ARTIGO

DOI: 10.20396/etd.v24i2.8660061

IDENTIFICAÇÃO DE DEFICIÊNCIAS EM TEXTOS EDUCACIONAIS COM A APLICAÇÃO DE PROCESSAMENTO DE LINGUAGEM NATURAL E APRENDIZADO DE MÁQUINA

IDENTIFICATION OF DISABILITIES IN EDUCATIONAL TEXTS WITH THE APPLICATION OF NATURAL LANGUAGE PROCESSING AND MACHINE LEARNING

IDENTIFICACIÓN DE DEFICIENCIAS EN TEXTOS EDUCATIVOS CON LA APLICACIÓN DE PROCESAMIENTO DEL LENGUAJE NATURAL Y APRENDIZAJE AUTOMÁTICO

Cíntia Maria de Araújo Pinho[1]; Amanda Ferreira de Moura[2]; Marcos Antonio Gaspar[3]; Domingos Márcio Rodrigues Napolitano[4]

RESUMO

A correção de textos educacionais como redações e questões discursivas é uma tarefa importante; além disso, diversas escolas têm exigido a intensificação da atividade da escrita para a evolução do discente. O esforço despendido para a correção, entretanto, pode aumentar a carga de trabalho do professor ou até mesmo gerar custos adicionais, bem como um longo tempo de correção para instituições como o MEC (Ministério da Educação), que é responsável pela aplicação do ENEM (Exame Nacional do Ensino Médio). Em 2019 foi anunciada pelo MEC a tendência de o ENEM se tornar digital, trazendo novas possibilidades para a análise e avaliação das redações elaboradas pelos estudantes. Nesse contexto, algumas técnicas de inteligência Artificial para análise de textos educacionais têm-se revelado úteis no processo de avaliação automática da linguagem escrita. Assim, o objetivo desta pesquisa é analisar textos, empregando para tanto as técnicas de Processamento de Linguagem Natural e Aprendizado de Máquina para identificar deficiências em textos educacionais. Esta pesquisa experimental consistiu na classificação de 695 redações elaboradas em língua portuguesa em 20 temas. Os resultados demonstraram que as técnicas empregadas possibilitaram a identificação de redações cujo conteúdo foge à temática proposta na prova, entre outras informações importantes para que o docente possa identificar falhas na escrita da redação, tais como a coesão textual ou texto insuficiente. Os resultados esperados com a aplicação da solução desenvolvida neste experimento buscam otimizar o trabalho do professor, bem como reduzir o tempo e o custo do processo de avaliação de textos educativos.

[1] Mestranda do Programa de Pós-graduação em Informática e Gestão do Conhecimento - Universidade Nove de Julho (UNINOVE). Docente na ETEC na área de Gestão e Informática - Centro Estadual de Educação Tecnológica Paula Souza, ETEC Prof. Maria Cristina Medeiros. Ribeirão Pires, SP - Brasil. **E-mail:** cintia.pinho@uni9.edu.br

[2] Mestranda do Programa de Pós-graduação em Informática e Gestão do Conhecimento - Universidade Nove de Julho (UNINOVE). São Paulo, SP - Brasil. Pós-graduada em Educação a Distância, pós-graduada em Educação Matemática, graduada em Análise e Desenvolvimento de Sistemas. Tutora de Ensino a Distância - Universidade Nove de Julho (UNINOVE). São Paulo, SP - Brasil. **E-mail:** a.f.moura@uni9.edu.br

[3] Doutor em Administração - Universidade de São Paulo (USP). São Paulo, SP - Brasil. Docente-permanente e pesquisador - Programa de Pós-graduação em Informática e Gestão do Conhecimento - Universidade Nove de Julho (UNINOVE). São Paulo, SP - Brasil. **E-mail:** marcos.antonio@uni9.pro.br

[4] Doutor em Informática e Gestão do Conhecimento - Universidade Nove de Julho (UNINOVE). São Paulo, SP - Brasil. Docente-permanente e pesquisador - Docente-permanente e pesquisador - Programa de Pós-graduação em Informática e Gestão do Conhecimento - Universidade Nove de Julho (UNINOVE). São Paulo, SP - Brasil.
E-mail: d.napolitano@uni9.pro.br

Submetido em: 14/06/2020 - **Aceito em:** 29/06/2021.

ARTIGO

DOI: 10.20396/etd.v24i2.8660061

PALAVRAS-CHAVE: Desenvolvimento educacional. Gestão do Conhecimento. Redação. Inteligência Artificial. Tecnologia.

ABSTRACT

The correction of educational texts such as essays and discursive questions is an important task, in addition, several schools have demanded the intensification of the activity of writing for the evolution of the student. However, the effort spent on correction can increase the workload of the teacher or even generate additional costs and a long correction time for institutions such as the MEC (Ministry of Education), which is responsible for the application of ENEM (National Examination for Education Medium). In 2019, MEC announced the trend of ENEM to become digital, bringing new possibilities for evaluating and analyzing the essays prepared by students. In this context, some artificial intelligence techniques for analyzing educational texts have proven to be useful in the process of automatic assessment of written language. Thus, the objective of this research is to analyze texts using the techniques of Natural Language Processing and Machine Learning to identify deficiencies in educational texts. This experimental research consisted of the classification of 695 essays prepared in Portuguese in 20 themes. The results showed that the techniques employed made it possible to identify essays whose content differs from the theme proposed in the test, among other important information so that the teacher can identify flaws in the writing of the essay, such as textual cohesion or insufficient text. The expected results with the application of the solution developed in this experiment seek to optimize the work of the teacher, reducing the time and cost of the process of evaluating educational texts.

KEYWORDS: Educational development. Knowledge management. Essay. Artificial intelligence. Technology.

RESUMEN

La corrección de textos educativos como redacciones y preguntas discursivas es una tarea importante, especialmente porque varias escuelas han exigido la intensificación de la actividad de la escritura para la evolución del estudiante. El esfuerzo dedicado a la corrección puede aumentar la carga de trabajo del maestro o incluso generar costos adicionales, además del largo tiempo de corrección para instituciones como el MEC (Ministerio de Educación), que es responsable de aplicar el ENEM (Examen Nacional del escuela secundaria). En 2019, MEC anunció la tendencia de ENEM a convertirse en digital, brindando nuevas posibilidades para evaluar y analizar las redacciones preparadas por los estudiantes. En este contexto, algunas técnicas de Inteligencia Artificial para el análisis automático de textos educativos han demostrado ser útiles en el proceso de evaluación automática del lenguaje escrito. El objetivo de esta investigación es analizar textos utilizando las técnicas de Procesamiento del lenguaje natural y Aprendizaje automático para identificar deficiencias en los textos educativos. Esta investigación experimental consistió en la clasificación de 695 redacciones en portugués en 20 temas. Los resultados mostraron que las técnicas empleadas permitieron identificar salas de redacción cuyo contenido difiere del tema propuesto en la prueba, entre otra información importante para que el maestro pueda identificar fallas en la redacción, como la cohesión textual o texto insuficiente. Los resultados esperados con la aplicación de la solución desarrollada en este experimento buscan optimizar el trabajo del profesor, reduciendo el tiempo y el costo del proceso de evaluación de textos educativos.

PALAVRAS-CLAVE: Desarrollo educativo. Gestión del conocimiento. Redacción. Inteligencia Artificial. Tecnología.

ARTIGO
DOI: 10.20396/etd.v24i2.8660061

***

1 INTRODUÇÃO

Um dos principais desafios do professor de Língua Portuguesa é conseguir que seus alunos aprendam a escrever bem. Para isso, um dos principais meios utilizados em sala de aula é a prática da redação. Entretanto a tarefa de correção de textos e redações é árdua. No Brasil, o Exame Nacional do Ensino Médio (ENEM) possui a avaliação da redação, que compõe uma das cinco dimensões de avaliação consideradas nessa prova, único item discursivo do referido exame. A maior preocupação dos candidatos é que lhes seja atribuída nota zero. Para que não seja atribuída essa nota, a redação não poderá apresentar as seguintes características (BRASIL, 2019): 1) fuga total ao tema; 2) não obediência à estrutura dissertativo-argumentativa; 3) extensão total de até 7 linhas; 4) cópia integral de texto(s) da Prova de Redação e/ou do Caderno de Questões; 5) impropérios, desenhos e outras formas propositais de anulação, em qualquer parte da folha de redação; números ou sinais gráficos fora do texto e sem função clara; 6) parte deliberadamente desconectada do tema proposto; 7) assinatura, nome, apelido, codinome ou rubrica fora do local devidamente designado para a assinatura do participante; 8) texto predominante ou integralmente em língua estrangeira e, por fim, 9) folha de redação em branco, mesmo que haja texto escrito na folha de rascunho. Considerando-se tais problemas, identificou-se que é possível minimizar o tempo e esforços gastos em correções, empregando-se técnicas de Inteligência Artificial (IA), como é o caso da mineração de textos. Sua aplicação pode permitir a identificação de características no texto que já apontariam a identificação de problemas existentes na redação.

A aplicação das técnicas de IA num contexto de avaliações em papel teria sua eficácia diminuída. Porém, em julho de 2019, foi anunciado pelo MEC (Ministério da Educação) que a avaliação do ENEM deixará de ser aplicada na versão em papel até 2026 (MEC, 2019). A implantação do Enem Digital será progressiva, com início em 2020 e previsão de consolidação em 2026. Em 2018, o INEP (Instituto Nacional de Estudos e Pesquisas) corrigiu cerca de 4.122.423 provas de redação, segundo indicado por Rocha e Moreno (2019). O uso de técnicas inteligentes no âmbito da Inteligência Artificial, baseadas no Processamento de Linguagem Natural (PLN) e Aprendizado de Máquina (AM), constitui uma alternativa interessante para reduzir o esforço e, principalmente, o tempo para a identificação das principais causas que acarretam o baixo desempenho dos estudantes, seja em redações, seja em questões discursivas.

A avaliação automática do conteúdo de redações é um tema abordado por diferentes autores como Newman *et al.* (2010), Shermis *et al.* (2010) e Vilallon e Calvo (2009). Em especial, sobre o desenvolvimento de técnicas de mineração e análise de texto em português,

DOI: 10.20396/etd.v24i2.8660061

destacam-se os trabalhos de Bazelato e Amorim (2013), Epstein e Reategui (2015), Nobre e Pellegrino (2010) e, mais recentemente, Cândido e Webber (2018). Diante disso, na presente pesquisa considerou-se a seguinte questão-problema: Como analisar deficiências de conteúdo em textos educacionais empregando técnicas de Processamento de Linguagem Natural (PLN) e Aprendizado de Máquina (AM)?

Visando responder a essa questão foi realizado um processo de descoberta do conhecimento numa base de redações, ou seja, extraíram-se dados disponíveis no repositório do Portal UOL, o qual dispõe de um conjunto de 695 redações com notas já atribuídas a redações do ENEM. Os experimentos tiveram como finalidade aplicar o Aprendizado de Máquina e, posteriormente, gerar conhecimento útil aos docentes. Num primeiro momento, os experimentos demonstrados neste artigo voltaram seu foco para os motivos que levam à atribuição de nota zero na redação do ENEM. Assim, inicialmente foi tratada a questão de fuga ao tema proposto para a redação. Dessa forma, nesta pesquisa tem-se como objetivo analisar textos empregando técnicas de PLN e AM para identificar deficiências de conteúdo em textos educacionais. O conhecimento adquirido decorrente da análise e experimentos efetuados nos textos educacionais utilizará o processo KDD (*Knowledge Discovery in Databases* – descoberta de conhecimento em bases de dados) em sua metodologia.

2 REFERENCIAL TEÓRICO

2.1 Gestão do Conhecimento e Descoberta de Conhecimento em Bases de Dados (KDD)

O aumento na quantidade de pesquisas abordando o tema Gestão do Conhecimento (GC) tem sido observado desde a década passada por autores como Brewer e Brewer (2010) e Serenko *et al.* (2010). O trabalho de Bhojaraju (2019) traz contribuições sobre a gestão do conhecimento (GC), definições, necessidades, ativos e o desafio em iniciar suas práticas em qualquer organização. O autor mostra que os conhecimentos de uma organização perfazem um importante ativo estratégico na atual Sociedade do Conhecimento. Por causa disso, Gunjal (2019) reforça a importância do estabelecimento de práticas de Gestão do Conhecimento nas organizações contemporâneas. Já Gaspar *et al.* (2016) afirmam que ferramentas de tecnologia da informação e comunicação viabilizam os processos de gestão do conhecimento nas organizações. Nonaka e Takeuchi (1997) observam a 'hierarquia' na qual o conhecimento é gerado a partir dos dados interpretados, tornando-se informação que, por sua vez, com base nas análises dos resultados coletados é transformada em conhecimento.

ARTIGO
DOI: 10.20396/etd.v24i2.8660061

Segundo Moraes e Ambrósio (2007), descobrir conhecimento significa identificar, receber informações relevantes e poder processá-las e agregá-las ao conhecimento prévio de seu usuário, mudando o estado de seu conhecimento atual, a fim de que determinada situação ou problema possa ser resolvido. Nesse sentido, observa-se que o processo de descoberta de conhecimento está fortemente relacionado com a forma pela qual informações e conhecimentos são processados.

Surge, então, o processo *Knowledge Discovery in Databases* (KDD) - descoberta de conhecimento em bases de dados, uma proposta de Fayyad *et al.* (1995) para se obterem respostas que não podem ser detectadas quando se aplicam métodos tradicionais na análise de dados para posterior tomada de decisão. Isso porque, em sua maioria, os métodos tradicionais são capazes de verificar apenas as relações explícitas existentes nos bancos de dados. Brachman *et al.* (1996) descrevem a evolução decorrente da descoberta de conhecimento em bancos de dados (KDD) e técnicas de mineração de dados. Os processos KDD são usados para se verificar uma hipótese de um usuário ou para se descobrirem novos modelos e relações entre informações e conhecimentos existentes. O processo de geração do conhecimento pode ser visualizado na Figura 1, na qual se considera o *framework* do processo KDD.

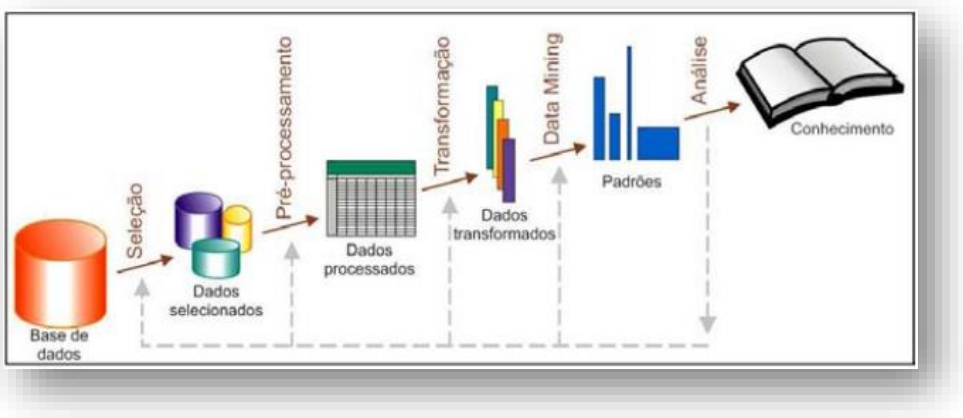

Figura 1 - *Framework* do processo KDD
Fonte: Fayyad *et al.* (1995).

Souza (2019), em sua pesquisa "Mineração de dados educacionais para avaliar os fatores que influenciam o desempenho de candidatos do ENEM", mostra que, usando-se de maneira correta os passos do processo de Descoberta de Conhecimento em Base de Dados (KDD), é possível analisar e caracterizar alunos e escolas de acordo com o desempenho obtido, tornando possível o conhecimento do perfil dos alunos e escolas submetidos a um determinado exame.

DOI: 10.20396/etd.v24i2.8660061

2.2 Processamento de Linguagem Natural (PLN) e Aprendizado de Máquina (AM) aplicados à avaliação de textos educacionais

Shermis *et al.* (2010) relatam em sua pesquisa que, desde 1966, já havia estudos para análise da escrita e acesso à tecnologia para fornecer *feedback* rápido aos professores. A tecnologia para tais aplicações, porém, experimentou grande evolução a partir da década de 1990, quando os microcomputadores permitiram a geração de textos eletrônicos. Num estudo realizado por Araújo (2011), discutem-se os movimentos da evolução na educação no último século e a incorporação de Tecnologias de Informação e Comunicação para entender se as tecnologias podem ajudar a promover maior qualidade e êxito na educação. Mungioli (2009), por sua vez, busca entender as transformações ocorridas com a chegada do computador à sala de aula. Neste estudo, foi possível entender que a nova geração lida com a linguagem e o movimento de forma diferente dos adultos e perceber uma adaptação muito fácil ao ambiente de aprendizagem com o uso da tecnologia.

As técnicas de PLN e outros mecanismos computacionais são cada vez mais pesquisados a fim de auxiliar professores no processo de correção e identificação de problemas no aprendizado. Cândido e Webber (2018) entendem que é um desafio tratar a coesão de um texto de forma automática; não obstante, em sua pesquisa, os autores descrevem as possibilidades de se tratar com assertividade a coerência e coesão de redações com uso de ferramentas de PLN. No estudo realizado por eles utilizam-se os elementos linguísticos e técnicas computacionais para a avaliação de uma redação. Nos experimentos, eles compararam a análise executada por um *software* e as avaliações feitas por dois especialistas humanos. Foram encontrados resultados convergentes em 70% dos casos analisados no experimento. Consideram-se tais resultados iniciais promissores para o desenvolvimento de solução para avaliação automática de redações, abrindo-se novas possibilidades de pesquisa.

Outros experimentos sobre análise automática de coesão textual em redações foram realizados por Nobre e Pellegrino (2010). Em seus estudos, identificou-se de forma automática problemas de coesão em 90% dos textos argumentativos e dissertativos analisados. Os resultados da solução automatizada aplicada no experimento foram compatíveis com as notas atribuídas em correções feitas por avaliadores humanos. Os autores afirmam, ainda, que uma correção realizada por um programa de computador não sofre interferências externas como fadiga e alteração de humor, permitindo avaliar e analisar sempre de forma equânime. Entretanto percebe-se a necessidade de revisão das expressões regulares visando detectar problemas não identificados pelo computador. Assim, o processo automatizado diminui a carga de trabalho do avaliador humano e mostra-se uma ferramenta para apoio ao processo de correção executado por um avaliador humano.

DOI: 10.20396/etd.v24i2.8660061

O aprendizado de máquina (AM) é um campo da Inteligência Artificial (IA) que permite a aplicação do PLN por meio de aprendizado baseado em experiências anteriores. Dessa forma, os algoritmos empregados no aprendizado de máquina extraem o conhecimento em bases de dados reais e tomam decisões fundamentadas em classificações existentes. Neste artigo aplica-se essa técnica num banco de redações já corrigidas e classificadas por avaliadores humanos.

2.3 Mineração de Dados aplicada à avaliação de textos educacionais

Segundo Epstein e Reategui (2015), a compreensão textual é um importante aspecto da educação do aluno. Se não for apropriadamente aprimorada no decorrer da formação educacional, poderá trazer problemas futuros para o indivíduo tanto na área pessoal, quanto na profissional. Os autores argumentam ainda que diferentes estudos demonstram ser possível auxiliar o processo de compreensão textual por meio do resumo de texto e visualização de conceitos-chaves. Esses mecanismos auxiliam o leitor a compreender mais facilmente o texto e realizar a análise crítica do que foi lido.

Em experimentos realizados por Macedo, Behar e Azevedo (2014), utilizaram-se técnicas de mineração de textos utilizando-se o *software* Sobeck em fóruns de ensino a distância. O sistema faz uma análise estatística de conceitos utilizados em textos escritos por alunos e traz como resultado os grafos com os conceitos principais na dissertação. O uso dessa ferramenta proporcionou redução no tempo dedicado pelo docente à leitura de todo fórum, pois a rede de conceitos pode proporcionar indicadores valiosos, como: relevância das postagens e pertinência do que foi escrito, disponibilizando um tempo maior ao professor para direcionar auxílio aos discentes que registraram poucas contribuições nos indicadores citados.

Considerando-se o processo de mineração de dados (*data mining*), é feita a exploração, análise e extração de informações de um determinado volume de dados. Para Epstein e Reategui (2015, p. 1), “o minerador identifica os conceitos mais relevantes de um texto, as relações entre esses conceitos e apresenta os resultados numa visualização de grafo”. Assim, nesta pesquisa empregou-se o processo de mineração de dados para diagnosticar a proporção de redações com problemas tais como: pontuação, coerência e dispersão do tema proposto.

2.4 Classificação de textos utilizando algoritmos não supervisionados

Segundo Rossi (2015), a classificação automática de textos atribui rótulos a documentos textuais ou porções de texto. Uma forma viável de realizar essa classificação dá-se por meio de algoritmos de aprendizado de máquina, que são capazes de ‘aprender’,

DOI: 10.20396/etd.v24i2.8660061

generalizar ou extrair padrões dos dados trabalhados. Na visão de Porto Filho (2017), o Aprendizado de Máquina não supervisionado é realizado com dados não rotulados, ou seja, quando não se sabe quantas ou quais classes existem. O objetivo de utilizar essa técnica volta-se para a classificação das entidades em grupos baseados na similaridade entre as instâncias, sendo a *cluster*ização uma das técnicas utilizadas nesse processo.

*Cluster*ização é o agrupamento automático de instâncias similares, ou seja, uma classificação não-supervisionada dos dados. Os grupos gerados por essa classificação são chamados *clusters* (agrupamentos). Muitas vezes, a similaridade entre os dados é encontrada pela aplicação de métricas de distância entre os dados em análise. Um dos algoritmos mais básicos para a *cluster*ização chama-se *K-means*. Esse algoritmo encontra k *clusters* diferentes no conjunto de dados em análise. O centro de cada *cluster* é chamado 'centroide' e apresenta a média dos valores num determinado *cluster* (HONDA, 2017).

A similaridade do algoritmo *K-means* é calculada com base numa função de distância, que pode estar relacionada com a própria distância euclidiana ou a alguma medida de similaridade, como um coeficiente de correlação. As coordenadas dos objetos para a plotagem dos gráficos correspondentes podem ser obtidas por escalonamento multidimensional, conforme indicam Rêgo (2016) e Fernandez e Marques (2019). O escalonamento multidimensional (MDS) é uma técnica de interdependência que permite mapear distâncias entre objetos. O MDS é especialmente apropriado para representar graficamente 'n' elementos num espaço de dimensão menor do que o original, levando-se em conta a distância ou a similaridade que os elementos têm entre si (FÁVERO *et al.*, 2009).

Nesse tipo de representação, quanto mais próximos estiverem os objetos, mais semelhantes entre si eles serão. O escalonamento multidimensional é uma técnica exploratória usada para a obtenção de avaliações comparativas entre objetos, para identificar comportamentos não observados por meio da aplicação de outras análises (HAIR JR. *et al.*, 2009). A título de ilustração, na Figura 2 é exemplificada uma projeção do mapa MDS utilizando o algoritmo *K-means* para demonstrar a classificação de filmes com base nos textos de suas sinopses. Cada *cluster* é representado por uma cor, de acordo com o gênero do filme.

ARTIGO
DOI: 10.20396/etd.v24i2.8660061

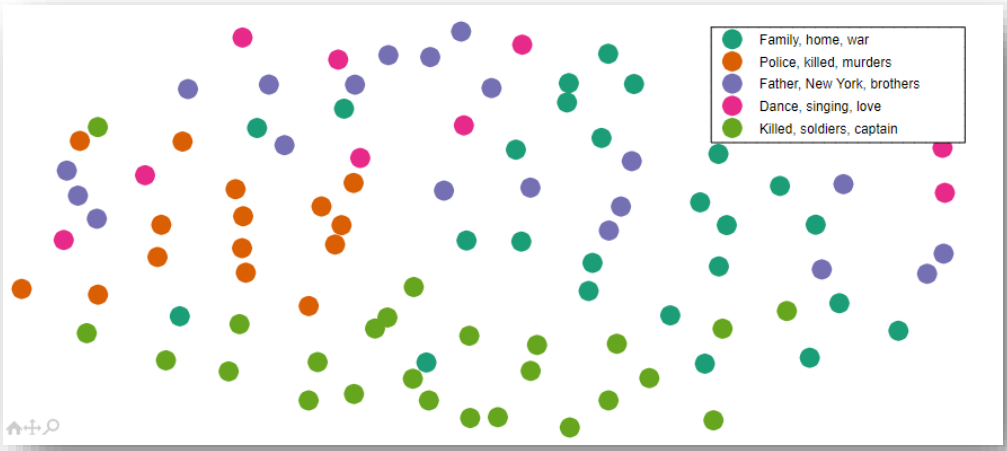

Figura 2 - Mapa MDS para classificação de filmes
Fonte: Brandon (2014).

3 MÉTODO E MATERIAIS

Para atingir o objetivo proposto nesta pesquisa quantitativa experimental, foram realizados experimentos computacionais utilizando-se como ferramenta principal a linguagem *Python* versão 3.7. O computador para processar as informações foi um Acer com processador AMD A12 *Quad-Core*, memória RAM de 8GB e HD de 1TB. Os experimentos realizados empregaram as seguintes bibliotecas do *Python*: NLTK para normalização do texto, *Numpy* para álgebra linear e operações com matrizes, pandas para gerenciamento de dados, *Scikit-learn* para extração de atributos e algoritmos de aprendizado de máquina e *Matplotlib* para visualização dos dados.

A proposta do experimento voltou-se para a aplicação de técnicas de Processamento de Linguagem Natural (PLN) e Aprendizado de Máquina (AM) utilizando-se algoritmos para mensurar a medida de similaridade entre as redações analisadas. Mais especificamente, empregou-se a remoção de *stop words, token*ização e *stemming* para a normalização dos textos (MARTIN; JURAFSKY, 2009). A seguir foi analisada a frequência dos vocábulos encontrados, empregando-se a Matriz TD IDF, conforme abordagem proposta por Salton e Yang (1973), que visa medir o grau de importância de uma palavra em relação a um conjunto de documentos. Para tanto, utilizou-se a frequência (TF) da palavra e sua relevância no conjunto de textos (IDF). Isso foi feito ao se extrair a palavra do texto normalizado e gerar uma matriz de sentenças ‘x palavras’, indicando-se o valor de TF-IDF de cada palavra para cada sentença. A similaridade das sentenças então foi calculada por meio do cosseno da distância entre os vetores TF-IDF dos pares de sentenças. O cálculo dessas distâncias foi

DOI: 10.20396/etd.v24i2.8660061

processado por meio do algoritmo *KMeans* (LLOYD, 1982), resultando em 20 agrupamentos (*clusters*) encontrados.

Para o delineamento da condução dos experimentos da pesquisa foi empregado o *framework* do processo KDD proposto por Fayyad *et al.* (1995). Assim, a proposta do experimento é obter respostas que não se podem detectar quando se aplicam métodos tradicionais na análise de dados para posterior tomada de decisão, pois, em sua grande maioria, os métodos tradicionais são capazes de verificar apenas as relações explícitas nos bancos de dados. O Quadro 1 demonstra a aplicação de cada etapa do experimento realizado.

Quadro 1 - Descrição dos métodos empregados no experimento realizado

Etapa	Descrição
Aquisição e seleção de dados	base com 695 redações; 20 temas; 20 descrições e todas as notas de cada redação.
Pré-processamento	remoção das *StopWords*; *tokenização* de palavras; *tokenização* de sentenças e *stemming;*
Formatação e transformação dos dados	construção da Matriz TFIDF (verificação da frequência de termos por documento);
Mineração de dados e padrões visualizados	Histogramas; Mapas de Calor; *Cluster Map;* Matriz MDS; *Kmeans* (classificação dos textos);
Interpretação e conhecimento	análise dos resultados e conclusão.

A pesquisa quantitativa utilizou uma amostra de 695 redações com 20 temas diferentes. Os dados disponibilizados para análise no experimento executado foram: número da redação; texto integral da redação escrita pelo aluno; título dado pelo aluno; texto motivador; tema da redação; nota da avaliação de cada competência (total de 5) e, por fim, nota total.

Após a aquisição dos dados, foram empregadas as técnicas de remoção de *stop words, token*ização *e stemming* para a preparação e normalização dos textos. Em seguida, foram construídos vetores para a verificação de frequências de palavras, sentenças e notas e, com base nesses resultados, gerados os histogramas.

No processo de formatação e transformação dos dados, realizou-se a contagem de ocorrência de palavras por documento, obtendo-se a correspondente matriz TF-IDF. Nesta etapa, ocorreu a ponderação inversa à frequência das palavras no documento, ou seja, as palavras que ocorrem com frequência num documento, mas não no *corpus*, receberam

DOI: 10.20396/etd.v24i2.8660061

ponderação mais alta, pois supõe-se que essas palavras contenham mais significado em relação ao documento. Nas 695 redações analisadas, foram levantados os vocábulos de maior significado e, dessa forma, as palavras comuns a todas as redações tiveram peso menor.

Depois que os dados foram vetorizados por meio do *TFIDF Vectorizer* (biblioteca do *Scikit-learn* para Aprendizado de Máquina), o próximo passo foi aplicar o algoritmo *K-means* para considerar os 20 *clusters* (temas das redações) e calcular/recalcular os centroides num processo iterativo, até que o algoritmo alcançasse convergência. Quando os dados convergiram e o algoritmo alcançou determinado nível de aprendizagem por meio da base de redações, o próximo passo executado foi a aplicação de escalonamento multidimensional (MDS).

No escalonamento multidimensional (MDS), foram mapeadas as distâncias entre os principais termos dos documentos e as notas aplicadas pelos avaliadores para identificar a possível fuga ao tema. Para cada tema de redação foi gerado um Mapa MDS evidenciando as redações que estavam mais aderentes ao respectivo tema, bem como as respectivas notas aferidas pelos avaliadores. As diferentes cores das elipses geradas em cada *cluster* demonstram a indicação de possível fuga ao tema da redação. Em complemento, o tamanho das elipses demonstra o alto ou baixo desempenho do grupo de estudantes naquele tema em especial.

Após a aplicação da *cluster*ização, que é um algoritmo de aprendizado de máquina não supervisionado para visualização de dados, procurou-se identificar o comportamento da base de redações. Para tanto, foram gerados mapas e gráficos que serão expostos nos resultados deste experimento no tópico a seguir.

ARTIGO
DOI: 10.20396/etd.v24i2.8660061

4 APRESENTAÇÃO E ANÁLISE DOS RESULTADOS

Após aplicação das técnicas mencionadas, foi possível averiguar que o experimento atingiu seu objetivo, ou seja, foi possível gerar informações úteis ao avaliador da redação aplicando-se técnicas de IA: TFIDF, Escalonamento Multidimensional e *KMeans* demonstradas a seguir.

4.1 Análise a partir de Histogramas

Os histogramas a seguir expostos têm como objetivo entender a base de dados de redações e identificar os primeiros desvios de escrita. Esses primeiros resultados foram extraídos após a realização da normalização dos dados, aplicando-se para tanto as técnicas de PLN e mineração de textos na fase de pré-processamento.

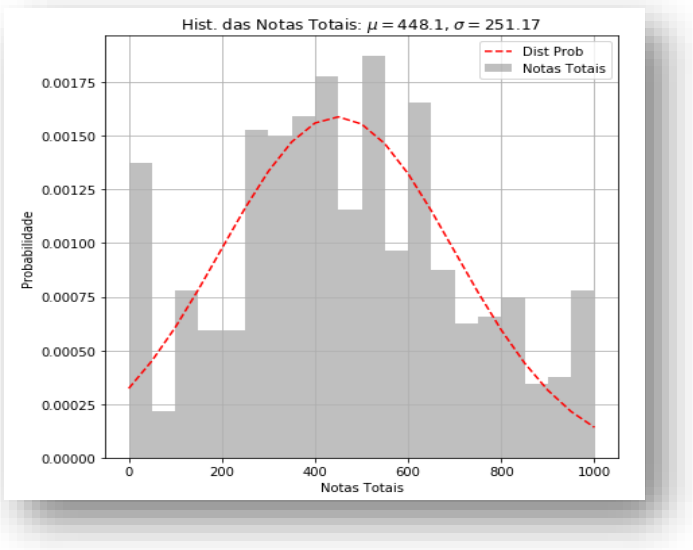

Figura 3a - Histórico de notas por redação

No primeiro histograma (Figura 3a), é possível visualizar o histórico das notas aplicadas, no qual foram identificadas as notas médias aplicadas entre 251 e 448 pontos, em sua maioria. Pôde-se observar a grande quantidade de alunos com notas inferiores a 100 pontos. Esse resultado preliminar já é capaz de proporcionar ao professor a indicação de um ponto de atenção em relação ao desempenho dos estudantes, principalmente quando o professor tem uma grande quantidade de trabalhos a corrigir.

DOI: 10.20396/etd.v24i2.8660061

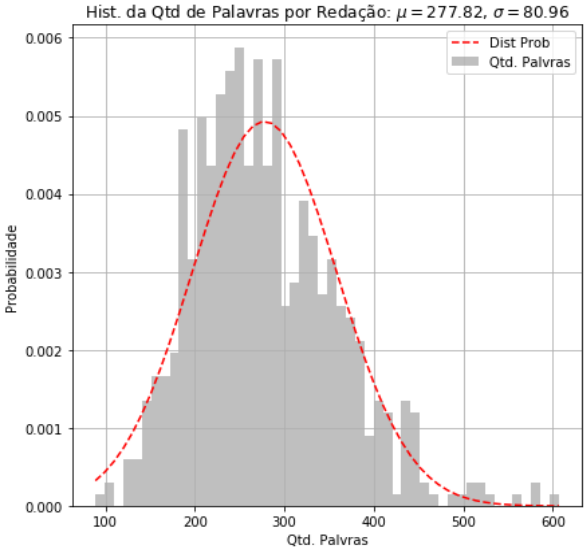

Figura 3b - Quantidade de palavras por redação

No segundo histograma (Figura 3b), é possível visualizar a quantidade de palavras por redação; a maioria das redações apresentou entre 80 e 277 palavras. Na Figura 3c é demonstrada a quantidade de sentenças na redação, com uma média de dez sentenças por redação. Entretanto também foi identificado que há redações com menos de 2 sentenças, o que aponta textos com quantidade de linhas insuficientes para a avaliação da redação.

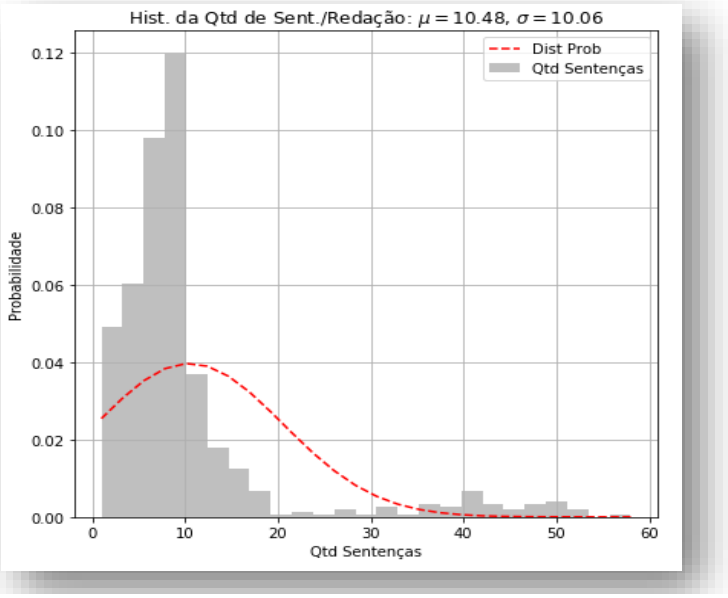

Figura 3c - Quantidade de sentenças por redação

DOI: 10.20396/etd.v24i2.8660061

O último histograma (Figura 3d) apresenta um cruzamento entre as sentenças e palavras por sentenças; a análise média é de 35 a 43 palavras por sentença. Foram encontradas sentenças com até 300 palavras, indicando insuficiência de pontuação e possível incoerência textual.

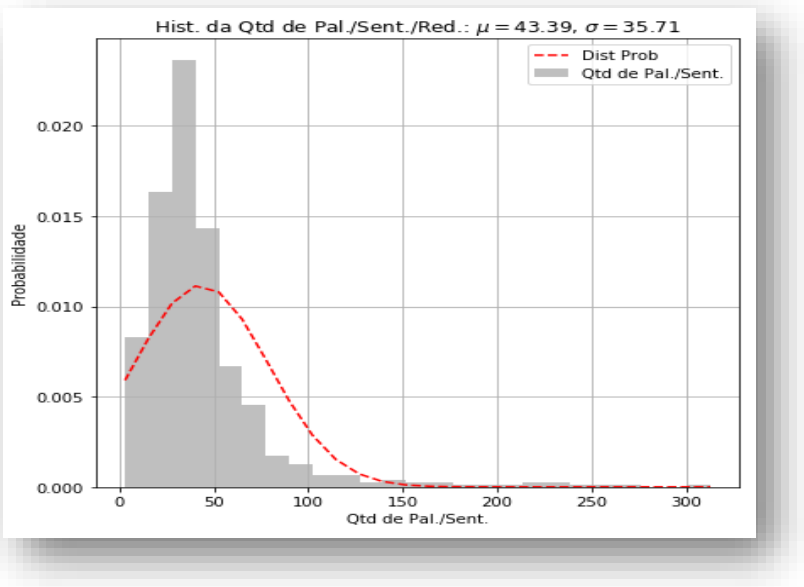

Figura 3d - Quantidade de palavras por sentença

Algumas análises podem ser levantadas a partir desses histogramas. Critérios diferenciados apontam para a possibilidade de estabelecimento de correlação entre a nota atribuída numa redação e a quantidade de sentenças. Tal achado poderia auxiliar o avaliador a responder às seguintes perguntas: Há um padrão entre a quantidade de sentenças e a nota aplicada pelo avaliador? Uma sentença muito longa indica a falta de pontuação? Neste caso, ao se aplicarem as primeiras técnicas de PLN são gerados indicadores que permitem a visualização de erros de pontuação e, consequentemente, problemas de coesão textual. Na aplicação desse experimento, é possível avaliar hábitos de escrita, como: regras de pontuação, repetição de palavras e sentenças e coesão textual. Tais achados estão em consonância com os estudos de Bazelato e Amorim (2013), Epstein e Reategui (2015) e Nobre e Pellegrino (2010).

4.2 Detecção de padrões pela aplicação de aprendizado não supervisionado

Na próxima etapa, após o pré-processamento, tratamento dos dados e plotagem dos histogramas, foi realizado o processo para detectar padrões entre as redações em análise. Para tanto, esse procedimento aplicou técnicas de aprendizado não supervisionado (*cluster*ização) que podem auxiliar os professores a entenderem os temas

DOI: 10.20396/etd.v24i2.8660061

nos quais os alunos encontram maior dificuldade na escrita, ou seja, aqueles nos quais os alunos obtiveram notas mais baixas.

No processo de *cluster*ização, foi utilizado o algoritmo *K-Means*, um algoritmo disponível na biblioteca *Scikit-learn* para aprendizado de máquina. Nesse experimento, tal algoritmo foi aplicado para calcular/recalcular os centroides que relacionam os temas com os principais termos das redações. Nesta etapa, há iteração até que o algoritmo alcance a convergência. Os resultados demonstrados a seguir estão em conformidade com as indicações feitas por Rêgo (2016), Honda (2017) e Fernandez e Marques (2019).

As redações foram divididas em 20 *clusters*, que são os temas correspondentes trabalhados nas redações em análise. Em seguida examinou-se a similaridade entre os temas, para verificar se as palavras destacadas como de maior relevância num tema estavam insertas em seu próprio *cluster*. Na Figura 4ª, são expostos os resultados desse processo, com a apresentação de um exemplo da análise do *cluster* referente ao tema "Universidade em Crise: quem paga a conta". Em complemento, na Figura 4b é exposto outro exemplo da análise do *cluster* referente ao tema "Cantar ou não cantar o Hino Nacional, eis a questão…". Por fim, a Figura 4c expõe um exemplo da análise do *cluster* referente ao tema "Redes Sociais e Manipulação Política".

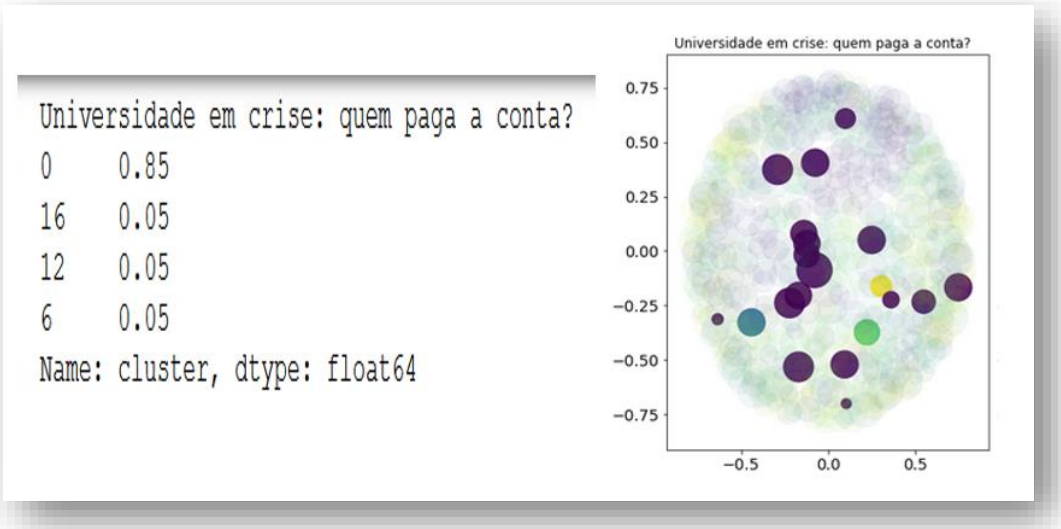

Figura 4a – Nível de Similaridade e Clusterização – Exemplo 1

Ao se realizar a análise dos resultados expostos na Figura 4a, entende-se que o professor pode utilizar essa solução para diagnosticar quanto os alunos se desviaram do tema solicitado. O tamanho das elipses mostra a nota do aluno: quanto maior a elipse maior a nota. Os índices da esquerda demonstram quanto os textos estavam consoantes

DOI: 10.20396/etd.v24i2.8660061

com o tema proposto. Neste caso, 85% dos textos foram classificados como aderentes ao tema. Em complemento, o algoritmo executado classificou os outros 15% de textos em outros temas, indicando possível fuga ao tema proposto na redação.

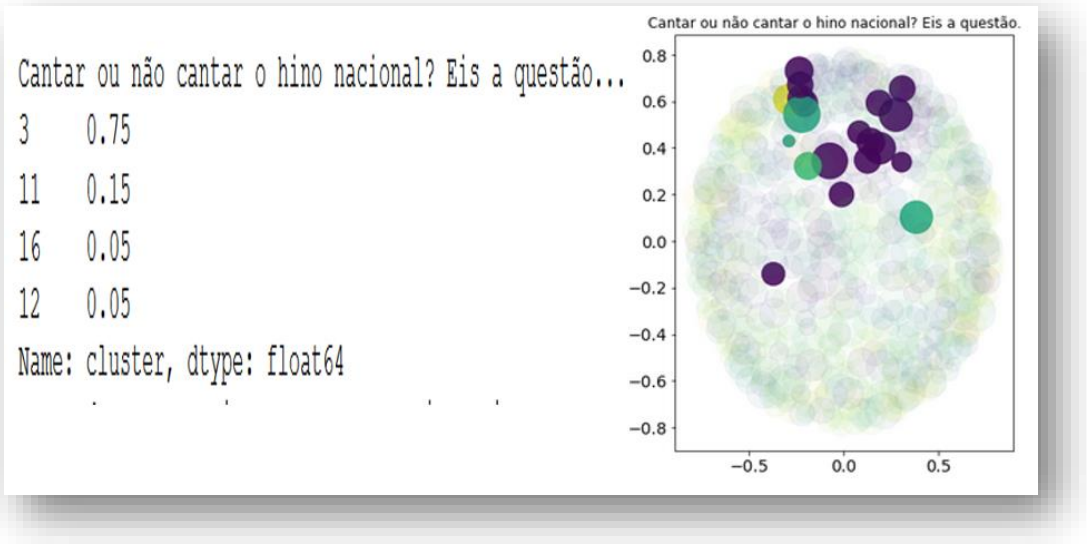

Figura 4b – Nível de Similaridade e Clusterização – Exemplo 2

Os resultados expostos na Figura 4b também demonstram fuga ao tema. Nesse *cluster* houve ocorrência maior da fuga do tema proposto na redação, indicada pela maior variação de cores nele. Verificou-se a ocorrência de quatro *clusters* neste agrupamento. A análise dos resultados indica que, nos textos escritos pelos alunos, apenas 75% das redações foram classificadas no próprio *cluster*. Outro resultado importante é o tamanho das elipses geradas no experimento, o que demonstra diferentes notas atribuídas pelo avaliador. Vale ressaltar que as notas atribuídas podem ser utilizadas em novos experimentos de classificação para determinar futura atribuição automática de notas pela solução desenvolvida.

ARTIGO
DOI: 10.20396/etd.v24i2.8660061

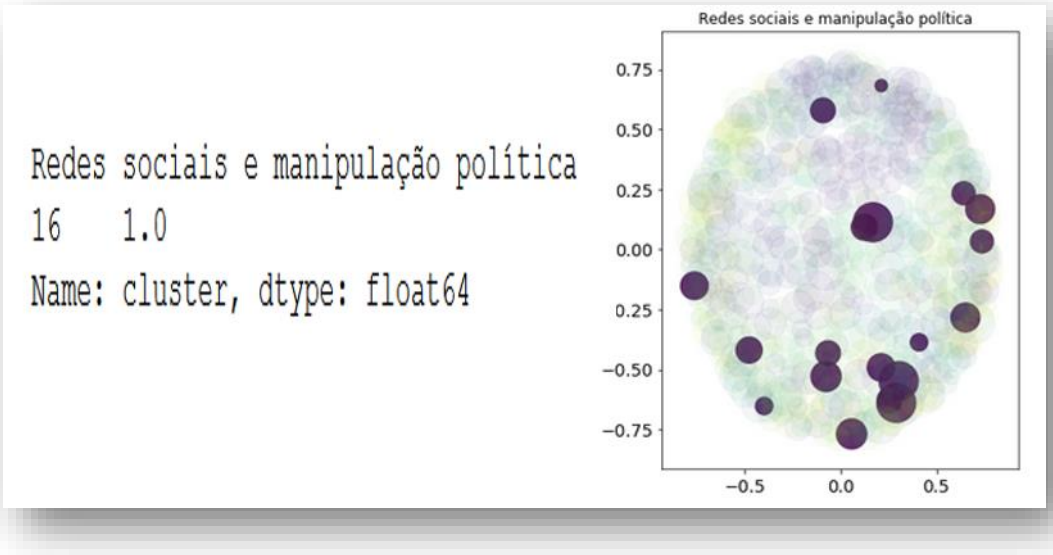

Figura 4c – Nível de Similaridade e *Cluster*ização – Exemplo 3

Os resultados da Figura 4c mostram um agrupamento de cor predominantemente azul, um exemplo de que os alunos conseguiram redigir suas redações consoantes com o tema proposto. É possível chegar a essa conclusão, pois as 20 elipses estão dentro do mesmo *cluster*, e cada uma delas representa uma redação analisada. Nesse caso, quando o professor analisa tais resultados, pode entender que não há a necessidade de revisão das redações quanto ao tema proposto. Em seu trabalho, Cândido e Webber (2018) também desenvolvem técnicas de mineração e análise de texto para aplicação em textos, verificando os agrupamentos ocorridos.

A correlação entre os *clusters* e os temas de redação são expostos na Figura 5. Essa matriz foi gerada como etapa de finalização dos experimentos realizados. Os resultados nela indicados possibilitam identificar quais temas tiveram maior ocorrência de fuga ao tema proposto para a redação e, consequentemente, maior dificuldade de escrita demonstrada pelos alunos.

Ao observar os resultados do experimento da Figura 5, identificou-se que apenas cinco temas não geraram fuga ao proposto: ‘Direitos Humanos: em defesa de quem?’; ‘Informação no rótulo de produtos transgênicos’; ‘O Brasil paralisado: o que você pensa sobre a greve dos caminhoneiros’; ‘O direito ao foro privilegiado’ e, por fim, ‘Por que os jovens querem deixar o Brasil?’.

ARTIGO
DOI: 10.20396/etd.v24i2.8660061

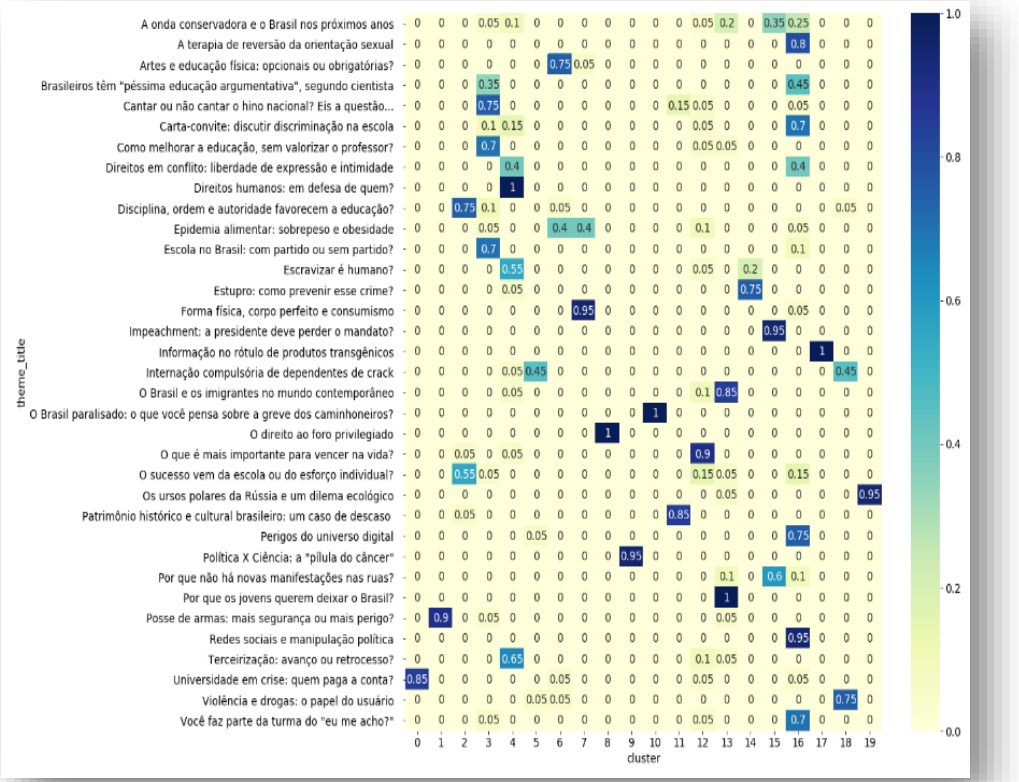

Figura 5 - Resultados dos experimentos com os 20 temas

Levando-se em consideração os experimentos desta pesquisa, foi possível identificar várias possibilidades para avaliação das técnicas de escrita e/ou argumentação dos alunos. Partindo-se desse princípio, entende-se que é possível mensurar o número de redações aderentes à proposta de tema, o que pode trazer importante conhecimento ao avaliador ou docente quanto à evolução dos alunos. Assim, é possível vislumbrar as primeiras benfeitorias ao trabalho de docentes e avaliadores no processo de correção de textos educacionais.

DOI: 10.20396/etd.v24i2.8660061

5 CONCLUSÕES

Na presente pesquisa, buscou-se identificar desvios de escrita em redações. As principais teorias utilizadas baseiam-se em algoritmos apresentados no estudo de Porto Filho (2017), que buscou aplicar técnicas de aprendizado de máquina não supervisionadas com o processo de *cluster*ização. Os experimentos nesta pesquisa levaram em conta as instruções da documentação do *Scikit-learn* e de Brandon (2014), porém aplicadas à classificação de redações quanto à aderência ao tema proposto ou à fuga dele.

A aplicação das técnicas de Processamento de Linguagem Natural (PLN), em especial a mineração de textos e AM, permitiu não apenas identificar padrões no conteúdo de cada redação analisada, como também encontrar relações de similaridade entre as redações sobre um determinado tema. Os resultados encontrados no experimento demonstraram existirem temas de redação com elevada coesão nos textos produzidos, ou seja, todas as redações analisadas encontram-se num mesmo *cluster* (agrupamento).

Também foi possível verificar a existência de outros temas com maior dispersão e isso demonstra que os alunos podem não ter a mesma compreensão sobre o tema proposto para a redação, uma vez que seus textos apresentaram conteúdos que fugiram ao tema da prova. Dessa forma, foi possível cumprir o objetivo desta pesquisa, ao se identificarem deficiências de conteúdo em textos educacionais. Com a aplicação das técnicas de Processamento de Linguagem Natural (PLN) e Aprendizado de Máquina (AM), foi possível ainda mensurar as redações aderentes ao tema originalmente proposto. Assim, dos 20 temas nos quais foram aplicados os experimentos, apenas cinco não sofreram fuga ao tema indicado, enquanto nos demais 15 verificou-se fuga parcial ou total ao tema proposto.

Em complemento, outros experimentos realizados neste trabalho apontaram a quantidade e o tamanho das sentenças nas redações analisadas. Como principais resultados, foi possível não só identificar redações com menos de duas sentenças, como também encontrar sentenças com até 300 palavras. A primeira ocorrência aponta para texto insuficiente, enquanto a segunda para possíveis erros de pontuação e falta de coesão textual na construção do texto. Assim, o processo aplicado neste trabalho possibilita identificar e analisar deficiências de conteúdo em textos educacionais de forma automatizada, o que está em consonância com trabalhos prévios empregados na fundamentação teórica desta pesquisa.

As principais contribuições deste estudo buscam permitir ao avaliador, professor ou empresas que aplicam processos seletivos avaliar as redações com menor esforço, otimizando seu o trabalho e reduzindo tempo e custo do processo de avaliação de textos educativos. Esta colaboração pode ser primordial na aplicação do ENEM digital,

DOI: 10.20396/etd.v24i2.8660061

proporcionando ao avaliador auxílio na identificação das falhas de escrita, minimizando interferências como fadiga e alteração de humor do avaliador, sintomas que podem afetar a correção de um texto dissertativo. O cumprimento do objetivo proposto denota ainda contribuição na perspectiva acadêmica, ao servir de base para estudos que, uma vez alinhados aos conhecimentos dos profissionais de ensino, possam gerar novas abordagens que facilitem capacitar os alunos a redigirem textos coesos com o tema proposto.

Outros autores procuraram contribuir para a análise automática de textos e identificação de desvios de escrita, tais como Nobre e Pellegrino (2010) e Cândido e Webber (2018), que realizaram estudos quanto à coesão textual. Já Epstein e Reategui (2015) procuraram empregar a mineração de textos para a compreensão textual e análise da dispersão do texto em relação ao tema proposto. Em seu estudo, Macedo *et al.* (2014) buscaram auxiliar os professores ao extraírem conceitos de redações. É importante ressaltar que em nenhum dos trabalhos citados aplicam-se as mesmas técnicas utilizadas neste estudo. Esses trabalhos, entretanto, têm como objetivo auxiliar docentes na tarefa de avaliar os alunos de forma mais ágil e menos cansativa. Assim, entende-se que não apenas esta pesquisa, mas também os estudos anteriores confirmam a importância da aplicação de Inteligência Artificial na área da Educação.

Conclui-se que técnicas de IA podem ser de grande valia para apoiar os professores, poupando-lhes esforço e tempo ao conferir maior eficácia ao processo de correção de conteúdos produzidos pelos alunos. A pesquisa experimental executada apresenta algumas limitações, dentre as quais destaca-se que ainda não foi possível identificar todas as possibilidades que acarretam a nota zero, de acordo com as regras do Exame Nacional do Ensino Médio. Outra limitação volta-se à escolha intencional por parte dos pesquisadores das técnicas e ferramentas utilizadas nos experimentos realizados neste estudo.

Como sugestões para estudos futuros, indica-se a possibilidade de aplicação de outras técnicas para aprimorar os resultados das informações e conhecimentos oriundos da automação proposta nos experimentos realizados neste estudo. Uma alternativa é a identificação de cópia de textos motivadores e desrespeito aos direitos humanos na dissertação, entre outros motivos que não são aceitos e desclassificam as redações produzidas pelos alunos. Outra indicação com potencial para aprimoramento dos resultados é a utilização de uma base maior de dados de redações, inclusive com maior diversidade de temas. Dessa forma, o processo de aprendizado de máquina poderia trazer maior inferência na acurácia dos experimentos realizados em estudos futuros. Outra perspectiva vislumbrada para trabalhos futuros consiste em se empregarem essas técnicas construindo bases de dados com redações, em colaboração com docentes, podendo-se observar, assim, como as técnicas

DOI: 10.20396/etd.v24i2.8660061

empregadas poderiam apoiar o trabalho de análise das redações e o aprimoramento das técnicas de IA com o apoio de especialistas.

REFERÊNCIAS

ARAÚJO, U. A quarta revolução educacional: a mudança de tempos, espaços e relações na escola a partir do uso de tecnologias e da inclusão social. **ETD - Educação Temática Digital**, v. 12, n. esp., p. 31-48, 2011.

BAZELATO, B. S.; AMORIM, E. C. F. A bayesian classifier to automatic correction of portuguese essays. In: CONGRESSO INTERNACIONAL DE INFORMÁTICA EDUCATIVA (TISE), 18., 2013. **Anais...** Porto Alegre: CCC, 2013, p. 1-13.

BHOJARAJU, G. Knowledge management: why do we need it for corporates. **Malaysian Journal of Library & Information Science**, p. 1-14, 2019.

BRACHMAN, R. J. Mining business databases. **Communications of the ACM**, p. 42, nov. 1996.

BRANDON, R. **Document clustering with Python - Top 100 films of all time**. 2014. Disponível em: http://brandonrose.org/clustering . Acesso em: 03 maio 2021

BRASIL - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). **A redação no Enem 2019:** cartilha do participante. Brasília: INEP, 2019.

BREWER, P. D.; BREWER, K. L. Knowledge management, human resource management, and higher education: a theoretical model. **Journal of Education for Business**, v. 85, n. 6, p. 330-335, 2010.

CÂNDIDO, T.; WEBBER, C. Avaliação da coesão textual: desafios para automatizar a correção de redações. **RENOTE**, v. 16, n. 1, p. 103-112, 2018.

SOUZA, C. **Mineração de dados educacionais para avaliar os fatores que influenciam no desempenho de candidatos do ENEM**. Trabalho de Conclusão de Curso (Análise e Desenvolvimento de Sistemas) - Universidade Federal Fluminense. Rio de Janeiro: UFF, 2019.

EPSTEIN, D.; REATEGUI, E. B. Uso de mineração de textos no apoio à compreensão textual. **RENOTE**, v. 13, n. 1, p. 1-10, 2015.

FÁVERO, L. P.; BELFIORE, P.; SILVA, F. L.; CHAN, B. L. **Análise de dados:** modelagem multivariada para tomada de decisões. Rio de Janeiro: Elsevier-Campos, 2009.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. **Advances in knowledge discovery & data mining**. Menlo Park: American Association for Artificial Intelligence, 1995.

DOI: 10.20396/etd.v24i2.8660061

FERNANDEZ, P. J.; MARQUES, P. C. F. **Data science, marketing & business**. São Paulo: Insper, 2019.

GASPAR, M. A.; SANTOS, S. A.; DONAIRE, D.; KUNIYOSHI, M. S.; MAGALHÃES, F. L. F. de. Gestão do conhecimento em empresas atuantes na indústria de *software* no Brasil: um estudo das práticas e ferramentas utilizadas. **Informação & Sociedade: Estudos**, v.26, n.1, p. 151-166, jan./abr. 2016.

GUNJAL, B. Knowledge management: why do we need it for corporates. **Malaysian Journal of Library & Information Science**, p. 1-14, Apr 2019.

HAIR JUNIOR, J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E.; TATHAM, R. L. **Análise multivariada de dados**. 6.ed. Porto Alegre, Bookman, 2009.

HONDA, H. **Introdução básica à clusterização**. Brasília: UNB, 2017.

LLOYD, S. P. Least squares quantization in PCM. **IEEE Trans. Inf. Theory**, v. 28, p. 129-136, 1982.

MACEDO, A. L.; BEHAR, P. A.; AZEVEDO, B. F. T. Acompanhamento da interação e produção textual coletiva por meio de mineração de textos. **ETD - Educação Temática Digital**, .v. 16, n. 1, p. 67-83, 2014.

MARTIN, J H.; JURAFSKY, D. **Speech and language processing**: an introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River: Pearson/Prentice Hall, 2009.

MEC - Ministério da Educação. **Enem terá aplicação digital em fase piloto em 2020 e deixará de ter versão em papel em 2026**. Brasília: MEC, 03 jul. 2019.

MORAIS, E. A. M.; AMBRÓSIO, A. P. L. **Mineração de textos**. Relatório técnico. Goiânia: UFGO, 2007.

MUNGIOLI, M. C. P. Jogando com o narrador: estratégias narrativas na produção de textos em ambientes escolares informatizados. **ETD - Educação Temática Digital**, v. 10, n. 1, p. 24-48, 2009.

NEWMAN, D.; LAU, J. H.; GRIESER, K.; BALDWIN, T. Automatic evaluation of topic coherence. In: Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2010. **Proceedings...** Association for Computational Linguistics, 2010, p. 100-108.

NOBRE, J. C. S.; PELLEGRINO, S. R. M. ANAC: um analisador automático de coesão textual em redação. In: BRAZILIAN SYMPOSIUM ON COMPUTERS IN EDUCATION - SBIE, 2010. **Anais...** Porto Alegre: SBC, 2010, p. 1-12.

DOI: 10.20396/etd.v24i2.8660061

NONAKA, I; TAKEUCHI, H. **Criação de conhecimento na empresa**. Rio de Janeiro: Elsevier, 1997.

PORTO FILHO, C. H. **Técnicas de aprendizado não supervisionado baseadas no algoritmo da caminhada do turista**. Dissertação (Mestrado em Bioengenharia) – Programa de Pós-graduação em Bioengenharia - Universidade de São Paulo. São Carlos, USP, 2017.

RÊGO, A. S. da C. **Aprendizado automático de relações semânticas entre tags de folksonomias.** Tese (Doutorado em Ciência da Computação) - Programa de Pós-graduação em Ciência da Computação - Universidade Federal de Campina Grande. Campina Grande, UFCG, 2016.

ROCHA, G.; MORENO, A. C. **Enem 2018:** número de redações nota mil volta a crescer, e cai o número de notas zero. Rio de Janeiro: Portal G1, 18 jan. 2019.

ROSSI, R. G. **Classificação automática de textos por meio de aprendizado de máquina baseado em redes**. 2015. Tese (Doutorado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo. São Carlos, USP, 2015.

SALTON, G.; YANG, C. S. On the specification of term values in automatic indexing. **Journal of Documentation**, v. 29, n. 4, p. 351-372, 1973.

SERENKO, A.; BONTIS, N.; BOOKER, L. D.; SADEDDIN, K. W. A scientometric analysis of knowledge management and intellectual capital academic literature (1994-2008). **Journal of Knowledge Management**, v. 14, n. 1, p. 3-23, 2010.

SHERMIS, M. D.; BURSTEIN, J.; HIGGINS, D.; ZECHNER, K. Automated essay scoring: Writing assessment and instruction. **International Encyclopedia of Education**, v. 4, n. 1, p. 20-26, 2010.

VILLALON, J.; CALVO, R. A. Concept extraction from student essays, towards concept map mining. In: IEEE INTERNATIONAL CONFERENCE ON ADVANCED LEARNING TECHNOLOGIES, 9th , 2009. **Proceedings...** IEEE, jul. 2009, p. 221-225.

Revisão gramatical realizada por: Iris Gardino
E-mail: irisg@fia.com.br