AVALIAÇÃO DO TD-BERT COM DIFERENTES MODELOS DE REPRESENTAÇÃO TEXTUAL PARA TAREFAS DE CLASSIFICAÇÃO DE TEXTOS

Luiz Henrique Dutra Martins[1]; Rodrigo Neves Trindade[2];
Geraldo Nunes Correa[3]; Camilla Côrtes Carvalho-Heitor[4]; Ivan José dos Reis Filho[5]

Resumo
A quantidade de dados gerados na internet cresceu exponencialmente na última década. Técnicas de Mineração de Dados (MD) e modelos de aprendizado de máquina são utilizados para obter conhecimento útil utilizando um grande volume de dados. Nesse contexto, a Mineração de Textos (MT), uma das principais atividades da MD, é o processo que busca descobrir conhecimento útil e padrões ocultos a partir de um grande volume de textos. Inicialmente, modelos de matriz atributo-valor foram apresentados na literatura para gerar representações vetoriais de textos. No entanto, as matrizes possuem alta dimensionalidade e não representam recursos semânticos dos textos. Atualmente, modelos com base na arquitetura Transformers são considerados como o estado-da-arte para representações textuais que consideram aspectos semânticos. No entanto, esses modelos geram vetores singulares e difíceis de serem compreendidos. Recentemente, uma representação denominada TD-BERT foi apresentada na literatura, considerando aspectos semânticos de dados textuais. Entretanto, o estudo foca na avaliação de uma função de rotulagem e não considera diferentes conjuntos de dados. Dessa forma, este trabalho propõe uma avaliação aprimorada do TD-BERT considerando seis representações vetoriais de textos para três conjuntos de dados de diferentes domínios. A metodologia deste trabalho avalia diferentes modelos de representação textual aplicados em tarefas de classificação. As principais atividades concentram-se nas etapas de pré-processamento e avaliação experimental, em que foram selecionados quatro algoritmos de diferentes paradigmas de aprendizado de máquina. Além disso, quatro modelos de representação textual foram utilizados para avaliar o desempenho preditivo em relação ao TD-BERT e sua variação TD-DistilBERT. Conclui-se que a abordagem TD-BERT apesar de obter desempenho poucamente inferior, se mostrou eficaz e obteve resultados similares aos demais.
Palavras-chave: Mineração de Textos; Aprendizado de Máquina; Representações Textuais.

Abstract
The amount of data generated on the internet has grown exponentially in the last decade. Data Mining (DM) techniques and machine learning models are used to obtain helpful knowledge using a large volume of data. In this context, Text Mining (MT), one of the main activities of

[1] Graduando em Sistemas de Informação pela Universidade do Estado de Minas Gerais-UEMG. E-mail: luiz.1093701@discente.uemg.br.
[2] Graduando em Sistemas de Informação pela Universidade do Estado de Minas Gerais-UEMG. E-mail: rodrigo.1093795@discente.uemg.br.
[3] Doutor em Engenharia Mecânica pela Universidade de São Paulo-USP, professor da Universidade do Estado de Minas Gerais-UEMG, unidade Frutal-MG. E-mail: geraldo.correa@uemg.br.
[4] Mestra em Ciências Ambientais pela Universidade Brasil - Campus Fernandópolis/S, professora do Curso de Sistemas de Informação da Universidade do Estado de Minas Gerais-UEMG, unidade Frutal-MG. E-mail: camilla.heitor@uemg.br.
[5] Doutorando pelo Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação (ICMC) da Universidade de São Paulo-USP-São Carlos, professor adjunto do Departamento de Ciências, Exatas e da Terra da Universidade do Estado de Minas Gerais-UEMG, unidade Frutal-MG e coordenador do curso de Sistemas de Informação e do Núcleo de Práticas em Sistemas de Informação (NUPSI) pela mesma Instituição. E-mail: ivan.filho@uemg.br.

DM, is the process that seeks to discover helpful knowledge and hidden patterns from a large volume of texts. Initially, attribute-value matrix models were presented in the literature to generate vector representations of texts. However, the matrices have high dimensionality and do not represent the semantic resources of the texts. Currently, models based on the Transformers architecture are considered state-of-the-art for textual representations that consider semantic aspects. However, these models generate unique vectors that are difficult to understand. Recently, a representation called TD-BERT was presented in the literature, considering semantic aspects of textual data. However, the study focuses on evaluating a labeling function and does not consider different data sets. Thus, this work proposes an improved evaluation of TD-BERT, considering six vector representations of texts for three sets of data from different domains. The methodology of this work evaluates different models of textual representation applied in classification tasks. The main activities focus on the pre-processing and experimental evaluation stages, in which four algorithms from different machine learning paradigms were selected. Furthermore, four textual representation models were used to evaluate the predictive performance of the TD-BERT and its variation TD-DistilBERT. It is concluded that despite having a slightly inferior performance, the TD-BERT approach proved to be effective and obtained similar results to the others.
keywords: Text Mining; Machine Learning; Textual Representations.

1 Introdução

Nos últimos anos, a quantidade de dados gerados e disponibilizados na internet cresceu de modo exponencial (JANEV et al., 2020). Os avanços tecnológicos na computação tornaram possível transformar dados em informações em conhecimento útil para auxiliar na tomada de decisão em diversas áreas do conhecimento. Nesse contexto, a MT, uma das principais atividades da MD, é o processo que busca descobrir conhecimento útil e padrões ocultos a partir de um grande volume de textos.

Os processos da MT podem ser divididos em cinco etapas: Identificação do Problema, Pré-Processamento, Extração de Padrões, Pós-Processamento e uso do Conhecimento (REZENDE *et al*., 2003). Cada etapa pode ser instanciada de acordo com a necessidade dos usuários e da aplicação.

O presente trabalho está situado na etapa de pré-processamento, onde se encontra a principal diferença entre os processos de MT e MD (AGGARWAL, 2018). A MD geralmente utiliza dados estruturados, isto é, valores numéricos que possuem uma estrutura bem definida. A MT utiliza de dados não estruturados, os quais não possuem uma formatação específica e são mais difíceis de serem processados. Dessa forma, a principal tarefa na etapa de pré-processamento na MT é oferecer uma representação vetorial que descreve informações textuais. Essa representação precisa estar adequada para a próxima etapa de Extração de Padrões.

Tradicionalmente, representações vetoriais de textos com base em *Bag-of-Words* (BoW) foram inicialmente propostos. Esse modelo associa textos em vetores que indicam o número de ocorrências de cada palavra em uma sentença (AGGARWAL, 2015). Em geral, abordagens

com base na BoW apresentam bom desempenho em tarefas simples que não requerem análise sintática e semântica dos textos (SINOARA et al., 2019). Por exemplo, o texto d1 “O Brasil perdeu de 7x1 para Alemanha”, e o texto d2 “A Alemanha perdeu de 7x1 para o Brasil” tem a mesma representação vetorial no modelo BoW (mesmas palavras com índices e mesma ocorrência dos termos d1 para d2), mas indicam significados opostos. No entanto, quando considerados recursos semânticos, o texto d3 “A Alemanha ganhou de 7x1 do Brasil” possui maior similaridade para d1 do que para d2. Além disso, estruturas com base na BoW geram representações esparsas (muitos zeros no vetor) e de alta dimensionalidade (AGGARWAL, 2014).

Recentemente na literatura, representações que consideram a semântica textual, estrutura linguística e entidades relacionadas ao contexto, têm sido propostos para superar as limitações da BoW. Inicialmente, *word2vec* foi apresentado na literatura como um modelo que gera vetores considerando palavras livres de contexto (MIKOLOV et al., 2013). Por exemplo, as palavras “rei” e “rainha” recebem representações vetoriais tão similares quanto às palavras “homem” e “mulher”, respectivamente. No entanto, as representações com *word2vec* são imutáveis e não alteram conforme as relações contextuais das palavras. Considerando o exemplo anterior, a palavra “rei” é mais similar para o contexto de “reino” do que para o contexto de “futebol”. No entanto, a palavra “rei” pode ser naturalmente usada em uma notícia do esporte para representar uma superioridade técnica, ao invés de representar a majestade de um reino.

Considerando os aspectos apresentados anteriormente, novas abordagens foram propostas para considerar relações contextuais de palavras e sentenças (DEVLIN et al., 2018, LIU et al., 2021). Os modelos com base na arquitetura *Transformers* são considerados como o estado-da-arte para representações textuais. Essa arquitetura considera a relação de contexto de palavras em sentenças, como o *Bidirectional Encoder Representations From Transformers* (BERT) (DEVLIN et al., 2018). Considerando o exemplo anterior, o modelo BERT pode compreender que a palavra “rei” é usada para indicar superioridade técnica dentro do contexto futebolístico. Nesse contexto, modelos pré-treinados do BERT geram vetores com dimensões previamente estabelecidas, nos quais esses vetores podem sofrer alterações conforme o conjunto de palavras e sentenças. No entanto, as representações vetoriais com base na arquitetura *Transformers* são singulares, isto é, valores estocásticos que são dificilmente compreendidos por um analista do domínio.

No ano de 2022, um trabalho propôs um modelo de representação vetorial de textos para atribuir pesos entre termos (palavras) e sentenças (textos) que utiliza como base BoW e

considera os modelos pré-treinados do BERT, Denominada *Term Distance From BERT* (TD-BERT) (FILHO et al., 2022). No entanto, o estudo proposto utiliza uma abordagem em que o foco foi avaliar o desempenho de uma função de rotulagem de notícias do agronegócio. Além disso, o estudo não considera diferentes conjuntos de dados para uma avaliação mais abrangente do TD-BERT.

Dessa forma, este trabalho propõe uma avaliação aprimorada do TD-BERT considerando seis representações vetoriais de textos para três *datasets* de diferentes domínios. Para comparar os resultados da representação TD-BERT são utilizadas técnicas tradicionais de TF (*Term Frequency*), TF-IDF (*Term Frequency - Inverse Document Frequency*) e vetores extraídos por meio do BERT. Além disso, quatro modelos de diferentes paradigmas de aprendizado são utilizados para as tarefas de classificação.

Esse trabalho está organizado da seguinte forma: a Seção 1 apresenta a introdução do trabalho e a Seção 2 o desenvolvimento, e, por fim, na Seção 3 as conclusões são apresentadas.

1.1 Trabalhos Relacionados

A mineração de Textos pode ser vista como a aplicação de um conjunto de técnicas usadas para analisar dados não estruturados e descobrir padrões que não eram previamente conhecidos (AGGARWAL, 2015). Em geral, o objetivo da mineração de textos é transformar documentos em representações vetoriais que possam ser utilizadas por algoritmos de aprendizado de máquina, com o intuito de extrair padrões úteis no processo de tomada de decisão (AGGARWAL, 2014).

A *Bag-of-Words* (BoW) é o método tradicional para representar textos por meio de um modelo espaço vetorial, no qual as palavras são indexadas e ponderadas de acordo com a ocorrência da palavra no texto (TURNEY; PANTEL, 2010). Muitos trabalhos utilizam a BoW como uma representação base para comparação de outros modelos espaço vetorial (SIONARA, 2018). Alguns estudos exploram diferentes abordagens para a representação de textos, analisando desde os modelos mais clássicos da BoW até os mais recentes de linguagem neural (KILANI et al., 2019, ARAUJO et al., 2020, ARAUJO et al., 2022, FILHO et al., 2022). Os trabalhos mostram que a BoW ainda obtêm resultados competitivos. No entanto, modelos de linguagem neural mostram-se mais vantajosos, pois obtêm bom desempenho de classificação, proporcionam redução significativa da dimensionalidade e lidam mais adequadamente com a proximidade semântica entre os textos.

A representação BoW é estruturada com base em palavras independentes e não expressam o relacionamento entre as palavras, sintaxe ou semântica textual (SINOARA et al., 2019). Para

superar as limitações das BoW, modelos que consideram aspectos semânticos e recursos linguísticos foram apresentados na literatura. Inicialmente, o modelo *word2vec* foi apresentado para calcular a proximidade vetorial de palavras livres de contexto (MIKOLOV et al., 2013). Modelos de linguagens neurais foram apresentados para lidar com estrutura sequencial das palavras dentro dos textos. Nesse aspecto, modelos com base na arquitetura *Transformers* são considerados como o estado-da-arte para representações vetoriais, que consideram recursos semânticos e dependência contextual de documentos de textos (DEVLIN et al. 2018, SANH et al. 2019, LIU et al. 2021). No entanto, esses modelos pré-treinados geram vetores singulares com valores que dificilmente são compreendidos na distribuição espacial.

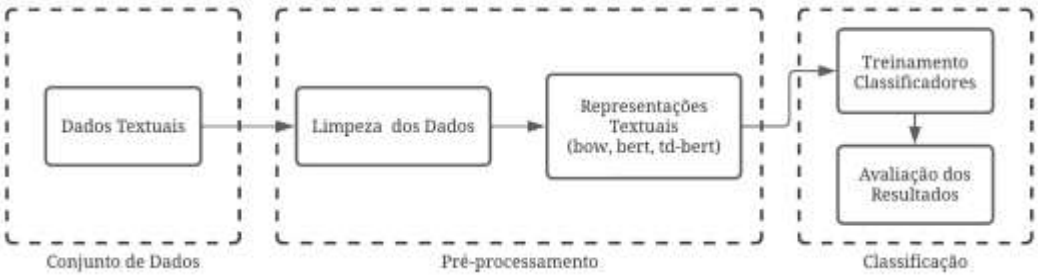
Recentemente, um trabalho propõe uma representação vetorial de textos com base na BoW e que considera a dissimilaridade de cosseno entre documentos e termos por meio de modelos pré-treinados do BERT (TD-BERT) (FILHO et al., 2022). Nove representações textuais foram utilizadas em cinco modelos de classificação. Modelos BERT e TD-BERT tiveram melhores desempenho e apresentaram resultados promissores. No entanto, o foco do trabalho foi de avaliar uma função para rotular notícias de modo automático, em que os rótulos atribuídos nas notícias podem ser imprecisos e não balanceados. Dessa forma, este estudo propõem uma avaliação fidedigna do TD-BERT, usando diferentes *datasets* balanceados e tradicionais na literatura.

2 Desenvolvimento

2.1 Métodos

Este trabalho avalia diferentes modelos de representação textual aplicados em tarefas de classificação. As principais atividades concentram-se nas etapas de obtenção do conjunto de dados, pré-processamento e avaliação experimental. Na Figura 1, é ilustrado as etapas realizadas para o desenvolvimento deste trabalho.

Figura 1 - Etapas da Avaliação Experimental



Fonte: elaborada pelos autores.

Na primeira etapa são definidos os conjuntos de dados. Em seguida, na etapa de pré-processamento são realizadas a limpeza dos dados e a criação das representações vetoriais dos

textos com BoW, modelos pré-treinados do BERT e o TD-BERT. Por fim, algoritmos de diferentes paradigmas de aprendizado são utilizados para avaliar o desempenho preditivo em tarefas de classificação. Na presente seção focou-se na etapa de pré-processamento.

Algoritmos que utilizam aprendizado de máquina para classificação de textos, exigem uma estrutura vetorial adequada para as etapas de treinamento e teste. Com o propósito de construir uma representação estruturada é necessário aplicar o pré-processamento dos dados. O processo de **Limpeza dos Dados** consiste na tarefa de tratamento, limpeza e redução no volume dos dados textuais para obter vetores que representam textos. Dessa forma, algumas técnicas são aplicadas, tais como: *i)* remoção das *stopwords*, visando eliminar alguns artigos, preposições, pronomes e conjunções que não trazem informação relevante ao contexto do documento; *ii)* a normalização, cujo objetivo é eliminar variações que a palavras podem assumir; *iii)* conversão de letras maiúsculas para minúsculas; e, *iv)* remoção de pontuação, caracteres especiais e alfanuméricos.

A próxima tarefa é gerar **Representações Textuais** usando a BoW, vetores gerados por meio do *sentence transformers* do BERT e o TD-BERT. A estrutura da *Bag-of-Words* normalmente pode ser representada por um modelo espaço vetorial, em que as palavras são indexadas e ponderadas por medidas de ocorrência entre termos e documentos (Aggarwal, 2014). Inicialmente é estabelecido uma coleção de documentos $D = \{d_1, d_2, d_3, \ldots, d_n\}$ e um conjunto de termos $T = \{t_1, t_2, t_3, \ldots, t_m\}$ extraído de D. A Tabela 1 apresenta a estrutura vetorial da BoW. Os valores dos pesos *w* são calculados na frequência dos termos t em relação ao documento d, podendo ser utilizadas as medidas mais comuns, como: *i)* binária, onde é representada de forma binária como 0 e 1, 0 como ausência e 1 como presença do termo no documento; *ii)* Frequência do Termo (*Term Frequency* (TF)) que define a frequência do termo no documento; *iii)* Frequência de Termo - Frequência de Documento Inverso (*Term Frequency-Inverse Document Frequency* (TF-IDF)) analisa o TF com a sua frequência inversa do termo no documento. As medidas TF e TF-IDF são utilizadas na avaliação deste trabalho.

Tabela 1 - Representação básica de BoW de dimensão (n, m)

	t_1	t_2	…	t_{m-1}	t_m
d_1	$w(d_1), t_1$	$w(d_1), t_2$	…	$w(d_1), t_{m-1}$	$w(d_1), t_m$
d_2	$w(d_2), t_1$	$w(d_2), t_2$	…	$w(d_2), t_{m-1}$	$w(d_2), t_m$
d_3	$w(d_3), t_1$	$w(d_3), t_2$	…	$w(d_3), t_{m-1}$	$w(d_3), t_m$
…	…	…	…	…	…
d_n	$w(d_n), t_1$	$w(d_n), t_2$	…	$w(d_n), t_{m-1}$	$w(d_n), t_m$

Fonte: elaborada pelos autores.

Na tabela 1, é ilustrado o modelo de espaço vetorial com base na BoW para uma coleção com n documentos e m atributos. Em relação aos modelos pré-treinados do BERT, a representação textual D com o *sentence transformers* é definida como $DS = ([B_1], [B_2], [B_3], \ldots, [B_n])$, em que cada vetor B é um vetor BERT de h posições, representando um documento d. Neste trabalho, o conjunto de vetores D é utilizado como entrada para os modelos de classificação, representando os modelos pré-treinados do BERT e DistilBERT (DEVLIN et al., 2018, SANH et al., 2019). Para gerar a representação TD-BERT, é necessário obter a representação vetorial para o conjunto de termos T. Dessa forma, a representação de termos com o *sentence transformers* é definida como $TS = ([W_1], [W_2], [W_3], \ldots, [W_m])$, em que W_m é um vetor BERT de h posições que representam um termo *t*. O conjunto de documentos é representado como uma matriz documento-termo constituído pela distância de cosseno c de cada vetor $n$composto de m dimensões, conforme apresentado na Tabela 2.

Tabela 2 - Representação matriz de similaridade termo-documentos

	t_1	t_2	…	t_{m-1}	t_m
d_1	$c(1,t_1)$	$c(1,t)$	…	$c(1,t_{m-1})$	$c(1,t_m)$
d_2	$c(2,t_1)$	$c(2,t_2)$	…	$c(2,t_{m-1})$	$c(2,t_m)$
d_3	$c(3,t_1)$	$c(3,t_2)$	…	$c(3,t_{m-1})$	$c(3,t_m)$
…	…	…	…	…	…
d_n	$c(n,t_1)$	$c(n,t_2)$	…	$c(n,t_{m-1})$	$c(n,t_m)$

Fonte: elaborada pelos autores.

Na representação apresentada na Tabela 2, *d* e *t* correspondem para cada documento e termo, respectivamente com vetores do *sentence transformers*. Portanto, $c(n,t_m)$ representa o cálculo da distância de cosseno entre documentos e termos. Neste trabalho, consideramos os modelos pré-treinados do BERT e DistilBERT para gerar o TD-BERT.

2.2 Configuração e Avaliação Experimental

Nessa seção é apresentada a configuração experimental utilizada para avaliar o desempenho de classificação de textos, usando seis representações vetoriais de textos. Foram considerados três conjuntos de dados para realizar os experimentos. Modelos de aprendizado de máquina de diferentes paradigmas de aprendizado foram considerados para avaliar o desempenho preditivo das representações textuais, os quais são: *Multi-Layer Perceptron* (MLP); *K-Nearest Neighbors* (KNN); *Gaussian Naive Bayes* (GNB) e *Support Vector Machine* (SVM). Nas subseções a seguir, discutimos cada etapa apresentada na Figura 1.

2.2.1 Conjuntos de Dados

O desempenho preditivo das representações textuais foi avaliado considerando três conjuntos de dados, em que foi utilizado um *dataset* de análise de sentimentos para competições de classificação disponibilizado no *Kaggle*[2] e dois *datasets* criados a partir de dados do repositório BTCSR2 (*Benchmarking of text collections from Solange, Ricardo and Rafael*)[3] (ROSSI; MARCACINI; REZENDE, 2013).

O conjunto de dados de análise de sentimentos (AS) possui índices de polaridade positiva (1) e negativa (0). Devido a grande quantidade de documentos textuais (24.888), foram extraídas mil amostras para realizar os experimentos, uma vez que o custo computacional do TD-BERT é alto com o grande volume de dados textuais. Além disso, o balanceamento dos dados foi realizado, considerando a divisão exata para a polaridade positiva (500) e negativa (500). O repositório BTCSR2 possui dados textuais coletados de diferentes segmentos. Neste trabalho, utilizamos o conjunto de textos classificados como *Industry Sector* (IS) e *Dmoz-Computers* (CP) a fim de manter a avaliação binária em alinhamento aos demais conjuntos de dados usados.

Os *datasets* disponíveis no repositório BTCSR2 possuem diversas classes de diferentes áreas. Portanto, neste trabalho selecionamos as classes *technology* e *transportation* para os experimentos (*dataset* IS), com 505 e 495 amostras, respectivamente. Utilizamos também, coleções de documentos com classes *Artificial* e *Education* do *dataset* CP, ambos com 500 amostra de cada classe.

2.2.2 Pré-processamento

O estudo apresenta a comparação entre representações de BoW, modelos pré-treinados de linguagens neurais e a abordagem proposta TD-BERT. Nessa etapa da investigação foi realizada a limpeza dos dados e a criação das representações textuais. Uma vez que os dados foram pré-processados as representações textuais são geradas com TF, TF-IDF, BERT, DistilBERT, TD-BERT e TD-DistilBERT, em que TD-DistilBERT refere-se ao uso do modelo pré-treinado DistilBERT no algoritmo.

As representações TF, TF-IDF, TD-BERT, TD-DistilBERT foram geradas considerando bigramas, isto é, cada termo (T_j) refere-se a duas palavras que ocorrem na sequência do texto. Após realizar experimentos iniciais, constatou-se que o uso de unigramas obteve representações

[2] https://github.com/GoloMarcos/BTCSR2

[3] https://www.kaggle.com/competitions/sentiment-analysis-pmr3508/overview

com menos relação contextual entre documento e termo, e por isso, não foi considerado na etapa de avaliação. As representações do TD-BERT e TD-DistilBERT foram criadas utilizando os mesmos modelos pré-treinados do BERT e DistilBERT, respectivamente. As representações vetoriais dos modelos pré-treinados de linguagens neurais foram geradas utilizando as versões multilíngue do BERT e DistilBERT. Para os modelos pré-treinados não foram aplicadas a limpeza dos dados, uma vez que o texto no formato de origem é fundamental para os modelos *Transformers*, que consideram o contexto para criar a representação vetorial do texto.

As representações vetoriais dos textos tiveram diferentes dimensões (*m*) para cada *dataset* , O TD-BERT e TD-DistilBERT apresentaram vetores de dimensão (91.823) no *dataset* AS, (114.003) em IS e (10.636) em CP. As representações TF e TF-IDF apresentaram (93.443), (114.932) e (10.574) para AS, IS e CP, respectivamente. Por fim, as representações geradas com os modelos pré-treinados tiveram vetores pré-fixados de 768 posições.

2.2.3 Classificação e Avaliação

Quatro algoritmos foram utilizados para as tarefas de classificação: *Multi-Layer Perceptron* (MLP), *K-Nearest Neighbors* (KNN), *Gaussian Naive Bayes* (GNB) e *Support Vector Machine* (SVM). Além disso, os parâmetros padrão disponibilizados pela biblioteca *scikit-learn* (PEDREGOSA et al., 2011) foram considerados para cada modelo de aprendizado de máquina.

Como estratégia de avaliação foi utilizada a validação cruzada. A técnica permite que o modelo seja avaliado com diferentes amostras de testes, em que o conjunto de dados é subdividido em grupos de tamanhos iguais, comumente chamados de *folds*. Em seguida, cada grupo é treinado e mensurado o erro pela métrica de avaliação previamente estabelecida. A avaliação dos modelos classificadores foi realizada utilizando como métrica a acurácia (Equação 1), considerada mais indicada para dados balanceados. Devido ao uso da técnica de validação cruzada, foi considerada a média entre todas as avaliações (10 folds) e o desvio padrão (Equação 2).

$$Accuracy = \frac{tp+tn}{tp+tn+fp+fn} \quad (1)$$

$$\sigma = \sqrt{\quad} \quad (2)$$

A Equação 1 apresenta o cálculo para se obter a acurácia do modelo preditivo, em que *tp* e *tn* referem-se aos valores verdadeiros positivos e verdadeiros negativos, e *fp* e *fn* representam os falsos positivos e falsos negativos. A Equação 2 refere-se ao cálculo para o desvio padrão

dos dados, em que σ representa o desvio padrão, N e μ representam o tamanho e a média do conjunto, e por fim, x_i para cada amostra dos dados.

2.3 Resultados e Discussão

Neste trabalho é apresentada uma investigação comparativa entre modelos de representação textual para a classificação de textos. Além disso, uma abordagem para a criação de representações que consideram aspectos semânticos é apresentada, o TD-BERT. Na Tabela 3 é apresentado os resultados considerando o *dataset* AS. Os valores em negrito representam o melhor resultado para cada algoritmo de classificação e os resultados sublinhados para cada representação textual.

Tabela 3 - Resultados obtidos com o conjunto de dados AS

Modelo	TF	TF-IDF	BERT	Dist	TD-BERT	TD-Dist
MLP	0.70 (0.06)	0.70 (0.07)	**0.76 (0.05)**	0.75 (0.03)	0.54 (0.08)	0.50 (0.01)
SVM	0.53 (0.03)	0.55 (0.06)	0.75 (0.05)	**0.76 (0.04)**	0.75 (0.04)	0.73 (0.05)
KNN	0.52 (0.02)	0.62 (0.05)	**0.68 (0.05)**	0.64 (0.03)	0.67 (0.04)	0.60 (0.03)
GNB	0.67 (0.05)	0.67 (0.04)	**0.70 (0.03)**	0.70 (0.05)	0.59 (0.05)	0.55 (0.05)

Fonte: elaborada pelos autores.

Considerando os resultados apresentados na Tabela 3, as representações pré-treinadas BERT e DistilBERT obtiveram a melhor acurácia em relação às demais. BERT apresentou 0.76 no modelo MLP, 0.68 no KNN e 0.70 para GNB. DistilBERT foi superior com 0.76 no modelo SVM. Nesse experimento, a representação TD-BERTs apresentou resultados similares aos modelos de linguagens neurais, TD-BERT obteve 0.75 e TD-DistilBERT 0.73, ambos com o modelo SVM. Na Tabela 4 é apresentado os resultados considerando o *dataset* CP.

Tabela 4 - Resultados obtidos com o conjunto de dados CP

Modelo	TF	TF-IDF	BERT	Dist	TD-BERT	TD-Dist
MLP	0.91 (0.02)	*0.92 (0.02)*	**0.97 (0.02)**	0.97 (0.03)	0.72 (0.24)	0.63 (0.21)
SVM	0.83 (0.04)	0.92 (0.02)	**0.97 (0.01)**	0.96 (0.01)	0.95 (0.01)	0.95 (0.01)
KNN	0.57 (0.03)	0.91 (0.02)	0.95 (0.01)	**0.96 (0.02)**	0.94 (0.01)	0.94 (0.02)
GNB	0.90 (0.03)	0.90 (0.03)	0.95 (0.02)	**0.96 (0.01)**	0.92 (0.02)	0.91 (0.01)

Fonte: elaborada pelos autores.

Conforme os resultados apresentados na Tabela 4, BERT e DistilBERT tiveram o melhor desempenho preditivo, em que BERT obteve 0.97 de acurácia com os algoritmos MLP e SVM. O DistilBERT foi superior para os modelos KNN e GNB com 0.96 de acurácia. Os resultados obtidos da representação BoW tiveram desempenho inferior aos demais resultados. Por outro lado, o TD-BERT e TD-DistilBERT apresentaram resultados similares aos modelos pré-treinados, em que tiveram 0.95 e 0.94 nos algoritmos SVM e KNN, respectivamente. Na Tabela 5 é apresentado os resultados considerando o *dataset* IS.

Tabela 5 - Resultados obtidos com o conjunto de dados IS

Modelo	TF	TF-IDF	BERT	Dist	TD-BERT	TD-Dist
MLP	**0.96 (0.01)**	0.96 (0.02)	0.90 (0.01)	0.89 (0.02)	0.55 (0.12)	0.52 (0.05)
SVM	0.75 (0.04)	**0.92 (0.02)**	0.86 (0.03)	0.85 (0.02)	0.82 (0.03)	0.81 (0.04)
KNN	0.66 (0.03)	0.56 (0.02)	**0.80 (0.03)**	0.80 (0.04)	0.76 (0.04)	0.75 (0.04)
GNB	0.96 (0.03)	**0.96 (0.01)**	0.78 (0.04)	0.75 (0.02)	0.63 (0.04)	0.59 (0.06)

Fonte: elaborada pelos autores.

Conforme os resultados apresentados na Tabela 5, TF e TF-IDF tiveram o melhor desempenho preditivo entre todas as representações nos algoritmos MLP, SVM e GNB com acurácia acima de 0.90. Para o modelo BERT, o modelo KNN alcançou o melhor resultado com 0.80 de acurácia. TD-BERT e TD-DistilBERT apresentaram resultados similares entre eles, em que o melhor desempenho foi obtido no algoritmo SVM com 0.82 de acurácia.

De modo geral, os modelos de linguagens neurais alcançaram o melhor desempenho preditivo entre todas as representações textuais para os *datasets* AS e CP. Por outro lado, o *dataset* IS obteve resultados melhores com as representações TF e TF-IDF. Além disso, os algoritmos MLP e SVM se sobressaíram aos demais nos três conjuntos de dados. No entanto, TD-BERT e TD-DistilBERT não apresentaram bons resultados com o algoritmo MLP, em que obtiveram menos de 0.70 de acurácia média.

3 Conclusões

Foi apresentada uma avaliação comparativa de representações textuais que considera recursos semânticos. Três conjuntos de dados de diferentes temas foram utilizados para avaliar o desempenho preditivo em tarefas de classificação. Para a avaliação, foram selecionados quatro algoritmos de diferentes paradigmas de aprendizagem e seis modelos de representação textual, considerando a abordagem proposta TD-BERT.

Em geral, as representações com base em modelos de linguagens neurais apresentam desempenho superior aos modelos de BoW. No entanto, a abordagem mais recente (TD-BERTs) se mostrou eficaz e obteve resultados similares aos demais, sendo superior às representações BoW na maioria dos casos. Contudo, assim como as representações de BoW, a abordagem proposta se limita a utilização de dados textuais curtos, visto que esses modelos consideram todas as palavras presentes no conjunto de dados. Além disso, devido ao uso de modelos neurais, utilizar unigramas (uma palavra como termo) pode não apresentar similaridade semântica entre documentos e termos.

Em trabalhos futuros, técnicas para a redução da dimensionalidade das representações do TD-BERT podem ser desenvolvidas, a fim de obter atributos semanticamente mais

representativos para o domínio textual. Além disso, a representação TD-BERT pode ser utilizada para avaliar a explicabilidade do modelo preditivo, realizando avaliações comparativas em diferentes abordagens.

Agradecimentos: Os autores agradecem ao Centro Nacional de Desenvolvimento Científico e Tecnológico (CNPq), à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) (Processo PCRH BPG-00054-210) e ao Programa de Bolsas de Produtividade em Pesquisa da Universidade do Estado de Minas Gerais (PQ/UEMG).

Referências

AGGARWAL, C. C. **Data Classification: Algorithms and Applications.** 1. ed. [S.l.]: Chapman & Hall/CRC, 2014.

AGGARWAL, C. C. Mining text data. In: SPRINGER. **Data mining**. [S.l.], 2015. p. 429–455.

AGGARWAL, C. Machine Learning for Text. 1st. ed. United States: **Springer Publishing Company**, Incorporated, 2018.

ARAUJO, A. *et al*. From bag-of-words to pre-trained neural language models: Improving automatic classification of app reviews for requirements engineering. In: SBC. **Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional.** [S.l.], 2020. p. 378–389. Disponível em:https://sol.sbc.org.br/index.php/eniac/article/view/12144. Acesso em: 26 ago. 2023.

ARAUJO, A. *et al*. Opinion mining for app reviews: an analysis of textual representation and predictive models. **Automated Software Engineering**, Springer, v. 29, n. 1, p. 1–30, 2022. Disponível em:https://doi.org/10.1007/s10515-021-00301-1. Acesso em: 26 ago. 2023.

DEVLIN, J. *et al*. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018. Disponível em: https://doi.org/10.48550/arXiv.1810.04805. Acesso em: 26 ago. 2023.

FILHO, I. J. *et al*. Sequential short-text classification from multiple textual representations with weak supervision. In: **Brazilian Conference on Intelligent Systems**. Cham: Springer International Publishing, 2022. p. 165-179. Disponível em: https://link.springer.com/chapter/10.1007/978-3-031-21686-2_12. Acesso em: 27 ago. 2023.

JANEV, V. *et al*. **Knowledge graphs and big data processing**. Cham-Suíça: Springer Nature, 2020. Disponível em: https://link.springer.com/book/10.1007/978-3-030-53199-7. Acesso em: 27 ago. 2023.

KILANI, N. A. *et al*. Automatic classification of apps reviews for requirement engineering: Exploring the customers need from healthcare applications. In: IEEE. **2019 sixth international conference on social networks analysis, management and security (SNAMS)**, Granada, Spain, 2019, pp. 541-548. Disponível em: https://ieeexplore.ieee.org/document/8931820. Acesso em: 27 ago. 2023.

LIU, Z. *et al*. A robustly optimized BERT pre-training approach with post-training. In: **China National Conference on Chinese Computational Linguistics**. Cham-Suiça: Springer

International Publishing, 2021. p. 471-484. Disponível em: https://link.springer.com/chapter/10.1007/978-3-030-84186-7_31. Acesso em: 27 ago. 2023.

MIKOLOV, T. *et al*. Distributed representations of words and phrases and their compositionality. **Advances in neural information processing systems**, v. 26, 2013. Disponível em: https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf. Acesso em: 27 ago. 2023.

PEDREGOSA, F. *et al*. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011. Disponível em: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?ref=https:/. Acesso em: 27 ago. 2023.

REZENDE, S. O. *et al*. Mineração de dados**.** In: REZENDE, S. O. (Org.). **Sistemas Inteligentes:** Fundamentos e Aplicações. Barueri-SP: Manole, 1ª edição, 2003. Cap. 12, p. 307–335.

ROSSI, R. G.; MARCACINI, R. M.; REZENDE, S. O. **Benchmarking text collections for classification and clustering tasks**. São Carlos-SP: Instituto de Ciências Matemáticas e de Computação-IMC2, Icmc Technical Report n° 393, 2013. Disponível em:https://repositorio.usp.br/bitstreams/342060e9-eebc-4530-8074-bd60bb8b125e. Acesso em: 27 ago. 2023.

SANH, V. *et al*. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. **arXiv preprint arXiv:1910.01108**, Cornell University, 2019. Disponível em: https://arxiv.org/abs/1910.01108. Acesso em: 27 ago. 2023.

SINOARA, R. A. *et al*. Knowledge-enhanced document embeddings for text classification. **Knowledge-Based Systems**, Elsevier, v. 163, p. 955–971, 2019. Disponível em: https://doi.org/10.1016/j.knosys.2018.10.026. Acesso em: 27 ago. 2023.

SIONARA, R. A. **Aspectos semânticos na representação de textos para classificação automática**. Tese (Doutorado Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional-PPG/CCMC) - Universidade de São Paulo-USP São Carlos, 2018. Disponível em: https://www.teses.usp.br/teses/disponiveis/55/55134/tde-10102018-143520/publico/RobertaAkemiSinoara_revisada.pdf. Acesso em: 27 ago. 2023.

TURNEY, P. D.; PANTEL, P. From frequency to meaning: Vector space models of semantics. **Journal of artificial intelligence research**, v. 37, p. 141–188, 2010. Disponível em: https://doi.org/10.1613/jair.2934. Acesso em: 27 ago. 2023.