Adapting Large Language Models for Topic Modeling Tasks

Daniel Carvalho[1], **Antônio Pereira**[1], **Elisa Tuler**[1], **Diego Dias**[2]
Washington Cunha[3], **Leonardo Rocha**[1]

[1] Universidade Federal de São João del-Rei (UFSJ)

[2]Universidade Federal do Espirito Santo (UFES)

[3]Universidade Federal de Minas Gerais (UFMG)

{danielcarvalho,antoniopereira}@aluno.ufsj.edu.br, etuler@ufsj.edu.br

diego.dias@ufes.br, washingtoncunha@dcc.ufmg.br, lcrocha@ufsj.edu.br

Abstract. *This work presents a proposal for adapting Large Language Models (LLMs) to the unsupervised task of Topic Modeling (TM). Our proposal consists of three stages: document summarization, characterization of topics, and definition of topics. We instantiated our proposal with two LLMs, one open-source (Llama3) and the other proprietary (GPT 3.5), comparing them with four state-of-the-art (SOTA) strategies in TM. Our results demonstrated that the approach is very promising, having been able to define topics as coherent as SOTA strategies but still with room for improvement in terms of organizational structure.*

1. Introdução

Modelagem de Tópicos (MT) consiste no conjunto de abordagens não-supervisionadas mais proeminentes e úteis para extrair e organizar informações de grandes conjuntos de dados textuais (Luiz et al. 2018). As abordagens de MT visam extrair e identificar "tópicos semânticos" subjacentes[1] de documentos textuais, que podem ser usados por outras aplicações, tais como mecanismos de busca e sistemas de recomendação, para aumentar a eficácia das mesmas (Porturas and Taylor 2021; Aziz et al. 2022). A MT envolve a seleção de conjuntos de p palavras que representam os principais k tópicos latentes presentes em um conjunto de D documentos (Júnior et al. 2022). Cada tópico latente deve (idealmente) agrupar subconjuntos de documentos em "temas" comuns (Viegas et al. 2018).

As estratégias de MT podem ser classificadas, basicamente, em probabilísticas e não-probabilísticas. Com respeito às probabilísticas, a mais utilizada na literatura é a LDA (*Latent Dirichlet Allocation*), a qual aplica uma distribuição de *Dirichlet* para inferir estruturas subjacentes nos dados, representando os documentos como uma combinação de tópicos e gerando palavras com base nesses tópicos. Como relação aos métodos não-probabilísticos, o método mais clássico e que apresenta excelentes resultados é o NMF (*Non-negative Matrix Factorization*) (Viegas et al. 2019), uma abordagem baseada em fatoração de matrizes que utiliza, originalmente, uma representação tradicional *Bag of Words* (BoW), ponderada pelo fator de correção TF-IDF(*Term Frequency–Inverse Document Frequency*). O avanço das estratégias de representação semântica, em especial as *word embeddings* (Mikolov et al. 2018), teve um impacto muito positivo na área de MT. Novas propostas vêm sendo apresentadas utilizando essas representações para

[1]Definido pela semântica ou significado das palavras contidas nos tópicos.

enriquecer o processo de construção de tópicos mais semanticamente coesos, como são os casos das estratégias CluWords (Viegas et al. 2019) e o BERTopic (Grootendorst 2022).

A evolução dessas representações semânticas, impulsionada pelos expressivos avanços de arquiteturas de redes neurais para tarefas de aprendizado profundo aplicadas a processamento de linguagem natural (PLN), são os Grandes Modelos de Linguagem (*Large Language Models - LLMs*). De fato, a literatura demonstra que essas arquiteturas (e.g., GPT4 (Achiam et al. 2023) e LLama3 (Touvron et al. 2023)) se destacam como o estado-da-arte em diversos domínios supervisionados, alcançando resultados notáveis em diversas tarefas, incluindo sistemas de recomendação (Ma et al. 2021) *ranking* (MacAvaney et al. 2020), perguntas-e-respostas (Singhal et al. 2023), dentre outras. Assim, nosso objetivo é responder à seguinte pergunta de pesquisa *PP1: As LLMs atuais poderiam ser adaptadas para a tarefa não-supervisionada de Modelagem de Tópicos? Qual o desempenho dessa adaptação frente às estratégias estado-da-arte?*

Para responder essa pergunta, propomos neste trabalho uma estratégia de adaptação de LLMs para a tarefa de MT, composta de três etapas básicas. Na etapa 1, dada uma coleção de documentos, cada documento é encaminhado para a LLM para que a mesma resuma o mesmo em um conjunto fixo de palavras. Na etapa 2, todas as palavras que definem cada um dos documentos da coleção são encaminhadas de volta a LLM para que ela defina os N (parâmetro) tópicos relacionados a esses conjuntos de palavras, retornando as M (parâmetro) palavras que definem cada um dos N tópicos gerados. Por fim, na etapa 3, para cada documento d da coleção, tomamos a representação vetorial (baseada na própria LLM) das palavras que o representam e calculamos a similaridade desses vetores com os vetores de cada palavra de cada um tópico t. A similaridade de d e t é dada pela maior similaridade entre todos os pares de palavras que representam cada um deles. Repetimos esse processo para os N tópicos e o documento é assinalado ao tópico de maior similaridade. Nossa abordagem é independente da LLM utilizada.

Avaliamos nossa proposta instanciando-a utilizando duas LLMs distintas: uma opensource (LLama3 (Touvron et al. 2023)) e outra proprietária (GPT-3.5 (Achiam et al. 2023)), comparando-as com quatro estratégias (LDA, NMF, CluWords e BERTopic), considerando seis métricas distintas: três tradicionalmente utilizadas na avaliação de MT, que qualificam a definição dos tópicos e a coerência dos assuntos (NPMI (Bouma 2009) e TF-IDF *Coherence* (Nikolenko et al. 2015), WEP (Dieng et al. 2019)) e três utilizadas em *clustering* que avaliam a organização estrutural dos tópicos (Coeficiente de Silhoueta (Rousseeuw 1987), Calinski-Harabasz (Caliński and Harabasz 1974) e Beta-CV (Ezugwu et al. 2022)). Consideramos três coleções de dados (i.e., ACM, 20News e WebKb).

Na avaliação considerando as métricas tradicionais, os resultados apontaram para um bom desempenho das instanciações propostas, equiparando-as às estratégias estado-da-arte (i.e. CluWords) na definição de tópicos coerentes e concisos. Por outro lado, sob a perspectiva da organização estrutural dos documentos, a estratégia proposta ainda precisa ser melhorada no processo de assinalamento de tópicos aos documentos. Retomando nossa pergunta de pesquisa, podemos afirmar que adaptação de LLMs para a tarefa de Modelagem de Tópicos é promissora, mas ainda há bastante espaço para melhorias.

2. Referencial Teórico

2.1. Modelagem de Tópico

As abordagens de Modelagem de Tópicos (MT) passaram por avanços significativos nas últimas décadas. Essas abordagens podem ser classificadas em probabilísticas e não-probabilísticas. As abordagens probabilísticas têm como marco principal a *PLSA - Probabilistic Latent Semantic Analysis* (Hofmann 1999). Neste método, os z tópicos são tratados como variáveis latentes, enquanto a coocorrência de palavras e documentos (w, d) é considerada uma variável observável. O ponto crucial desta abordagem é estabelecer uma associação entre as variáveis latentes e observáveis. Isso é feito modelando a probabilidade de cada coocorrência como uma mistura de distribuições multinomiais condicionalmente independentes. Os parâmetros são estimados maximizando a probabilidade dos dados, usando o algoritmo expectativa-maximização (EM). Este algoritmo visa descobrir estruturas subjacentes nos dados, representando documentos como uma combinação de tópicos e gerando palavras com base nesses tópicos. Esta abordagem foi fundamental para o desenvolvimento de modelos probabilísticos mais robustos, como o *LDA - Latent Dirichlet Allocation* (Viegas et al. 2019). Especificamente, o **LDA** substitui a distribuição multinomial pela de *Dirichlet*, permitindo uma melhor modelagem das distribuições de tópicos em documentos e palavras em tópicos.

Quanto às abordagens não-probabilísticas, estratégias de fatoração de matrizes, como Fatoração de matrizes Não-Negativas (*Non-Negative Matrix Fatorization* - NMF), tornaram-se muito populares. O **NMF** fatora a representação tradicional *Bag of Words* (BoW), ponderada pelo fator de correção TF-IDF, em duas novas matrizes: uma matriz H de documentos por tópico e uma matriz W de tópicos por palavras. Da matriz H é possível extrair quais tópicos estão relacionados a cada documento, e da matriz W é possível definir as principais palavras que descrevem cada tópico.

Com os avanços em representações contextuais, como os *embeddings* de palavras (Mikolov et al. 2018), as estratégias de MT também vêm evoluindo por meio do uso dessas representações, destacando-se duas abordagens: **CluWords** (Viegas et al. 2019) e **BERTopic** (Grootendorst 2022). CluWords substitui a representação TF-IDF convencional por uma representação orientada à semântica, construída usando agrupamentos de palavras próximas em um espaço contextual de *word-embeddings* e o NMF extrai os tópicos. BERTopic (Grootendorst 2022) usa *word-embeddings* contextuais resumidos à nível de documento para agrupar documentos em um espaço vetorial de *embeddings*. BERTopic também usa uma modificação do TF-IDF para extrair as principais palavras p de cada tópico. Essas duas abordagens são atualmente consideradas o que há de mais moderno em MT, mas existem vários outros algoritmos e abordagens que estão disponíveis na literatura, cada um com características e cenários de aplicação específicos.

2.2. Grandes Modelos de Linguagem

Atualmente, LLM's como GPT (Achiam et al. 2023) e LLama3 (Touvron et al. 2023) vêm se destacando em diversas atividades relacionadas a Processamento de Linguagem Natural (PLN). De acordo com Andrew Ng (Ng 2017), há duas razões principais para os resultados em aplicações tão bem-sucedidas. A primeira é a quantidade de dados usados para pré-treinar esses modelos – o modelo GPT-3 (Brown et al. 2020), por exemplo, foi

pré-treinado em 45 TB de dados textuais. A segunda razão é a possibilidade de reutilizar o modelo geral pré-treinado em múltiplas tarefas apenas realizando consultas (*prompts*).

Um exemplo desse bem sucedido uso de LLMs está em sistemas de recomendação (Ma et al. 2021), mais especificamente, recomendação de *hiperlinks* no qual os autores propõem aproveitar os *hiperlinks* em grande escala e os textos âncora para pré-treinar LLMs. Como os textos âncora são criados por *webmasters* e geralmente resumem o documento de destino, a proposta apresentada auxiliou a construir amostras de pré-treinamento mais precisas e confiáveis. Na área médica os exemplos de sucesso também são observados. Em (Singhal et al. 2023) é apresentado o Med-PaLM 2, que combina LLMs recentes com ajuste fino do domínio médico e estratégias de estímulo, incluindo uma nova abordagem de refinamento de conjunto, que é capaz de recuperar conhecimento médico, raciocinar sobre ele e responder perguntas médicas de forma comparável às dos médicos. Mesmo para a tarefa de *ranking* observamos uma proposta na literatura que contorna as questões de custo computacional de treinamento de LLMs para utilizá-las de forma eficiente (MacAvaney et al. 2020).

2.3. Trabalhos Relacionados

Embora a aplicação de LLMs tenha sido explorada em diversos cenários (Liang et al. 2023), seu uso ainda está em estágio inicial no contexto de Modelagem de Tópicos (MT). O trabalho de (Mu et al. 2024) discute algumas das complexidades relacionadas ao uso de LLM na tarefa de MT. Além disso, apresentam uma estratégia, mas que, no entanto, ainda dependem do uso de rótulos pré-existentes nas bases de dados, o que não é viável do ponto de vista prático. Por fim, a estratégia proposta não discute como os documentos são assinalados aos tópicos. Nossa proposta mitiga ambas as questões.

Os trabalhos de (Rijcken et al. 2023; El-Gayar et al. 2024) utilizam LLMs para gerar representações interpretáveis dos resultados de MT, mas não aplicam esses modelos na extração dos tópicos. Em contraste, os estudos de (Pham et al. 2023) propõem uma estratégia inovadora para a construção de tópicos utilizando LLMs, onde tópicos são gerados de forma interativa a partir de um subconjunto de documentos e prompts complexos que incluem exemplos. Posteriormente, esses tópicos são filtrados e atribuídos a documentos, também por meio da LLM. Os autores concluem que LLMs de código aberto não são adequadas para a extração de tópicos e recomendam o uso do GPT-4 como uma solução eficaz, embora financeiramente onerosa. Nossa estratégia, além de ser baseada em prompts mais simples, também considera LLMs de código aberto mais recentes, como LLama3, mais acessível e com potencial comparável em termos de desempenho.

3. Estratégia Proposta

Nessa seção apresentamos nossa proposta para adaptação de Grandes Modelos de Linguagem (*Large Language Models* - LLMs) para Modelagem de Tópicos. Nossa proposta é dividida em três etapas distintas e com objetivos específicos.

Na **etapa 1**, a qual denominamos de **Sumarização dos Documentos**, cada documento da coleção para a qual se deseja organizar por meio de MT é encaminhado para a LLM para sua sumarização em até X palavras. Essa sumarização por documento é armazenada e todas palavras retornadas para toda coleção é consolidada em um dicionário. No Algoritmo 1, linhas 1-9, ilustramos o funcionamento dessa etapa e nas Tabelas 1 e 2 ilustramos os *prompts* para duas LLMs distintas.

Algorithm 1: Modelagem de Tópicos por LLMs

Input: dataset, #palavras, M, N
Output: TopicoPorDocumento

```
1  ============================
2  Etapa 1 - Sumarização dos Documentos
3  ============================
4  dicionario ← ∅;
5  summ_doc ← {}
6  for d ∈ dataset do
7  |   summ_doc[d] ← SumarizaDocLLM(d, #palavras);
8  |   dicionario ← dicionario ∪ summ_doc[d];
9  end
10 ============================
11 Etapa 2 - Caracterização dos Tópicos
12 ============================
13 Lista_Topicos ← obterTopicosLLM(dicionario, N, M);
14 ============================
15 Etapa 3 - Definição dos Tópicos
16 ============================
17 TopicoPorDocumento ← {}
18 for d ∈ dataset do
19 |   doc_vec ← obterEmbeddingsLLM(summ_doc[d]);
20 |   MaxSim = −1;
21 |   for t ∈ Lista_Topicos do
22 |   |   for p ∈ Lista_Topicos[t] do
23 |   |   |   p_vec ← obterEmbeddingsLLM(summ_doc[d]);
24 |   |   |   if simCos(doc_vec, p_vec) ≥ MaxSim then
25 |   |   |   |   TopicoPorDocumento[d] ← t;
26 |   |   |   |   MaxSim = simCos(doc_vec, p_vec);
27 |   |   |   end
28 |   |   end
29 |   end
30 end
```

A **etapa 2** é denominada de **Caracterização dos Tópicos**. O dicionário de palavras distintas obtidas na etapa 1 é encaminhado para LLM para que a mesma defina N tópicos. A saída dessa etapa são M palavras que definem cada um dos N tópicos, correspondendo à uma listagem gerada pela LLM de seu entendimento sobre o conjunto de documentos, suas principais palavras e tópicos. Esse processo corresponde a linha 13 do Algoritmo 1 e os *prompts* utilizados são apresentados na Tabelas 3 e 4.

Summarize the content of the document in unique word. The word must be contained in the document. The result should be just the word in lowercase and nothing else.
<input> I love you.
<output> love.
<input> The product is bad.
<output> product.
<input> {**Evaluate Text**}
<output> {**Response from LLM**}

Tabela 1. Prompt para Sumarização dos Documentos - LLAMA 3

Select one word, in the document, to summarize the document. The result should be just the word in lowercase and nothing else.
<input> I love you.
<output> love.
<input> The product is bad.
<output> product.
<input> {**Evaluate Text**}
<output> {**Response from LLM**}

Tabela 2. Prompt para Sumarização dos Documentos - GPT 3.5.

A **etapa 3**, denominada **Definição dos Tópicos**, consiste em definir quais documentos pertencem a cada um dos tópicos gerados. As LLMs atuais ainda não estão totalmente preparadas para executar grandes tarefas em apenas um *prompt* de comando. Além disso, existem limitações de hardware e arquitetura (número máximo de tokens por *prompt*), e, no caso das LLMs pagas, limitações de quantidade de arquivos a serem enviados através das *APIs* e, por fim, custo financeiro para a execução. Assim, para a etapa 3, não é possível encaminhar todos os documentos e conjuntos de palavras de

cada tópico para que a LLM faça a distribuição dos documentos entre os tópicos. Dessa forma, nossa estratégia para essa etapa consiste em obter uma representação semântica a partir da própria LLM (*word embedding*) para cada um dos documentos da coleção, bem como a representação semântica de cada uma das palavras representantes de cada tópico. Calculamos a similaridade de cosseno entre a representação semântica de um documento d e cada uma das M palavras de um tópico específico t. A similaridade entre d e t é dada pelo maior valor encontrado dentre as similaridades do documento diante cada palavra do tópico em questão. O processo é repetido para todos os N tópicos e o documento d é assinalado ao tópico de maior similaridade. Todo o processo é feito para cada um dos documentos da coleção. Ilustramos esse funcionamento no Algoritmo 1 nas linhas 17-30.

Cluster all input words into number of groups semantic groups. Represent each group with 10 (ten) close words in from the entry. All words in lowercase. <input> computer science network car auto bike... <output> computer science car auto bike <input> {**Evaluate Text**} <output> {**Response from LLM**}

Tabela 3. Prompt para Caracterização dos Tópicos - LLAMA 3.

Cluster all the words into only number of groups semantic groups. Show group label and top 10 (ten) most representative words by group. <input> computer science network car auto bike... <output> computer science car auto bike <input> {**Evaluate Text**} <output> {**Response from LLM**}

Tabela 4. Prompt para Caracterização dos Tópicos - GPT 3.5.

O final das três etapas resulta em N tópicos, definidos por M palavras, com cada um dos documentos da coleção pertencendo a um dos tópicos, exatamente como a saída definida pelas estratégias de modelagem de tópicos tradicionais. A seguir, avaliamos o desempenho dessa estratégia frente às estratégias de MT consideradas estado-da-arte.

4. Avaliação Experimental

4.1. Ambiente Experimental

Instanciamos a estratégia proposta na seção anterior de duas formas diferentes, considerando duas LLMs distintas: LLama3 (Touvron et al. 2023) e GPT-3.5 (Achiam et al. 2023). Compararemos essas instanciações de nossa proposta com outras quatro estratégias da literatura, duas tradicionais: LDA (abordagem probabilística) e NMF (abordagem não-probabilística) e outros dois métodos recentes, considerados o estado-da-arte em MT, baseados em representações semânticas: CluWords e BERTopic. Todos esses métodos estão destacados e descritos na Seção 2.

Consideraremos seis métricas distintas para avaliar o desempenho das estratégias. As três primeiras correspondem a métricas tradicionalmente utilizadas em MT, sendo duas que avaliam a perspectiva sintática, NPMI (Bouma 2009) e TF-IDF *Coherence* (Nikolenko et al. 2015) e uma sob a perspectiva semânticas, WEP (Dieng et al. 2019). O NPMI (Bouma 2009) é uma medida que quantifica o ganho de informação ao considerar a ocorrência conjunta de duas palavras em relação às suas ocorrências individuais. O TF-IDF *Coherence* (Nikolenko et al. 2015) avalia a facilidade de interpretação dos tópicos com base na coocorrência de palavras nos documentos. Essa métrica leva em consideração a frequência dos termos (TF) e sua importância no *corpus* (IDF), penalizando a coocorrência de palavras muito frequentes que possuem baixo poder discriminativo. Por fim, a WEP (Dieng et al. 2019) se concentra na interpretação semântica das palavras dos tópicos avaliando a proximidade semântica entre as palavras que descrevem cada tópico. Para todas elas, quanto maior o valor, melhor a qualidade.

A outras três métricas são normalmente aplicadas no cenário de *clustering*, que visam avaliar estruturalmente os tópicos: Coeficiente de Silhoueta (Rousseeuw 1987), Calinski-Harabasz (Caliński and Harabasz 1974) e Beta-CV (Ezugwu et al. 2022). Silhoueta avalia a qualidade dos tópicos com base na coesão e separação entre grupos de documentos. Calinski-Harabasz avalia a qualidade de um tópico por meio da relação entre a dispersão intra-cluster e a dispersão inter-cluster. Por fim, o Beta-CV avalia a qualidade do tópico aplicando um compromisso entre a coesão e a separação entre tópicos. Para as duas primeiras, quanto maior o valor melhor a qualidade, enquanto para Beta-CV é o contrário.

Consideramos em nossos experimentos três coleções que vêm sendo utilizadas na avaliação de tarefas de classificação de texto e modelagem de tópicos (Júnior et al. 2023): 1) ACM, que compreende artigos científicos publicados na Biblioteca Digital da ACM; 2) 20News, composta por postagens de grupos de notícias; e 3) WOS, que contém artigos científicos publicados na plataforma *Web of Science*. Na Tabela 5, resumimos as informações dessas coleções. Todas elas passaram por um processo de pré-processamento (Júnior et al. 2022). Assim, foram aplicadas as seguintes etapas: (i) conversão das palavras do texto para minúsculas, (ii) remoção de *stopwords*, (iii) remoção de termos numéricos e (iv) remoção de palavras com menos de três caracteres.

Coleção	#Classes	#Palavras	#Documentos
ACM	11	50.660	24.897
20News	20	82.547	18.286
WOS	33	48.973	11.967

Tabela 5. Características das Coleções de Dados

Por fim, em termos de parametrização, adotamos na etapa 1 que cada documento fosse sumarizado pela LLM por apenas uma palavra. Realizamos uma avaliação preliminar com cinco palavras, contudo, apesar do aumento significativo do custo computacional (e financeiro no caso do GPT), não identificamos melhoras em termos de resultados. O número de tópicos (N) gerados para cada coleção corresponde ao número de classes de cada uma delas e o número de palavras (M) que definem cada tópico foi 10. Seguindo trabalhos anteriores (Júnior et al. 2023), padronizamos as escolhas dessas 10 palavras por aquelas de maior TF-IDF dentro de cada tópico, para todas as estratégias.

4.2. Análise dos Resultados

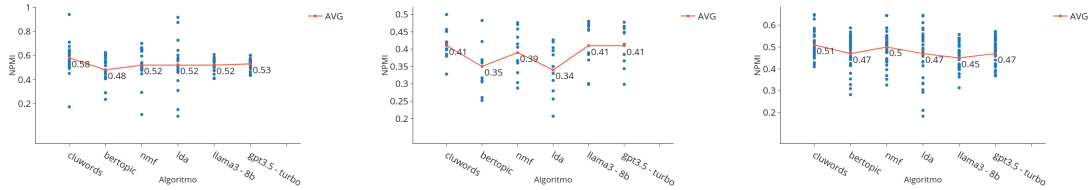
Dividimos nossas análises em duas perspectivas. Na primeira consideramos as métricas tradicionais de avaliação de MT que avaliam os tópicos sob as perspectivas sintática (NPMI e TF-IDF Coherence) e semântica (WEP) em relação à definição dos tópicos. Na segunda análise focamos na avaliação estrutural dos tópicos (organização dos documentos entre tópicos) por meio da adaptação de métricas tradicionalmente usadas no cenário de *clustering* (Coeficiente de Silhoueta, Calinski-Harabasz e Beta-CV).

4.2.1. Avaliações Sintáticas e Semânticas

Iniciamos nossas análises pelas métricas que avaliam a consistência sintática das definições dos tópicos gerados considerando as métricas NPMI e TF-IDF Coherence (Coherence), Figuras 1 e 2 respectivamente. Conforme podemos observar para os resultados relacionados ao NPMI, a estratégia CluWords segue sendo a que apresenta os melhores resultados, o que é coerente com o reportado na literatura (Viegas et al. 2019;

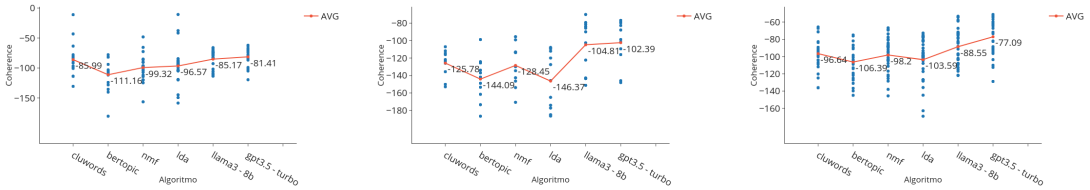
Viegas et al. 2020; Viegas et al. 2022). Por outro lado, podemos observar que estratégias baseadas em LLMs também apresentaram resultados bem competitivos nas três coleções, inclusive melhores que as demais estratégias de MT. Com relação aos resultados relacionados à TF-IDF Coherence, já observamos que as estratégias baseadas em LLMs apresentaram os melhores resultados nas três coleções, sendo, em alguns casos, bem superior às demais estratégias (i.e. coleção ACM). Para ambas as métricas, observamos um desempenho melhor para a estratégia baseada no GPT3.5, o que pode ser considerado esperado pelo tamanho dos dados utilizados em seu treinamento (Achiam et al. 2023), porém com um custo computacional também maior, além do custo financeiro. Dessa forma, do ponto de vista sintático, temos que as estratégias baseadas em LLMs, apesar de ainda simples, foram capazes de definir bem os temas relacionados à cada um dos tópicos, com resultados melhores e/ou equivalentes aos métodos estado-da-arte.

O desempenho das estratégias sob a perspectiva semântica - métrica WEP - pode ser observado na Figura 3. Conforme podemos observar, para as coleções 20News e WOS, Figuras 3 (a) e (c) respectivamente, a estratégia CluWords apresentou os melhores resultados. Já para a coleção ACM, Figuras 3 (b), as estratégias baseadas em LLMs apresentaram um desempenho superior.



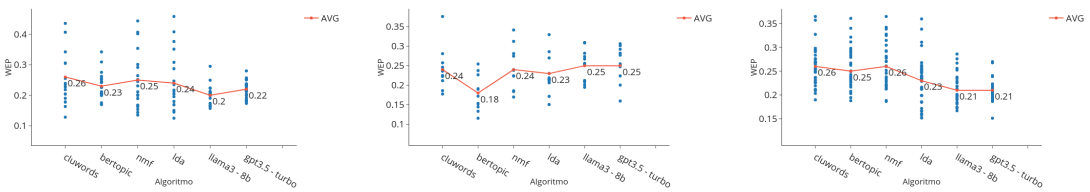
(a) Coleção 20News (b) Coleção ACM (c) Coleção WOS

Figura 1. NPMI: a estratégia Cluwords apresenta os melhores resultados, com as estratégias baseadas em LLMs com valores muito próximos e competitivos.



(a) Coleção 20News (b) Coleção ACM (c) Coleção WOS

Figura 2. TF-IDF Coherence: as estratégias baseadas em LLMs apresentam resultado bem melhores, com uma vantagem para estratégia baseada em GPT.



(a) Coleção 20News (b) Coleção ACM (c) Coleção WOS

Figura 3. WEP: as estratégias baseadas em LLMs apresentam os melhores resultados para coleção ACM, mas piores nas demais coleções. Destaque em todas é a CluWords.

Resumimos nossas descobertas sobre o comportamento de todas as estratégias analisadas, nas três coleções de dados e considerando as três métricas tradicionais de avaliação de MT na Tabela 6. Nela, contabilizamos o número de vezes que cada estratégia alcançou o melhor desempenho. Conforme podemos observar, as duas melhores estratégias, empatadas em termos de número de vezes em que teve o melhor desempenho são CluWords (considerada estado-da-arte em MT) e a estratégia baseada na LLM proprietária GPT 3.5, seguida da estratégia baseada na LLAMA 3. Esses resultados apontam que soluções baseadas em LLMs para tarefa de MT são promissoras, uma vez que as estratégias aqui apresentadas ainda são apenas as primeiras propostas de soluções, ainda com espaço para diversas melhorias. Na próxima seção apresentamos uma avaliação dessas estratégias sob a perspectiva da organização estrutural dos tópicos.

Método	Métrica			$\sum$
	NPMI	Coherence	WEP	(Soma)
CluWords	**3**	0	**2**	**5**
GPT 3.5 Turbo	1	**3**	1	**5**
LLAMA 3	1	0	1	2
NMF	0	0	1	1
BERTopic	0	0	0	0
LDA	0	0	0	0

Tabela 6. Número de vezes em que cada estratégia foi a melhor para as métricas de avaliação sintática e semântica.

4.2.2. Avaliações Estrutural

Para a avaliação da organização estrutural dos tópicos, ou seja, a distribuição dos documentos entre os tópicos, utilizamos métricas tradicionalmente aplicadas no cenário de *clustering*. Sendo assim, apresentamos os resultados para as métricas BetaCV, Calinski e Silhoueta para as coleções 20News, ACM e WOS nas Tabelas 7, 8 e 9, respectivamente.

Conforme podemos observar, em termos de organização estrutural, novamente a estratégia CluWords apresenta os melhores resultados no geral. Por outro lado, as estratégias de MT baseadas em LLMs não apresentaram bons resultados em nenhuma das coleções. Em termos de BetaCV, os resultados obtidos pelas estratégias LLMs conseguiram até apresentar resultados próximos ao resultado obtido pela melhor estratégia (20News - CluWords, ACM - LDA e WOS - CluWords). No entanto, quando consideramos os resultados relacionados à métrica Calinski, as estratégias baseadas em LLMs ficam significativamente aquém do esperado.

Métrica	CluWords	BERTopic	NMF	LDA	LLAMA3	GPT3.5
BetaCV ↓	0.93	0.97	0.93	0.97	0.99	0.99
Calinski ↑	142.73	84.01	136.50	72.25	3.54	3.49
Silhouette ↑	0.03	-0.02	0.03	-0.04	-0.02	-0.03

Tabela 7. Métricas de *clustering* - Coleção 20News

Métrica	CluWords	BERTopic	NMF	LDA	LLAMA3	GPT3.5
BetaCV ↓	0.97	0.98	0.97	0.96	1.00	1.00
Calinski ↑	183.39	62.27	170.91	126.62	1.57	1.43
Silhouette ↑	-0.01	-0.04	-0.01	-0.01	-0.01	-0.01

Tabela 8. Métricas de *clustering* - Coleção ACM

Métrica	CluWords	BERTopic	NMF	LDA	LLAMA3	GPT3.5
BetaCV ↓	0.86	0.96	0.87	0.91	1.00	1.00
Calinski ↑	116.12	78.28	107.70	78.73	0.95	0.99
Silhouette ↑	0.05	-0.03	0.04	-0.04	-0.02	-0.02

Tabela 9. Métricas de *clustering* - Coleção WOS

Resumimos nossas descobertas sobre o comportamento de todas as estratégias analisadas, nas três coleções de dados e considerando as três métricas tradicionais de *clustering* na Tabela 10. Nela, contabilizamos o número de vezes que cada estratégia foi

considerada como o melhor desempenho. Nela confirmamos que a estratégia CluWords é, de longe, a mais efetiva sob a perspectiva de organização estrutural dos documentos, sendo considerada a melhor em 8 cenários. As estratégias baseadas no GPT3.5 e Llama3 foram melhores apenas uma vez cada.

Método	Métrica			∑
	BetaCV	Calinski	Silhouette	(Soma)
CluWords	**2**	**3**	**3**	**8**
NMF	1	0	2	3
LDA	1	0	1	2
LLAMA 3	0	0	1	1
GPT 3.5	0	0	1	1
BERTopic	0	0	0	0

Tabela 10. Número de vezes em que cada estratégia foi a melhor para as métricas de avaliação estrutural.

Esse desempenho aquém do esperado pode estar relacionado com a etapa 3 de nossa abordagem. Conforme mencionamos na Seção 3, essa etapa é feita quase que completamente fora da LLM selecionada, usando apenas a representação semântica de palavras (*word embeddings*) fornecidas por ela. Isso foi necessário em função das limitações das LLMs atuais em executar toda a etapa 3 em apenas um *prompt* de comando repassado para elas. Nesse caso, uma possível solução seria repassar, via *prompt*, a lista de tópicos com as respectivas palavras que os definem, e os conjuntos de documentos para que os mesmos fossem atribuídos aos tópicos pelas próprias LLMs. Existem limitações de hardware e arquitetura (número máximo de tokens por *prompt*), e, no caso das LLMs pagas, limitações de quantidade de arquivos a serem enviados através das *APIs* e, por fim, custo financeiro para a execução que representam um impeditivo para uma solução como essa. Portanto, a etapa 3 será foco de nossos estudos futuros.

Retomando nossa pergunta de pesquisa, podemos afirmar que as LLMs atuais podem ser adaptadas para a tarefa não-supervisionada de Modelagem de Tópicos. Apesar de apresentarem resultados promissores quanto à definição de tópicos, destacando e identificando temas tão coerentes e significativos quanto às estratégias estado-da-arte, sob a perspectiva de organização estrutural essas estratégias ainda precisam ser melhoradas.

5. Conclusões e Trabalhos Futuros

Neste trabalho propomos uma abordagem de adaptação de LLMs para a tarefa de Modelagem de Tópicos, composta de três etapas básicas. Na primeira (**sumarização dos documentos**), a LLM sumariza cada documento de uma coleção em um conjunto fixo de palavras. Na segunda (**caracterização dos tópicos**), todo o conjunto de palavras extraído da etapa anterior é passado para LLM definir os tópicos. Na terceira (**definição dos tópicos**), os documentos são organizados entre os tópicos. Instanciamos nossa proposta com duas LLMs, uma código aberto (Llama3) e outra proprietária (GPT 3.5), comparando-os com quatro estratégias estado-da-arte em MT (i.e. LDA, NMF, BERTopic e CluWords), considerando três coleções de documentos e seis métricas de avaliação. Nossos resultados demonstraram que a abordagem proposta é muito promissora, tendo sido capaz de definir tópicos tão coerentes quanto aqueles definidos pelas estratégias estado-da-arte, mais ainda com um bom espaço para melhorias termos de estrutura organizacional dos tópicos. Como trabalho futuro, nossa meta é melhorar a terceira etapa da abordagem para que a distribuição dos documentos entre os tópicos também seja realizada pela LLM.

Agradecimentos Este trabalho foi apoiado pelo CNPq, CAPES, FAPEMIG e AWS.

Referências

[Achiam et al. 2023] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

[Aziz et al. 2022] Aziz, S., Dowling, M., Hammami, H., and Piepenbrink, A. (2022). Machine learning in finance: A topic modeling approach. *EFM*, 28(3):744–770.

[Bouma 2009] Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30:31–40.

[Brown et al. 2020] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems*.

[Caliński and Harabasz 1974] Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.

[Dieng et al. 2019] Dieng, A. B., Ruiz, F. J. R., and Blei, D. M. (2019). Topic modeling in embedding spaces. *CoRR*, abs/1907.04907.

[El-Gayar et al. 2024] El-Gayar, O., Al-Ramahi, M., Wahbeh, A., Nasralah, T., and Elnoshokaty, A. (2024). A comparative analysis of the interpretability of lda and llm for topic modeling: The case of healthcare apps. In *AMCIS 2024 Proceedings*.

[Ezugwu et al. 2022] Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., and Akinyelu, A. A. (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges. *Engineering Applications of Artificial Intelligence*.

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

[Hofmann 1999] Hofmann, T. (1999). Probabilistic latent semantic indexing. In *SIGIR*.

[Júnior et al. 2022] 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. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*, WebMedia '22, page 191–201.

[Júnior et al. 2023] Júnior, A. P. D. S., Viegas, F., Gonçalves, M. A., and Rocha, L. (2023). Evaluating the limits of the current evaluation metrics for topic modeling. In *Proceedings of the 29th Brazilian Symposium on Multimedia and the Web, WebMedia 2023, Ribeirão Preto, Brazil, October 23-27, 2023*, pages 119–127. ACM.

[Liang et al. 2023] Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C. A., Manning, C. D., Re, C., Acosta-Navas, D., Hudson, D. A., Zelikman, E., Durmus, E., Ladhak, F., Rong, F., Ren, H., Yao, H., WANG, J., Santhanam, K., Orr, L., Chatterji, N. S., Khattab, O., Henderson, P., Huang, Q., Chi, R. A., Xie, S. M., Santurkar, S., Ganguli, S., Hashimoto, T., Icard, T., Zhang, T., Chaudhary, V., Wang, W., Li, X., Mai, Y., Zhang, Y., and Koreeda, Y. (2023). Holistic evaluation of language models. *TMLR*. Featured Certification, Expert Certification.

[Luiz et al. 2018] Luiz, W., Viegas, F., Alencar, R., Mourão, F., Salles, T., Carvalho, D., Gonçalves, M. A., and Rocha, L. (2018). A feature-oriented sentiment rating for mobile app reviews. In *Proceedings of the 2018 world wide web conference*.

[Ma et al. 2021] Ma, Z., Dou, Z., Xu, W., Zhang, X., Jiang, H., Cao, Z., and Wen, J.-R. (2021). Pre-training for ad-hoc retrieval: hyperlink is also you need. In *CIKM*.

[MacAvaney et al. 2020] MacAvaney, S., Nardini, F. M., Perego, R., Tonellotto, N., Goharian, N., and Frieder, O. (2020). Efficient document re-ranking for transformers by precomputing term representations. In *SIGIR*.

[Mikolov et al. 2018] Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. (2018). Advances in pre-training distributed word representations. In *LREC*.

[Mu et al. 2024] Mu, Y., Dong, C., Bontcheva, K., and Song, X. (2024). Large language models offer an alternative to the traditional approach of topic modelling. In *LREC-COLING*.

[Ng 2017] Ng, A. (2017). Machine learning yearning. *URL: http://www. mlyearning. org/(96)*, 139.

[Nikolenko et al. 2015] Nikolenko, S., Koltsov, S., and Koltsova, O. (2015). Topic modelling for qualitative studies. *Journal of Information Science*, 43.

[Pham et al. 2023] Pham, C. M., Hoyle, A., Sun, S., and Iyyer, M. (2023). Topicgpt: A prompt-based topic modeling framework. *arXiv preprint arXiv:2311.01449*.

[Porturas and Taylor 2021] Porturas, T. and Taylor, R. A. (2021). Forty years of emergency medicine research: Uncovering research themes and trends through topic modeling. *The American Journal of Emergency Medicine*, 45:213–220.

[Rijcken et al. 2023] Rijcken, E., Scheepers, F., Zervanou, K., Spruit, M., Mosteiro, P., and Kaymak, U. (2023). Towards interpreting topic models with chatgpt. The 20th World Congress of the International Fuzzy Systems Association, IFSA ; Conference date: 20-08-2023 Through 24-08-2023.

[Rousseeuw 1987] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

[Singhal et al. 2023] Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., et al. (2023). Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

[Touvron et al. 2023] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). Llama: Open and efficient foundation language models.

[Viegas et al. 2019] 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 *WSDM*, pages 753–761.

[Viegas et al. 2020] Viegas, F., Cunha, W., Gomes, C., Pereira, A., Rocha, L., and Goncalves, M. (2020). Cluhtm-semantic hierarchical topic modeling based on cluwords. In *ACL*, pages 8138–8150.

[Viegas et al. 2022] Viegas, F., Júnior, A. P. D. S., Cecilio, P., Tuler, E., Jr., W. M., Gonçalves, M. A., and Rocha, L. (2022). Semantic academic profiler (SAP): a framework for researcher assessment based on semantic topic modeling. *Scientometrics*, 127(8):5005–5026.

[Viegas et al. 2018] Viegas, F., Luiz, W., Gomes, C., Khatibi, A., Canuto, S., Mourão, F., Salles, T., Rocha, L., and Gonçalves, M. A. (2018). Semantically-enhanced topic modeling. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 893–902.