UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

FILIPE GOMES DE LIMA

ANÁLISE DE LÉXICO ARGUMENTATIVO COMO RECURSO PARA MINERAÇÃO TEXTUAL

CAMPINA GRANDE - PB

2021

FILIPE GOMES DE LIMA

ANÁLISE DE LÉXICO ARGUMENTATIVO COMO RECURSO PARA MINERAÇÃO TEXTUAL

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Professor Dr. Cláudio Campelo.

CAMPINA GRANDE - PB

2021

L732a Lima, Filipe Gomes de.
Análise de léxico argumentativo como recurso para mineração textual. / Filipe Gomes de Lima. – 2021.

11 f.

Orientador: Prof. Dr. Cláudio Elízio Calazans Campelo.
Trabalho de Conclusão de Curso - Artigo (Curso de Bacharelado em Ciência da Computação) - Universidade Federal de Campina Grande; Centro de Engenharia Elétrica e Informática.

1. Mineração de argumentos. 2. Léxicos argumentativos. 3. Distância semântica. 4. Word Mover's Distance. 5. Textos – identificação de argumentos. 6. Mineração textual. I. Campelo, Cláudio Elízio Calazans. II. Título.

CDU:004.912(045)

Elaboração da Ficha Catalográfica:

Johnny Rodrigues Barbosa
Bibliotecário-Documentalista
CRB-15/626

FILIPE GOMES DE LIMA

ANÁLISE DE LÉXICO ARGUMENTATIVO COMO RECURSO PARA MINERAÇÃO TEXTUAL

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Professor Dr. Cláudio Campelo

Orientador – UASC/CEEI/UFCG

Professora Dr. Francisco Brasileiro

Examinador – UASC/CEEI/UFCG

Professor Tiago Lima Massoni

Professor da Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em 25 de maio de 2021.

CAMPINA GRANDE - PB

RESUMO

Mineração de Argumentos é uma área de pesquisa que visa identificar estruturas argumentativas em textos a partir da identificação dos componentes argumentativos e das relações entre eles. A identificação automática das estruturas pode ser considerada um problema pelo fato de não existirem padrões ou regras para esta etapa. Além disso, a escassez de trabalhos direcionados ao idioma português brasileiro dificulta tal atividade. Dessa forma, este trabalho propõe discutir e analisar a influência de léxicos argumentativos na etapa de identificação de argumentos em sentenças argumentativas. Para isto, foi utilizada a técnica de distância semântica WMD entre textos com a finalidade de analisar a presença de léxicos argumentativos em sentenças que contém argumentatividade ou não. Este trabalho visa contribuir como uma análise exploratória entre diferentes tipos de léxicos argumentativos.

Análise de Léxico Argumentativo como Recurso para Mineração Textual

Trabalho de Conclusão de Curso

Filipe Gomes de Lima (Aluno), Cláudio Campelo (Orientador)
Departamento de Sistemas e Computação
Universidade Federal de Campina Grande
Campina Grande, Paraíba - Brasil

RESUMO

Mineração de Argumentos é uma área de pesquisa que visa identificar estruturas argumentativas em textos a partir da identificação dos componentes argumentativos e das relações entre eles. A identificação automática das estruturas pode ser considerada um problema pelo fato de não existirem padrões ou regras para esta etapa. Além disso, a escassez de trabalhos direcionados ao idioma português brasileiro dificulta tal atividade. Dessa forma, este trabalho propõe discutir e analisar a influência de léxicos argumentativos na etapa de identificação de argumentos em sentenças argumentativas. Para isto, foi utilizada a técnica de distância semântica WMD entre textos com a finalidade de analisar a presença de léxicos argumentativos em sentenças que contém argumentatividade ou não. Este trabalho visa contribuir como uma análise exploratória entre diferentes tipos de léxicos argumentativos.

PALAVRAS-CHAVE

Mineração de Argumentos, Léxicos Argumentativos, Distância Semântica, WMD

1 INTRODUÇÃO

A argumentação é o desenvolvimento de um raciocínio com o fim de defender ou repudiar uma tese ou ponto de vista, para convencer alguém ou a nós próprios [1]. Em um texto, uma estrutura argumentativa é formada por componentes argumentativos (i.e. afirmação e premissas) e diferentes tipos de relacionamentos que podem ser utilizados na organização desses componentes [7].

A Mineração de Argumentos (do inglês, Argument Mining) é uma área que tem como objetivo a identificação automática das estruturas argumentativas presente nos textos. Para isto, são utilizadas técnicas de Inteligência Artificial para identificar componentes argumentativos e suas relações, sendo essas as duas atividades do processo de identificação das estruturas argumentativas.

Na literatura são apresentadas diferentes abordagens para a identificação dos componentes argumentativos [5, 7, 21]. Um mecanismo que pode auxiliar na identificação das estruturas argumentativas é a verificação da ocorrência de indicadores argumentativos, ou seja, termos que podem indicar argumentatividade em fragmentos de texto [2] (i.e. “portanto”, “assim”, “daí”, “porque”). Esses termos podem ter mais de um significado, dependendo do seu contexto, portanto, não é possível afirmar que o texto analisado é argumentativo, apenas por conter um indicador argumentativo.

Diante disso, este trabalho propõe uma análise comparativa entre diferentes conjuntos de indicadores argumentativos (léxicos argumentativos), com o objetivo de verificar se esses léxicos são semanticamente similares às sentenças argumentativas, tornando-se potenciais recursos para a tarefa de Mineração de Argumento. Para este fim, utilizamos a técnica chamada Word Mover’s Distance (WMD) [6] que nos permite medir a distância semântica entre um léxico argumentativo e uma sentença, seja previamente anotada como argumentativa ou não. Nós utilizamos 3 léxicos (conjunto de termos) de trabalhos distintos, criamos 3 léxicos a partir de plataformas relacionadas à educação, e criamos 1 léxico que contém a junção de todos os léxicos argumentativos coletados.

Foram conduzidos experimentos considerando duas bases de dados: a primeira, uma base de redações de gênero dissertativo-argumentativo [3], com sentenças anotadas com Tese, Proposta de Intervenção, Argumento e Não Argumento. A segunda base vem do *UKP Sentential Argument Mining Corpus* [19] traduzido do inglês para português de forma automática por Rodrigues & Branco [14], com sentenças anotadas como Argumento e Não Argumento.

A partir dos experimentos realizados, observamos que todos os léxicos citados anteriormente conseguem distinguir sentenças argumentativas e não argumentativas. Também notamos que utilizando a base de redações, a qual foi construída em português brasileiro, os léxicos não traduzidos parecem ter um potencial maior para distinção de argumento e não argumento, ou seja, para serem usados como recursos para Mineração de Argumentos.

Esse documento está organizado como se segue: na Seção 2, discutimos os trabalhos relacionados. Na Seção 3 é apresentada a metodologia utilizada no trabalho. A Seção 4 apresenta os resultados dos experimentos realizados para validar a abordagem. Por fim, na Seção 5, é descrita as considerações finais do estudo realizado.

2 TRABALHOS RELACIONADOS

Conforme descrito anteriormente, o processo realizado para a Mineração de Argumentos, em nível macro, é composto pelas atividades de identificação dos componentes argumentativos e das relações entre eles. Nesses processos, técnicas de Processamento de Linguagem Natural (PLN) e Aprendizagem de Máquina (AM) podem ser aplicadas, porém, a literatura mostra que não há consenso no que diz respeito à identificação dos componentes argumentativos, tendo em vista a diversidade de estudos e abordagens [9]. O uso

dos conjuntos de termos (léxicos) argumentativos é uma técnica que pode auxiliar esse processo de identificação [15].

Em Stab & Gurevych [16] foi descrita uma metodologia para identificação de trechos argumentativos em redações escritas no idioma inglês. Para isso, foram utilizados dois léxicos argumentativos para identificar os trechos argumentativos dos textos. Ressalta-se que os próprios autores afirmam que podem existir trechos que não contém termos do léxico, mas que podem ser argumentativos. Desse modo, é considerado o casamento de padrão exato dos termos, descartando suas possíveis variações e permitindo que alguns trechos não tenham sido identificados. Como resultado, os autores criaram uma base de dados para a Mineração de Argumentos[1] que foi utilizada em pesquisas posteriores [13, 17].

Em Stab & Gurevych [18] os autores propuseram um conjunto de termos em inglês que podem ser utilizados na argumentação. Os termos foram agrupados nas classes *(i) forward, (ii) backward, (iii) thesis* e *(iv) rebuttal*, e foram utilizados como uma das características de um classificador. Por sua vez, esse trabalho serviu como base para o trabalho de Nau [12], que investigou a Mineração de Argumentos no idioma português brasileiro por meio da adaptação do classificador criado, inclusive com a tradução dos termos de forma automática, resultando nas classes (i) avançados, (ii) regressivos, (iii) de tese e (iv) de refutação.

Os trabalhos apresentados são exemplos de como os léxicos, de modo geral, podem auxiliar o processo de identificação de trechos argumentativos. Porém, não há uma regra única de como a argumentação deve ser expressa. Dessa forma, a comparação exata em busca dos termos nos textos poderá apresentar limitações, sendo uma possível solução a análise da similaridade semântica entre léxico e texto. Como exemplo desta solução, em outro contexto de aplicação, Lima et al. [8] utilizou um conjunto de termos referentes à subjetividade para serem analisados em textos de notícias e comentários, coletados de diferentes portais de esporte e política. Aplicou-se a técnica de WMD que calcula a distância semântica entre dois documentos, sendo possível verificar, não se os termos estão contidos nos textos ou não, mas sim a distância entre os termos e o texto analisando.

Corroborando com essa ideia, Jeronimo et al. [4] utilizou a técnica baseada no WMD para mensurar a subjetividade presente em notícias reais e falsas. Para isso, foi utilizado um léxico de subjetividade (com as dimensões de pressuposição, argumentação, modalização, sentimento e valoração) e notícias reais e falsas escritas no idioma português brasileiro. Como resultado, foram apresentadas as distâncias de subjetividade calculada nos textos e, além disso, o valor foi utilizado como característica (*feature*) de um classificador de notícias falsas.

De modo geral, verifica-se que o uso de léxicos podem ser úteis no processo de identificação ou classificação de textos em diferentes contextos de aplicação, desde que exista um conjunto de termos do contexto em questão. Além disso, nota-se que a comparação exata pode ser um fator limitante das abordagens propostas, tendo em vista a diversidade de escrita. Logo, o uso de abordagens para mensurar a similaridade semântica entre documentos, considerando o léxico como um desses, pode ser uma decisão mais acertada. Diante disso, a metodologia proposta para esse trabalho utilizará os conceitos expostos para alcançar os objetivos propostos nessa pesquisa.

[1]Disponível em: https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2421

3 METODOLOGIA

Nessa seção apresenta-se a metodologia adotada para conduzir os experimentos supracitados. Serão descritas as características das bases de dados e léxicos argumentativos utilizados, o pré-processamento do *Word Embedding* com base em artigos de enciclopédia e o cálculo da similaridade usando o WMD.

3.1 Base de Dados

Nos trabalhos de Nau [12] e Rodrigues & Branco [14] foram utilizados corpus de textos escritos em português para a aplicação de técnicas de Mineração de Argumentos. Logo, visto que as bases possuem rótulos para identificar os trechos argumentativos, ambas foram escolhidas para serem utilizadas na análise proposta.

3.1.1 Redações dissertativa-argumentivas. No trabalho de Nau [12] utilizou-se uma base de redações de gênero dissertativo-argumentativo, extraídas do Brasil Escola[2], originalmente criadas como treinamento para o Exame Nacional do Ensino Médio (ENEM)[3]. A base utilizada no estudo é composta por 50 redações, que tiveram suas sentenças extraídas e classificadas manualmente por especialistas em três classes: tese, proposta de intervenção, e argumento. Este procedimento resultou em uma base de dados composta por 458 sentenças, sendo 77 classificadas como tese, 116 como proposta de intervenção e 265 como argumento.

Ao analisar a base de dados e seus respectivos textos (redações que tiveram os trechos extraídos e anotados), verificou-se que alguns trechos das redações não foram classificados em nenhuma das categorias, caracterizando-se como sendo trechos não argumentativos. A identificação desses trechos não argumentativos é de suma importância para o escopo desse trabalho, sendo assim, foi implementado um algoritmo para classificá-los. Este algoritmo, assim como a abordagem utilizada por Nau [12], segmenta as redações por sentenças e, em seguida, verifica se a sentença está contida na base anotada. Caso negativo, a sentença é anotada como "Não Argumento". A Tabela 1 mostra a quantidade de sentenças rotuladas em cada categoria.

Tese	PI	Argumento	Não Argumento
77	116	265	207

Tabela 1: Quantidade de sentenças em cada categoria

Para a execução dos experimentos e análises, além das quatro categorias descritas na Tabela 1, também foi considerada uma proposta binária representada por todos os trechos argumentativo e não argumentativos. Destaca-se que esta base tem como proposta o gênero dissertativo-argumentativo para o ENEM, no qual o autor precisa apresentar um ponto de vista sobre um determinado tema, expor argumentos para defender ou refutar a tese, além de uma proposta de intervenção, ou seja, respectivamente os trechos classificados como Tese, Proposta de Intervenção (PI) e Argumentos.

[2]Brasil Escola - https://brasilescola.uol.com.br
[3]ENEM - https://enem.inep.gov.br/

Com base nisto, foi criada uma adaptação da base original, na qual as três categorias foram unificadas como Argumento, resultando em 458 registros, e foram preservados os 207 registros da categoria Não Argumento.

3.1.2 UKP Sentential Argument Mining Corpus. – Devido à escassez de bases de dados deste tópico em português, o trabalho de Rodrigues & Branco [14] realiza uma tradução automática do *UKP Sentential Argument Mining Corpus* [19] do inglês para o português. Tal base é composta por 23097 sentenças, sendo 12994 classificadas como Não Argumento e 10103 como Argumento. É importante destacar que os trechos foram coletados a partir de notícias, editoriais, blogs, fóruns de debate e artigos de enciclopédia.

O uso desse corpus, apesar de apresentar ameaças à validade devido à tradução automática, apresenta-se como sendo uma alternativa para a verificação da abordagem proposta nesse trabalho, uma vez que seus autores obtiveram sucesso ao utilizarem como base de treinamento para um classificador de sentenças argumentativas.

3.2 Léxicos Argumentativos

Os léxicos utilizados na abordagem foram criados a partir de termos presentes na literatura e coletados na internet, totalizando 6 fontes.

Nau [12] utiliza um conjunto de termos para representar a argumentação no contexto das redações, porém, tais termos foram traduzidos do inglês para o português. Tomando como base essa proposta de tradução e os termos traduzidos, também foi feita a tradução automática dos termos apresentados em Stab et al. [19], totalizando duas bases traduzidas. Além desses, buscou-se inserir na base de léxicos termos escritos para o contexto brasileiro. Dessa forma, utilizamos o léxico de "Argumento" apresentado em Jeronimo et al. [4], e coletados termos referentes à argumentação disponível no Brasil Escola[4], Acrobata das Letras[5] e Mundo Educação[6], conforme detalhado na Tabela 2. É importante salientar que, os léxicos coletados na literatura atingiram resultados positivos em seus respectivos trabalhos de origem.

Para a criação do léxico "Todos", os conjunto de termos das 6 fontes foram unificados e pré-processados com a finalidade de ajustar variantes e escrita, e remover duplicatas, *stopwords* e termos presentes em diversos contextos além da argumentação, ou seja, termos considerados *gerais* e que podem apresentar ambiguidade quando usados isoladamente.

Foram realizados ajustes para adicionar dois termos: (i) "na minha opinião" – variante do termo "em minha opinião" e (ii) "no que diz respeito" – variante generalista do termo "no que me diz respeito" já existente. Também foi removido o pronome pessoal "eu" de alguns termos, por exemplo, "eu concordo que" tornou-se "concordo que".

Para a remoção dos termos gerais, selecionamos todos os unigramas (termos que contém apenas uma palavra) e aplicamos cada um deste na função *most_similar*[7]. Esta função calcula a similaridade do cosseno entre o peso do vetor da palavra fornecida e os vetores de cada palavra no modelo, retornando um ranking das palavras mais similares, partindo da premissa de que, ao passar um termo de léxico argumentativo, a função retorne palavras semanticamente similares (i.e., outros termos de léxicos argumentativos). Consideramos um termo geral (retirado do léxico) aquele que retorna as 5 primeiras palavras não pertencentes à lista unificada de léxicos.

A quantidade de palavras igual a 5 foi definida a partir de experimentos realizados com diferentes quantidades. Nos experimentos variando entre 3 até 12 palavras, para uma quantidade de palavras alta, foram capturados termos como "acho", "primeiro" e "segundo", que julgamos como gerais. Para uma quantidade de palavras baixa, foram capturados termos como "similarmente" e "posteriormente", que julgamos ser mais específicos na elaboração de um argumento. Então foi considerado um valor médio de 5 palavras. Além disso, também foram removidos termos que não existem no corpus do *Word Embedding*, ou seja, palavras sem representações vetoriais.

Como vimos, os termos dos léxicos podem ser compostos por mais de uma palavra, o que torna-se um problema ao usar algoritmos que utilizam um *Word Embedding*. Isso acontece pois o *Word Embedding* é constituído por palavras únicas em que cada uma delas é representada por um vetor. Para o uso do WMD é necessário que cada palavra seja representado por um vetor de termos simples e, por isso, foram colocados "_" (*underscores*) para substituir os " " (espaços em branco), para podermos transformar termos compostos (e.g., "no entanto") em *tokens* únicos (e.g., "no_entanto").

O mesmo algoritmo de pré-processamento foi executado para cada uma das fontes de léxico individualmente. Vale salientar que, para o pré-processamento individual, não houve adição de termos.

3.3 *Word Embedding*

O uso da abordagem baseada no WMD é dependente de um arquivo *Word Embedding* pré-treinado para que ocorra os cálculos das distâncias entre documentos. Para o treinamento do arquivo, com o corpus de 01 de março de 2021 dos artigos em português da Wikipédia[8], foi utilizado o algoritmo Word2Vec [10, 11].

A particularidade deste *Word Embedding* é que o seu corpus precisou ser modificado de modo que cada termo composto (i.e., termo com mais de uma palavra), presente nos artigos, se tornasse um *token* para posteriormente ser representado por um único vetor. Por exemplo, todas as ocorrências do termo "de certa forma", no corpus da Wikipédia, são modificadas para "de_certa_forma".

Também fizeram parte da etapa de pré-processamento dos artigos da Wikipédia: (i) remoção de pontuações; (ii) e-mails e URLs transformados em tokens; (iii) números transformados em 0 (zero) e (iv) conversão do texto para minúsculo. Após esta etapa, o treinamento ocorre de forma não-supervisionada, gerando um modelo no qual será a base dos cálculos do WMD.

3.4 Análise da Similaridade com WMD

O WMD *(Word Mover's Distance*) [6] é uma técnica de comparação semântica, já consolidada no campo de PLN que, dado um *Word Embedding* pré-treinado, calcula a similaridade entre dois documentos de texto como a distância mínima em que os vetores das palavras de um documento precisam se "mover" para alcançar os vetores das palavras do outro documento.

[4]Brasil Escola: https://brasilescola.uol.com.br/redacao/operadores-argumentativos.htm
[5]Acrobata das Letras: https://www.acrobatadasletras.com.br/2017/06/operadores-argumentativos-para-redacao.html
[6]Mundo Educação: https://mundoeducacao.uol.com.br/redacao/operadores-argumentativos.htm
[7]Disponível em: https://radimrehurek.com/gensim/models/keyedvectors.html
[8]https://www.wikipedia.org/

Fonte	Descrição	Quantidade
(Nau, 2020)	Léxicos classificados por refutação, tese, avançado e regressivo.	54
(Stab & Gurevych, 2017)	Léxicos classificados por afirmação e premissa.	42
(Jeronimo et al., 2019)	Léxico de argumentação.	90
Acrobata das Letras	Léxico criado a partir do site do Acrobata das letras	71
Brasil Escola	Léxico criado a partir do site do Brasil Escola	46
Mundo Educação	Léxico criado a partir do site do Mundo Educação	57
Todos	Resultado da concatenação de todos os léxicos acima.	295

Tabela 2: Origem dos léxicos

Supondo que tomamos dois textos separadamente e chamamos de A e B, a distância entre os dois textos é calculada através da distância em que os vetores do texto A precisam percorrer para corresponder exatamente com os vetores do texto B. O algoritmo mede a diferença entre dois textos com o propósito de encontrar o caminho mais curto entre suas palavras.

No contexto desse trabalho, essa técnica foi utilizada para comparação dos documentos contendo léxicos argumentativos e textos argumentativos. Espera-se que, ao comparar os dois documentos, a similaridade (distância) semântica usando o WMD para os trechos argumentativos seja mais próximo a 0 (zero), diferentemente dos trechos não argumentativos, pois acredita-se que os termos do léxico estejam mais presentes em trechos que contém argumentatividade do que em trechos que não contém.

Após o cálculo da similaridade semântica, para cada léxico de argumentação utilizado, teremos distribuições de distância WMD. As quantidades de elementos contidos nessas distribuições correspondem ao número de sentenças de cada categoria de seus respectivos corpus. Para cada distribuição, é verificada a sua normalidade e, além disso, o Teste T [20] é aplicado para verificar, estatisticamente, se os valores médios das duas distribuições são diferentes. Quando o p-valor retornado pelo Teste T é abaixo do limiar 0.05, a hipótese nula de que os valores médios das distribuições são iguais é rejeitada, ou seja, as distâncias médias dos trechos que contém argumentatividade são, de fato, diferentes das distâncias dos trechos que não contém argumentatividade.

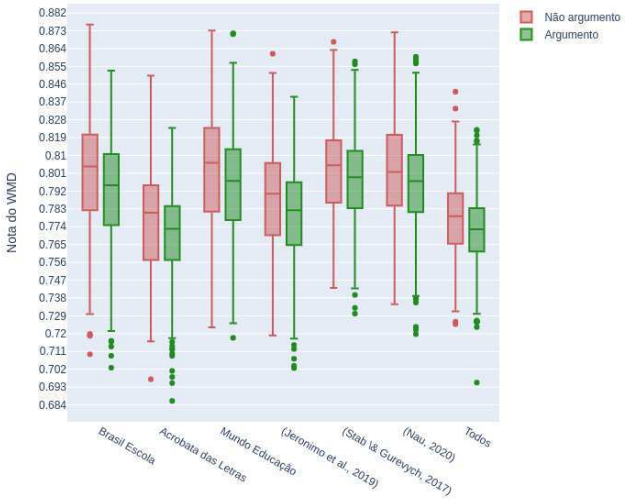

Figura 1: Distâncias do WMD com a base de redações rotuladas com argumento (contendo PI, Tese e Argumento) e não argumento

4 RESULTADOS

Nesta seção são apresentados os resultados dos experimentos realizados. No intuito de estendermos os cenários da análise comparativa, utilizamos duas bases de dados diferentes.

4.1 Experimento 1

Nesse experimento avaliamos os valores das distâncias semânticas dos diferentes léxicos com a base de textos dissertativo-argumentativos. Inicialmente, será discutido as distribuições de distâncias referentes às sentenças marcadas como "Argumento" e "Não Argumento".

Na Figura 1 o eixo X é representado pelos 7 léxicos utilizados no estudo. Para cada um, são apresentadas duas distribuições de distâncias nas cores vermelha e verde que, respectivamente, representam as distribuições de distâncias das sentenças rotuladas como "Não Argumento", e as distâncias das sentenças rotuladas como "Argumento". Nota-se que, para todos os léxicos, a mediana das sentenças de "Argumento" é menor que a mediana das sentenças de "Não Argumento". Esse resultado era esperado, pois acreditávamos que os trechos que contém argumentatividade, aparentemente, são semanticamente mais próximos do léxico argumentativo do que os trechos que não contém.

A Figura 2 mostra, para cada léxico, a diferença absoluta entre as medianas das distribuições de "Argumento" e "Não Argumento". Para este experimento, os 4 léxicos não traduzidos, ou seja, que foram construídos originalmente na língua portuguesa, obtiveram as maiores diferenças de mediana entre as distribuições. Um possível motivo desse resultado, quando comparado aos léxicos traduzidos, é que os textos das redações também foram construídos originalmente na língua portuguesa.

Ao aplicar o Teste T vemos que as distribuições de "Argumento" e "Não Argumento" retornam um *p-valor* menor que 0.05, ou seja, estas distribuições são estatisticamente diferentes. O teste foi aplicado para cada léxico e o resultado mostra que, independente do léxico ser traduzido ou não, as distribuições de "Argumento" e "Não Argumento" tem valores médios diferenciáveis, conforme a Tabela 3.

Também foi realizado um experimento considerando as classes Tese, PI, Argumento e Não Argumento, conforme a Figura 3. Como

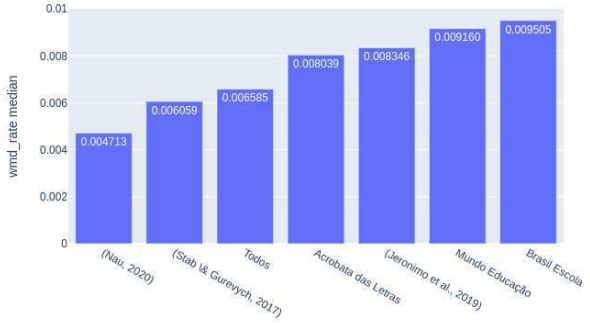

Figura 2: Diferença entre as medianas das distribuições de Argumento e Não Argumento na base de redações

Léxico	*p-valor*
Brasil Escola	0.00042318
Acrobata das Letras	0.00080507
Mundo Educação	0.00056519
(Jeronimo et al., 2019)	0.00023435
(Stab & Gurevych, 2017)	0.00705241
(Nau, 2020)	0.00313812
Todos	2.97681e-05

Tabela 3: Resultado do *p-valor* do Teste T para as classes "Argumento" e "Não argumento"

resultado, nota-se que as distribuições de "Não Argumento", em vermelho, apresentaram medianas maiores que as outras distribuições, exceto quando observado o léxico de Stab & Gurevych [14]. Ao analisar os valores apresentados na Tabela 4, verifica-se que os *p-valores* referente ao par "Não Argumento/PI" do léxico de Stab & Gurevych [14], o valor está acima de 0.05, não rejeitando a hipótese de que os valores médios destas duas distribuições são iguais, ou seja, mesmo com a mediana da "Proposta de Intervenção" (PI) acima da mediana de "Não Argumento", ambas as distribuições podem ter os valores médios iguais.

Além disso, apenas para as distribuições de "Argumento" (não unificado) e "Não Argumento", todos os léxicos obtiveram um *p-valor* abaixo de 0.05, assim como esperado. Por outro lado, em todos os léxicos o par de distribuições de "Argumento/Tese" obteve *p-valor* acima de 0.05, não rejeitando a hipótese de que as duas distribuições têm valores médios diferentes. Como dito anteriormente, Tese, Argumento e PI são partes de uma estrutura argumentativa, por isso, os pares de distribuições destes rótulos podem apresentar valores médios iguais/próximos.

4.2 Experimento 2

Para validar a abordagem em outro cenário, foi conduzido o mesmo experimento com as distribuições de Argumento e Não Argumento presente na base de dados traduzida do *UKP Sentential Argument Mining*.

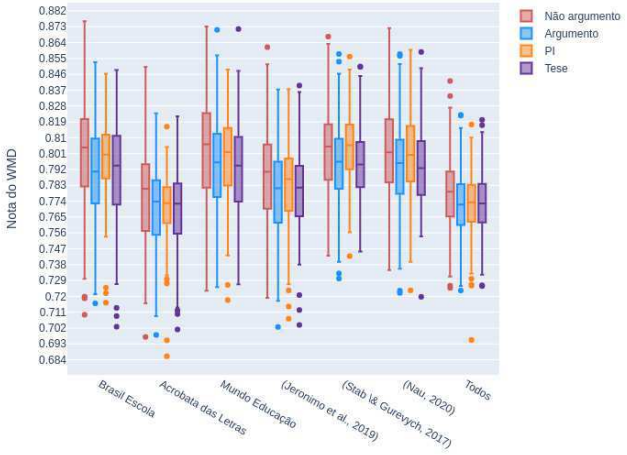

Figura 3: Distâncias do WMD com a base de redações rotuladas com Argumento, PI, Tese e Não Argumento

A Figura 4 mostra as distâncias do WMD aplicadas na base em questão. Assim como nas redações (Figura 1), todas as medianas das distribuições de "Argumento" apresentaram-se abaixo das medianas de "Não Argumento" indicando que, para todos os léxicos utilizados, as sentenças que contém argumentatividade estão semanticamente mais próximas do léxico argumentativo do que as sentenças que não contém.

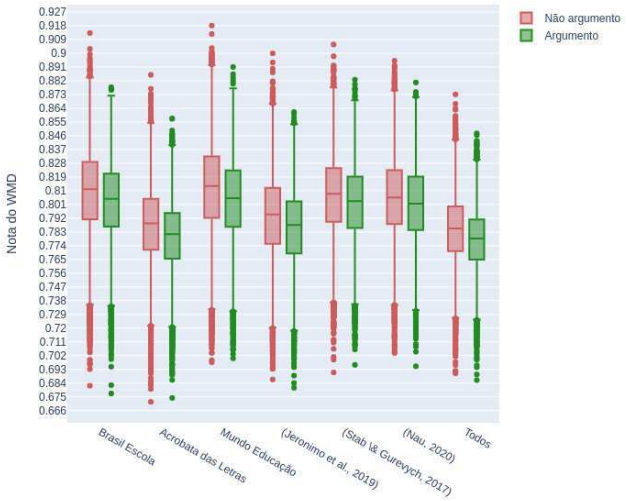

Figura 4: Distâncias do WMD com a base traduzida do UKP Sentential Argument Mining rotuladas com Argumento e Não Argumento

Léxico	p-valor					
	não_arg / arg	**não_arg / pi**	**não_arg / tese**	**arg / pi**	**arg / tese**	**pi / tese**
Brasil Escola	9.06056e-05	0.469068	0.0124389	0.00523007	0.985073	0.0459811
Acrobata das Letras	0.00432028	0.0191694	0.0275386	0.959885	0.680573	0.73358
Mundo Educação	0.000344639	0.146822	0.0537193	0.102689	0.607727	0.437537
(Jeronimo et al., 2019)	0.000123436	0.081495	0.0355272	0.159935	0.648346	0.539791
(Stab & Gurevych, 2017)	0.000226491	0.416805	0.0373518	4.19129e-05	0.627611	0.00590629
(Nau, 2020)	0.000380144	0.775973	0.0431679	0.00576202	0.668927	0.088279
Todos	5.66305e-05	0.00546516	0.0484811	0.672671	0.426827	0.704006

Tabela 4: P-valor do Teste T de todas as combinações de categorias para cada léxico

Léxico	*p-valor*
Brasil Escola	1.30918e-71
Acrobata das Letras	7.12453e-119
Mundo Educação	1.10611e-87
(Jeronimo et al., 2019)	6.39107e-109
(Stab & Gurevych, 2017)	3.31811e-44
(Nau, 2020)	1.21113e-32
Todos	3.50206e-126

Tabela 5: *p-valor* do Teste T para Argumentos e Não argumentos

Na Figura 5, diferente da base das redações, os léxicos traduzidos não foram excepcionalmente os que tiveram as maiores distâncias entre medianas. A característica do texto traduzido, além das fontes coletadas para que a base fosse construída, podem ter influenciado neste resultado.

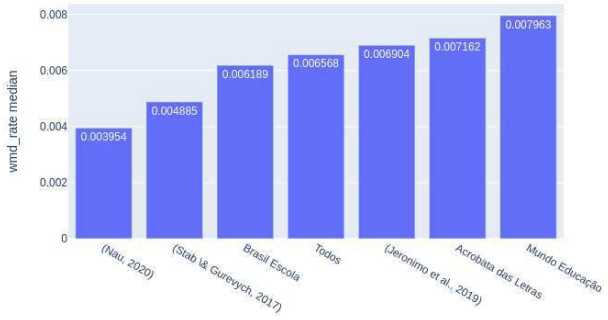

Figura 5: Distâncias do WMD do *dataset UKP Sentential Argument Mining* rotuladas com Argumento e Não Argumento

Na aplicação do Teste T, assim como na base de redações, todos os léxicos obtiveram um *p-valor* abaixo de 0.05 quando comparou-se as distribuições entre Argumento e Não Argumento, conforme a Tabela 5. Isso mostra que as distribuições de "Argumento" e "Não Argumento" são estatisticamente diferentes.

5 CONCLUSÃO

Este trabalho apresentou um estudo com a finalidade de analisar se léxicos argumentativos são potencialmente úteis na tarefa de Mineração de Argumentos no idioma português. Para isso, considerou-se a técnica de WMD para obter a distância semântica entre léxicos argumentativos e sentenças anotadas como Argumento ou Não Argumento. Partimos da premissa de que obteríamos distâncias menores para sentenças argumentativas, indicando que os termos argumentativos e suas variações de escrita estariam mais presentes em textos com argumentatividade.

Com a base de sentenças de redações identificamos que para todos os pares de distribuições estatisticamente diferentes (p-valores menores que 5%), a mediana da distribuição de Não Argumento foi sempre mais alta, ou seja, sempre mais distante do léxico argumentativo. Vimos que para esta base, os léxicos criados com a língua portuguesa obtiveram as maiores distâncias entre medianas, concluindo que para os léxicos argumentativos em português, as sentenças de argumento e não argumento são mais diferenciáveis. Além disso, observamos que as distâncias de Tese e Argumento são bastante próximas (estatisticamente idênticas), o que nos mostra que sentenças com Tese também contém palavras semanticamente próximas dos léxicos argumentativos.

Além desse, foi conduzido outro experimento utilizando os mesmos léxicos e uma base de dados traduzida do inglês para português, e o resultado obtido foi o mesmo da base de redações, com exceção do ranking de medianas, em que desta vez, os léxicos não traduzidos ficaram distribuídos no ranking.

De modo geral, no contexto dos experimentos realizados, foi possível identificar que sentenças argumentativas são semanticamente mais próximas dos léxicos argumentativos, porém, ressalta-se uma possível ameaça à validade devido à tradução automática da base utilizada em um dos experimentos e dos termos argumentativos.

Em trabalhos futuros pretende-se aplicar o resultado das distribuições de distâncias do WMD no âmbito das redações dissertativa-argumentativas, e usar as distâncias como uma *feature* para um classificador de sentenças argumentativas, podendo ser útil em diferentes aplicações que utilizem a técnica de Mineração de Argumentos.

REFERÊNCIAS

[1] Carlos Ceia. [n. d.]. Argumentação. Retrieved Oct 27, 2020 from https://edtl.fcsh.unl.pt/encyclopedia/argumentacao/
[2] Gary Curtis. [n. d.]. Arguments. Retrieved Oct 20, 2020 from https://gohighbrow.com/arguments/

[3] Matthew Van Gundy, Davide Balzarotti, and Giovanni Vigna. 2007. Catch me, if you can: Evading network signatures with web-based polymorphic worms. In *Proceedings of the first USENIX workshop on Offensive Technologies (WOOT '07)*. USENIX Association, Berkley, CA, Article 7, 9 pages.

[4] Caio Libanio Melo Jeronimo, Leandro Balby Marinho, Claudio E. C. Campelo, Adriano Veloso, and Allan Sales da Costa Melo. 2019. Fake News Classification Based on Subjective Language. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3366030.3366039

[5] Johannes Kiesel, Khalid Al-Khatib, Matthias Hagen, and Benno Stein. 2015. A Shared Task on Argumentation Mining in Newspaper Editorials. In *Proceedings of the 2nd Workshop on Argumentation Mining*. Association for Computational Linguistics, Denver, CO, 35–38. https://doi.org/10.3115/v1/W15-0505

[6] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*. 957–966.

[7] John Lawrence and Chris Reed. 2020. Argument Mining: A Survey. *Computational Linguistics* 45, 4 (01 2020), 765–818. https://doi.org/10.1162/coli_a_00364 arXiv:https://direct.mit.edu/coli/article-pdf/45/4/765/1847520/coli_a_00364.pdf

[8] D. F. Lima, A. S. C. Melo, and L. B. Marinho. 2019. An Analysis of Subjectivity in Brazilian News. In *Anais do VII Symposium on Knowledge Discovery, Mining and Learning*. SBC, Porto Alegre, RS, Brasil, 81–88. https://doi.org/10.5753/kdmile.2019.8792

[9] Marco Lippi and Paolo Torroni. 2016. Argumentation Mining: State of the Art and Emerging Trends. 16, 2 (2016). https://doi.org/10.1145/2850417

[10] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. arXiv:cs.CL/1301.3781

[11] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. arXiv:cs.CL/1310.4546

[12] Jonathan Nau. 2020. *Processamento de Discurso em Textos Dissertativos-argumentativos.* Master's thesis. Universidade do Vale do Itajaí.

[13] Huy Nguyen and D. Litman. 2016. Improving Argument Mining in Student Essays by Learning and Exploiting Argument Indicators versus Essay Topics. In *FLAIRS Conference*.

[14] João Rodrigues and António Branco. 2020. Argument Identification in a Language Without Labeled Data. In *Computational Processing of the Portuguese Language*, Paulo Quaresma, Renata Vieira, Sandra Aluísio, Helena Moniz, Fernando Batista, and Teresa Gonçalves (Eds.). Springer International Publishing, Cham, 335–345.

[15] Patrick Saint Dizier. 2020. The Lexicon of Argumentation for Argument Mining: methodological considerations. *Anglophonia. French Journal of English Linguistics* 29 (2020).

[16] Christian Stab and Iryna Gurevych. 2014. Annotating Argument Components and Relations in Persuasive Essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, Dublin, Ireland, 1501–1510. https://www.aclweb.org/anthology/C14-1142

[17] Christian Stab and Iryna Gurevych. 2014. Identifying Argumentative Discourse Structures in Persuasive Essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 46–56. https://doi.org/10.3115/v1/D14-1006

[18] Christian Stab and Iryna Gurevych. 2017. Parsing Argumentation Structures in Persuasive Essays. *Computational Linguistics* 43, 3 (09 2017), 619–659.

[19] Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic Argument Mining from Heterogeneous Sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 3664–3674. https://doi.org/10.18653/v1/D18-1402

[20] B. L. WELCH. 1947. THE GENERALIZATION OF 'STUDENT'S' PROBLEM WHEN SEVERAL DIFFERENT POPULATION VARLANCES ARE INVOLVED. *Biometrika* 34, 1-2 (01 1947), 28–35. https://doi.org/10.1093/biomet/34.1-2.28 arXiv:https://academic.oup.com/biomet/article-pdf/34/1-2/28/553093/34-1-2-28.pdf

[21] Adam Wyner, Raquel Mochales, Marie-Francine Moens, and David Milward. 2010. Approaches to Text Mining Arguments from Legal Cases. 60–79. https://doi.org/10.1007/978-3-642-12837-0_4