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

Ana Cláudia Machado[1], **Celso França**[2], **Ian Nunes**[1]
Marcos André Gonçalves[2], **Leonardo Rocha**[1]

[1] Universidade Federal de São João del Rei, Brasil
[2] Universidade Federal de Minas Gerais, Brasil
{anaclaudiamachado211, iannunes}@aluno.ufsj.edu.br
{celsofranca,mgoncalv}@dcc.ufmg.br,lcrocha@ufsj.edu.br

Resumo. *Este trabalho propõe o uso de representações contextual-esparsas na tarefa de Modelagem de Tópicos (MT), com o objetivo de combinar a interpretabilidade das representações esparsas com o poder semântico das representações contextuais. Utilizando o modelo SPLADE, buscamos representar documentos de forma sensível ao contexto por meio da expansão e ponderação de termos. Avaliamos empiricamente essa abordagem em comparação com outras representações. Os resultados indicam que a ponderação permite uma MT eficaz, enquanto a expansão, embora promissora, apresenta limitações devido à incompatibilidade entre o vocabulário da representação e os textos originais.*

Abstract. *This work proposes using contextual-sparse representations in Topic Modeling (TM), aiming to combine the interpretability of sparse representations with the semantic power of contextual ones. Using the SPLADE model, we seek to represent documents in a context-sensitive manner through term expansion and weighting. We empirically evaluate this approach in comparison with other representations. The results indicate that term weighting enables effective TM, while term expansion, although promising, presents limitations due to the mismatch between the representation's vocabulary and the original texts.*

1. Introdução

A Modelagem de Tópicos (MT) é uma técnica não supervisionada de aprendizado de máquina que identifica padrões temáticos latentes em grandes coleções textuais [Viegas et al. 2019]. Ao agrupar termos semanticamente relacionados, reduz a dimensionalidade dos dados e facilita sua organização e análise [Churchill and Singh 2022, Abdelrazek et al. 2023a]. As abordagens variam de modelos probabilísticos, como o *Latent Dirichlet Allocation* (LDA) [Blei et al. 2003], a métodos determinísticos, como a Fatoração de Matrizes Não Negativas (NMF) [Kuang et al. 2015], além de técnicas recentes baseadas em representações densas, como CluWords [Viegas et al. 2019, Viegas et al. 2025] e BERTopic [Grootendorst 2022].

As técnicas de modelagem de tópicos diferem principalmente na forma de representação textual. Métodos tradicionais, como LDA e NMF, utilizam representações esparsas e estáticas baseadas em frequência de termos, limitando a captura de relações semânticas. O CluWords introduz vetores densos e estáticos que incorporam similaridade semântica, porém ainda descontextualizados. O BERTopic representa um avanço ao empregar vetores densos e contextuais derivados de modelos *transformers*, capazes de

captar o significado contextual das palavras. Contudo, persiste uma lacuna quanto à exploração de representações simultaneamente esparsas e contextuais, que poderiam aliar interpretabilidade e riqueza semântica.

Este trabalho propõe o uso de representações contextual-esparsas na MT, com o objetivo de capturar sensibilidade ao contexto, manter interpretabilidade e promover maior coerência temática. Para isso, emprega-se o SPLADE [Formal et al. 2022], uma abordagem baseada em *Masked Language Modeling* (MLM) que transforma *embeddings* densos de modelos *transformer* em representações esparsas. O MLM mascara *tokens* no texto e treina o modelo para predizê-los a partir do contexto, gerando distribuições que refletem a relevância contextual dos termos. Essas distribuições atribuem maior peso à palavra mascarada e a termos semanticamente relacionados, permitindo expansão semântica. A proposta combina dois componentes: (i) expansão contextual de termos e (ii) ponderação contextual, que ajusta a importância dos termos com base em sua relevância contextual. A partir disso, investigam-se 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 mostram que as representações contextuais-esparsas, quando combinadas com a NMF, produzem resultados de MT que são tão efetivos quanto ou superiores aos melhores métodos da literatura em diferentes coleções de dados. Observamos que esse bom desempenho está associado à qualidade das **ponderações**. Por outro lado, ainda são necessárias novas estratégias que permitam explorar melhor a expansão, que atualmente introduz muito ruído. Nossa análise desse componente mostrou que a limitação decorre da incompatibilidade entre os *subtokens* do vocabulário do modelo e os termos dos documentos originais, um importante desafio a ser enfrentado.

2. Revisão da Literatura

Na Tabela 1 comparamos as representações vetoriais comumente utilizadas na MT, organizadas em dois eixos principais: esparsidade e sensibilidade ao contexto.

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

Representações estático-esparsas, como o TF-IDF, destacam-se pela alta interpretabilidade e excelente escalabilidade [Gao et al. 2024], mas sofrem com alta dimensionalidade e baixa coerência temática, pois tratam os termos isoladamente e ignoram relações semânticas. Isso pode resultar em agrupamentos frequentes, porém semanticamente incoerentes [Abdelrazek et al. 2023b, Doogan and Buntine 2021]. Já

representações estático-densas, como word2vec, fastText e GloVe, reduzem a dimensionalidade e incorporam similaridade semântica, mas permanecem insensíveis ao contexto [Arora et al. 2020], o que limita sua eficácia em MT e reduz a interpretabilidade. Modelos como o CluWords, baseados nessas representações, requerem estratégias adicionais, como expansão de sinônimos e filtragem de ruído [Viegas et al. 2019].

Representações contextuais-densas, como as de BERT e RoBERTa, capturam semântica de forma contextualizada, promovendo maior coerência temática, mas com limitações de interpretabilidade, escalabilidade e alta demanda computacional [Liang et al. 2022]. Métodos como BERTopic dependem de múltiplas etapas de pós-processamento — como redução de dimensionalidade e clusterização — para obter bons resultados. Já representações contextuais-esparsas, como as do SPLADE [Formal et al. 2022], aliam interpretabilidade e eficiência da esparsidade à riqueza semântica contextual, viabilizando expansão semântica de documentos. A exploração dessa abordagem, ainda pouco investigada em MT, representa um avanço relevante, ao buscar um equilíbrio mais eficaz entre coerência temática e interpretabilidade.

3. Abordagem Proposta

Nossa proposta baseia-se no uso de representações contextuais-esparsas em MT, com ênfase no método SPLADE. O funcionamento do pipeline proposto é ilustrado na Figura 1. A partir dos textos, geramos representações contextuais-esparsas que formam a matriz $V \in \mathbb{R}^{n \times m}_{>0}$, onde n representa o número de documentos e m, a quantidade de termos do vocabulário. Essa matriz é decomposta por meio da técnica de NMF, permitindo a extração de k tópicos. A seguir, detalhamos cada um dos componentes do processo.

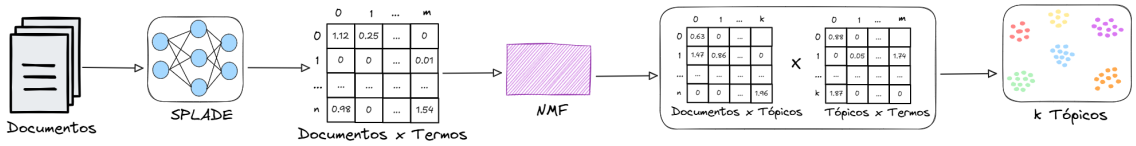


Figura 1. Pipeline de modelagem de tópicos.

3.1. Representação Contextual-Esparsa

O SPLADE [Formal et al. 2022] é um modelo baseado em *transformers* que utiliza uma estratégia de MLM para gerar *embeddings* contextuais, conforme ilustrado na Figura 2. No MLM, partes do texto são mascaradas e o modelo é treinado para prever os *tokens* ausentes com base no contexto [Devlin et al. 2019]. A geração das representações ocorre em três etapas principais. Na primeira, o texto é *tokenizado* e alguns *tokens* são aleatoriamente substituídos por [MASK]. A sequência resultante é convertida em IDs numéricos e processada pelas camadas do *transformer*. Por meio do mecanismo de *self-attention* bidirecional, são gerados *embeddings* contextuais densos (EMB-1, EMB-2, ..., EMB-6) para cada *token*, considerando suas interações com os demais da sequência. Na segunda etapa, esses *embeddings* são passados pela *MLM Head*, que produz, para cada posição, uma distribuição de probabilidade sobre o vocabulário, indicando a probabilidade de cada *token* estar presente, dado o contexto. Por fim, na terceira etapa, essas distribuições são agregadas em uma única representação vetorial esparsa por meio de uma função de ativação com saturação logarítmica [Formal et al. 2022].

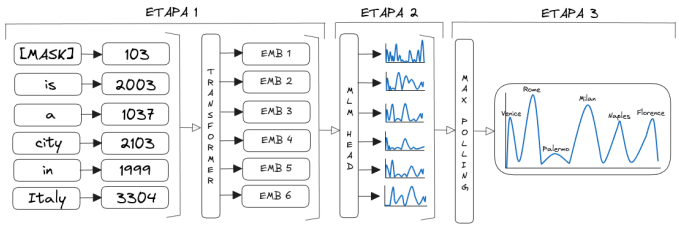


Figura 2. Representações contextuais-esparsas são geradas a partir dos *embeddings* de um *transformer* (etapa 1), processadas por uma *MLM Head* (etapa 2) e agregadas por *max pooling* (etapa 3).

O SPLADE assume que *tokens* com alta probabilidade de ocorrência são semanticamente relevantes. *Tokens* não presentes no texto original podem compor a representação vetorial, promovendo expansão semântica. O vetor final tem a mesma dimensionalidade do vocabulário e indica a importância de cada *token* no documento.

3.2. Modelagem de Tópicos

Na etapa de modelagem, utilizamos a técnica de NMF devido à sua capacidade de gerar representações interpretáveis, ao impor restrições de não negatividade que favorecem a identificação de componentes latentes coerentes, como os tópicos [Lee and Seung 1999]. Como entrada para o NMF, fornecemos o número k de tópicos a serem inferidos e a matriz V, que contém as representações contextuais-esparsas dos documentos. O NMF decompõe V em duas matrizes: $W \in \mathbb{R}_{>0}^{n \times k}$, que associa documentos aos tópicos, e $H \in \mathbb{R}_{>0}^{k \times m}$, que relaciona tópicos aos termos do vocabulário, de forma que $V \approx W \times H$, 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 nas avaliações experimentais.

Base de Dados: Utilizamos três coleções de dados amplamente conhecidas na literatura: *ACM* (artigos científicos publicados na *ACM Digital Library* - 11 classes), *20News* (postagens de grupos de notícias - 20 classes) e *WOS* (artigos científicos publicados em *Web of Science Platform* - 33 classes). O número de tópicos ideal nos experimentos é determinado pela quantidade de classes existentes na base original.

Pré-processamento: Aplicamos os seguintes passos em todas as bases: (1) conversão para minúsculas; (2) remoção de *stopwords* em inglês; (3) remoção de números e de pontuação; e (4) remoção de palavras com menos de três caracteres [Júnior et al. 2022].

Algoritmos de MT: Avaliamos diferentes abordagens de modelagem de tópicos: LDA e NMF com vetores estáticos e esparsos; Cluwords com vetores estáticos e densos; e BERTopic com vetores contextuais e densos. Para o NMF, utilizamos inicialização via *Nonnegative Double SVD* [Boutsidis and Gallopoulos 2008]; para o LDA, o método *Online Variational Bayes* [Ghahramani and Attias 2000]. No Cluwords, seguimos as instruções do trabalho original [Viegas et al. 2019, Viegas et al. 2025], e no BERTopic usamos os parâmetros padrão com o modelo *all-MiniLM-L6-v2* e K-Means para clusterização.

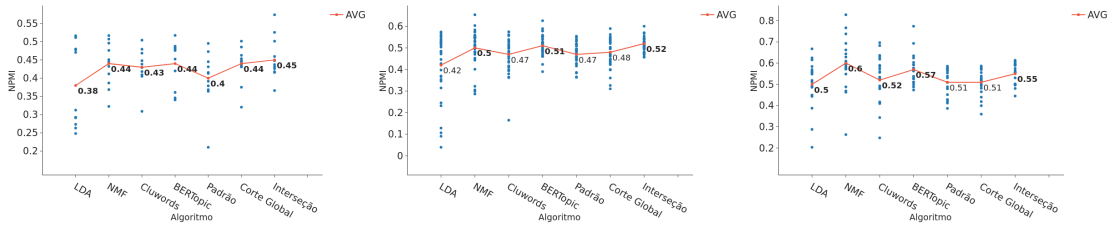
Nossa proposta: Combinamos NMF com representações contextuais-esparsas (via SPLADE), avaliando três cenários: (i) **Padrão**: utiliza todas as ponderações e expansões dos documentos; (ii) **Corte Global**: descarta os 40% menores valores da matriz; e (iii) **Interseção**: considera apenas termos presentes no documento original. Em relação

ao Corte Global, realizamos experimentos variando essa porcentagem e observamos que descartar os 40% dos valores mais baixos proporcionou um bom equilíbrio entre a redução de ruído e a preservação das informações relevantes. Nosso objetivo é entender de forma isolada o impacto das ponderações e das expansões. Ao restringir os valores não negativos aos termos presentes no documento, como ocorre na interseção, conseguimos isolar a ponderação do SPLADE e garantir que nenhuma expansão seja considerada. Ao realizar um corte nos menores valores presentes na matriz, como ocorre no Corte Global, visamos observar se esses valores podem ser considerados ruídos na representação.

Métrica e comparação estatística: Utilizamos uma métrica tradicional para avaliar a coerência dos tópicos sob a perspectiva sintática, o *Normalized Pointwise Mutual Information* (NPMI) [Bouma 2009], o qual é baseada no ganho de informação proveniente da coocorrência de palavras em um documento. Cada tópico é representado por 10 palavras e a significância estatística é avaliada utilizando o teste t com correção de Bonferroni.

5. Resultados Experimentais

A Figura 3 apresenta resultados da comparação de algoritmos de MT baseados em diferentes representações com as nossas propostas contextuais-esparsas (QP1), utilizando a métrica *NPMI*. As médias dos modelos são destacadas em vermelho, e os valores por tópico, em azul. Resultados em negrito indicam os melhores desempenhos por coleção, conforme o teste t pareado; empates estatísticos também são destacados. Observa-se que os resultados do BERTopic estão em conformidade com a literatura, mantendo-se como o estado-da-arte em MT. Quando comparado a LDA, NMF e CluWords, o BERTopic apresenta os maiores valores de *NPMI*, ou empata estatisticamente com o melhor. Destaca-se ainda sua baixa dispersão por tópico, indicando maior consistência temática.



(a) ACM (b) WOS (c) 20NG

Figura 3. Valores de NPMI para diferentes algoritmos/representações de MT

Em relação às abordagens propostas, a representação **Interseção** apresentou os melhores valores absolutos de *NPMI* nas coleções ACM e WOS, além de desempenho comparável ao melhor método na 20NG (NMF), posicionando-se ao nível equivalente ao estado-da-arte em MT. Observa-se também menor dispersão nos valores por tópico em comparação ao BERTopic, indicando maior consistência temática. A Tabela 2 ilustra qualitativamente essa análise por meio de três exemplos de tópicos bem definidos, um por coleção, gerados pela abordagem **Interseção**. *Esses resultados indicam que representações contextuais-esparsas são eficazes e comparáveis às melhores abordagens da literatura em diferentes cenários, respondendo à QP1.*

Comparando o desempenho das três abordagens contextuais-esparsas, observamos que as estratégias que realizam algum tipo de expansão (i.e. Padrão e Corte Global)

não apresentam ganhos em relação à Interseção e, consequentemente, em relação aos demais métodos. Isso sugere que, *apesar da ponderação de termos fornecida pela representação contextual-esparsa permitir realizar uma boa modelagem de tópicos, a expansão ainda não é eficaz em agregar informações contextuais que melhorem o desempenho da abordagem proposta, respondendo à QP2*. Para compreender esse comportamento, analisamos detalhadamente essa expansão.

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

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

Para cada documento de cada coleção, selecionamos os 50 termos (*tokens*) com maior peso atribuído pelo SPLADE e contabilizamos a parcela desses *tokens* que não pertenciam ao vocabulário da coleção. O resultado dessa análise é apresentado na Figura 4, onde o eixo X representa a parcela de *tokens* que não pertenciam ao vocabulário e o eixo y representa a frequência de documentos relacionada a cada parcela. Esses gráficos evidenciam que uma parcela considerável dos *tokens* relevantes não está no vocabulário do conjunto de documentos. Entre esses *tokens* há uma ocorrência frequente de *subtokens* - unidades menores de uma palavra, resultantes da *tokenização* de termos ausentes no vocabulário. O mapeamento reverso de *subtokens* para palavras completas é ambíguo e ruidoso, já que um mesmo *subtoken* pode compor diversas palavras diferentes. Podemos citar, por exemplo, o subtoken *"##ing"*, que aparece em diversas palavras em inglês, como *reconfiguring*. Esse *subtoken*, isoladamente, não possui significado completo e pode estar presente em múltiplas palavras distintas. Isso dificulta a reconstrução precisa do vocabulário original e compromete a interpretação semântica dos tópicos.

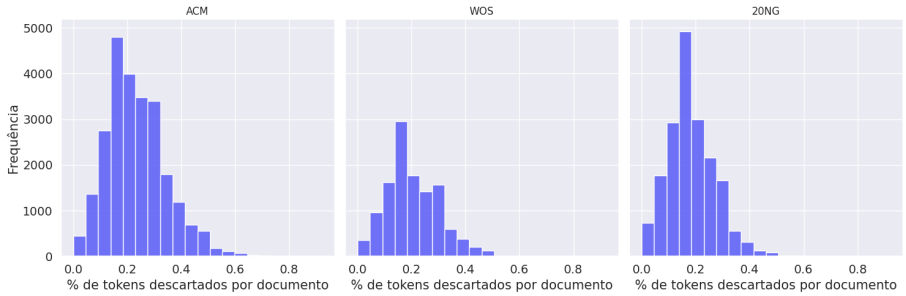


Figura 4. % de *tokens* descartados por documento em uma distribuição de frequência por coleção

6. Conclusão e Direções Futuras

Este trabalho investigou o uso de representações contextuais esparsas na MT, com foco em dois mecanismos: expansão e ponderação de termos. Os resultados mostram que a ponderação melhora a efetividade, enquanto a expansão teve desempenho limitado, em grande parte pela incompatibilidade entre os vocabulários de *subtokens* dos modelos e os documentos originais. Como trabalho futuro, propomos adaptar a saída do SPLADE para gerar distribuições de probabilidade compatíveis com vocabulários específicos, substituindo o vocabulário original do *transformer*. Para isso, planejamos treinar representações esparsas diretamente sobre vocabulários-alvo, usando tarefas de mascaramento não supervisionado, permitindo que o modelo produza saídas ajustadas a cada conjunto de dados.

Agradecimentos

Este trabalho foi apoiado por CNPq, Capes, Fapemig, Fapesp, AWS, NVIDIA, CIIA-Saúde e INCT-TILDIAR (408490/2024-1).

Referências

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

Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., and Hassan, A. (2023b). 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*.

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.

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 the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2353–2359, New York, NY, USA. Association for Computing Machinery.

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.

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*.

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., 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.