http://journal.unoeste.br/index.php/ce/index
DOI: 10.5747/ce.2023.v15.e399
ISSN on-line 2178-8332

Colloquium Exactarum

Submetido: 22/06/2021 Revisado: 04/08/2022 Aceito: 17/08/2022

CORREÇÃO AUTOMATIZADA DE QUESTÕES DISSERTATIVAS UTILIZANDO MEDIDAS DE SIMILARIDADE E PROCESSAMENTO DE LINGUAGEM NATURAL

AUTOMATED CORRECTION OF ESSAY QUESTIONS USING SIMILARITY MEASURES AND NATURAL LANGUAGE PROCESSING

Aislan Neves de Lima Dias, Francisco Assis da Silva, Leandro Luiz de Almeida, Mário Augusto Pazoti

Universidade do Oeste Paulista –UNOESTE, Faculdade de Informática de Presidente Prudente, Presidente Prudente, SP.
E-mail: aislanneves29@gmail.com, chico@unoeste.br, llalmeida@unoeste.br, mario@unoeste.br

RESUMO – O Brasil tem vivenciado na última década um grande aumento da procura pelo Ensino a Distância, com o advento da pandemia essa tendência ficou ainda mais acentuada com a adaptação forçada pela realidade pandêmica aos professores, embora difícil inicialmente isso contribuiu para uma visão clara de como ferramentas tecnológicas podem proporcionar um auxílio maior na prática de atividades pedagógicas e de ensino aos discentes, auxiliando e prevenindo avaliações incorretas durante um período de estresse físico ou emocional, ao mesmo passo que otimizaria o tempo dedicado a sua atividade profissional. Utilizando técnicas de Processamento de Linguagem Natural para obter a representação de conhecimento a partir resposta pré-processado do aluno, foi possível obter uma acurácia entre 50 e 77 por cento quando comparado a nota atribuída por um avaliador humano e o sistema computacional em si, bastando que o avaliador forneça ao sistema uma resposta de referência ou um conjunto de palavras-chave ou diversas amostras de cada questão daquilo que se espera que esteja abordado dentro da resposta avaliada.
Palavras-chave: PLN; Correção Automática; questões dissertativas;

ABSTRACT – Brazil has experienced in the last decade, a great increase in demand for Distance Learning, with the advent of the pandemic, this trend was even more accentuated with the adaptation forced by the pandemic reality to teachers, although difficult initially, contributed to a clear vision of how technological tools can provide greater assistance in the practice of their pedagogical and teaching activities to students, helping and preventing incorrect assessments during a period of physical or emotional stress, while optimizing the time dedicated to their professional activity. Using Natural Language Processing techniques to obtain the representation of knowledge from the student's pre-processed response, it was possible to obtain an accuracy between 50 and 77 per cent when compared to the grade assigned by a human evaluator and the computational system itself, simply by the evaluator providing the system a reference answer or a set of keywords or several samples of each question of what is expected to be covered within the evaluated answer.
Keywords: NLP; automatic evaluation; essay questions;

1. INTRODUÇÃO

O tempo dedicado pelo professor no planejamento das aulas, elaboração e correção de atividades avaliativas é reconhecido como hora-atividade, em convenções coletivas e acordos coletivos de trabalho, recebido como um adicional de 5% tanto na educação básica como no ensino superior (FEPESP, 2018). Segundo Parisi (2017), o tempo gasto pelo professor em trabalho fora de sala de aula é demasiadamente significativo, pois são funções fundamentais dentro do contexto educacional, tais como correções de provas, trabalhos e atividades dos discentes, por exemplo.

Dentro deste contexto, o uso da tecnologia, como a inteligência artificial (IA), tem crescido cada vez mais nesse âmbito educacional e que, conforme Carvalho, Nevado e Menezes (2005), vem sendo adotada por professores na realização do ensino, apresentando resultados importantes e produtivos na elaboração de práticas em sala de aula.

Estudos realizados dentro de uma análise de pequeno e médio prazo apontam que o uso da Inteligência Artificial será preponderante e possui uma potencialidade promissora na educação, sendo um fator significativo e importante no processo ensino-aprendizagem, com o avanço tecnológico e pesquisas realizadas nessa área evidenciando o comportamento mundial nesta direção (VICARI, 2018).

Quando voltamos ao processo de correção textual sendo feito muitas vezes de forma manual, podemos encontrar aplicações de correção de redações, estabelecendo critérios avaliativos, que poderiam servir como suporte ao examinador em tomadas de decisão, proporcionando um auxílio importante em sua avaliação (COSTA; OLIVEIRA; CASTRO JÚNIOR, 2020). Trabalhos como de Sirotheau *et al*. (2019) destacam como uma aplicação que processa linguagem natural na correção de questões dissertativas apresenta maturidade suficiente para que possam ser utilizadas em Ambientes Virtuais de Aprendizagem.

Desta forma, mesmo havendo ferramentas voltadas à correção automática de textos, estudos mais recentes sobre a língua portuguesa ainda são escassos e proprietários, evidenciando um espaço para novas soluções de código aberto. Além disso, o tempo gasto pelo docente numa tarefa manual de correção de questões dissertativas é significativo e pode afetar o planejamento de outras atividades. Além disso, o aumento das atividades remotas e o crescimento do EAD, traz novos problemas, como por exemplo, o tempo despendido pelo docente para atender uma demanda alta de alunos no seu processo avaliatório. O tempo gasto por um trabalhador em horas-extras não garante a mesma confiabilidade do que aquele realizado em horas normais, visto que, o cansaço e o estresse emocional começam a se acentuar ainda mais após o seu período normal de atividades, situação essa, que se pode refletir nas avaliações de questões dissertativas prejudicando o processo avaliativo.

Na ferramenta desenvolvida pelos autores deste estudo pretende-se realizar uma comparação da resposta obtida pelo avaliador humano e a resposta atribuída pelo avaliador automático e com isso obter um coeficiente Kappa Moderado e Substancial, isto é, entre 0,41 e 0,6 e entre 0,61 e 0,8, tratando-se de uma métrica utilizada para avaliar a concordância entre dois avaliadores distintos.

O trabalho foi ordenado da seguinte forma. Na seção 2 é descrita a fundamentação teórica com a explicação das técnicas utilizadas neste trabalho. Na Seção 3 é apresentado a metodologia como o TF-IDF, LSA e CNN. Por sua vez, na Seção 4 é apresentada a interface do sistema desenvolvido. Na Seção 5 os experimentos realizados e os resultados obtidos. Por fim, é apresentada a conclusão do trabalho com os resultados alcançados e respectivas contribuições.

2. FUNDAMENTAÇÃO TEÓRICA

2.1. TF-IDF

O Term Frequency-Inverse Document Frequency (TF-IDF) é uma abordagem que consiste em identificar a ocorrência dos termos em um determinado documento e a frequência inversa dos documentos, sendo feito de forma bruta, com o produto proporcionado por duas medidas estatísticas TF e IDF (RAJARAMAN; ULLMAN, 2011).

Para o cálculo do TF é utilizado a equação (1) onde é feita uma contagem de quantas vezes uma determinada palavra aparece em um documento, buscando assim encontrar a frequência desta palavra, ao passo que o cálculo da frequência inversa é utilizado para que palavras que raramente ocorram possam ter um peso considerável, evitando que seja desconsiderada sua importância dentro do contexto (JONES, 1972).

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}} \quad (1)$$

O objetivo do cálculo da frequência inversa (IDF) definido pela equação (2) é detectar discrepâncias tangíveis no documento, onde palavras comuns e raras, obtenham um peso menor ou maior em relação ao documento que estão contextualizadas. Palavras com uma frequência próxima de zero podem ser descartadas, já que indicam uma importância quase nula no documento levado em questão (JONES, 1972).

$$idf(w) = \log\left(\frac{N}{df_t}\right) \quad (2)$$

A partir da obtenção das duas medidas TF e IDF, é feita a multiplicação entre elas, buscando-se obter o valor TF-IDF de forma ponderada ao mesmo passo que sua frequência é alta ela é balanceada pelo valor da infrequência, IDF.

$$w_{i,j} = tf_{i,j} \cdot \log\left(\frac{N}{df_i}\right) \quad (3)$$

2.2. LSA

O LSA (*Latent Semantic Analysis*) trata-se de uma técnica utilizada para realizar a filtragem de informação, especialmente desenvolvida para que se obtenha a partir de uma quantidade de palavras-chave, documentos que sejam relevantes dentro de um espaço vetorial, a intuição principal desta abordagem é através da utilização da Decomposição em Valores Singulares, em inglês, *Singular Value Decomposition* (SVD), que obtenha-se uma matriz multidimensional onde as dimensões mais superiores representam o significado de forma mais enxuta e expressiva, removendo os ruídos e reduzindo a dimensionalidade de um documento o que, por consequência do tamanho menor, impacta na diminuição do tempo de processamento (LANDAUER; LAHAM; FOLTZ, 1997).

As dimensões são ordenadas das mais significativas até as menos significativas, sendo assim, as dimensões superiores representam os dados mais importantes obtidos do documento enquanto as dimensões inferiores apresentam informações classificadas como de menor importância e ruidosa dentro do contexto avaliado pela técnica (LANDAUER, LAHAM; FOLTZ, 1997).

Segundo Wiemer-Hastings (2004) os passos mais comuns dentro de uma abordagem LSA caracterizam-se em cinco etapas, sendo estas:

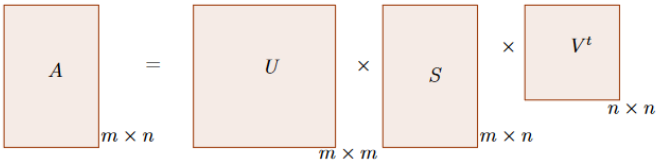
- Primeiro é necessário obter uma matriz de co-ocorrência de termos por documentos, que represente em forma vetorial cada documento desse corpus, por exemplo, um vetor de documentos, onde cada posição possui um respectivo vetor de contagem de palavras dentro desse documento em específico.
- Ao obter essa matriz, pode-se aplicar um peso dentro de cada palavra de cada documento específico, buscando representar cada documento de forma a destacar as informações mais importantes que se deseja buscar.
- A Decomposição em Valores Singulares (Singular Value Decomposition - SVD) é aplicada obtendo-se uma representação de conhecimento de acordo com o tamanho da dimensão k especificada.

A lógica utilizada pela Decomposição em Valores Singulares principal fator por trás dessa abordagem, permite que sejam encontradas Estruturas Latentes Semânticas (escondidas) dentro da estrutura de palavras através do documento, o SVD (Singular Value Decomposition) realiza a decomposição da matriz termo documento como demonstrado na equação (4) (LANDAUER; LAHAM; FOLTZ, 1997).

$$M = U \sum V_T \quad (4)$$

A transformação utilizando o SVD, utilizada pelo LSA embora à princípio tenha sido feita utilizando a abordagem tradicional da álgebra linear multiplicando-se as três matrizes geradas pela Decomposição Singular de Valores, é feita em sua maioria utilizando vetores esparsos e algoritmos que façam a computação de forma otimizada multiplicando apenas os termos mais importantes de cada matriz (WIEMER-HASTINGS, 2004).

Figura 1. SVD da matriz A



Fonte: (Santos *et al*., 2016).

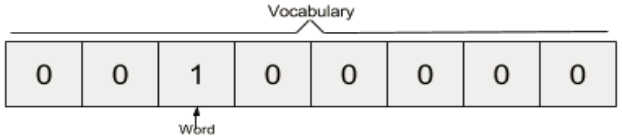
A ideia por trás do SVD é que os dados possam ser rotacionados e escalonados de forma a ocupar um espaço menor, obtendo-se no final do processo uma matriz documento por tópicos Z, diminuindo ao máximo a quantidade de ruídos e a dimensionalidade.

2.3. Incorporação de Palavras (Word2Vec)

O Word2Vec trata-se de um modelo de Rede Neural customizado para que possa ser treinado para reconstruir o contexto linguístico das palavras.

Este modelo é criado a partir de uma entrada de texto, assinalando vetores específicos para cada palavra dentro do corpus de forma que as palavras possam ser definidas dentro de um espaço vetorial, normalmente multidimensional, palavras similares, que compartilham de um mesmo contexto dentro do corpus do documento, tipicamente tem vetores posicionados pertos um do outro, pois possuem um significado sintaticamente e semanticamente similar, caso contrário, os vetores são posicionados longe um do outro o que indica uma baixa similaridade, entre duas palavras distintas. Na Figura 2 é ilustrado um exemplo de um vetor que representa uma palavra dentro de um vocabulário de palavras qualquer.

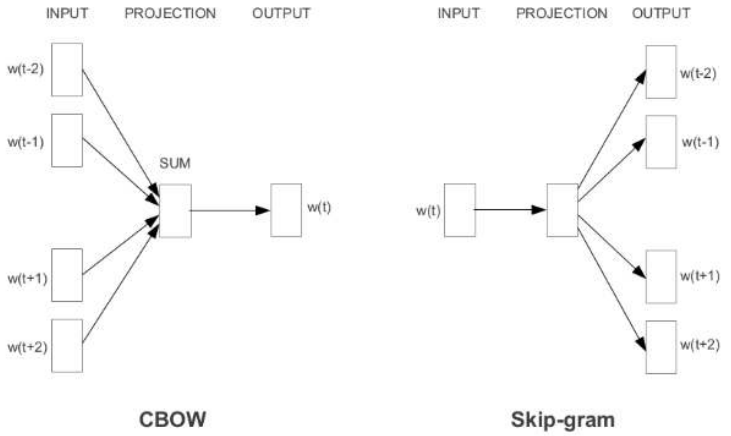
Figura 2. Posição da palavra na dimensão vocabulário



Fonte: (ALI, 2019)

O Word2Vec utiliza dois modelos de arquitetura, para realizar esse representação distribuída de palavras: Continuous Bag-of-Words (BOW) e Continuous Skip-Gram (SKIP-GRAM), a principal diferença entre os dois modelos, consiste em que no primeiro caso pretende-se prever o contexto (palavras vizinhas da palavra de entrada) a partir de uma palavra alimentada ao modelo, enquanto no segundo pretende-se prever uma palavra alvo a partir do contexto alimentado ao modelo, a Figura 3 é um exemplo ilustrativo dos dois tipos de abordagem (MIKOLOV *et al*., 2013).

Figura 3. CBOW e SKIP-GRAM



Fonte: (Mikolov *et al.* , 2013)

2.4. CNN (Convolutional Neural Networks)

Uma Rede Neural Convolucional ou Convolutional Neural Network (CNN), trata-se de uma Rede Neural Convencional que utiliza convolução, foi criada com o objetivo principal de ser utilizada para encontrar padrões em imagens.

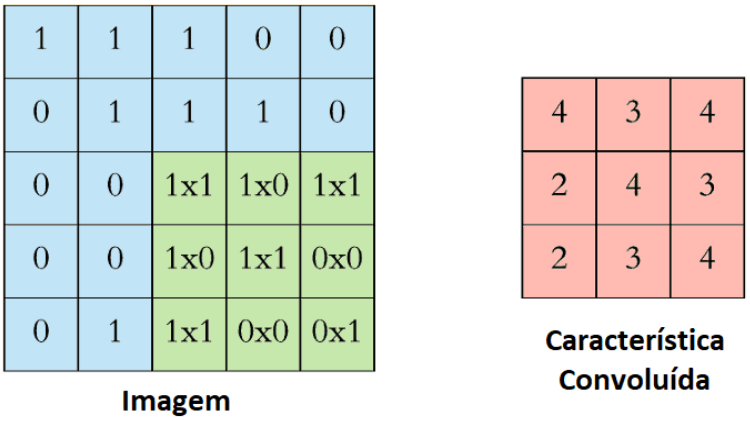
A abordagem tradicional consiste de realizar após cada camada de convolução, a aplicação de uma camada de pooling, sempre procurando reduzir pela metade a entrada, a importância deste tipo de arquitetura é que apesar da redução da dimensionalidade, as características mais importantes são preservadas e como o tamanho do filtro se mantém sempre o mesmo, padrões maiores são detectados, um exemplo ilustrativo é o aprendizado em detecção de bordas na primeira camada, em seguida numa segunda camada, membros e na terceira, a face de uma pessoa, no caso de imagens (ALBAWI; MOHAMMED; AL-ZAWI, 2017).

2.4.1. Convolução

A ideia intuitiva de uma convolução consiste em a partir de uma imagem de entrada e um filtro, também conhecido como *kernel*, obter uma imagem de saída.

O símbolo matemático utilizado para uma operação convolucional é a estrela ou asterisco, não confundir com o operador matemático de multiplicação utilizado na programação, pode-se destacar dois exemplos de aplicações famosas da convolução como detecção de bordas e o desfoque de uma imagem, através de um filtro gaussiano. Na Figura 4, pode-se observar um exemplo de uma operação convolucional, o filtro, em verde, sobrepondo a matriz em azul e a característica convolvida que consiste no resultado da convolução (ALBAWI; MOHAMMED; AL-ZAWI, 2017).

Figura 4. Exemplo de convolução



Fonte: (Dertat, 2017)

Na equação (5), a abordagem convolucional de Aprendizado Profundo efetivamente utilizada em Redes Neurais Convolucionais, neste caso, é a relação cruzada.

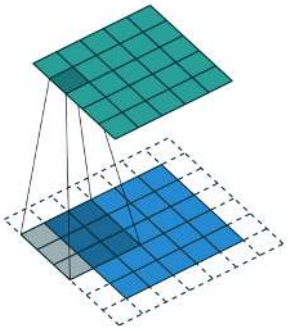
$$(A * w)_{ij} = \sum_{i'=0}^{K-1} \sum_{j'=0}^{K-1} A(i+i', j+j')\, w(i', j') \quad (5)$$

2.4.2. Padding

O movimento do filtro em uma camada de convolução é delimitado pelas bordas da imagem, o modo padrão utilizado é o conhecido como válido, onde a saída é sempre menor que a entrada.

O padding é uma técnica utilizada para manter o tamanho dos dados de entrada, consistindo em adicionar zeros imaginários ao redor da matriz, sendo uma forma de evitar a perda de informação durante a etapa de convolução (DERTAT, 2017).

Figura 5. Padding



Fonte: (Dertat, 2017)

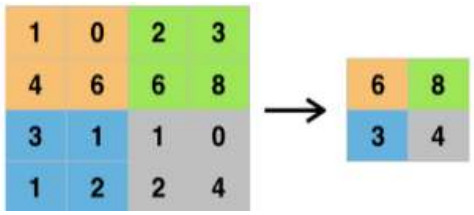
2.4.3. Pooling

Sua principal função é reduzir a dimensionalidade dos dados de entrada, mantendo as características importantes daquilo que está sendo analisado e diminuindo informações como onde a característica foi encontrada.

Existem dois tipos de pooling, Average Pooling e sendo o Max Pooling o mais utilizado, o método mais comum consiste em diminuir a dimensionalidade da matriz de características pela metade, a Figura 6 abaixo é um exemplo de como o processo de Max Pooling funciona, no primeiro caso o maior valor de cada subdivisão é considerado enquanto no segundo a média dos valores dentro de cada área analisada é utilizado (ALBAWI; MOHAMMED; AL-ZAWI, 2017).

A ideia básica por trás da utilidade do Pooling, é que a quantidade de dados para processar é diminuída e o princípio da invariância translacional, onde o importante é que a característica foi encontrada e não onde foi localizada (ALBAWI; MOHAMMED; AL-ZAWI, 2017).

Figura 6. Max Pooling

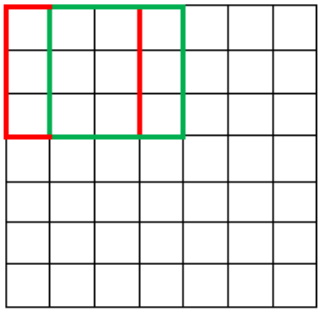


Fonte: (Deep AI, 2022)

2.4.4. Stride

O stride, similar ao pooling, consiste na ideia de sobreposição do filtro, podendo ser utilizado como uma forma de redução na dimensionalidade de entrada, possuindo mais eficiência, consiste no estabelecimento de um valor de quantos passos a janela deslizante deve dar de um lado para o outro ao passar na imagem, a eficiência por trás dessa técnica vem da ideia de que imagens geralmente irão possuir pixels muito similares ao seu redor e, portanto, redundantes e ruidosos (DEEP AI, 2022).

Figura 7. Stride



Fonte: (Deep AI, 2022)

2.5. Medidas de similaridade

Entre as medidas de similaridade utilizadas, para determinar o quão similar é um termo em comparação ao outro, pode-se citar a distância euclidiana e do cosseno.

2.5.1. Distância Euclidiana

A distância euclidiana é uma medida que calcula a distância entre dois pontos no plano cartesiano, é uma medida utilizada para avaliar a distância entre dois pontos em um espaço vetorial, a equação (6) refere-se a distância euclidiana para planos com duas dimensões, que pode ser utilizado para encontrar a distância relativa, de acordo com os pontos x e y (SMITH, 2011; COHEN, 2004).

$$d_{ij} = \sqrt{\left(x_{i1} - x_{j1}\right)^2 + \left(x_{i2} - x_{j2}\right)^2} \quad (6)$$

2.5.2. Similaridade e Distância do Cosseno

A similaridade do cosseno é uma medida muito utilizada em mineração de textos, para comparação de quão similar é um vetor numérico de um determinado termo em comparação a outro. Qualquer que seja o tamanho dos documentos analisados, a fórmula (7) refere-se à similaridade entre os vetores, enquanto para se calcular a distância é necessário subtrair o valor obtido de 1, x e y são os pontos do vetor no espaço (NIST, 2017; SINGHAL *et al.,* 2001).

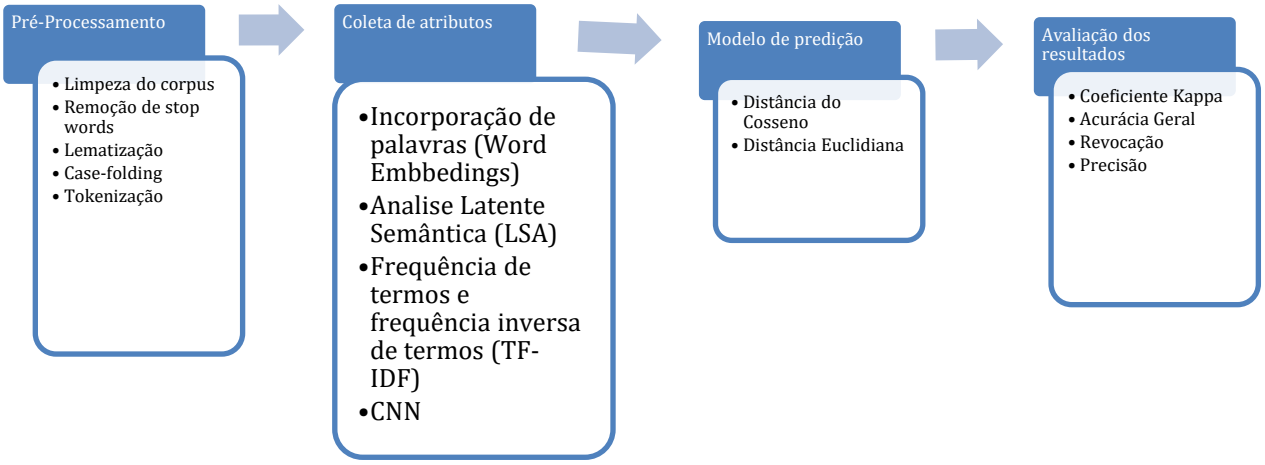
$$\frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \quad (7)$$

3. METODOLOGIA

Esta seção descreve a abordagem proposta consistindo em 5 etapas: (1) Pré-Processamento, (2) Coleta de atributos, (3) Modelo de predição, sendo a última etapa (5) dedicada à avaliação dos resultados.

A metodologia apresentada pode ser observada na Figura 8.

Figura 8. Metodologia



Fonte: Os autores

3.1. Base de Dados

A base de dados foi montada buscando responder a seguinte pergunta: “Explique o que é aprendizagem por Reforço”, a nota foi atribuída seguindo a seguinte lógica, respostas completas que abordassem todos os conceitos foram pontuadas com a pontuação máxima, respostas parciais receberam metade da pontuação, enquanto respostas incorretas receberam a pontuação mínima, ou seja, zero.

No Quadro 1 é ilustrado uma amostra das respostas da base de dados utilizada para o teste.

Quadro 1. Amostra da Base de Dados

Classificação	Resposta
Referência	A aprendizagem por reforço é muito utilizada em sistemas de decisão em tempo real. Neste caso, não existe um grande conjunto de dados históricos para que seja gerado um conjunto de treinamento e consequentemente seja possível criar um modelo de aprendizado. O tempo todo novos dados são criados e processados servindo de combustível para decisões. Uma boa ilustração de aplicação são sistemas que operam em ambientes incertos e mutáveis, esse tipo de aprendizagem faz com que robôs por exemplo consigam trafegar por ambientes nunca explorados anteriormente
Correta	É muito utilizada em sistemas que fazem a decisão em tempo real, não existe um conjunto de dados para ser treinado e criar o modelo de aprendizado, a todo momento novos dados são criados e processados para alimentar novas decisões dentro do algoritmo. Uma forma de visualizar esse funcionamento é através de sistemas que podem operar ambientes que são incertos e que mudam constantemente, robôs dessa forma podem trafegar por caminhos que antes eles não caminhavam.
	É uma forma de aprendizado que é utilizado para sistemas que fazem decisão em tempo real, conjunto de dados de treinamento não é possível ser criado para gerar um modelo de aprendizado que possa ser utilizado para criar um modelo de aprendizado, consiste na criação de novos dados que mais tarde podem ser utilizados para alimentar o sistema como combustível para novas decisões. Um exemplo são aplicações que operam em solo que muda e é sempre incerto, sistemas computacionais podem dessa forma explorar e descobrir novos horizontes que antes eles não conseguiam.
	São aplicações que operam ambientes incertos e que mudam constantemente, sistemas computacionais aqui podem tomar decisões baseado no conjunto de dados anteriormente definido, consiste no aprendizado a partir de um conjunto de decisões em tempo real, como não é possível criar um modelo de aprendizado devido à falta de dados históricos, o sistema precisa criar dados constantemente.
Parcial	Como não existe uma forma de colocar dados de entrada neste tipo de sistema, não é possível fazer o treinamento como um modelo de aprendizado tradicional.
	As decisões são tomadas em tempo real a partir de dados que estão sendo gerados constantemente.
	O modelo de treinamento convencional não se aplica a esse tipo de sistema.
Incorreta	O uso do conhecimento aqui é feito com base naquilo que ele forma com base naquilo que é formado utilizando o conhecimento proporcionado pelo conjunto de treinamento, por se tratar de um sistema que faz uso do conjunto de treinamento tudo funciona com base naquilo que foi coletado.
	Os novos aprendizados são possíveis pois o sistema faz o uso do conhecimento que é coletado com base num conjunto de treinamento convencional, não podendo ser utilizado para tomar decisões em tempo real, sendo utilizado apenas para formar conhecimento único e diferenciado.

	O uso do conhecimento é feito com base naquilo que está sendo formado e, portanto, construído com o sistema em si, esse tipo de conhecimento é gerado utilizando o conjunto de treinamento e para isso é feito uma análise que não pode ser utilizada em sistemas de tempo real.
	Formatos de conhecimento aqui gerados não fazem uso de tempo real e prejudicam a forma como o sistema funciona em tese, pois o sistema se alimenta com a utilização de um conjunto de treinamento e não se alimenta com um conjunto de treinamento personalizado.

Fonte: Os autores.

3.2. Pré-Processamento

Na primeira etapa o corpus da resposta dissertativa passa por um processo que consiste na eliminação de acentuação, transformação da caixa do texto em caixa-baixa e remoção de pontuação, buscando eliminar divergências que possam
atrapalhar que a frequência e importância de uma palavra seja mensurada dentro de um contexto, uma vez que, a pontuação por ao aparecer com mais frequência pode obter um peso maior dentro de algoritmos como o TF-IDF que avaliam a frequência. A eliminação da pontuação, acentuação e transformação em caixa baixa, são alternativas que buscam por sua vez eliminar as possíveis diferentes formas que uma palavra pode aparecer dentro de um determinado corpus e prejudicar a avaliação de sua importância, ao considerar palavras iguais, porém apresentadas de formas distintas de acordo com os formalismos da Língua Portuguesa.

Além disso, técnicas como a Lematização foram aplicadas, buscando substituir as diversas aparições morfológicas de uma mesma palavra dentro do corpus apresentado, a Lematização consiste em um processo onde são eliminados os sufixos de uma palavra e a mesma é substituída por uma palavra que a representa no modo infinitivo, por exemplo a palavra comereis, ao passar por um processo de Lematização tem seu sufixo eliminado, neste caso -eis, diferentemente do *stemming*, aqui a palavra é substituída por sua forma no infinitivo que realmente existe dentro do vocabulário da Língua Portuguesa, para esta questão em específico, “comer”.

Outra técnica de Pré-Processamento utilizado em sistemas de PLN é a remoção de stopwords, palavras que são repetitivas dentro de textos dissertativos em linguagem natural e podem confundir sistemas computacionais que avaliam a importância de uma palavra dentro de uma resposta com base na frequência, exemplos de termos repetitivos e que não acrescentam representação de conhecimento.

Quadro 2. Exemplo de pré-processamento

Etapas	Resposta
Corpus Original	As queimadas nos estados do norte do Brasil, especialmente na Amazônia
Case Folding	as queimadas nos estados do norte do brasil, especialmente na amazônia.
Remoção de Stop Words	as queimadas nos estados do norte do brasil, especialmente na amazônia.
Eliminação de pontuação	queimadas estados norte brasil especialmente, amazônia
Eliminação de acentos	queimadas estados norte brasil especialmente amazônia
Resumo textual	queimadas norte brasil amazonia
Tokenização	[queimadas, norte, brasil, amazonia]

Fonte: Os autores.

3.3. Coleta de Atributos

A etapa de coleta de atributos consiste em obter-se a representação da estrutura semântica da forma mais performática e fidedigna possível através de técnicas de extração e mineração de texto como TF-IDF baseada na extração de *features* consistindo em unigramas, bigramas e trigramas, conversão em

vetores como *Word Embeddings* (Word2Vec) e técnicas de diminuição do espaço semântico como é o caso do LSA (Latent Semantic Analysis).

Foram realizados testes utilizando três abordagens, a primeira seria a coleta de atributos com base em uma resposta de referência, onde foi utilizado TF-IDF, Word2Vec e LSA, enquanto a segunda baseia-se na utilização de um conjunto de treinamento utilizando-se principalmente de uma CNN.

A partir da resposta pré-processada, ela passa por uma processo de extração de *features* consistindo de unigramas, bigramas e trigramas da resposta de referência utilizando-se principalmente TF-IDF, a partir das *features* obtidas é então realizado uma conversão das respostas dos alunos em vetores de frequência, obtendo-se representações numéricas que possam ser processadas computacionalmente, este vetor obtido passa por mais uma etapa buscando eliminar ruídos e redundâncias que atrapalhem na etapa de atribuição de nota e não foram eliminados na etapa de pré-processamento, utilizando-se do LSA, para esta abordagem foram escolhidas as 6 camadas superiores do modelo proposto, a ideia por trás da utilização do LSA além da redução do espaço semântico o que impacta num tempo de processamento menor, outra vantagem é a solução do problema dos sinônimos (quando múltiplas palavras têm um mesmo significado) e da polissemia (uma mesma palavra que pode ter diferentes significados).

Uma segunda técnica utilizada foi uma abordagem de Incorporação de Palavras utilizando Word2Vec, através de um modelo pré-treinado, realizando a conversão das palavras em vetores multidimensionais que são representáveis dentro de um espaço vetorial, a partir da média desses vetores é possível obter uma representação da resposta em si para ser analisada posteriormente.

A terceira técnica consiste na utilização de uma CNN separando 50% para o conjunto de treino e 50% para o conjunto de testes, utilizou-se uma arquitetura padrão, *early callbacks*, alimentando a rede com o texto convertido em *sequences*.

3.4. Distância do Cosseno e Euclidiana

Para avaliar a similaridade entre os vetores, foi utilizado a distância euclidiana e do cosseno, realizando uma média entre as duas medidas, a utilização da distância do cosseno tem por objetivo resolver problemas que podem acontecer quando apenas é considerado a distância relativa entre os vetores sem considerar o ângulo, neste caso, sentenças que citam uma palavra-chave em quantidade maior podem ter seu vetor da sentença posicionado de forma distante em relação a uma sentença que cita a mesma palavra porém em uma quantidade menor devido ao tamanho da resposta.

3.5. Critério para avaliação das respostas

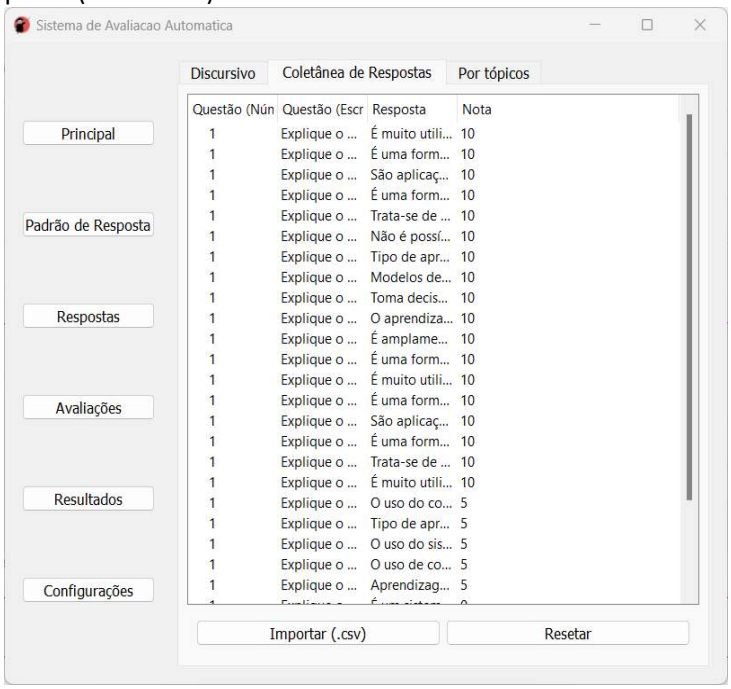
Utilizou-se o coeficiente Kappa, uma métrica estatística muito utilizada para avaliar a concordância entre dois avaliadores distintos, neste caso o avaliador humano e o avaliador automático que no caso seria a aplicação desenvolvida, também foi analisado a acurácia, revocação e precisão, o critério utilizado para que uma resposta seja considerada correta foi quando seu índice de similaridade é maior que 65%, meio certa quando é maior que 50% e está abaixo de 65% e incorreta quando possui índice menor que 50%, esse tipo de classificação ocorre pois respostas que contenham as palavras-chave podem possuir um índice de similaridade médio, porém não o suficiente para obter um índice maior como 75%, devido ao contexto desconexo com o padrão de resposta.

4. INTERFACE

A ferramenta desenvolvida consiste em cinco módulos, sendo estes: Padrão de Resposta, Respostas, Avaliações e Resultados. Cada módulo tem uma função específica, que é descrita abaixo:

Padrão de Resposta: Nesta aba do sistema, o avaliador humano entrará com o(s) padrão(ões) de resposta para cada questão alimentada ao sistema, cabendo ao mesmo decidir qual tipo pretende-se utilizar, podendo ser desde um padrão de resposta único de referência, no caso da opção pelo padrão por palavras-chave ou discursivo (resposta ótima) ou outra opção, no caso um padrão de resposta por coletânea que consiste na entrada de diversas amostras de respostas para uma mesma questão.

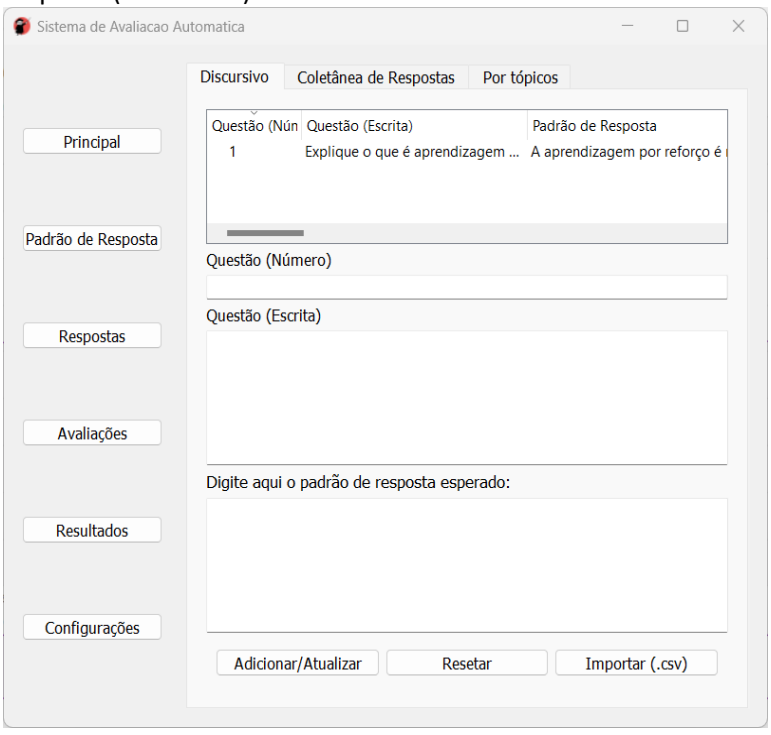
Figura 9. Padrão de Resposta (Coletânea)



Fonte: Os autores.

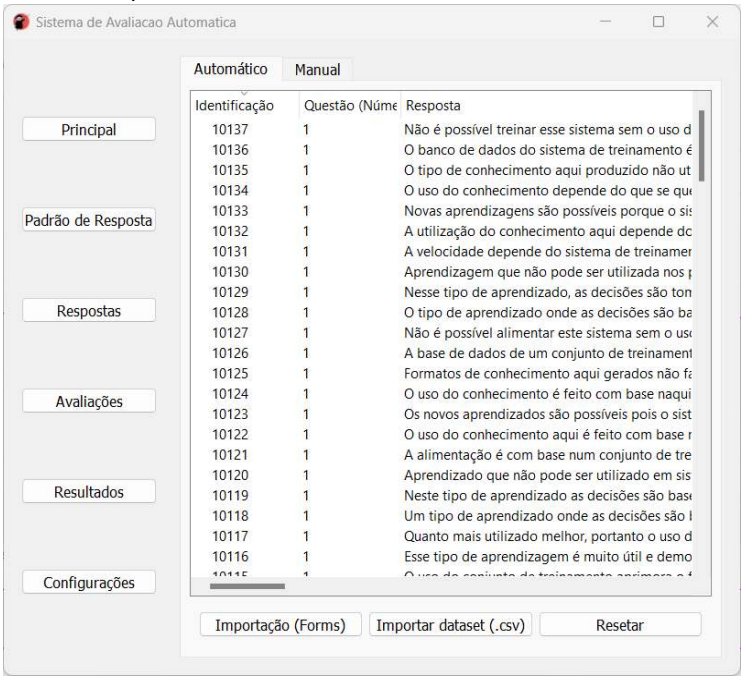
Respostas: Nesta seção do sistema, o avaliador alimentará a ferramenta com as respostas dos alunos, tanto de forma automática através do Google Forms, como por importação de um arquivo CSV.

Figura 10. Padrão de Resposta (Discursivo)



Fonte: Os autores.

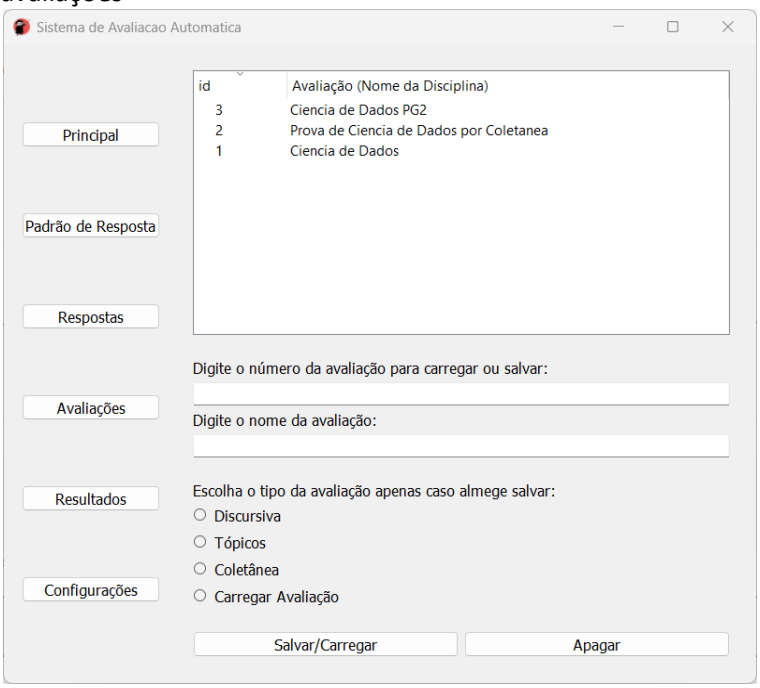
Figura 11. Respostas (Automático)



Fonte: Os autores.

Avaliações: Nesta seção a avaliação pode ser salva, excluída e carregada dentro do sistema sem a necessidade de o professor realizar carregamento todas as vezes que for corrigir uma mesma prova.

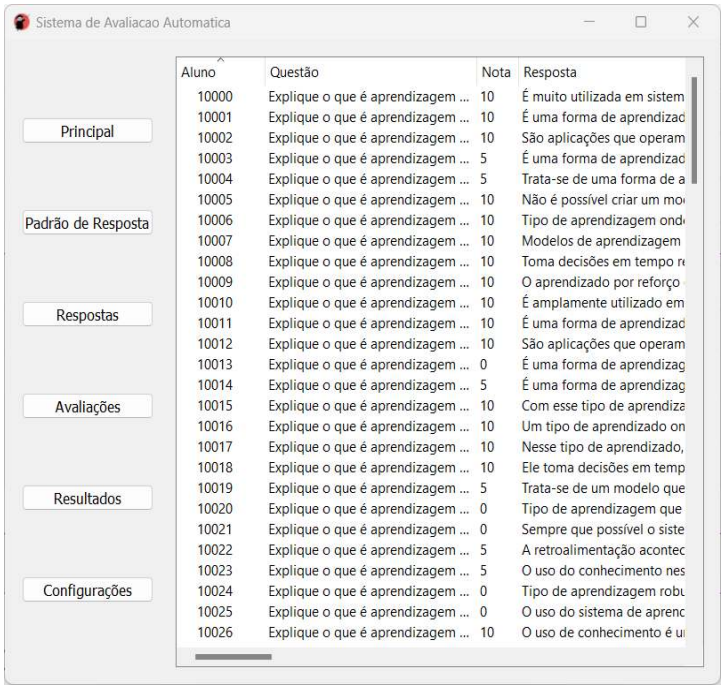
Figura 12. Módulo de avaliações



Fonte: Os autores.

Resultados: Na seção de resultados, é possível verificar a nota atribuída a cada aluno, sendo três notas possíveis, 10 (Correta), 5 (Parcial) e 0 (Incorreta), outras informações importantes como a identificação, enunciado da questão, nota e resposta também são visíveis.

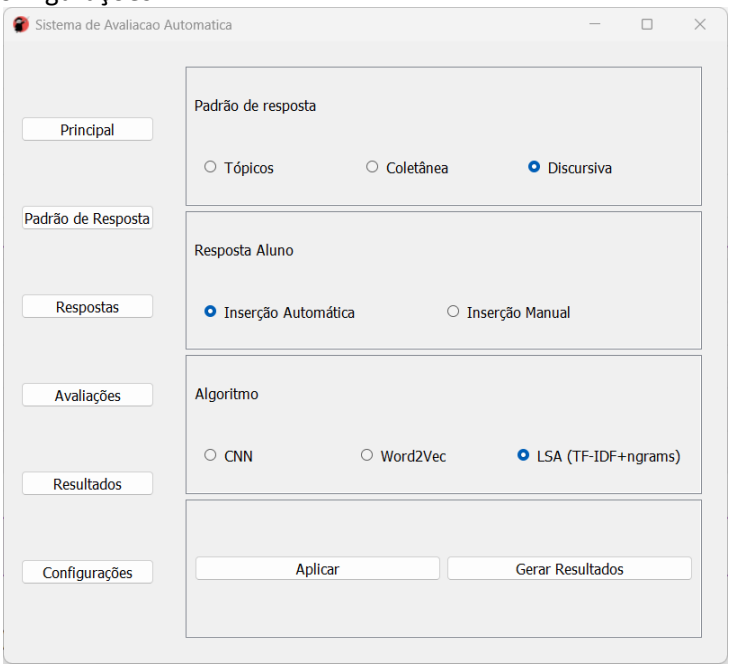
Figura 13. Módulo de resultados



Fonte: Os autores

Configurações: Seção da ferramenta onde o avaliador escolhe o que definiu previamente nas seções anteriores, recomenda-se o uso de CNN para os padrões de resposta por coletânea e Word2Vec ou LSA para os padrões de resposta que utilizam um único padrão de resposta.

Figura 14. Módulo de configurações



Fonte: Os autores.

5. RESULTADOS

Foram conduzidos três experimentos, o primeiro, utilizando uma abordagem utilizando LSA, que consiste em avaliar as respostas dos alunos de acordo com a cobertura da resposta apresentada. O segundo experimento tem como entrada uma única resposta de referência da qual a pontuação é atribuída de acordo com a distância da resposta do aluno. O terceiro experimento utilizou-se de palavras-chave. O quarto experimento conduzido baseia-se no padrão de resposta por coletânea através do treinamento de uma Rede Neural Convolucional (CNN), separando 50% para treino e 50% para teste. Resultados com

coeficiente Kappa menor que 0,20, foram considerados insuficientes pela baixa concordância com o avaliador humano.

Quadro 3. Resultados

Base de Dados	Acurácia	Kappa	Experimento
Geografia	54,46%	0,36 Razoável	LSA
Filosofia	50,27%	0,41 Moderada	LSA
Ciência de Dados	**76,25%**	**0,76 Substancial**	LSA
Geografia	52,83%	0,42 Moderada	Word2Vec
Filosofia	**60,98%**	**0,51 Moderada**	Word2Vec
Ciência de Dados	58,27%	0,36 Moderada	Word2Vec
Geografia	67,82%	0,59 Moderada	CNN
Filosofia	**73,95%**	**0,60 Moderada**	CNN
Ciência de Dados	**78,15%**	**0,51 Moderada**	CNN

Fonte: Os autores

5.1 Experimento 1: Resposta de Referência utilizando LSA

Quadro 4. Ciência de Dados

#		Única Resposta de Referência			Recall
		Predito			
	#	C	P	E	
Real	C	26	9	3	68,42%
	P	0	17	4	80,95%
	E	1	16	63	78,75%
Precisão		96,29%	40,47%	90%	**Geral 76,25%**

Fonte: Os autores.

Quadro 5. Filosofia

#		Única Resposta de Referência			Recall
		Predito			
	#	C	P	E	
Real	C	62	41	7	56,36%
	P	17	17	12	36,95%
	E	3	11	13	48,14%
Precisão		75,60%	24,63%	40,62%	**Geral 50,27%**

Fonte: Os autores.

Quadro 6. Geografia

#		Única Resposta de Referência			Recall
		Predito			
	#	C	P	E	
Real	C	19	19	0	50%
	P	27	75	5	70,09%
	E	10	41	28	35,44%
Precisão		33,92%	55,55%	84,84%	Geral **54,46%**

Fonte: Os autores.

5.2 Experimento 2: Resposta de Referência utilizando Word2Vec

Quadro 7. Ciência de Dados

#		Única Resposta de Referência			Recall
		Predito			
	#	C	P	E	
Real	C	19	18	1	50%
	P	1	7	13	33,33%
	E	1	24	55	68,75%
Precisão		90,47%	14,28%	79,71%	Geral **58,27%**

Fonte: Os autores.

Quadro 8. Filosofia

#		Única Resposta de Referência			Recall
		Predito			
	#	C	P	E	
Real	C	79	21	4	75,96%
	P	25	17	8	34%
	E	5	8	15	53,57%
Precisão		72,47%	36,95%	55,55%	Geral **60,98%**

Fonte: Os autores.

Quadro 9. Geografia

#		Única Resposta de Referência			Recall
		Predito			
	#	C	P	E	
Real	C	17	19	1	45,94%
	P	26	68	19	52,83%
	E	4	39	36	45,56%
Precisão		36,17%	53,96%	64,28%	Geral **52,83%**

Fonte: Os autores.

5.3 Experimento 3: Utilizando Palavras-Chave

Quadro 10. TF-IDF+N-grams

#		Única Resposta de Referência			Recall
		Predito			
	#	C	P	E	
Real	C	27	10	1	71,05%
	P	0	10	11	47,61%
	E	3	15	62	77,50%
Precisão		90%	28,57%	83,78%	Geral **71,22%**

Fonte: Os autores.

5.4 Experimento 4: Padrão por Coletânea

Quadro 11. Ciência de Dados

#		Coletânea			Recall
		Predito			
	#	C	P	E	
Real	C	22	0	10	68,75%
	P	0	8	9	47,05%
	E	3	4	63	90,00%
Precisão		88%	66,66%	76,82%	Geral **78,15%**

Fonte: Os autores.

Quadro 12. Filosofia

#		Coletânea			Recall
		Predito			
	#	C	P	E	
Real	C	52	0	2	96,29%
	P	13	8	5	30,76%
	E	5	0	11	68,75%
Precisão		74,28%	100%	61,11%	Geral **73,95%**

Fonte: Os autores.

Quadro 13. Geografia

#		Coletânea			Recall
		Predito			
	#	C	P	E	
Real	C	12	7	0	63,15%
	P	6	46	14	69,69%
	E	1	9	20	66,66%
Precisão		63,15%	74,19%	58,82%	Geral **67,82%**

Fonte: Os autores.

6. DISCUSSÃO

Constatou-se que a base de dados de Ciência de Dados apresentou resultados significativos na maioria dos testes realizados, uma hipótese para isso seria pela forma como foi pontuado pelo avaliador humano, neste caso, pela proximidade da resposta do aluno com a resposta de referência, enquanto nos outros casos, as bases de dados públicas utilizadas como uma referência externa, tiveram um critério de avaliação multifatorial, desde a presença de determinadas palavras-chave como uma pontuação mais alta na composição da nota final, pela abordagem do respondente de determinados tópicos do padrão de

resposta oficial o que pode conferir-lhe uma nota maior, independentemente do tamanho da resposta e da abrangência dos demais tópicos, dificultando a forma como o sistema avalia, já que o que foi desenvolvido trata a composição do escore atribuído de forma homogênea.

Dentro dessa temática, estudos correlatos demonstram como a utilização de tecnologias que processam linguagem natural podem agilizar os processos de correção automática, em redações como o do Exame Nacional do Ensino Médio (ENEM) (NAU *et al*., 2019), em correção de erros gramaticais em textos de língua inglesa (BRYANT; BRISCOE, 2018; OMELIANCHUK *et al*., 2020), em erros de ortografia e escrita (HU *et al.*, 2020), bem como, analisar a escrita e a correta utilização de termos da literatura médica (LÓPEZ-HERNÁNDEZ; ALMELA; VALENCIA-GARCÍA, 2019), com coeficientes Kappa satisfatórios, ou seja, acima de 0,6.

Embora os resultados obtidos através dos métodos LSA e Word2Vec tenham sido menores nestas bases do que os obtidos no caso da base de dados de Ciência de Dados, a abordagem que utiliza uma Rede Neural Convolucional (CNN) apresentou os melhores resultados para estes casos, graças a sua capacidade de aprender características e a vantagem de possuir diversas amostras para determinar a pontuação final, o que não foi o caso das outras abordagens, onde pelo fato de possuir apenas a resposta ótima a única variável avaliada para determinar a pontuação foi a similaridade.

7. CONSIDERAÇÕES FINAIS

A partir dos resultados apresentados pelo modelo proposto pode-se concluir que um sistema de avaliação automatizada para questões dissertativas pode ser utilizado para prover um resultado auxiliar como ferramenta de apoio, agilizando o processo de correção de avaliadores humanos e proporcionando uma segunda checagem, quando levamos em consideração o baixo nível de falsos positivos e falsos negativos, também é importante destacar que todos os modelos apresentaram um coeficiente Kappa que demonstra a efetividade e eficácia dos modelos propostos, variando entre concordância moderada e substancial, com exceção do resultado obtido na base de dados de Geografia no experimento 1, e os resultados obtidos tanto da base de Geografia e Filosofia no caso da experimentação por palavras-chave, este último experimento citado aqui, evidenciando a importância da análise contextual.

Como sugestão para trabalhos futuros, seria interessante a realização de uma integração em um ambiente de aprendizagem *WEB*, possibilidade de atribuir pesos maiores para determinados tópicos, correção heterogênea, incluir neste modelo uma escala maior de avaliação à resposta do aluno e a possibilidade de incluir conceitos pedagógicos como por exemplo, Taxonomia de Bloom.

REFERÊNCIAS

ALBAWI, S.; MOHAMMED, T. A.; AL-ZAWI, S. Understanding of a convolutional neural network. *In*: INTERNATIONALCONFERENCE ON ENGINEERING AND TECHNOLOGY (ICET). 2017. https://doi.org/10.1109/ICEngTechnol.2017.8308186.

ALI, Z. A simple Word2vec tutorial Medium, 7 jan. 2019. Disponível em: https://medium.com/@zafaralibagh6/a-simple-word2vec-tutorial-61e64e38a6a1. Acesso em: 24 abr. 2022.

BRYANT, C.; BRISCOE, T. Language model based grammatical error correction without annotated training data. *In*: WORKSHOP ON INNOVATIVE USE OF NLP FOR BUILDING EDUCATIONAL APPLICATIONS. 13., 2018. New Orleans, Louisiana. **Proceedings [**...]. New Orleans: Association for Computational Linguistics,2018. p. 247-253. https://doi.org/10.18653/v1/W18-0529

CARVALHO, M. J. S.; NEVADO, R. A.; MENEZES, C. S. Arquiteturas pedagógicas para educação à distância: concepções e suporte telemático. *In*: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO-SBIE, 16., 2005, Juiz de Fora. **Anais** [...]. Juiz de Fora, 2005. p. 351-360.

COHEN, D. **Precalculus**: A Problems-Oriented Approach , Cengage Learning,2004.

COSTA, L.; OLIVEIRA, E. H T.; CASTRO JÚNIOR, A. Corretor Automático de Redações em Língua Portuguesa: um mapeamento sistemático de literatura. *In*: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 31.,

2020. Porto Alegre, **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2020. p. 1403-1412. DOI: https://doi.org/10.5753/cbie.sbie.2020.1403.

NIST. COSINE DISTANCE, COSINE SIMILARITY, ANGULAR COSINE DISTANCE, ANGULAR COSINE SIMILARITY, 2017. Disponível em: https://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/cosdist.htm. Acesso em: 17 abr. 2022.

DERTAT, A. Applied Deep Learning - Part 4: Convolutional Neural Networks. **Towards Data Science**, nov. 2017. Disponível em: https://towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2. Acesso em: 11 dez. 2022.

FEPESP. **Hora-atividade.** São Paulo: FEPESP, 2018. Disponível em: http://fepesp.org.br/noticia/hora-atividade/. Acesso em: 10 out. 2021.

HU, Y.; JING, X.; KO, Y.; RAYZ, J. T. Misspelling Correction with Pre-trained Contextual Language Model. *In*: INTERNATIONAL CONFERENCE ON COGNITIVE INFORMATICS & COGNITIVE COMPUTING (ICCI* CC). 19., IEEE, 2020. p. 144-149. https://doi.org/10.1109/ICCICC50026.2020.9450253

JONES, K. S. A statistical interpretation of term specificity and its application in retrieval. **Journal of documentation**, 1972.

LANDAUER, T.; LAHAM, D.; FOLTZ, P. Learning human-like knowledge by singular value decomposition: A progress report. *In*: CONFERENCE ON ADVANCES IN NEURAL INFORMATION PROCESSING. NIPS'97. 1997. **Proceedings** [...]. 1997, v. 10. Disponível em: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://proceedings.neurips.cc/paper_files/paper/1997/file/cec6f62cfb44b1be110b7bf70c8362d8-Paper.pdf. Acesso em: 30 jun. 2023.

LÓPEZ-HERNÁNDEZ, J., ALMELA, Á., VALENCIA-GARCÍA, R. Automatic Spelling Detection and Correction in the Medical Domain: A Systematic Literature Review. *In*: VALENCIA-GARCÍA, R. *et al.* (eds.) **Technologies and Innovation**. CITI 2019. Communications in Computer and Information Science. [S.l.]: Springer, Cham, 2019, v.1124. https://doi.org/10.1007/978-3-030-34989-9_8

MIKOLOV, T. *et al.* Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 [cs], 6 set. 2013.

NAU, J.; HAENDCHEN FILHO, Aluizio; DAZZI, R. L. S. Identificação e avaliação automática da proposta de intervenção em textos dissertativos-argumentativos: uma revisão sistemática da literatura. In: COMPUTER ON THE BEACH, 10., 2019, Itajaí. **Anais** [...]. Itajaí, 2019. p. 493-501.

OMELIANCHUK, K.; ATRASEVYCH, V.; ARTEM, C.; SKURZHANSKYI, O. GECToR--Grammatical Error Correction: Tag, Not Rewrite. arXiv preprint arXiv:2005.12592, 2020. https://doi.org/10.18653/v1/2020.bea-1.16

PARISI, J. M. **A remuneração da hora-atividade e dos períodos de janela na jornada de trabalho do docente.** Trabalho de Conclusão de Curso (Graduação) - Faculdade de Direito, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2017.

RAJARAMAN, A.; ULLMAN, J. D. **Mining of massive datasets.** [S.l.]: Cambridge University Press, 2011. https://doi.org/10.1017/CBO9781139058452

SANTOS, J. C. A. *et al.* **Avaliação automática de questões discursivas usando LSA**. 2016. Tese (Doutorado em Engenharia Elétrica) - Instituto de Tecnologia, Universidade Federal do Pará, Belém, 2016.Disponível em: http://repositorio.ufpa.br/jspui/handle/2011/7485. Acesso em: 30 jun. 2023.

SMITH, K. J. **Precalculus**: A functional approach to graphing and problem solving. [S. l.]: Jones & Bartlett Publishers, 2011.

SINGHAL, A. *et al.* Modern information retrieval: A brief overview. **IEEE Data Eng. Bull**., v. 24, n. 4, p. 35-43, 2001.

SIROTHEAU, S.; SANTOS, J.; FAVERO, E.; FREITAS, S. N. Avaliação Automática de respostas discursivas curtas baseado em três dimensões linguísticas. *In:*
CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 8., SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 30., 2019, Brasília. **Anais** [...]. Brasília: 2019. https://doi.org/10.5753/cbie.sbie.2019.1551

WIEMER-HASTINGS, P. Latent semantic analysis. *In:* INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE. 16., 2004**. Proceedings** [..], 2004, p. 1-14.

VICARI, R. M. **Tendências em inteligência artificial na educação no período de 2017 a 2030**: sumário executivo. Porto Alegre: SENAI/RS, 2018.