Revista Eletrônica de Iniciação Científica em Computação, 2026, 24:1
ISSN: 3085-8461 • doi: 10.5753/reic.2026.8472

Artigo de Pesquisa/Research Paper

Pondere e Expanda: Impacto e Limitações de Representações Contextual-Esparsas na Modelagem de Tópicos

Weigh and Expand: Impact and Limitations of Contextual Sparse Representations in Topic Modeling

Ana Cláudia Machado [Universidade Federal de São João del Rei | *anaclaudiamachado211@aluno.ufsj.edu.br*]
Celso França [Universidade Federal de Minas Gerais | *celsofranca@dcc.ufmg.br*]
Marcos André Gonçalves [Universidade Federal de Minas Gerais | *mgoncalv@dcc.ufmg.br*]
Leonardo Rocha [Universidade Federal de São João del Rei | *lcrocha@ufsj.edu.br*]

Departamento de Computação, Universidade Federal de São João del Rei, Rodovia 494, s/n, Bairro Colônia do Bengo, São João del-Rei, MG, 36301-360, Brasil.

Resumo. Este trabalho investiga o uso de estratégias baseadas em representações contextual-esparsas na tarefa de Modelagem de Tópicos (MT), buscando conciliar a interpretabilidade das representações esparsas com a riqueza semântica proporcionada pelo contexto. Para isso, empregamos o modelo SPLADE, que representa documentos de maneira sensível ao contexto por meio de mecanismos de expansão e ponderação de termos. A abordagem é avaliada empiricamente em comparação com outras formas de representação, utilizando uma métrica tradicional e considerando conjuntos de dados distintos. Os resultados evidenciam que a etapa de ponderação favorece uma MT eficaz, enquanto a expansão, apesar de promissora, ainda apresenta restrições decorrentes da incompatibilidade entre o vocabulário da representação e os textos originais.

Abstract. This work investigates the use of strategies based on contextual-sparse representations for the task of Topic Modeling (TM), aiming to reconcile the interpretability of sparse representations with the semantic richness provided by context. To this end, we employ the SPLADE model, which represents documents in a context-sensitive manner through mechanisms of term expansion and weighting. The approach is empirically evaluated in comparison with other forms of representation, using a traditional metric and considering distinct datasets. The results show that the weighting stage supports effective TM, whereas expansion, although promising, still presents limitations arising from the incompatibility between the representation vocabulary and the original texts.

Palavras-chave: Modelagem de Tópicos, Representações Contextual-Esparsas, Ponderação de Termos, Expansão de Termos

Keywords: Topic Modeling, Contextual Sparse Representations, Term Weighting, Term Expansion

Recebido/Received: 15 June 2026 • **Aceito/Accepted:** 15 June 2026 • **Publicado/Published:** 10 July 2026

1 Introdução

A Modelagem de Tópicos (MT) é uma técnica não supervisionada de aprendizado de máquina voltada à identificação de padrões temáticos latentes em grandes coleções textuais [Viegas *et al*., 2019]. Ao agrupar termos semanticamente relacionados, essa abordagem reduz a dimensionalidade dos dados e facilita sua organização, exploração e análise [Churchill and Singh, 2022; Abdelrazek *et al*., 2023]. As estratégias de MT abrangem desde modelos probabilísticos, como o *Latent Dirichlet Allocation* (LDA) [Blei *et al*., 2003], até métodos determinísticos, como a Fatoração de Matrizes Não Negativas (NMF) [Kuang *et al*., 2015], além de abordagens mais recentes fundamentadas em representações densas, como CluWords [Viegas *et al*., 2019, 2020, 2025], BERTopic [Grootendorst, 2022] e Contextualized Topic Model(CTM) [Bianchi *et al*., 2021].

As técnicas de modelagem de tópicos diferem principalmente quanto à forma de representação textual adotada. Métodos tradicionais, como LDA e NMF, baseiam-se em representações esparsas e estáticas fundamentadas na frequência de termos, o que limita a captura de relações semânticas. O CluWords introduz representações densas e estáticas capazes de incorporar similaridade semântica, porém ainda descontextualizados. O BERTopic, por sua vez, representa um avanço ao empregar vetores densos e contextuais derivados de modelos *transformer*, capazes de capturar o significado das palavras em função do contexto em que ocorrem. Apesar desses progressos, permanece uma lacuna na exploração de representações que sejam simultaneamente esparsas e contextuais, potencialmente capazes de conciliar interpretabilidade e riqueza semântica.

Este trabalho propõe o uso de representações contextual-esparsas na MT, com o objetivo de incorporar sensibilidade ao contexto, preservar interpretabilidade e promover maior coerência temática. Para isso, empregamos o SPLADE [Formal *et al*., 2022], uma abordagem baseada em *Masked Language Modeling* (MLM) que transforma *embeddings* densos provenientes de modelos *transformer* em representações esparsas interpretáveis. O MLM consiste em mascarar *tokens* do texto e treinar o modelo para predizê-los a partir do contexto, produzindo distribuições que refletem a relevância contextual dos termos. Tais distribuições atribuem maior peso à palavra mascarada e a termos semanticamente relacionados, pos-

sibilitando expansão semântica. A proposta fundamenta-se em dois componentes principais: (i) expansão contextual de termos e (ii) ponderação contextual, que ajusta a importância dos termos conforme sua relevância no contexto. A partir disso, investigamos as seguintes questões de pesquisa:

QP1 Qual a efetividade do uso de representações contextual-esparsas comparadas a outras representações da literatura em Modelagem de Tópicos?

QP2 Qual o impacto da expansão de termos e da ponderação de termos da representação contextual-esparsa na Modelagem de Tópicos?

Nossos resultados experimentais demonstram que as representações contextual-esparsas, quando integradas à NMF, alcançam desempenho em MT que são tão efetivos quanto ou superiores aos melhores métodos da literatura em diferentes coleções de dados. Verificamos que esse desempenho está fortemente relacionado à qualidade das **ponderações contextuais** atribuídas aos termos. Em contrapartida, a exploração do componente de expansão ainda demanda o desenvolvimento de estratégias mais eficazes, uma vez que sua aplicação atual tende a introduzir muito ruído. A análise conduzida indica que essa limitação decorre, sobretudo, da incompatibilidade entre os *subtokens* do vocabulário do modelo e os termos presentes nos documentos originais, configurando um desafio relevante e aberto para investigações futuras. Esse trabalho de iniciação científica resultou na publicação de um artigo no SBBD 2025 (A4) Machado *et al*. [2025].

2 Revisão da Literatura

Na Tabela 1, apresentamos uma comparação das representações vetoriais empregadas em MT, organizando-as de acordo com dois eixos fundamentais: esparsidade e sensibilidade ao contexto.

Representações estático-esparsas, como o TF-IDF, destacam-se pelas elevadas interpretabilidade e escalabilidade [Gao *et al*., 2024]; contudo, apresentam limitações importantes, como alta dimensionalidade e baixa coerência temática, decorrentes do tratamento isolado dos termos e da desconsideração de relações semânticas. Como consequência, podem produzir agrupamentos baseados em frequência, porém semanticamente incoerentes [Abdelrazek *et al*., 2023; Doogan and Buntine, 2021]. Por outro lado, representações estático-densas, como *word2vec*, *fastText* e *GloVe*, reduzem a dimensionalidade e capturam similaridades semânticas, mas permanecem insensíveis ao contexto [Arora *et al*., 2020]. Essa limitação restringe sua eficácia em MT e reduz sua interpretabilidade. Nesse cenário, modelos como o CluWords, fundamentados nessas representações, requerem estratégias adicionais, como expansão de palavras baseada em similaridade e filtragem de ruído, para melhorar seu desempenho [Viegas *et al*., 2019].

Representações contextuais-densas, como as derivadas de BERT e RoBERTa, capturam a semântica de forma contextualizada e tendem a promover maior coerência temática; entretanto, apresentam limitações relacionadas à interpretabilidade, à escalabilidade e à elevada demanda computacional [Liang *et al*., 2022]. Métodos como o BERTopic, por exemplo, dependem de múltiplas etapas de pós-processamento, como a redução de dimensionalidade dos embeddings semânticos para facilitar a organização dos dados, a clusterização de documentos semanticamente semelhantes para formar tópicos coerentes e a identificação de palavras-chave representativas em cada cluster, etapas que, em conjunto, são essenciais para produzir resultados interpretáveis e satisfatórios.

Por sua vez, representações contextuais esparsas, como as do SPLADE [Formal *et al*., 2022], conciliam a interpretabilidade e a eficiência características da esparsidade com a expressividade semântica derivada do contexto, permitindo a expansão semântica de documentos. A exploração dessa abordagem, ainda pouco investigada em MT, configura um avanço promissor ao buscar um equilíbrio entre coerência temática e interpretabilidade dos resultados.

3 Abordagem Proposta

Nossa proposta fundamenta-se no uso de representações contextuais-esparsas em MT, com ênfase no método SPLADE. O pipeline adotado, ilustrado na Figura 1, gera, a partir dos textos de entrada, uma matriz de representações $V \in \mathbb{R}_{>0}^{n \times m}$, na qual n denota o número de documentos e m o tamanho do vocabulário. Essa matriz é subsequentemente fatorada por meio da técnica de NMF, possibilitando a extração de k tópicos latentes. Na sequência, descrevemos detalhadamente cada um dos componentes do processo.

3.1 Representação Contextual-Esparsa

O SPLADE [Formal *et al*., 2022] é um modelo baseado em *transformers* que emprega uma estratégia de MLM para produzir *embeddings* contextuais, conforme ilustrado na Figura 2. No MLM, partes do texto são mascaradas, e o modelo é treinado para prever os *tokens* ausentes a partir do contexto [Devlin *et al*., 2019]. A geração das representações ocorre em três etapas principais. Primeiro, o texto é *tokenizado*, e alguns *tokens* são aleatoriamente substituídos por `[MASK]`. A sequência resultante é convertida em identificadores numéricos e processada pelas camadas do *transformer*, cujo mecanismo de *self-attention* bidirecional produz *embeddings* contextuais densos (EMB-1, EMB2 ... EMB-6) para cada *token*, considerando suas interações com toda a sequência. Em seguida, na segunda etapa, esses *embeddings* são projetados pela *MLM Head*, que gera, para cada posição, uma distribuição de probabilidade sobre o vocabulário, indicando a probabilidade de ocorrência de cada *token* dado o contexto.

Por fim, na terceira etapa, tais distribuições são agregadas em uma única representação vetorial esparsa por meio de uma função de ativação com saturação logarítmica combinada a uma operação de *max pooling*, resultando em uma estimativa de importância w_j para cada *token* j do vocabulário:

$$w_j = \max_{i \in t} \log(1 + \mathrm{ReLU}(w_{ij}))\,, \tag{1}$$

em que t denota o conjunto de *tokens* da sequência de entrada, enquanto i e j indexam, respectivamente, as posições dessa sequência e os *tokens* do vocabulário. O termo w_{ij} corresponde ao peso atribuído ao *token* j do vocabulário em função do contexto capturado a partir do *token* de entrada na posição i. Esse procedimento combina as distribuições produzidas pelo MLM em um único vetor esparso que expressa a relevância de cada *token* do vocabulário em

Propriedade	Representação Vetorial			
	Estática		Contextual	
Esparsidade	Esparso	Denso	Esparso	Denso
Exemplo	TF-IDF	Word2Vec fastText GloVE	SPLADE	BERT RoBERTa
Sensível ao contexto	Não	Não	Sim	Sim
Dimensionalidade	Muito alta	Média (50-300)	Muito alta	Médio-Alta (768-2048)
Escalabilidade	Excelente	Excelente	Moderado	Moderado
Aplicabilidade em Modelagem de Tópicos	Boa	Limitado	**Inexplorada**	Limitado
Interpretabilidade dos Tópicos	Alta	**Baixa**	**Inexplorada**	Baixa
Coerência do Tópico	Baixo	Médio	Alto potencial	Alta

Tabela 1. Comparação entre diferentes representações vetoriais

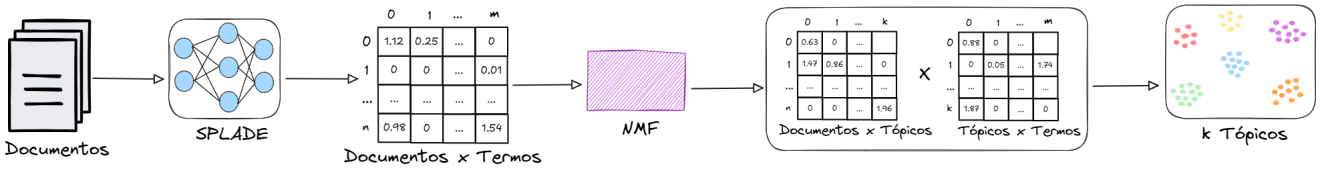

Figura 1. Pipeline de modelagem de tópicos.

relação ao texto de entrada, ao mesmo tempo em que controla a quantidade de valores não nulos na representação.

O SPLADE assume que *tokens* com maior probabilidade de ocorrência são semanticamente mais relevantes. Dessa forma, termos ausentes do texto original podem integrar a representação vetorial, promovendo expansão semântica. O vetor resultante, com dimensionalidade igual à do vocabulário, expressa a importância relativa de cada *token* na caracterização do documento.

3.2 Modelagem de Tópicos

Na etapa de modelagem, empregamos a técnica de NMF devido à sua capacidade de produzir representações interpretáveis, decorrente das restrições de não negatividade que favorecem a identificação de componentes latentes semanticamente coerentes, como os tópicos [Lee and Seung, 1999]. Como entrada para o NMF, são fornecidos o número k de tópicos a serem inferidos e a matriz V, composta pelas representações contextuais esparsas dos documentos. O NMF decompõe V em duas matrizes não negativas, $W \in \mathbb{R}_{>0}^{n \times k}$ e $H \in \mathbb{R}_{>0}^{k \times m}$, tais que

$$V \approx WH, \tag{2}$$

conforme ilustrado na Figura 1. A matriz W permite identificar os tópicos predominantes em cada documento, enquanto H revela os termos mais representativos de cada tópico.

4 Metodologia Experimental

Nesta seção, descrevemos as configurações utilizadas em nossos experimentos.

4.1 Bases de Dados e Pré-processamento

Utilizamos três coleções de dados amplamente conhecidas na literatura: **ACM**, composta por artigos científicos publicados na *ACM Digital Library* e organizados em 11 classes temáticas; *20News Group* - **20NG**, formada por postagens provenientes de grupos de notícias, distribuídas em 20 categorias; e *Web of Science Platform* - **WOS**, que reúne artigos científicos indexados na *Web of Science Platform*, abrangendo 33 classes. Nos experimentos, o número de tópicos considerado é definido de acordo com a quantidade de classes existente em cada base original, permitindo uma correspondência direta entre a estrutura temática inferida e a categorização de referência.

Em todas as bases, realizamos uma etapa padronizada de preparação textual composta por: (1) conversão dos termos para minúsculas; (2) remoção de *stopwords* em inglês; (3) eliminação de números e sinais de pontuação; e (4) descarte de palavras com menos de três caracteres [Júnior *et al.*, 2022]. Esses procedimentos reduzem o ruído lexical, tornam o vocabulário mais consistente e favorecem uma comparação mais adequada entre os métodos analisados.

4.2 Configuração Experimental

Avaliaremos diferentes abordagens de MT que se distinguem principalmente pelo tipo de representação vetorial utilizada. Especificamente, consideramos **LDA** e **NMF** baseados em vetores estáticos e esparsos; **Cluwords**, que emprega representações estáticas densas; e **BERTopic**, fundamentado em embeddings contextuais densos. Para o NMF, adotamos a estratégia de inicialização *Nonnegative Double SVD* [Boutsidis and Gallopoulos, 2008] e, para o LDA, o método *Online Variational Bayes* [Ghahramani and Attias, 2000]. No Cluwords, seguimos as configurações recomendadas nos trabalhos originais [Viegas *et al.*, 2019, 2025]. Já no BERTopic, utilizamos os parâmetros padrão, com o modelo *all-MiniLM-L6-v2* para geração dos *embeddings* e o algoritmo K-Means na etapa de agrupamento.

No que se refere à nossa proposta, combinamos o NMF com representações contextuais esparsas geradas pelo SPLADE e avaliamos três cenários distintos: (i) **Padrão**, que utiliza integralmente as ponderações e expansões atribuídas aos documentos; (ii) **Corte Global**, no qual são descartados

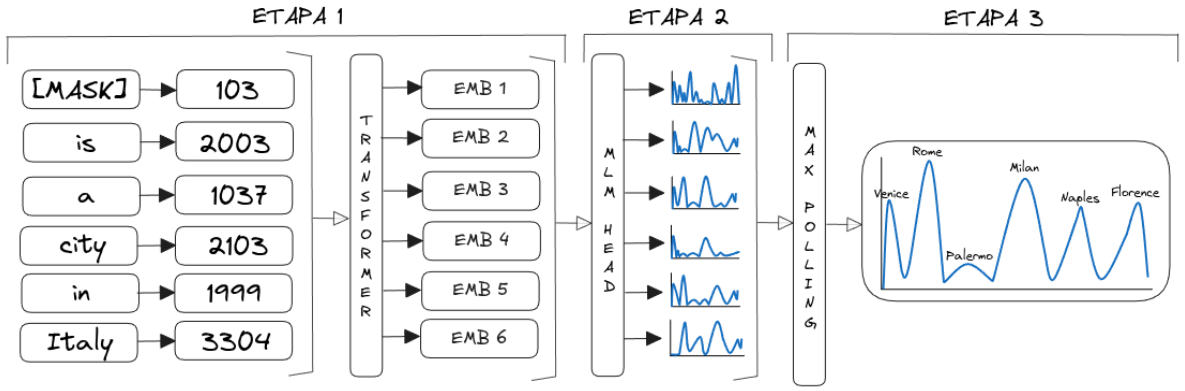

Figura 2. Representações contextuais-esparsas são geradas a partir dos *embeddings* de um *transformer* (etapa 1), processadas por uma *MLM Head* (etapa 2) e agregadas por *max pooling* (etapa 3).

os 40% menores valores da matriz de representação; e (iii) **Interseção**, que considera apenas os termos originalmente presentes em cada documento.

No cenário de Corte Global, conduzimos experimentos com diferentes percentuais de remoção e verificamos que a exclusão dos 40% menores valores oferece um bom equilíbrio entre a redução de ruído e a preservação de informações relevantes. O objetivo central dessa análise é compreender, de forma isolada, o impacto das ponderações e das expansões produzidas pelo SPLADE. Ao restringir os valores apenas aos termos presentes no documento, no cenário de Interseção isolamos o efeito das ponderações, assegurando que nenhuma expansão seja considerada. Por outro lado, ao aplicar um corte global nos menores valores da matriz, investigamos em que medida esses valores podem ser interpretados como ruído na representação.

4.3 Métrica de Avaliação e Validação Estatística

Para avaliar o desempenho dos algoritmos de TM, utilizamos uma métrica tradicional amplamente adotada na literatura, que mensura a relevância das palavras que compõem cada tópico a partir do ganho de informação associado à coocorrência de pares de palavras em documentos. Nesse contexto, empregamos o *Normalized Pointwise Mutual Information* (NPMI) [Bouma, 2009], definido pela Equação 3:

$$\text{NPMI}(w_1, w_2) = \frac{\log\left(\frac{P(w_1,w_2)}{P(w_1)P(w_2)}\right)}{-\log(P(w_1,w_2))} \tag{3}$$

em que w_1 e w_2 representam um par de palavras, $P(w_1)$ e $P(w_2)$ correspondem às probabilidades individuais de ocorrência dessas palavras, e $P(w_1, w_2)$ denota a probabilidade de coocorrência do par ao longo do corpus. O valor do NPMI varia no intervalo $[0, 1]$: valores próximos de 0 indicam independência entre as palavras, enquanto valores próximos de 1 sugerem forte associação decorrente de coocorrência frequente. Cada tópico é representado por um conjunto de 10 palavras, e a significância estatística dos resultados é avaliada por meio do teste t com correção de Bonferroni.

5 Resultados Experimentais

As Figuras 3, 4 e 5 apresentam os resultados da comparação entre algoritmos de MT baseados em diferentes representações e as propostas contextuais-esparsas (QP1), utilizando a métrica *NPMI*. As médias dos modelos estão destacadas em vermelho, enquanto os valores por tópico aparecem em azul. Os resultados em negrito indicam os melhores desempenhos em cada coleção, de acordo com o teste t pareado, sendo também destacados os casos de empate estatístico.

Constatamos que o desempenho do BERTopic está em conformidade com a literatura, mantendo-se como estado da arte em MT. Em comparação com LDA, NMF e CluWords, o BERTopic apresenta os maiores valores de *NPMI* ou empata estatisticamente com o melhor resultado. Além disso, destaca-se por apresentar baixa dispersão entre os tópicos, o que indica maior consistência temática.

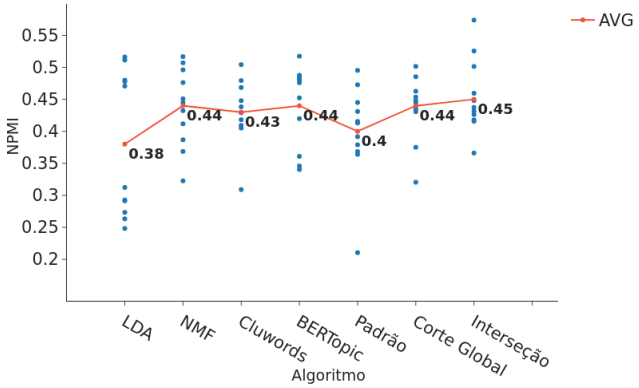

Figura 3. Valores de NPMI para diferentes algoritmos na base ACM.

No que se refere às abordagens propostas, a representação **Interseção** alcançou os maiores valores absolutos de *NPMI* nas coleções ACM e WOS, além de apresentar desempenho comparável ao melhor método na 20NG (NMF), posicionando-se, assim, em nível equivalente ao estado da arte em MT. Evidenciamos, ainda, menor dispersão dos valores por tópico em comparação ao BERTopic, o que sugere maior consistência temática. A Tabela 2 complementa essa análise ao ilustrar, de forma qualitativa, três exemplos de tópicos bem definidos, um para cada coleção, gerados pela abordagem **Interseção**. Em conjunto, esses resultados indicam que *representações contextuais-esparsas são eficazes e comparáveis às melhores abordagens reportadas*

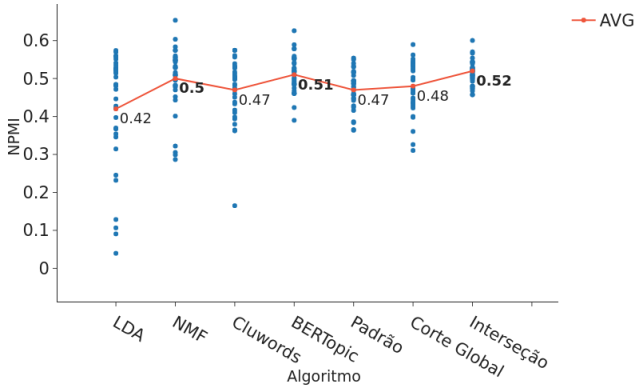

Figura 4. Valores de NPMI para diferentes algoritmos na base WOS.

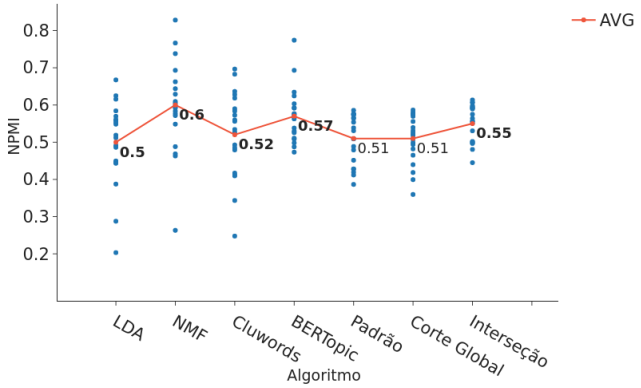

Figura 5. Valores de NPMI para diferentes algoritmos na base 20NG.

na literatura em diferentes cenários, respondendo à QP1.

Ao comparar o desempenho das três abordagens contextuais-esparsas, observamos que as estratégias que realizam algum tipo de expansão (isto é, Padrão e Corte Global) não apresentam ganhos em relação à Interseção e, por consequência, tampouco superam os demais métodos avaliados. Esse resultado sugere que, *embora a ponderação de termos fornecida pela representação contextual-esparsa permita uma modelagem de tópicos eficaz, o processo de expansão ainda não se mostra capaz de agregar informações contextuais que resultem em melhorias de desempenho na abordagem proposta, respondendo à QP2.* Para compreender esse comportamento, analisamos detalhadamente essa expansão.

Para cada documento de cada coleção, foram selecionados os termos (*tokens*) com pesos atribuídos pelo SPLADE e, em seguida, calculada a proporção daqueles que não pertenciam ao vocabulário da coleção. O resultado dessa análise é apresentado na Figura 6, em que o eixo X indica a parcela de *tokens* fora do vocabulário e o eixo Y representa a frequência de documentos associada a cada parcela. Os gráficos evidenciam que uma fração considerável dos *tokens* relevantes não integra o vocabulário do conjunto de documentos. Entre esses casos, observamos a ocorrência frequente de *subtokens*, isto é, unidades menores derivadas do processo de *tokenização* de palavras ausentes no vocabulário. O mapeamento reverso desses *subtokens* para palavras completas é, contudo, ambíguo e ruidoso, uma vez que um mesmo *subtoken* pode compor diferentes palavras. Como exemplo, destaca-se o *subtoken "##ing"*, presente em diversas palavras do inglês, como *reconfiguring* e *embedding*. Isoladamente, esse fragmento não carrega significado completo e pode ocorrer em múltiplos vocábulos distintos, o que dificulta a reconstrução precisa do vocabulário original e compromete a interpretação semântica dos tópicos.

Coleção	Tópico	Palavras
20NG	Religion.Christianism	god, christian, bible, jesus, church
ACM	Computing.Mathematics	algorithm, problem, time, linear, complexity
WOS	Network Security	network, communication, security, paper, detection

Tabela 2. Exemplo de tópicos por classe nas coleções 20NG, ACM e WOS

6 Conclusão e Direções Futuras

Este trabalho investigou o uso de representações contextuais-esparsas na MT, com ênfase nos mecanismos de ponderação e expansão de termos, utilizando o modelo SPLADE. Os resultados experimentais demonstraram que a ponderação contextual contribui de forma consistente para a efetividade da modelagem, permitindo alcançar desempenho comparável ao estado da arte em diferentes coleções de dados. Por outro lado, a etapa de expansão apresentou limitações, decorrentes principalmente da incompatibilidade entre o vocabulário baseado em *tokens* e *subtokens* dos modelos *transformer* e as palavras presentes nos documentos originais, o que introduz ruído semântico e compromete a interpretação dos tópicos.

Esses achados indicam que representações contextual-esparsas constituem uma alternativa promissora para conciliar interpretabilidade e riqueza semântica em MT, especialmente quando a informação contextual é explorada por meio de ponderações mais precisas. Ao mesmo tempo, evidenciam que a eficácia da expansão depende diretamente do alinhamento entre o espaço de representação do modelo e o vocabulário do domínio analisado, configurando um desafio ainda em aberto.

Como direções futuras, propomos investigar estratégias que mitiguem essa incompatibilidade entre vocabulários, incluindo a adaptação do vocabulário do modelo para domínios específicos em conjunto com o treinamento de representações esparsas diretamente sobre vocabulários-alvo. Para isso, planejamos explorar tarefas de mascaramento não supervisionado que permitam ao modelo aprender distribuições de probabilidade alinhadas ao vocabulário de cada coleção, produzindo representações ajustadas ao conjunto de dados e semanticamente mais adequadas ao domínio analisado.

Declarações complementares

Financiamento

Esta pesquisa foi financiada por: CNPq, Capes, Fapemig, Fapesp, AWS, NVIDIA, CIIA-Saúde e Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (INCT-TILD-IAR; 408490/2024-1).

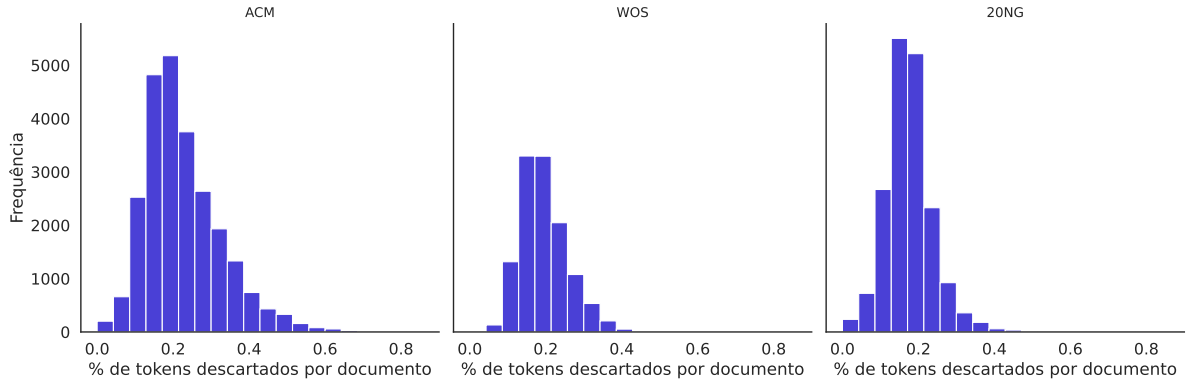

Figura 6. % de *tokens* descartados por documento em uma distribuição de frequência por coleção

Contribuições dos autores

A. C. Machado, C. França, M. A. Gonçalves e L. Rocha contribuíram para a concepção do estudo (Conceptualization) e análise formal dos resultados (Formal analysis). A. C. Machado foi a responsável pelo desenvolvimento dos códigos (Software), investigação (Investigation) e escrita do rascunho original (Writing – original draft). L. Rocha atuou na supervisão da pesquisa (Supervision). C. França, M. A. Gonçalves colaboraram na validação da metodologia (Validation). Todos os autores participaram da revisão (Writing – review & editing) e aprovaram o manuscrito final.

Conflitos de interesse

Os autores declaram que não têm nenhum conflito de interesses.

Disponibilidade de dados e materiais

Os conjuntos de dados (e/ou softwares) gerados e/ou analisados durante o estudo atual serão feitos mediante solicitação

Outras informações relevantes

Ferramentas de Inteligência Artificial Generativa foram utilizadas como suporte à revisão gramatical do texto. Os autores realizaram uma revisão completa do texto e assumem total responsabilidade pela integridade das informações apresentadas.

Referências

Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., and Hassan, A. (2023). Topic modeling algorithms and applications: A survey. *Information Systems*, 112:102131.

Arora, S., May, A., Zhang, J., and Ré, C. (2020). Contextual embeddings: When are they worth it? *arXiv preprint arXiv:2005.09117*.

Bianchi, F., Terragni, S., and Hovy, D. (2021). Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 759–766.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*.

Boutsidis, C. and Gallopoulos, E. (2008). Svd based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition*, 41(4):1350–1362. DOI: 10.1016/j.patcog.2007.09.010.

Churchill, R. and Singh, L. (2022). The evolution of topic modeling. *ACM Comput. Surv.*, 54(10s).

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American ACL: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Doogan, C. and Buntine, W. (2021). Topic model or topic twaddle? re-evaluating demantic interpretability measures. In *North American Association for Computational Linguistics 2021*, pages 3824–3848. ACL.

Formal, T., Lassance, C., Piwowarski, B., and Clinchant, S. (2022). From distillation to hard negative sampling: Making sparse neural ir models more effective. In *Proceedings of ACM SIGIR*, page 2353–2359.

Gao, X., Lin, Y., Li, R., Wang, Y., Chu, X., Ma, X., and Yu, H. (2024). Enhancing topic interpretability for neural topic modeling through topic-wise contrastive learning. In *2024 IEEE 40th ICDE*.

Ghahramani, Z. and Attias, H. (2000). Online variational bayesian learning. In *Slides from talk presented at NIPS workshop on Online Learning*.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Júnior, A. P. D. S., Cecilio, P., Viegas, F., Cunha, W., Albergaria, E. T. D., and Rocha, L. C. D. D. (2022). Evaluating topic modeling pre-processing pipelines for portuguese texts. WebMedia '22, page 191–201. DOI: 10.1145/3539637.3557052.

Kuang, D., Choo, J., and Park, H. (2015). *Nonnegative Matrix Factorization for Interactive Topic Modeling and Document Clustering*, pages 215–243. Springer International Publishing, Cham.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *nature*, 401(6755):788–791.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., *et al.* (2022). Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Machado, A. C., França, C., Nunes, I., Gonçalves, M. A., and Rocha, L. (2025). Pondere e expanda: Impacto e limitações de representações contextual-esparsas na modelagem de

tópicos. In *Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 928–934. SBC.

Viegas, F., Canuto, S., Gomes, C., Luiz, W., Rosa, T., Ribas, S., Rocha, L., and Gonçalves, M. A. (2019). Cluwords: exploiting semantic word clustering representation for enhanced topic modeling. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 753–761.

Viegas, F., Cunha, W., Gomes, C., Pereira, A., Rocha, L., and Goncalves, M. (2020). CluHTM - semantic hierarchical topic modeling based on CluWords. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8138–8150.

Viegas, F., Pereira, A., Cunha, W., França, C., Andrade, C., Tuler, E., Rocha, L., and Gonçalves, M. A. (2025). Exploiting contextual embeddings in hierarchical topic modeling and investigating the limits of the current evaluation metrics. *Computational Linguistics*, pages 1–41. DOI: 10.1162/coli_a_00543.