Investigação de métodos de similaridade textual no contexto da avaliação automática de questões discursivas

José Augusto O. da S. Almeida[1], **Raimundo Santos Moura**[1]

[1]Universidade Federal do Piauí (UFPI) – Teresina, PI - Brasil

{jose10augusto10, rsm}@ufpi.edu.br

Abstract. *This article addresses the automated assessment of essay questions, a significant challenge for educators due to their laborious nature. Despite advances in this area, challenges persist such as model fluctuation, the scarcity of datasets in Portuguese and the diversity of methods available without an established standard. The objective of this work is to study and evaluate different textual similarity techniques and tools for the automatic correction of discursive questions, aiming to offer solutions to the aforementioned challenges. In the experiments carried out, an average error of 13.9 points was obtained, considering a scale of 0 to 100, for the best model, which makes it encouraging to use such an approach in the educational context.*

Resumo. *Este artigo aborda a avaliação automatizada de questões discursivas, um desafio significativo para os educadores devido à sua natureza laboriosa. Apesar dos avanços nessa área, persistem desafios como a oscilação dos modelos, a escassez de conjuntos de dados em português e a diversidade de métodos disponíveis sem um padrão estabelecido. O objetivo deste trabalho é estudar e avaliar diferentes técnicas e ferramentas de similaridade textual para a correção automática de questões discursivas, visando oferecer soluções para os desafios mencionados. Nos experimentos realizados obteve-se um erro médio de 13,9 pontos, considerando uma escala de 0 a 100, para o melhor modelo, o que torna animador utilizar tal abordagem no contexto educacional.*

1. Introdução

No contexto educacional, a avaliação desempenha um papel crucial no processo de aprendizagem, fornecendo *insights* sobre o nível de assimilação do conteúdo pelos educandos. Segundo [Shields 2022], a avaliação em sala de aula desempenha um duplo propósito, se comportando como formativa, auxiliando na melhoria do aprendizado do aluno durante a instrução, e somativa, documentando a proficiência do aluno no final de uma unidade instrucional. Contudo, essa etapa, embora essencial, frequentemente impõe desafios significativos aos professores, responsáveis por corrigir avaliações e atividades. A sobrecarga resultante desse processo corretivo pode impactar negativamente a qualidade de vida desses profissionais, levando-os a realizar grande parte do trabalho em casa [Galhardi et al. 2020].

Diante desse cenário desafiador, esta pesquisa propõe a aplicação de métodos avançados de Processamento de Linguagem Natural(PLN), um subcampo da Ciência da Computação e da Inteligência Artificial(IA) que foca na interação entre computadores

e seres humanos em linguagem natural, de modo a permitir que tais computadores entendam, interpretem e gerem a linguagem humana [Dande and Pund 2023]. Assim, o objetivo principal é explorar a viabilidade desses métodos na automação do processo de correção, comparando a proximidade entre os resultados gerados por eles e os resultados fornecidos pelos professores.

Essa abordagem resulta em benefícios notáveis tanto para os docentes, aliviando a carga de trabalho, quanto para os discentes, que frequentemente aguardam um tempo considerável para receber seus resultados [Galhardi et al. 2020]. Outrossim, tendo em vista o processo de popularização de novas metodologias de aprendizagem associadas a tecnologias, como os Ambientes Virtuais de Aprendizagem(AVA) [GOMES and PIMENTEL 2021], o sucesso de tal abordagem corrobora ainda mais para o alavancar dessas novas metolodogias.

O restante desse trabalho se subdivide da seguinte forma: A seção 2 apresenta os trabalhos relacionados, contendo algumas pesquisas que foram desenvolvidas no decorrer do tempo; a seção 3 apresenta os modelos e experimentos utilizados na pesquisa em questão; na seção 4 os resultados obtidos são discutidos e, por fim, a seção 5 apresenta uma conclusão diante do que fora exposto anteriormente.

2. Trabalhos Relacionados

A tarefa de correção automática de questões discursivas tem evoluído cada vez mais com o decorrer do tempo, sendo impulsionada principalmente pelos avanços na área de PLN e IA. Diversos estudos foram e têm sido conduzidos, de modo a contribuir significativamente para o desenvolvimento de sistemas capazes de analisar e pontuar respostas textuais de maneira eficiente.

Ao se fazer uma análise de anterioridade de pesquisas, quando se fala em correção automatizada, destaca-se o contexto de redações, com uma quantidade bem maior de pesquisas feitas nesse sentido. [Burstein et al. 2001] conduziram a investigação de um software denominado *E-rater*, utilizado para produzir pontuações holísticas de redações, baseadas em *features* de escrita eficaz que os docentes normalmente já utilizavam, tais como organização, estrutura da sentença e conteúdo. A proposta dos autores foi expandir essa pesquisa para a área de correção automatizada de respostas curtas, através de um sistema chamado *C-rater*. Segundo eles, os resultados, mesmo naquela época, já haviam sido bem promissores, chegando a níveis de concordância de até 80% entre o resultado do *C-rate*r e o resultado do avaliador humano.

[Mohler and Mihalcea 2009] exploraram técnicas não-supervisionadas para realizar a tarefa de correção automatizada de questões discursivas. A ideia era utilizar métricas de similaridade textual baseadas em conhecimento e em *corpus*, juntamente com novas técnicas, a fim de melhorar o desempenho do sistema como um todo, fornecendo o feedback automatizado aos estudantes. Através das pesquisas feitas, eles observaram que as duas métricas testadas tiveram resultados bem semelhantes, porém, a métrica baseada em *corpus* passa a ter uma vantagem adicional pela possibilidade de se aumentar a sua dimensão, o que poderia trazer um desempenho maior.

Tendo por objetivo aprofundar os estudos e testar novos métodos referentes à similaridade entre sentenças, [Galhardi et al. 2020] fizeram uma combinação de diversas

abordagens a fim de construir um modelo final que pudesse realizar a correção automatizada de questões discursivas, se utilizando de técnicas relacionadas à aprendizagem de máquina. De modo geral, a pesquisa consistiu em avaliar 6 diferentes abordagens para fazer a modelagem das respostas, de maneira que cada uma dessas abordagens teve sua própria técnica de pré-processamento, algoritmo de machine learning, bem como seus próprios parâmetros internos. Após os experimentos, eles pegaram a melhor variação para cada grupo e realizaram uma combinação, que tinha por objetivo melhorar o desempenho geral do sistema. Por fim, os resultados advindos da combinação foram comparados com os resultados fornecidos pelos avaliadores humanos, obtendo um nível de concordância moderada de 0,4 a 0,6, considerando a métrica *Cohen Kappa*.

Uma outra pesquisa bastante relevante nessa área foi realizada por [de Oliveira et al. 2020], que consistiu em avaliar técnicas para correção automatizada de questões discursivas, tendo como finalidade construir um Sistema Tutor Inteligente(STI) chamado *MAZK*. A proposta utilizou métodos para cálculo de similaridade baseados em *word embeddings*, que permitem representar palavras como vetores numéricos de baixa dimensão, densos e contínuos, ou seja, como um vetor de palavras [Bao et al. 2022]. Desse modo, tendo posse dos vetores referente à resposta do aluno e a uma resposta modelo fornecida pelo professor, é possível calcular a similaridade utilizando diferentes técnicas, sendo que, no contexto da pesquisa em questão, foi testado o cálculo de similaridade por cosseno [Januzaj and Luma 2022] e por *Word Mover Distance*(*WMD*) [Yamagiwa et al. 2022]. Após realizar os experimentos, os pesquisadores chegaram a resultados bastante promissores, com uma acurácia de 88,7%.

Ante o exposto anteriormente, é perceptível que, diante do panorama dinâmico da correção automatizada de questões discursivas, os estudos mencionados evidenciam a progressão significativa que ocorreu ao longo do tempo. Desde as primeiras incursões, como o software *E-rater* em 2001, até às abordagens mais recentes, como as combinações de técnicas propostas por [Galhardi et al. 2020] e a implementação de *word embeddings* no STI *MAZK*, pode-se observar um panorama de evolução constante.

Em adição, vale pontuar algumas soluções mercadológicas que têm atuado na correção automática de questões discursivas. O primeiro é o *AvaliaSmart*[1], o qual afirma realizar a correção de questões discursivas em instantes, com um tempo de até 8 vezes mais rápido que o de um professor. O diferencial dessa pesquisa em relação à solução é ser gratuita e realizar a correção em um tempo ainda menor, uma vez que o melhor modelo testado é capaz de corrigir uma questão discursiva em questão de segundos. Outra solução existente é a do Instituto Avalia[2], o qual permite a realização de provas online, mas com uma correção que ocorre em até 48 horas, indicando que um processo automatizado não é utilizado. Por fim uma última proposta que vale mencionar é do sistema Multiprova[3], a qual realiza a correção automatizada apenas para questões objetivas. Assim, fica evidente que a pesquisa é relevante e traz um diferencial com relação a soluções consolidadas no mercado.

[1]https://www.avaliasmart.com.br/

[2]https://www.avalia.org.br/prova-online

[3]https://site.multiprova.ufrn.br/

3. Modelos e Experimentos

Durante o decorrer da pesquisa, optou-se por utilizar a linguagem *Python* através da plataforma *Colab* fornecida pela *Google*. Tal escolha se deu devido a algumas características da linguagem, como ser de alto nível, com regras claras, elegantes e concisas, além de ter suporte a um grande número de bibliotecas de terceiros, tornando-a amplamente aplicada em *web crawlers*, análise de dados, *machine learning*, IA, PLN e outros campos [Wang and Hu 2021]. A Figura 1 retrata os passos realizados ao longo dos experimentos realizados.

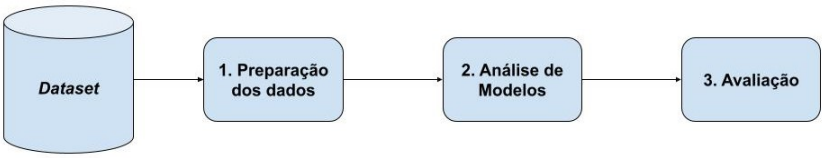

Figura 1. Visão geral das etapas do projeto

Como o objetivo foi realizar a investigação de métodos de similaridade textual no contexto de questões discursivas, foi necessário selecionar um conjunto adequado de dados, o qual contivesse o enunciado de uma determinada questão, as respostas dadas pelos alunos, além de pelo menos uma resposta de referência fornecida pelo professor ou responsável. Desse modo, foi considerado o *corpus* utilizado na pesquisa de [de Oliveira et al. 2020], o qual será denominado, a partir de agora, *Dataset_1*

É importante ressaltar algumas caracterísitcas do *Dataset_1*, as quais devem ser levadas em consideração para uma melhor compreensão dos experimentos realizados. Ele é composto por 6 questões relacionadas à área de IA, 1 resposta de referência para cada questão, além de 5 respostas de alunos para cada questão, apresentando um total de 30 respostas de alunos. As notas foram atribuídas pelos professores em um limiar de 0 a 100.

A primeira etapa consistiu na preparação dos dados. Nesta etapa, aplicou-se técnicas de pré-processamento, como a transformação da sentença para minúsuculo, remoção de *stopwords*, pontuações, caracteres não-alfanuméricos, caracteres únicos, números e espaços vazios. Outro ponto a se considerar é que, como os dados foram obtidos em português, optou-se por realizar um experimento extra em inglês, para verificar a sensibilidade dos modelos quanto às diferenças linguísticas. Já na etapa 2 foi feita uma análise dos modelos, onde investigou-se quatro métodos de similaridade textual disponíveis em bibliotecas *Python*, a saber: i) *SpaCy*; ii) *Gensim*; iii) *Transformers*; e iv) *Gemini*.

O *SpaCy* foi escolhido pois fornece uma ampla gama de recursos embutidos e é bastante eficiente para tarefas de processamento de texto e modelagem de linguagem [Amade et al. 2024], enquanto que a biblioteca *Gensim* é também muito famosa e permite obter vetores de palavras(do inglês, *word vectors*) de alta qualidade para a realização dos cálculos de similaridade [Savytska et al. 2022]. Já o terceiro método, baseado em *Transformers*, foi selecionado devido à sua capacidade avançada de compreensão de contexto e nuances linguísticas, proporcionando uma análise mais precisa em textos complexos [Vaswani et al. 2017]. Por fim, a inclusão do *Gemini* deve-se ao seu potencial inovador, além de uma maior acessibilidade e transparência, se comparada a outros modelos

de língua [McIntosh et al. 2023], de modo que seus recursos de *embeddings* podem ser utilizados facilmente.

Considerando isso, para cada modelo, foram feitos 4 experimentos. O primeiro utilizou o *corpus* pré-processado, ou seja, tanto as respostas de referência quanto às fornecidas pelos discentes passaram pelo pré-processamento. Além disso, o primeiro experimento foi feito com ambas as respostas comparadas estando na língua portuguesa. Já o segundo considerou as respostas novamente em português, porém sem o pré-processamento, o terceiro, respostas em inglês com pré-processamento e, no último, respostas em inglês sem o pré-processamento.

Dando continuidade, o primeiro modelo considerado foi o da biblioteca *SpaCy*, que utiliza para seus cálculos a similaridade por cosseno, a qual realiza a medida entre os dois vetores, sendo, no caso em questão, a resposta esperada e a fornecida. Assim, quanto menor for o ângulo formado entre esses dois vetores, maior é o nível de similaridade entre eles [Januzaj and Luma 2022]. Já para o segundo modelo, a biblioteca utilizada foi a *Gensim*[4] e a medida utilizada foi a *WMD*, que faz o cálculo da distância entre dois documentos baseado na *Earth Mover's Distance*(*EMD*), de maneira que, na prática, o que será calculado é a distância sobre o espaço vetorial de palavras [de Oliveira et al. 2020].

Ademais, o terceiro modelo contou com o uso de *Sentence Transformers*[5], o qual, antes de aplicar o cálculo da similaridade em si, utilizando cosseno por exemplo, os vetores são codificados em *embeddings* com base nos diferentes modelos disponibilizados no site *Hugging Face*[6]. Devido a quantidade considerável de modelos disponíveis no site, para o presente trabalho foram feitos os testes de 85 desses modelos para o respectivo *dataset*. Após a obtenção dos resultados, o melhor modelo foi definido.

O último experimento considerado contou com a aplicação das *embeddings* produzidas pela IA Generativa da Google, denominada *Gemini* [McIntosh et al. 2023]. Tendo posse das *embeddings*, o modo de realização do cálculo de similaridade foi igual aos métodos anteriores, novamente fazendo o uso da similaridade por cosseno. Por fim, a etapa 3 consistiu na avaliação dos modelos e será discutido na próxima seção.

4. Resultados e discussões

Após realizar os experimentos, foram obtidos resultados que revelaram algumas características interessantes para o conjunto de dados. Como as notas atribuídas estavam em uma escala contínua de 0 a 100, então, para se fazer uma avaliação honesta, foi necessário aplicar uma métrica de regressão. Desse modo, a métrica escolhida foi a de Erro Médio Absoluto(do inglês, *Mean Absolute Error - MAE*), a qual permite verificar o quão distante, em média, a resposta do aluno se encontra da resposta do docente. A fórmula é descrita na figura seguinte:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - p_i|$$

Figura 2. Fórmula da métrica MAE

[4]https://radimrehurek.com/gensim/

[5]https://www.sbert.net/

[6]https://huggingface.co/sentence-transformers

em que:

- n é o número de amostras
- y é o valor observado para cada amostra
- p é o valor previsto pelo modelo para cada amostra
- $|\ |$ representa o valor absoluto

Para o modelo do *SpaCy*, os testes realizados para o *Dataset_*1 revelaram que não houve uma diferença significativa entre utilizar os dados em português ou em inglês, sendo que, em algumas situações, os resultados em português foram até mesmo melhores, o que deixa evidente que nem sempre utilizar o inglês irá implicar em um melhor desempenho do modelo. Ademais, a etapa de pré-processamento demonstrou ser um diferencial, uma vez que diminuiu o nível do erro calculado pela métrica, se comparado aos testes sem essa etapa. Os valores de *MAE* obtidos para todos os modelos testados podem ser conferidos através da tabela 1

Tabela 1. Avaliação dos modelos utilizando a métrica MAE

	MAE			
	Português		Inglês	
	Com pré-processamento	Sem pré-processamento	Com pré-processamento	Sem pré-processamento
spaCy	13.99	21.86	14.25	18.52
Gensim	16.06	16.29	16.73	16.31
Transformers	16.63	15.75	17.92	13.92
Gemini	13.90	15.05	14.61	14.17

Observando a tabela, é possível notar que o melhor resultado, para o modelo *SpaCy*, foi aquele que realizou o cálculo da similaridade considerando as respostas pré-processadas e em português, com uma margem de erro absoluta de aproximadamente 14 pontos, em uma escala de 0 a 100. Apesar de, no geral, haver uma concordância até que bem próxima, algumas situações mais específicas acabam distoando bastante, havendo uma diferença considerável entre a nota fornecida pelo professor e a nota gerada através do cálculo de similaridade.

Dando continuidade, para os testes utilizando o modelo da biblioteca *Gensim* e cálculos utilizando *WMD*, os números obtidos para o *corpus* foram piores do que os obtidos para o modelo do *SpaCy*, levando em consideração os cenários em que os dados foram pré processados, tanto para o portugês quanto para o inglês. No entanto, sem fazer o pré-processmento, os resultados foram melhores, indicando que, provavelmente, o conjunto de *embeddings* utilizado conseguiu capturar melhor as informações semânticas. Ademais, vale ressaltar que tais *embeddings*, para o caso em português, foram obtidas através dos modelos disponibilizados pelo Núcleo Interinstitucional de Linguística Computacional(NILC) da Universidade de São Paulo(USP), de modo que, considerando o teste em questão, o modelo utilizado foi o *word2vec - skipgram* com 100 dimensões[Hartmann et al. 2017]. Já para o inglês, as *embeddings* procediam de notícias do Google[7], contendo um número de 300 dimensões.

[7]https://huggingface.co/fse/word2vec-google-news-300

No que se refere ao modelo utilizando *Transformers*, os experimentos realizados permitiram observar resultados positivos, mas que apresentam muita dependência em relação ao tipo de *Transformers* escolhido. Assim como explicitado na seção 3, foram testados 85 diferentes tipos para o *Dataset*_1, sendo que, dependendo da língua e da realização ou não de pré-processamentos, os tipos podem variar. Levando isso em consideração, os tipos escolhidos foram *all-roberta-large-v1*[8] para a língua portuguesa, *stsb-roberta-base*[9] para o inglês, com pré-processamento dos dados, e *nli-roberta-base-v2*[10] para o inglês, sem o pré processamento.

Os resultados, de modo geral, não apresentaram uma diferença tão relevante, considerando os modelos anteriores. Nesse sentido, o experimento que obteve maior concordância foi o que considerou o conjunto de dados em inglês, e sem realizar a etapa de pré-processamento, apresentando um erro de 13,92 pontos, algo bem próximo ao resultado obtido para o *SpaCy*, utilizando o *dataset* em português e com pré-processamento.

Por fim, quanto ao modelo do *Gemini*, novamente os valores de erro foram bastante semelhantes aos já apresentados. Esse último modelo foi o que apresentou o melhor resultado, dentre todos os outros, com um erro de 13,90 para o *corpus* em português com o pré-processamento. Apesar da semelhança com os melhores resultados para o modelo do *SpaCy* e *Transformers*, o *Gemini* acaba apresentando um potencial maior, pois não apresenta a mesma limitação de tipo que a arquitetura *Transformers*, e o seu modelo continua em constante aprimoramento, de modo que a tendência é que aumente a qualidade das *embeddings* do modelo. Os gráficos da Figura 3 retratam os resultados obtidos para cada modelo.

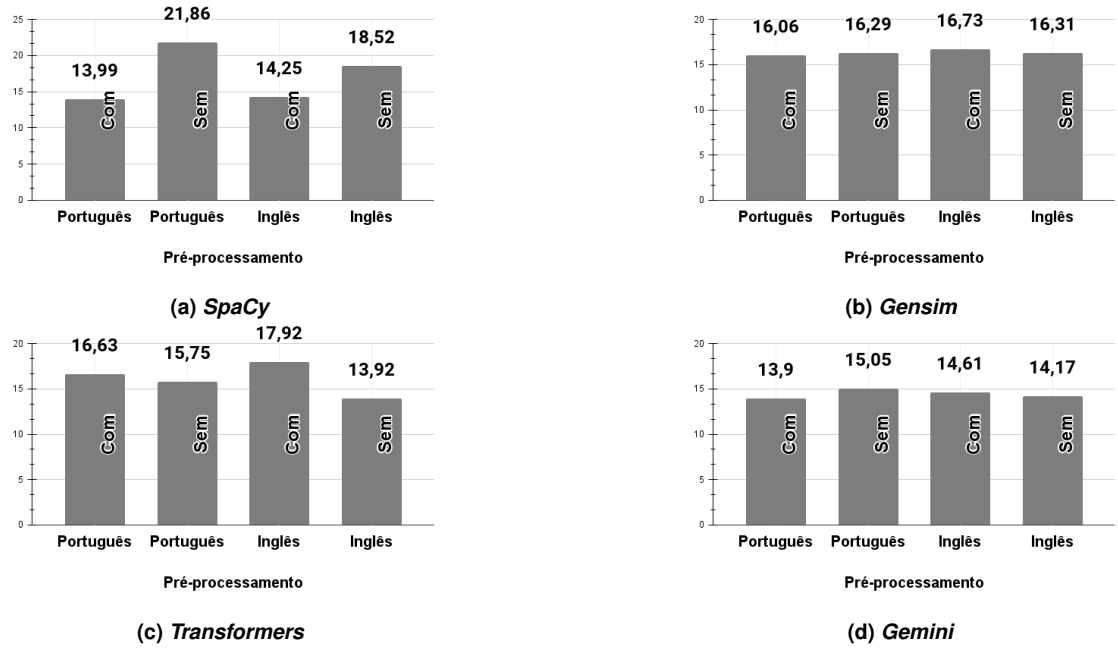

Figura 3. Graficos comparativos para cada modelo avaliado - Métrica MAE

[8]https://huggingface.co/sentence-transformers/all-roberta-large-v1
[9]https://huggingface.co/sentence-transformers/stsb-roberta-base
[10]https://huggingface.co/sentence-transformers/nli-roberta-base-v2

5. Conclusão

Este artigo explorou quatro diferentes modelos para verificação da similaridade textual no contexto de questões discursivas. Por meio dos experimentos, a análise crítica desses modelos revelou a complexidade inerente às particularidades presentes nas respostas discursivas, de modo que fica evidente a necessidade de se realizar aprimoramentos contínuos para atender às demandas crescentes de avaliação automatizada.

Os resultados obtidos evidenciaram a eficácia de alguns modelos, sendo que a escolha do método deve levar em consideração a diversidade semântica e a complexidade estrutural da resposta discursiva. Como conclusão, vale ressaltar os resultados obtidos através das embeddings provenientes da IA Generativa da *Google*, pois, se demonstrou o modelo mais promissor, obtendo erro de 13,9 considerando escala de 0 a 100.

Como trabalhos futuros, pretende-se investigar outros modelos de IA generativa, tais como o GPT e o Llama, além de investigar a aplicação dos métodos de similaridade semântica em outras áreas, por exemplo, letras e humanas, bem como em questões do ensino médio. Outrossim, deseja-se fazer os experimentos em ambientes reais, realizando integração com interface, de modo a permitir o desenvolvimento de um sistema avaliativo completo.

Referências

Amade, D., Chandra, R., Sinha, V., and Anand, D. (2024). Automatic text summarization using nltk spacy*. *SSRN Electronic Journal*.

Bao, H., Wang, Z. X., Cheng, X., Su, Z., Yang, Y.-H., Zhang, G.-Y., Wang, B., and Cai, H.-J. (2022). Using word embeddings to investigate human psychology: Methods and applications. *Xinli kexue jinzhan*, 31:887–887.

Burstein, J., Leacock, C., and Swartz, R. (2001). Automated evaluation of essays and short answers. *Proceedings of the 5th CAA Conference, Loughborough: Loughborough University*.

Dande, A. A. and Pund, D. M. A. (2023). A review study on applications of natural language processing. *International journal of scientific research in science, engineering and technology*.

de Oliveira, D., Pozzebon, E., and Santos, T. (2020). Aplicação das técnicas de processamento de linguagem natural cosine similarity e word movers distance para auxiliar na correção de questões discursivas em um tutor inteligente. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 1243–1252. SBC.

Galhardi, L., de Souza, R., and Brancher, J. (2020). Automatic grading of portuguese short answers using a machine learning approach. In *Anais Estendidos do XVI Simpósio Brasileiro de Sistemas de Informação*, pages 109–124. SBC.

GOMES, A. S. and PIMENTEL, E. P. (2021). Ambientes virtuais de aprendizagem para uma educação mediada por tecnologias digitais. *Informática na Educação: ambientes de aprendizagem, objetos de aprendizagem e empreendedorismo. Porto Alegre: Sociedade Brasileira de Computação*.

Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., and Aluisio, S. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025*.

Januzaj, Y. and Luma, A. (2022). Cosine similarity - a computing approach to match similarity between higher education programs and job market demands based on maximum number of common words. *International Journal of Emerging Technologies in Learning (ijet)*, pages 258–268.

McIntosh, T. R., Susnjak, T., Liu, T., Watters, P., and Halgamuge, M. N. (2023). From google gemini to openai q*(q-star): A survey of reshaping the generative artificial intelligence (ai) research landscape. *arXiv preprint arXiv:2312.10868*.

Mohler, M. and Mihalcea, R. (2009). Text-to-text semantic similarity for automatic short answer grading. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 567–575.

Savytska, L., Sübay, T., Vnukova, N., Bezugla, I., and Pyvovarov, V. (2022). Word2vec model analysis for semantic and morphologic similarities in turkish words. *CEUR-WS*.

Shields, J. A. E. (2022). Classroom assessment. In *International Encyclopedia of Education (Fourth Edition)*, pages 519–528. Elsevier eBooks.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Wang, M. and Hu, F. (2021). The application of nltk library for python natural language processing in corpus research. *Theory and Practice in Language Studies*.

Yamagiwa, H., Yokoi, S., and Shimodaira, H. (2022). Improving word mover's distance by leveraging self-attention matrix. *arXiv preprint arXiv:2211.06229*.