SIDEAS - Detectando a Similaridade Semântica de Discursos

Rita C. A. B. Costa[1], **Osmar O. Braz Júnior**[1,2], **Renato Fileto**[1]

[1] Universidade Federal de Santa Catarina (UFSC) | PPGCC / INE
Centro Tecnológico (CTC), Campus Trindade, Florianópolis-SC, Brasil

[2]Departamento de Educação Científica e Tecnológica
Universidade do Estado de Santa Catarina (UDESC)
Av. Madre Benvenuta, 2007, Itacorubi, Florianópolis-SC, Brasil

rita.alamino@posgrad.ufsc.br, osmar.braz@udesc.br, fileto@ufsc.br

Abstract. *Texts abundantly inserted in digital platforms nowadays may have meaning similarities whose automatic detection is essential for applications like plagiarism detection and analysis of social movements. However, detecting semantic similarity of discourses, which may convey analogous ideas using different lexical and syntactic constructions, is a challenge still under-explored. The main objective of this work is to compare approaches to measure and classify the semantic similarity of discourses in short texts. Firstly, it investigates the use of traditional and contextualized embeddings of corresponding structural components of discourses in short texts for measuring and classifying the semantic similarity of the discourses. Then it investigates the use of language models to asses the similarities of the discourses in the raw texts. The performance of these alternative methods was evaluated in experiments using 3 corpora. The experimental results show that properly prompting GPT allows higher performance than using word embeddings to compare discourse components.*

Resumo. *Textos abundantemente inseridos em plataformas digitais atualmente podem apresentar similaridades semânticas cuja detecção automática é essencial para aplicações como identificação de plágio e análise de movimentos sociais. No entanto, a detecção de similaridade semântica entre discursos, que podem transmitir ideias análogas usando diferentes construções léxicas e sintáticas, permanece um desafio pouco explorado. Este trabalho tem como objetivo principal comparar abordagens para medir e classificar a similaridade semântica de discursos em textos curtos. Primeiramente, investiga o uso de embeddings tradicionais e contextualizados de componentes estruturais correspondentes dos discursos. Em seguida, explora o uso de modelos de linguagem para medir e classificar as similaridades diretamente nos textos brutos. A eficácia dessas abordagens foi avaliada em experimentos utilizando 3 corpora distintos. Os resultados experimentais demonstram que o uso adequado de prompts no GPT permite obter um desempenho superior ao uso de embeddings de palavras na comparação de componentes do discurso, estabelecendo assim uma base comparativa para futuros estudos nesta área.*

1. Introdução

Discurso é qualquer expressão verbal ou escrita em um contexto comunicativo [Marcuschi et al. 2002]. A análise de discursos é um tema de pesquisa recente e ainda

pouco explorado, tanto na área de linguística quanto de processamento de linguagem natural (PLN). A análise de discurso observa inclusive a relação entre a língua e a ideologia. Uma ideia pode ser expressa de diversas maneiras. A materialidade da ideologia é o discurso, e a materialidade do discurso é a língua [Orlandi 2005].

O processamento computacional adequado dos discursos expressos em textos e áudio abundantemente postados em plataformas digitais atualmente pode viabilizar muitas aplicações. Esses discursos podem ter semelhanças de significado cuja detecção automática é essencial em aplicações como detecção de plágio e análise de movimentos sociais. A similaridade semântica de discursos é relevante também em tarefas de PLN como sumarização de texto, tradução automática, resposta a perguntas expressas em linguagem natural (em inglês *Question Answring - QA*), análise de sentimentos e avaliação de linguagem [Song and Liu 2020].

Este trabalho foca na similaridade de discursos em textos curtos, muitas vezes com uma única sentença. A similaridade de discurso considera os elementos discursivos presentes em sentenças curtas, analisando como cada uma carrega aspectos do discurso mais amplo, incluindo a intenção comunicativa, o contexto implícito e as relações retóricas, mesmo quando condensadas em uma única frase. Discursos frequentemente dependem de contexto que pode não estar explícito no texto, o que dificulta a identificação apenas por extrações estruturais. O propósito por trás das palavras é crucial na análise de discurso. Discursos podem usar sinônimos, metáforas ou referências indiretas que requerem compreensão semântica. Além disso, a forma como as ideias se conectam é importante no discurso, mesmo em sentenças individuais. Essas características tornam a similaridade de discursos mais sutil e desafiadora para detectar e medir do que a mera afinidade dos significados das palavras nos textos.

A Figura 1 ilustra esses desafios com sentenças em inglês análogas às utilizadas nos experimentos relatados neste trabalho. Note que as sentenças S1 a S6 usam léxicos diferentes para representar sentidos próximos (por exemplo, "*boy*", "*kids*", "*yougsters*" e "*teenagers*"), apesar de sentenças como S1 e S3 terem estruturas sintáticas análogas. Léxicos com sentido de criança ou jovem nas sentenças estão em magenta, criaturas ameaçadoras em amarelo e verbos com sentido de eliminar, ser valente ou ser machucado ou mordido em azul. As sentenças S1, S3 e S5, na coluna da esquerda, carregam a ideia de crianças ou jovens (poderiam também se referir a filhos ou a pessoas quaisquer) que enfrentam e eliminam as criaturas ameaçadoras, enquanto S2, S4 e S6 carregam a ideia de crianças ou jovens (poderiam também ser quaisquer indivíduos) vítimas de tais criaturas (ou de alguma coisa qualquer). Em outras palavras, S1, S3 e S5 carregam a ideologia do enfrentamento, enquanto S2, S4 e S6 colocam os agentes como vítimas, independentemente de quem sejam.

As propostas mais tradicionais para detectar similaridade semântica de textos baseiam-se em análise estrutural [Torkanfar and Azar 2020, Almuhaimeed et al. 2022], métodos estatísticos [Mehndiratta and Asawa 2020] e ontológicos ([Jha et al. 2022, Yang et al. 2021]. No entanto, como tem acontecido com diversas tarefas de PLN, há uma tendência recente de migração para modelos neurais profundos [Joty et al. 2019, Lv et al. 2021, Cao et al. 2022, Wang and Zhang 2021, Wang et al. 2021, Malkiel et al. 2022, An et al. 2020, Chen et al. 2023, Peng et al. 2021, Sonawane and Kulkarni 2022, Xiao et al. 2022]. Todavia, a vasta maioria dos métodos

Figura 1. Sentenças expressando ideias similares de forma diferente

S1 = "The boy who kills the snake is strong."	S2 = "The boy is injured by a snake."
S3 = "Kids who destroy dangerous creatures are brave."	S4 = "The kid was bitten by the serpent."
S5 = "Buff youngsters gun beasts."	S6 = "Those teenagers could not avoid being bitten by the limbless reptiles."

propostos na literatura para medir ou classificar similaridade semântica entre textos não contempla a análise semântica dos discursos. Assim, eles podem falhar na captura de similaridades de ideias, como exemplificado na Figura 1.

Estudos contemplando similaridade de discursos são escassos. Na nossa revisão de literatura, identificamos apenas dois trabalhos com esta temática: [Farouk 2020a] e [Farouk 2020b]. Eles usam representações da estrutura discursiva para identificar termos com papéis correspondentes, cuja semântica pode ser comparada para identificar similaridades de ideias. [Farouk 2020a] captura estruturas de discurso usando a DRT (do inglês, *Discourse Representation Theory*) [Kamp and Reyle 2013] e compara componentes correspondentes dos discursos usando *embeddings* do Word2Vec.

Neste artigo, primeiramente avaliamos variações da abordagem de [Farouk 2020a] baseada na DRT. Porém, além do Word2Vec, investigamos o uso de outros modelos de *embedding*, inclusive contextualizados do BERT. Posteriormente, investigamos o comportamento de um ajuste fino do BERT para classificar similaridades e versões atuais dos grandes modelos de linguagem (LLMs, do inglês, *Large Language Models*) GPT e o LLaMA para mensurar e classificar similaridades usando diretamente os textos.

Os resultados de experimentos com 3 conjuntos de textos curtos, anotados com medidas ou classes de similaridade, revelaram que LLMs podem atingir resultados superiores aos de métodos convencionais inspirados na DRT e usando representações vetoriais, tanto na mensuração quanto na classificação das similaridades dos discursos.

Até onde sabemos, este trabalho é o primeiro a comparar essas abordagens para mensurar e classificar a similaridade de discursos. Suas contribuições incluem: (i) comparação do desempenho de diferentes tipos de *embeddings* na mensuração de similaridades de discursos em implementação atual de método inspirado na DRT; (ii) ajuste fino do BERT para classificação de similaridades de discursos; (iii) elaboração e teste de prompts para efetuar a mensuração e a classificação de similaridades de discursos com LLMs atuais e (iv) comparação do desempenho dessas abordagens em experimentos que mostram a superioridade consistente dos LLM em corpora diversos.

Este artigo segue com os fundamentos necessários ao seu entendimento e a discussão de trabalhos relacionados na Seção 2. A Seção 3 descreve o fluxo de trabalho proposto e implementado em um framework para comparar alternativas para detecção de similaridade de discursos. A Seção 4 reporta os experimentos realizados e a Seção 5 apresenta e discute os resultados. Finalmente, a Seção 6 traça conclusões e enumera alguns temas para trabalhos futuros.

2. Fundamentos

2.1. Representação Formal de Discursos

A Teoria de Representação do Discurso (do inglês, *Discourse Representation Theory* - DRT) [Kamp and Reyle 2013], permite identificar os componentes de discursos expressos verbalmente ou em textos. Ela lida principalmente com a estrutura do discurso para suportar análises de seus significados, com ênfase em aspectos como referência anafórica, tempos verbais e presunções. A DRS (do inglês, *Discourse Representation Structure*) é uma ferramenta usada na DRT para representar a estrutura de um discurso e auxiliar na sua interpretação.

A estrutura do discurso pode ser representada através de um grafo para cada sentença, o qual mostra como ela é dividida em suas diferentes partes gramaticais, e como essas partes se relacionam umas com as outras [Lascarides and Asher 2007]. Há ferramentas para extrair automaticamente representações formais das relações entre componentes de discursos. Neste trabalho, tal como em [Farouk 2020b], que nos serve como linha base, usamos o Boxer conectado à saída do C&C Parser para realizarem tal tarefa.

O Boxer é um analisador estatístico avançado para extrair estruturas DRS [Curran et al. 2007]. Ele se distingue de outros analisadores de estrutura de discursos porque produz representações formais de significado compatíveis com a lógica de primeira ordem, enquanto alcança ampla cobertura [Bos 2015]. O Boxer usa como entrada para gerar a representação do discurso a derivação CCG (*Combinatory Categorial Grammar*) gerada pelo C&C Parser [Hockenmaier 2003], o qual usa a gramática extraída do CCGbank, uma versão CCG do Penn Treebank [Hockenmaier 2003]. Esta gramática tem 425 categorias lexicais, expressando informações de sub categorização, além de um pequeno número de regras que combinam as categorias.

2.2. Modelos de Linguagem de Grande Escala

Os Modelos de Linguagem de Grande Escala *(Large Language Models - LLMs)* são sistemas de inteligência artificial treinada para entender e gerar texto natural em uma ampla variedade de contextos. Esses modelos são baseados em arquiteturas de redes neurais profundas, particularmente a arquitetura Transformer, introduzida por Vaswani [Vaswani et al. 2017]. As aplicações dos LLMs são diversas, incluindo assistentes virtuais, tradução automática, análise de texto e geração de conteúdo em várias indústrias. [Radford et al. 2019]

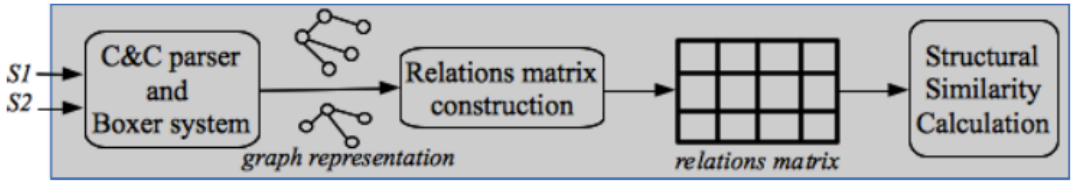
2.3. Mensuração da Similaridade de Componentes das Estruturas dos Discursos

[Farouk 2020a] propõe um método para calcular a similaridade entre os discursos de duas sentenças comparando a semântica de seus componentes estruturais. O autor sustenta que a semelhança semântica entre discursos pode ser determinada através da extração de componentes relevantes na estrutura dos discursos e combinação ponderada das semelhanças entre *embeddings* de componentes correspondentes extraídos dos discursos a comparar.

A Figura 2 ilustra o fluxo de trabalho proposto em [Farouk 2020a] para calcular a similaridade entre os discursos de duas sentenças, $S1$ e $S2$. O Boxer, conectado à saída do analisador sintático C&C Parser, extrai um grafo que representa a estrutura do discurso em cada sentença. Pares desses grafos, referentes a sentenças a comparar, são usados

para gerar a matriz das relações entre componentes das sentenças. A similaridade dos discursos é computada como uma média ponderada das similaridades entre *embeddings* de componentes correspondentes nas estruturas dos discursos.

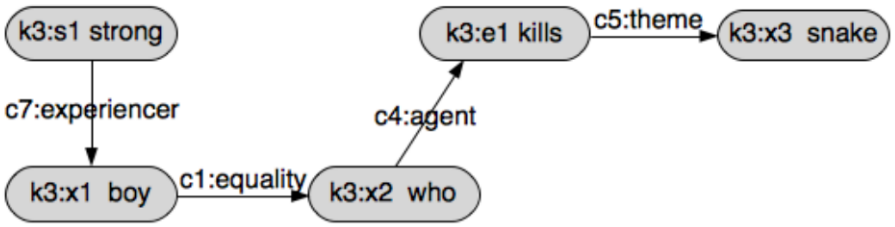
Figura 2. Solução de Farouk para medir similaridade entre discursos curtos.



Fonte: [Farouk 2020a]

O C&C Parser analisa cada sentença e gera a respectiva árvore sintática, a partir da qual o Boxer gera o grafo que representa o discurso da sentença. A Figura 3 mostra o grafo extraído para representar o discurso da sentença "*The boy who kills the snake.*". Nodos representam palavras e arestas representam relações semânticas entre palavras.

Figura 3. Representação da sentença em grafo



Fonte: [Farouk 2020a]

Os grafos gerados para cada par de sentenças são usados para construir uma matriz de relações, cujas linhas representam as relações do primeiro grafo e as colunas as do segundo grafo. Cada célula desta matriz mantém a similaridade entre a relação da linha i e a da coluna j. A Equação 1 calcula a similaridade $RelSim(R_1, R_2)$ entre duas relações R_1 e R_2, levando em conta a similaridade entre as palavras internas I, externas E e os nomes das relações. $NameSim(R_1, R_2)$ compara duas relações, retornando 1 se similares, 0 se não. Finalmente, a Equação 2 calcula a similaridade $Sim(S_1, S_2)$ entre os discursos das sentenças S_1 e S_2 como a soma ponderada das máximas similaridades de cada relação R_i de S_1 com relações de S_2, sendo n o número de relações da estrutura do discurso de S_1. Os pesos W_{Ri} refletem a importância de cada relação na estrutura do discurso. Estes podem ser ajustados para dar mais relevância a certos elementos do discurso. Ao enfatizar a relevância das conexões relacionadas ao verbo, é possível enfatizar a semelhança de seus significados no cálculo da similaridade do discurso. Distingue-se discursos com ideologia de enfrentamento (como "matar" e "destruir") dos com ideologia de vitimização (como "ser machucado" ou "ser mordido"), dando mais peso às relações em torno dos verbos do que aos sujeitos e objetos. Na proposta original de Farouk [Farouk 2020a] e

nos experimentos as relações $agent$ e $theme$ têm peso 8, $experiencer$ 6, a relação is peso 4, in peso 3 e outras relações peso 1.

$$RelSim(R_1, R_2) = \frac{Sim(I_{R_1}, I_{R_2}) + Sim(E_{R_1}, E_{R_2})}{2} * NameSim(R_1, R_2) \quad (1)$$

$$Sim(S_1, S_2) = \frac{\sum_i^n maxSim(R_i, S_2) * W_{R_i}}{\sum_i^n W_{R_i}} \quad (2)$$

2.4. Trabalhos Relacionados

Modelos de similaridade semântica tradicionais são fundamentadas em análise estrutural [Torkanfar and Azar 2020, Almuhaimeed et al. 2022], métodos estatísticos [Mehndiratta and Asawa 2020] e ontológicos ([Jha et al. 2022, Yang et al. 2021]. Porém, como tem acontecido com diversas tarefas de PLN, há uma tendência recente de migração para modelos neurais profundos [Joty et al. 2019, Lv et al. 2021, Cao et al. 2022, Wang and Zhang 2021, Wang et al. 2021, Malkiel et al. 2022, An et al. 2020, Chen et al. 2023, Peng et al. 2021, Sonawane and Kulkarni 2022, Xiao et al. 2022].

Enquanto a similaridade semântica é centrada na correspondência de significado de palavras ou frases, a similaridade de discurso amplia o foco para considerar a estrutura e a ideia geral permeando o texto [Farouk 2020a]. A similaridade de discurso envolve o estudo de estruturas linguísticas, sequências de ideias e como essas ideias se conectam para formar um argumento ou ponto de vista coerente [Song and Liu 2020, Joty et al. 2019]. Isso permite uma compreensão mais profunda da mensagem central, possivelmente subliminar, de crença em algum ideário – a ideologia – que o autor transmite, consciente ou inconscientemente, intencionalmente ou não, através do seu discurso.

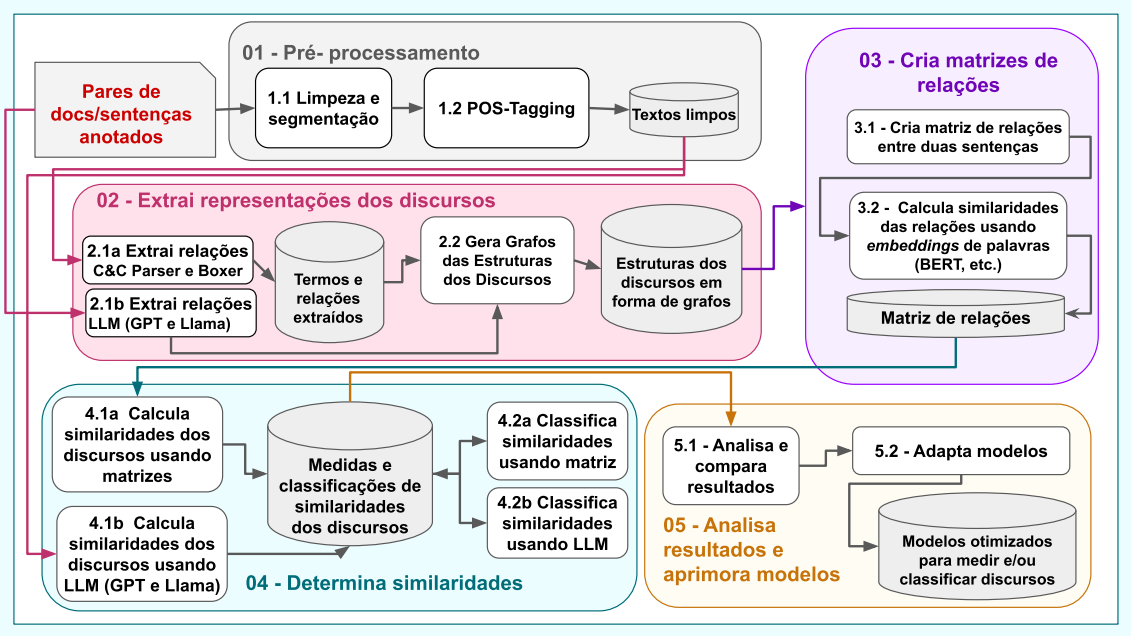
[Farouk 2020a] foi escolhido como base na nossa pesquisa por sua fundamentação na DRT e proximidade com os nossos objetivos de pesquisa em similaridade de discurso. Neste trabalho avançamos investigando o uso de diversos modelos de *embedding* para comparar componentes de discursos identificados com a DRT, ajuste fino do BERT para classificar similaridades e LLMs para mensurar e classificar similaridades de discursos. Experimentos mais extensivos com técnicas do estado da arte e diferentes conjuntos de dados nos permitem monstrar a superioridade dos LLMs também nessas tarefas.

3. O SIDEAS

O SIDEAS (*SImilarity of iDEAS*) é um framework desenvolvido no âmbito deste trabalho para avaliar alternativas para comparar a similaridade de discursos expressos em textos curtos. Sua versão atual permite desde a extração e comparação de componentes essenciais dos discursos, conforme a DRT, até o uso de LLMs para mensurar e classificar as similaridades dos discursos diretamente a partir dos textos. A Figura 4 ilustra o fluxo de processamento de dados do SIDEAS, o qual é dividido nas etapas descritas a seguir.

Etapa 1: Pré-processamento envolve a limpeza dos textos, remoção de elementos irrelevantes, segmentação de sentenças, tokenização e marcação das palavras com suas respectivas classes gramaticais (*PoS-Tagging - Part-of-Speech Tagging*).

Figura 4. Fluxo de trabalho do SIDEAS para detectar similaridades de discursos



Etapa 2: Extração de Representações dos Discursos transforma as sentenças individuais limpas em representações estruturais do discurso no formato de grafos. Embora estejamos trabalhando com sentenças curtas, essa representação captura elementos discursivos essenciais presentes em cada sentença. A versão desta etapa baseada na DRT usa o C&C Parser e o Boxer. Outra alternativa é utilizar LLMs para extrair as estruturas do discursos e gerar um grafo correspondente. Por limitação de espaço, esta alternativa não foi avaliada nos experimentos relatados neste trabalho, ficando para trabalhos futuros.

Etapa 3: Geração da Matriz de Relações cria uma matriz de relações a partir dos grafos que representam os discursos a comparar, onde cada célula mantém a similaridade cosseno entre de um par de relações das respectivas sentenças, calculada de acordo com a Equação 1, usando algum modelo de *embedding*.

Etapa 4: Determinação da Similaridade determina a similaridade entre pares de sentenças, podendo ser realizada de maneiras alternativas:

- **Cálculo de Similaridade Discursiva** usando a matriz de similaridade de relações através da Equação 2.
- **Classificador de similaridade de discurso** treinado mediante ajuste fino de modelo de linguagem pré-treinado usando pares de sentenças rotuladas.
- **Cálculo e Classificação de Similaridades por LLMs** usando prompts.

Etapa 5: Avaliação e Aprimoramento faz a avaliação dos resultados e a adaptação dos modelos, buscando otimizar a eficácia da representação discursiva.

4. Experimentos

Primeiramente, foram realizados experimentos de mensuração das similaridades de discursos em pares de sentenças, usando variações da abordagem proposta por Farouk [Farouk 2020a], com *embeddings* tradicionais do Word2Vec e Glove, e contextu-

alizados do BERT. Em seguida, foi feito um ajuste fino do BERT para classificar similaridades e explorados os LLMs GPT e LLaMA para calcular e classificar similaridades de discursos diretamente nos textos.

4.1. Conjuntos de Dados

Os experimentos usaram três conjuntos de dados da literatura, cujas características são resumidas na Tabela 1. Eles representam uma variedade domínios, estruturas discursivas e níveis de similaridade. Li2006 é um conjunto de dados extraídos do Collins Cobuild Dictionary, com 65 pares de sentenças anotados com escores de similaridade variando no intervalo contínuo $[0, 1]$, atribuídos por 32 avaliadores humanos. Cada sentença tem em média $15,52$ palavras, $4,22$ substantivos e $2,54$ verbos (incluindo principais e auxiliares).

Tabela 1. Estatísticas dos Conjuntos de Dados

Conjunto de dados	Li2006	SICK	MSRP
N° de Pares de Sentenças	65	10.000	5.800
Média de Palavras por Sentença/Doc	15,52	19,45	22,12
Média de Subst. p/ Sentença/Doc	4,22	5,23	4,27
Média de verbos + aux por documento	2,54	3,32	4,38
N° de Avaliadores	32	-	2
Escala de Similaridade	0,00 - 1,00	1 - 5	1 ou 0
Origem dos Dados	[O'Shea et al. 2008]	[Marelli et al. 2014]	[Dolan and Brockett 2005]

O SICK tem 10.000 pares de sentenças provenientes de fontes como 8K ImageFlick, SemEval-2012 STS e MSR-Video. Esses pares são rotulados com similaridades entre 1 a 5 em uma escala com variações de $0,5$. Cada sentença tem em média $19,45$ palavras, $5,23$ substantivos e $3,32$ verbos. Por fim, o conjunto MSRP tem 5.800 pares de sentenças extraídas de artigos de notícias, anotados como similares (1) ou dissimilares (0), por dois avaliadores humanos. Cada sentença tem em média $22,12$ palavras, $4,27$ substantivos e $4,38$ verbos. O MSRP foi projetado para treinar e avaliar modelos para a tarefa de classificação binária de paráfrase.

4.2. Comparação dos Discursos usando *embeddings* de Componentes DRT

O código fonte original de [Farouk 2020a], escrito em C++, não era facilmente replicável, dificultando a reprodução dos experimentos, adaptações para utilizar modelos de *embedding* recentes e a comparação com outros métodos. Para contornar essas limitações, o código foi convertido para Python utilizando bibliotecas atuais. Tal adaptação foi feita cuidadosamente, garantindo que a lógica e os cálculos subjacentes fossem mantidos. Essa adaptação não apenas torna o código mais acessível à comunidade científica, mas também permite uma integração mais fácil com bibliotecas populares de processamento de linguagem natural em Python.

- Word2Vec: Um modelo de *embeddings* de palavras treinado no corpus Google News proposta por [Mikolov et al. 2013].
- GloVe: Um modelo de representação vetorial de palavras treinado no corpus Common Crawl, utilizando a técnica proposta por [Pennington et al. 2014].
- Paragram: Uma variação do Word2Vec incorporando conhecimento de relações paradigmáticas [Wieting et al. 2015].

- BERT-base-multilingual-case: Um modelo contextualizado BERT pré-treinado em várias línguas [Devlin et al. 2019].

4.3. Ajuste Fino do BERT para Classificar Similaridades de Discursos

As faixas de valores de hiperparâmetros sugeridas pelos autores da variação do BERT utilizada nos nossos experimentos foram combinadas: taxa de aprendizagem de $1 * 10^{-5}$, $2 * 10^{-5}$, $3 * 10^{-5}$, $4 * 10^{-5}$, $5 * 10^{-5}$ (incluímos $1 * 10^{-5}$ e $4 * 10^{-5}$) e número de épocas em $\{1, 2, 3, 4, 5\}$. Dessa forma, foram realizadas 250 execuções (10 *folds* $\times$ 5 taxas de aprendizado $\times$ 5 épocas) para cada conjunto de dados avaliado.

4.4. Mensuração e Classificação de Discursos usando LLM

Os conjuntos de dados contêm muitos pares de sentenças. Não é viável processar todas de uma só vez. Assim, foram selecionados subconjuntos aleatórios de pares rotulados até o limite de tokens, para serem fornecidos aos LLMs com a hashtag de #treinamento. Isso permitiu expor os modelos a exemplos de pares de sentenças similares e dissimilares, auxiliando no aprendizado da tarefa.

Em seguida, na fase de avaliação, fornecemos ao LLM um novo conjunto de dados selecionando sentenças no tamanho do limite de tokens da requisição, contendo apenas pares de sentenças sem pontuações de similaridade, divididos em blocos menores e processados separadamente pelos LLMs, utilizando a hashtag de #avaliação. O LLM foi instruído a usar seu conhecimento adquirido no treinamento para prever a similaridade de cada par de sentenças. Os resultados dos LLMs foram comparados com as pontuações atribuídas por humanos, usando correlação de Pearson e Spearman para comparar mensurações de similaridade e as métricas acurácia e precisão para avaliar resultados de classificação.

5. Resultados

5.1. Mensuração da Similaridade dos Discursos

Os resultados apresentados a seguir demonstram a eficácia das abordagens propostas na detecção de similaridade de discurso em sentenças individuais. É notável como, mesmo em unidades textuais curtas, foi possível capturar e comparar elementos discursivos relevantes.

A Tabela 2 apresenta os resultados de mensuração da similaridade semântica entre discursos de pares de sentenças. Note que o GPT-4 obteve o melhor desempenho geral no conjunto de dados Li2006, alcançando coeficientes de $0,89$ e $0,93$, para correlação de Pearson de Spearman, respectivamente. No conjunto de dados SICK o GPT-4 ainda apresentou a melhor correlação de Pearson, mas o LLaMA 3 a melhor correlação de Spearman. Essa discrepância sugere que o melhor desempenho do LLaMA 3 na correlação de ranking se deve à sua maior capacidade de capturar a semântica das construções linguísticas mais complexas e diversificadas das sentenças do SICK.

A Figura 5 detalha correlações obtidas para o conjunto de dados Li2006, usando o método baseado na DRT com *embeddings* do Word2Vec, GloVe, Paragram e BERT, além do GPT-4 e do LLaMA 3. Esses resultados sugerem que os LLMs treinados em grandes volumes de dados de texto foram capazes de capturar com eficácia as nuances semânticas presentes nas sentenças avaliadas.

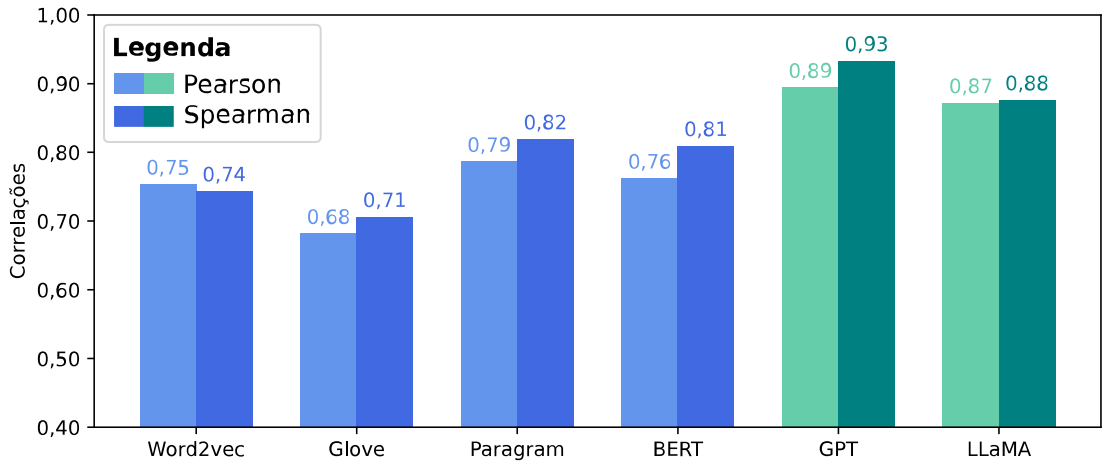
Tabela 2. Comparação dos resultados de mensuração de similaridade

Conjunto de dados	Modelo	Pearson	Spearman
Li2006	*Farouk**	*0,87*	*0,89*
	BERT	0,76	0,81
	GPT-3.5	0,80	0,80
	GPT-4	**0,89**	**0,93**
	LLaMA 3	0,87	0,88
SICK	GPT-3.5	0,42	0,50
	GPT-4	**0,67**	0,66
	LLaMA 3	0,57	**0,70**

∗ **Valor publicado em [Farouk 2020a]**

Note que Word2Vec e Glove levaram a desempenho inferior ao alcançados com *embeddings* do Paragram e contextualizados do BERT. Porém, os modelos de linguagem tiveram melhor desempenho do que o método baseado em DRT com quaisquer dessas incorporações.

Figura 5. Correlações da regra ouro de Li2006 com as medidas de similaridade obtidas usando DRT com diferentes *embeddings* e LLMs



5.2. Classificação da Similaridade dos Discursos

A Tabela 3 apresenta os resultados obtidos pelos LLMs, pelo classificador *BERT LG case* e pelo método de [Farouk 2020a] na classificação das similaridades dos pares de sentenças do conjunto de dados MSRP. Note que o GPT-4o alcançou o melhor desempenho geral, com acurácia de $0,94$ e a precisão perfeita de $1,0$. Esse resultados possivelmente se devem à maior capacidade do GPT-4o de capturar sutilezas semânticas dos discursos. O GPT-3.5 também apresentou um bom desempenho, com acurácia de $0,80$ e precisão de $0,95$. Por exemplo, em casos onde sentenças usavam sinônimos complexos ou estruturas sintáticas variadas para expressar ideias similares, o GPT-4o foi consistentemente capaz de identificar a similaridade subjacente.

Por outro lado, o modelo GPT-4 obteve uma acurácia de $0,81$ e uma precisão de $0,83$, desempenho inferior ao GPT-4o, mas ainda relativamente alto. O modelo LLaMA 3 obteve uma acurácia mais modesta de $0,73$, mas uma precisão perfeita de $1,0$. Essa discrepância entre acurácia e precisão sugere que o LLaMA 3 pode ter uma tendência a favorecer a identificação de instâncias positivas (sentenças similares), às custas de um maior número de falsos negativos. O *BERT LG Case* obteve uma acurácia de $0,84$ e uma precisão de $0,88$ enquanto o *BERT BASE* obteve a mesma acurácia mas com uma precisão de $0,89$, demonstrando seu potencial na tarefa de classificação de similaridade. O método de [Farouk 2020a] obteve uma acurácia de $0,71$ e uma precisão de $0,76$, desempenho inferior aos modelos LLM e *BERT LG Case*.

Tabela 3. Comparação dos resultados de classificação de similaridade

Conjunto de dados	LLM	Acurácia	Precisão	Classificador	Acurácia	Precisão
MSRP	GPT-3.5	0,80	0,95	*BERT LG Case*	0,84	0,88
	GPT 4	0,81	0,83	*BERT Base*	0,84	0,89
	LLaMA 3	0,73	1	Farouk*	0,71	0,76
	GPT-4o	**0,94**	**1**			

∗ **Valor publicado em [Farouk 2020a]**

6. Conclusões e Trabalhos Futuros

Este estudo foi dedicado ao desenvolvimento e à avaliação comparativa de modelos avançados para a mensuração e classificação da similaridade discursiva em textos. Partindo de soluções existentes, este trabalho propôs inovações significativas, incorporando técnicas e ferramentas contemporâneas de PLN, tais como diferentes modelos de *embedding* e o uso dos LLMs GPT-3.5, GPT-4, LLaMA 3 e GPT-4o na investigação das similaridades de discursos em textos curtos.

Os resultados obtidos demonstram o desempenho superior do GPT-4 e do GPT-4o, tanto na mensuração quanto na classificação da similaridade dos discursos. Isso provavelmente se deve ao vasto conhecimento capturado em seu treinamento e à capacidade de capturar nuances semânticas complexas dos LLMs. No entanto, observou-se que o desempenho pode variar de acordo com características como o domínio dos dados.

É importante notar que os três conjuntos de dados utilizados neste estudo (Li2006, SICK e MSRP) foram coletados antes de 2015, o que precede o período de treinamento dos modelos de linguagem como GPT e LLaMA. Embora isso possa sugerir que o desempenho superior desses modelos se deve à sua capacidade de generalização e compreensão semântica, não podemos descartar completamente a possibilidade de memorização parcial dos dados durante o treinamento. Estudos futuros poderiam abordar essa questão utilizando conjuntos de dados mais recentes ou criados especificamente para testar a capacidade de generalização desses modelos em contextos totalmente novos. Além disso, técnicas para detectar e mitigar os efeitos da memorização em modelos de linguagem poderiam ser exploradas para garantir a robustez dos resultados em aplicações de similaridade de discurso.

Temas para trabalhos futuros incluem (i) investigar extração automática de estruturas discursivas com LLMs; (ii) explorar design de prompts e outras técnicas para uso

efetivo de LLMs na captura e comparação de discursos; (iii) aplicar os métodos propostos para análise de discurso a grandes corpora de domínios variados, visando auxiliar em aplicações como detecção plágio ao nível de discursos e análises de movimentos sociais, (iv) experimentos com conjuntos de dados em língua portuguesa e multilíngues, explorando inclusive a capacidade cross-lingual dos LLMs e (v) estender a análise de similaridade de discurso para textos mais longos, investigando como as técnicas desenvolvidas para sentenças podem ser adaptadas e expandidas para parágrafos e documentos completos.

Agradecimentos

Agradecemos à Google Cloud pelo créditos de pesquisa concedidos para executar parte dos experimentos. Este trabalho também teve o apoio de projeto CNPq Universal concedido em 2022, do projeto FAPESC 2021TR1510, do Print CAPES-UFSC Automação 4.0 e do Céos, projeto financiado pelo Ministério Público de Santa Catarina (MPSC), que tem contribuído de maneira relevante para a melhoria do Laboratório LISA e da infraestrutura de processamento de alto desempenho da UFSC.

Referências

Almuhaimeed, A., Alhomidi, M. A., Alenezi, M. N., Alamoud, E., and Alqahtani, S. (2022). A modern semantic similarity method using multiple resources for enhancing influenza detection. *Expert Systems with Applications*, 193:116466.

An, H., Wu, D., and Li, Z. (2020). Hybrid self-interactive attentive siamese network for medical textual semantic similarity - proceedings of the 2020 4th international conf. on management engineering, software engineering and service sciences. page 52–56.

Bos, J. (2015). Open-domain semantic parsing with boxer. In *Proceedings of the 20th nordic conference of computational linguistics (NODALIDA 2015)*, pages 301–304.

Cao, S., Vo, H., Le, H. T.-T., and Dinh, D. (2022). Hybrid approach for text similarity detection in vietnamese based on sentence-bert and wordnet - proceedings of the 4th international conference on information technology and computer communications. page 59–63.

Chen, Q., Zhao, G., Wu, Y., and Qian, X. (2023). Fine-grained semantic textual similarity measurement via a feature separation network. *Applied Intelligence*.

Curran, J. R., Clark, S., and Bos, J. (2007). Linguistically motivated large-scale nlp with c&c and boxer. In *Proceedings of the 45th annual meeting of the ACL Companion volume proceedings of the demo and poster sessions*, pages 33–36.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the ACL*, pages 4171–4186, Minneapolis, Minnesota. ACL.

Dolan, B. and Brockett, C. (2005). Automatically constructing a corpus of sentential paraphrases. In *Third International workshop on paraphrasing (IWP2005)*.

Farouk, M. (2020a). Measuring sentences similarity based on discourse representation structure. *Computing and Informatics*, 39(3):464–480.

Farouk, M. (2020b). Measuring text similarity based on structure and word embedding. *Cognitive Systems Research*, 63:1–10.

Hockenmaier, J. (2003). Data and models for statistical parsing with combinatory categorial grammar.

Jha, A., Rakesh, V., Ch, rashekar, J., Samavedhi, A., Reddy, C., and an K. (2022). Supervised contrastive learning for interpretable long-form document matching. *ACM Trans. Knowl. Discov. Data*. Just Accepted.

Joty, S., Carenini, G., Ng, R., and Murray, G. (2019). Discourse analysis and its applications. In *Proceedings of the 57th Annual Meeting of the ACL: Tutorial Abstracts*, pages 12–17.

Kamp, H. and Reyle, U. (2013). *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory*, volume 42. Springer Science & Business Media.

Lascarides, A. and Asher, N. (2007). Segmented discourse representation theory: Dynamic semantics with discourse structure. In *Computing meaning*, pages 87–124. Springer.

Lv, C., Wang, F., Wang, J., Yao, L., and Du, X. (2021). Siamese multiplicative lstm for semantic text similarity - 2020 3rd international conference on algorithms, computing and artificial intelligence.

Malkiel, I., Ginzburg, D., Barkan, O., Caciularu, A., Weill, J., and Koenigstein, N. (2022). Interpreting bert-based text similarity via activation and saliency maps - proceedings of the acm web conference 2022. page 3259–3268.

Marcuschi, L. A. et al. (2002). Gêneros textuais: definição e funcionalidade. *Gêneros textuais e ensino*, 2:19–36.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., and Zamparelli, R. (2014). A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Mehndiratta, A. and Asawa, K. (2020). Spectral Learning of Semantic Units in a Sentence Pair to Evaluate Semantic Textual Similarity - big Data Analytics. 12581:49–59. Series Title: Lecture Notes in Computer Science.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Orlandi, E. P. (2005). Michel pêcheux e a análise de discurso (michel pêcheux et l'analyse de discours). *Estudos da Língua (gem)*, 1(1):9–13.

O'Shea, J., Bandar, Z., Crockett, K., and McLean, D. (2008). Pilot short text semantic similarity benchmark data set: Full listing and description. *Computing*.

Peng, D., Hao, B., Tang, X., Chen, Y., Sun, J., and Wang, R. (2021). Learning long-text semantic similarity with multi-granularity semantic embedding based on knowledge enhancement - proceedings of the 2020 1st international conference on control, robotics and intelligent system. page 19–25.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Sonawane, S. S. and Kulkarni, P. (2022). Concept based document similarity using graph model. *International Journal of Information Technology*, 14(1):311–322.

Song, W. and Liu, L. (2020). Representation learning in discourse parsing: A survey. *Science China Technological Sciences*, 63(10):1921–1946.

Torkanfar, N. and Azar, E. (2020). Quantitative similarity assessment of construction projects using wbs-based metrics. *Advanced Engineering Informatics*, 46:101179.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, K., Zeng, Y., Meng, F., Feiyu, and Yang, L. (2021). Comparison between calculation methods for semantic text similarity based on siamese networks - 2021 4th international conference on data science and information technology. page 389–395.

Wang, Z. and Zhang, B. (2021). Chinese text similarity calculation model based on multi-attention siamese bi-lstm - proceedings of the 4th international conference on computer science and software engineering. page 93–98.

Wieting, J., Bansal, M., Gimpel, K., and Livescu, K. (2015). From paraphrase database to compositional paraphrase model and back. *Transactions of the ACL*, 3:345–358.

Xiao, Q., Qin, Y., Li, K., Tang, Z., Wu, F., and Liu, Z. (2022). An unsupervised semantic text similarity measurement model in resource-limited scenes. *Information Sciences*, 616:444–460.

Yang, J., Li, Y., Gao, C., and Zhang, Y. (2021). Measuring the short text similarity based on semantic and syntactic information. *Future Generation Computer Systems*, 114:169–180.