades, ensinamentos, orientações e conhecimentos, que foram, e sempre serão fundamentais para meu crescimento pessoal, profissional e acadêmico;

Aos meus colegas de curso que, diretamente ou indiretamente, contribuíram para a realização deste trabalho com apoio. Em especial aos amigos Luís, Márcio, Anderson e Roger, pelos conhecimentos compartilhados e companheirismo;

A minha esposa, Tayane Ligia R. Muniz, com quem pude contar com apoio, compreensão e por estar ao meu lado sempre durante o desenvolvimento desse trabalho, mesmo quando não foi possível lhe dar atenção.

E aos meus pais que sempre estiveram ao meu lado, que mesmo com muito esforço reconhecido, me propiciaram um ambiente e condições de acesso a educação em todos os períodos de minha vida. Além da compreensão e apoio tanto em questões financeiras quanto familiares.

"*As crenças são muitas vezes os mais perigosos inimigos da verdade.*".
(Friedrich Nietzsche)

RESUMO

Na área de Processamento de Linguagem Natural, a avaliação da similaridade semântica textual é considerada como um elemento importante para a construção de recursos em diversas frentes de trabalho, tais como a recuperação de informações, a classificação de textos, o agrupamento de documentos, as aplicações de tradução, a interação através de diálogos, entre outras. A literatura da área descreve aplicações e técnicas voltadas, em grande parte, para a língua inglesa. Além disso, observa-se o uso prioritário de recursos probabilísticos, enquanto os aspectos linguísticos são utilizados de forma incipiente. Trabalhos na área destacam que a linguística possui um papel fundamental na avaliação de similaridade semântica textual, justamente por ampliar o potencial dos métodos exclusivamente probabilísticos e evitar algumas de suas falhas, que em boa medida são resultado da falta de tratamento mais aprofundado de aspectos da língua. Este contexto é potencializado no tratamento de frases curtas, que consistem no maior campo de utilização das técnicas de similaridade semântica textual, pois este tipo de sentença é composto por um conjunto reduzido de informações, diminuindo assim a capacidade de tratamento probabilístico eficiente. Logo, considera-se vital a identificação e aplicação de recursos a partir do estudo mais aprofundado da língua para melhor compreensão dos aspectos que definem a similaridade entre sentenças.

O presente trabalho apresenta uma abordagem para avaliação da similaridade semântica textual em frases curtas no idioma português brasileiro. O principal diferencial apresentado é o uso de uma abordagem híbrida, na qual tanto os recursos de representação distribuída como os aspectos léxicos e linguísticos são utilizados. Para a consolidação do estudo, foi definida uma metodologia que permite a análise de diversas combinações de recursos, possibilitando a avaliação dos ganhos que são introduzidos com a ampliação de aspectos linguísticos e também através de sua combinação com o conhecimento gerado por outras técnicas. A abordagem proposta foi avaliada com relação a conjuntos de dados conhecidos na literatura (evento PROPOR 2016) e obteve bons resultados.

Palavras-chave: Processamento de Linguagem Natural. Similaridade Semântica Textual. Linguística. Aprendizagem de Máquina. *Support Vector Machines*. *Word embeddings*. *Principal Component Analysis*.

ABSTRACT

One of the areas of Natural language processing (NLP), the task of assessing the Semantic Textual Similarity (STS) is one of the challenges in NLP and comes playing an increasingly important role in related applications. The STS is a fundamental part of techniques and approaches in several areas, such as information retrieval, text classification, document clustering, applications in the areas of translation, check for duplicates and others.

The literature describes the experimentation with almost exclusive application in the English language, in addition to the priority use of probabilistic resources, exploring the linguistic ones in an incipient way. Since the linguistic plays a fundamental role in the analysis of semantic textual similarity between short sentences, because exclusively probabilistic works fails in some way (e.g. identification of far or close related sentences, anaphora) due to lack of understanding of the language. This fact stems from the few non-linguistic information in short sentences. Therefore, it is vital to identify and apply linguistic resources for better understand what make two or more sentences similar or not.

The current work presents a hybrid approach, in which are used both of distributed, lexical and linguistic aspects for an evaluation of semantic textual similarity between short sentences in Brazilian Portuguese. We evaluated proposed approach with well-known and respected datasets in the literature (PROPOR 2016) and obtained good results.

Keywords: Natural Language Processing. Semantic Textual Similarity. Linguistic. Machine Learning. Support Vector Machines. Word embeddings. Principal Component Analysis.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

BOW	Bag of Words
CBOW	Continuous Bag of Words
CP	Correlação de Pearson
EQM	Erro Quadrático Médio
GLM	Generalized Linear Models
GloVe	Global Vectors for Word Representation
LSTM	Long-Short Term Memory Networks
MEV	Modelo de Espaço Vetorial
PCA	Principal Component Analysis
PLN	Processamento de Linguagem Natural
REM	Reconhecimento de Entidades Mencionadas
RNN	Recurrent Neural Networks
RNA	Redes Neurais Artificiais
STS	Semantic Textual Similarity
SG	Skip-gram
SVM	Support Vector Machines
TF-IDF	Term Frequency - Inverse Document Frequency

SUMÁRIO

1 INTRODUÇÃO

O Processamento de Linguagem Natural (PLN) é um ramo da Inteligência Artificial que faz uso de conceitos linguísticos para tratamento de informações, significados, elementos de textos e fala. (KAO; POTEET, 2007). A análise de similaridade semântica textual desempenha um papel cada vez mais importante nas pesquisas e aplicações relacionadas com a área de PLN, tais como a recuperação de informações, a classificação de texto, o agrupamento de documentos, entre outros. A capacidade de identificar a similaridade entre as palavras e textos é fundamental para muitas destas tarefas. (GOMAA; FAHMY, 2013; FREIRE; PINHEIRO; FEITOSA, 2016).

Boa parte dos métodos atuais para a avaliação de similaridade semântica textual, baseia-se principalmente na identificação de semelhança através das palavras compartilhadas entre as sentenças comparadas, representando-as como um vetor de termos e, desta forma, restringindo a análise apenas à informação sintática disponível. Ao adotar esta abordagem são desconsiderados, por exemplo, aspectos como a ordem das palavras e o significado das sentenças como um todo (FERREIRA et al., 2016). Estes problemas são relatados na literatura, indicando como as abordagens com foco predominante na sintaxe falham quando as sentenças não possuem termos em comum, devido à incompatibilidade do vocabulário. Quando isso ocorre, é atribuída uma pontuação de similaridade muito baixa, independente do quão relacionadas as sentenças sejam em seu significado. Além disto, o contexto também pode ser um problema para a avaliação de similaridade, pois enquanto documentos possuem um volume de texto que permite inferir o contexto de um termo, as frases curtas são limitadas neste aspecto (METZLER; DUMAIS; MEEK, 2007).

Independentemente do tamanho das sentenças, as três principais formas de se avaliar a similaridade dentre estas decorrem de análises léxicas, sintáticas ou semânticas. (GOMAA; FAHMY, 2013). No primeiro caso, é verificado se os conjuntos dos termos possuem a mesma sequência de caracteres. Já na segunda forma, avalia-se a sintaxe das frases, através da análise das classes gramaticais de cada palavra e também da estrutura sentencial. Na análise semântica, é avaliado o pertencimento dos termos à um mesmo contexto ou a um mesmo significado. (PRADHAN; GYANCHANDANI; WADHVANI, 2015). Para apoiar estas análises são utilizados tanto recursos ligados aos aspectos linguísticos, como técnicas com base probabilística.

É comum na literatura da área de PLN o emprego de técnicas que utilizam o WordNet (FELLBAUM, 1998), o FrameNet (BAKER; FILLMORE; LOWE, 1998) e o VerbNet (SCHULER, 2005) devido às relações linguísticas mapeadas nestes recursos, bem como pela possibilidade da aplicação destas no processamento da língua. De forma complementar, mais recentemente os Modelos de Espaço Vetorial (MEV) vem motivando estudos devido sua abordagem probabilística, independência de domínio e capacidade de obter automaticamente as relações semânticas entre textos, dado um espaço de contextos.

No presente trabalho, utilizou-se um conjunto de técnicas linguísticas e probabilísticas para melhor representar as sentenças, de modo a maximizar a assertividade da classificação de si-

milaridade entre pares de frases. Para tanto, utilizou-se as relações linguísticas de antonímia, hiperonímia, hiponímia e sinonímia, exploradas através das bases TeP (MAZIERO et al., 2008) e PULO (SIMÕES; GUINOVART, 2014), além dos recursos probabilísticos obtidos com uso de MEVs, com a métrica de TF-IDF e com análise de componentes principais.

1.1 Contextualização e problema

Atualmente o estado da arte para avaliação de similaridade textual vem recebendo bastante atenção (FERREIRA et al., 2016; AGIRRE et al., 2017; HARTMANN, 2016; BARBOSA et al., 2016; FREIRE; PINHEIRO; FEITOSA, 2016). Além disto, eventos como SemEval[1] e PROPOR[2], que possuem tarefas específicas para mensurar a similaridade semântica entre sentenças, estão adquirindo maior popularidade e estimulando o desenvolvimento de uma série de outras aplicações.

Conforme Gomaa e Fahmy (2013), a identificação de similaridade entre frases e textos é uma parte fundamental para muitas tarefas em PLN. A necessidade de desenvolvimento de técnicas e abordagens é um fato, exemplificado, já desde Agirre et al. (2012), quando o autor propõe um dos primeiros desafios públicos para avaliação de similaridade semântica entre textos escritos em inglês. Desde então, a tarefa foi introduzida nos demais eventos (AGIRRE et al., 2014, 2015, 2016, 2017). No caso da língua portuguesa brasileira, um destes desafios foi proposto na tarefa de Avaliação de Similaridade Semântica e Inferência Textual (ASSIN).

Parte significativa dos trabalhos na literatura atua com o objetivo de encontrar a similaridade semântica entre textos em inglês, como é visto em Ferreira et al. (2016), Brychcín e Svoboda (2016) e Kashyap et al. (2016). Além disto, muitos destes abordam o problema utilizando exclusivamente técnicas de aprendizagem de máquina ou então recursos linguísticos, de modo que poucos são aqueles que combinam as duas abordagens. Dentro do contexto de similaridade textual na língua brasileira, este número torna-se ainda menor.

Um dos trabalhos de interesse, neste sentido, voltado para português do Brasil é o de Hartmann (2016), pois a abordagem do autor utiliza aprendizagem de máquina, medidas de distâncias e relações de sinonímia para identificação da similaridade semântica entre pares de frases curtas. Outros trabalhos como os de Barbosa et al. (2016) e Freire, Pinheiro e Feitosa (2016), destacam que conseguiram bons resultados em obter a similaridade utilizando abordagens diferenciadas com uso de recursos simbólicos ou probabilísticos.

Portanto, observa-se que boa parte dos métodos atuais para esta tarefa utilizam prioritariamente a similaridade entre as palavras, representando-as como um vetor de termos. Além disso, uma parte significativa dos trabalhos restringem suas análises e utilizam-se de poucos recursos linguísticos.

[1] *International Workshop on Semantic Evaluation*
[2] *International Conference on the Computational Processing of Portuguese*

1.2 Questão de pesquisa

Diante do contexto apresentado, foi elaborada a seguinte questão de pesquisa:

> A similaridade semântica textual, em português do Brasil, avaliada através de uma abordagem híbrida, a qual faz uso de recursos probabilísticos e linguísticos, pode apresentar melhores resultados do que uma abordagem apenas probabilística? E se for o caso, qual a influência de recursos linguísticos como antonímia, hiperonímia, hiponímia e sinonímia para o seu desempenho?

1.3 Objetivos

O objetivo geral deste trabalho é propor e avaliar o uso de uma abordagem híbrida para a tarefa de mensurar a similaridade semântica textual entre frases curtas em português do Brasil, através do uso de recursos linguísticos e probabilísticos. Para isto, alguns objetivos específicos foram identificados, sendo vistos como pré-requisitos para se alcançar o objetivo geral:

- Identificar os principais recursos linguísticos que podem contribuir para o estudo;
- Propor um *corpus* para utilização na geração de um Modelo de Espaço Vetorial;
- Obter o modelo de espaço vetorial que será utilizado na abordagem proposta;
- Propor uma abordagem para avaliação de similaridade semântica, entre frases curtas, para português do Brasil;
- Implementar e avaliar o desempenho da abordagem proposta, através da comparação dos resultados entre as diversas técnicas escolhidas.

1.4 Metodologia proposta

Analisando os fatos descritos nesta seção, foi encontrada uma lacuna no que tange a língua de avaliação da similaridade semântica entre frases curtas, bem como na quantidade, qualidade e complexidade das relações linguísticas utilizadas pelos trabalhos. Será realizada uma pesquisa explicativa (KOCHE, 2011) na qual inicialmente é estudado e descrito com o suficiente detalhe um determinado cenário no qual se identifica o problema a ser tratado. A partir deste ponto de partida, serão analisadas e demonstradas possibilidades para uma solução, considerando o estudo de um caso específico. Para compor o método de pesquisa, foram definidas sete etapas com o objetivo de avaliar a técnica proposta, sendo estas as seguintes:

I. Realizar um estudo com profissionais da linguística para identificar os principais recursos léxicos, sintáticos e ou semânticos que podem contribuir para o estudo;

II. Identificar trabalhos relacionados, que utilizam uma abordagem híbrida, aplicado em português do Brasil, e descrever suas técnicas;

III. Identificar e avaliar os principais algoritmos de aprendizado de máquina empregados no estado da arte, que serão utilizados na composição da abordagem proposta;

IV. Realizar o processamento do *corpus* elaborado e obter o Modelo de Espaço Vetorial que será utilizado como um dos recursos probabilísticos para avaliação das frases curtas;

V. Extrair as informações léxicas, sintáticas, semânticas e probabilísticas das frases curtas;

VI. Realizar experimentos com os recursos linguísticos e algoritmos de aprendizagem de máquina escolhidos;

VII. Comparar os resultados dos experimentos com o estado da arte.

A primeira etapa consistiu em buscar, através de consultoria com um grupo de pesquisadores do Programa de Pós-Graduação em Linguística Aplicada da Universidade do Vale do Rio dos Sinos, os recursos descritos na literatura que podem ser utilizados para inferir similaridade entre sentenças. Após, buscou-se identificar os trabalhos relacionados, com aplicação para português do Brasil e suas principais abordagens.

A terceira etapa consistiu em identificar e avaliar os principais algoritmos de aprendizagem de máquina, que são descritos na literatura da área. A próxima etapa foi responsável por obter e realizar o processamento do *corpus*[3] gerado, além de utilizar-se deste para treinamento do MEV. Na sequência, serão obtidos os atributos léxicos, sintáticos, semânticos e probabilísticos dos pares de sentenças, pois através destes será composta a abordagem proposta e avaliadas as contribuições de cada atributo para com a similaridade do par de sentenças.

Na sexta etapa foram realizados diversos experimentos para buscar o melhor conjunto de atributos que descreva a similaridade entre o par de sentenças. Por fim, foram realizadas comparações entre os resultados obtidos e as contribuições dos recursos (probabilísticos e linguísticos) utilizados.

1.5 Organização do texto

O presente trabalho está organizado da seguinte maneira: no Capítulo 2 são apresentados aspectos gerais dos principais assuntos abordados. O Capítulo 3 consiste da descrição dos trabalhos relacionados que contribuíram para o este trabalho. No Capítulo 4 é apresentada a técnica proposta para resolver o problema em questão. Na sequência, o Capítulo 5 aborda os experimentos e resultados obtidos com o uso da abordagem descrita neste trabalho. Por fim, os trabalhos futuros e as conclusões desse trabalho são abordados no Capítulo 6.

[3]Coleção de documentos. (MANNING; SCHÜTZE, 2000).

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo é apresentado o embasamento teórico que foi necessário para a realização deste trabalho, de modo a proporcionar o melhor entendimento e justificativa para a abordagem proposta. Na seção 2.1 serão apresentadas as informações acerca da linguística e suas contribuições para com o atual trabalho, bem como os recursos da língua que foram explorados na abordagem proposta. A seção 2.2 apresenta a área de Processamento de Linguagem Natural, além das técnicas léxicas, semânticas e probabilísticas empregadas no trabalho. Na sequência, é apresentado na Seção 2.3 a área de aprendizagem de máquina e os algoritmos utilizados na avaliação de similaridade textual.

2.1 Linguística

No início do século XX, a linguística passou a ser reconhecida como estudo científico após a divulgação de trabalhos da Universidade de Genebra. Seu principal objetivo é a investigação científica das línguas naturais, que são a forma de comunicação mais altamente desenvolvida e de maior uso. Embora se ocupe da expressão escrita, a linguística também compreende estudo da língua falada. (FIORIN, 2010).

De acordo com Fiorin (2010):

> Uma pintura, uma dança, um gesto podem expressar, mesmo que sob formas diversas, um mesmo conteúdo básico, mas só a linguagem verbal é capaz de traduzir com maior eficiência qualquer um desses sistemas semióticos. As línguas naturais situam-se numa posição de destaque entre os sistemas sígnicos porque possuem, entre outras, as propriedades de flexibilidade e adaptabilidade, que permitem expressar conteúdos bastante diversificados: emoções, sentimentos, ordens, perguntas, afirmações, como também possibilitam falar do presente, passado ou futuro.

Utilizaremos neste trabalho a descrição da linguística proposta por Cançado (2012):

> A descrição da linguística tem diferentes níveis de análise: o léxico, que é o conjunto de palavras da língua; a fonologia, que é o estudo dos sons de uma língua e de como esses sons se combinam para formar as palavras; a morfologia, que é o estudo das construções e palavras; a sintaxe, que é o estudo de como as palavras podem ser combinadas em sentenças; e a semântica, que é o estudo do significado das palavras e sentenças.

Conforme destacado por Oliveira et al. (2015); Alves, Rodrigues e Oliveira (2016), o emprego de recursos de linguísticos é essencial para a compreensão das sentenças. Deste modo, ao longo dos estudos desenvolvidos com um grupo de pesquisadores do Programa de Pós-

Graduação em Linguística Aplicada da Universidade do Vale do Rio dos Sinos, foi identificado que as relações de antonímia, hiperonímia, hiponímia e sinonímia continham um maior potencial para colaborar na identificação da similaridade entre as sentenças e por isso deveriam ser incorporadas no trabalho atual.

2.1.1 Antonímia e sinonímia

Duas facetas bastante conhecidas das relações semânticas são a antonímia e sinonímia. O conceito de antonímia é baseado na propriedade de dois lexemas possuírem uma oposição de sentidos e serem suscetíveis à gradação. (LYONS, 1970, 1977). De acordo com Fellbaum (1998, p. 48), essa relação pode ser considerada a relação semântica mais (básica) importante entre os termos. No entanto, Lyons (1977, p. 32) afirma que há uma falta de entendimento no que diz respeito às relações de oposição.

Segundo Lyons (1970), no que tange os estudos da antonímia, muitos semanticistas tendem a tratar dessa relação do mesmo modo que tratam as relações de sinonímia; o que, de acordo com Palmer (1977), seria um problema por serem duas relações que possuem conceitos diferentes. Nesse sentido, Lyons (1977, p. 149) define a relação, que referimos como antonímia, como sendo uma relação estrutural básica, paradigmática e léxico-semântica em todas as línguas.

Leech (1981) trata das relações de oposição através de uma reflexão sobre a conceitualização dessas relações e problematiza relações de antonímia popularmente conhecidas tais como homem e mulher. Ainda, o autor afirma que, quando pensamos no antônimo de mulher, se considerarmos que tal relação de oposição se apega ao aspecto do significado do termo, a resposta poderia ser tanto menina quanto homem. Para Lyons (1970, 1977), a antonímia pode ser subdividida da seguinte forma: relações de oposição binárias e não-binárias. Assim, categorizando estas em:

i **complementaridade** - a afirmação de um termo anula a afirmação do outro;

ii **antonímia** - elementos opostos gradualmente;

iii **reciprocidade** - elementos que se complementam;

iv **direcionalidade** - (contrários ortogonais e contrários antipodais) - elementos opostos que tomam um ponto como referencial;

v **múltipla incompatibilidade** - elementos que geram contrariedade.

Dentre vários dos semanticistas que tratam do tema antonímia, Lyons traz uma maior e mais substanciosa contribuição para a conceitualização desta. Isso, pois o autor, em comparação aos outros, é o único que trata da distinção entre as relações de oposição e de contraste.

Já no que tange a relação de sinonímia, Lyons (1970, p. 348) define-a como sendo "uma relação de equivalência de sentido matemático do termo"; ou seja, "expressões com o mesmo

significado". (LYONS, 1995, p. 60). Ainda, Lyons (1992, p. 447) afirma que sinônimos são lexemas que "podem se substituir em qualquer contexto dado, sem a mínima mudança no aspecto cognitivo ou emotivo". No entanto, autores como Palmer (1981, p. 89), afirmam que sinônimos como os descritos por Lyons (1992) são inexistentes, visto que "não existem duas palavras com o mesmo significado".

Para Cruse (1986, p. 267), sinônimos são "itens lexicais cujo sentidos são idênticos no que diz respeito a seus traços semânticos "centrais", mas diferem, se for o caso, no que diz respeito ao que pode ser descrito como traços "menores"ou "periféricos". O autor afirma que "um sinônimo é frequentemente empregado como uma explicação ou esclarecimento do significado de outra palavra". (CRUSE, 1986, p. 267).

De acordo com Cruse (1986, p. 265) existe uma escala de sinonímia. Para o autor "alguns sinônimos são mais sinônimos que outros". Nesse sentido, portanto, há três tipos distintos de sinônimos:

i **Sinonímia Total (Sinonímia Absoluta)** - Cruse (1986, p. 270) afirma que esse tipo de sinônimo é raro, visto que "línguas naturais abominam sinônimos absolutos". No entanto, o autor não descarta a existência desta classe de sinônimos e afirma que existem diferentes graus de sinonímia;

ii **Sinonímia Cognitiva (Sinonímia Proposicional)** - Para Cruse (1995, p. 158) "se dois itens lexicais forem sinônimos proposicionais, eles podem ser substituídos em qualquer expressão com propriedades condicionais de verdade sem efeito sobre essas propriedades";

iii **Quase-sinonímia (Parassinonímia)** - Para Lyons (1995, p. 60), quase-sinônimos são "expressões que são mais ou menos similares, mas não são idênticas em significado". De acordo com Barros (2004):

> Os parassinônimos são termos que podem ser considerados como tendo o mesmo sentido, mas cuja distribuição não é exatamente equivalente. O conceito de parassinonímia se distingue, assim, da sinonímia, que recobre os termos tendo o mesmo sentido e a mesma distribuição, isto é, são comutáveis em todos os contextos e em todas as situações. Como não existem sinônimos perfeitos, é preferível falar de parassinônimos ou de sinônimos em discurso.

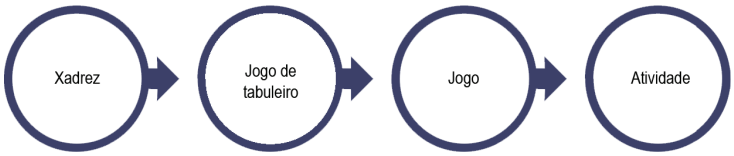
2.1.2 Hiperonímia e hiponímia

A hiponímia é uma relação linguística que estrutura o léxico das línguas em classes, onde o sentido de uma palavra está incluído em outra. (CANÇADO, 2012). Observando a Figura 1 (elaborada com uso da DBpedia[1]), o item lexical mais específico, o qual permanece à esquerda

[1]Conteúdo estruturado extraído a partir da Wikipédia, disponível em `http://www.dbpedia.pt/`.

e contém todas as propriedades das classes seguintes é chamado de hipônimo. Enquanto que o item à direita, ou seja, o que está contido nos demais itens da cadeia, é o hiperônimo. De acordo com Cançado (2012), a relação de hiponímia é assimétrica, pois as propriedades do hiperônimo estão contidas no hipônimo, mas o inverso não é verdadeiro.

Figura 1: Modelo da relação de hiperonímia e hiponímia extraídos da DBpedia



Fonte: Elaborado pelos autores.

Deste modo é possível afirmar que **Xadrez** é hipônimo de **Jogo**, assim como **Atividade** é hiperônimo de **Xadrez**. Logo, todo **Xadrez** é uma **Atividade**, porém o inverso não é válido.

2.1.3 Similaridade textual

A similaridade pode ser tomada como um critério para a identificação de diferentes propriedades semânticas. Sob o ponto de vista onomasiológico e o pensamento de relações paradigmáticas, o fenômeno típico que evidencia a semelhança é a sinonímia, que reflete a construção da máxima identidade semântica entre dois itens lexicais distintos. A relação da hiponímia também é comumente vista como um fator que evidencia similaridade de algum tipo. Na perspectiva semasiológica, a polissemia é o fenômeno que está diretamente relacionado à semelhança. A identificação da semelhança entre os significados associados ao mesmo item léxico ou à mesma categoria lexical é considerada como o principal critério para caracterizar a polissemia. Em ambos os casos, a explicação para a semelhança, em sua quase totalidade, gira em torno da noção de metáfora de Cruse (1995). Isso, pois o autor baseia seus estudos em Lakoff, que diz que “a essência da metáfora é entender e experimentar um tipo de coisa em termos de outra”. (LAKOFF e JOHNSON, 2007, p. 41). Ou seja, como afirma Cruse (1986, p. 41), a "metáfora induz o ouvinte (ou leitor) a ver algo, objetos, e qualquer outra coisa como outra". Desse modo, ela induz o leitor ou o ouvinte a buscar um sentido conotativo por trás daquele (outro) uso da palavra.

Mais elementos para a caracterização da similaridade encontramos ao olharmos para o fenômeno como um princípio cognitivo responsável pela construção de aproximações entre diferentes entidades. Nesse contexto, podemos explicar a similaridade em termos dos referidos princípios Gestalt, que explicam os mecanismos perceptuais inconscientes responsáveis pela construção de ‘todos’ ou de ‘gestalts’ a partir do processamento de inputs que são incompletos. (EVANS; GREEN, 2006). De acordo com este critério, podemos encontrar (se houver) um elemento unificador relevante para a interpretação dos sentidos em comparação. Tal associação pode ocorrer em diferentes termos: com base em fatores objetivos e ou fatores subjetivos. Os fatores objetivos assumem que as entidades percebidas como semelhantes em uma cena

compartilham características físicas, como tamanho, forma ou cor, e que são percebidas como pertencentes a um grupo. (EVANS; GREEN, 2006). Olhando para os sentidos, esse tipo de associação exige que as palavras denotem entidades similares objetivamente. (HIRSCH, 1996).

Além de variar em termos de objetividade, subjetividade e intensidade, a similaridade pode ser construída com menor ou maior grau de linearidade ou regularidade. A partir dessas observações, entendemos que olhar para a noção de semelhança como um princípio cognitivo responsável pela construção de relações sensoriais de diferentes tipos ajuda a descobrir de forma menos intuitiva o que se pretende comunicar ao reconhecer a existência da similaridade entre dois sentidos.

2.1.4 Contribuição da linguística no trabalho proposto

A linguística possui um papel fundamental a análise de similaridade entre frases curtas, pois apesar de existirem trabalhos exclusivamente probabilísticos na literatura, são observadas falhas em alguns pontos devido a falta de compreensão da língua (e.g. identificação de sentenças muito ou pouco similares, anáforas). Tal fato decorre da pequena quantidade de informações não linguísticas existentes nas frases curtas. Logo, é vital a identificação e aplicação de recursos do estudo das línguas para melhor compreensão do que difere ou torna similar duas ou mais sentenças.

Outros autores (OLIVEIRA et al., 2015; ALVES; RODRIGUES; OLIVEIRA, 2016) também apontam que o emprego de recursos linguísticos é essencial para a compreensão das sentenças. Podemos propor verificar a utilidade do entendimento da língua através dos exemplos:

O mercado possui muitos itens guardados.
Existem muitos produtos em estoque no mercado.

Considerando as sentenças acima, suas representações sintáticas seriam:

artigo + substantivo + verbo + advérbio + substantivo + adjetivo.
verbo + advérbio + substantivo + verbo + substantivo + substantivo

Em uma análise semântica das sentenças de exemplo, onde seriam avaliados os contextos e significados destas, podemos representa-las como:

entidade + ação + intensidade + alvo + local
ação + intensidade + alvo + local + entidade

Por fim, a representação de ambas as sentenças em vetores de TF-IDF (Seção 2.2.1), somente faria a avaliação de termos presentes em ambas, permitindo a representação como:

0 1 0 1 0 0
0 1 0 0 0 0 1

Como é possível observar nos exemplos acima, através das análises léxicas, sintáticas e semânticas fica evidenciada a similaridade entre as sentenças, pois trata-se da mesma entidade,

com a mesma intensidade e a mesma ação.

O mesmo não é possível perceber em uma abordagem probabilística mais simplista, onde os vetores de representação dos termos apresentam-se de forma completamente diferente. Apesar dos recentes avanços nas áreas de PLN e inteligência artificial alcançados (MIKOLOV et al., 2013a; PENNINGTON; SOCHER; MANNING, 2014), é possível apenas obter uma aproximação através de termos comuns ou contextos probabilísticos das sentenças, o que por sua vez destaca ainda mais a necessidade de entender as relações entre os termos e a dependência do vocabulário compartilhado entre as frases. Logo, como foi observado acima, a utilização de relações linguísticas pode auxiliar tanto na aproximação quanto no distanciamento de termos através da identificação e processamento das relações entre as palavras das frases, o que por sua vez possibilita melhores resultados para as técnicas probabilísticas.

2.2 Processamento de linguagem natural

A área de Processamento de Linguagem Natural teve seu início em meados de 1950 como a interseção das áreas de linguística e inteligência artificial. Inicialmente foi diferenciada da área de extração de informação (EI), pois esta empregava alto percentual de técnicas baseadas em estatística. Foi somente no ano de 1980 que ocorre a reorientação do PLN para utilização de métodos estatísticos, tornando assim a área muito próxima a EI e consequentemente levando a fusão de ambas para a criação do PLN como é conhecido. (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011).

Atualmente, o PLN é um ramo de aplicação da área de inteligência artificial que faz uso de conceitos linguísticos para extração de informação, significados e representações de textos e fala. Em sua utilização, é comum o emprego de várias representações de conhecimento, como léxico de palavras, regras gramaticais, ontologias, sinônimos, hipônimos, hiperônimos e outros. (KAO; POTEET, 2007).

2.2.1 Term frequency - inverse document frequency

A técnica *Term Frequency - Inverse Document Frequency* (TF-IDF) foi originalmente proposta por Jones (2004), com o objetivo fornecer uma estatística numérica que indique o quão importante uma palavra ou *token*[2] é para um documento em um *corpus*. (RAJARAMAN; ULLMAN, 2011). Segundo Jones (2004), a abordagem utilizada não considera a ordem das palavras no documento, mas sim a sua importância para este. De acordo com Rajaraman e Ullman (2011), supondo a coleção de documentos D, o número de ocorrências da palavra p no documento d como f_{pd}, a frequência do termo TF_{pd} é dada por:

$$TF_{pd} = \frac{f_{pd}}{max_k f_{pd}} \tag{2.1}$$

[2]Nome atribuído a uma palavra isolada.

Isto é, a frequência do termo p no documento d é f_{pd} normalizado através a divisão pelo número máximo de ocorrências de qualquer termo k dentro do documento. A Equação 2.1 sofre com o uso da língua, pois as palavras com maiores frequências serão aquelas utilizadas para construção de ideias e conexão de elementos (e.g. "e", "a", "de", "dos"). (RAJARAMAN; ULLMAN, 2011). De modo que palavras utilizadas apenas para conexão ou construção de ideias tendem a não apresentar contribuição para extração de informação. Sua remoção é quase sempre feita nesta etapa através da eliminação de palavras populares (conhecidas como *stopwords*) ou através de etiquetagem semântica.

Para tratar dos problemas apontados, em que palavras com menor frequência mas em um maior número de documentos apresentam mais significância, o inverso da frequência do termo (IDF) pode ser explorado. De acordo com Rajaraman e Ullman (2011), supondo que a palavra p aparece em d_p de D documentos no *corpus*, o valor do IDF_p é dado por:

$$IDF_p = \log_2(\frac{D}{d_p}) \tag{2.2}$$

Por fim, o $TF\text{-}IDF_{pd}$ é obtido através de:

$$TF_{pd} \times IDF_p \tag{2.3}$$

Onde as palavras com a maior pontuação de $TF\text{-}IDF$ são frequentemente os termos que melhor caracterizam o assunto do documento. (RAJARAMAN; ULLMAN, 2011).

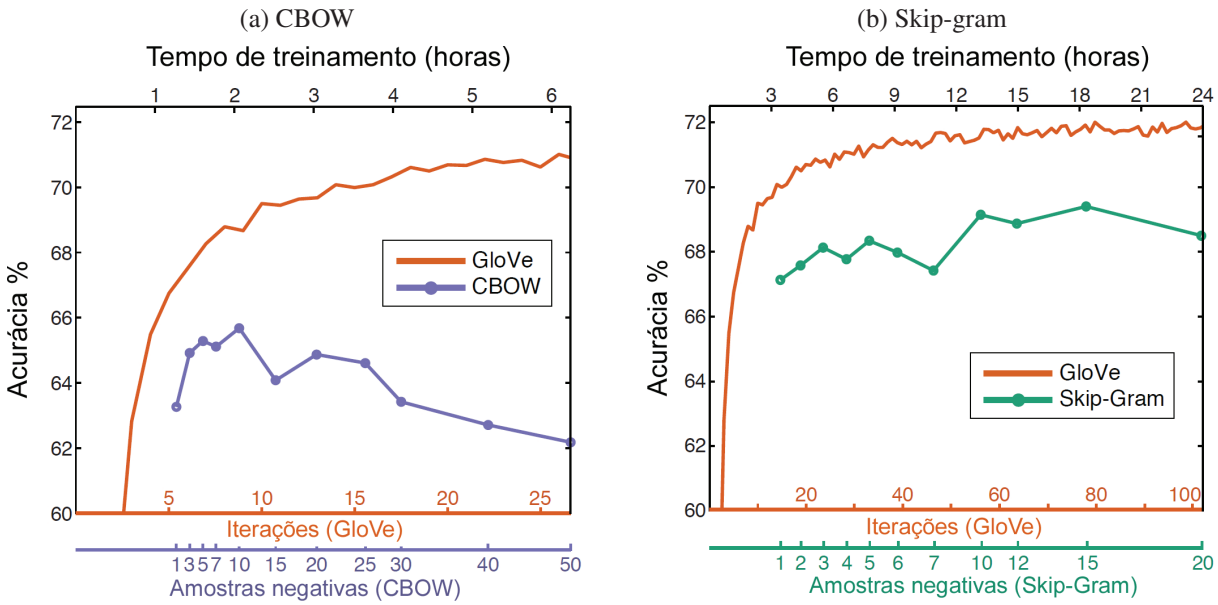
2.2.2 Modelos de espaço vetorial

Os modelos de espaço vetorial foram batizados por Bengio et al. (2003) como *word embeddings* e induzidos por suas pesquisas sobre modelos de linguagem. Bengio et al. (2003) propõe uma representação distribuída para vetores de palavras, onde cada uma destas presente no *corpus* é mapeada para um vetor de atributos, o qual representa diferentes contextos da palavra e considera cada um destes como um ponto no espaço de vetores. O funcionamento do modelo deve-se ao fato de ser esperado que palavras similares (sintática e/ou semanticamente) possuam vetores de atributos similares, baseado na hipótese de Harris (1954). Desta forma, uma vez que todas as sentenças são compostas de palavras e estas possuem seus valores de contextos, é possível representar uma frase através de uma matriz de contextos.

Recentemente os *word embeddings* demonstraram capacidade em capturar a semântica das sentenças e devido ao sucesso da técnica, surgiram vários algoritmos capazes de criar modelos para representação distribuída de palavras. Dentre eles o mais conhecido atualmente é o *word2vec*[3] de Mikolov et al. (2013b). Em seu trabalho, o autor trouxe a definição de duas técnicas com propostas semelhantes para aprendizado de *word embeddings*, o *Continuous Bag-of-Words* (CBOW) e o *Skip-Gram* (SG). No CBOW o algoritmo busca predizer palavras dado

[3]Disponível em `https://code.google.com/archive/p/word2vec/`

Figura 2: Comparação do *GloVe* com os algoritmos CBOW e SG



Fonte: Pennington, Socher e Manning (2014)

o contexto de origem (e.g. "Quem não tem cão caça com", a palavra predita seria "gato"). Enquanto que no SG é feito o caminho inverso, onde dada uma palavra o algoritmo tenta predizer as palavras próximas na sentença.

No presente trabalho foi utilizado o algoritmo *GloVe*[4] de Pennington, Socher e Manning (2014) para modelagem do espaço de vetores e obtenção dos *word embeddings*, devido a disponibilidade da técnica *word2vec* para a linguagem R[5]. Apesar do modelo utilizado diferir do apresentado por Mikolov et al. (2013b), pois o primeiro é baseado na contagem de elementos e o segundo é um modelo de linguagem neural.

É possível observar nos experimentos de Pennington, Socher e Manning (2014) (Figuras 2a e 2b), o desempenho do *GloVe* em capturar a semântica das palavras frente aos algoritmos CBOW e SG do *word2vec* ainda nas primeiras interações.

2.2.3 Bases semânticas

As relações semânticas são um aspecto fundamental a ser levado em conta quando se pretende construir programas de computador capazes de lidar com conteúdo textual, pois estabelecem associações de sentido entre palavras. Tais relações podem ser obtidas através de bases de conhecimento léxico-semântico. (OLIVEIRA et al., 2015). Na Tabela 1, apresentamos os recursos mais conhecidos na literatura para português do Brasil.

É comum neste tipo de recurso a existência de relações específicas como: antonímia, hi-

[4]Disponível em `https://nlp.stanford.edu/projects/glove/`
[5]Disponível em `https://www.r-project.org/`

Tabela 1: Bases léxico-semânticas para português do Brasil

Base	Autores	Pública
WordNet.PT	Marrafa (2002)	Não
WordNet.BR	Silva, Oliveira e Moraes (2002)	Não
MultiWordNet.PT	Pianta, Bentivogli e Girardi (2002)	Não
Onto.PT	Oliveira e Gomes (2010)	Sim
OpenWordNet-PT	Paiva, Rademaker e Melo (2012)	Sim
UfesWordNet.BR	Gomes, Beltrame e Cury (2013)	Sim
TeP	Maziero et al. (2008)	Sim
PULO	Simões e Guinovart (2014)	Sim

Fonte: Elaborado pelos autores.

peronímia, hiponímia, meronímia e sinonímia. (OLIVEIRA et al., 2015). Contudo, as bases não são completas e algumas já não recebem mais atualizações. Seguindo por este caminho, a escolha mais óbvia de utilização aponta para os últimos trabalhos desenvolvidos, desde que estes possuam um número significativo de relações. Neste sentido, o destaque seria para a base UfesWordNet.BR, porém Gomes, Beltrame e Cury (2013) comentam que o projeto é bastante preliminar, pois foi o trabalho de conclusão de curso de um aluno da graduação e construída baseando-se na tradução automática da WordNet.Pr[6], através do *Google Translate*[7]. Deste modo, optou-se pela escolha da PULO, a qual contém mais de 17 mil sentidos, referentes a 13 mil *synsets*[8] diferentes e 48 mil relações.

Apesar da PULO ser uma base recente e com um número razoável de relações, considerou-se uma maior abrangência em sua utilização conjunta com o TeP, pois este contém mais de 44 mil itens lexicais, organizados em 19 mil *synsets*, que por sua vez estão ligados através de 4 mil relações de antonímia.

2.3 Aprendizagem de máquina

Nesta seção serão descritos elementos das principais técnicas matemáticas e de aprendizado de máquina utilizadas na atual pesquisa.

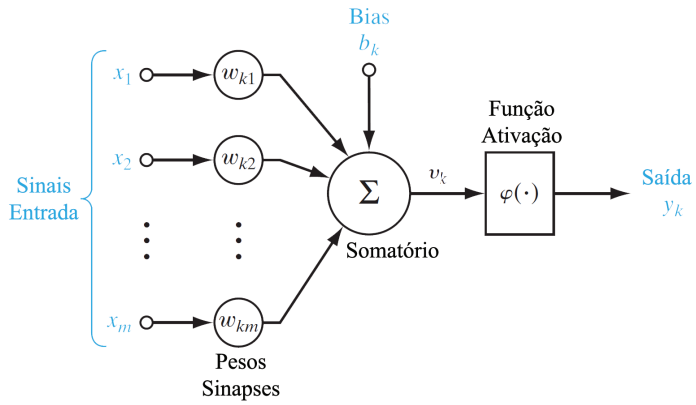
[6]Disponível em `https://wordnet.princeton.edu/`
[7]Disponível em `https://translate.google.com/`
[8]Conjuntos de sentidos de palavras que expressam o mesmo significado. (ALVES et al., 2014).

2.3.1 Redes neurais artificiais

As redes neurais artificiais (RNA) são modelos matemáticos que possuem sua inspiração no funcionamento do cérebro e nas teorias psicológicas de aprendizado, especialmente a "Teoria do Aprendizado Neural" de Hebb (1949). (LUGER, 2013). Em sua forma mais geral, as informações de uma rede neural são implícitas tanto na organização quanto nos pesos de um conjunto de processadores conectados e o aprendizado envolve tanto a reorganização quanto a modificação do peso geral dos nós e da estrutura do sistema. (LUGER, 2013).

De acordo com Haykin (2008) e Luger (2013), dois dos principais elementos das rede neurais são as conexões (conhecidas como sinapses) e os neurônios. Na Figura 3 podemos observar o funcionamento de uma unidade (neurônio), onde é realizado o processamento de sinais de entrada recebidos por de um conjunto de sinapses com diferentes intensidades (pesos) e o cálculo da função de ativação.

Figura 3: Modelo e neurônio não linear



Fonte: Adaptado de Haykin (2008)

Onde $x_1, x_2, \cdots x_m$ são sinais de entrada, bem como $w_{k1}, w_{k2} \cdots w_{km}$ os pesos das sinapses do neurônio k. Além disso, o somatório pode ser representado pela Equação 2.4.

$$\sum = net_k = \sum ni = 0x_i \cdot w_{ki} \tag{2.4}$$

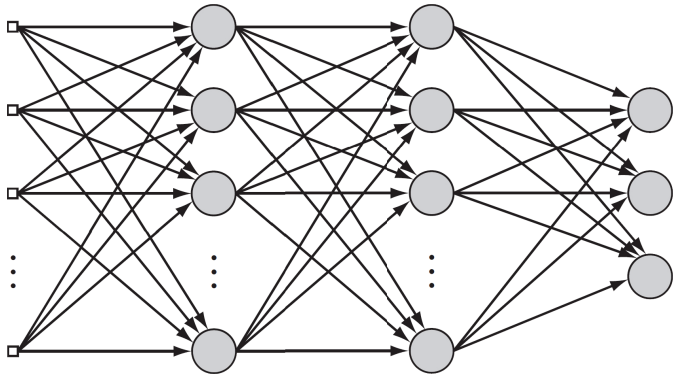
Segundo Haykin (2008), o modelo ainda conta com um bias b_k, cujo papel é aumentar ou diminuir a entrada líquida da função de ativação φ, de modo que a saída y do neurônio k é obtida através da Equação 2.5.

$$y_k = \varphi(b_k + net_k) \tag{2.5}$$

No presente trabalho foi utilizado o modelo clássico de RNA conhecido como *multilayer perceptron* (MLP) *feedfoward* devido a capacidade desta no tratamento de informações não lineares. As três características de maior destaque para as redes MLP são (1) alto grau de conectividade utilizado, onde (2) existem uma ou mais camadas que estão escondidas dos nós

de entrada e saída, além de (3) cada neurônio na rede incluir uma função de ativação não linear que é diferenciável. (HAYKIN, 2008). Já a arquitetura *feedfoward* é caracterizada pela não existência de conexões entre os neurônios de uma mesma camada, de modo que estes estão conectados apenas com outros nas camadas adjacentes. Um exemplo de arquitetura das redes MLP *feedfoward* pode ser visto na Figura 4.

Figura 4: Rede neural *multilayer perceptron feedfoward*



Fonte: Adaptado de Haykin (2008)

2.3.2 Modelos lineares generalizados

O termo modelo linear generalizado é devido a Nelder e Wedderburn (1972) que estenderam o método para estimar a máxima verossimilhança em famílias exponenciais. (MCCULLAGH, 1984). Apesar de ser uma extensão do modelo linear geral padrão, os modelos lineares generalizados ou ainda *Generalized Linear Models* (GLM) como são conhecidos, necessitam desta distinção devido as peculiaridades de cada técnica. (MCCULLAGH; NELDER, 1983).

Os GLMs unificam uma grande quantidade de métodos estatísticos como regressão, ANOVA e modelos para processamento dados categóricos. (AGRESTI, 2006). A capacidade da técnica advém de sua constituição por um componente aleatório, cuja responsabilidade é identificar a variável de resposta Y e assumir uma distribuição de probabilidade para esta; um componente sistemático, onde são especificadas as variáveis explicativas para o modelo; e também por uma função de ligação, a qual especifica uma função do valor esperado que o GLM relaciona com as variáveis explicativas através de uma equação de predição com forma linear. (AGRESTI, 2006). De um modo geral, o modelo é constituído de uma variável aleatória, um conjunto de variáveis explicativas e a especificação de uma distribuição de probabilidade para a variável aleatória, a qual pode assumir uma forma particularmente simples. (NELDER; BAKER, 2006).

Em especial nos casos de regressão com dados contínuos, a técnica assume inicialmente uma distribuição normal para a variável Y e realiza a modelagem diretamente ao redor da média usando a função de ligação $g(\mu) = \mu$. Além disso, são aplicados dois passos muito importantes para os dados categóricos: (1) permite ao Y possuir outra distribuição que não normal e (2) permite a modelagem da função ao redor da média. (AGRESTI, 2006). Uma vez

que as variáveis explicativas são especificadas pelo componente sistemático, estas são utilizadas como preditores no lado direito da equação do modelo para determinar os elementos x da Equação 2.6.

$$\alpha + \beta_1 x_1 + \cdots + \beta_k x_n \tag{2.6}$$

Já na função de ligação, é especificada uma função para $g(\cdot)$ que possua relação com o preditor linear $g(\mu)$. Uma vez que $g(\mu) = \mu$, ocorre a modelagem pela média através da Equação 2.7, a qual é forma utilizada para obter as respostas contínuas utilizando as variáveis explicativas $x_1, x_2, \cdots, x_n$. (AGRESTI, 2006)

$$\mu = \alpha + \beta_1 x_1 + \cdots + \beta_k x_n \tag{2.7}$$

2.3.3 Máquinas de vetores de suporte

Introduzidas por Vapnik (1995), as máquinas de vetores de suporte (do inglês, *Support Vector Machines*) utilizam o princípio de minimização do risco estrutural (SRM) e a teoria do aprendizado estatístico de Vapnik e Chervonenkis (1971). No qual, Joachims (1998) afirma que a ideia do SRM é encontrar a hipótese h para qual é possível garantir o menor risco empírico. Ainda, de acordo com o autor, o algoritmo é capaz de aprender qualquer conteúdo independentemente da dimensionalidade dos dados, sendo que tal fato torna-o promissor para tarefas de classificação textual.

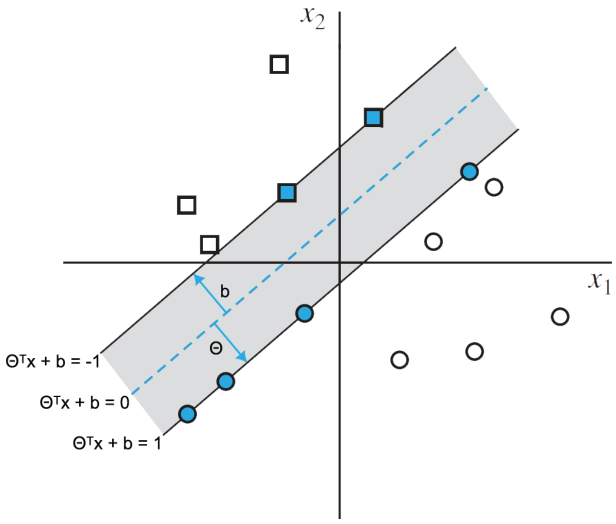
De acordo com Haykin (2008), o SVM é uma técnica de aprendizagem de máquina que pode ser usada tanto para problemas de classificação quanto para de regressão linear. Quando aplicada em tarefas de classificação, a técnica tem como objetivo construir um hiperplano como superfície de decisão de tal forma que a margem de separação entre amostras positivas e negativas seja máxima, como é possível observar Figura 5 .

Quando a técnica é aplicada para tarefas de regressão linear, o objetivo do algoritmo é aprender a função $f(x) = ax + b$ que melhor representa a relação entre as variáveis independentes e a variável dependente no conjunto de treinamento. Segundo Haykin (2008), podemos considerar o modelo de regressão linear onde a dependência da variável θ sobre x é expresso através da Equação 2.8.

$$h_\theta(x) = \theta^T x + b \tag{2.8}$$

Onde θ é o vetor de elementos de entrada e b, o bias. Considerando C como a constante que determina a troca entre o erro de treinamento e a penalidade $||\theta||^2$. Para cada elemento do vetor esperado de saída y, é computado h_θ para o elemento correspondente no vetor de entrada x com

Figura 5: Hiperplano ótimo para um problema linearmente separável



Fonte: Adaptado de Haykin (2008)

o objetivo de minimizar a função:

$$\frac{1}{2}||\theta||^2 + C\sum_{i=1}^{N}|h_\theta(x^i) - y^i| \tag{2.9}$$

2.3.4 Análise de componentes principais

A técnica de Análise de Componentes Principais[9] (PCA) foi proposta em sua forma geométrica por Pearson (1901), no trabalho em que o autor buscava encontrar linhas e planos que melhor se encaixavam em um conjunto de pontos no espaço p-dimensional. Além desta, a forma algébrica da técnica como conhecemos hoje foi proposta por Hotelling (1933).

De acordo com Jolliffe (2002), a aplicação principal da PCA é reduzir a quantidade de dimensões de uma matriz ou conjunto de dados, onde existe um grande número de atributos inter-relacionados, de modo a preservar a variação presente nestes. Para tanto, a abordagem elabora um novo conjunto de atributos, os componentes principais (PCs), que são combinações lineares não correlacionadas dos atributos originais e ordenados de acordo com a quantidade de variação destes. Isto é, o primeiro PC possui grande parte da variação presente nos atributos de origem.

Supondo que temos uma matriz m, com p colunas (atributos) e k linhas. A PCA realiza a padronização de $p_1.p_2.\cdots,p_n$ para obter média zero e variância unitária. Desta forma, a técnica busca as combinações lineares desses atributos que maximizem a variância. Tal fato leva a definição dos componentes principais pelos autovalores e autovetores unitários da matriz de correlação de Hotelling (1933), em vez da matriz de covariância, pois a última é sensível a diferenças de escalas e ruído entre as variáveis. Quando isso ocorre, os primeiros componentes

[9] do inglês, *Principal Component Analysis*.

não conseguem expressar informações superiores a uma análise das variâncias das variáveis originais. (JOLLIFFE, 2014).

O método de PCA é bastante utilizado na literatura para redução de dimensionalidade no conjunto de dados. Um dos seus usos pode ser visto em Arora, Liang e Ma (2017), onde os autores calculam a média ponderada dos *word embeddings* da sentença e então removem suas projeções dos vetores médios obtidos de seu primeiro componente principal. Segundo Arora, Liang e Ma (2017), tal abordagem obteve melhor desempenho do que a média não ponderada dos vetores em uma variedade de tarefas de similaridade textual, mesmo os que utilizam alguns modelos de aprendizagem profunda como as *Long-Short Term Memory Networks* (LSTM).

3 TRABALHOS RELACIONADOS

Neste capítulo serão apresentados os trabalhos com temas relacionados com o trabalho proposto. A revisão do estado da arte buscou identificar, com pesquisas de referências em revistas científicas qualificadas e em eventos da área, exemplos de técnicas e abordagens linguísticas, probabilísticas ou híbridas com o objetivo de mensurar a similaridade semântica entre frases curtas. Para tanto, foram pesquisados trabalhos nos repositórios digitais *ACM*[1], *ALC Anthology*[2], *Google Scholar*[3], *IEE*[4], *ScienceDirect*[5] e . Foram estudados trabalhos que continham um ou mais dos seguintes termos: "similaridade semântica textual", "relações linguísticas", "recursos linguísticos"e "aprendizado de máquina". Além disso, de modo a obter uma maior abrangência do estado atual da arte na área, também foram realizadas buscas com as variações em inglês dos termos anteriores: "*semantic textual similarity*", "*portuguese semantic textual similarity*", "*linguistic relations*", "*linguistic resources*"e "*machine learning*".

A filtragem dos trabalhos ocorreu pelo ano de publicação e através do tipo de relações linguísticas e recursos probabilísticos empregados, bem como da forma utilizada para a modelagem destas informações. Entretanto, a aplicação de tal filtro evidenciou que apenas uma pequena parte dos trabalhos encontrados foram aplicados para português, os quais em sua maioria pertencem ao evento PROPOR de 2016.

A seguir destacamos detalhes dos trabalhos de Ferreira et al. (2016), Hartmann (2016), Alves, Rodrigues e Oliveira (2016) e Cavalcanti et al. (2017) devido à sua maior proximidade do objetivo, da abordagem ou das técnicas propostas no atual trabalho.

3.1 Assessing sentence similarity through lexical, syntactic and semantic analysis

No trabalho é proposta uma técnica não supervisionada para avaliação de similaridade léxica, sintática e semântica entre frases curtas. Ferreira et al. (2016) apresenta soluções para problemas de comparação do contexto e da ordem entre as palavras, utilizando três matrizes de similaridade e penalização no caso de diferença entre o tamanho das frases.

Inicialmente as sentenças são divididas em *tokens* e após são removidas as *stopwords* propostas por Dolamic e Savoy (2010), as quais consistem de artigos, pronomes e a pontuação do texto. Na sequência, o autor aplica a lematização das palavras resultantes utilizando o *Stanford CoreNLP*[6]. Ainda nesta etapa, Ferreira et al. (2016) cria outra matriz para análise de similaridade entre os números presentes nas sentenças, pois de acordo com o autor, os números contêm informações específicas das frases. A matriz de similaridade é então criada utilizando a mé-

[1]Disponível em: `https://dl.acm.org/`.
[2]Disponível em: `http://aclweb.org/anthology/`.
[3]Disponível em: `https://scholar.google.com.br/`.
[4]Disponível em: `http://ieeexplore.ieee.org/Xplore/home.jsp`.
[5]Disponível em: `http://www.sciencedirect.com/`.
[6]Disponível em `https://stanfordnlp.github.io/CoreNLP/`.

trica *Path Length*[7](PL), porém nos casos em que $PL(conceito_1, conceito_2) < 0.1$ é utilizado a variação da distância de *Levenshtein*[8]. Considerando L o número de caracteres do termo mais longo, a similaridade entre os conceitos é dada pela Equação 3.1.

$$Sim_{lev} = 1 - \frac{Levenshtein(conceito_1, conceito_2)}{L} \tag{3.1}$$

Uma vez obtidas as matrizes de similaridade, o autor itera sobre cada linha da matriz, armazenando a maior semelhança de cada tupla em uma variável k. Ao final, a similaridade da matriz é dada por $Total_s = \frac{k}{i}$, onde i é o número de iterações realizadas. Após obtidos valores totais para as matrizes de caracteres e números, é então aplicada a Equação 3.2 para calcular o coeficiente de penalização referente ao tamanho da sentença de cada uma das matrizes.

$$SDPC = \begin{Bmatrix} \frac{|t_{s1} - t_{s2}| \times Total_s}{t_{s1}}, & se(t_{s1} > t_{s2})) \\ \frac{|t_{s1} - t_{s2}| \times Total_s}{t_{s2}}, & caso\ contrário \end{Bmatrix} \tag{3.2}$$

Onde, t_{s1} e t_{s2} correspondem ao número de *tokens* da primeira e segunda sentença. Ao final desta etapa, considerando p a matriz de palavras e n a matriz de números, a similaridade entre as sentenças é dada através de:

$$\begin{aligned} Sim_p &= Total_p - SDPCp \\ Sim_n &= Total_n - SDPCn \end{aligned}$$

$$Sim_{s1,s2} = \begin{Bmatrix} \frac{(t_p \times Sim_p + (t_n \times Sim_n)}{t_p + t_n} & se(Sim_n = 1) \\ Sim_p - (1 - Sim_n) & caso\ contrário \end{Bmatrix} \tag{3.3}$$

Onde, t_p e t_n representam o total de *tokens* de palavras e números.

A camada sintática recebe os *tokens* gerados na etapa anterior e gera um grafo representando a ordem das relações entre eles. Na sequência o autor remove da árvore de dependência os nós envolvendo relações de preposições e conjunções. Após, é feito o cálculo de similaridade seguindo as mesmas formulações da camada léxica, porém ao invés dos *tokens* são utilizados os nós do grafo gerado.

A terceira e última camada proposta por Ferreira et al. (2016) é a semântica, a qual recebe a sequência de *tokens* extraídos na camada léxica e aplica a técnica *Semantic Role Annotation*[9] (SRA) para definir os papeis de cada uma das entidades e identificar seus contextos nas frases. O objetivo principal desta camada é extrair informações do discurso na medida em que desdobra a

[7]A distância entre dois conceitos no WordNet para marcar sua semelhança. (FERREIRA et al., 2016).

[8]O número mínimo de operações de inserção, exclusão ou substituição de um único caractere necessário para transformar uma palavra em outra. (FERREIRA et al., 2016).

[9]Detecção de elementos e papéis semânticos.

ordem das ações e identifica as entidades contidas nas sentenças. O autor fez uso do FrameNet de Fillmore (2003), o qual fornece quadros semânticos, tais como uma descrição de um tipo de evento, relação ou entidade e seus agentes. Além disto é utilizado o Semafor, proposto por Das et al. (2010) para processar as sentenças e obter sua informação semântica do FrameNet. Entretanto, o cálculo de similaridade ocorre com as mesmas formulações das camadas léxicas e sintáticas.

No final do processo, as três similaridades são combinadas utilizando a equação 3.4, onde $S1$ e $S2$ representam as sentenças utilizadas como entrada para a técnica.

$$similaridade(S1, S2) = \frac{(lex_n \times lex_s) + (sin_n \times sin_s) + (sem_n \times sem_s)}{lex_n + sin_n + sem_n} \tag{3.4}$$

De acordo com Ferreira et al. (2016), lex_n representa o total de palavras da camada léxica de ambas as sentenças, enquanto que sin_n e sem_n correspondem a soma do número de triplas das camadas sintáticas e semânticas de ambas as sentenças.

O autor utilizou três conjuntos de dados abertos (dos quais um pertencia ao SemEval 2012), para comparação com trabalhos do estado da arte. De acordo com Ferreira et al. (2016), o método desenvolvido fica entre os dez primeiros resultados da competição, alcançando 0.6548 no coeficiente de *Pearson* (Seção 4.3), quando o cálculo das matrizes de similaridade é realizado com o *Path Length*. Além disto, o autor comenta que os melhores resultados surgiram de combinações da camada léxica com as medidas de similaridade do caminho, Lin e Resnik.

3.2 Solo Queue at ASSIN: combinando abordagens tradicionais e emergentes

O trabalho de Hartmann (2016) ocupa o primeiro lugar na tarefa de Avaliação de Similaridade Semântica e Inferência Textual (ASSIN) do PROPOR em 2016, cujo o objetivo é a identificação da similaridade semântica e classificação entre pares de frases curtas. Em seu trabalho, o autor atenta para o problema da esparcidade dos dados provocada por técnicas exclusivamente matemáticas ou léxicas e ainda aponta para a dependência de domínio causada por ferramentas como o WordNet, as quais ele alega que restringem a aplicação do método apenas a uma determinada língua, devido às características únicas de linguagem encontradas nos recursos.

Inicialmente o autor utiliza a técnica *word2vec* (Seção 2.2.2) para obter os *word embeddings*, através do CBOW e SG utilizando uma janela de 600 dimensões e um *corpus* contendo três bilhões de *tokens* em português brasileiro, composto por textos dos sites G1, Wikipédia e o *corpus* PLN-Br de Bruckschen et al. (2008).

De acordo com Hartmann (2016), todas as palavras do *corpus* foram mapeadas para caixa baixa afim de reduzir esparsidade dos dados neste. Além disso, as palavras não encontradas no vocabulário ou com apenas uma ocorrência no *corpus*, foram mapeadas para um *token* gené-

rico UNK, na tentativa de aprimorar a generalização de contextos do *word2vec*. Na sequência, Hartmann (2016) utiliza a técnica de Mikolov et al. (2013b), a qual o autor alega que ao somar os vetores dos *word embeddings* correspondentes as palavras de uma frase, temos como resultado um vetor de contextos que representam a sentença. Por fim, Hartmann (2016) faz uso da similaridade do cosseno entre a soma dos vetores dos pares de sentenças para o algoritmo de regressão linear SVM (Seção 2.3.3).

Ao ponto em que obtém a similaridade através dos *word embeddings*, Hartmann (2016) utiliza a clássica técnica do TF-IDF (Seção 2.2.1) para criar a representação sentencial. O autor tentou solucionar o problema da esparsidade das sentenças, através da aplicação de *stemming* e também da expansão destas com as relações de sinonímia contidas no TeP (Seção 2.2.3). Segundo Hartmann (2016), tal operação ocorreu somente para palavras que possuíam até dois sinônimos, pois ele indica que ao adicionar mais *tokens* na sentença, estas tornavam-se representações demasiado genéricas. Na sequência são utilizadas as duas representações como entradas para regressão linear através do algoritmo SVM (Seção 2.3.3).

O autor especula que o uso dos *word embeddings* contribuiu para a captura da semântica das sentenças, principalmente nos casos em que o significado do contexto possuía relevância, pois somente o TF-IDF não consegue capturar tal informação.

A soma dos *word embeddings* atua como uma representação alternativa dos contextos das sentenças, porém a operação pode resultar em um valor próximo a outros contextos e então o valor palavra-contexto perde o sentido esperado. Conforme é demonstrado no exemplo[10] abaixo.

Considerando as duas sentenças "indicador acumula alta"e "indicador subiu neste mês", a matriz de *word embeddings* obtida é:

$$\texttt{indicador} = \{0.5858, -0.4446, -0.0255, \cdots, w_d\}$$
$$\texttt{acumula} = \{0.4388, 0.2616, 0.7027, \cdots, w_d\}$$
$$\texttt{alta} = \{-0.3280, 0.2855, 0.2288, \cdots, w_d\}$$

$$\texttt{indicador} = \{0.5858, -0.4446, -0.0255, \cdots, w_d\}$$
$$\texttt{subiu} = \{-0.2608, -0.2656, 0.2669, \cdots, w_d\}$$
$$\texttt{neste} = \{0.0969, 0.0510, -0.1568, \cdots, w_d\}$$
$$\texttt{mes} = \{0.0876, -0.6151, -0.6870, \cdots, w_d\}$$

Onde w_d representa o valor da palavra w na dimensão d.

Considerando C como o contexto da sentença s, a variável d como o número dimensões dos vetores de palavras e w como a palavra no espaço de contextos. A representação da sentença

[10] O exemplo foi implementado em linguagem R com o pacote *text2vec* utilizando o algoritmo de Pennington, Socher e Manning (2014) para treinamento de *word embeddings* com janelas de 50 elementos, 10 iterações, semente 1 e mensurados utilizando distância do cosseno.

ocorre através da Equação 3.5 abaixo:

$$C_s = \sum_{i=0}^{d} w_i \tag{3.5}$$

Desta maneira, os valores de C_{s1} e C_{s2} seriam:

$C_{s1} = \{0.6966392, 0.1024995, 0.9061006, \cdots, Cs_d\}$
$C_{s2} = \{0.5094654, -1.2744304, -0.6024548, \cdots, Cs_d\}$

Onde Cs_d representa o valor do contexto da sentença s na dimensão d.

Tal como é advertido por Hartmann (2016), a soma dos *word embeddings* das sentenças cria uma representação genérica da frase e não reflete as informações individuais. Desta forma, apesar dos resultados obtidos pelo trabalho, se faz necessário buscar outras formas de representação que preservem tais informações.

3.3 ASAPP: alinhamento semântico automático de palavras aplicado ao português

O trabalho de Alves, Rodrigues e Oliveira (2016) traz duas abordagens, a primeira exclusivamente simbólica (apelidada de Reciclagem), baseada em heurísticas sob redes léxico-semânticas para a língua portuguesa. Enquanto que a segunda (apelidada de ASAPP) utiliza recursos de aprendizagem automática supervisionada e foi inspirada no trabalho de Alves et al. (2014). A opção do autor por criar duas abordagens dá-se pela possibilidade de comparação dos resultados, mesmo que de forma indireta entre abordagem exclusivamente heurística com a variação que faz uso de aprendizado de máquina.

Inicialmente foram contabilizados os grupos nominais, verbais e preposicionais em cada uma das frases de cada par, bem como calculado o valor absoluto da diferença para cada tipo de grupo. Na sequência, foi aplicado Reconhecimento de Entidades Mencionadas (REM) e para cada tipo de entidade encontrada foi calculado o valor absoluto da diferença da contagem em ambas as sentenças. (ALVES; RODRIGUES; OLIVEIRA, 2016).

Em sua abordagem heurística, Alves, Rodrigues e Oliveira (2016) aplica atomização e etiquetagem semântica como pré-processamento das sentenças. Na sequência, o autor faz uso da lematização através do LemPort[11] e também o REM através do Apache OpenNLP[12].

Uma vez obtidas as características, são utilizadas de nove redes léxico-semânticas para o cálculo de similaridade, através da semelhança mais elevada entre palavras vizinhas de cada sentença. Para tanto, as redes são utilizadas de modo a obter cinco tipos de relações: antonímia, hiperonímia, hiponímia, sinonímia e outros (o último faz referência ao agrupado de todas as demais relações existentes). Considerando $Viz(x)$ como a união de todas as relações de uma palavra através da medida de Banerjee e Pedersen (2003), um dos cálculos de similaridade

[11] Proposto por Rodrigues, Oliveira e Gomes (2014).
[12] Disponível em `http://opennlp.apache.org/`

utilizados por Alves, Rodrigues e Oliveira (2016) é apresentado na Equação 3.6.

$$Sim_{max}(t,h) = \sum_{i=1}^{|t|} max(Sim(Viz(T_i), Viz(H_j))) \ : H_j \ \epsilon \ H \tag{3.6}$$

De acordo com o autor, cada relação foi mensurada através contagem de quatro métricas de distância: Lesk (Equação 3.6), Jaccard, Overlap[13] e Dice[14]. Além disso, outras três heurísticas foram utilizadas, a Distância Média, *Personalized PageRank*[15] de Agirre e Soroa (2009), bem como a presença em *synsets* difusos da CONTO.PT. De modo que, cada heurística descrita até o momento é utilizada pelo autor na abordagem Reciclagem, e no final, gera um valor de similaridade específico entre as sentenças.

Em sua abordagem supervisionada, Alves, Rodrigues e Oliveira (2016) utiliza o conjunto ASSIN para realizar três experimentos com algoritmos de regressão linear disponíveis no software WEKA[16]: o primeiro é proposto por Friedman (2002), conhecido como regressão aditiva e *boosting*; o segundo é um esquema múltiplo para seleção proposto por Hall et al. (2009); e o terceiro foi o processo gaussiano de Mackay (1998). Por fim, o autor afirma que seu melhor resultado foi através s regressão por *boosting*, com a base TeP, Sim_{max} e distância de *overlap*.

Ao final do trabalho, os resultados obtidos com a abordagem ASAPP superam com cerca de 9% a Reciclagem, independente do algoritmo de regressão ou rede léxico-semântica utilizada. Além disso, é importante ressaltar que, todas as redes semânticas obtiveram resultados próximos (dentro de uma margem de 0.1). O que é justificado pelo autor, como prova de relevância da heurística aplicada frente o conteúdo da rede utilizada.

3.4 Statistical and Semantic Features to Measure Sentence Similarity in Portuguese

O trabalho de Cavalcanti et al. (2017) apresenta uma abordagem hibrida para calcular a semelhança semântica entre frases escritas em português, a qual obteve resultados superiores aos de Hartmann (2016) no conjunto de dados da ASSIN. De acordo com o autor, sua técnica supera o problema do significado das sentenças combinando as técnicas de TF-IDF, o tamanho das frases e a semelhança dos *word embeddings* obtida utilizando tanto um método baseado em matriz de similaridade quanto em matriz binária.

A utilização das técnicas resulta em um conjunto de atributos para cada par de sentenças, o qual é utilizado para treinar um algoritmo de regressão linear. A similaridade do TF-IDF (Seção 2.2.1) é calculada de forma similar ao realizado por Hartmann (2016), onde o autor faz uso da técnica em conjunto com a aplicação de *stemming* e também da expansão das relações

[13]Medida de similaridade que mede a sobreposição entre dois conjuntos.
[14]Medida estatística da proporção de termos compartilhados.
[15]Nos experimentos foi utilizado $n = 50$.
[16]Disponível em `http://www.cs.waikato.ac.nz/ml/weka/`

de sinonímia contidas no TeP (Seção 2.2.3). Já o valor de similaridade dos *word embeddings* é obtido através de dois métodos, o primeiro baseado em matriz de similaridade e o segundo utilizando uma matriz binária. Em ambos, Cavalcanti et al. (2017) aplica a remoção de *stopwords* e realiza a lematização das palavras.

No método baseado em matriz de similaridade, similar ao proposto por Ferreira et al. (2016), o autor repete os passos a seguir até que não existam mais palavras para calcular a semelhança: (1) calcula o valor de similaridade entre cada palavra das duas sentenças através do *word2vec* (Seção 2.2.2) e então (2) remove os termos que obtiveram os maiores valores. Uma vez que não restem mais palavras, é então realizada a média entre os valores mais altos de similaridade obtidos entre as frases através da Equação 3.7.

$$Similaridade(A, B) = \frac{\sum_{i=1}^{n} MaxSim(a_n, b_n)}{n} \tag{3.7}$$

Enquanto que no método que faz uso da matriz binária, os valores de similaridade são obtidos através da Equação 3.8 e no final a similaridade entre o par de frases é obtida através média das semelhanças entre as palavras.

$$Similaridade(A, B) = \begin{cases} 1, & se\ as\ palavras\ forem\ iguais \\ 0, & caso\ contrario \end{cases} \tag{3.8}$$

Além destes, o autor também utiliza o tamanho de cada par de sentenças como atributo, o qual é calculado através da divisão do número de palavras da frase menor pelo número de palavras da frase mais longa.

3.5 Comparativo e considerações

A seguir são descritos de forma breve alguns trabalhos relacionados não comentados até este ponto do texto. São também relacionados aspectos considerados relevantes para o atual trabalho, usados como elementos de comparação entre todos os trabalhos estudados. Estes aspectos são sumarizados em uma tabela comparativa, ao final desta seção.

No decorrer da revisão da literatura, notou-se que atualmente o interesse na tarefa de avaliação de similaridade textual vem recebendo bastante atenção (FERREIRA et al., 2016; AGIRRE et al., 2017; HARTMANN, 2016; BARBOSA et al., 2016; FREIRE; PINHEIRO; FEITOSA, 2016; CAVALCANTI et al., 2017). Tal fato também é observado em eventos como o SemEval STS[17] e ASSIN, os quais são tarefas para mensurar a similaridade semântica entre sentenças, tanto para a língua inglesa quanto para portuguesa.

Muitos dos trabalhos encontrados no estado da arte são voltados para avaliação de similaridade exclusivamente na língua inglesa ou bilíngue. Dentre estes, destacamos o trabalho de

[17] Trilha *Semantic Textual Similarity* do evento.

Hänig et al. (2015) e Kashyap et al. (2016), pois os autores utilizaram abordagens similares com as da atual dissertação e obtiveram bons resultados em seus experimentos. Em Kashyap et al. (2016), o autor obteve o segundo lugar no evento SemEval 2014 para avaliação de similaridade da língua inglesa e espanhola. A técnica proposta busca mensurar a semelhança entre palavras e sentenças, através da combinação do *Hyperspace Analog to Language*[18] (HAL) de Burgess, Livesay e Lund (1998) em conjunto com medidas de similaridade extraídas do WordNet. Já em Hänig et al. (2015), são utilizadas entidades nomeadas e expressões temporais, bem como uma série de medidas de distâncias e manipulação de negação para avaliar a similaridade entre sentenças. Além disto, em conjunto com os recursos anteriores, o autor faz uso de antonímia, hiperonímia, hiponímia e sinonímia contidas em uma variação do WordNet para compor os atributos utilizados no SVM.

No trabalho de Barbosa et al. (2016) são criadas métricas com *word embeddings* e *Inverse Document Frequency* (IDF) para utilização em uma rede siamesa (*Siamese Networks*) de Chopra, Hadsell e Y. (2005). Em Freire, Pinheiro e Feitosa (2016) é um proposto um *framework* de três sistemas: STS_MachineLearning, STS_HAL e STS_WORDNET_HAL. O primeiro utiliza a similaridade entre palavras pelo coeficiente Dice e pelo WordNet, enquanto que os demais utilizam a abordagem simbólica com o cálculo da similaridade de palavras através da *Latent Semantic Analysis* (LSA).

A Tabela 2 foi elaborada com base no estudo realizado, onde são descritas algumas das características dos principais trabalhos encontrados. Devido ao baixo volume de pesquisas objetivando a avaliação de similaridade semântica em português do Brasil, a comparação também inclui estudos de aplicação na língua inglesa. Apesar da maioria dos trabalhos na tabela utilizarem recursos linguísticos, o mesmo não se repete na literatura. Além disto, possível observar alguns estudos apenas com o uso de recursos probabilísticos ou então heurísticos para avaliação da similaridade entre sentenças (BRYCHCÍN; SVOBODA, 2016; FERREIRA et al., 2016; BARBOSA et al., 2016).

É observado que a abordagem probabilística escolhida difere nos trabalhos, mas que a maioria utiliza os *word embeddings* em conjunto com outros recursos. Além disto, destaca-se o uso do SVM como algoritmo utilizado para a regressão. Acreditamos que o uso deste ainda é dominante na tarefa devido aos bons desempenhos acumulados com a utilização no decorrer dos anos, além da dificuldade em modelar as informações linguísticas e probabilísticas nos algoritmos de aprendizagem profunda. Entretanto, são observados trabalhos mais recentes como Barbosa et al. (2016), Brychcín e Svoboda (2016) e Mueller (2016) quebrando este paradigma.

Dos estudos que fazem uso de recursos linguísticos, nota-se que a maioria utiliza a contagem ou então realiza o cálculo de distância para apenas uma das relações observadas. Tal fato acaba por não explorar completamente o potencial da língua e muitas vezes não implica em similaridade. Vejamos o exemplo considerando apenas as relações de hiperonímia e hiponímia contidas na OpenWordNet-PT (Seção 2.2.3) para as sentenças em português do Brasil:

[18]Espaço semântico de co-ocorrências de palavras. (GOMAA; FAHMY, 2013).

Eu vendi um **carro.**
A empresa efetuou a venda de um **trem.**

É de conhecimento que tanto **trem** quanto **carro** possuem uma relação direta de hiperônimo com **veículo**. Tal fato pode levar a acreditar que existe uma similaridade apenas pela presença desta relação, porém como é possível observar nas sentenças de exemplo, isso não implica em concluir que uma frase é similar a outra. Logo, o trabalho aqui apresentado é motivado pelo interesse na integração destas classes de recursos visando aproveitar de maneira mais efetiva e aprofundada cada uma de suas potencialidades.

Tabela 2: Comparação com trabalhos relacionados e similares

	Recursos linguísticos	Forma de utilização dos recursos linguísticos	Recursos probabilísticos	Algoritmo	Conjunto de dados	Linguagem de análise
Atual	Antonímia, hiperonímia, hiponímia, sinonímia	Contagem e substituição	*Word embeddings* e TF-IDF	SVM, RNA e GLM	ASSIN 2016	Português
Hartmann (2016)	Sinonímia	Substituição	*Word embeddings* e TF-IDF	SVM	ASSIN 2016	Português
Cavalcanti et al. (2017)	Sinonímia	Substituição	*Word embeddings* e TF-IDF	Regressão linear	ASSIN 2016	Português
Ferreira et al. (2016)	Papeis semânticos, árvore sintática e sinonímia	Substituição e medidas de similaridade	-	-	SemEval 2012	Inglês
Alves, Rodrigues e Oliveira (2016)	Antonímia, hiperonímia, hiponímia, sinonímia	Contagem e medidas de similaridade	-	Regressão aditiva, esquema múltiplo para seleção e regressão gaussiana	ASSSIN 2016	Português
Fialho et al. (2016)	Polaridade de sentenças e negação	Medidas de similaridade	TF-IDF	SVM	ASSIN 2016	Português
Kashyap et al. (2016)	Hiperonímia, hiponímia, sinonímia e sentido	Verificação de existência	HAL	SVM	SemEval 2014	Inglês
Hänig et al. (2015)	Antonímia, hiperonímia, hiponímia, sinonímia e negação	Medidas de similaridade	*Word embeddings*	SVM	SemEval 2015	Inglês
Barbosa et al. (2016)	-	-	*Word embeddings*	SVM e redes neurais siamesas	ASSIN 2016	Português
Brychcín e Svoboda (2016)	-	-	*Word embeddings*, *paragraph2vec* e TF-IDF	SVM, *perceptron*, árvores de decisão, regressão gaussiana e LSTM	SemEval 2016	Inglês

Fonte: Elaborado pelos autores.

Na Tabela 2 nota-se as diferenças do atual trabalho frente aos comparados, pois com exceção de Kashyap et al. (2016), apenas os estudos de Alves, Rodrigues e Oliveira (2016) e Hänig et al. (2015) utilizaram mais de três recursos linguísticos. Ainda, é possível descartar Kashyap et al. (2016), pois apenas os outros dois lançaram mão dos recursos de antonímia, hiperonímia, hiponímia e sinonímia, sendo que destes, somente Alves, Rodrigues e Oliveira (2016) foi aplicado na língua portuguesa. Tal fato reforça uma das premissas deste trabalho, de que a maioria dos estudos voltados para o português faz pouco uso dos recursos linguísticos.

Comparando o trabalho atual com Alves, Rodrigues e Oliveira (2016), observa-se que neste trabalho não são utilizados recursos probabilísticos obtidos das sentenças. Além disso, é dado enfoque nas medidas de similaridade obtidas através das distâncias entre nós e na contagem das relações linguísticas compartilhadas entre as frases. Ao contrário do autor, buscamos explorar a substituição de palavras por seus sinônimos em comum e também generalizamos os hipônimos compartilhados para seus hiperônimos. Deste modo, as sentenças tornaram-se mais próximas gramaticalmente, pois as palavras que compartilhavam relações de hiperonímia, hiponímia ou sinonímia tiveram sua forma generalizada em ambas as frases, causando um aumento no número de termos em comum.

4 ABORDAGEM PROPOSTA

Neste capítulo será apresentada a abordagem proposta, de modo a proporcionar as descrições e justificativas para as escolhas realizadas. Além disso, serão descritos os principais elementos da abordagem, bem como a forma de utilização dos recursos probabilísticos e linguísticos para realizar a avaliação de similaridade entre sentenças curtas em português do Brasil.

A abordagem proposta considera o processo geral para identificar a similaridade entre as sentenças a partir de uma etapa inicial que aborda a representação das frases, seguida de uma etapa que implementa um conjunto de medidas de similaridade entre as frases, a serem aplicadas sobre os recursos de representação escolhidos. Estas medidas de similaridade estão diretamente relacionadas ao tipo de informação usada na etapa inicial. Por fim, tanto as representações das sentenças como os resultados das medidas de similaridade são utilizadas como entrada para algoritmos de classificação.

Conforme citado no capítulo de introdução, são utilizados recursos linguísticos e probabilísticos na abordagem proposta para melhor representar as sentenças. No que tange o estudo de relações linguísticas, foram utilizados os conceitos de antonímia, hiperonímia, hiponímia e sinonímia, através das bases TeP (MAZIERO et al., 2008) e PULO (SIMÕES; GUINOVART, 2014). Já no âmbito probabilístico, foram explorados os conceitos de Modelo de Espaço Vetorial (MEV), TF-IDF e Análise de Componentes Principais (PCA). Para realizar a classificação das sentenças foram utilizados algoritmos como SVM, RNA e GLM. Destaca-se que a abordagem permite a utilização de diferentes modelos de aprendizagem de máquina e também a inclusão de outros recursos léxico-semânticos ou probabilísticos além dos citados.

Um dos diferenciais desta abordagem é o uso exclusivo de recursos abertos, pois todos os itens elaborados nesta pesquisa serão disponibilizados em *open-source*, de modo a permitir a reprodutibilidade desta e promover novos ferramentais para área de Processamento de Linguagem Natural.

O capítulo está organizado da seguinte maneira: inicialmente será abordada a metodologia proposta na Seção 4.1, descrevendo seus aspectos gerais. Na Seção 4.2 está a descrição do *corpus* elaborado para suportar algumas das análises, bem como os processamentos realizados sobre este. A Seção 4.3 consiste da descrição do conjunto de dados utilizado para comparação dos resultados deste trabalho, para com os obtidos no estado da arte. Por fim, na Seção 4.4 são apresentadas em detalhe as técnicas que são utilizadas na abordagem proposta na atual dissertação.

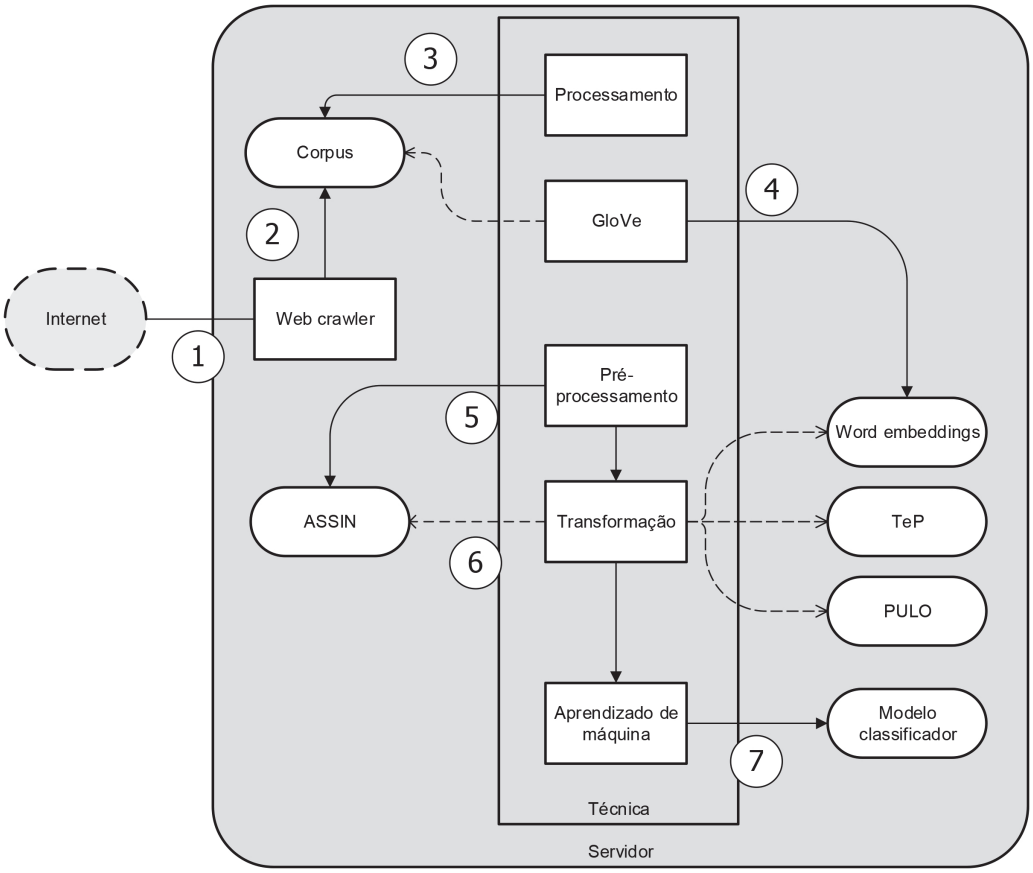
4.1 Metodologia geral

Nesta seção é apresentada a metodologia geral da técnica proposta nesse trabalho, bem como a descrição de suas etapas e da arquitetura elaborada. É possível afirmar que o objetivo desta é prover suporte ao processo descrito na Seção 4.4, o qual necessita de uma série de configurações

e recursos obtidos previamente para seu correto funcionamento.

Para proporcionar melhor condições de descrição e de análise da abordagem proposta, a metodologia adotada foi dividida em duas etapas: composição e aplicação. A etapa denominada de composição é apresentada na Figura 6. Esta etapa possui como objetivo utilizar recursos de aprendizagem de máquina para gerar um modelo com capacidade para classificar a similaridade entre duas sentenças em linguagem natural. Um dos recursos necessários é um corpus para apoiar aspectos da abordagem. Para geração deste recurso, o primeiro passo considerado é a captura de textos em páginas de notícias disponíveis na Web através de um *Web Crawler*[1] desenvolvido como parte da pesquisa. Na medida em que ocorre a coleta, a cada página visitada este software realiza a extração dos elementos textuais, a remoção de marcações *HTML* e grava o texto em um arquivo contendo um parágrafo por linha. Após a coleta de um número considerado como suficiente de páginas, é formado o *corpus* descrito na Subseção 4.2, como pode ser visto na Figura 6, no passo 2.

Figura 6: Metodologia proposta para a etapa de composição



Fonte: Elaborado pelos autores.

Ainda conforme a Figura 6, no passo 3 são aplicadas operações para preparação do *corpus* elaborado, a partir das quais ocorre uma filtragem do texto para obter melhoras com o uso deste como entrada para o *GloVe* (Seção 2.2.2). Ao final do passo 4, o algoritmo de Pennington,

[1] Software destinado a coleta e captura de textos na internet.

Socher e Manning (2014) gera os *word embeddings* e armazena-os no servidor, em um arquivo no formato *Comma Separated Values* (CSV). A opção por gerar os modelos de espaço vetorial decorre da falta de um repositório consolidado na literatura para *word embeddings* em Português do Brasil. Já a opção por gerar o arquivo CSV decorre da popularidade do padrão e também da facilidade de manuseio do arquivo por softwares e linguagens de programação.

Uma vez que os recursos de *word embeddings* foram gerados, ocorre a etapa de pré processamento (passo 5) do conjunto de dados utilizado para a avaliação e treinamento do modelo. Neste caso, trata-se do corpus disponibilizado pelo ASSIN 2016. Esta etapa busca reduzir a esparcidade da informação, através da remoção da pontuação, transformação do texto para caixa baixa e remoção de dados numéricos. No passo 6 são utilizados os recursos léxico-semânticos (PULO e TeP) para incorporar as relações linguísticas e realizar tanto as substituições de sinônimos, bem com a generalização dos hipônimos.

Buscando apresentar um breve resumo destes passos da abordagem geral (os passos 5 e 6), veremos um exemplo de transformação considerando as sentenças originais (números 1 e 2) descritas a seguir:

1. A comissão apura denúncias de abuso e exploração sexual em meninas da comunidade quilombola.
2. O grupo apura denúncias de abusos e exploração sexual de crianças da Comunidade Quilombola.

Após os passos de pré processamento número 5 e 6, são obtidas as seguintes sentenças transformadas (números 1.1 e 2.2) , a partir deste exemplo:

1.1. comissao apurar denunciar abusar exploracao sexual criancas comunidade quilombola
2.1. comissao apurar denunciar abusos exploracao sexual criancas comunidade quilombola

Destaca-se nas frases resultantes a influência de relações de hiperônimo, com as palavras "crianças"e "meninas". Além disso, observa-se a substituição de "grupo"por "comissão", no caso de relação de sinonímia.

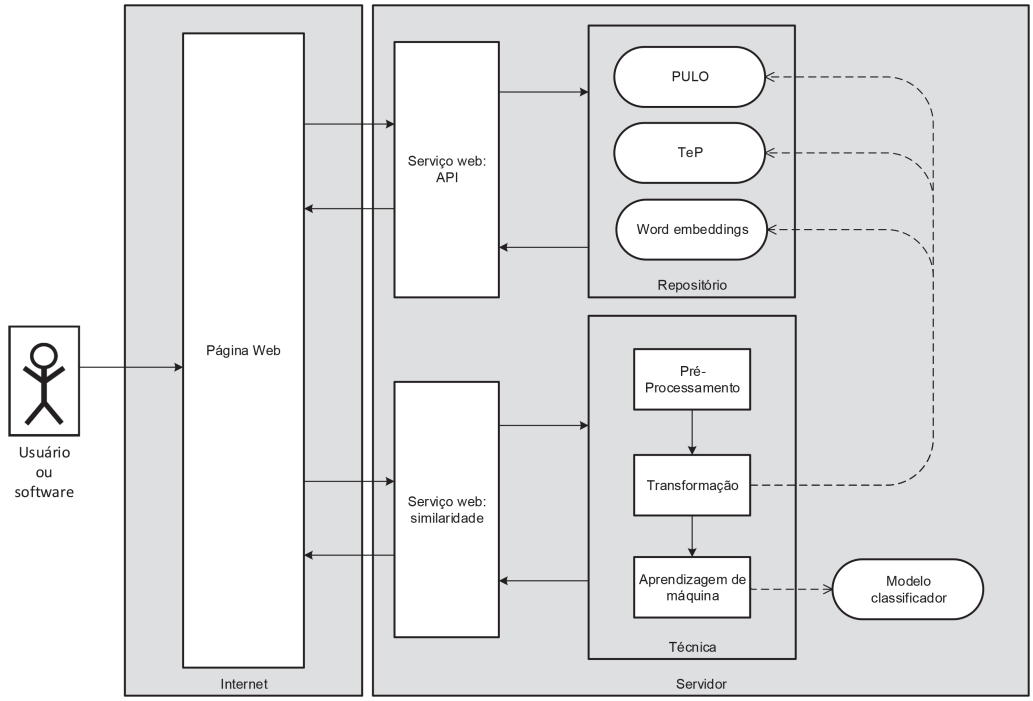
Por fim, os dados resultantes do processo até este momento são utilizados como entrada para treinamento e teste dos algoritmos de aprendizagem de máquina, onde serão gerados os modelos classificadores de similaridade (identificado no passo 7).

Já a etapa denominada de aplicação da metodologia, apresentado na Figura 7 consiste nos recursos necessários para a avaliação de similaridade entre sentenças em um ambiente na web, com acesso disponível a qualquer usuário. É possível observar na mesma figura e através da compreensão da abordagem, que não são necessários o serviço ou página web para o pleno funcionamento desta. De modo que este trabalho não tem intenções de constituir ou explorar o desenvolvimento destes recursos, pois são vistos apenas como facilitadores de acesso e para disponibilização como interfaces online.

A etapa de aplicação da metodologia prevê dois fluxos diferentes a partir da interação de um usuário ou software: a análise de similaridade ou solicitações para recursos léxico-semânticos. No primeiro, é necessário que o usuário submeta duas sentenças através de uma página web. A partir deste ação, a aplicação fará a chamada do componente **Serviço web: similaridade** para acionar a técnica proposta neste trabalho. Em um primeiro momento, as frases passarão pelas etapas de pré processamento, transformação e aprendizagem de máquina para obter a similaridade entre estas. No segundo fluxo, o usuário deverá informar duas palavras e o tipo de relação (*embeddings*, sinonímia, hiperonímia ou hiponímia) que está buscando na página web, após esta fará a requisição para o componente **Serviço web: API** que buscará as informações nos recursos *word embeddings*, PULO e TeP.

Como é possível observar nas Figuras 6 e 7, a metodologia geral proposta é facilmente adaptável para atuar com diferentes modelos de aprendizagem de máquina ou incluir outros recursos léxico-semânticos. De modo que basta a configuração destes na técnica (disponível online) para utilização dos novos algoritmos, bases e relações.

Figura 7: Metodologia proposta para etapa de aplicação



Fonte: Elaborado pelos autores.

Nas próximas seções serão detalhados: o *corpus* (Subseção 4.2) obtido pelo Web Crawler e o conjunto de dados utilizado para treinamento dos algoritmos de aprendizagem de máquina (Subseção 4.3). Além disso, veremos em detalhes todas as etapas e recursos envolvidos na abordagem proposta (Subseção 4.4).

4.2 Corpus elaborado

No presente trabalho foi desenvolvido um *Web Crawler*[2] na linguagem Java para captura de textos em páginas de notícias como Google News e Wikipédia. Durante o processo de coleta, a cada página visitada o software realiza a extração dos elementos textuais, a remoção de marcações HTML e a gravação do texto em um arquivo físico contendo um parágrafo por linha. Ao final do processo, é obtido o *corpus* que será utilizado pelo algoritmo *GloVe* (Seção 2.2.2) para gerar os *word embeddings*.

Ao inspecionar manualmente o *corpus* constatamos que a maior parte do conteúdo textual foi extraído da Wikipédia, além da existência de frases contendo apenas uma palavra ou somente caracteres especiais. Um exemplo destas sentenças pode ser observado no Apêndice A. Este tipo de ocorrência pode causar uma alteração na contagem realizada pelo algoritmo de Pennington, Socher e Manning (2014), pois este utiliza como premissa a teoria de Harris (1954). Uma vez que tais sentenças muitas vezes não contêm texto, mas apenas marcadores ou indicadores das páginas web e consequentemente não acrescentariam informação útil para o treinamento dos *word embeddings*, realizamos o processamento do *corpus* para a remoção de frases compostas somente com números ou que continham menos de cinco palavras.

Outro ponto que foi observado no texto, é a quantidade de caracteres não imprimíveis. Para solucionar este caso, optamos por não remover toda a pontuação, pois Manning e Schütze (2000) enfatizam a existência de informação contida em tal grupo. Logo, somente os caracteres de pontuação diferente de $\{., ; ?! - \}$ foram removidos da versão processada do *corpus*.

De modo a tornar reprodutível esta pesquisa e também contribuir com os trabalhos futuros na área, disponibilizamos o *corpus* elaborado (em sua forma original) e os *word embeddings* através de uma página na Universidade[3]. Além disso, acreditamos que tal ato contribui para o aumento da disponibilidade de recursos na área de PLN e possibilita que outros projetos possam fazer uso destes em diferentes domínios da computação.

4.3 Conjunto de dados anotados

O presente trabalho utilizou como base para comparação de resultados o conjunto de dados disponibilizado pela tarefa ASSIN, pertencente ao evento PROPOR 2016. O objetivo desta é a identificação da similaridade semântica e classificação entre pares de frases curtas disponibilizados no conjunto de dados. Segundo Fonseca et al. (2016), o conjunto de dados disponibilizado foi anotado pelo total de 36 pessoas que participaram em diferentes quantidades, sendo que cada frase foi avaliada por 4 pessoas. É possível observar na Tabela 4, alguns exemplos de sentenças anotadas.

Ao final, 11.3% dos pares de sentenças foram descartados por não conterem três julgamen-

[2]Disponível em `https://github.com/albarsil/WebCrawler`.
[3]Disponível em `https://www.projeto.unisinos.br/pipca_sts`.

tos iguais quanto à implicação. O resumo das estatísticas do *corpus* pode ser observado na Tabela 3.

Tabela 3: Estatísticas de similaridade do ASSIN

Similaridade	Quantidade de sentenças
4,0 - 5,0	2.410
3,0 - 3,75	2.872
2,0 - 2,75	3.814
1,0 - 1,75	904

Fonte: Fonseca et al. (2016)

Tabela 4: Exemplos de sentenças da ASSIN

Frases	Relação	Similaridade
Prometi ao estúdio que entregaria uma última trilogia para terminar a saga Prometi ao estudo que faria uma última trilogia para finalizar a saga	Paráfrase	5
Fechados há três semanas, os bancos gregos reabriram na manhã desta segunda-feira Os bancos gregos reabriram as portas três semanas depois de terem fechado	Nenhuma	4
O caminhão vinha do estado de São Paulo e foi abordado na BR-116 em frente a Unidade Operacional Taquari O caminhão vinha do estado de São Paulo	Inferência	3
De acordo com a PM, por volta das 10h30 havia 2 mil militantes no local O protesto encerrou por volta de 12h15 (horário local)	Inferência	2
O prazo terminara dia 28 de julho Foi antes da meia-noite, as 23h40m	Nenhuma	1

Fonte: Elaborado pelos autores.

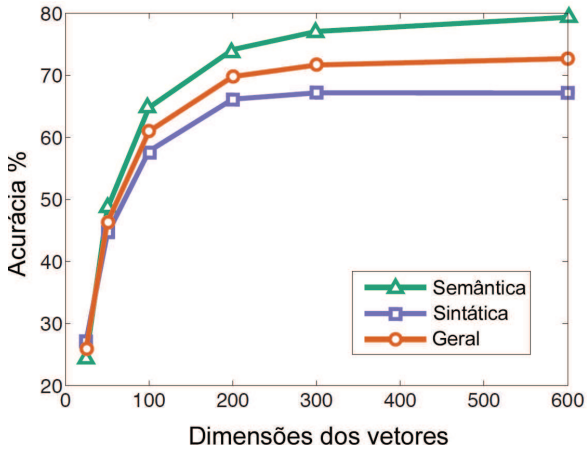
O conjunto de dados conta com 10.000 pares de sentenças coletadas através do Google News (divididos igualmente para o português do Brasil e de Portugal), destes 6.000 registros são dados para treinamento e os demais para teste, ambos os conjuntos contendo o valor de similaridade entre os pares de sentenças no intervalo $[1, 5]$. De acordo com Fonseca et al. (2016), a avaliação dos trabalhos submetidos para a tarefa deu-se através da Correlação de *Pearson* (CP) e do Erro

Quadrático Médio (EQM), onde o primeiro mensura o quão linearmente estão relacionados o resultado e o valor esperado, enquanto que o segundo estima o erro ao classificar à similaridade.

4.4 Técnica

No presente trabalho foi utilizado o algoritmo *GloVe* (Seção 2.2.2) para modelagem do espaço de vetores e obtenção dos *word embeddings* através da utilização do *corpus* descrito na Seção 4.2. Para tanto, o *GloVe* foi treinado durante 10 épocas com 6 elementos na janela de contexto, 100 co-ocorrências e taxa de aprendizagem de 0.15. Além disso, o tamanho dos vetores foi definido para 600 posições, pois notou-se nos testes realizados por Pennington, Socher e Manning (2014) (Figura 8) o aumento da acurácia do algoritmo em capturar a semântica das sentenças com o valor de posições escolhido.

Figura 8: Acurácia em função da dimensão dos vetores



Fonte: Pennington, Socher e Manning (2014)

Inicialmente foi realizada a composição de cada frase através dos *word embeddings* correspondentes a cada palavra e desta maneira foi obtida uma matriz de contextos com 782.129 linhas e 600 colunas. Neste ponto, assim como nos trabalhos de Hartmann (2016) e Mikolov et al. (2013b), criou-se um atributo através da similaridade do cosseno entre a soma da matrize de contextos de cada sentença. Contudo, Hartmann (2016) comenta que a soma da matriz de *word embeddings* cria uma representação genérica da frase e acaba por não refletir seus contextos. Desta forma, optou-se por aplicar a técnica PCA (Seção 2.3.4) para redução de dimensionalidade e posterior cálculo da distância euclidiana entre o primeiro componente de cada sentença, o qual contém os itens com maior variação na matriz de contextos.

Além dos atributos que fazem uso dos *word embeddings*, foram elaboradas outros sete que utilizam aspectos léxicos e semânticos das sentenças, totalizando o conjunto de dez atributos que é apresentado na Tabela 5. Os detalhes da obtenção e os algoritmos de cada um destes, estão detalhados nos Apêndices B e C.

Os atributos 1, 2 e 3 (Tabela 5) utilizam relações léxicas e semânticas das sentenças, os

Tabela 5: Lista de atributos elaborados

Índice	Atributo
1	Substituição de sinônimos
2	Substituição dos hipônimos e hiperônimos nas sentenças
3	Contagem de antônimos nas sentenças
4	Proporção de palavras diferentes entre as sentenças
5	Proporção de *ngramas* em comum das sentenças
6	Proporção de palavras em comum entre as sentenças
7	Coeficiente de penalização pelo tamanho das sentenças
8	Similaridade do cosseno entre a soma dos *word embeddings*
9	Distância euclidiana entre o primeiro componente principal de cada sentença
10	Similaridade do cosseno entre os vetores TF-IDF de cada sentença

Fonte: Elaborado pelos autores.

quais foram obtidos das bases PULO e TeP para a substituição de hiperônimos, hipônimos e sinônimos nas sentenças. Os três atributos seguintes (4, 5 e 6) foram obtidos usando a contagem de palavras e a busca por unigramas, bigramas ou trigramas em ambas as sentenças, através da ferramenta WEKA de Witten et al. (2016) para encontrar termos compostos e comuns com pelo menos uma ocorrência. Em uma análise empírica, notamos que muitas das entidades nomeadas foram contabilizadas neste atributo.

Utilizou-se a equação indicada por Ferreira et al. (2016) para o cálculo da penalização de sentenças com tamanhos diferentes como atributo 7, porém o valor da similaridade usado na equação do autor foi substituído pela média aritmética das similaridades entre os *word embeddings* e TF-IDF. A opção por realizar a alteração na conjuntura original ocorre como uma alternativa para atenuar a diferença de vocabulário entre as sentenças, sem priorizar os valores de contexto ou as palavras compartilhadas entre as mesmas. A adaptação da fórmula pode ser vista na Equação 4.1, onde T corresponde ao tamanho das sentenças e $Sim(frase)$ é o valor da média.

$$Penalizacao = \begin{cases} \frac{|T(frase_1) - T(frase_2)| \times Sim(frase)}{T(frase_1)} & se\ T(frase_1) > T(frase_2) \\ \frac{|T(frase_1) - T(frase_2)| \times Sim(frase)}{T(frase_2)} & caso\ contrario \end{cases} \quad (4.1)$$

Os atributos 8, 9 e 10 foram obtidos através do GloVe e aplicação de TF-IDF (Seção 2.2.1) com o uso tanto das orações originais quanto com as variações obtidas através das substituições. A opção por realizar as modificações nas sentenças, dá-se como tentativa para redução da

esparsidade dos dados, pois a abordagem TF-IDF utiliza em seu cálculo a contagem de palavras compartilhadas entre as sentenças. Logo, quanto mais elementos compartilhados entre os textos, maior será a similaridade entre ambos.

5 RESULTADOS

Este capítulo descreve os experimentos realizados e analisa os resultados obtidos. Foi realizada uma série de experimentos para avaliar a contribuição dos diversos atributos estudados, tanto os derivados de relações linguísticas, como elementos probabilísticos e os *word embeddings*, na obtenção da similaridade semântica. Devido a natureza do problema que é estimar um valor a partir da comparação entre as sentenças, foram utilizados os algoritmos SVM (Seção 2.3.3), RNA (Seção 2.3.1) e GLM (Seção 2.3.2) para regressão linear, porém a abordagem proposta não está limitada a estes e possibilita o emprego de outros para a tarefa. De modo a padronizar a abordagem e permitir a comparação com os trabalhos relacionados na literatura, foi utilizado o conjunto de dados da ASSIN (Seção 4.3) para treinamento e teste dos modelos.

Em todos os casos tratados, devido a capacidade de processamento exigida, foi utilizado um servidor com dois processadores *E5-2620* versão 4 2.1GHz, 128 gigabytes RDIMM (2400MT/s) e placa de vídeo *Matrox G200eR2* com 16 megabytes. Além disso, também foi configurado um ambiente *R Studio Server*[1], cujas principais atribuições são tanto a geração dos modelos de aprendizado de máquina, como também o pré processamento e transformação dos dados.

Como se pode observar na Tabela 6, os resultados obtidos com os *word embeddings* isolados não foram suficientes para um bom desempenho dos modelos de aprendizagem de máquina, o que também é observado no trabalho de Hartmann (2016). Entende-se que a utilização de PCA ao invés de soma para obtenção da similaridade a partir dos *word embeddings* mantém o desempenho não satisfatório porque a redução de dimensionalidade destes pode levar à perda das nuances e peculiaridades das sentenças, ocasionando assim a perda do contexto.

Analisando os resultados da Tabela 6, observa-se que a maior Correlação de Pearson (CP) e o menor Erro Quadrático Médio (EQM) foram obtidos através dos experimentos com utilização de recursos linguísticos, tais como os antônimos e as relações de hiponímia. Entretanto, ao analisar a quantidade de antônimos por tuplas no conjunto de dados utilizado (Seção 4.3), notou-se que em raros casos foram identificadas uma ou mais relações de antonímia na mesma sentença, o que é justificado pelo baixo volume de registros da relação na base PULO. Tal fato dificultou a utilização das relações linguísticas e repercutiu no desempenho da técnica para uso dos atributos de antônimos e hiponímia. Além disso, nota-se o baixo desempenho do atributo de penalização pela diferença de tamanho entre as sentenças. Após aplicada uma análise estatística, foi constatada a não existência de correlação da penalização com o valor esperado de similaridade ($p > 0.05$).

Na Tabela 7 são apresentados os melhores resultados no estado da arte para avaliação de similaridade semântica, os quais são comparados com o atual trabalho através do conjunto de dados do ASSIN (Seção 4.3). Apesar de ser possível observar na mesma tabela que este trabalho não obteve o melhor resultado para CP ou EQM, ressaltamos que o número de *tokens* no *corpus* usado para obtenção dos *word embeddings* foi extremamente reduzido. De forma que foram

[1]Disponível em `https://www.rstudio.com/products/rstudio`.

Tabela 6: Experimentos e resultados

Índice	Atributos *	Correlação de Pearson	Erro Quadrático Médio
1	8	0.3165	0.6847
2	9	0.2641	0.7226
3	10	0.4448	0.6174
4	3	0.0355	0.7754
5	6	0.6213	0.5057
6	7,9	0.2672	0.7087
8	5,6,8,10	0.6364	0.4535
9	4,5,6,8,10	0.5782	0.5102
10	5,6,9,10	0.6357	0.4543
11	4,5,6,9,10	0.6343	0.4622
12	5,6,10	0.6160	0.4790
13	1,2,3,4,5,6,8,10	0.6394	0.4499
14	1,3,4,5,6,8,10	0.6370	0.4522
15	1,3,4,6,8,10	0.6408	0.4482
16	1,2,4,6,8,10	0.6410	0.4479
17	1,2,5,6,8,10	0.6625	**0.4302**
18	1,2,4,6,7,10	0.6625	0.4303
19	1,2,4,5,6,8,10	**0.6626**	0.4304

* A segunda coluna representa o índice dos atributos descritos na Tabela 5 .

Fonte: Elaborado pelos autores.

resultados próximos ao estado da arte, utilizando uma abordagem de aprendizado de máquina, com o conjunto de treinamento do MEV muito inferior aos trabalhos comparados.

Nos trabalhos de Cavalcanti et al. (2017) e Hartmann (2016), os autores utilizaram os *word embeddings* treinados em um *corpus* contendo cerca de 300 milhões de *tokens* coletados dos sites G1 e Wikipédia, além da utilização do *corpus* PLN-Br de Bruckschen et al. (2008) no caso de Hartmann (2016). Enquanto que foram utilizados apenas $1.584.492$ *tokens* para o treinamento dos *word embeddings* no atual trabalho, o que corresponde cerca de $0,005\%$ do que foi usado por Cavalcanti et al. (2017) e Hartmann (2016). Acreditamos que a quantidade de *tokens* pode ser uma das causas para o desempenho dos experimentos com os atributos derivados dos *word embeddings*. Porém, nossos resultados foram superiores aos de Fialho et al. (2016) para português do Brasil quando observado apenas o EQM e obtivemos resultados superiores aos de Alves, Rodrigues e Oliveira (2016) mesmo sem uma análise sintática ou reconhecimento de

Tabela 7: Comparação com o estado da arte

	Abordagem	CP	EQM
Técnica proposta	Embeddings com PCA	0.30	0.69
	Soma dos *word embeddings*	0.30	0.68
	TF-IDF	0.44	0.61
	Embeddings com PCA + TF-IDF	0.46	0.59
	Soma dos *word embeddings* + TF-IDF	0.55	0.52
	Melhor resultado da Tabela 6 *	0.66	0.43
Hartmann (2016)	Soma dos *word embeddings*	0.58	0.50
	TF-IDF	0.68	0.41
	Soma dos *word embeddings* + TF-IDF	0.70	0.38
Cavalcanti et al. (2017)	TF-IDF Matriz de similaridade Matriz binária Tamanho das sentenças	0.71	**0.37**
Fialho et al. (2016)	Soft TF-IDF Similaridades entre palavras Sobreposição de *ngramas*	**0.73**	0.63
Alves, Rodrigues e Oliveira (2016)	ASAPP	0.65	0.44
	Reciclagem	0.59	1.31

* A linha com o título "Melhor resultado da Tabela 6"corresponde à combinação dos atributos: Soma dos *word embeddings*, proporção de palavras em comum, TF-IDF, proporção de palavras diferentes, contagem de antônimos, substituição de sinônimos e relações de hiponímia.

Fonte: Elaborado pelos autores.

entidades nomeadas.

Os melhores resultados obtidos pelo presente trabalho foram com o uso da RNA configurada com 15 neurônios, taxa de aprendizagem de 0.12 , função de ativação sigmoide e treinamento durante 1000 épocas, além de realizar a substituição dos sinônimos e relações de hiponímia das sentenças. Tal recurso não afeta o sentido da frase e permite a comparação direta entre ocorrências de palavras comuns em ambas as sentenças. A abordagem descrita maximizou os resultados da técnica TF-IDF, agregando para esta um papel fundamental na obtenção da similaridade. Entretanto, é visto que apesar da métrica dos antônimos não apresentar correlação com o valor esperado de similaridade ($p > 0.05$), este demonstrou bom desempenho quando utilizado em conjunto com outros atributos, tal como é possível observar na Tabela 6. Apesar

da tentativa de usar o PCA para reduzir a dimensionalidade dos *word embeddings* e preservar sua linearidade, os resultados obtidos com o uso desta técnica foram piores do que a soma dos vetores sugerida por (MIKOLOV et al., 2013a). Portanto, acreditamos que a PCA suprime alguns contextos e não pode capturar toda a informação semântica oculta nos *word embeddings*.

Todos os experimentos utilizam um conjunto de combinações dos atributos mostrados na Tabela 5, incluindo técnicas de normalização. Em nossa experiência, a normalização dos atributos obteve seus melhores resultados com RNA, mas quando consideramos apenas os algoritmos SVM e GLM, as técnicas de MaxMin e Z-score obtiveram resultados próximos ao conjunto original. Como podemos ver na Tabela 8, a generalização dos sinônimos melhorou os resultados para a CP, demonstrando que os as estimativas de similaridade ficaram próximas ao objetivo no conjunto de dados, porém o aumento no EQM alerta que ocorreram mais erros em estimar a similaridade.

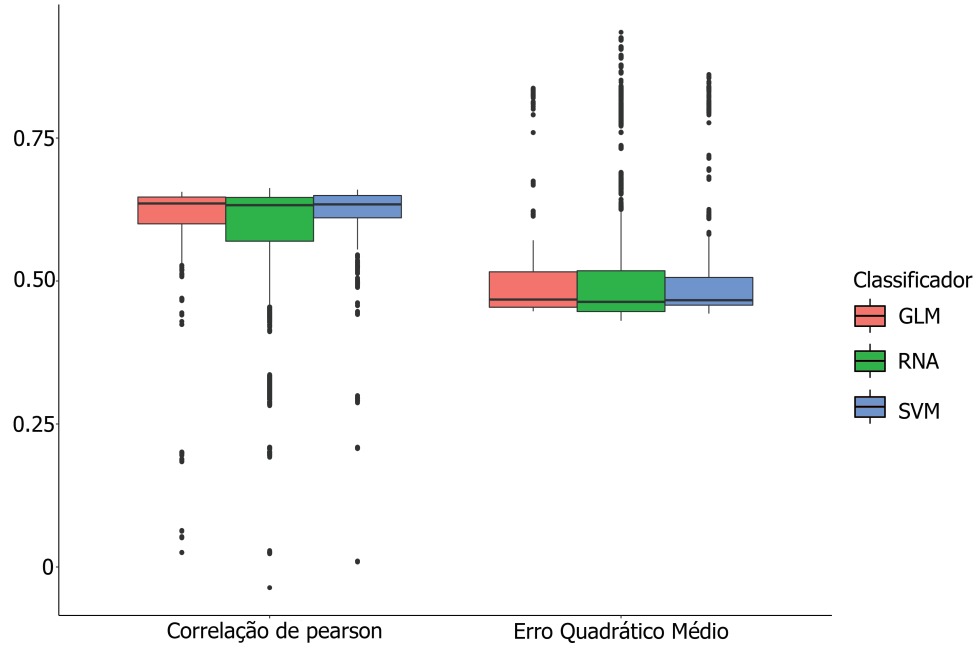
Tabela 8: Comparação de resultados com recursos linguísticos

	Correlação de Pearson		Erro Quadrático Médio	
	Média	Desvio padrão	Média	Desvio padrão
Sentenças originais	0.5864	0.0649	0.5020	0.0587
Generalização de sinônimos	0.5939	0.0674	0.5197	0.0837
Generalização de sinônimos, hiperônimos e hipônimos	0.5969	0.0676	0.5163	0.0828

É possível observar na Figura 9, a comparação dos resultados obtidos pelos algoritmos de aprendizado de máquina utilizados, onde é apresentado o melhor desempenho da RNA em relação a SVM e GLM. Embora a RNA obteve os melhores resultados em CP e EQM, o SVM teve uma mediana próxima com outros algoritmos e mostrou menor variação em seus resultados. Portanto, destacamos que o uso da RNA parece ser promissor devido a capacidade desta em modelar funções não lineares, mas são necessários outros experimentos considerando um conjunto maior de parâmetros, uma vez que nossa análise usou apenas combinações entre três épocas (500, 700 e 1000), sete neurônios ocultos (5, 7, 10, 15, 19, 22 e 30) e três taxas de aprendizagem (0.012, 0.01 e 0.12).

É necessário o destaque de alguns aspectos que representaram limitações nesta pesquisa, os quais serão abordados nos trabalhos futuros. Existem alguns recursos de apoio ao PLN em Português brasileiro que não foram aplicados, mas podem ser importantes para melhorar a qualidade dos recursos linguísticos. Um desses recursos é a OpenWordnet-PT, a qual recentemente incorporou diversas bases léxico-semânticas distintas e atualizações. Outra limitação destacada é o tamanho do corpus usado para tratar os *word embeddings*, pois acreditamos que o recurso probabilístico pode representar uma boa possibilidade de obter melhores resultados, mas até agora o tamanho do corpus utilizado é extremamente pequeno quando comparado com outros

Figura 9: Comparação entre os algoritmos de aprendizagem de máquina utilizados



Fonte: Elaborado pelos autores.

trabalhos. Uma alternativa ao treinamento dos *word embeddings* é o trabalho publicado por Hartmann et al. (2017) recentemente, onde foram disponibilizados diversos modelos de espaço vetorial obtidos com diferentes algoritmos.

6 CONCLUSÕES

Neste trabalho foi apresentada uma abordagem híbrida para avaliar a similaridade semântica entre frases curtas. Para tanto, foram integrados recursos como Modelos de Espaço Vetorial, TF-IDF, PCA e também as relações linguísticas de antonímia, hiperonímia, hiponímia e sinonímia. A integração destes recursos na abordagem proposta nos permitiu obter um conjunto de características, as quais foram utilizadas em experimentos com classificadores SVM, RNA e GLM. Os melhores resultados obtidos em todos os algoritmos de classificação utilizados contaram com o uso de aspectos linguísticos e probabilísticos, o que responde a questão de pesquisa motivadora deste trabalho e demonstra a importância da integração de informações linguísticas para avaliação de similaridade semântica no português do Brasil. Entretanto, destacamos que o melhor resultado surgiu com o uso do algoritmo RNA, o qual obteve a maior correlação de Pearson e o menor Erro Quadrático Médio registrado em todos os experimentos. Além disso, deve ser destacado que a quantidade de *tokens* no *corpus* utilizado para treinamento do algoritmo *GloVe* pode ter influenciado diretamente na proximidade das palavras e portanto, na similaridade das sentenças. Apesar disso, nota-se que foi possível obter resultados próximos ao estado da arte, mesmo com um *corpus* limitado para o treinamento do MEV, através do emprego de recursos linguísticos.

Como trabalhos futuros, destacamos a necessidade de explorar outros recursos léxico-semânticos como a o OpenWordnet (PAIVA; RADEMAKER; MELO, 2012) e avaliar o desempenho dos algoritmos SVM, RNA e GLM frente as *Long-Short Term Memory Networks* (LSTM), pois já são vistos em outros trabalhos como Mueller (2016), a capacidade destas para tratar representações e modelagens semânticas complexas com o objetivo de mensurar a similaridade entre sentenças.

6.1 Contribuições

De modo a promover a reprodutibilidade desta pesquisa, todas as ferramentas utilizadas tinham código aberto e todos os recursos elaborados a partir do desenvolvimento deste trabalho foram disponibilizados na web, bem como o código fonte contendo os procedimentos realizados para obtenção dos resultados alcançados. Deste modo, o presente trabalho trouxe como contribuição em forma de ferramentas ou recursos para área de Processamento de Linguagem Natural: um *corpus* não anotado contendo $1.584.492$ *tokens*, os *word embeddings*[1] obtidos através do *GloVe* e um serviço web[2] para consulta de informações léxico-semânticas.

Destacamos como principais contribuições deste trabalho: uma abordagem aplicável para mensurar a similaridade semântica entre frases curtas; resultados de experimentos com os algoritmos SVM, RNA, GLM e redução de dimensionalidade; e uma metodologia para disponi-

[1] Disponível em `https://github.com/albarsil/unists`.
[2] Disponível em `https://github.com/albarsil/unists-webservice-textmining`.

bilização de recursos na web. Além disso, destaca-se resultados da aplicação de contagem em relações de antonímia, da aplicação de penalização de diferença de tamanho entre sentenças, bem como do uso de relações de hiperonímia, hiponímia e sinonímia no apoio de modelos de espaço vetorial para análise de similaridade.

REFERÊNCIAS

AGIRRE, E.; BANEA, C.; CER, D.; DIAB, M.; GONZALEZ-AGIRRE, A.; MIHALCEA, R.; RIGAU, G.; WIEBE, J. Semeval-2016 task 1: semantic textual similarity, monolingual and cross-lingual evaluation. **International workshop on semantic evaluation**, San Diego, California, p. 497–511, 2016.

AGIRRE, E.; BANEAB, C.; CARDIEC, C.; CERD, D.; DIABE, M.; GONZALEZ-AGIRREA, A.; GUOF, W.; LOPEZ-GAZPIOA, I.; MARITXALARA, M.; MIHALCEAB, R.; OTHERS. Semeval-2015 task 2: semantic textual similarity, english, spanish and pilot on interpretability. In: INTERNATIONAL WORKSHOP ON SEMANTIC EVALUATION, 2015, Denver, Colorado. **Anais...** Association for Computational Linguistics, 2015. p. 252–263.

AGIRRE, E.; CER, D.; DIAB, M.; GONZALEZ-AGIRRE, A. SemEval-2012 Task 6: a pilot on semantic textual similarity. In: INTERNATIONAL WORKSHOP ON SEMANTIC EVALUATION, 2012, Montréal, Canada. **Anais...** Association for Computational Linguistics, 2012. n. 3, p. 385–393.

AGIRRE, E.; CER, D.; DIAB, M.; LOPEZ-GAZPIO, I.; SPECIA, L. SemEval-2017 Task 1: semantic textual similarity multilingual and crosslingual focused evaluation. In: INTERNATIONAL WORKSHOP ON SEMANTIC EVALUATION, 2017, Vancouver, Canada. **Anais...** Association for Computational Linguistics, 2017. p. 1–5.

AGIRRE, E.; SOROA, A. Personalizing PageRank for word sense disambiguation. In: CONFERENCE OF THE EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 2009, Athens, Greece. **Anais...** Association for Computational Linguistics, 2009. p. 33–41.

AGRESTI, A. Generalized Linear Models. In: **An Introduction to Categorical Data Analysis**. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2006. p. 65–98.

ALVES, A. O.; FERRUGENTO, A.; LOURENÇO, M.; RODRIGUES, F. ASAP: automatic semantic alignment for phrases. In: INTERNATIONAL WORKSHOP ON SEMANTIC EVALUATION, 2014, Dublin, Ireland. **Anais...** Association for Computational Linguistics and Dublin City University, 2014. n. SemEval, p. 104–108.

ALVES, A. O.; RODRIGUES, R.; OLIVEIRA, H. G. ASAPP: alinhamento semântico automático de palavras aplicado ao português. **Linguamática**, [S.l.], v. 8, n. 2, p. 43–58, 2016.

ARORA, S.; LIANG, Y.; MA, T. A simple but tough to beat baseline for sentence embeddings. In: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS, 2017, Palais des Congrès Neptune, Toulon, France. **Anais...** [S.l.: s.n.], 2017. p. 1–14.

BAKER, C. F.; FILLMORE, C. J.; LOWE, J. B. The Berkeley FrameNet Project. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS -, 36., 1998, Morristown, NJ, USA. **Proceedings...** Association for Computational Linguistics, 1998. v. 1, p. 86.

BANERJEE, S.; PEDERSEN, T. Extended gloss overlaps as a measure of semantic relatedness. In: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, 2003, Acapulco, Mexico. **Anais...** Morgan Kaufmann Publishers Inc., 2003. p. 805–810.

BARBOSA, L.; CAVALIN, P.; GUIMARÃES, V.; KORMAKSSON, M. Blue Man Group at ASSIN: using distributed representations for semantic similarity and entailment recognition. **Linguamática**, [S.l.], v. 8, n. 2, p. 15–22, 2016.

BARROS, L. A. **Curso básico de terminologia**. [S.l.]: Edusp, 2004. (Acadêmica (Edusp)).

BENGIO, Y.; DUCHARME, R.; VINCENT, P.; JANVIN, C. A Neural Probabilistic Language Model. **The Journal of Machine Learning Research**, [S.l.], v. 3, p. 1137–1155, 2003.

BRUCKSCHEN, M.; MUNIZ, F.; GUILHERME, J.; De Souza, C.; FUCHS, J. T.; INFANTE, K.; MUNIZ, M.; GONÇALVES, P. N.; VIEIRA, R.; ALUÍSIO, S. **Anotação Linguística em XML do Corpus PLN-BR**. São Carlos, Brasil: Universidade de São Paulo, 2008.

BRYCHCÍN, T.; SVOBODA, L. UWB at SemEval-2016 Task 1: semantic textual similarity using lexical, syntactic, and semantic information. In: INTERNATIONAL WORKSHOP ON SEMANTIC EVALUATION, 2016, San Diego, California. **Anais...** Association for Computational Linguistics, 2016. p. 588–594.

BURGESS, C.; LIVESAY, K.; LUND, K. Explorations in context space: words, sentences, discourse. **Discourse Processes**, [S.l.], v. 25, n. 2-3, p. 211–257, 1998.

CANÇADO, M. **Manual de Semântica: noções básicas e exercícios**. Linguistic. ed. [S.l.]: Contexto, 2012. 183 p.

CAVALCANTI, A. P.; MELLO, R. F. L. de; FERREIRA, M. A. D.; ROLIM, V. B.; TENORIO, J. V. S. Statistical and Semantic Features to Measure Sentence Similarity in Portuguese. In: BRAZILIAN CONFERENCE ON INTELLIGENT SYSTEMS (BRACIS), 2017., 2017. **Anais...** IEEE, 2017. p. 342–347.

CHOPRA, S.; HADSELL, R.; Y., L. Learning a similiarty metric discriminatively, with application to face verification. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2005, Washington, USA. **Anais...** IEEE Computer Society, 2005. p. 539–546.

CRUSE, D. A. **Lexical Semantics**. [S.l.]: Cambridge University Press, 1986. 310 p. (Cambridge Textbooks in Linguistics).

CRUSE, D. A. Polysemy and related phenomena from a cognitive linguistic viewpoint. In: SAINT-DIZIER, P.; VIEGAS, E. (Ed.). **Computational Lexical Semantics**. [S.l.]: Cambridge University Press, 1995. p. 33–49. (Studies in Natural Language Processing).

DAS, D.; SCHNEIDER, N.; CHEN, D.; SMITH, N. A. N. Probabilistic frame-semantic parsing. **Annual Conference of the North American Chapter of the Association for Computational Linguistics**, Los Angeles, California, v. 3, n. June, p. 948–956, 2010.

DOLAMIC, L.; SAVOY, J. Brief communication: when stopword lists make the difference. **Journal of the American Society for Information Science and Technology**, New York, USA, v. 61, n. 1, p. 200–203, 2010.

EVANS, V.; GREEN, M. **Cognitive Linguistics: an introduction**. [S.l.]: L. Erlbaum, 2006.

FELLBAUM, C. **WordNet: an electronic lexical database**. [S.l.: s.n.], 1998. 423 p. v. 71, n. 3.

FERREIRA, R.; LINS, R. D.; SIMSKE, S. J.; FREITAS, F.; RISS, M. Assessing sentence similarity through lexical, syntactic and semantic analysis. **Computer Speech & Language**, London, United Kingdom, v. 39, p. 1–28, 2016.

FIALHO, P.; MARQUES, R.; MARTINS, B.; COHEUR, L.; QUARESMA, P. INESC-ID@ASSIN: medição de similaridade semântica e reconhecimento de inferência textual. **Linguamática**, [S.l.], v. 8, n. 2, p. 33–42, 2016.

FILLMORE, C. J. Background to Framenet. **International Journal of Lexicography**, [S.l.], v. 16, n. 3, p. 235–250, 2003.

CONTEXTO (Ed.). **Introdução à Linguística: objetos teóricos**. 6. ed. São Paulo, Brasil: Contexto, 2010. 230 p.

FONSECA, E. R.; BORGES, L.; SANTOS, D.; CRISCUOLO, M.; ALUÍSIO, S. M. Visão Geral da Avaliação de Similaridade Semântica e Inferência Textual. **Linguamática**, [S.l.], v. 8, n. 2, p. 15–22, 2016.

FREIRE, J.; PINHEIRO, V.; FEITOSA, D. LEC_UNIFOR no ASSIN: flexsts um framework para similaridade semântica textual. **Linguamática**, [S.l.], v. 8, n. 2, p. 23–31, 2016.

FRIEDMAN, J. H. Stochastic gradient boosting. **Computational Statistics & Data Analysis**, Amsterdam, Netherlands, v. 38, n. 4, p. 367–378, 2002.

GOMAA, W.; FAHMY, A. A Survey of Text Similarity Approaches. **International Journal of Computer Applications**, [S.l.], v. 68, n. 13, p. 13–18, apr 2013.

GOMES, M. M.; BELTRAME, W.; CURY, D. Automatic Construction of Brazilian Portuguese WordNet. In: NATIONAL MEETING ON ARTIFICIAL AND COMPUTATIONAL INTELLIGENCE, 2013, Fortaleza, Brasil. **Anais...** IEEE Computer Society Press, 2013.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. The WEKA data mining software: an update. **SIGKDD Explorations**, New York, USA, v. 11, n. 1, p. 10–18, 2009.

HÄNIG, C.; REMUS, R.; DE, X.; PUENTE, L. ExB Themis: extensive feature extraction from word alignments for semantic textual similarity. In: INTERNATIONAL WORKSHOP ON SEMANTIC EVALUATIONVALUATION, 2015, Denver, USA. **Anais...** Association for Computational Linguistics, 2015. p. 264–268.

HARRIS, Z. S. Distributional Structure. **WORD**, [S.l.], v. 10, n. 2-3, p. 146–162, aug 1954.

HARTMANN, N.; FONSECA, E.; SHULBY, C.; TREVISO, M.; RODRIGUES, J.; ALUISIO, S. Portuguese Word Embeddings: evaluating on word analogies and natural language tasks. **Symposium in Information and Human Language Technology**, [S.l.], aug 2017.

HARTMANN, N. S. Solo Queue at ASSIN : combinando abordagens tradicionais e emergentes. **Linguamática**, [S.l.], v. 8, n. 2, p. 59–64, 2016.

PEARSON (Ed.). **Neural Networks and Learning Machines**. New Jersey, USA: Pearson, 2008. 906 p. v. 3.

HEBB, D. O. **The Organization of Behavior - A Neuropsychological Theory**. New York, USA: JOHN WILEY & SONS, Inc., 1949. 365 p.

HIRSCH, E. **Dividing Reality**. [S.l.]: Oxford University Press, 1996. (Oxford paperbacks).

HOTELLING, H. Analysis of a complex of statistical variables into principal components. **Journal of Educational Psychology**, [S.l.], v. 24, n. 6, p. 417–441, 1933.

JOACHIMS, T. Text categorization with Support Vector Machines: learning with many relevant features. In: **Machine Learning**. London, United Kingdom: Springer-Verlag, 1998. v. 1398, n. LS-8 Report 23, p. 137–142.

JOLLIFFE, I. **Principal Component Analysis**. 2nd ed.. ed. New York, USA: Springer-Verlag, 2002. 487 p. (Springer Series in Statistics).

JOLLIFFE, I. Principal Component Analysis. In: **Wiley StatsRef Statistics Reference Online**. Chichester, United Kingdom: John Wiley & Sons, Ltd, 2014.

JONES, K. S. A statistical interpretation of term specificity and its application in retrieval. **Journal of Documentation**, London, United Kingdom, v. 60, n. 1, p. 493–502, 2004.

KAO, A.; POTEET, S. R. (Ed.). **Natural Language Processing and Text Mining**. London, United Kingdom: Springer, 2007. 1–265 p.

KASHYAP, A.; HAN, L.; YUS, R.; SLEEMAN, J.; SATYAPANICH, T.; GANDHI, S.; FININ, T. Robust semantic text similarity using LSA, machine learning, and linguistic resources. **Language Resources and Evaluation**, [S.l.], v. 50, n. 1, p. 125–161, 2016.

MATOS, A. P.; SILVA, J. M.; PERETTI, L.; OLENIKI, M. L. (Ed.). **Fundamentos da Metodologia Cientifica**. Digital. ed. Petrópolis, Brasil: Editora Vozes Ltda., 2011. 185 p.

LEECH, G. N. **Semantics: the study of meaning**. [S.l.]: Penguin, 1981. 383 p. (A pelican original).

LUGER, G. F. **Inteligencia Artificial**. [S.l.]: Pearson, 2013.

LYONS, J. **Linguistique générale: introduction à la linguistique théorique**. [S.l.]: Larousse, 1970. 384 p. (Collection Langue et langage).

LYONS, J. **Semantics**. [S.l.]: Cambridge University Press, 1977. 371 p. v. 1.

LYONS, J. **Introduction to Theoretical Linguistics**. [S.l.]: Cambridge University Press, 1992. 519 p.

LYONS, J. **Linguistic Semantics: an introduction**. [S.l.]: Cambridge University Press, 1995. 396 p.

MACKAY, D. J. C. Introduction to Gaussian processes. **Neural Networks and Machine Learning**, Berlin, Germany, v. 168, n. 1996, p. 133–165, 1998.

MANNING, C. D.; SCHÜTZE, H. **Foundations of Natural Language Processing**. USA: MIT Press, 2000. 678 p.

MARRAFA, P. Portuguese WordNet: general architecture and internal semantic relations. **DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada**, São Paulo, Brasil, v. 18, n. SPE, p. 131–146, 2002.

MAZIERO, E. G.; PARDO, T. a. S.; Di Felippo, A.; SILVA, B. C. Dias-da. A base de dados lexical e a interface web do TeP 2.0. In: BRAZILIAN SYMPOSIUM ON MULTIMEDIA AND THE WEB, 2008, New York, USA. **Anais...** ACM Press, 2008. p. 390.

MCCULLAGH, P. Generalized linear models. **European Journal of Operational Research**, [S.l.], v. 16, n. 3, p. 285–292, jun 1984.

MCCULLAGH, P.; NELDER, J. A. **Generalized Linear Models**. [S.l.]: Chapman and Hall, 1983. 274 p. (Monographs on Statistics and Applied Probability).

METZLER, D.; DUMAIS, S.; MEEK, C. Similarity Measures for Short Segments of Text. In: EUROPEAN CONFERENCE ON IR RESEARCH, 2007, Rome, Italy. **Anais...** Springer-Verlag Berlin Heidelberg, 2007. v. 4425, p. 16–27.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. **Efficient Estimation of Word Representations in Vector Space**. [S.l.]: ArXiv, 2013. 104–108 p. n. 1.

MIKOLOV, T.; SUTSKEVER, I.; CHEN, K.; CORRADO, G.; DEAN, J. Distributed Representations of Words and Phrases and their Compositionality. In: INTERNATIONAL CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 2013, Lake Tahoe, USA. **Anais...** Curran Associates Inc, 2013. p. 3111–3119.

MUELLER, J. Siamese Recurrent Architectures for Learning Sentence Similarity. In: CONFERENCE ON ARTIFICIAL INTELLIGENCE, 2016, Phoenix, USA. **Anais...** AAAI Press, 2016. n. 2012, p. 2786–2792.

NADKARNI, P. M.; OHNO-MACHADO, L.; CHAPMAN, W. W. Natural language processing: an introduction. **Journal of the American Medical Informatics Association**, [S.l.], v. 18, n. 5, p. 544–551, sep 2011.

NELDER, J. A.; BAKER, R. J. **Generalized Linear Models**. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2006.

NELDER, J. A.; WEDDERBURN, R. W. M. **Generalized Linear Models**. [S.l.]: General J. R. Statist. Soc. A, 1972. 370–384 p. v. 13517213, n. 3.

OLIVEIRA, H. G.; GOMES, P. Onto. PT: automatic construction of a lexical ontology for portuguese. In: STARTING AI RESEARCHERS SYMPOSIUM, 2010, Amsterdam, Netherlands. **Anais...** IOS Press Amsterdam, 2010. v. 222, p. 199–211.

OLIVEIRA, H. G.; PAIVA, V. de; FREITAS, C.; RADEMAKER, A.; REAL, L.; SIMÕES, A. As wordnets do Português. **Oslo Studies in Language**, [S.l.], v. 7, n. 1, p. 97–424., 2015.

PAIVA, V.; RADEMAKER, A.; MELO, G. OpenWordNet-PT: an open brazilian wordnet for reasoning. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, 2012, Mumbai, India. **Anais...** COLING 2012 Organizing Committee, 2012. p. 353–360.

PALMER, F. R. **Semantics: a new outline**. [S.l.]: Cambridge University Press, 1977. 164 p.

PALMER, F. R. **Semantics**. [S.l.]: Cambridge University Press, 1981. 232 p.

PEARSON, K. LIII. On lines and planes of closest fit to systems of points in space. **Philosophical Magazine Series 6**, Londres, United Kingdom, v. 2, n. 11, p. 559–572, nov 1901.

PENNINGTON, J.; SOCHER, R.; MANNING, C. Glove: global vectors for word representation. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2014, Stroudsburg, USA. **Anais...** Association for Computational Linguistics, 2014. p. 1532–1543.

PIANTA, E.; BENTIVOGLI, L.; GIRARDI, C. MultiWordNet: developing an aligned multilingual database. In: INTERNATIONAL CONFERENCE ON GLOBAL WORDNET, 2002, Mysore, India. **Anais...** [S.l.: s.n.], 2002. n. 1996, p. 21–25.

PRADHAN, N.; GYANCHANDANI, M.; WADHVANI, R. A Review on Text Similarity Technique used in IR and its Application. **International Journal of Computer Applications**, [S.l.], v. 120, n. 9, p. 29–34, jun 2015.

RAJARAMAN, A.; ULLMAN, J. D. **Mining of Massive Datasets**. New York, USA: Cambridge University Press, 2011. 328 p. v. 67.

RODRIGUES, R.; OLIVEIRA, H. G.; GOMES, P. LemPORT: a high-accuracy cross-platform lemmatizer for portuguese. In: SYMPOSIUM ON LANGUAGES, APPLICATIONS AND TECHNOLOGIES, 2014, Bragança, Portugal. **Anais...** [S.l.: s.n.], 2014. p. 267.

SCHULER, K. K. VerbNet: a broad-coverage, comprehensive verb lexicon. **Dissertation Abstracts International, B: Sciences and Engineering**, [S.l.], v. 66, n. 6, 2005.

Semeval-2014 task 10: Multilingual semantic textual similarity, 2014, Dublin, Ireland. **Anais...** Association for Computational Linguistics and Dublin City University, 2014. p. 81–91.

SILVA, B. C. Dias-da; OLIVEIRA, M. F. de; MORAES, H. R. de. Groundwork for the Development of the Brazilian Portuguese Wordnet. In: **Advances in natural language processing**. Berlin, Germany: Springer, 2002. v. 2389, p. 189–196.

SIMÕES, A.; GUINOVART, X. Bootstrapping a Portuguese WordNet from Galician, Spanish and English Wordnets. In: INTERNATIONAL CONFERENCE OF ADVANCES IN SPEECH AND LANGUAGE TECHNOLOGIES FOR IBERIAN LANGUAGES, 2014, Cham, Germany. **Anais...** Springer International Publishing, 2014. p. 239–248.

SPRINGER (Ed.). **The Nature of Statistical Learning Theory**. New York, USA: Springer, 1995. 314 p.

VAPNIK, V. N.; CHERVONENKIS, A. Y. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. **Theory of Probability & Its Applications**, [S.l.], v. 16, n. 2, p. 264–280, jan 1971.

KAUFMANN, M. (Ed.). **Data Mining: practical machine learning tools and techniques**. 4. ed. [S.l.]: Morgan Kaufmann, 2016. 654 p.

APÊNDICE A EXEMPLOS DE SENTENÇAS

"●"

"● ●"

", uma sobre o em , ."

"... () não possuía na sua composição original?"

"... há de em ?"

"... os das de pertencentes ao não se desenvolvem na ?"

"... nos livros originais de , a frase "Elementar, meu caro Watson"nunca aparece?"

"... o é proveniente da ?"

"A () é uma sobre o , em , na , famosa por ter uma quantidade de lojas (principalmente ourivesarias e joalharias) ao longo de todo o tabuleiro."

"Acredita-se que tenha sido construída ainda na e era feita originalmente de madeira. Foi destruída pelas cheias de 1333 e reconstruída em 1345, com projecto da autoria de . Consiste em três arcos, o maior deles com 30 metros de diâmetro. Desde sempre alberga lojas e mercadores, que mostravam as mercadorias sobre bancas, sempre com a autorização do , a autoridade municipal de então. Diz-se que a palavra teve ali origem. Quando um mercador não conseguia pagar as dívidas, a mesa () era quebrada () pelos soldados. Essa prática era chamada ."

"Durante a , a ponte não foi danificada pelos alemães. Acredita-se que tenha sido uma ordem direta de ."

"O corpo de funcionários da sede é composto por delegações nacionais dos países membros e inclui escritórios civis e militares e missões diplomáticas e diplomatas de países parceiros, bem como o Secretariado Internacional e Estado-Maior Internacional cheias de membros do serviço das forças armadas de estados membros. não Governamentais grupos de cidadãos também têm crescido em apoio da NATO, em geral, sob a bandeira do movimento Conselho Atlântico/Associação do Tratado Atlântico."

"As operações militares da organização são dirigidas pelo Presidente do Comitê Militar da OTAN e divididos em dois Comandos Estratégicos comandados por um oficial sênior dos e um oficial sênior francês, assistida por funcionários de toda a OTAN. Os Comandantes Estratégicos são responsáveis ●●perante o Comitê Militar pela direção geral e coordenação de todos os assuntos militares da Aliança dentro de suas áreas de comando."

"A delegação de cada país inclui um representante militar, um oficial superior das forças armadas de cada país, apoiado pelo Estado-Maior Militar Internacional. Juntos, os representantes militares formam o Comitê Militar, órgão responsável por recomendar às autoridades políticas da OTAN as medidas consideradas necessárias para a defesa comum da área de responsabilidade da organização. Sua função principal é fornecer orientação e aconselhamento sobre política e estratégia militar. Ele fornece a orientação sobre assuntos militares aos Comandantes Estratégicos da OTAN, cujos representantes participam nas suas reuniões e são responsáveis pela condução geral dos assuntos militares da Aliança, sob a autoridade do conselho. O presidente

do Comitê Militar da OTAN é o , desde 2012."

"Como o Conselho, de vez em quando o Comitê Militar também atende a um nível superior, ou seja, ao nível dos Chefes de Estado-Maior, o oficial militar mais graduado das forças armadas de cada país membro. Até 2008, o Comitê Militar atuava sem a presença da , devido à decisão do país de retirar-se da estrutura militar integrada da OTAN em 1966. Os franceses voltaram ao comitê em 1995. Até a França voltar para a OTAN, o país não era representado na Comissão de Planejamento de Defesa e isso levou a conflitos entre os franceses e os membros da Aliança, como foi o caso da liderança da . O trabalho operacional do comitê é suportado pelo Estado-Maior Militar Internacional ."

"A estrutura de comando da OTAN evoluiu durante a e no período posterior. Uma estrutura militar integrada da OTAN foi estabelecida pela primeira vez em 1950, quando se tornou claro que a organização precisava melhorar suas defesas a longo prazo contra um potencial ataque . Em abril de 1951, o Comando Aliado da Europa e sua sede foram estabelecidos; depois, quatro sedes subordinadas foram adicionados no norte, no centro, no sul da e na região ."

"De 1997-2003 os Comandantes Estratégicos foram o Comandante Supremo Aliado da Europa e o Comandante Supremo Aliado do Atlântico, mas o arranjo atual é o de separar a responsabilidade de comando entre o Comando Aliado da Transformação, responsável pela transformação e formação das forças da OTAN e de operações do Comando Aliado, responsável pelas operações da OTAN no mundo inteiro. A partir do final de 2003, a organização tem reestruturado como ela comanda e implanta as suas tropas através da criação de vários destacamentos rápidos, como os , bem como altos forças navais de prontidão, que todos os relatórios para as Operações do Comando Aliado."

"(, de - Rio de Janeiro, de) foi um de , chegando ao 9º (vermelha)."

"O Fadda nasceu, viveu e morreu em . Homem humilde, conhecedor profundo do Jiu jitsu e o pioneiro a levar a "arte suave"para o . Quando, aos 17 anos, era da , Oswaldo Fadda começou a treinar Jiu-Jitsu e foi o melhor pupilo do professor , que fez parte do pequeno grupo de alunos de , conhecido como Conde Koma, introdutor do jiu jitsu no , em , na cidade de Belém, no estado do ."

"O Grande Mestre Oswaldo Baptista Fadda nasceu no Rio de Janeiro no bairro de Bento Ribeiro filho de chegados ao no início do . Praticamente respirou jiu-jitsu e era muito conhecido também por ser um homem de família e bem humilde e de um conhecimento imenso da "arte suave". Fez muitas amizades quando vivo e, sendo o primeiro a iniciar suas aulas no bairro, gerou o título de pioneiro da arte no Rio e adjacências do da ."

APÊNDICE B DETALHAMENTO DE ATRIBUTOS

O presente capítulo descreve minuciosamente os passos para obtenção de cada atributo desenvolvido, presente na Tabela 5 e que compõe a técnica proposta na Seção 4.4. De modo a promover um melhor entendimento a respeito, serão apresentados exemplos práticos e teóricos de utilização. Para acompanhar os passos da parte prática, serão utilizadas as seguintes sentenças:

1. `A comissão apura denúncias de abuso e exploração sexual em meninas da comunidade quilombola.`
2. `O grupo apura denúncias de abusos e exploração sexual de crianças da Comunidade Quilombola.`

Dados os exemplos acima, inicialmente é realizada a substituição de sinônimos através da verificação se cada palavra (ou seus sinônimos) de uma frase está contida em outra. De modo a obter a generalização de ambas as sentenças, tanto o item 1, quanto o item 2 são submetidos para execução no Algoritmo 1.

Algoritmo 1: SUBSTITUIÇÃO DE SINÔNIMOS

Entrada: $frase1, frase2$
Saída: $frase1$ generalizada

```
1  início
2      para cada palavra ∈ frase1 faça
3          se palavra ∉ frase2 então
4              para cada sin ∈ Sinonimos(palavra) faça
5                  se sin ∈ frase2 então
6                      frase1[palavra] ← sin
7                  fim
8              fim
9          fim
10     fim
11 fim
12 retorna frase1
```

Uma vez que ambas as sentenças passam pela primeira etapa de generalização de sinônimos, ocorre a generalização as relações de hiponímia e hiperonímia nas sentenças. Como pode ser visto no Algoritmo 2, temos uma otimização das verificações, pois se uma palavra possuí hiperônimos, então ela é um hipônimo destes. Além disso, para solucionar os casos em que os termos de uma frase são hiperônimos, mas não estão contidos na outra, optamos por submeter ambas as sentenças para generalização. Deste modo, também é verificado se uma ou mais palavras da segunda frase possuem hiperônimos na primeira.

Algoritmo 2: SUBSTITUIÇÃO DE HIPERÔNIMOS

Entrada: $frase1, frase2$

Saída: $frase1$ generalizada

```
1  início
2      para cada palavraF1 ∈ frase1 faça
3          para cada palavraF2 ∈ frase2 faça
4              se palavraF2 ∉ frase1 então
5                  se palavraF2 ∈ Hiperonimos(palavraF1) então
6                      frase1[palavraF1] ← palavraF2
7                  fim
8              fim
9          fim
10     fim
11 fim
12 retorna frase1
```

Considerando as sentenças de exemplo, após o processamento através dos Algoritmos 1 e 2 temos como resultado as frases abaixo:

1. `comissao apurar denunciar abusar exploracao sexual criancas comunidade quilombola`
2. `comissao apurar denunciar abusos exploracao sexual criancas comunidade quilombola`

A verificação de antônimos segue o mesmo princípio que os sinônimos, como é possível observar no Algoritmo 3. Para tanto, é realizado uma varredura em cada palavra de uma frase, verificando a existência de um antônimo desta na outra sentença. Caso exista um ou mais, o valor de contagem é acrescido em uma unidade.

Algoritmo 3: CONTAGEM DE ANTÔNIMOS

Entrada: $frase1, frase2$

Saída: A quantidade de antônimos presentes em ambas as sentenças

```
1  início
2      count = 0
3      para cada palavraF1 ∈ frase1 faça
4          para cada palavraF2 ∈ frase2 faça
5              se PathLength(palavraF1, PalavraF2, antonimo) == 1 então
6                  count ← count + 1
7              fim
8          fim
9      fim
10 fim
11 retorna count
```

Onde a função $PathLength$ retorna o caminho mais curto entre as duas palavras, através de uma relação da base PULO (Seção 2.2.3).

As proporções de palavras em comum, diferentes e de *ngramas* deram-se através das bibliotecas da ferramenta WEKA para encontrar termos com pelo menos uma ocorrência em ambas as sentenças. Considerando $T(frase1)$ e $T(frase2)$ como a quantidade de palavras das frases 1 e 2, bem como Uni como os *unigramas* compartilhados. As proporções de palavras iguais ($PropI$) e diferentes ($PropD$) entre as sentenças são obtidas através das Equações B.1 e B.2 respectivamente.

$$PropI = \frac{Uni}{T(frase1) + T(frase2)} \tag{B.1}$$

$$PropD = 1 - PropI \tag{B.2}$$

A proporção de *ngramas* em comum ocorreu através da busca por bigramas e trigramas compartilhados, com pelo menos uma ocorrência em ambas as sentenças. Em uma análise empírica, notamos que algumas entidades nomeadas foram contabilizadas neste atributo.

Outra medida utilizada foi a penalização por diferença entre os tamanhos das sentenças, a qual foi realizada através da Equação 4.1. Porém notamos que este atributo não trouxe contribuições significativas para o aumento das métricas de qualidade que foram utilizadas para mensurar o desempenho da técnica na avaliação de similaridade.

Uma parte fundamental para a técnica proposta é a criação da matriz *word embeddings* das sentenças, a qual é obtida através do Algoritmo 4. Além dessa, também são utilizados os vetores TF-IDF (Apêndice C) para obtenção das medidas descritas na Seção 4.4.

Algoritmo 4: COMPOSIÇÃO DA MATRIZ DE *word embeddings*

Entrada: $word\_embeddings, frase$

Saída: A matriz de *word embeddings* da $frase$

```
1  início
2      frase_matriz;
3      para cada palavra ∈ frase faça
4          se palavra ∈ word_embeddings então
5              frase_matriz[palavra] ← word_embeddings[palavra]
6          fim
7          senão
8              frase_matriz[palavra] ← 0
9          fim
10     fim
11 fim
12 retorna frase_matriz
```

Em porte da matriz de *word embeddings*, é aplicada a Equação 3.5, também descrita no Algoritmo 5.

Algoritmo 5: VETORES DE *word embeddings* DA SENTENÇA

Entrada: $frase\_matriz$

Saída: O vetor de *word embeddings* da $frase$

```
1  início
2      frase_vetor;
3      para cada palavra ∈ frase_matriz faça
4          frase_vetor[palavra] ← Soma(frase_matriz[palavra])
5      fim
6  fim
7  retorna frase_vetor
```

Uma vez que todos as transformações e medidas foram realizadas, são gerados os atributos descritos na Tabela 5, bem como apresentados na Figura 10. Através dos dados numéricos obtidos e com exemplos apresentados na imagem, é treinado e testado um modelo de regressão linear para obter um valor continuo de similaridade.

No presente capítulo foram apresentadas as transformações básicas necessárias para os cálculos de similaridade, os quais são obtidos a partir dos vetores de *word embeddings* (Tabela 10), dos principais componentes (Tabela 11) e dos vetores TF-IDF (Tabela 12). Uma descrição detalhada dos algoritmos utilizados para os cálculos de similaridades é apresentada no Apêndice C.

Figura 10: Exemplos de valores dos atributos gerados

embeddings_sum	embeddings_pca	tfidf	penality	ngram_proportion	common_word_proportion	uncommon_word_proportion	antonym	target
155.6328	14.231249	0.009394397	1.8596693	0.00000000	0.16666667	0.8333333	0	3.50
1117.8370	16.375038	0.009273956	2.1397087	0.10714286	0.39285714	0.6071429	1	4.75
1361.8454	5.358973	0.009590114	0.7006795	0.04545455	0.22727273	0.7727273	0	3.50
739.3958	22.291296	0.014796176	2.9129218	0.00000000	0.20000000	0.8000000	0	3.00
1038.5972	10.381955	0.011241159	1.3569492	0.09090909	0.13636364	0.8636364	0	2.50
656.6937	21.602487	0.005058522	2.8223053	0.00000000	0.17391304	0.8260870	0	3.00
749.3962	10.162141	0.005748320	1.3278757	0.16666667	0.20833333	0.7916667	0	2.25
584.1145	20.128620	0.011742997	2.6302078	0.17647059	0.29411765	0.7058824	0	3.75
824.6538	12.675939	0.005604612	1.6562486	0.00000000	0.15789474	0.8421053	0	1.25
840.0257	19.680160	0.010275122	2.5715288	0.04166667	0.16666667	0.8333333	2	2.75
1184.9493	7.905620	0.008540361	1.0332843	0.00000000	0.16666667	0.8333333	0	1.00
406.3241	18.414806	0.011624381	2.4063212	0.10000000	0.25000000	0.7500000	0	2.75
1166.1271	5.081505	0.007991762	0.6643288	0.08333333	0.29166667	0.7083333	1	3.25
513.7139	15.528182	0.008387855	2.0290245	0.04761905	0.19047619	0.8095238	0	2.75
1116.3463	7.057018	0.004981877	0.9221974	0.00000000	0.10000000	0.9000000	0	2.00
247.2650	15.716737	0.006165073	2.0535106	0.00000000	0.11111111	0.8888889	0	3.50
1043.9155	13.347094	0.006323152	1.7439698	0.04000000	0.24000000	0.7600000	0	2.00

Fonte: Elaborado pelos autores.

APÊNDICE C TRANSFORMAÇÃO TF-IDF E CÁLCULOS DE SIMILARIDADE

Os códigos-fonte demonstrados neste capítulo foram desenvolvidos em linguagem R[1], utilizando os pacotes *text2vec*[2] e ou *stats*[3]. A opção pelo uso de ferramentas já existentes deve-se a ampla utilização destas no meio acadêmico, bem como devido as suas facilidades de manuseio.

No primeiro exemplo (Tabela 9), é apresentado o código utilizado para obter dos vetores TF-IDF de ambas as sentenças. Nas Tabelas 10, 11 e 12 são apresentados os trechos de códigos elaborados para obtenção das similaridades dos vetores de *word embeddings*, primeiros componentes de cada sentença e dos vetores TF-IDF, respectivamente.

Tabela 9: Implementação da transformação TF-IDF

Código

```
require(text2vec)
local_sentence_token <- itoken(
  c(frase1, frase2), tokenizer = space_tokenizer)

local_sentence_vectorizer = vocab_vectorizer(
  create_vocabulary(local_sentence_token), TRUE)

local_dtm <- create_dtm(
  local_sentence_token,
  local_sentence_vectorizer
)

tfidf_matrix <- fit_transform(local_dtm, TfIdf$new())

tfidf_matrix <- as.matrix(tfidf_matrix)
tfidf_matrix <- data.frame(
  frase1 = paste(as.character(tfidf_matrix[1,]), collapse=";"),
  frase2 = paste(as.character(tfidf_matrix[2,]), collapse=";")
)
```

Fonte: Elaborado pelos autores.

[1]Disponível em https://cran.r-project.org/.

[2]Disponível em https://cran.r-project.org/web/packages/text2vec/index.html.

[3]Disponível em https://stat.ethz.ch/R-manual/R-devel/library/stats/html/stats-package.html.

Tabela 10: Implementação de cálculo de similaridade através dos vetores de *word embeddings*

Código

```
require(text2vec)
sim2(t(frase1_embeddings_vector), t(frase2_embeddings_vector),
        method = "cosine", norm = "none")
```

Fonte: Elaborado pelos autores.

Tabela 11: Implementação de cálculo de similaridade através de PCA

Código

```
require(stats)
require(text2vec)

x <- princomp(x, cor = FALSE)
y <- princomp(y, cor = FALSE)

x <- as.matrix(x$scores)
y <- as.matrix(y$scores)

x <- x[,1]
y <- y[,1]

dist2(t(x), t(y), method = "euclidean", norm = "none")
```

Fonte: Elaborado pelos autores.

Tabela 12: Implementação de cálculo de similaridade através dos vetores TF-IDF

Código

```
require(text2vec)

sim2(t(frase1_tfidf_vector), t(frase2_tfidf_vector),
     method = "cosine", norm = "none")
```

Fonte: Elaborado pelos autores.

APÊNDICE D AMOSTRAS DE EXPERIMENTOS REALIZADOS

Neste capítulo são apresentadas amostras de resultados obtidos, os quais foram escolhidos empiricamente a partir de uma planilha de experimentos elaborada pelos autores. Considerando o total de 49726 experimentos executados, as estatísticas de resultados obtidos podem ser observadas na Tabela 13. Para a melhor observação dos resultados alcançados, foram selecionados

Tabela 13: Estatísticas dos experimentos realizados

	SVM		RNA		GLM	
	μ	σ	μ	σ	μ	σ
Correlação de Pearson (CP)	0.5948	0.1023	0.5915	0.1011	0.5863	0.1210
Erro Quadrático Médio (EQM)	0.5184	0.1128	0.4972	0.0880	0.5183	0.1145

μ corresponde a média.
σ corresponde ao desvio padrão.
Fonte: Elaborado pelos autores.

quatro casos considerados melhores, médios e piores (dentro do intervalo de desvio padrão para CP e EQM), os quais podem ser vistos na Tabela 14.

Tabela 14: Amostra de resultados de experimentos

	Atributos *	Correlação de Pearson	Erro Quadrático Médio	Normalização**
Melhores	1,2,4,5,6,8,10	0.6626	0.4304	MaxMin
	1,2,4,6,7,10	0.6625	0.4303	MaxMin
	1,2,5,6,8,10	0.6625	0.4302	MaxMin
	1,3,4,6,8,10	0.6410	0.4479	Não
Médios	1,3,4,6,8,10	0.6393	0.4504	MaxMin
	1,4,6,8,10	0.6392	0.4505	MaxMin
	1,3,4,6,10	0.6385	0.4504	Não
	1,3,6,8,10	0.6385	0.4510	MaxMin
Piores	3,9	-0.0361	0.7604	MaxMin
	3	0.0087	0.7665	Não
	7,9	0.2896	0.8569	Não
	9	0.3023	0.9254	MaxMin

* A segunda coluna representa o índice dos atributos descritos na Tabela 5.
** 'MaxMin' refere-se a aplicação de normalização por máximo e mínimo nos atributos.
Fonte: Elaborado pelos autores.