Aplicação de técnicas de processamento de linguagem natural na automatização de correção de questões discursivas

Douglas Camilo de Oliveira[1], Eliane Pozzebon[1], Tatiana Nilson dos Santos[1]

[1]Laboratório de Tecnologias Computacionais – LabTeC
Departamento de Computação - Centro de Ciências, Tecnologias e Saúde
Universidade Federal de Santa Catarina (UFSC) – Araranguá – SC – Brasil

{douglas.dco.oliveira, epozzebon, tatiana.nilon}@gmail.com

Abstract. *The evaluation of students through discursive questions is one of the most traditional forms of the education system, but its correction is not a trivial task. In classes with a high number of students, this task occupies a large part of the teacher's working time. This problem is aggravated when we talk about an Intelligent Tutor System (STI) where the number of students can grow considerably and, automatically, the number of discursive questions for correction grows with each exercise proposed for the class. Carrying out studies in the area of Natural Language Processing (PLN) and using the similarity measures Cosine similarity and Word Mover's Distance, a solution was proposed for the creation of a system for automatic correction of discursive questions. Tests were carried out to evaluate the efficiency of each of the similarity measures, with Cosine Similarity standing out, once chosen to integrate the solution implemented in STI MAZK. In the tests performed, within the STI, MAZK's automatic correction system obtained a relative error of 15% and an accuracy of 88.7%, interesting and encouraging results.*

Resumo. *A avaliação dos alunos por meio de questões discursivas é uma das formas mais tradicionais do sistema de ensino, mas sua correção não é uma tarefa trivial. Em turmas com um número elevado de alunos essa tarefa ocupa boa parte do tempo de trabalho do professor. Este problema se agrava quando falamos de um Sistema Tutor Inteligente (STI) onde o número de alunos pode crescer consideravelmente e, automaticamente, o número de questões discursivas para correção cresce a cada exercício proposto para a turma. Realizando estudos na área de Processamento de Linguagem Natural (PLN) e utilizando as medidas de similaridade Cosine similarity e Word Mover's Distance, foi proposta uma solução para a criação de um sistema de correção automática de questões discursivas. Foram realizados testes para avaliar a eficiência de cada uma das medidas de similaridade sendo que a Cosine Similarity se destacou, uma vez escolhida para integrar a solução implementada no STI MAZK. Nos testes realizados, dentro do STI, o sistema de correção automática do MAZK obteve um erro relativo de 15% e uma acurácia de 88,7%, resultados interessantes e animadores.*

1. Introdução

Os Sistemas Tutores Inteligentes (STI) são sistemas computacionais voltados para o ensino e representam uma parte considerável da Inteligência Artificial na Educação (IAEd) (POZZEBON *et al.*, 2008). Uma das características dos STI's é serem capazes de representar

DOI: 10.5753/cbie.sbie.2020.1243

conhecimento de um especialista em uma determinada área de conhecimento, tendo a capacidade de resolver os problemas apresentados aos estudantes.

Os STI's no meio de aprendizagem possuem conteúdos que estão à disposição do estudante bem como avaliações, através de exercícios e problemas para serem resolvidos, visando coletar informações sobre a aprendizagem do estudante para assim adaptar sua estratégia de ensino. O MAZK é um tutor inteligente para ensino e aprendizagem de diversos domínios, no qual os professores poderão incluir os materiais e os estudantes poderão aprender sobre um determinado conteúdo exemplos, explanações e exercícios (MAZK, 2019).

O MAZK foi desenvolvido no Laboratório de Tecnologias Computacionais (LabTeC) da Universidade Federal de Santa Catarina (UFSC) e oferece quatro tipos de usuário: coordenador, o professor, o estudante e administrador (CAMARGO *et al.*, 2018).

A utilização de questões discursivas no processo de ensino permite ao professor, avaliar os processos cognitivos mais elevados quando comparado a aplicação de questões objetivas, além de incentivar as habilidades de comunicação e expressão do estudante (BURROWS; GUREVYCH; STEIN, 2015).

Uma solução para esse problema é a automatização da correção de questões discursivas. Para isso, será utilizado o Processamento de Linguagem Natural (PLN) que é uma área da Ciência da Computação dedicada a encontrar formas para os computadores serem capazes de analisar, reconhecer e/ou gerar textos em linguagens humanas, ou linguagens naturais (VIEIRA; LOPES, 2010). Sabendo que o PLN permite que o computador consiga processar dados em forma de textos, estudar suas técnicas para encontrar uma solução para automatização de questões discursivas se torna uma das opções mais viáveis.

Tendo em vista os impactos positivos e as técnicas disponíveis atualmente o presente estudo estabelece como problema de pesquisa: é possível realizar a correção automática de questões discursivas utilizando as medidas de similaridade entre textos provindas da PLN? Assim, o presente trabalho apresenta a implementação e a comparação da eficiência entre as medidas de similaridade estudadas.

2. Processamento de linguagem natural

O PLN vem sendo estudado e aplicado desde meados dos anos 50, quando houve a necessidade de se trabalhar com linguagem natural na comunicação com computadores. (SETZER, 2015). Contudo, umas das principais barreiras é a compreensão da língua humana de uma forma que os computadores consigam reconhecer e interpretar essa mesma linguagem (TERESO, 2019). As principais técnicas que a PLN utiliza para realizar esta interpretação serão detalhadas a seguir.

2.1. Pré-Processamento

Para que os dados sejam processados é necessário que eles sejam tratados. O pré-processamento consiste na preparação desses dados (SILVA, 2014). Esse procedimento é necessário, pois a linguagem natural é muito complexa, dessa forma são extraídas dos textos apenas as informações relevantes, que são tratadas a ponto de o computador conseguir processá-las.

Dentro deste pré-processamentos, algumas etapas são aplicadas como, por exemplo:

- *Tokenização* que transforma os textos em sequências de tokens, sendo estes responsáveis por realizar a segmentação do texto identificando e separando em unidades, podendo ser por caractere, palavra, sentenças, entre outros;
- Normalização que é utilizada para padronizar todos os dados de entrada, colocando todo o texto em letras minúsculas, removendo tokens indesejados como espaços em branco, caracteres especiais (ALVARENGA, 2019);
- Remoção de *stopwords*, ou seja, a remoção de palavras consideradas irrelevantes, que não contribuem para o significado geral da frase (HARDENIYA *et al.*, 2016).

Após o pré-processamento é necessário representar o conjunto de dados de um documento de forma que seja possível extrair informações. Entre os principais modelos que buscam realizar essa função, pode-se citar o modelo *word embedding* que busca representar textos em forma de vetor. A forma que ele realiza esse processo é representando uma palavra dentro de um espaço vetorial, no qual palavras com sentidos similares estão em posições próximas umas das outras (CORDEIRO, 2019).

Segundo Russel (2011), em um espaço vetorial multidimensional, cujos vetores representam documentos, a similaridade entre eles é representada pela distância entre esses vetores. Entre os modelos que realizam esta função são o *Cosine Similarity*, que é uma medida que utiliza o cosseno do ângulo entre vetores em uma representação vetorial realizando o cálculo. Quanto maior for a similaridade entre os dois vetores que estão sendo comparados, menor será o ângulo e consequentemente maior será o cosseno (VILLAÇA *et al.*, 2013).

Outro modelo é o *word mover's distance* (WMD), ele faz o cálculo da distância entre os documentos baseado na *Earth Mover's Distance* (EMD), que é a distância calculada sobre o espaço vetorial de palavras (RUBNER; TOMASI; GUIBAS, 1998). Resumidamente, o WMD mede a distância entre as palavras em um espaço vetorial, quanto menor a distância, mais próximos semanticamente elas estão.

As medidas de similaridade foram escolhidas seguindo as seguintes justificativas, ambas se beneficiariam do conceito de representação das palavras em um espaço vetorial da *word embedding*, sendo o *cosine similarity* uma das mais utilizadas. Outro ponto, é que existem poucos estudos sobre a eficiência do *Word Mover's Distance*, na aplicação de correção automática de questões discursivas, valendo assim o estudo para observar os resultados.

3. Metodologia - Proposta de Solução

Para que a solução proposta seja possível, será necessário que, quando o professor cadastrar seu material no MAZK, cadastre uma sugestão de resposta correta para cada questão discursiva presente nos exercícios, ou seja, o aluno responderá à questão discursiva que junto com a resposta do professor irão passar por um pré-processamento, sendo aplicadas a normalização, *tokenização* e remoção de *stopwords*. Logo em seguida as respostas serão enviadas para uma função de medida de similaridade, que fará consultas a *Word Embedding* para obter os valores semânticos de cada palavra que está sendo comparada. No final dos cálculos a função retornará a nota para o aluno, as quais podem variar de 0 a 100.

Contudo apenas um, dos dois, algoritmos de similaridade poderão ser implementado no sistema MAZK, visando não afetar no tempo de resposta do sistema e atrapalhar a

experiência dos usuários. Para isso foi realizado um pequeno teste de definição entre as medidas, para identificar qual delas obtém melhores resultados, e assim se tornar a medida oficial da solução proposta.

O teste de definição que foi aplicado em uma turma com 17 alunos e teve como objetivo criar um pequeno cenário com apenas uma questão discursiva, para analisar os comportamentos e resultados entre as medidas de similaridade realizado diretamente no ambiente MAZK. E por último foi realizado o teste principal que tem o objetivo de aplicar a solução diretamente no sistema e observar seu desempenho, não só dos resultados, mas também da experiência do professor utilizando a solução, aplicando no STI.

Serão utilizadas as técnicas de processamento de linguagem natural, como pré-processamento, vetores de características e medidas de similaridade como é possível observar na Figura 1.

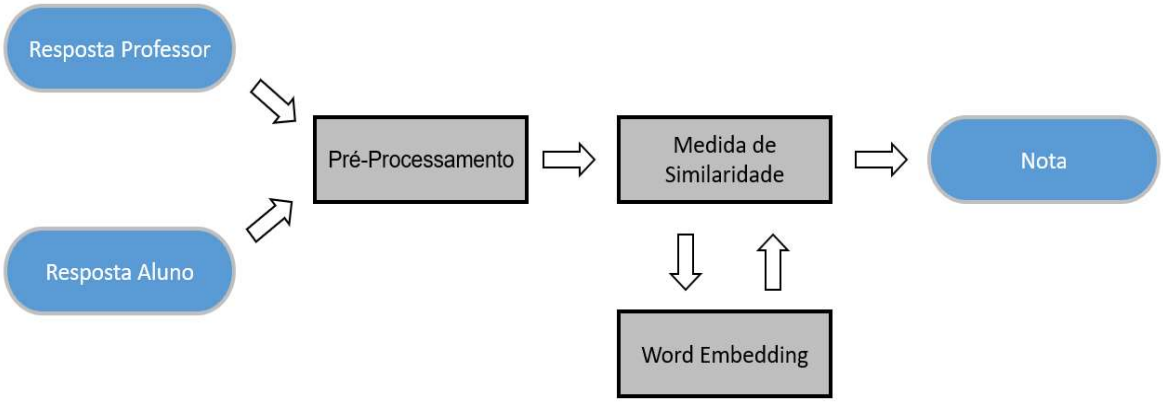

Figura 1. Fluxo da solução proposta

4. Resultados

Após a implantação do pré-processamento dos dados, das medidas de similaridade combinadas com o modelo *word embeddings* e utilizando a linguagem Python, a execução dos primeiros testes pode ser aplicada, realizando a correção automática de uma questão discursiva aplicada em uma turma.

4.1. Teste de definição

Foi solicitado a um professor usuário do STI MAZK, que disponibilizasse uma questão e resposta esperada para essa pergunta, junto com as respostas dos alunos e a nota que ele mesmo atribuiu para cada aluno. O cenário foi montado com a seguinte questão e resposta esperada do professor:

- *Questão: Defina Inteligência Artificial (IA)*
- *Resposta Esperada: A IA é um campo que busca construir entidades inteligentes, ou seja, ela procura sistematizar e automatizar as tarefas intelectuais. Basicamente, ela procura simular em uma máquina o comportamento humano.*

Com base nessas informações foram aplicadas as duas medidas de similaridades estudadas até o momento. Para otimizar a representação dos dados apresentados neste artigo, a Tabela 1 apresenta 5 dos 17 resultados.

Tabela 1. Apresentação das respostas do Teste de Definição

Aluno	Resposta	WMD	Cosine	Professor
1	Inteligência Artificial é a capacidade de um sistema computacional de realizar tarefas de maneira inteligente, eficiente e lógica, costumeiramente usando de base a inteligência humana.	49,5	83,5	90
2	É a capacidade da máquina realizar tarefas como um ser humano.	49,25	80	90
3	É a parte da computação que estuda e tenta desenvolver algoritmos que tenham características de pensamentos humanos.	45,2	78	85
4	Um campo da computação, que procura um modo de resolver problemas com possibilidades diversas, implementando em diversas áreas, usando softwares integrando em hardwares.	47,67	80,5	70
5	É a capacidade de uma máquina resolver determinados problemas que lhe são propostos através de um conhecimento adquirido através de padrões.	45	70	60

No teste aplicado foi possível observar que a resposta com maior e menor pontuação do professor, também foi a com maior pontuação das medidas de similaridade. Uma observação interessante é a resposta do segundo aluno, mesmo não tendo várias palavras parecidas com as utilizadas pelo professor as medidas deram uma maior classificação para ela, identificando que mesmo sem possuir o mesmo tamanho de resposta ou palavras parecidas elas estavam no mesmo contexto.

Mas da mesma forma que temos uma observação positiva é possível identificar que quando utilizamos muitas palavras voltadas a área da computação as medidas acabam classificando com boas notas é o caso do aluno 4, que mesmo não sendo direto na resposta da questão, utilizou palavras da área tendo uma boa pontuação, chegando a receber uma nota de 80,5 por uma das medidas.

Algumas conclusões podem ser avaliadas pelas informações encontradas, como a de que a medida de similaridade *word mover's distance*, não teve um bom desempenho no teste realizado, já o *cosine similar* obteve melhores resultados se comparados com o WMD, chegando a ficar próximo das respostas do professor. Sendo assim escolhida para a implementação na plataforma MAZK e realização do teste principal no sistema.

4.2. Teste principal

Para que o teste fosse viável foi necessário realizar um repasse de conhecimento para a equipe de desenvolvimento do MAZK, no qual foi apresentada a solução e seu funcionamento no ponto de vista técnico. Após esta etapa, foi confirmada a viabilidade da aplicação e se deu início a implantação da solução no sistema para realização do teste.

Com o sistema em funcionamento o cenário de teste foi criado; a turma da disciplina de inteligência artificial do curso de Tecnologia da Informação e Comunicação do semestre 2019/2 foi escolhida, obtendo cerca de 80 questões respondidas.

Primeiro os alunos tiveram acesso ao material disponibilizado pelo professor na sala criada e logo em seguida uma série de exercícios foram aplicados, entre eles algumas questões discursivas. Serão apresentadas as questões, algumas respostas selecionadas de forma aleatória e notas selecionadas para discussão.

- *Questão 1: Qual é a diferença que existe entre aprendizado supervisionado e não supervisionado?*
- *Resposta Esperada: Aprendizado supervisionado: A rede neural recebe um conjunto de dados de entrada e seus padrões de saída correspondentes. Por isso ele é supervisionado. Aprendizado não supervisionado: Trabalha os dados de maneira a determinar algumas propriedades dos conjuntos de dados. Não existe para cada entrada uma saída desejada.*

A Tabela 2 mostra os cinco resultados selecionados. É possível analisar que o sistema, assim como os professores, atribuiu boas notas para os alunos 1 e 2, mas com uma diferença considerável. Com os alunos 3 e 5, as notas do sistema e do professor se assemelharam muito, o que não foi o caso do aluno 4.

Tabela 2. Respostas da Primeira Questão do Teste Principal

Aluno	Resposta	Nota Sistema	Nota Professor
1	Supervisionado é quando a rede neural recebe um conjunto de dados de entrada e seus padrões de saída correspondentes. O não supervisionado é quando não existe para cada entrada uma saída desejada.	91	100
2	Supervisionado: Quando há um "professor" intermediando o aprendizado, especificando se ele está correto ou errado. Não Supervisionado: é quando para fazer modificações nos valores das conexões sinápticas não se usa informações sobre se a resposta da rede foi correta ou não no caso não existe para cada entrada uma saída desejada.	86	100
3	No supervisionado, o professor indica o bom e o mal comportamento. A rede neural recebe conjunto de dados de entrada e padrões de saída correspondente. Não supervisionado, para fazer modificações nos valores das conexões sinápticos sem informações sobre resposta da rede correta ou não.	91	90
4	Supervisionado o professor indica explicitamente um comportamento bom ou ruim já o não supervisionado é quando para fazer modificações nos valores das conexões sinápticas não se usa informações sobre se a resposta da rede foi correta ou não.	79	50
5	Supervisionado: no qual um professor indica qual o resultado esperado ou saída esperada. Não supervisionado: Não se sabe-se se o resultado obtido é mesmo o correto.	49	50

Nesta primeira questão o sistema continua tendo características do teste de definição, acompanhando o professor e avaliando positivamente e negativamente os alunos, mas ampliando a quantidade de resposta algumas diferenças consideráveis começam a aparecer.

- *Questão 2: O que são sinapses?*
- *Resposta Esperada: As sinapses são junções entre a terminação de um neurônio e a membrana de outro neurônio. São elas que fazem a conexão entre células vizinhas, dando continuidade à propagação do impulso nervoso por toda a rede neuronal.*

A tabela 3 mostra as respostas para a questão 2, no qual alguns padrões continuam se repetindo, como a maior classificação e a menor, também ocorreu um acerto sem erro entre a nota do professor e do sistema. Outro ponto que se repetiu, esse de forma negativa, é a diferença entre as notas que aumentou em alguns casos, como a dos alunos 1,2 e 4.

Tabela 3. Respostas da Segunda Questão do Teste Principal

Aluno	Resposta	Nota Sistema	Nota Professor
1	Sinapses é a parte do neurônio que conecta as extremidades do axônio com os dendritos de outros neurônios, que servem para memorização e armazenamento da informação.	85	100
2	São onde agem os neurotransmissores, que transmitem os impulsos entre neurônios, ou seja, por onde os neurônios conversam.	71	100
3	As sinapses têm um papel fundamental na memorização da informação e são principalmente as do córtex cerebral e algumas vezes de partes mais profundas do cérebro que armazenam esta informação.	80	80
4	Atividade inibitória que previne a excitação do neurônio	81	50
5	São conexões que memorizam informações e interconectam todos os neurônios	62	50

O caso do aluno 4 é um dos mais alarmantes, pois o sistema elevou de forma considerável a nota comparada com a do professor, isso é resultado da proximidade das palavras no espaço vetorial e um ponto negativo para ser observado.

- *Questão 3: Cite as funções mais comumente utilizadas como funções de saída em neurônios?*
- *Resposta Esperada: A Função Linear, A Função Sigmoidal ou Logística e A Função Tangente Hiperbólica.*

Continuando, a tabela 4 apresenta as respostas referentes a questão 3. Esses resultados se diferenciam, pois nesse caso alguns alunos usaram fórmulas junto com a resposta que por serem consideradas como caracteres especiais são eliminadas no pré-processamento, levando em consideração apenas o texto.

Tabela 4. Respostas da Terceira Questão do Teste Principal

Aluno	Resposta	Nota Sistema	Nota Professor
1	A Função Linear, A Função Logística ou Sigmoidal e a Função Tangente Hiperbólica.	100	100
2	Função Linear, Função Sigmoidal e Função Tangente Hiperbólica.	98,6	100
3	Essencialmente, qualquer função contínua e monotômica crescente, tal que x E R e y(x) E [-1,1], pode ser utilizada como função de saída na modelagem neural. Alguns exemplos são: â€¢ Função Linear y(x) = ax, Função Sigmoidal ou Logística y(x) = 1 / (1 + e-kx), Função Tangente Hiperbólica y(x) = tanh(kx) = (1 - e-kx) / (1 + e-kx).	87	100
4	Linear, Sigmoidal ou Logística e Tangente Hiperbólica.	89	90
5	Função continua ou monotônica crescente.	62	30

Este é o caso do aluno 3, que mesmo sendo muito específico na sua resposta, o sistema não conseguiu identificar isso. É importante ressaltar que o caso do aluno 4 da questão anterior, se repete nesta questão, com o aluno 5.

- *Questão 4: Quais são os elementos básicos de um Neurônio Artificial?*
- *Resposta Esperada: Entradas, A Combinação das Entradas - O "Net", Função de Ativação (fa) e Função de Saída (fs).*

Dentre as respostas da questão 4, um fato curioso pode ser destacado, ou seja, as respostas dos alunos 1 e 2, mesmo sendo muito parecidas o sistema só considera a questão que está idêntica a esperada pelo professor com nota máxima. Este fato pode ser observado na Tabela 5 a seguir.

Tabela 5. Respostas da Quarta Questão do Teste Principal

Aluno	Resposta	Nota Sistema	Nota Professor
1	Entradas, a Combinação das Entradas - O "Net", Função de Ativação (fa), Função de Saída (fs)	100	100
2	Entradas, a combinação de entradas ou o "Net", a função de ativação e a função de saída	96	100
3	Entradas - Pesos Sinápticos - Função Soma - Função Transferência – Saída	88,61	100
4	Entradas, Pesos, Bias, somador, funções de ativação, Saídas.	71,75	100
5	Entrada, somatório, saída	60	50

- *Questão 5: Onde está armazenado o conhecimento de uma rede neural?*
- *Resposta Esperada: Nos valores das suas conexões sinápticas.*

Na questão 5, os resultados seguem o padrão das questões anteriores, mas vale a pena destacar o aluno 5. Ele acertou a questão, mas algumas palavras estavam um pouco mais afastadas no espaço vetorial, afetando o cálculo do sistema.

Tabela 6. Respostas dos Alunos da Quinta Questão do Teste Principal

Aluno	Resposta	Nota Sistema	Nota Professor
1	Nos valores das suas conexões sinápticas.	100	100
2	Está armazenado nos valores das suas conexões sinápticas.	92,45	100
3	Está armazenando nos valores de suas conexões.	87,6	100
4	O conhecimento de uma rede neural está armazenado nos valores das suas conexões sinápticas.	84,5	100
5	Estão dispersados por toda a rede através dos neurônios e seus pesos sinápticos.	64	100

O teste principal, busca analisar o comportamento da solução em uma escala maior, além do seu funcionamento dentro do STI MAZK, encontrando assim características que não ficaram visíveis no primeiro teste. Nos resultados dos cálculos a combinação com menor erro foi a medida de similaridade *cosine similarity* obtendo um valor percentual de 15%.

Em uma visão geral, os testes revelaram várias características com relação ao desempenho das medidas de similaridade. A primeira é que na maioria dos casos as avaliações do professor e do sistema convergiam quanto a classificação, das melhores para as piores. A segunda é que mesmo convergindo, existiam algumas diferenças entre as notas e que uma resposta só será classificada com a nota máxima se estiver idêntica a resposta esperada pelo professor.

Analisando as tabelas é possível observar que ao mesmo tempo que a solução classificava as notas da mesma forma que o professor, no sentido de maior para a menor, existia uma diferença entre as notas na grande parte das vezes, sendo umas menores, mas outras muito grandes.

CONSIDERAÇÕES FINAIS

Este trabalho teve como objetivo desenvolver uma proposta de solução capaz de realizar a correção automática de questões discursivas no STI Mazk utilizando as medidas de similaridade entre textos. A solução visava auxiliar os professores, usuário do Mazk, nesta atividade que muitas vezes acaba tomando muito do seu tempo, além de trazer para o aluno, em implementações futuras, um feedback quase que instantâneo dos seus resultados

A solução proposta utilizava a comparação entre as respostas do aluno com a resposta esperada do professor, que seria cadastrada junto com cada pergunta. Ambas as respostas passam pelo pré-processamento e em seguida é calculada a medida de similaridade entre elas, por fim é retornada a nota do aluno.

Ao final, foi possível atingir o objetivo geral deste trabalho, com a construção de um corretor automático de questões discursivas no STI Mazk, mostrando através de testes que a utilização do *Cosine.* Porém sua eficiência precisa ser melhorada, o erro relativo ainda é alto e quando se trata de avaliar um aluno, uma pequena porcentagem de erro pode reprová-lo, mesmo merecendo a aprovação ou aprová-lo, mesmo merecendo a reprovação.

No ponto de vista do professor, com relação a usabilidade e resposta do sistema, houve um retorno positivo, destacando que o sistema respondia de forma rápida e não afetava as outras funcionalidades do STI, além de demonstrar seu apoio aos estudos na busca da automatização da correção de questões discursivas. O objetivo foi alcançado, vários comportamentos diferentes foram encontrados, alguns bons outros ruins, mas todos necessários para avaliar a eficiência da solução.

Já na análise dos dados o resultado não foi positivo, houve um aumento considerável no percentual do erro relativo, que no primeiro teste era de 7% e no segundo cenário de teste foi para 15%, mesmo tendo resultados positivos a diferença entre a avaliação do professor e do aluno, em alguns casos, era consideravelmente alta. Para diminuir o erro relativo é possível avaliar algumas soluções, como por exemplo estudar outras medidas de similaridade

REFERÊNCIAS

ALVARENGA, João Paulo Reis. Avaliação de métodos de transferência de aprendizado aplicados a problemas de processamento de linguagem natural em textos da língua portuguesa. 2019.

BURROWS, Steven; GUREVYCH, Iryna; STEIN, Benno. The eras and trends of automatic short answer grading. International Journal of Artificial Intelligence in Education, v. 25, n. 1, p. 60-117, 2015.

CORDEIRO, Bernardo Cardoso. BERT E WORD2VEC: UMA ANÁLISE INFERENCIAL E COMPUTACIONAL NA CLASSIFICACAO DE TEXTOS COM REDES NEURAIS CONVOLUCIONAIS. 2019. Tese de Doutorado. Universidade Federal do Rio de Janeiro.

PASSERO, Guilherme; HAENDCHEN FILHO, Aluizio; DAZZI, Rudimar. Avaliaçao do uso de métodos baseados em lsa e wordnet para correçao de questoes discursivas. In: Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE). 2016. p. 1136.

POZZEBON, Eliane. Um modelo para suporte ao aprendizado em grupo em sistemas tutores inteligentes. Tese de doutorado do PGEEL, UFSC, 2008.

RUBNER, Yossi; TOMASI, Carlo; GUIBAS, Leonidas J. A metric for distributions with applications to image databases. In: Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271). IEEE, 1998. p. 59-66.

RUSSELL, Matthew A. Mining the social web: Analyzing data from Facebook, Twitter, LinkedIn, and other social media sites. " O'Reilly Media, Inc.", 2011.

SILVA, M. d AO. Pré-Processamento em Mineração de Dados como método de suporte à modelagem algorítmica. 2014. Tese de Doutorado. Curso de Pós-Graduação em Modelagem Computacional de Sistemas da Universidade Federal do Tocantins, Palmas.

TERESO, Marco. ANÁLISE DE SENTIMENTO A COMPANHIAS AÉREAS NORTE AMERICANAS. ISLA Multidisciplinary e-Journal, v. 2, n. 1, p. 52-65, 2019.

VIEIRA, Renata; LOPES, Lucelene. PROCESSAMENTO DE LINGUAGEM NATURAL E O TRATAMENTO COMPUTACIONAL DE LINGUAGENS CIENTÍFICAS. EM CORPORA, p. 183, 2010.

VILLAÇA, Rodolfo da Silva et al. Hamming DHTe HCube: arquiteturas distribuídas para busca por similaridade. 2013.