CENTRO UNIVERSITÁRIO FEI

LUCAS MATEUS DE MORAES - RA: 22.220.004-0

ORIENTADOR: PROF. DR. CHARLES HENRIQUE PORTO FERREIRA

APLICAÇÃO DE TÉCNICAS DE PROCESSAMENTO DE LINGUAGEM NATURAL PARA AVALIAÇÃO AUTOMÁTICA DE QUESTÕES DISSERTATIVAS

São Bernardo do Campo

2023

RESUMO

Este trabalho propõe uma abordagem para a avaliação automática de respostas dissertativas em ambientes educacionais. A métrica proposta pelo presente trabalho combina a presença de palavras-chave, a frequência de sentidos de palavras, a análise da sintaxe do texto e a distância semântica através do cosseno para oferecer uma solução que simplifica o processo de correção manual. Utilizando o *DataSet ASQA* do *Google* e dados de respostas do ENADE, a pesquisa busca validar a eficácia da técnica proposta considerando fatores relevantes para processamento de linguagem natural.

A proposta possui seus desafios de pesquisa, estudo e desenvolvimento, porém, não possui problemas significativos em relação a obter bases de dados ou lidar com dados sensíveis. Além de que, possui uma valor de contribuição valioso para a presente instituição de ensino e outras universidades, também sendo capaz de contribuir com o ensino de forma geral, pois, o uso de inteligência artificial (*AI*) e processamento de linguagem natural (*NLP*) aplicados a educação ainda é um tópico pouco explorado e de grande relevância, podendo assim, tornar a universidade e o presente trabalho uma possível referência no assunto, abrindo espaço para aplicações reais caso seu desenvolvimento futuro seja bem sucedido.

Palavras-chave: Semantic Similarity; Paraphrase Detection; Natural Language Processing; Evaluation of Descriptive Answers

ABSTRACT

This paper proposes an innovative approach for the automated evaluation of essay-type responses in educational environments. In contrast to the preference for objective questions, the proposed metric combines keyword presence, word sense frequency, syntax analysis, and semantic distance using cosine similarity to offer a solution that streamlines the manual correction process. Utilizing the Google's ASQA DataSet and data from ENADE's exam, the research aims to validate the effectiveness of the technique, considering relevant factors for natural language processing.

The proposal presents challenges in terms of research, study, and development; however, it does not encounter significant issues regarding obtaining databases or handling sensitive data. Moreover, it holds valuable contribution potential for the current educational institution and other universities. It is also capable of contributing to education in general, as the use of artificial intelligence (*AI*) and natural language processing (*NLP*) applied to education is still a relatively unexplored and highly relevant topic. Therefore, the university and the present work could potentially become a reference in the field, opening space for real applications if its future development is successful.

Keywords: Semantic Similarity; Paraphrase Detection; Natural Language Processing; Evaluation of Descriptive Answers

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Na atualidade das instituições de ensino, uma variedade de métodos de avaliação são empregados para mensurar o aprendizado dos alunos, destacando-se entre eles as questões dissertativas, nas quais os alunos devem fornecer suas respostas de maneira textual. Esse método proporciona ao professor uma compreensão mais aprofundada da linha de raciocínio do aluno durante a correção, possibilitando assim, uma avaliação mais precisa do nível de aprendizado alcançado e um melhor acompanhamento da evolução do aluno ao longo do tempo (OLIVEIRA; SANTOS, 2005).

Embora a precisão da avaliação seja uma vantagem desse método, o mesmo exige maior tempo de leitura, análise e compreensão de cada questão por parte do professor, uma vez que ele precisará analisar integralmente o conteúdo dos textos fornecidos como resposta pelos alunos. Tal dificuldade pode levar o docente a preferir questões objetivas, que, por possuírem um gabarito, demandam um tempo menor de correção. Entretanto, é importante ressaltar que as questões objetivas não atingem o mesmo nível de precisão na avaliação em comparação com as questões dissertativas, uma vez que possibilitam que o aluno escolha uma alternativa de maneira aleatória e possa obter a resposta correta mesmo sem ter nenhum conhecimento da mesma, ao contrário das questões dissertativas, em que essa possibilidade não existe.

A utilização de avaliações dissertativas é utilizado em apenas 30% das formas de avaliação aplicadas aos alunos, conforme demonstrado em um estudo realizado por (OLIVEIRA; SANTOS, 2005). Este estudo aborda as vantagens das avaliações dissertativas e sua maior adequação para a avaliação do desempenho e acompanhamento do progresso dos alunos no processo de aprendizado. Considerando esse contexto, seria benéfico para as instituições de ensino adotar mais frequentemente avaliações dissertativas. No entanto, como mencionado anteriormente, um dos principais impeditivos da adoção desse método seria o aumento na carga de trabalho dos professores responsáveis pelas correções. Dessa forma, o presente trabalho tem por objetivo propor uma abordagem para realizar avaliação automática das respostas dissertativas dos alunos, comparando-as com uma resposta padrão fornecida por um especialista da área como modelo do conhecimento esperado para aquela questão.

De forma geral, a abordagem proposta irá avaliar o "grau de similaridade" entre as respostas das questões levando em consideração propriedades estatística e semântica dos textos. Sendo assim, pretende-se propor uma métrica que possa mensurar a similaridade entre as respostas. Para que essa avaliação seja efetivamente realizada, uma série de fatores, os quais podemos considerar

como elementos relevantes na composição de uma resposta, devem ser considerados. Entre esses fatores pode-se citar: presença de palavas chaves, número de modificações para transformar uma resposta na outra, ordem das palavras, similaridade de texto, grau de parafraseamento, entre outros.

1.1 OBJETIVO

O objetivo final deste trabalho é o desenvolvimento e implementação de uma abordagem automatizada de avaliação para respostas dissertativas em ambientes educacionais. A proposta busca simplificar o processo de correção manual dessas respostas, promovendo uma análise precisa do desempenho dos alunos e reduzindo a carga de trabalho sobre os professores. As metas planejadas podem ser especificadas nos seguintes tópicos:

a) Desenvolver um algoritmo para mensurar a similaridade semântica entre respostas dissertativas e uma resposta padrão.
b) Considerar fatores como a presença de palavras-chave, quantidade de repetições de palavras, grau de parafraseamento, distância entre cossenos dos textos e ordem das palavras na avaliação automática.
c) Validar a eficácia da técnica por meio de estudos de caso e comparações com dados de avaliações já corrigidas.
d) Como última meta, caso as anteriores sejam alcançadas com sucesso, planejar a integração da técnica com plataformas educacionais existentes ou com um protótipo, testando seu funcionamento na prática, o que pode torná-la acessível e aplicável em ambientes reais no futuro.

1.2 QUESTÕES DE PESQUISA

O tema abordado levanta importantes questões de pesquisa, nas quais o presente trabalho buscará responder questões tais como:

a) Toda resposta correta pode ser considerada, em algum grau, uma paráfrase de uma resposta padrão?
b) A partir de qual ponto pode-se dizer que o grau de paráfrase entre os textos da resposta de um aluno e a resposta padrão indica corretude? Ou, em vez de uma saída binária, o resultado deve ser representado por uma escala variável?

c) Quais parâmetros podem ser considerados como componentes relevantes para pontuar a similaridade semântica entre as respostas?

d) Esses parâmetros podem definir uma pontuação que funcione de maneira geral quando aplicada a casos reais e práticos de correções dissertativas?

2 CONCEITOS

Nesta seção serão apresentados alguns conceitos fundamentais para o entendimento da porposta desse projeto. Serão abordados conceitos de processamento de lingugagem natural, formas de representação de texto em formato númerico e metricas para comparação de textos e avalialção de desempenho. Assim livros como *Introduction to natural language processing* (EISENSTEIN, 2019) e *Machine Learning: An Algorithmic Perspective* (MARSLAND, 2014), são importantes fontes para o aprofundamento nesses tópicos.

2.1 PROCESSAMENTO DE LINGUAGEM NATURAL

O Processamento de Linguagem Natural (PLN) refere-se à aplicação de técnicas computacionais para a interpretação e manipulação de linguagem humana. Envolve o desenvolvimento de algoritmos e modelos que capacitam computadores a compreender, analisar e gerar texto de maneira semelhante ao entendimento humano.

2.2 REPRESENTAÇÃO DE TEXTOS

A Representação de Textos é crucial para permitir que algoritmos compreendam palavras e documentos. Duas técnicas comuns são TFIDF (Term Frequency-Inverse Document Frequency) e Word Embedding. O TFIDF avalia a importância de uma palavra em um documento, enquanto o Word Embedding mapeia palavras em vetores contínuos, capturando relações semânticas.

2.2.1 Word embedding

O Word Embedding é uma técnica que mapeia palavras em vetores de números reais, capturando relações semânticas e contextuais. Essa representação densa permite que algoritmos de processamento de linguagem natural compreendam a similaridade e a semântica entre palavras. Considere as palavras "rei"e "rainha."Se estiverem bem representadas por embeddings, a subtração dos vetores "rei"e "homem"deve ser aproximadamente igual à subtração dos vetores "rainha"e "mulher,"refletindo a relação semântica de gênero.

2.3 SIMILARIDADE DE TEXTOS

A Similaridade de Textos é fundamental para comparar documentos ou palavras. Diversas métricas são empregadas, como Palavras-Chave, Frequência de Palavras, Distância de Jaccard, Distância de Cosseno, Distância Euclidiana e Modelos de Linguagem.

2.3.1 Palavras-chave

A similaridade pode ser avaliada considerando as palavras-chave mais relevantes em documentos. A sobreposição ou relevância compartilhada entre essas palavras indica o grau de similaridade.Considere dois documentos sobre inteligência artificial. Se ambos compartilharem palavras-chave como "aprendizado de máquina", "algoritmos"e "processamento de linguagem natural,"é provável que sejam semanticamente similares.

2.3.2 Frequência de Sentidos de Palavras

A frequência de sentidos de lalavras (SFD) é uma métrica que avalia a distribuição de frequência dos diferentes sentidos ou significados associados a uma palavra ao longo de um conjunto de documentos. Em outras palavras, visa entender como a polissemia (múltiplos significados) de uma palavra se distribui em contextos específicos. Essa métrica é relevante para a detecção de mudanças semânticas em textos ao longo do tempo. Considerando a palavra "bateria"por exemplo que pode ser um instrumento musical ou um dispositivo eletrônico para armazenar energia elétrica. Se, ao longo do tempo, a frequência de uso de "bateria"em contextos relacionados a música diminuir enquanto o uso em contextos de armazenamento de energia aumentar, a SFD refletirá essa mudança semântica

2.3.3 Distância de Jaccard

A Distância de Jaccard avalia a similaridade entre conjuntos, medindo a proporção de elementos comuns entre dois conjuntos. Para texto, representa a sobreposição de palavras entre dois documentos. Considere dois conjuntos de palavras em dois documentos. Se o Conjunto A contiver as palavras {a, b, c} e o Conjunto B as palavras {b, c, d}, a Distância de Jaccard seria de 50% de similaridade, como na demonstrado na Equação 1.

$$\frac{|A \cap B|}{|A \cup B|} = \frac{2}{4} = 0.5 \tag{1}$$

2.3.4 Distância de Cosseno

A Distância de Cosseno mede o ângulo entre dois vetores de palavras, representando a similaridade direcional entre documentos. Quanto menor o ângulo, maior a similaridade. Considere dois vetores de palavras representando documentos. Se esses vetores apontarem na mesma direção, a distância de cosseno será próxima de zero, indicando alta similaridade. Se apontarem em direções opostas, a distância será próxima de 1, indicando baixa similaridade.

2.3.5 Distância Euclidiana

A Distância Euclidiana calcula a distância geométrica entre pontos em um espaço vetorial. Em texto, representa a dissimilaridade entre as distribuições de palavras. Considere dois documentos representados como pontos em um espaço vetorial. Se os pontos (representando os documentos) estiverem próximos no espaço, a distância euclidiana será pequena, indicando alta similaridade. Se estiverem distantes, a distância será grande, indicando baixa similaridade.

2.3.6 Distância Jensen-Shannon

A Distância Jensen-Shannon (JSD) é uma medida de divergência estatística entre duas distribuições de probabilidade que é utilizada em processamento de linguagem natural para avaliar a similaridade entre textos com base na distribuição de frequência das palavras. O cálculo da JSD envolve a criação de uma distribuição média ponderada e o uso da entropia de Kullback-Leibler. Ela considera não apenas a presença ou ausência de palavras, mas também a probabilidade de ocorrência dessas palavras nos textos. Quanto menor a distância obtida, maior é a similaridade semântica entre os textos.

2.3.7 Language Models

Os Modelos de Linguagem, como os de Parafraseamento, buscam entender a similaridade semântica entre frases ou documentos, indo além da análise baseada em palavras. Supondo que um modelo de linguagem deve prever palavras em frases. Na frase "O gato está na", o modelo de linguagem pode prever as palavras "casa", "árvore"e rua por exemplo. Com base em um conjunto de dados de treinamento a probabilidade da palavra "casa"pode ser maior.

2.3.7.1 Parafraseamento

Modelos de Parafraseamento são específicos para avaliar a similaridade entre frases ou documentos que expressam a mesma ideia de maneira diferente. Esses modelos buscam capturar nuances semânticas e estruturais, identificando relações de equivalência entre diferentes formulações. As frases "O clima estava agradável para um passeio no parque"e "O parque oferecia um ambiente agradável para passeios"devem ser reconhecidas por um modelo de parafraseamento já que ambas as frases têm uma intenção semelhante, apesar das diferenças na escrita

2.4 MÉTRICAS DE AVALIAÇÃO

As Métricas de Avaliação quantificam o desempenho de modelos de processamento de linguagem natural. A acurácia é uma medida fundamental, representando a proporção de predições corretas em relação ao total. Outras métricas, como precisão, revocação e F1-Score, oferecem insights adicionais sobre o desempenho do modelo em diferentes aspectos da classificação ou similaridade.

2.4.1 Acurácia

A acurácia é uma métrica fundamental de avaliação, comumente usada para medir o desempenho geral de modelos de processamento de linguagem natural. Representa a proporção de predições corretas em relação ao total de predições. Embora seja uma medida direta, a Acurácia pode ser limitada em cenários desbalanceados, sendo complementada por métricas adicionais, como precisão, revocação e F1-Score, para avaliação mais abrangente do desempenho do modelo. No entanto, em cenários desbalanceados, a acurácia pode ser limitada. Para avaliação mais abrangente, métricas adicionais, como BLEU (Bilingual Evaluation Understudy) e ROUGE (Recall-Oriented Understudy for Gisting Evaluation), podem ser utilizadas.

2.4.1.1 BLEU (Bilingual Evaluation Understudy)

O BLEU é comumente usado para avaliar a qualidade de traduções automáticas em tarefas de processamento de linguagem natural. Ele calcula a sobreposição de palavras entre a tradução gerada pelo modelo e a tradução de referência. Quanto mais sobreposição, maior é o escore BLEU.

2.4.1.2 ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

O ROUGE é empregado para avaliar a qualidade de resumos automáticos, focando na recordação (revocação) das palavras-chave. Ele mede a sobreposição de n-gramas (sequências contínuas de n palavras) entre o resumo gerado e o resumo de referência. Maior sobreposição resulta em um escore ROUGE mais alto.

3 REVISÃO BIBLIOGRÁFICA

Para realização da revisão bibliográfica foram utilizados as ferramentas de buscas de artigos científicos do *Google Scholar* (https://scholar.google.com/), *IEEE Xplore* (https://ieeexplore.ieee.org/Xplore/home.jsp), *ScienceDirect* (https://www.sciencedirect.com/) e *Semantic Scholar* (https://www.semanticscholar.org/).

Como palavras-chaves na busca foram utilizados termos em inglês, sendo eles, *"semantic similarity between texts"*, *"measure degree of paraphrase"*, *"paraphrase detection"*, *"natural language processing"*, *"measure semantic similarity between answers"* e *"evaluation of descriptve answers"*. Os termos que trouxeram os melhores resultados e estavam presentes nos melhores artigos selecionados foram *"semantic similarity"*, *"natural language processing"* e *"evaluation of descriptve answers"*.

Inicialmente foram selecionados 26 artigos que poderiam ser relevantes para o presente trabalho com base nos temas, a sumarização dos artigos pode ser vista na Tabela 1 em que os artigos selecionados no final estão destacados na cor cinza.

Título	Fonte	Referência
Determining Degree of Relevance of Reviews Using a Graph-Based Text Representation	IEEE	(RAMACHANDRAN; GEHRINGER, 2011)
A Chinese text paraphrase detection method based on dependency tree	IEEE	(JIANG; HAO; ZHU, 2016)
Nowhere to Hide: Finding Plagiarized Documents Based on Sentence Similarity	IEEE	(GUSTAFSON; PERA; NG, 2008)
Enhanced Text Matching Based on Semantic Transformation	IEEE	(ZHANG et al., 2020)
Semantic similarity based assessment of descriptive type answers	IEEE	(MEENA; LAWRANCE, 2016)
A comparative analysis of various approaches for automated assessment of descriptive answers	IEEE	(KAUR; SASIKUMAR, 2017)
A reliable approach to automatic assessment of short answer free responses	SSL	(BACHMAN et al., 2002)
A Descriptive Answer Evaluation System Using Cosine Similarity Technique	IEEE	(THALOR, 2021)
An Intelligent System for Evaluation of Descriptive Answers	IEEE	(BAGARIA et al., 2020)
Application Research of Similarity Algorithm in the Design of English Intelligent Question Answering System	IEEE	(ZHANG, 2022)
Near duplicate text detection using graph depiction	IEEE	(POULOS, 2016)
Recognition of Parallelism Sentence Based on Recurrent Neural Network	IEEE	(DAI et al., 2018)
LSGC: An Interactive Text Matching Model Combined with Enhanced Encoding	IEEE	(WANG et al., 2022)
A software system for determining the semantic similarity of short texts in Serbian	IEEE	(BATANOVIĆ; FURLAN; NIKOLIĆ, 2011)
Arabic Semantic Textual Similarity Identification based on Convolutional Gated Recurrent Units	IEEE	(MAHMOUD; ZRIGUI, 2021)
A Chinese text paraphrase detection method based on dependency tree	IEEE	(JIANG; HAO; ZHU, 2016)
Using paraphrases to improve tweet classification: Comparing WordNet and word embedding approaches	IEEE	(LI et al., 2016)
SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection	SSL	(SCHLECHTWEG et al., 2020)
SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation	SSL	(CER et al., 2017)
Use of Syntactic Similarity Based Similarity Matrix for Evaluating Descriptive Answer	IEEE	(PAUL; PAWAR, 2014)
Chapter 16 - Semantic similarity–based descriptive answer evaluation	SCD	(SHAUKAT et al., 2021)
A Study of Automated Evaluation of Student's Examination Paper using Machine Learning Techniques	IEEE	(SANUVALA; FATIMA, 2021)
Online Examination with short text matching	IEEE	(KUDI et al., 2014)
Towards Automated Evaluation of Handwritten Assessments	IEEE	(ROWTULA; OOTA; C.V, 2019)
Automatic Short Answer Grading (ASAG) using Attention-Based Deep Learning MODEL	IEEE	(AMUR; HOOI; SOOMRO, 2022)
Semantic similarity–based descriptive answer evaluation	SSL	(SHAUKAT et al., 2021)

Tabela 1 – Tabela de revisão bibliográfica sumarizada.

Após uma filtragem com base nos resumos e palavras-chaves, foram selecionados quatro artigos como base para a revisão bibliográfica como consta na Tabela 2, considerando sua proximidade com a proposta deste trabalho para leitura integral de seu conteúdo.

Base	Total encontrados	Após remoção dos Duplicados	Após análise do resumo
IEEE	21	21	2
Science Direct	1	1	1
Semantic Scholar	4	3	1

Tabela 2 – Tabela do funil de leitura.

O artigo proposto por (SCHLECHTWEG et al., 2020), faz o uso de embeddings de tipo (type embeddings) e embeddings contextualizados (token embeddings) para representar as palavras. Primeiro, é introduzida a distribuição de frequência de sentido (SFD), e a detecção de mudança binária é definida em termos de limiares de frequência. Após isso, a distância de Jensen-Shannon (JSD) entre as distribuições normalizadas de frequência é utilizada para medir a mudança efetuada.

No artigo proposto por (PAUL; PAWAR, 2014), podemos destacar que o trabalho utiliza a técnica de Análise Semântica Latente (LSA), que é comumente usada para determinar a similaridade de documentos, mas ressalta suas limitações em documentos curtos. O artigo destaca a ausência de abordagens anteriores que se concentrem na avaliação automática de respostas descritivas usando vetores de ordem de palavras. O método proposto utiliza uma matriz de similaridade entre vetores de ordem de palavras para avaliar respostas descritivas. A similaridade entre os vetores é calculada por meio de uma métrica de similaridade sintática, ou seja baseada na ordem das palavras. Os resultados indicam que a abordagem baseada em ordem de palavras é promissora para a avaliação automática de respostas descritivas. A matriz de similaridade é apresentada como uma ferramenta eficaz para computar as notas de cada pergunta.

Na proposta do artigo escrito pelos autores (SHAUKAT et al., 2021), pode-se destacar que a pesquisa faz uso de Processamento de Linguagem Natural (NLP) para automatizar o processo de avaliação, especialmente a similaridade de cosseno e índices de similaridade, são empregadas para atribuir notas às respostas.

Os autores (SANUVALA; FATIMA, 2021) incorporam uma abordagem que emprega ferramentas de Reconhecimento Óptico de Caracteres (OCR) para extrair texto de respostas manuscritas digitalizadas. A ênfase principal, no entanto, recai sobre o emprego de técnicas avançadas de processamento de linguagem natural para aprimorar a avaliação. O estudo destaca a importância de etapas como a tokenização, remoção de stop words e verificação de sinônimos e antônimos no pré-processamento das respostas. Além disso, aborda a criação de modelos semânticos e o cálculo de similaridade sem mencionar explicitamente as métricas de Machine Learning utilizadas.

No contexto da revisão bibliográfica, alguns conceitos importantes foram retirados dos artigos selecionados, para contribuir com o presente trabalho. Dentre esses, destacam-se a análise da similaridade semântica com cossenos, a consideração da frequência de Sentidos de Palavras (SFD) e a utilização de vetores de ordem de palavras como elementos-chave.

A análise de similaridade semântica com cossenos é uma técnica importante, conforme evidenciado nos artigos revisados. Essa abordagem é frequentemente empregada para medir a proximidade semântica entre textos.

O uso da Frequência de Sentidos de Palavras (SFD) em um dos artigos indica a importância específica da distribuição de frequência das palavras. Esse conceito pode contribuir para o trabalho, sendo relevante para aspectos da sintaxe e da semântica do texto.

A utilização de vetores de ordem de palavras, como abordado em um dos artigos, resalta a importância da ordem das palavras na avaliação de respostas descritivas. Essa técnica pode superar limitações associadas à Análise Semântica Latente (LSA) em documentos curtos, oferecendo uma abordagem promissora para a avaliação automática.

Em síntese, a revisão bibliográfica ressalta a importância de conceitos como a análise de similaridade semântica com cossenos, a distribuição de frequência de sentidos de palavras e da exploração de vetores de ordem de palavras como pontos relevantes nos estudos revisados.

4 METODOLOGIA

4.1 ABORDAGEM DE PESQUISA:

Levando em conta que o presente trabalho possui como objetivo realizar avaliação automática de respostas dissertativas, pretende-se que o resultado final do mesmo seja uma proposta de resolução desse problema. Assim, podemos classificar sua abordagem de metodologia como experimental, haja vista que serão realizados experimentos com os dados e as técnicas, que serão descritas no decorrer do texto, afim de atingir o objetivo proposto.

4.2 BASES DE DADOS:

Como fonte principal de dados, utilizaremos o *DataSet ASQA (Answer Summaries for Questions which are Ambiguous)* da *Google Research* em parceria com as universidade de *Duke University* e *Cornell University*, o *DataSet* contém respostas dissertativas para questões com perguntas ambíguas, isso permitirá a construção dos experimentos e refino das técnicas que serão utilizadas no presente trabalho (STELMAKH et al., 2022).

Além da fonte principal de dados, planejamos também utilizar dados das respostas dissertativas dos alunos de Ciência da Computação do Centro Universitário FEI para questões dissertativas do ENADE.

Ademais, existe a possibilidade de que um futuro questionário seja confeccionado para que, com a ajuda voluntária dos alunos da mesma instituição, mais dados sejam coletados e usados no presente trabalho. Essa possibilidade será investigada caso seja necessário.

4.3 MATERIAIS:

Serão utilizados computadores para o desenvolvimento da experimentação e a documentação do processo. Os computadores principais utilizados serão de posse própria, caso seja necessário, também podem ser utilizados computadores disponibilizados no campus do Centro Universitário FEI.

Se no decorrer do processo de experimentação, dificuldades em relação as capacidades de processamento e memórias dos computadores forem identificados, abre-se também a possibilidade do uso de plataformas disponibilizadas *online*, como o *Google Colab* (https://colab.google/)

e a *Weights & Biases* (https://wandb.ai/site/), ambas oferecem possibilidade de uso gratuito para estudantes que utilizarem apenas com fins acadêmicos.

4.4 MÉTODOS:

Serão utilizados métodos de NLP, tais como mineração de texto, medição de distância de cossenos (para mensurar similaridade de textos), frequência de sentidos de palavras (SFD), presença de palavras-chaves (KW), ordem de sequência das palavras, entre outros. Para compor toda a metodologia, pretende-se elaborar uma métrica de avaliação do grau de similaridade semântica entre dois conteúdos.

Assumindo que dois textos possuem exatamento o mesmo conteúdo, sua similaridade poderia ser descrita em uma escala percentual como 100%, por outro lado, haverão textos que possuirão conteúdo sintaticamente diferentes, logo, espera-se que seu percentual de similaridade diminua, porém, ainda assim seria possivel a existência de similaridade semântica entre esse texto e a respota correta, sendo assim, é importante avaliar aspectos semânticos do texto.

Desta forma, a hipótese é que, para avaliarmos tal similaridade, será necessário avaliar uma série de fatores importantes que irão compor a métrica de avaliação proposta. Esses fatores seriam:

a) **As propriedades da sintaxe do texto:** Tal fator será levado em consideração tendo em vista que os trabalhos avaliados na revisão bibliográfica demonstraram a importância de propriedades sintáticas como a ordem em que as sequências de palavras estão posicionadas no texto.

b) **Frequência de Sentidos de Palavras (SFD):** Pois também foi demonstrado nos trabalhos avaliados na revisão bibliográfica que essa medida estatística é relevante para a avaliação das respostas.

c) **A medição da distância de cossenos:** Essa medida tem relevância para a métrica pois mensura a similaridade semântica entre dois textos através de suas representações vetoriais.

d) **A presença de palavras-chave (*keywords*):** Serão palavras definidas para cada questão pelo seu confeccionador como termos com um grau mais elevado de relevância para a resposta esperada do aluno. determinadas como parte da resposta padrão para aquela questão.

Todos os fatores apresentados acima já foram utilizados de maneira isolada por diferentes trabalhos como uma forma de avaliação. No caso do presente trabalho pretende-se realizar

uma junção dos fatores em uma métrica única para avaliação, averiguando-se ao longo do desenvolvimento do trabalho qual a relevância de cada um deles, para que um peso ajustado seja atribuído para cada fator da métrica.

Caso a métrica proposta acima apresente bons resultados, pode-se cogitar a elaboração de uma segunda métrica, em que dois outros fatores seriam adicionados como uma possível melhoria à primeira, sendo eles:

a) A verificação da possibilidade de um texto ser classificado como paráfrase de outro, já que o trabalho possui como hipótese que toda resposta correta é em algum grau uma paráfrase da resposta modelo. Com a identificação da paráfrase, como pode ser feita a medição do grau de parafraseamento entre os dois textos.

b) A quantidade de modificações necessárias para que a resposta fornecida seja transformada lexicalmente e sintaticamente na resposta modelo.

Esses dois fatores não foram considerados nos trabalhos revisados anteriormente, na eventual proposta de uma segunda métrica, ter-se-á como hipótese que os dois fatores resultarão em uma melhoria no resultado.

A proposta da métrica pode ser matemáticamente descrita como na Equação 2.

$$Métrica1 = \frac{fator1 \times peso_1 + fator2 \times peso_2 + fator3 \times peso_3 + fator4 \times peso_4}{\sum_{i=1}^{4} peso_i} \quad (2)$$

Em que *fator1* são as propriedades sintáticas do texto, o *fator2* representa o SFD, o *fator3* é a distância de cossenos e o *fator4* é referente as *keywords*.

Como resultado a métrica deverá fornecer um número no intervalo de 0 100 que, quanto maior, indica maior proximidade semântica entre os textos e uma maior probabilidade de acerto da resposta fornecida por um aluno.

No caso da segunda métrica, o diferencial seria adição dos dois outros fatores para mensurar o grau de parafraseamento e a quantidade de modificações, sendo respectivamente o *fator5* e o *fator6* como descrito na Equação 3.

$$Métrica2 = \frac{fator1 \times peso_1 + fator2 \times peso_2 + fator3 \times peso_3 + fator4 \times peso_4 + fator5 \times peso_5 + fator6 \times peso_6}{\sum_{i=1}^{6} peso_i} \quad (3)$$

4.5 AVALIAÇÃO:

Para avaliação do desempenho das técnicas propostas, as bases de dados serão divididas em duas, uma primeiramente para treinamento e refino dos experimentos. Após os resultados

estarem satisfatórios, a segunda parte dos dados será utilizada para a validação dos resultados obtidos pelo algoritmo, verificando se estão condizente com os resultados fornecidos pelas avaliações registradas nos dados e se seu funcionamento é geral ou se há algum viés nos resultados. Uma representação visual das etapas gerais imaginadas para o processo pode ser vista na Figura 1.

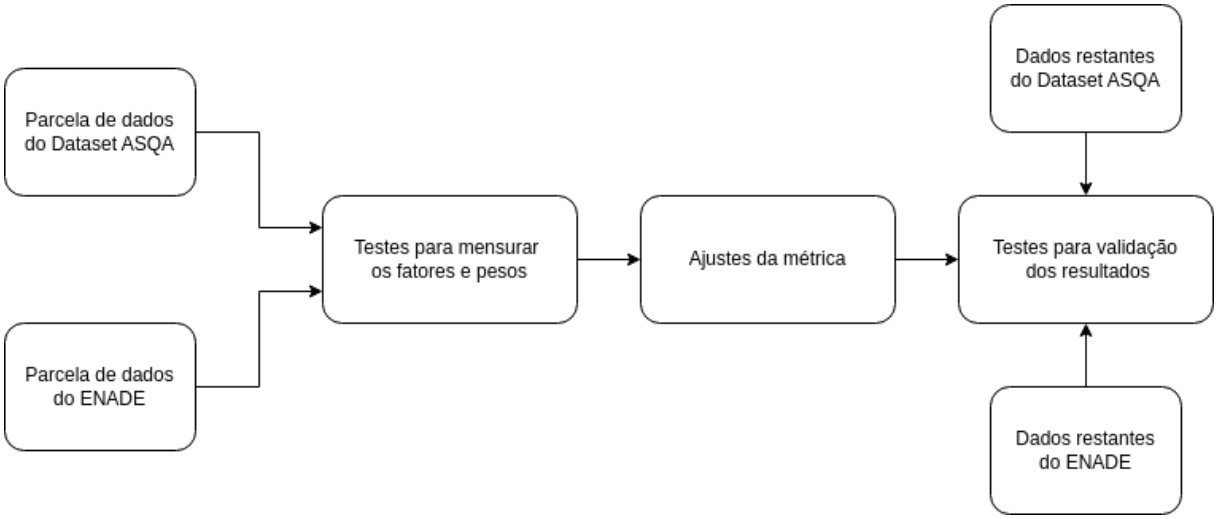

Figura 1 – Diagrama geral das etapas do processo.

5 PROPOSTA EXPERIMENTAL

Como proposta experimental, será realizado a avalição das questões dissertativas através de um algoritmo que leva em consideração a métrica de pontuação proposta nos capitulos anteriores do texto, tanto com seus fatores em conjunto, quanto com eles individualmente, o resultado das avaliações realizadas pelo algoritmo será comparado aos valores de avaliação dos professores que estão presentes nas bases de dados específicadas anteriormente no capítulo de metodologia.

A proposta experimental pode ser dividida nos seguintes tópicos:

a) **Avaliação individual da presença de palavras chaves:** O algoritmo proposto deve avaliar se as palavras chaves determinadas para aquela questão, estão presentes no texto e pontuar com base na quantidade de palavras-chave encontradas. Tendo como hipótese que essa avaliação individualemte não fornecerá um resultado com alta precisão.

b) **Avalição individual da frequência de palavras (SFD):** O algoritmo realizará a medição da frequência das palavras no texto da resposta padrão e da resposta fornecida pelo aluno, cada palavra terá sua frequência em ambos os textos e a diferença entre as duas será medida em porcentagem. Caso palavras diferentes apareçam no texto da respota do aluno, podemos levar em consideração inicialmente a possibilidade de descartá-las, outras possibilidades futuras também podem ser consideradas caso melhores abordagens sejam encontradas na literatura científica.

c) **Avalição individual da sintaxe do texto:** Utilizando vetores de ordem de palavras o algoritmo deve realizar a medição das métricas de sintaxe do texto, primeiramente da resposta padrão, em sequência, das respotas dos alunos. O vetor de ordem de palavras da respota padrão será comparado com os vetores de ordem de palavras que representem as respostas dos alunos.

d) **Avaliçao semantica através da distância de cossenos:** Na avaliação da distância de cossenos os textos serão representados em vetores de palavras que representam a similaridade direcional, não a ordem como no item anterior, conforme os ângulos calculado entre os vetores diminui, sabemos que o grau de similaridade entre os textos aumenta.

Além das avaliações individuais definidas acima, também serão avaliados os fatores em diferentes conjuntos para que os resultado obtidos através dessas combinções sejam comparados

com a junção de todos os quatro fatores em uma só métrica, tendo como hipótese que as métricas com um conjunto de fatores combinados funcionarão melhor do que um fator individualmente e que a métrica de todos os fatores juntos terá o melhor resultado.

Em todas as métricas combinadas, os diferentes fatores devem possuir pesos que representem o quão relevante cada elemento é para uma avalição da resposta. Com base nas avaliações individuais realizadas inicialmente, serão obtido quais dos fatores possuem uma proximidade maior com os valores das avalições dos professores disponíveis nas bases de dados, podemos, inicialmente, propor que quanto maior a proximidade dos valores individuais obtidos com as avaliações nas bases de dados, mais relevantes esse fator deve ser em seu peso.

Com o objetivo de obter a melhor precisão possível, os pesos inicias serão modificados para valores com maior ou menor relevância, afim de que o resultado final obtido seja o mais próximo possível dos resultados das avaliações feitas pelos professores disponíveis nas bases de dados. Além disso, métricas como BLEU e ROUGE podem ser empregadas para avaliar a qualidade da pontuação em termos de sobreposição de palavras entre as respostas.

A proposta experimental busca não apenas avaliar a viabilidade do algoritmo, mas também otimizar sua precisão por meio da ponderação adequada dos fatores. A abordagem modular, combinada com a análise iterativa dos pesos, visa criar um sistema capaz de fornecer resultados de avaliação de respostas dissertativas com alta concordância em relação às avaliações humanas.

5.1 CRONOGRAMA

Na Tabela 3 está estruturado o cronograma proposta para as fases que já concluídas e as que deverão ser concluídas afim de alcaçar os objetivos propostos.

	2023						2024					
Atividade / Mês	07	08	09	10	11	12	01	02	03	04	05	06
Definição do tema e orientador	■	■										
Revisão bibliográfica			■	■								
Planejamento da metodologia				■	■							
Finalização do relatório parcial						■						
Aprofundamento nos conceitos de NLP							■					
Experimentação do algoritmo para avaliação da relevância dos fatores								■				
Experimentação para aprimoramento dos valores dos pesos na métrica									■			
Validação dos resultados obtidos com os dados de avaliações de professores										■		
Adição dos fatores de parafraseamento para implementar a segunda métrica											■	
Ajustes finais para refinamento do algoritmo e finalização do projeto												■

Tabela 3 – Cronograma de execução do projeto

6 CONCLUSÃO

Diante da diversidade de métodos de avaliação empregados nas instituições de ensino contemporâneas, as questões dissertativas destacam-se por proporcionarem uma compreensão mais profunda da linha de raciocínio dos alunos. Embora a precisão seja uma vantagem desse método, sua efetividade muitas vezes é prejudicada pelo tempo exigido para leitura e análise por parte dos professores. A preferência por questões objetivas, embora economize tempo, diminui a precisão da avaliação.

Este trabalho propõe uma abordagem para amenizar esse problema, buscando automatizar o processo de avaliação de respostas dissertativas. A métrica proposta leva em consideração a presença de palavras-chave, a frequência de sentidos de palavras, a sintaxe do texto e a distância semântica através do cosseno, visando simplificar o processo de correção manual, possibilitando analisar o desempenho dos alunos.

Ao utilizar o DataSet ASQA e dados de respostas para questões do ENADE, este trabalho planeja realizar experimentos para validar a eficácia da abordagem proposta. Avaliando quais fatores seriam relevantes para compor a métrica de avaliação. Esta pesquisa busca não apenas simplificar o processo de avaliação, mas também abrir a possibilidade de uma integração futura entre plataformas educacionais e a técnica proposta, eventualmente, proporcionando uma solução prática para a área da educação.

As questões de pesquisa levantadas durante a revisão bibliográfica e estabelecidas nos objetivos, relacionadas à paráfrase, grau de similaridade semântica e parâmetros relevantes para pontuação, serão abordadas na fase experimental, visando contribuir não apenas com avanços teóricos, mas também com aplicações práticas e melhorias efetivas na avaliação automática de respostas dissertativas.

Em última análise, este trabalho almeja não apenas uma solução para as limitações atuais na avaliação de respostas dissertativas, mas também contribuir na determinação de quais parâmetros são relevantes em uma métrica para esse problema de pesquisa. Ademais, o resultado prático do presente trabalho, caso os objetivos sejam alcançados, deverá oferecer uma contribuição significativa os processos educacionais, proporcionando aos professores uma ferramenta valiosa para avaliação e acompanhamento do progresso dos alunos.

REFERÊNCIAS

AMUR, Zaira Hassan; HOOI, Yew Kwang; SOOMRO, Gul Muhammad. Automatic Short Answer Grading (ASAG) using Attention-Based Deep Learning MODEL, p. 1–7, 2022. DOI: 10.1109/ICDI57181.2022.10007187. Disponível em: https://ieeexplore.ieee.org/document/10007187.

BACHMAN, Lyle F. et al. A Reliable Approach to Automatic Assessment of Short Answer Free Responses. In: INTERNATIONAL Conference on Computational Linguistics. [S.l.: s.n.], 2002. Disponível em: https://api.semanticscholar.org/CorpusID:27889503.

BAGARIA, Vinal et al. An Intelligent System for Evaluation of Descriptive Answers, p. 19–24, 2020. DOI: 10.1109/ICISS49785.2020.9316110. Disponível em: https://ieeexplore.ieee.org/document/9316110.

BATANOVIĆ, Vuk; FURLAN, Bojan; NIKOLIĆ, Boško. A software system for determining the semantic similarity of short texts in Serbian, p. 1249–1252, 2011. DOI: 10.1109/TELFOR.2011.6143778. Disponível em: https://ieeexplore.ieee.org/document/6143778.

CER, Daniel Matthew et al. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation, 2017. Disponível em: https://api.semanticscholar.org/CorpusID:4421747.

DAI, Yange et al. Recognition of Parallelism Sentence Based on Recurrent Neural Network, p. 148–151, 2018. DOI: 10.1109/ICSESS.2018.8663734. Disponível em: https://ieeexplore.ieee.org/document/8663734.

EISENSTEIN, Jacob. **Introduction to natural language processing**. [S.l.]: MIT press, 2019.

GUSTAFSON, Nathaniel; PERA, Maria Soledad; NG, Yiu-Kai. Nowhere to Hide: Finding Plagiarized Documents Based on Sentence Similarity. v. 1, p. 690–696, 2008. DOI: 10.1109/WIIAT.2008.16. Disponível em: https://ieeexplore.ieee.org/document/4740531.

JIANG, Yipeng; HAO, Yu; ZHU, Xiaoyan. A Chinese text paraphrase detection method based on dependency tree, p. 1–5, 2016. DOI: 10.1109/ICNSC.2016.7479003. Disponível em: https://ieeexplore.ieee.org/document/7479003.

KAUR, Amarjeet; SASIKUMAR, M. A comparative analysis of various approaches for automated assessment of descriptive answers. In: 2017 International Conference on Computational Intelligence in Data Science(ICCIDS). [S.l.: s.n.], 2017. P. 1–7. DOI: 10.1109/ICCIDS.2017.8272650. Disponível em: https://ieeexplore.ieee.org/document/8272650.

KUDI, Pooja et al. Online Examination with short text matching, p. 56–60, 2014. DOI: 10.1109/GCWCN.2014.7030847. Disponível em: https://ieeexplore.ieee.org/document/7030847.

LI, Quanzhi et al. Using paraphrases to improve tweet classification: Comparing WordNet and word embedding approaches, p. 4014–4016, 2016. DOI: 10.1109/BigData.2016.7841094. Disponível em: https://ieeexplore.ieee.org/document/7841094.

MAHMOUD, Adnen; ZRIGUI, Mounir. Arabic Semantic Textual Similarity Identification based on Convolutional Gated Recurrent Units, p. 1–7, 2021. DOI: 10.1109/INISTA52262.2021.9548576. Disponível em: https://ieeexplore.ieee.org/document/9548576.

MARSLAND, Stephen. **Machine Learning: An Algorithmic Perspective, Second Edition**. 2nd. [S.l.]: Chapman & Hall/CRC, 2014. ISBN 1466583282.

MEENA, K; LAWRANCE, R. Semantic similarity based assessment of descriptive type answers. In: 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16). [S.l.: s.n.], 2016. P. 1–7. DOI: 10.1109/ICCTIDE.2016.7725366. Disponível em: https://ieeexplore.ieee.org/document/7725366.

OLIVEIRA, Katya Luciane de; SANTOS, Acácia Aparecida Angeli dos Santos. Avaliação da aprendizagem na universidade. pt. **Psicologia Escolar e Educacional**, scielopepsic, v. 9, p. 37–46, jun. 2005. ISSN 1413-8557. Disponível em: http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1413-85572005000100004&nrm=iso.

PAUL, Dimple V.; PAWAR, Jyoti D. Use of Syntactic Similarity Based Similarity Matrix for Evaluating Descriptive Answer, p. 253–256, 2014. DOI: 10.1109/T4E.2014.60. Disponível em: https://ieeexplore.ieee.org/document/7009583.

POULOS, Marios. Near duplicate text detection using graph depiction, p. 1–6, 2016. DOI: 10.1109/IISA.2016.7785368. Disponível em: https://ieeexplore.ieee.org/document/7785368.

RAMACHANDRAN, Lakshmi; GEHRINGER, Edward F. Determining Degree of Relevance of Reviews Using a Graph-Based Text Representation, p. 442–445, 2011. DOI: 10.1109/ICTAI.2011.72. Disponível em: https://ieeexplore.ieee.org/document/6103362.

ROWTULA, Vijay; OOTA, Subba Reddy; C.V, Jawahar. Towards Automated Evaluation of Handwritten Assessments, p. 426–433, 2019. DOI: 10.1109/ICDAR.2019.00075. Disponível em: https://ieeexplore.ieee.org/document/8977982.

SANUVALA, Ganga; FATIMA, Syeda Sameen. A Study of Automated Evaluation of Student's Examination Paper using Machine Learning Techniques, p. 1049–1054, 2021. DOI: 10.1109/ICCCIS51004.2021.9397227.

SCHLECHTWEG, Dominik et al. SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. **ArXiv**, abs/2007.11464, 2020. Disponível em: https://api.semanticscholar.org/CorpusID:220686630.

SHAUKAT, Mohammad Shaharyar et al. Chapter 16 - Semantic similarity–based descriptive answer evaluation. Edição: Sarika Jain, Vishal Jain e Valentina Emilia Balas. Academic Press, p. 221–231, 2021. DOI: https://doi.org/10.1016/B978-0-12-822468-7.00014-6. Disponível em: https://www.sciencedirect.com/science/article/pii/B9780128224687000146.

STELMAKH, Ivan et al. Asqa: Factoid questions meet long-form answers. **arXiv preprint arXiv:2204.06092**, 2022. Disponível em: https://aclanthology.org/2022.emnlp-main.566.pdf.

THALOR, Meenakshi A. A Descriptive Answer Evaluation System Using Cosine Similarity Technique. In: 2021 International Conference on Communication information and Computing Technology (ICCICT). [S.l.: s.n.], 2021. P. 1–4. DOI: 10.1109/ICCICT50803.2021.9510170. Disponível em: https://ieeexplore.ieee.org/document/9510170.

WANG, Chenglin et al. LSGC: An Interactive Text Matching Model Combined with Enhanced Encoding, p. 1–7, 2022. DOI: 10.1109/IJCNN55064.2022.9889792. Disponível em: https://ieeexplore.ieee.org/document/9889792.

ZHANG, Jing. Application Research of Similarity Algorithm in the Design of English Intelligent Question Answering System. In: 2022 IEEE 2nd International Conference on Mobile Networks and Wireless Communications (ICMNWC). [S.l.: s.n.], 2022. P. 1–4. DOI: 10.1109/ICMNWC56175.2022.10031708.

ZHANG, Shutao et al. Enhanced Text Matching Based on Semantic Transformation. **IEEE Access**, v. 8, p. 30897–30904, 2020. DOI: 10.1109/ACCESS.2020.2973206. Disponível em: https://ieeexplore.ieee.org/document/8993831.