Modelagem de Tópicos em Textos Curtos: uma Avaliação Experimental*

Annie Amorim[1], Nils Murrugarra-Llerena[2], Vítor Silva[3], Daniel de Oliveira[1], Aline Paes[1]

[1]Instituto de Computação, Universidade Federal Fluminense (UFF), Niterói, RJ, Brasil

annieamorim@id.uff.br, {danielcmo,alinepaes}@ic.uff.br

[2]Weber State University, Ogden, Utah, Estados Unidos

nmurrugarrallerena@weber.edu

[3]Snap Inc., Santa Monica, California, Estados Unidos

vsilvasousa@snap.com

Resumo. *As redes sociais são utilizadas para expressar opiniões ou interagir com outras pessoas. Diante do amplo escopo de assuntos publicados e a linguagem informal presente nas postagens, a busca de informações é significativamente desafiadora. Assim, descobrir automaticamente os tópicos tratados nos textos ruidosos e com pouco contexto postados é primordial. Dado este cenário, este artigo contribui com uma análise comparativa de métodos de modelagem de tópicos, incluindo os baseados em abordagens probabilísticas e neurais. Ademais, esse artigo contribui com um método para rotular automaticamente os tópicos, permitindo uma análise qualitativa dos tópicos descobertos.*

Abstract. *People use social networks to express opinions or interact with other people. However, information retrieval is significantly challenging in the face of the broad scope of posted topics and the informal language in posts. Thus, automatically discovering topics from the noisy and short texts posted on social networks is paramount. Given this scenario, this paper contributes with a comparative analysis of topic modeling methods, comparing them with classical probabilistic and recent neural approaches. Also, this paper contributes with a technique for labeling topics automatically, allowing a qualitative analysis of the discovered topics.*

1. Introdução

As redes sociais têm desempenhado um papel fundamental para a propagação e disseminação de informações [Oraby et al. 2019]. Tais informações podem variar desde a necessidade de um indivíduo compartilhar um sentimento (*e.g.*, luto, alegria, raiva, *etc*.) até um posicionamento político. Assim, ferramentas de redes sociais, junto com a participação e utilização dos usuários, são verdadeiras geradoras de grandes volumes de conteúdo [Vermelho et al. 2014]. A escala com que os dados são produzidos pode ser de fato intimidadora: o Twitter, por exemplo, possuiu um total de 1,3 bilhões de contas no ano de

*O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. A pesquisa foi também apoiada parcialmente por CNPq e FAPERJ. Os autores também agradecem a Snap Inc.

2022, sendo 330 milhões de usuários altamente ativos. Todos os dias são enviados aproximadamente 500 milhões de *tweets*[1] Se levarmos em consideração eventos especiais e datas comemorativas, a velocidade e volume de postagens podem aumentar em até 25 vezes[2].

A extração de conhecimento útil deste volume de dados não é uma tarefa trivial [Likhitha et al. 2019]. Em especial, um dos desafios para analisar textos em redes sociais consiste na identificação e categorização do assunto tratado. Essa categorização é importante para diversas áreas como o jornalismo, onde os especialistas devem ser capazes de detectar *Breaking News* e avaliar *Fake news*. Para abordar essa tarefa, existem diversos trabalhos na área de aprendizado de máquina, em especial os que focam na tarefa de Modelagem de Tópicos (MT) [Likhitha et al. 2019]. Esses trabalhos visam encontrar assuntos representados por grupos de palavras (tópicos) em um conjunto de textos [Qiang et al. 2020]. Contudo, o volume de dados a ser analisado é somente um dos problemas a ser considerado.

Uma das maneiras de lidar com o volume de dados gerados em redes sociais consiste em limitar a quantidade de caracteres que um usuário pode utilizar em uma postagem na rede social. No caso do Twitter, um *tweet* é atualmente limitado a 280 caracteres, incluindo *hashtags*, *links* e *emojis*. Ao mesmo tempo em que essa restrição na quantidade de caracteres promove a criatividade dos usuários, ela acaba trazendo mais desafios para os algoritmos que podem ser aplicados para categorizar esses textos. Em geral os *tweets* possuem poucas palavras, muitas *hashtags*, ruídos e são sensíveis ao tempo [Li et al. 2018]. Tais características acabam levando à falta de contexto e conteúdo para a categorização do assunto tratado em cada publicação.

Abordagens probabilísticas para modelagem de tópicos, tais como os modelos clássicos *Probabilistic Latent Semantic Analysis* (PLSA) [Hofmann 2013] e *Latent Dirichlet Allocation* (LDA) [Blei et al. 2003], são amplamente adotadas para descobrir os assuntos latentes em uma coleção de documentos e fornecem de maneira simples a possibilidade de analisar grandes volumes de textos sem a necessidade de anotação prévia ou etiquetagem de documentos. Os tópicos são formados por agrupamentos de palavras, que frequentemente ocorrem em conjunto nos documentos [Likhitha et al. 2019]. No entanto, métodos tradicionais de modelagem de tópicos enfrentam dificuldades ao tentar identificar os assuntos tratados em textos curtos, principalmente pela falta de coocorrência de palavras [Costa and Duarte 2019]. Mais recentemente, novas abordagens foram desenvolvidas que se apóiam na semântica distribucional e na utilização de vetores densos de palavras para compor tópicos [Egger and Yu 2022]. Entretanto, problemas como a falta de coocorrência de palavras, ausência de informações nos textos e o uso de termos inseridos apenas no contexto das redes sociais ainda se mostram como desafios [Qiang et al. 2020].

Dessa forma, este artigo avalia um conjunto de abordagens para modelagem de tópicos no contexto de textos curtos. São comparados quatro métodos de modelagem de tópicos usados com frequência na literatura, para mostrar experimentalmente as vantagens e desvantagens de cada uma das abordagens. Para a análise experimental, foram escolhidos o LDA [Blei et al. 2003], o *Gibbs Sampling Dirichlet Multinomial Mixture* (GSDMM) [Yin and Wang 2014], o *Pseudo-document based Topic Model* (PTM) [Zuo et al. 2016] e o BERTopic [Grootendorst 2022]. Uma vez que tais métodos não geram como saídas tópicos rotulados (*i.e.*, o tópico identificado é composto de múltiplas palavras), pode ser difícil ana-

[1]https://www.websiterating.com/pt/research/twitter-statistics/referências

[2]https://blog.twitter.com/engineering/en$_us/a/2013/new-tweets-per-second-record-and-how$

lisar a semântica de tais tópicos. Assim, o artigo também explora um método para rotulá-los, de forma a tornar a comparação entre os diferentes métodos não apenas quantitativa, mas também qualitativa. De forma resumida, este artigo possui as seguintes contribuições: (i) uma análise comparativa de diferentes métodos de modelagem de tópicos em um conjunto de dados com textos curtos, (ii) uma metodologia para rotulação automática de tópicos usando uma fonte de textos externa, *e.g.*, *Wikipédia*, e (iii) uma metodologia para seleção do melhor método de modelagem de tópicos a partir de análises quantitativa e qualitativa, de métricas clássicas e dos rótulos obtidos.

Esse artigo se encontra no contexto de um projeto de pesquisa entre a empresa *Snap Inc.* e a Universidade Federal Fluminense e está organizado em cinco seções, além desta introdução. A Seção 2 discute os métodos para a modelagem de tópicos considerados em nossa análise. A Seção 3 descreve a metodologia proposta para a avaliação experimental dos métodos de modelagem de tópicos. A Seção 4 apresenta e discute os resultados experimentais. A Seção 5 descreve os trabalhos relacionados. Finalmente, a Seção 6 apresenta conclusões e trabalhos futuros.

2. Métodos para a Modelagem de Tópicos

Esta seção descreve os métodos de modelagem de tópicos utilizados na avaliação experimental: o *Latent Dirichlet Allocation* (LDA) [Blei et al. 2003], o *Gibbs Sampling Dirichlet Multinomial Mixture* (GSDMM) [Yin and Wang 2014], o *Pseudo-document based Topic Model* (PTM) [Zuo et al. 2016] e o BERTopic [Grootendorst 2022]. Apesar de existirem outros métodos de modelagem de tópicos na literatura [Qiang et al. 2020], os métodos selecionados para a análise comparativa do presente artigo foram escolhidos de forma a termos uma variedade de métodos. Enquanto os métodos LDA e BERTopic têm o foco em textos de formatos e estilos genéricos, os métodos GSDMM e PTM foram desenvolvidos para lidar com textos curtos. O método LDA é um dos mais adotados na literatura e é uma abordagem probabilística generativa baseada na coocorrência de palavras. O BERTopic se baseia em *embeddings*, sendo uma alternativa para os problemas de coocorrência de palavras em textos curtos, enfrentados pelos demais métodos.

LDA. O LDA induz modelos probabilísticos generativos para coleções de documentos, sendo utilizado em diversas áreas de pesquisa [Jelodar et al. 2019]. Os tópicos são definidos como distribuições de probabilidade sobre um vocabulário fixo de palavras (definido de acordo com os documentos disponíveis). Um determinado tópico é identificado sempre que um mesmo subconjunto de palavras se mostra presente em múltiplos documentos de forma frequente. O LDA se vale de dois hiper-parâmetros, que são oriundos da distribuição de Dirichlet: α_{LDA} que representa a densidade de tópicos por documento e β_{LDA} que representa a densidade de termos por tópico. O modelo não foi desenvolvido para textos curtos [Hong and Davison 2010], portanto, identificar tópicos a partir deste tipo de dados pode ser desafiador, devido ao número reduzido de palavras por documento (*e.g.*, *tweet*) e aos contextos limitados, o que acarreta na falta de informação de coocorrência de palavras em comparação com documentos maiores, dificultando a execução de métodos probabilísticos [Wu et al. 2020].

GSDMM. É um algoritmo de agrupamento de textos curtos adaptado a partir do LDA em que cada documento contém apenas um tópico. Trabalhos anteriores [Qiang et al. 2020] apontam que com uma seleção adequada de valores para os hiperparâmetros, o modelo gerado pelo GSDMM converge rapidamente, o que o caracteriza como uma opção eficiente

e escalável [Yin and Wang 2014]. O GSDMM depende de dois hiperparâmetros: α_{GSDMM} que estabelece a probabilidade de um documento ser agrupado em um tópico; e β_{GSDMM} que quantifica a influência das palavras utilizadas, visando manter documentos com palavras distintas em tópicos distintos. Uma outra particularidade do GSDMM é que não se considera a repetição de palavras em um determinado documento, partindo da ideia de que textos curtos tendem a ser menos repetitivos. Esse método não é capaz de agrupar totalmente palavras que são semanticamente relacionadas, mas que raramente ocorrem simultaneamente.

PTM. O PTM utiliza a auto-agregação para combinar textos curtos em pseudodocumentos maiores. Esse processo melhora as informações de simultaneidade de palavras e soluciona o problema da escassez de dados. Assim, a modelagem de distribuições de tópicos para textos curtos é transformada na modelagem de tópicos de pseudodocumentos. Embora o PTM aprenda as distribuições de tópicos a partir de pseudodocumentos, seu processo de inferência envolve a amostragem de tópicos e a agregação de texto, os quais podem ser muito complexos e demorados.

BERTopic. O BERTopic é um método que engloba algoritmos para busca automática de tópicos densos em uma coleção de documentos, assumindo que documentos semanticamente semelhantes formam tópicos. Diferente do LDA, o BERTopic fornece a modelagem de tópicos contínua em vez de discreta [Alcoforado et al. 2022]. Sendo assim, ele aproveita os *embeddings* do *framework Sentence*-BERT (SBERT) [Reimers and Gurevych 2019], além de utilizar a valoração de *Term Frequency-Inverse Document Frequency* (TF-IDF) [Aizawa 2003] para criar tópicos, permitindo tópicos facilmente interpretáveis e mantendo palavras relevantes nas descrições dos tópicos. O algoritmo executa em três etapas: (i) cada documento é convertido para a sua representação de *embedding* utilizando um modelo linguístico pré-treinado, (ii) antes de agrupar os *embeddings*, a sua dimensionalidade é reduzida, e (iii) a partir dos tópicos de documentos, as representações de tópicos são extraídas usando uma variação personalizada de TF-IDF. Como principais desvantagens, o BERTopic permite a construção de uma representação contextual dos documentos através de seus modelos de linguagem baseados em transformadores, porém os tópicos são gerados a partir de *bag-of-words* [Grootendorst 2022]. Como resultado, palavras em um tópico podem ser semelhantes umas às outras e podem ser redundantes para a interpretação do tópico. Além disso, o modelo assume que cada documento contém apenas um único tópico.

3. Metodologia para a Avaliação Experimental da Modelagem de Tópicos

Esta seção descreve as etapas da metodologia empregada neste artigo para a avaliação experimental dos métodos de modelagem de tópicos. O processo, que pode ser visualizado na Figura 1, é composto de quatro macro-atividades, a saber: (I) Preparação de Dados, (II) Aplicação dos Métodos, (III) Avaliação Quantitativa, e (IV) Avaliação Qualitativa dos Tópicos. Na etapa (I) é realizada a coleta dos dados (*tweets* no contexto desse artigo) dentro de uma janela de tempo definida e o pré-processamento para a aplicação de cada um dos métodos. Em seguida, na etapa (II) são selecionados os diferentes métodos de modelagem de tópicos e seus hiperparâmetros são ajustados. Já nas etapas (III) e (IV) são avaliados os resultados de forma quantitativa e qualitativa de forma a identificar o melhor método. Ainda como parte da etapa da avaliação qualitativa, os tópicos construídos recebem rótulos com o intuito de examinar o desempenho dos métodos. Muitas das decisões quanto à metodologia foram baseadas nos dados oriundos do Twitter (que utilizamos como estudo de caso). Porém, em um contexto diferente, com outros tipos de dados, modificações pontuais

na metodologia podem ser necessárias, em especial nas atividades de preparação dos dados. A seguir, discutimos com mais detalhes cada uma das macro-atividades citadas.

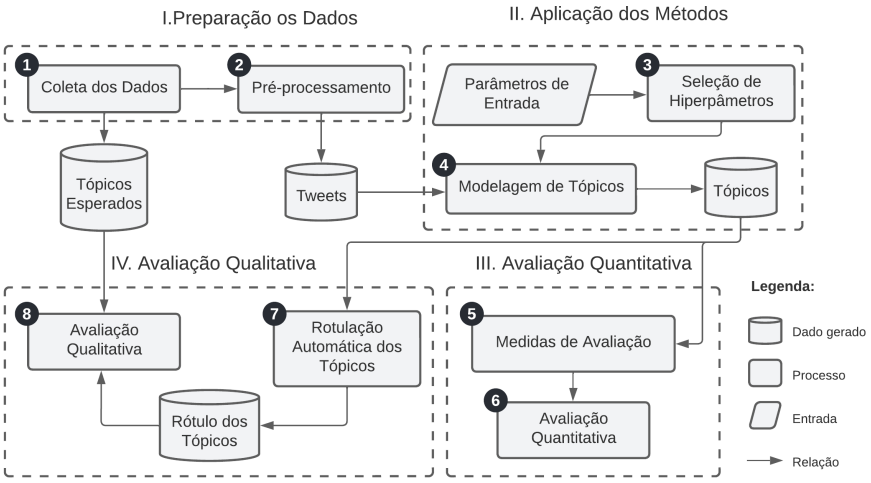

Figura 1. Metodologia para a avaliação experimental da modelagem de tópicos.

3.1. Preparação dos Dados

A macro-atividade de Preparação dos Dados pode ser decomposta em duas atividades: (i) Coleta dos Dados e (ii) Pré-processamento. A atividade de Coleta dos Dados é o momento em que os *tweets* são coletados por meio da API *Twitter Stream*[3]. Em nossos experimentos, nos baseamos em um conjunto de *tweets* publicados na janela de tempo entre 13/02/2022 e 15/02/2022. Foram considerados somente *tweets* em língua inglesa (en-us). Essa janela de tempo foi escolhida de forma que pudéssmos realizar a análise dos dados com foco no *Valentine's Day* nos EUA (14/02/2022). De forma a definir qual o subconjunto de *tweets* dessa janela que seriam considerados no experimento, foram selecionadas 10 categorias de assuntos mais relevantes de acordo com os termos que eram tendências durante a janela de tempo (*i.e.*, *trending terms*) no *Twitter*[4] e no *Google Trends*[5]. Assim, as seguintes categorias foram consideradas: (a) Esporte, (b) Filmes e Séries, (c) Música, (d) Política, (e) Saúde, (f) Romance, (g) Religião, (h) Comida, (i) Jogos e (j) Clima. A categoria "Romance" se encontra diretamente relacionada ao evento *Valentine's Day*. Cada categoria contém oito palavras-chave coletadas no *Twitter Trends* e no *Open Directory Project* (ODP) [6]. Assim, o conjunto de dados utilizado é composto de *tweets* postados durante a janela de tempo definida e que continham pelo menos uma das categorias. Na atividade de coleta de dados ainda são excluídos *retweets*, comentários ou citações. Enquanto *retweets* e citações são repetições dos *tweets* originais, comentários podem ser uma digressão do assunto principal, introduzindo um ruído nos dados. O conjunto final de dados contém um total de 45.097 *tweets*.

Uma vez definido o conjunto de dados que será considerado no experimento, inicia-se a atividade de pré-processamento. Na atividade de pré-processamento, os *tweets* são normalizados com a remoção de menções de usuários, *links*, *hashtags*, caracteres e

[3]https://developer.twitter.com/en
[4]https://trends24.in/
[5]https://trends.google.com.br/trends/?geo=US
[6]http://www.odp.org/homepage.php

pontuações, números e espaços, além de transformar *emojis* em textos representativos (*e.g.*, uma carinha triste na *string* '*sad*'). Em seguida, é realizada a tokenização por palavra de cada *tweet* e a remoção das *stopwords* usando o Gensim [Rehurek and Sojka 2011][7] e o NLTK [Bird et al. 2009][8] (*Natural Language ToolKit*), respectivamente. O modelo BERTopic contém tokenização própria, mas assim como os demais modelos, pode receber as palavras tokenizadas. Além disso, termos ou palavras muito frequentes (*i.e.*, presentes em mais de 50% dos *tweets*), assim como termos que aparecem apenas uma vez no conjunto de dados, foram removidos.

3.2. Aplicação dos Métodos

Definir valores de hiperparâmetros em experimentos de aprendizado de máquina pode ser uma tarefa complexa. Diferentes combinações de hiperparamêtros podem ser exploradas com o apoio de uma varredura de parâmetros (*i.e.*, *grid search*). Os hiperparâmetros considerados no experimento apresentado neste artigo são descritos a seguir, uma vez que, exceto pelo número de tópicos que é um parâmetro comum a todos os métodos, cada método de modelagem de tópicos possui seu próprio conjunto de hiperparâmetros. A métrica de coerência dos tópicos foi utilizada para identificar a melhor combinação de hiperparâmetros para cada método durante a varredura de parâmetros realizada. Como 10 categorias de tópicos foram consideradas na coleta de dados (Seção 3.1), os seguintes valores para o número de tópicos foram considerados: 9, 10, 11 e 12. Cada decisão do processo de seleção de hiperparâmetro é discutida nos parágrafos seguintes.

No caso do LDA[9], os valores de hiperparâmetros definidos são α_{LDA} = *asymmetric* (o hiperparâmetro *asymmetric* para α_{LDA} define uma distribuição a priori normalizada e fixa com o valor de $1.0/(\text{indice\_do\_topico} + \sqrt{\text{numero\_de\_topicos}})$), β_{LDA} = 1,0 e número de tópicos (*K*) = 9. No caso do GSDMM[10], os valores de hiperparâmetros definidos são *K* = 12, α_{GSDMM} = 0,01 e β_{GSDMM} = 1,0. No caso do PTM[11], os valores de hiperparâmetros definidos são α_{PTM} = 1,0, β_{PTM} = 0,01 e *K* = 11, com 100 iterações. O número de pseudo-documentos foi fixado em 1.000 [Zuo et al. 2016], além de definir *TermWeight-ONE* [Wilson and Chew 2010] como o esquema uniforme de ponderação de termos. Finalmente, no caso do BERTopic[12] o método demonstrou melhor desempenho com a combinação dos modelos SBERT[13] (*embedding* de sentenças) e RoBERTa, na versão *large*[14] (*embedding* de palavras). Além disso, melhor valor para *K* nesse método foi 12.

3.3. Avaliação Quantitativa

O desempenho dos métodos de modelagem de tópicos explorados neste artigo é avaliado por meio de duas métricas, a saber, a coerência [Röder et al. 2015] e a diversidade dos tópicos. A coerência dos tópicos é avaliada por meio da média da *Coherence* C_V e da *Coherence* NPMI. Ambas medem a relação entre as dez primeiras palavras de um tópico, considerando a frequência das palavras no *corpus* original. A *Coherence C_V* é baseada em

[7]https://radimrehurek.com/gensim/
[8]https://www.nltk.org/
[9]https://radimrehurek.com/gensim/models/ldamodel.html
[10]https://github.com/rwalk/gsdmm.git
[11]https://bab2min.github.io/tomotopy/v0.12.3/en/tomotopy.PTModel
[12]https://github.com/MaartenGr/BERTopic.git
[13]https://www.sbert.net/
[14]https://huggingface.co/roberta-large

uma janela deslizante, na segmentação de um conjunto das principais palavras e em uma medida de confirmação indireta que usa *Normalized Pointwise Mutual Information* (NPMI) [Bouma 2009] e a similaridade de cosseno. Já a *Coherence* NPMI é baseada apenas na janela deslizante e no cálculo do NPMI de todos os pares de palavras das principais palavras fornecidas.

A métrica de diversidade dos tópicos permite medir a interpretabilidade dos tópicos, avaliando o quão diversos são os tópicos gerados por um modelo. A diversidade é calculada por meio da média das medidas de diversidade de tópicos e *Rank-Biased Overlap* invertido (*Inverted* RBO). O *Inverted* RBO é uma medida de desarticulação entre os tópicos ponderados nas classificações de palavras. A métrica *Inverted RBO* avalia o quão diversos são os tópicos gerados por um método, comparando os tópicos dois a dois. O cálculo das duas métricas de diversidade é feito sobre as 10 principais palavras de cada tópico. O resultado igual a 0,0 indica que os tópicos são considerados idênticos e 1,0 que os tópicos são completamente diferentes.

3.4. Rotulação

O processo para avaliar os tópicos induzidos de forma qualitativa baseia-se na análise dos rótulos identificados de forma automática para os tópicos latentes. O objetivo de gerar tais rótulos é associar ideias ou conceitos formados pela conexão de palavras ou expressões que não se encontram presentes nos principais termos dos tópicos. Idealmente, um bom rótulo nos permite identificar o assunto de cada tópico. O processo de avaliação qualitativa dos tópicos a partir da tarefa de identificação de rótulos é composto das seguintes macro-atividades (Figura 2): (I) Coleta dos Dados, (II) Construção dos Candidatos, (III) Vetorização e (IV) Seleção dos Rótulos. A macro-atividade (I) coleta títulos de artigos da *Wikipédia*[15] em inglês, usando para busca os termos associados a um determinado tópico. Em seguida, na macro-atividade (II), os possíveis candidatos a rótulo de cada tópico são construídos a partir da sequência de N palavras (*n-gram*) dos substantivos e adjetivos extraídos dos títulos coletados da *Wikipedia*. Por fim, são selecionados os X rótulos candidatos com o maior valor de PMI, calculado a partir das palavras no rótulo. A macro-atividade (III) consiste em identificar a similaridade entre as palavras dos candidatos e os termos dos tópicos, usando os *embeddings* de palavras obtidos pelo método *fastText*[16] [Bojanowski et al. 2017], dada sua habilidade de manipular subpalavras, o que é essencial para textos com ruído oriundos de redes sociais. Por fim, a macro-atividade (IV) verifica a similaridade de cada candidato com os tópicos, por meio do cálculo normalizado da média ponderada das seguintes métricas: similaridade de cosseno, distância Euclidiana, distância de Manhattan e *word mover's distance* (WMD) [Huang et al. 2016]. De acordo com a revisão da literatura realizada, a similaridade de cosseno é a métrica mais utilizada na avaliação dos tópicos gerados. Portanto, a similaridade de cosseno foi utilizada com peso superior às demais médias em nossa análise.

4. Avaliação Experimental

Nessa seção discutimos os resultados obtidos a partir da aplicação da metodologia descrita na Seção 3 para o conjunto de *tweets* coletados na janela de tempo entre 13/02/2022 e 15/02/2022. Os resultados experimentais são avaliados sob duas óticas: (i) quantitativa,

[15]https://en.wikipedia.org/wiki/English$_Wikipedia$

[16]https://fasttext.cc/

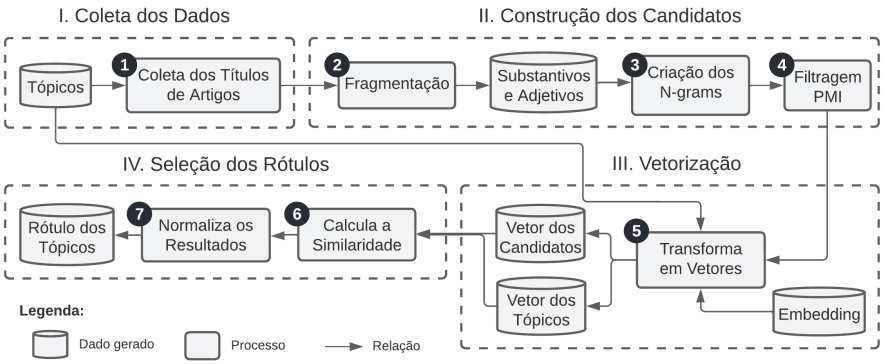

Figura 2. Rotulação automática dos tópicos

tomando como base as métricas discutidas na Seção 3.3, e (ii) qualitativa, tendo como apoio o método de rotulação discutido na Seção 3.4.

4.1. Análise Quantitativa

Conforme apresentado na Tabela 1, o método GSDMM apresentou a melhor coerência e um resultado aceitável para a diversidade dos tópicos identificados. Não é uma surpresa que o GSDMM tenha obtido o resultado com melhor coerência, uma vez que este método foi projetado para lidar com textos curtos. Entretanto, os resultados obtidos com o LDA se mostraram surpreendentes, pois mesmo sendo o método mais antigo e baseado em matrizes de coocorrência, ele obteve resultados de diversidade e coerência superiores a métodos mais novos como o BERTopic.

Tabela 1. Comparação quantitativa dos modelos.

Modelos	Coerência C_V	Coerência NPMI	Média das Coerências	Diversidade	Inverted RBO	Média das Diversidades
LDA	0,597	0,123	0,360	**1,000**	**1,000**	**1,000**
GSDMM	**0,623**	**0,209**	**0,416**	0,833	0,973	0,903
PTM	0,607	0,142	0,374	0,882	0,979	0,931
BERTopic	0,492	-0,104	0,194	0,983	0,999	0,991

Entretanto, o LDA apresentou uma coerência inferior aos métodos específicos para textos curtos como o PTM e o GSDMM. O BERTopic demonstrou uma boa diversidade dos tópicos, mas ao mesmo tempo apresentou o pior resultado de coerência dentre os métodos avaliados. O LDA e o BERTopic são dois métodos que não são específicos para textos curtos, porém ambos apresentaram os melhores resultados para a diversidade dos tópicos.

4.2. Análise Qualitativa

Na análise qualitativa, os tópicos gerados pelos métodos e seus respectivos rótulos foram relacionados com os tópicos esperados. Os tópicos esperados correspondem aos 10 tópicos ou categorias originalmente considerados na coleta de *tweets*. A Tabela 2 descreve os resultados qualitativos. Para cada tópico identificado (com as *Top* 10 palavras do tópico), apresentamos a categoria em que se enquadra, o rótulo aplicado e o método usado para identificação do tópico.

Conforme a Tabela 2 indica, o LDA identificou tópicos para apenas quatro das categorias selecionadas, enquanto os outros métodos foram capazes de identificar tópicos associados aos demais temas. Ainda, a maioria dos tópicos identificados pelo LDA apresentaram palavras chaves ou conteúdo de outros tópicos. Alguns tópicos identificados pelo

LDA também receberam o mesmo rótulo, sendo referenciados ao mesmo tópico esperado. O LDA é o único método que identificou menos tópicos que a quantidade de categorias.

Tabela 2. Comparação qualitativa dos tópicos identificados pelos métodos.

Categoria	Tópico	Rótulo	Modelo
Clima	Reports, Station, Gusts, Sustained, Canyon, Winds, Fremont, Temp, Mesonet, Inch	Storm Wind Speeds	LDA
	Mph, Humidity, Wind, Pressure, Gust, Feels, Temperature, Rain, Today, Barometer	Storm Speeds	GSDMM
	Mph, Humidity, Wind, Pressure, Gust, Temperature, Rain, Today, Barometer, Rising	Cyclone Speed Climatology	PTM
	Mph, Humidity, Wind, Gust, Pressure, Barometer, Temperature, Slowly, Rain, Rising	Thunderstorm Surface Weather	BERTopic
Comida	Cookery, Raspberries, Charred, Graffiti, Maple, Glazed, Chop, Carrots, Pork, Strawberry	Chicken Bacon	LDA
	Food, Pizza, Cheese, Amp, Bowl, Valentine, Super, Day, Poisoning, Chicken	Breakfast Foods Steak	GSDMM
	Food, Pizza, Amp, Cheese, Dinner, Eat, Chicken, Poisoning, Eating, Restaurant	Breakfast Foods Steak	PTM
Esporte	Bowl, Super, Football, Sunday, Like, Time, Game, Today, Halftime, Football	Monday Night Baseball	GSDMM
	Bowl, Super, Sunday, Football, Halftime, Game, Football, Rams, Win, Bengals	College Playoff	PTM
	Basketball, Nba, Varsity, Def, Pts, Tournament, College, Grade, Lakers, Warriors	Championship Volleyball	BERTopic
Música	Pick, Song, Favorite, Enter, App, Learn, Hop, Insert, Hip, West	Rip Ride	GSDMM
	Like, Music, Happy, Song, Movie, Know, People, Cry, Laughing, Time	Think Dance	PTM
	Eminem, Snoop, Hip, Hop, Dogg, Dre, Kendrick, Rap, Lamar, Blige	George Floyd	BERTopic
Política	Trump, Health, Russia, Biden, Ukraine, Christ, Job, Jesus, President, Latest	America Immigration	LDA
	Trump, Russia, Ukraine, Biden, Health, Amp, Putin, Like, People, President	Saudi Arabia	GSDMM
	Trump, Ukraine, United, Russia, Biden, Putin, President, Russian, Amp, War	America Union Soviet	PTM
	Ukraine, Russia, Putin, Russian, Invasion, War, Invade, Troops, Olympics, Nato	Europe Eurasia	BERTopic
Religião	Music, Church, Amp, New, Love, Jesus, God, Song, Movie, Photo	Romance Songs Saint	GSDMM
	Jesus, God, Christ, Life, Love, Church, Lord, Children, John, World	America Prayer Catholic	PTM
	Christ, Shall, Lord, Anan, Father, Pray, Elohim, Psalms, Amen, Hebrews	Son God	BERTopic
Romance	Valentine, Day, Happy, Heart, Love, Music, Mph, Amp, Bowl, Super	Good Friday Prayer	LDA
	Valentine, Day, Happy, Heart, Love, Today, Hearts, Affection, Amp, Heart	Good Friday Prayer	GSDMM
	Day, Valentine, Heart, Love, Happy, Affection, Hearts, Heart, Rose, Crush	Saint Kiss	PTM
	Kiss , Love, Infatuation, Kiss, Wine, Check, Valentines, Fingers, Prohibited, Smiling	Happy Song Kisses	BERTopic

Os tópicos identificados pelos métodos GSDMM e PTM se encaixam em sete das categorias de tópicos esperados. Entretanto, a Tabela 2 nos mostra que o tópico identificado na categoria "Romance" do método GSDMM não recebeu rótulos relacionados ao assunto, apesar das palavras pertencerem ao tema. Ambos os métodos apresentaram o tema "Esporte" em dois tópicos com assuntos distintos. Nos demais, identificaram tópicos com termos bastante semelhantes, mas, apesar da proximidade, esses tópicos foram associados com rótulos diferentes. Conforme pode ser visto, os métodos associaram rótulos bem relacionados ao tema dos tópicos identificados, porém, o método PTM obteve um desempenho superior ao método GSDMM.

O BERTopic identificou seis tópicos relacionados às categorias escolhidas, e apresentou mais repetições dos temas nos rótulos que os demais modelos executados. O método gerou três tópicos para o tema 'Religião', dois para "Clima" e dois em "Romance". Apesar de não obter uma grande variedade dos temas, o método construiu a maioria dos tópicos interpretáveis por meio dos rótulos bem relacionados aos temas. Entretanto, a categoria "Música" apresentou um rótulo não relacionado ao tópico, *i.e.*, "George Floyd".

Comparando os métodos, o modelo LDA apresentou os piores tópicos e rótulos. O GSDMM e o PTM foram os métodos que reconheceram a maior quantidade de temas das categorias definidas. Entretanto, comparando os dois métodos, o método PTM foi capaz de associar mais rótulos relacionados aos temas das categorias. O modelo GSDMM utilizou palavras mais genéricas nos tópicos. Já os tópicos do modelo BERTopic foram estruturados com palavras específicas ao tema, portanto, apresentando bons rótulos. Os modelos PTM e BERTopic apresentaram os melhores desempenhos para os dados analisados.

As categorias "Filmes e Séries", "Saúde" e "Jogos" não aparecem em nenhum dos tópicos identificados, diferente dos temas "Clima", "Romance" e ''Política" que tiveram tópicos identificados por todos os métodos. Esse resultado pode estar relacionado à quantidade de palavras pertencentes a cada tema, e se essas palavras aparecem de forma simultânea nos *tweets*, permitindo que os métodos identifiquem a correlação entre elas. Os tópicos ba-

seados em "Comida", "Clima", ''Política", "Esporte" e "Religião" possuem a melhor junção de termos relacionados ao tema específico e, consequentemente, melhores rótulos.

5. Trabalhos Relacionados

Algoritmos já bem conhecidos para a modelagem de tópicos, *e.g.*, *Probabilistic Latent Semantic Analysis* (PLSA) [Hofmann 2013] e *Latent Dirichlet Allocation* (LDA) [Blei et al. 2003], são adotados para a descoberta da estrutura semântica latente a partir de textos, sem a necessidade de qualquer anotação prévia. Entretanto, tais métodos não foram desenvolvidos com foco na análise de textos curtos e podem apresentar dificuldades ao modelar esse tipo de texto. Assim, novos métodos têm sido propostos com foco específico em textos curtos. Trabalhos anteriores já compararam o desempenho de diferentes métodos de modelagem de tópico propostos. Nessa seção discutimos os principais. Em [Qiang et al. 2020] os autores investigam três categorias de métodos para a modelagem de tópicos com textos curtos: métodos baseados na *Dirichlet Multinomial Mixture* (DMM) (GSDMM, LF-DMM, GPU-DMM e GPU-PDMM), métodos baseados na ocorrência simultânea de palavras (BTM e WNTM) e métodos baseados na auto-agregação (SATM e PTM). Na análise experimental apresentada no artigo, os métodos baseados em DMM alcançam o melhor desempenho na coerência dos tópicos em todos os conjuntos de dados analisados. O WNTM tem um desempenho melhor do que o BTM com o uso de dados do *Twitter*. Já o PTM apresentou resultados promissores no conjunto de dados com *tweets*.

[Costa and Duarte 2019] avaliaram métodos de modelagem de tópicos para textos curtos que usam abordagens probabilísticas. Nesse estudo, quatro métodos foram selecionados: BTM, PTM, SATM e WNTM. Considerando métricas de coerência dos tópicos, conjuntos de dados e número de tópicos, as execuções do BTM obtiveram os resultados mais coerentes em mais cenários do que as outras abordagens. Em [Albalawi et al. 2020], os autores também investigaram experimentalmente a eficácia de métodos de modelagem de tópicos utilizando textos curtos. Nesse estudo, os métodos LDA e NMF (*Non-Negative Matrix Factorization* [Lee and Seung 1999]) geraram os resultados com maior qualidade. O NMF usa uma abordagem de álgebra linear para extração de tópicos. [Egger and Yu 2022] avaliam e comparam o desempenho de quatro técnicas de modelagem de tópicos: LDA, NMF, Top2Vec e BERTopic. Além dos métodos já citados anteriormente, o Top2Vec usa *embeddings* para representar os textos curtos. Este estudo utiliza postagens do *Twitter* restritas à viagens e à pandemia da COVID-19. Os autores identificaram que o modelo LDA produziu tópicos mais genéricos e irrelevantes. Ao comparar o BERTopic ao NMF, reconheceram que uma grande deficiência do NMF gira em torno de sua baixa capacidade de identificar significados embutidos dentro de um *corpus*. Por fim, avaliaram que o modelo BERTopic foi capaz de gerar novas perspectivas usando a abordagem de *embeddings*.

Ao analisar os trabalhos apresentados nessa seção, percebemos que alguns autores elaboraram um estudo exploratório de métodos específicos para textos curtos [Qiang et al. 2020, Costa and Duarte 2019], enquanto outros focaram em avaliar o desempenho de métodos de modelagem de tópicos para textos genéricos, mas aplicados no contexto de textos curtos [Albalawi et al. 2020]. Por fim, as abordagens apresentadas verificaram o desempenho apenas de métodos de textos genéricos com e sem o uso de *embeddings* [Egger and Yu 2022]. Nenhum dos trabalhos relacionados explorou o que este artigo propõe: uma análise comparativa entre métodos tradicionais (LDA), modelos para textos curtos (GSDMM e PTM) e novas abordagens para textos genéricos com o uso de *embeddings* (BERTopic). Além disso, com relação aos estudos já realizados, este artigo contribui

com uma técnica de rotulação automática, permitindo aperfeiçoar a etapa de avaliação qualitativa dos tópicos identificados. A Tabela 3 sumariza os trabalhos relacionados de acordo com as suas características principais.

Tabela 3. Características dos trabalhos relacionados

Artigo	Texto Curto	*Embedding*	Redução de Dimensionalidade	Misturas Multinomial Dirichlet	Agregação de Pseudodocumento	Baseado em Coocorrência
[Qiang et al. 2020]	✓	✓		✓	✓	✓
[Costa and Duarte 2019]	✓				✓	✓
[Lossio-Ventura et al. 2021]	✓			✓		✓
[Mazarura and De Waal 2016]	✓			✓		
[Agarwal et al. 2020]	✓			✓		
[Dimitriadis 2020]	✓			✓		
[Omurca et al. 2021]	✓			✓		
[Egger and Yu 2022]		✓	✓			
[Albalawi et al. 2020]			✓			

6. Conclusões e Trabalhos Futuros

Descobrir estruturas semânticas, *e.g.* tópicos, a partir de textos curtos é uma tarefa desafiadora, considerando principalmente o estilo de escrita ruidoso empregado nas redes sociais. Os desafios a serem superados incluem a falta de simultaneidade das palavras, ausência de contexto, estilo de escrita próprio e uso excessivo de *emojis* e gírias. Portanto, uma análise exploratória dos métodos de modelagem de tópicos é essencial para identificar quais são os métodos mais adequados para lidar com os problemas mencionados.

Neste artigo, realizamos uma avaliação experimental dos métodos de modelagem de tópicos LDA, GSDMM, PTM e BERTopic. O LDA não foi desenvolvido especificamente para textos curtos, mas é um dos métodos mais utilizados para a modelagem de tópicos na literatura. O GDSMM e o PTM foram desenvolvidos com o foco em textos curtos, sendo que o último é um método baseado na auto-agregação de pseudodocumentos. Finalmente, o BERTopic usa a representação vetorial do BERT ou outra técnica de *embedding* para lidar melhor com textos ruidosos. Além disso, reduz a dimensionalidade dos *embeddings* e agrupa os documentos semanticamente semelhantes. Usando esses métodos, elaboramos uma análise comparativa com o objetivo de identificar e avaliar os tópicos gerados por meio de um conjunto de dados de textos curtos.

O LDA apresentou um bom resultado de diversidade, mas quando se trata da avaliação qualitativa, não obteve um bom desempenho. O BERTopic apresentou resultados promissores de diversidade, mas o pior valor de coerência dos tópicos. Entretanto, esse método identificou bons rótulos e tópicos de fácil compreensão na análise qualitativa. Já o GSDMM e o PTM foram os métodos que reconheceram a maior quantidade de temas das categorias definidas. Ambos também elaboraram rótulos bem relacionados aos tópicos. Contudo, comparando os métodos GSDMM e PTM de forma qualitativa, o método PTM foi capaz de associar mais rótulos relacionados aos temas das categorias. Algumas das palavras dos tópicos gerados pelo GSDMM também são mais genéricas que as utilizadas na construção dos tópicos do PTM. Portanto, pode-se concluir que os modelos PTM e BERTopic apresentaram os melhores desempenhos para os dados analisados pelo experimento. Com isso, podemos citar algumas direções para trabalhos futuros: (i) desenvolver tais processos para avaliar os dados em diferentes idiomas, (ii) avaliação dos tópicos gerados por meio de ferramentas de visualização, (iii) investigar métodos de rotulação de tópicos e (iv) executar a metodologia com um conjunto maior de dados.

Referências

Agarwal, N., Sikka, G., and Awasthi, L. K. (2020). Evaluation of web service clustering using dirichlet multinomial mixture model based approach for dimensionality reduction in service representation. *IP&M*, 57(4):102238.

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. *Inf. Proc. & Management*, 39(1):45–65.

Albalawi, R., Yeap, T. H., and Benyoucef, M. (2020). Using topic modeling methods for short-text data: A comparative analysis. *Frontiers in Artificial Intelligence*, 3:42.

Alcoforado, A., Ferraz, T. P., Gerber, R., Bustos, E., Oliveira, A. S., Veloso, B. M., Siqueira, F. L., and Costa, A. H. R. (2022). Zeroberto–leveraging zero-shot text classification by topic modeling. *arXiv*.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc.".

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

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguistics*, 5:135–146.

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

Costa, M. and Duarte, D. (2019). Avaliação de abordagens probabilísticas de extração de tópicos em documentos curtos. In *Anais da XV Escola Regional de Banco de Dados*, pages 51–60. SBC.

Dimitriadis, N. S. (2020). Applying topic modelling algorithms on twitter messages in greek language. *Graduate Thesis. Aristotle University of Thessaloniki*.

Egger, R. and Yu, J. (2022). A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in Sociology*, 7.

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

Hofmann, T. (2013). Probabilistic latent semantic analysis. *arXiv preprint arXiv:1301.6705*.

Hong, L. and Davison, B. D. (2010). Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*, pages 80–88.

Huang, G., Guo, C., Kusner, M. J., Sun, Y., Sha, F., and Weinberger, K. Q. (2016). Supervised word mover's distance. *NeurIPS*, 29.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., and Zhao, L. (2019). Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *M. Tools and App.*, 78(11):15169–15211.

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

Li, X., Wang, Y., Zhang, A., Li, C., Chi, J., and Ouyang, J. (2018). Filtering out the noise in short text topic modeling. *Information Sciences*, 456:83–96.

Likhitha, S., Harish, B., and Kumar, H. K. (2019). A detailed survey on topic modeling for document and short text data. *Int. J. of Computer App.*, 178(39):1–9.

Lossio-Ventura, J. A., Gonzales, S., Morzan, J., Alatrista-Salas, H., Hernandez-Boussard, T., and Bian, J. (2021). Evaluation of clustering and topic modeling methods over health-related tweets and emails. *Artificial Intelligence in Medicine*, 117:102096.

Mazarura, J. and De Waal, A. (2016). A comparison of the performance of latent dirichlet allocation and the dirichlet multinomial mixture model on short text. In *PRASA-RobMech*, pages 1–6. IEEE.

Omurca, S. İ., Ekinci, E., Yakupoğlu, E., Arslan, E., and Çapar, B. (2021). Automatic detection of the topics in customer complaints with artificial intelligence. *BJECE*, 9(3):268–277.

Oraby, S., Bhuiyan, M., Gundecha, P., Mahmud, J., and Akkiraju, R. (2019). Modeling and computational characterization of twitter customer service conversations. *ACM Trans. Interact. Intell. Syst.*, 9(2–3).

Qiang, J., Qian, Z., Li, Y., Yuan, Y., and Wu, X. (2020). Short text topic modeling techniques, applications, and performance: a survey. *IEEE Transactions on Knowledge and Data Engineering*.

Rehurek, R. and Sojka, P. (2011). Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408.

Vermelho, S. C., Velho, A. P. M., Bonkovoski, A., and Pirola, A. (2014). Refletindo sobre as redes sociais digitais. *Educação & sociedade*, 35(126):179–196.

Wilson, A. and Chew, P. A. (2010). Term weighting schemes for latent dirichlet allocation. In *human language technologies: The 2010 conf. of the N. American Chap. of the Assoc. for Comp. Linguistics*, pages 465–473.

Wu, X., Li, C., Zhu, Y., and Miao, Y. (2020). Short text topic modeling with topic distribution quantization and negative sampling decoder. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1772–1782.

Yin, J. and Wang, J. (2014). A dirichlet multinomial mixture model-based approach for short text clustering. In *ACM SIGKDD*, pages 233–242.

Zuo, Y., Wu, J., Zhang, H., Lin, H., Wang, F., Xu, K., and Xiong, H. (2016). Topic modeling of short texts: A pseudo-document view. In *ACM SIGKDD*, pages 2105–2114.