



Texto Livre: Linguagem e Tecnologia  
E-ISSN: 1983-3652  
revista@textolivre.org  
Universidade Federal de Minas Gerais  
Brasil

Barreto Moreira, Leonard; Del Real Tamariz, Annabell; Vieira Fettermann, Joyce  
O uso da mineração de textos no suporte a correções de questões discursivas em uma  
instituição de educação superior  
Texto Livre: Linguagem e Tecnologia, vol. 11, núm. 3, september-december, 2018, pp.  
213-227  
Universidade Federal de Minas Gerais

Disponível em: <https://www.redalyc.org/articulo.oa?id=577163619012>

- Como citar este artigo
- Número completo
- Mais artigos
- Home da revista no Redalyc

redalyc.org

Sistema de Informação Científica  
Rede de Revistas Científicas da América Latina, Caribe, Espanha e Portugal  
Projeto acadêmico sem fins lucrativos desenvolvido no âmbito da iniciativa Acesso Aberto

## O USO DA MINERAÇÃO DE TEXTOS NO SUPORTE A CORREÇÕES DE QUESTÕES DISCURSIVAS EM UMA INSTITUIÇÃO DE EDUCAÇÃO SUPERIOR

### THE USE OF TEXTS MINING IN THE SUPPORT TO CORRECTIONS OF DISCURSIVE QUESTIONS IN A HIGHER EDUCATION INSTITUTION

Leonard Barreto Moreira

Universidade Estadual do Norte Fluminense

[leonardbarreto@gmail.com](mailto:leonardbarreto@gmail.com)

Annabell Del Real Tamariz

Universidade Estadual do Norte Fluminense

[annabell@uenf.br](mailto:annabell@uenf.br)

Joyce Vieira Fettermann

Universidade Estadual do Norte Fluminense

[joycejvieira@gmail.com](mailto:joycejvieira@gmail.com)

**RESUMO:** A presente pesquisa tem como objetivo principal o desenvolvimento de um modelo computacional com uso de técnicas de Mineração de Textos para a tarefa de correção de questões dissertativas em ambientes *online*, possibilitando, por sua vez, a diminuição da subjetividade na avaliação das questões discursivas dos discentes. O conjunto de dados utilizados para os experimentos baseia-se em 15 questões discursivas de computação pertencentes ao ciclo básico de cursos da área das Engenharias. A metodologia proposta é apoiada em três grandes fases: 1) Aplicação de técnicas de pré-processamento de textos e representação dos documentos segundo a abordagem “Saco de palavras”, com esquema de ponderação *term-frequency*; 2) Realização do processamento dos textos por meio da comparação dos termos contidos nas respostas com os do gabarito por intermédio de medidas baseadas em termos e edição; 3) Confrontação dos resultados numéricos obtidos com as notas da correção do avaliador, ao investigar a hipótese de que as médias das notas reais e estimadas são iguais por meio do Teste T, assim como análise do erro médio absoluto percentual (MAPE) entre tais subconjuntos. Os resultados obtidos indicaram uma alta aderência à hipótese de que as médias dos dados reais vs estimados são iguais, principalmente para as medidas baseadas em *tokens*. A acurácia foi da ordem de 84,2% para Coseno no modelo *bigram*. Assim, o principal resultado deste trabalho é a concepção de um modelo de MT para o apoio ao processo avaliativo de questões discursivas em ambiente EaD.

**PALAVRAS-CHAVE:** aprendizado de máquina; mineração de texto; sistemas de ensino inteligentes.

**ABSTRACT:** The present research has as main objective the computational development with the use of techniques of Texts Mining for the task of correcting the dissertative questions online, making it possible to provide the diminution of the subjectivity in the evaluation of the discursive questions of the students. The set of data used for the experiments is based on 15 discursive computational questions belonging to the basic

course cycle of the Engineering area. The proposed methodology is supported by three major phases: 1) Application of pre-processing techniques and representation of documents according to the “Bag of words” approach, with term-frequency weighting scheme; 2) Carrying out the processing of texts by comparing the terms contained in the answers with those of the template by means of measures based on terms and editing; 3) Confrontation of the numerical results obtained with the evaluator's correction notes, investigating the hypothesis that the means of the real and estimated scores are equal by means of the T-Test, as well as analysis of the percentage absolute mean error (MAPE, in Portuguese) between such subsets. The results obtained indicated a high adherence to the hypothesis that the averages of the actual vs. estimated data are the same, especially for the tokens-based measurements. The accuracy was of the order of 84.2% for Cosine in the bigram model. Thus, the main result of this work is the design of a TM model to support the evaluation process of discursive issues in the distance learning environment.

**KEYWORDS:** machine learning; text mining; smart education systems.

## 1 Introdução

Em meados da década de 1970, o emprego dos computadores no espaço escolar de vários países foi um marco no uso de tecnologias na educação, originando assim um termo comumente referenciado como *computadores na educação*. Com a popularização e o uso das redes de computadores pelas escolas, em particular a Internet, um novo termo foi cunhado: Tecnologias da Informação e Comunicação (TIC). Em suma, as TIC remetem à capacidade de sistemas em manipular e distribuir informações por meio de uma pluralidade de tecnologias (ANDERSON, 2010).

No Brasil, as TIC, no domínio da Educação a Distância (EaD), têm sido um importante agente facilitador no contexto das ações pedagógicas voltadas à aprendizagem. Integrada ao sistema de educação formal no ano de 1996, a modalidade EaD surgiu como uma alternativa de diversificação do sistema de ensino nacional (GOMES, 2013; SOARES-LEITE; NASCIMENTO-RIBEIRO, 2012). A partir desse ano, uma série de iniciativas públicas e privadas foram observadas para a inserção das TIC no cotidiano escolar, seja sob o discurso de democratização do ensino ou através da expansão de cursos sob a ótica da otimização dos lucros, respectivamente (GOMES, 2013).

Historicamente, a avaliação da aprendizagem se caracteriza como um dos aspectos mais desafiadores da prática pedagógica, com dificuldades que vão desde sua concepção até a elaboração dos mais simples instrumentos (D'ANTOLA, 1992). Conceitualmente, a avaliação pode ser definida como apreciação qualitativa sobre os dados relevantes do processo de ensino e aprendizagem, que auxilia o professor a tomar decisões sobre o seu trabalho. A apreciação qualitativa desses dados, mediante a análise de provas, exercícios, trabalhos, respostas dos alunos, realização de tarefas, entre outros, permite uma tomada de decisão para o que deve ser feito (LUCKESI, 1996). Assim, a incorporação de novas tecnologias apresenta-se como um ferramental promissor no que tange às novas formas de avaliação da aprendizagem, abrindo novas perspectivas para essa área de conhecimento.

A crescente adoção de sistemas informatizados no âmbito da educação,

particularmente relacionados à modalidade de ensino a distância (EaD), impõe uma série de desafios no processo de ensino-aprendizagem. O estudo feito pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) sobre dados relacionados ao Censo da Educação Superior de 2016 indicou um crescimento do número de estudantes da ordem de 7,2% (INEP, 2016) em comparação com o levantamento anterior, refletindo, assim, uma tendência internacional (FREITAS *et al.*, 2018; SHESHASAYEE; NAZREEN BEE, 2018).

A oferta e o aumento de instrumentos educativos, bem como os bancos de dados com informações sobre os estudantes, têm criado grandes repositórios de dados, refletindo como os alunos aprendem e potencializando, assim, esse problema (BAKER, 2010; ROMERO; VENTURA, 2007), abrindo novas perspectivas de pesquisa nessa área. Os processos pedagógicos relativos a métodos avaliativos, mais particularmente associados com questões discursivas, surgem como uma área promissora para a aplicação de métodos computacionais voltados ao tratamento e à interpretação de tais informações.

Na história moderna, o desenvolvimento de estratégias computacionais inteligentes tem sido objeto de pesquisa nos mais variados campos de conhecimento. Cunhado na década de 1950, o termo “Inteligência Artificial” (IA) pode ser considerado um precursor das atuais técnicas de mineração de dados (MD), cuja tarefa principal concentra-se em sistematizar e automatizar tarefas intelectuais. Na área da educação, a aplicação de MD em dados educacionais originou o termo conhecido como *Educational Data Mining* (EDM), que tem como objetivo principal o desenvolvimento de métodos para explorar os tipos exclusivos de dados em ambientes educacionais e, usando esses métodos, propiciar um melhor entendimento dos alunos e dos ambientes de aprendizagem utilizados por eles (ROMERO; VENTURA, 2013).

As novas tecnologias da Web 3.0, também chamadas de web semântica, por meio da análise de textos e objetos na Internet, juntamente com o seu significado, elevam o processo de ensino-aprendizagem a um novo patamar (PRABHA; SHANAVAS, 2014). Recursos como filtragem colaborativa inteligente, computação em nuvem, *big data*, dados vinculados, interoperabilidade e mobilidade inteligente são algumas terminologias desse novo ecossistema, um ambiente que preconiza o desenvolvimento de tecnologias que permitem ao computador compartilhar e manipular as informações contidas na web de forma adequada e inteligente (HUSSAIN, 2012; KURILOVAS; KUBILINSKIENE; DAGIENE, 2014). No contexto educacional, tal característica pode ser observada pelos trabalhos de Anderson (2016), Kurilovas; Kubilinskiene; Dagiene (2014) e Temple *et al.* (2018).

Em relação às avaliações de dados em formato livre, as características inerentes aos textos impõem uma série de desafios para lidar com informações não estruturadas. A riqueza dos textos, sua dimensionalidade, além da complexidade inerente à linguagem humana, podem ser citadas como os principais aspectos envolvidos na tarefa de uniformização textual. De caráter interdisciplinar, a Mineração de Textos (MT) é um campo de conhecimento fundamentado principalmente nas áreas relacionadas à recuperação da informação, aprendizagem de máquina, mineração de dados, estatística e linguística computacional (HAN; KAMBER; PEI, 2011). Dentre os algoritmos de aprendizagem de máquina, destacam-se os métodos e as técnicas de Processamento de Linguagem Natural (PLN), que se baseiam na interpretação, tradução e representação de alguma

língua natural que possa ser utilizada pelo computador, e as técnicas da área de Recuperação de Informação (RI) para a redução de dimensionalidade, representação e ponderação de documentos (BAEZA-YATES; RIBEIRO-NETO, 2013).

Dentre as tarefas típicas de MT, destacam-se as relacionadas ao agrupamento de textos, que, de forma resumida, tentam avaliar o grau de similaridade (ou dissimilaridade) em função de alguma medida de distância. A medição da similaridade entre palavras, frases, parágrafos e documentos é um componente importante em várias tarefas, como o agrupamento de documentos. Comumente, essa é uma etapa usada como um estágio primário para avaliação de similaridade de frases, parágrafos e documentos (SRIVASTAVA; SAHAMI, 2009; WEISS; INDURKHYA; ZHANG, 2010), geralmente precedida por uma série de técnicas da área de PLN para o tratamento sintático e semântico dos textos (AGGARWAL; ZHAI, 2012; YANG; PEDERSEN, 1997). Adicionalmente, os modelos de MT podem se beneficiar do uso de ontologias, ou seja, de conceitos específicos de um domínio de conhecimento, por meio de modelos que fornecem a estrutura necessária para a representação semântica da informação textual (FODEH; PUNCH; TAN, 2011; MANNING; SCHÜTZE, 1999).

Atualmente, com a informatização, diversos sistemas manipulam e armazenam grandes quantidades de dados em formato digital. Em larga escala, torna-se inviável sua análise através de métodos manuais convencionais, tornando-se necessária uma alta disponibilidade de tempo e, ainda, estando sujeitos a diferentes interpretações por parte do analisador (KNAFLIC, 2015). Dessa forma, é necessária alguma técnica e/ou ferramenta capaz de auxiliar os especialistas na tarefa de analisar em tempo útil esses dados. Assim, diversos segmentos da sociedade podem se beneficiar das técnicas de MT para a descoberta de informações em textos, possibilitando, dessa forma, a extração de padrões para o auxílio no processo de tomada de decisão (WEISS; INDURKHYA; ZHANG, 2010).

Diante desse cenário, a presente pesquisa consiste na aplicação de conhecimentos do campo de Mineração de Dados, mais especificamente da subárea de Mineração de Textos, na tarefa de correção de questões dissertativas em ambientes EaD, e tem por objetivo principal a diminuição da subjetividade na avaliação das questões discursivas dos discentes. Complementarmente, espera-se que a adoção de agentes inteligentes propicie agilidade e uniformidade nas correções dos tipos de questão mencionados na referida pesquisa. Tendo em vista que uma maior frequência de questões discursivas em processos avaliativos é desejada, aumenta-se dessa forma a qualidade em todas as etapas do processo pedagógico.

Considerando uma vasta quantidade de fatores limitantes no âmbito dos problemas relacionados à manipulação com textos, é de suma importância a identificação dos fatores e características que podem ser determinantes para a otimização no trabalho dos docentes de correção de avaliações discursivas, possibilitando, assim, ampliar o uso de tal tipo de questão no processo de avaliação do processo de ensino e aprendizagem. Dada a complexidade e a quantidade de restrições impostas pelo problema, será proposto um modelo computacional para suporte à tomada de decisão, valendo-se de técnicas de Mineração de Dados, especificamente relacionados à Mineração de Textos para a temática em questão. A presente pesquisa justifica-se diante da atualidade do tema e da necessidade de reflexão sobre as consequências socioeconômicas da correlação suscitada.

Dessa forma, este artigo está organizado da seguinte maneira: na próxima seção, é descrita a metodologia utilizada para o processo de mineração, bem como o conjunto de dados utilizados nos experimentos; a terceira seção apresenta as métricas utilizadas para avaliação dos resultados, os pressupostos de aderência, ou seja, em que medida os dados estimados têm comportamento estatístico semelhante aos dados originais, e os respectivos valores para cada uma das abordagens tratadas na pesquisa; e, por fim, na seção seguinte, são tecidas as conclusões do trabalho realizado e perspectivas de estudos futuros.

## 2 Metodologia

A pesquisa foi realizada tendo como suporte as informações de instrumentos avaliativos relacionadas a questões discursivas de discentes e os respectivos gabaritos produzidos pelos docentes, a partir de registros armazenados eletronicamente. Dentre os elementos observados, seu foco será restrito à avaliação de questões dissertativas contidas em um banco de questões (BDQ) de uma disciplina de computação do ciclo básico de engenharias de uma empresa do ramo de ensino superior no município de Campos dos Goytacazes. Ademais, o processo de mineração não contempla o peso de critérios ortográficos e gramaticais, comumente adotados pelos avaliadores, na composição da nota. Cabe salientar que o modelo proposto não realiza análise em códigos e fórmulas matemáticas.

Assim, serão avaliados diversos aspectos de suma importância para o problema em questão, compondo os parâmetros do modelo computacional a ser desenvolvido com apoio de técnicas de Mineração de Textos (MT), através de rotinas desenvolvidas e disponíveis na ferramenta *Python*.

Na etapa relacionada ao pré-processamento de textos, foram aplicados métodos e técnicas de Processamento de Linguagem Natural (PLN), que se baseiam na interpretação, tradução e representação de alguma língua natural que possa ser utilizada pelo computador.

Utilizou-se nesta fase um conjunto de tarefas da biblioteca NLTK (PERKINS, 2014) do *Python*. No primeiro momento, realizou-se a conversão dos textos para caixa baixa (*case folding*), bem como a substituição de siglas, abreviações e acrônimos por seus respectivos termos. O procedimento relacionado à substituição de sinônimos (*thesaurus*) teve como base um vocabulário estruturado denominado *OpenWordNet-PT* (DE PAIVA; RADEMAKER; DE MELO, 2012). O passo seguinte foi altamente dependente do conjunto de medidas de distâncias adotadas. Para as medidas baseadas em edição (*edit-based*), os textos já se encontravam preparados para o processamento, não necessitando, portanto, de nenhuma adequação textual. Já nas métricas baseadas em *tokens*, foi aplicada uma série de tarefas para normalização textual, tal como a remoção de pontuações, caracteres especiais e palavras irrelevantes. A lista de palavras consideradas irrelevantes para o processo de mineração, *stopword lists* ou *stop lists*, foi obtida a partir do dicionário na língua portuguesa disponível no pacote NLTK. A redução dos termos aos respectivos radicais (*stemming*) foi possível com o uso do algoritmo de *stemming* de Porter (PORTER, 1980), mais especificamente pela sua adaptação à língua portuguesa. Em seguida, as palavras consideradas para a etapa de mineração foram então agrupadas



em *tokens* com uso do algoritmo *ngrams* do pacote *scikit-learn* do *Python*. Um dos parâmetros importantes nesse algoritmo é relacionado ao número mínimo e máximo de termos, configurado entre 2 e 5 *tokens*, respectivamente.

Na fase de transformação, com o apoio de técnicas Recuperação da Informação (RI), os documentos foram representados por uma abordagem conhecida como “Saco de palavras” (*Bag of words*), que consiste, essencialmente, na transformação do conjunto de documentos em um vetor de características por meio da divisão de cada documento em palavras ou termos (BAEZA-YATES; RIBEIRO-NETO, 2013). Visto que se buscou avaliar a similaridade das respostas com o gabarito, o esquema de ponderação de termos utilizado foi a Frequência de Termos (*Term Frequency* – TF).

Na etapa intitulada processamento, buscou-se investigar o uso das métricas de distância baseadas em edição e termos, de modo a identificar a medida mais representativa para a coleção. Deste primeiro conjunto, avaliou-se a similaridade dos textos de acordo com as medidas de distâncias Jaro (JARO, 1989), Damerau (DAMERAU, 1964) e Levenshtein (LEVENSHTEIN, 1966). As métricas Coseno (Cosine), Euclideana e Manhattan foram utilizadas para estimar a similaridade em função de cada esquema de divisão de termos (*ngram*). Tal etapa mostrou-se, particularmente, desafiadora, tendo em vista a pluralidade de métricas, muitas altamente dependentes dos objetivos e da dimensionalidade do problema.

Por fim, de maneira a assegurar, estatisticamente, a qualidade dos resultados obtidos a partir dos modelos previamente estabelecidos, a avaliação foi apoiada em técnicas da estatística inferencial, mais especificamente o Teste T e o erro percentual absoluto médio (*Mean Absolute Percentage Error* – MAPE), nos subconjuntos de dados normalizados. Devido às suas vantagens de independência de escalas e interpretabilidade, a MAPE tem sido considerada como uma das medidas de precisão de previsão mais amplamente utilizadas (KIM; KIM, 2016), motivo pelo qual foi adotada nesta pesquisa. Dessa forma, buscou-se verificar a qualidade das notas estimadas pelas diversas métricas em função do gabarito, possibilitando determinar a melhor métrica para a composição da “nota”.

A Figura 1 sintetiza as tarefas para a descoberta de conhecimento em textos (esquerda) e as respectivas tarefas (direita) em cada uma das etapas.

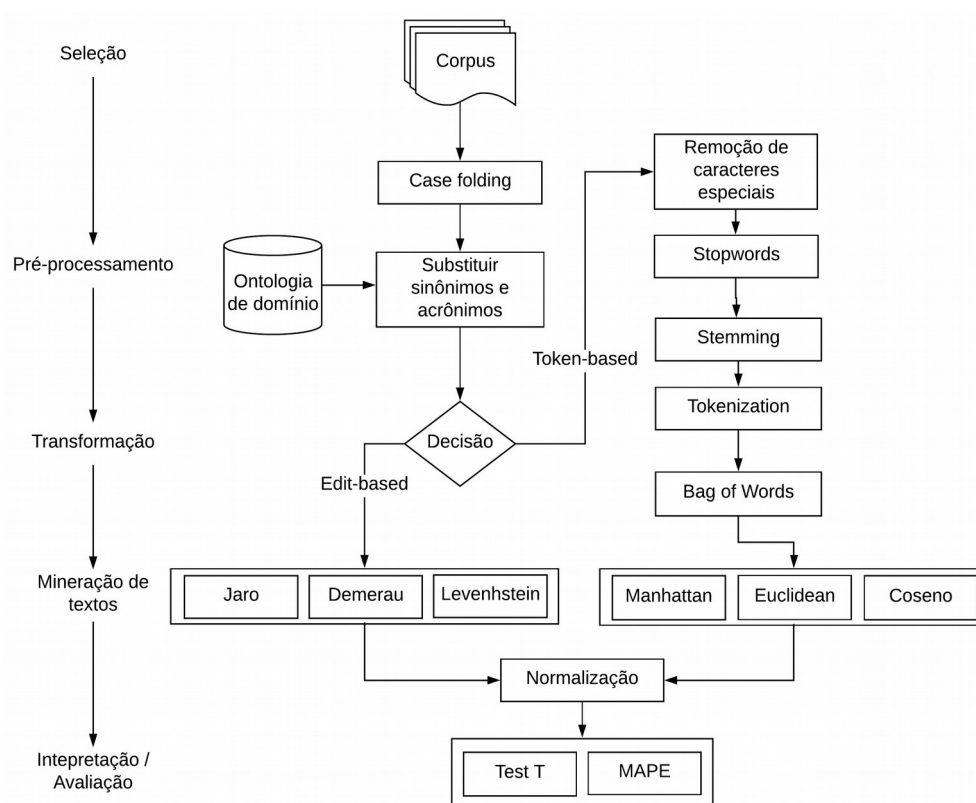


Figura 1: Metodologia de MT proposta para avaliação das questões dissertativas  
Fonte: Os autores (2018)

Assim, serão apresentadas diferentes soluções para o problema, permitindo que, dentre elas, a melhor do ponto de vista metodológico seja escolhida.

### 3 Resultados

Conforme previamente mencionado, os experimentos foram conduzidos com um subconjunto de 15 respostas e o respectivo gabarito. Em função do tipo de medida de distância utilizada, realizou-se um pré-processamento dos textos para fins de normalização textual. A título de ilustração, o quadro a seguir apresenta a resposta original (1) e a resultante (2) do pré-processamento. O resultado de cada tarefa pode ser visualizado através dos termos sucedidos por uma sequência numérica entre parênteses. Por exemplo, o termo processado “exist<sup>(1)</sup>” (Quadro 1 (2)) é resultante da aplicação da etapa referente à “Remoção de caracteres especiais” à palavra “Existem<sup>(1)</sup>”, contida na resposta original (Quadro 1(1)).

Quadro 1: Texto (2) resultante do pré-processamento aplicado a uma resposta (1) do conjunto de dados.

1	O desvio no programa ocorre sempre que o programa sai da sequência normal de execução. Existem <sup>(1)</sup> dois tipos de desvio: incondicional e condicional. Desvio incondicional: ocorre quando o programa sai da sequência normal de execução <sup>(2)</sup> e é desviado para <sup>(3)</sup> outro ponto da memória do programa independente <sup>(4)</sup> de
---	---



	qualquer condição. Desvio condicional: uma condição deve ser satisfeita para o desvio ocorrer.
2	desvi program ocorr sempr program sai sequenc norm execuca exist <sup>(1)</sup> doi tip desvi incondic condic desvi incondic ocorr program sai sequenc norm execuca <sup>(2)</sup> desvi <sup>(3)</sup> outr pont memor program independ <sup>(4)</sup> qualqu condica desvi condica condica dev satisfeit desvi

Notas: Marcadores e respectivas tarefas: <sup>(1)</sup> Case Folding; <sup>(2)</sup> Remoção de caracteres especiais; <sup>(3)</sup> Stopword; <sup>(4)</sup> Stemming

Fonte: Os autores (2018).

Posteriormente, após processados, os resultados obtidos foram avaliados mediante análise da variância média e do erro médio absoluto percentual.

O Quadro 2 resume as métricas utilizadas para comparar a distância entre os textos, bem como o domínio dos dados e as respectivas faixas e melhores valores. Para mais detalhes sobre as métricas, consultar: Aggarwal; Zhai (2012), Baeza-Yates; Ribeiro-Neto (2013), Srivastava; Sahami (2009).

Quadro 2: Métricas, domínios, intervalos e melhores valores de cada medida utilizada na pesquisa.

Métrica	Domínio	Intervalo	Melhor valor
Euclideana	<i>Term-based</i>	[0,+inf]	0
Coseno	<i>Term-based</i>	[0,1]	1
Manhattan	<i>Term-based</i>	[0,+inf]	0
Levenshtein	<i>Edit-based</i>	[0,+inf]	0
Damerau	<i>Edit-based</i>	[0,+inf]	0
Jaro	<i>Edit-based</i>	[0,1]	1

Fonte: Os autores (2018).

Dadas as diferentes faixas e melhores valores para cada medida observada no Quadro 2, fez-se necessário um ajuste nas diferentes escalas, de maneira que os valores ficassem em pequenos intervalos, como de 0 a 1. Exceto as métricas Jaro e Coseno, utilizou-se a equação (1) para efetuar a normalização de cada um dos subconjuntos de dados estimados pelas diferentes medidas.

$$i - \min(i) / \max(i) - \min(i) \quad (1)$$

No entanto, devido à divergência entre os valores que representam a similaridade/dissimilaridade dos documentos, realizou-se uma adequação, através do seu complementar, nas métricas Levenshtein, Damerau, Euclideana e Manhattan, onde 1 significa totalmente similar e 0 completamente dissimilar.

A validação dos experimentos entre os dois subconjuntos de notas (real e calculado) foi conduzida pelo Teste T para amostras não pareadas. Por definição, Teste T,

também chamado Teste T de *Student*, compara duas médias de forma a identificar se elas são diferentes uma da outra (SNEDECOR; COCHRAN, 1989). Em outras palavras, a hipótese assumida como verdadeira para a construção do teste, hipótese nula ( $H_0$ ), parte do pressuposto que as médias são iguais, conforme observado na equação (2).

$$H_0: \mu_1 = \mu_2 \quad (2)$$

$$H_1: \mu_1 \neq \mu_2 \quad (3)$$

Outro aspecto tratado pela referida métrica é a importância de tais diferenças, ou seja, evidencia se essas diferenças podem ter acontecido por acaso. Tal análise é realizada mediante observação dos valores obtidos por *T-Values* e *p-values*, respectivamente.

Se for observado um p-valor grande, por exemplo, maior que 0,05 ou 0,1, então não será possível rejeitar a hipótese nula de pontuações médias idênticas. Se o valor p for menor que o limite, por exemplo, 1%, 5% ou 10%, então rejeita-se a hipótese nula de médias iguais (SNEDECOR; COCHRAN, 1989). Comumente, na literatura utiliza-se o nível de significância de 5%, valor este adotado no presente trabalho.

Os pressupostos do Teste T foram verificados antes da análise. Dado o pequeno conjunto de dados, o teste de normalidade foi apoiado no método denominado Shapiro-Wilk (SHAPIRO; WILK, 1965). Utilizou-se o Teste F para o cálculo da variância entre os dois subconjuntos de dados. A hipótese de independência foi verificada mediante a análise do coeficiente de correlação de Pearson (SNEDECOR; COCHRAN, 1989).

Os Quadros 3 e 4 sintetizam os valores p (*p-value*) de cada uma das métricas utilizadas na pesquisa. Cabe ressaltar que os valores foram obtidos através da comparação do gabarito com as respostas. Neste caso, para as medidas baseadas em *tokens*, investigou-se o melhor esquema de divisão dos termos (*n-grams*), conforme observado no Quadro 3.

Quadro 3: Valores p (p-values) em função de cada métrica para os diferentes esquemas de divisão de textos. Melhor valor destacado em negrito.

Métrica	<i>n-grams</i>			
	2	3	4	5
Euclideana	0,0082577244	0,0063413715	0,0048307403	0,0037329371
Coseno	<b>0,5809614649</b>	0,3533874903	0,2280959278	0,1539552384
Manhattan	0,118316822	0,0933775266	0,0724749343	0,0567628861

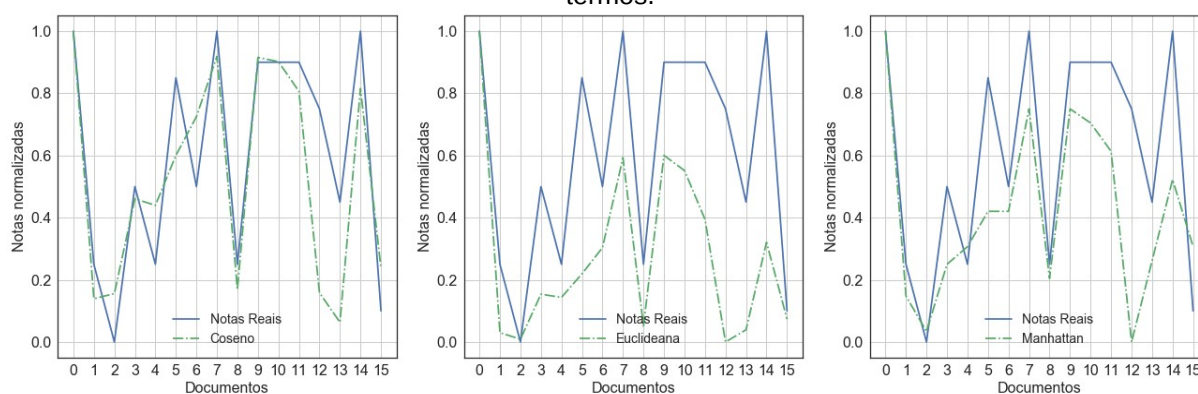
Fonte: Os autores (2018).

De acordo com o referido quadro, observa-se que o esquema de segmentação de frases em *bigrams* (*n-grams*=2), ou seja, pares de palavras adjacentes que ocorrem em um texto, apresentou melhor desempenho em relação aos demais cenários, independentemente da medida utilizada. Dentre as métricas analisadas, rejeita-se a hipótese nula para a distância Euclideana em todos os esquemas de divisão. Para as

demais medidas, em todos os cenários, aceita-se a hipótese de que as médias são iguais ( $H_0$ ). Desse subconjunto de resultados, a medida Coseno com o modelo *bigram* apresenta uma forte evidência (~ 58%) de aceitação da hipótese nula.

O Gráfico 1 apresenta as notas normalizadas (reais e estimadas) dos 15 documentos submetidos ao processo de MT, onde os valores normalizados se encontram no eixo das ordenadas e os documentos no eixo das abscissas. As linhas contínuas representam as notas estipuladas pelo avaliador e as linhas tracejadas constituem os valores estimados pelas medidas baseadas em *token* Coseno, Euclideana e Manhattan, considerando o esquema *bigram*.

Gráfico 1: comparativo entre as notas reais e as estimadas pelas medidas de distância baseadas em termos.



Fonte: Os autores (2018).

Conforme ilustrado pelo Gráfico 1, nota-se uma similaridade entre as linhas que representam as Notas Reais e as Estimadas pela métrica Coseno (gráfico à esquerda), contrapondo com os gráficos, considerando as distâncias Euclideana (centro) e Manhattan (direita).

Em relação às métricas baseadas em edição, os valores estimados pelo teste T encontram-se sumarizados no Quadro 4. Conforme valores p estimados, observa-se que somente para a métrica Jaro é aceita a hipótese de nulidade.

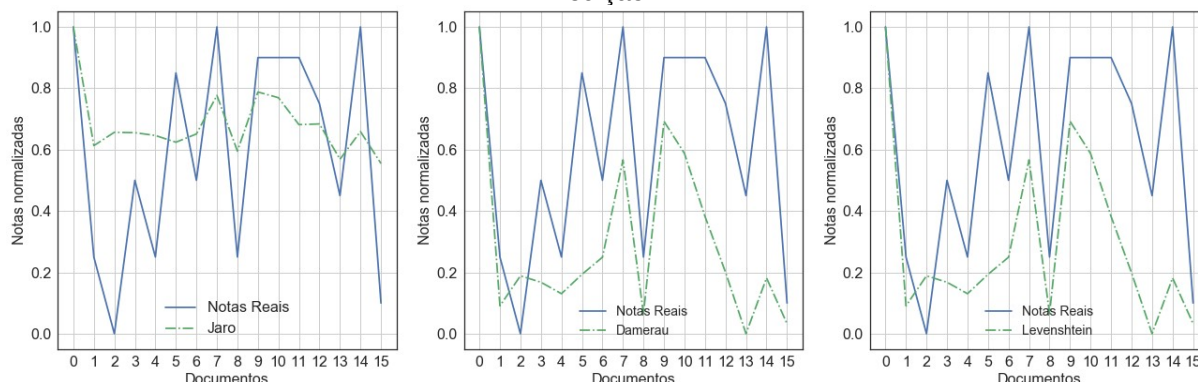
Quadro 4: Valores p (p-values) de cada métrica. Melhor valor destacado em negrito.

Métrica	Valor p (p-value)
Jaro	<b>0,376316892</b>
Damerau	0,010868374
Levenshtein	0,0108396187

Fonte: Os autores (2018).

De maneira análoga à análise efetuada nas medidas baseadas em termo, a Figura 3 apresenta os gráficos das notas normalizadas estimadas para cada uma das métricas baseadas em edição – Jaro, Damerau e Levensthein. Cabe salientar que, dada a característica do referido subconjunto de medidas, nenhum esquema de ponderação, e.g. *bigram*, foi considerado.

Gráfico 2: comparativo entre as notas reais e as estimadas pelas medidas de distância baseadas em edição.



Fonte: Os autores (2018).

Os valores obtidos pelo método Jaro não refletem a premissa de  $H_0$ , conforme observado no gráfico (esquerda) do Gráfico 2. Apesar de rejeitarmos a hipótese nula para as medidas Damerau e Levenshtein, os respectivos gráficos (Gráfico 2 – centro e direita) sugerem uma melhor predição das notas.

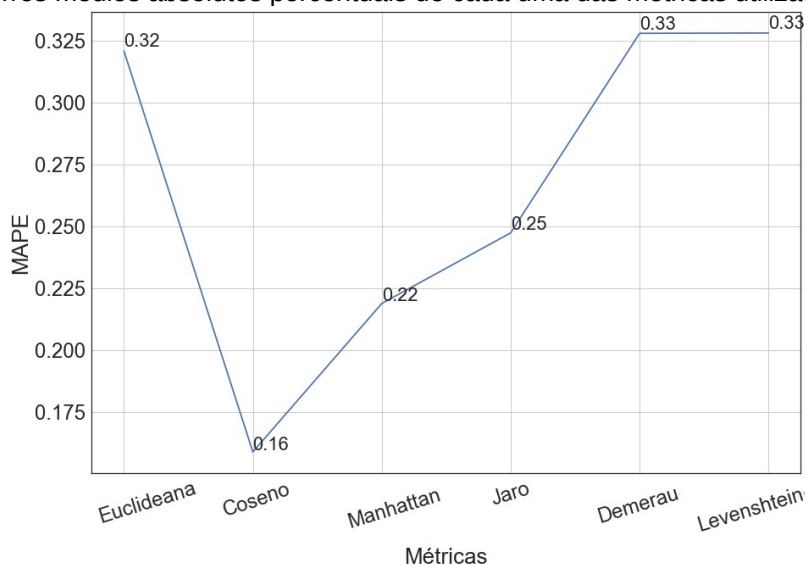
De forma a complementar a análise dos resultados estimados pelos diferentes métodos utilizados na pesquisa, a avaliação da precisão foi pautada por meio do erro percentual absoluto médio (*Mean absolute percentage error* – MAPE). Também conhecida como desvio percentual médio absoluto (*Mean absolute percentage deviation* – MAPD), a métrica é definida pela seguinte fórmula:

$$M = \frac{100}{n} \sum_{t=1}^n \frac{A_t - F_t}{A_t} \quad (4)$$

onde  $A_t$  é o valor atual,  $F_t$  é o valor previsto, e  $n$ , o número de pontos.

Os melhores modelos são aqueles que apresentam menores valores de  $M$ . O Gráfico 3 apresenta os erros estimados para cada um dos subconjuntos tratados na pesquisa.

Gráfico 3: erros médios absolutos percentuais de cada uma das métricas utilizadas no trabalho.



Fonte: Os autores (2018).

A Gráfico 3 evidencia o menor erro médio absoluto (~16%) por meio da medida Coseno, corroborando assim os resultados obtidos a partir do Teste T (Quadro 2), bem como as estimativas representadas no gráfico (esquerdo) do Gráfico 2.

Em relação às medidas baseadas em edição, os resultados do método Jaro confirmam os valores obtidos no teste T (Quadro 3). No entanto, as demais métricas desse conjunto apresentaram comportamento ligeiramente inferior, apesar da significativa diferença observada nos valores  $p$  ilustrados no referido quadro.

#### 4 Considerações finais

A finalidade do presente estudo foi a proposição de um sistema de suporte à decisão na tarefa de correção de questões dissertativas em ambientes *online*. Os modelos de mineração utilizados, bem como os modelos gerados, foram apresentados para estimar as notas atribuídas pelos avaliadores. O tema é justificado pela sua relevância, pela crescente adoção de tecnologias na educação, particularmente no que tange aos processos avaliativos.

Acredita-se que a maior contribuição do trabalho foi propor uma metodologia envolvendo a mineração de textos com conhecimento do domínio do campo de exatas, mais especificamente da área de engenharias com ênfase em computação.

Considerando o conjunto de dados utilizados, assim como as diferentes métricas avaliadas nas múltiplas etapas de extração de conhecimento em textos adotadas ao longo da pesquisa, conclui-se que, majoritariamente, as medidas baseadas em termos obtiveram as melhores taxas de sucesso, observadas por intermédio do Teste T e do MAPE. Desse subconjunto, os melhores valores obtidos para o valor  $p$  foram alcançados pela métrica Coseno, considerando o modelo de segmentação de palavras *bigram*. Em relação à MAPE, nota-se um comportamento similar à medida anterior, onde o menor erro percentual absoluto médio é atribuído novamente à medida Coseno.

Cabe salientar que a presente pesquisa não considerou critérios ortográficos e gramaticais dos revisores na composição das notas. Outro aspecto peculiar não abrangido diz respeito à existência de respostas compostas somente por termos específicos, possibilitando, assim, incidir na ocorrência de falsos positivos.

Ao analisar a hipótese de melhora da tarefa de correção de questões discursivas com auxílio de modelos computacionais, conclui-se que as ferramentas de PLN em conjunto com técnicas da área de RI podem auxiliar profissionais e gestores da educação nos processos avaliativos, particularmente os que envolvem dados não estruturados. Sua adoção no âmbito dos sistemas EaD pode possibilitar a diversificação e ampliação de tais instrumentos avaliativos no sistema educacional.

Entende-se que a maior contribuição deste artigo consistiu na proposição de uma metodologia que visa auxiliar pesquisadores na escolha do conjunto de tarefas na etapa de pré-processamento dos dados, assim como do tipo de medida de similaridade mais adequada para o problema tratado na referida pesquisa, podendo esta experiência servir como fonte para a elaboração de outros trabalhos relacionados ao tratamento de questões discursivas.

## Referências

AGGARWAL, C. C.; ZHAI, C. *Mining Text Data*. 1. ed. New York, NY, USA: Springer-Verlag New York, 2012.

ANDERSON, J. *ICT transforming education: a regional guide*. Bangkok: 2010.

ANDERSON, T. Theories for learning with emerging technologies. In: VELETSIANOS, G. (Ed.). *Emergence and innovation in digital learning: foundations and applications*. Athabasca University: Edmonton, 2016. p. 35-50.

BAEZA-YATES, R.; RIBEIRO-NETO, B. *Recuperação de informação: conceitos e tecnologia das máquinas de busca*. 2a. ed. Porto Alegre: Bookman Editora, 2013.

BAKER, R. S. J. D. Data mining for education, *International Encyclopedia of Education*, v. 7, p. 112-118, 2010.

D'ANTOLA, A. *A Prática Docente Na Universidade*. 1. ed. [s.l.] Epu, 1992.

DAMERAU, F. J. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, v. 7, n. 3, p. 171–176, 1964.

DE PAIVA, V.; RADEMAKER, A.; DE MELO, G. OpenWordNet-PT: An Open Brazilian Wordnet for Reasoning. Proceedings of COLING 2012: Demonstration Papers. *Anais...* Mumbai, India: The COLING 2012 Organizing Committee, 2012. Disponível em: <<http://www.aclweb.org/anthology/C12-3044>>. Acesso em: 27 nov. 2018.

FODEH, S.; PUNCH, B.; TAN, P.-N. On ontology-driven document clustering using core



semantic features. *Knowledge and Information Systems*, v. 28, n. 2, p. 395-421, 2011.

FREITAS, H. M. R. et al. Pesquisa em Sistemas de Informação no Brasil: 27 Anos sob uma Ótica Internacional. *Revista de Gestão e Projetos-GeP*, São Paulo, v. 9, n. 1, p. 58-86, 2018.

GOMES, L. F. EAD no Brasil: perspectivas e desafios. *Avaliação: Revista da Avaliação da Educação Superior*, Campinas, v. 18, n. 1, p. 13-22, 2013.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: concepts and techniques*. 3rd. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011.

HUSSAIN, F. E-Learning 3.0= E-Learning 2.0+ Web 3.0?. *IADIS International Conference on Cognition and Exploratory Learning in Digital Age (CELDA 2012)*, n. Celda, p. 11-18, 2012.

INEP – Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Ministério da Educação. *Censo da Educação Superior 2016*. 17 p. Disponível em: <<http://portal.inep.gov.br/web/guest/microdados>>. Acesso em: 27 nov. 2018.

JARO, M. A. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association*, v. 84, n. 406, p. 414-420, 1989.

KIM, S.; KIM, H. A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*, v. 32, n. 3, p. 669-679, 2016.

KNAFLIC, C. N. *Storytelling with data: a data visualization guide for business professionals*. New Jersey: Wiley, 2015.

KURILOVAS, E.; KUBILINSKIENE, S.; DAGIENE, V. Web 3.0--Based personalisation of learning objects in virtual learning environments. *Computers in Human Behavior*, v. 30, p. 654-662, 2014.

LEVENSHTAIN, V. I. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, v. 10, n. 8, p. 707-710, 1966.

LUCKESI, C. C. *Avaliação da aprendizagem escolar: estudos, proposições*. 3. ed. São Paulo: Cortez, 1996.

MANNING, C. D.; SCHÜTZE, H. *Foundations of Natural Language Processing*. 1. ed. Cambridge, MA, USA: MIT Press, 1999.

PERKINS, J. *Python 3 Text Processing With NLTK 3 Cookbook*. Birmingham: Packt Publishing, 2014.

PORTER, M. F. An algorithm for suffix stripping. *Program: electronic library and*

information systems, v. 14, n. 3, p. 130-137, 1980.

PRABHA, S. L.; SHANAVAS, A. R. M. Educational Data Mining Applications. *Operations Research and Applications: An International Journal*, v. 1, n. 1, p. 23-29, 2014.

ROMERO, C.; VENTURA, S. Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, v. 33, n. 1, p. 135-146, 2007.

ROMERO, C.; VENTURA, S. Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, v. 3, n. 1, p. 12-27, 2013.

SHAPIRO, S. S.; WILK, M. B. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, v. 52, n. 3/4, p. 591, 1965.

SHESHASAYEE, A.; NAZREEN BEE, M. E-learning: Mode to improve the quality of educational system. Smart Innovation, Systems and Technologies. *Anais...* V. 78, 2018.

SNEDECOR, G. W.; COCHRAN, W. G. *Statistical Methods*. [s.l.] Affiliated East-West Press, 1989.

SOARES-LEITE, W. S.; NASCIMENTO-RIBEIRO, C. A. A inclusão das TICs na educação brasileira: problemas e desafios. *Revista Internacional de Investigación en educación*, v. 5, n. 10, p. 173-187, 2012.

SRIVASTAVA, A.; SAHAMI, M. *Text Mining, classification, clustering, and applications*. 1. ed. [s.l.] CRC Press, 2009.

TEMPLE, C. A. et al. *All children read: teaching for literacy in today's diverse classrooms*. [s.l.] Pearson, 2018.

WEISS, S.; INDURKHYA, N.; ZHANG, T. *Fundamentals of predictive text mining*. 2. ed. New York, NY, USA: Springer, 2010.

YANG, Y.; PEDERSEN, J. O. A comparative study on feature selection in text categorization. *Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97)*, p. 412-420, 1997.

Recebido em dia 25 de junho de 2018.

Aprovado em dia 19 de junho de 2018.