Contributions to Social Media Analysis Based on Topic Modelling

Giordanno Brunno Bergamini Gomes, Romis Attux

Universidade Estadual de Campinas, Brazil
g146244@dac.unicamp.br, attux@unicamp.br

Abstract. This work proposes a computational approach to a task deeply related to human sciences, that of employing natural language processing in text analysis. Researchers working in that field are often faced with the need for extracting information from large masses of textual data. One of such applications is topic modelling, a task that requires the discovery of the topics discussed in texts - to deal with it, there are several available techniques, such as Latent Dirichlet Allocation (LDA), Biterm Topic Model (BTM), Topic Bidirectional Encoder Representation from Transformers (BERTopic) and Non-negative Matrix Factorization (NMF). In this work, we design a methodological setup and perform a comparative analysis of the aforementioned techniques over data retrieved from Twitter. Through this social media, we seek to contribute to the study of political, economic and social issues, as well as to assess the relative merits of topic modelling techniques. The results indicate a higher topic coherence performance for BERTopic, second for NMF, followed by BTM and, lastly, by LDA.

CCS Concepts: • **Computing methodologies** → **Information extraction**.

Keywords: computational human sciences, natural language processing, social media analysis, topic modelling

1. INTRODUÇÃO

Uma vertente que aproxima as ciências humanas da computação é, sem dúvida, o processamento de linguagem natural (PLN), já que textos dos mais variados tipos são amplamente utilizados nas humanidades. Há vários exemplos interessantes dessa confluência [Robila and Robila 2020], dentre os quais citamos, a título de exemplo, o trabalho de Sumikawa et al. [2018]. Nele, foi feita a extração de referências temporais de textos da rede social Twitter ao longo de 11 meses, nos anos de 2016 e 2017. Essa extração foi feita por meio de uma ferramenta computacional baseada em regras chamada HeidelTime. Com essas referências temporais, foi possível analisar memórias coletivas [Halbwachs 1950] no Twitter.

Esse exemplo é um claro indicativo de que a *modelagem de tópicos* surge, organicamente, como um caminho para determinar os assuntos discutidos em grandes blocos de textos ou em textos de grande porte. Exemplos de modelos são: a simples contagem da frequência total de palavras no Bag of Words [Nisha and Kumar R 2019], LSA (do inglês Lantent Semantic Analysis) [Deerwester et al. 1990], NMF [Paatero and Tapper 1994], pLSA (do inglês Probabilistic Latent Semantic Analysis) [Hofmann 1999], LDA [Blei et al. 2003], Hierarchical Dirichlet Processes [Teh et al. 2004], STM (do inglês Structural Topic Model) [Roberts et al. 2013], BTM [Yan et al. 2013], W2V-GMM (do inglês word2vec Gaussian mixture model) [Sridhar 2015], CorEx [Gallagher et al. 2017], Top2Vec [Angelov 2020] e BERTopic [Grootendorst 2022b].

As mídias sociais são fontes de grande importância para as humanidades, e uma rede social que se destaca pela grande disponibilidade de textos é o Twitter, que possuía cerca de 368 milhões de

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior – Brasil (CAPES) – Código de Financiamento 001 e CNPq (Proc. 308811/2019-4).

usuários ativos em 2022.[1] Por meio dela, é possível coletar uma grande quantidade de dados para estudar o que está sendo dito por muitas pessoas sobre uma enorme variedade de assuntos. Também é possível estudar os dados por país e outras divisões geográficas.

Neste trabalho, apresentamos uma nova metodologia para análise de mídias sociais, buscando, assim, contribuir para o uso de PLN no contexto das humanidades com foco na realidade brasileira e no idioma português. Também temos como objetivo comparar modelos de diferentes abordagens (probabilística, matricial, especialista em textos curtos, neural de aprendizado profundo) para a modelagem de tópicos nessas mídias e para avaliar aquele que possui maior desempenho para essa tarefa. Como abordagem probabilística, usamos o LDA, como abordagem matricial, temos os NMF, como metodologia especializada em textos curtos, trabalhamos com o BTM, e, como abordagem neural, utilizamos o BERTopic.

O trabalho está estruturado da seguinte forma: a seção 2 traz uma breve revisão dos modelos utilizados; a seção 3 apresenta a metodologia proposta, os resultados estão na seção 4 e as conclusões são expostas na seção 5.

2. BREVE REVISÃO

O modelo de alocação de Dirichlet latente (LDA, do inglês latent Dirichlet allocation) foi proposto por Blei et al. [2003], e se tornou, ao longo dos anos, um dos modelos mais conhecidos e utilizados em modelagem de tópicos. Ele corresponde a um modelo probabilístico generativo de um *corpus*, em que os documentos são representados como misturas aleatórias em tópicos latentes, sendo cada tópico caracterizado por uma distribuição de palavras. Seus hiperparâmetros α e η são referentes à distribuição Dirichlet. O primeiro representa uma suposição *a priori* na distribuição de tópicos de documentos. O segundo representa uma assunção *a priori* na distribuição tópico-palavra. Eles têm a necessidade de serem ajustados, processo que é descrito na seção Metodologia.

O modelo LDA possui limitações: segundo Jónsson [2016], LDA não é um modelo com desempenho notável para modelagem de tópicos em textos curtos. Como um dos objetivos do trabalho é explorar o Twitter, e, sendo ele constituído essencialmente por textos curtos, torna-se evidente que esse problema é importante em se trabalhar. Portanto, foi necessário trabalhar com algum modelo direcionado a textos curtos.

Uma proposição com esse direcionamento é o modelo de tópico bitermo (BTM, do inglês biterm topic model), que foi apresentado por Yan et al. [2013]. Nele, os tópicos são extraídos ao se modelar a geração de padrões de coocorrência de palavras. A justificativa de escolha desse modelo foi por seu desempenho superior em Jónsson [2016].

A Fatoração de Matriz Não-Negativa (NMF, do inglês *Non-Negative Matrix Factorization*) foi introduzida por Paatero and Tapper [1994] sob o conceito de Fatoração de Matriz Positiva, de acordo com Wang and Zhang [2013]. Entretanto, a abordagem não era, então, utilizada para modelagem de tópicos, o que veio a ocorrer posteriormente. A NMF é um procedimento matemático no qual uma matriz formada por valores não-negativos é decomposta em duas novas matrizes, de forma que o produto dessas duas novas matrizes seja igual à matriz original [Churchill and Singh 2022]. A matriz que é decomposta é a *matriz documento-palavra*, que consiste em um conjunto de documentos em que cada documento é representado por um vetor de palavras. As duas matrizes menores resultantes são a *matriz tópico-palavra*, que pode ser interpretada como a distribuição de tópicos em relação às palavras, e a *matriz tópico-documento*, que pode ser interpretada como a distribuição de tópicos em relação aos documentos [Churchill and Singh 2022].

Como último modelo para experimentação, comparação e análise em textos curtos do Twitter, lançamos mão de um modelo mais recente que utiliza métodos estado da arte para tarefas de pro-

[1] https://www.statista.com/statistics/303681/twitter-users-worldwide/

cessamento de linguagem natural. Esse é o BERTopic, uma variação de Representações de Codificador Bidirecional de Transformadores (BERT, do inglês Bidirectional Encoder Representations from Transformers) e proposto por Grootendorst [2022b]. BERTopic é um modelo de linguagem baseado em *transformer* pré-treinado que gera representações de documentos, agrupa-as, e, ao fim, gera representações de tópicos com o procedimento TF-IDF baseado em classe.

3. METODOLOGIA

Esta seção apresenta a metodologia empregada neste trabalho do ponto de vista de: 1) avaliação da quantidade dos dados; 2) coleta, pré-processamento dos dados e 3) inserção dos dados nos modelos para análise de sensibilidade aos parâmetros e análise de desempenho. As subseções estabelecem uma sequência e detalham cada uma das etapas.

3.1 Delimitação do Idioma dos Tweets

O primeiro passo da metodologia consistiu em avaliar de que maneira seria possível concentrar nossa análise no conteúdo produzido no contexto brasileiro, uma vez que esse enfoque era uma motivação central da pesquisa. Embora o Twitter possua um mecanismo de busca que inclui a localização apontada por quem posta uma mensagem, percebemos que havia um número significativo de mensagens marcadas como tendo sido postadas no Brasil que não eram relevantes para nós, como mensagens postadas por visitantes atraídos por eventos como a Copa do Mundo de 2014 ou a Olimpíada de 2016.

Ponderamos então que poderíamos utilizar como crivo o idioma português, uma vez que tínhamos a percepção de que o número de usuários brasileiros seria amplamente majoritário nesse recorte lusófono. Essa ponderação foi confirmada pelo ranking de países com maior quantidade de usuários no Twitter presente na plataforma Statista.[2] Também pela consideração de que o número de usuários brasileiros (19,05 milhões) é mais do que 13 vezes o número de usuários de Portugal (1,40 milhões),[3] terceiro país com maior quantidade de pessoas lusófonas no mundo.[4] Angola como segundo país com mais falantes de português, possui apenas 71,4 mil usuários no Twitter em 2022.[5]

3.2 Coleta de Dados do Twitter

Nas coletas, o tamanho de amostra é uma questão fundamental: é necessário investigar a quantidade de *tweets* que podem ser representativos de determinado assunto que se quer estudar. Entretanto, a distribuição estatística dos dados deste trabalho era-nos desconhecida *a priori*. Não havia, assim, um modelo que determinasse um tamanho de amostra que pudesse garantir uma população para uma definida confiança e margem de erro. Há, portanto, uma dificuldade em se definir a quantidade de *tweets* para que se possa trabalhar.

Uma estratégia apresentada por Krippendorff [2018] é realizar experimentos de amostragem para descobrir um tamanho do conjunto. A primeira coleta adotada neste trabalho consistiu em obter três amostras de 50 mil tweets no ano de 2021 e realizar comparações entre elas para analisar a adequação do tamanho. As três amostras coletadas foram compostas pela palavra-chave "comunismo", isto é, continham "comunismo" e/ou "#comunismo". A opção por essa palavra se dá por sua centralidade no discurso da extrema-direita, que ocupou grande espaço no debate político brasileiro nos últimos dez anos [Machado and Colevati 2021] e também desempenha função de ser um exemplo de aplicação da metodologia desenvolvida neste trabalho em um assunto social, político e econômico.

[2]https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/
[3]https://datareportal.com/reports/digital-2022-portugal#: :text=Numbers%20published%20in%20Twitter's%20advertising,total%20population%20at%20the%20time.
[4]https://www.worlddata.info/languages/portuguese.php
[5]https://datareportal.com/reports/digital-2022-angola

Para realizar a coleta, utilizamos a API do Twitter.[6] Como ela retorna no máximo 500 *tweets* por resposta, a cada conjunto com esse máximo retornado, eram fornecidos à API uma data e um horário pseudoaleatórios no ano de 2021. Todavia, nem sempre o máximo de 500 foi retornado por haver *tweets* indisponíveis. Outra questão é que o Twitter possui diferentes tipos de *tweets*[7], eles são: *tweets* gerais, *tweets* de status, *retweets*, *replies*, menções e comentários. A coleta abrangeu todos esses tipos.

3.3 Pré-Processamento

Inspecionadas as distribuições das amostras, realizamos nos textos dos *tweets* um pré-processamento de acordo com o indicado em Allahyari et al. [2017]. Ele consistiu nas seguintes etapas: 1) tratar os caracteres e outros processos; 2) quebrar o texto em unidades individuais (o que é conhecido como *tokenizing* em inglês); 3) remoção de palavras de parada (conhecidas como *stopwords* em inglês); 4) lematização; por fim, 5) outra remoção de palavras de parada geradas pela lematização.

No tratamento dos caracteres, houve a remoção dos itens não-alfanuméricos, ou seja, aqueles que não são letras latinas e algarismos indo-arábicos. Retiramos usuários mencionados no texto e URLs. Por meio de uma lista manual, substituímos as abreviações por suas palavras completas correspondentes. Após essa substituição, juntamente com a quebra dos textos em unidades individuais (*tokenizing*), todas as letras foram transformadas em minúsculas e removemos as pontuações.

Palavras de parada são aquelas muito frequentes em um texto e que não possuem um valor semântico muito grande, como preposições e conjunções. A remoção delas foi feita por meio da lista da biblioteca NLTK [Bird et al. 2009] e de outras palavras adicionadas manualmente a essa lista.

A lematização consiste em transformar palavras em sua forma mais simples, como colocá-las no singular, masculino ou, no caso de verbo, colocar no infinitivo, i.e., agrupar as várias formas flexionadas de uma palavra para que possam ser analisadas como um único item considerando a análise morfológica [Allahyari et al. 2017]. Esse processo foi realizado utilizando um modelo treinado de rede neural convolucional da biblioteca spaCy [Honnibal and Montani 2017], sendo a versão pequena desse modelo a usada por conta de seu menor custo computacional e de memória. Por fim, encerrando o pré-processamento, foi realizada mais uma remoção de palavras de parada porque mais delas foram geradas pela lematização.

3.4 Experimentos em Tamanho de Amostra com LDA

Foram realizados três experimentos com o modelo LDA para avaliar o tamanho de amostra. Eles estão na subseções a seguir, LDA nas Amostras de 50 mil tweets, LDA com Técnica de Divisão pela Metade e LDA com Aumento de Tamanho de Amostra. Uma razão por trabalhar com LDA, além de ser o modelo mais clássico para modelagem de tópicos, é por sua implementação ser a mais rápida em tempo de execução de código. Ela se chama *LdaMulticore* e pertence à biblioteca Gensim [Rehůřek and Sojka 2010]. Logo, todos os experimentos em tamanho de amostra com LDA foram realizados com a configuração de cinco tópicos, α e η simétricos de acordo com a equação 1.

3.4.1 *LDA nas Amostras de 50 Mil Tweets.* Analisados os *bag of words*, aplicamos o modelo LDA com cinco tópicos para cada uma das três amostras em busca de comparar seus resultados e analisar a representatividade do tamanho de amostra. Entretanto, a comparação entre tópicos gerados por diferentes modelos de LDA é complexa devido aos seguintes fatores: 1) caráter estocástico do modelo, sendo necessário executar várias vezes cada modelo e fazer uma comparação entre médias; 2) caráter não supervisionado do modelo, não existindo, portanto, tópicos gerados esperados e sendo necessária uma métrica adequada para análise do desempenho; 3) necessidade de escolher seus hiperparâmetros

[6] https://developer.twitter.com/en/docs/twitter-api
[7] https://help.twitter.com/en/using-twitter/types-of-tweets

para se obter tópicos gerados ótimos; 4) necessidade de um especialista no tema dos tweets para comparar melhor os tópicos gerados pelos modelos.

Tendo em vista essas complexidades, é necessária uma métrica adequada para análise de desempenho do modelo. A utilizada foi a métrica coerência de tópicos C_V, indicada como a versão com correlação mais forte com avaliações humanas [Röder et al. 2015]. Sendo essa métrica utilizada, primeiramente, para avaliar o tamanho de amostra e, depois, também para o desempenho dos outros modelos (BTM, BERTopic e NMF) que foram utilizados para assim conseguir ser possível realizar uma comparação entre todos os modelos.

Para realizar a comparação entre esses modelos, utilizamos a média de coerência de tópicos C_V de 10 execuções para cada amostra como métrica. Resultando, portanto, em 30 execuções ao todo. Também utilizamos a distância de Jaccard entre os tópicos gerados pelos diferentes modelos em cada execução e calculamos a média de distância para avaliar o tamanho de amostra. Essa métrica foi utilizada somente nas análises de tamanho de amostra.

3.4.2 *Técnica de Divisão pela Metade.* Essa é uma técnica proposta por Krippendorff [2018], também chamada de *Split-Half Technique*, em que dividem-se as amostras ao meio para analisar se elas apresentam o mesmo comportamento em relação ao conjunto inteiro. Realizamos essa técnica nas três amostras.

Após a divisão foi feito o cálculo da frequência das palavras no *bag of words*. Nessa etapa, as metades apresentaram classificações em ordem de palavras mais frequentes bem semelhantes. Indicando, então, um indício de que o comportamento se manteve ao se dividir a amostra pela metade. Analisadas as frequências de palavras, executamos 10 vezes o modelo LDA com 5 tópicos em cada metade. Realizamos novamente o cálculo de coerências C_V e da distância de Jaccard entre os tópicos gerados pelas metades.

3.4.3 *Experimento de Aumento de Tamanho de Amostra.* Como o comportamento das amostras mudaram ao dividir pela metade, realizamos um experimento de amostragem também proposto por Krippendorff [2018], que é realizar outra coleta com tamanho diferente de amostra para comparação.

Fizemos 3 coletas de 100 mil *tweets*, o dobro do tamanho de amostra inicial. O intuito foi avaliar as médias de coerência C_V e distância de Jaccard entre os tópicos para analisar se esse tamanho de amostra altera o comportamento dessas métricas em relação ao tamanho de amostra inicial. Também foram realizadas 10 execuções para cada amostra.

3.5 Escolha dos Hiperparâmetros do LDA

Terminados os experimentos de tamanho de amostra, decidimos trabalhar com 50 mil *tweets* por ter obtido uma coerência C_V maior e ter um custo computacional menor do que 100 mil *tweets*. Como o modelo LDA e seus métodos de otimização apresentam vários hiperparâmetros, é necessário realizar a afinação, isto é, buscas para encontrar aqueles que apresentam o melhor desempenho.

Antes dessa afinação é necessário definir quais hiperparâmetros são mais relevantes para o desempenho, visto que uma busca de todos se torna inviável computacionalmente. Logo, realizamos a afinação de hiperparâmetros escolhidos e trabalhados na literatura [Panichella 2021]. Esses são: número de tópicos, α e η.

Primeiro realizamos uma busca grossa, ou seja, com passos grandes de valores para ter uma amplitude maior do espaço de busca. Depois realizamos uma busca fina, isto é, com passos pequenos de números de tópicos enquanto os valores de α e η estavam fixos. Baseado em Panichella [2021], os valores de hiperparâmetros na busca grossa foram:

—número de tópicos (k): 2, 42, 82, 122, 162

—α: 'symmetric', 'asymmetric', 0.01, 0.1, 1, 10

—η: 'symmetric', 'auto', 0.1, 1, 10

Sendo 'symmetric' equivalente ao inverso de k, 'asymmetric' equivalente ao inverso da soma do índice do tópico com a raiz quadrada de k e 'auto' há um aprendizado assimétrico do corpus. Essas equivalências são apresentadas nas equações 1 e 2.

$$symmetric = \frac{1}{k}; \tag{1}$$

$$asymmetric(t) = \frac{1}{t + \sqrt{k}}; \tag{2}$$

Em que t é o índice do tópico, assumindo valores de 0 a $k-1$. Finalizada a busca grossa, os valores de hiperparâmetros na busca fina foram:

—número de tópicos: [2, 4, 6, ..., 158, 160, 162];

—α e η com maiores valores encontrados de coerência C_V na busca grossa como também combinações propostas na literatura.

3.6 Aplicação do BTM, NMF e BERTopic nas amostras e escolha de seus hiperparâmetros

Após os experimentos com o LDA, realizamos a afinação do modelo BTM para as 3 amostras. A justificativa de escolha desse modelo foi por ter obtido melhor desempenho em Jónsson [2016]. Nesse processo foi necessário também definir quais hiperparâmetros buscar. Para essa questão utilizamos como referência Jónsson [2016] em que os hiperparâmetros buscados foram: número de tópicos, α e β (equivalente a η). O pacote BTM utilizado foi o bitermplus[8], que implementa Yan et al. [2013] em Cython. As combinações de valores testadas também foram como Jónsson [2016], i.e., com as seguintes configurações:

—número de tópicos (k): [10 50 100 200];

—α: [1/k 50/k 100/k];

—β: [0,001 0,01 0,5]

Com o BERTopic, aplicamos a modelagem de tópicos para as amostras. Em busca de um número de tópicos com maior coerência de tópico C_V, realizamos a redução desse número utilizando a implementação da função *reduce_topics*, que realiza a mesclagem recursa de pares de tópicos de dados de amostra. Foi utilizada a implementação presente no repositório Grootendorst [2022a] para aplicar o modelo BERTopic em sua configuração padrão, modelo de *embedding* multilíngua *paraphrase-multilingual-MiniLM-L12-v2*.

Depois dos experimentos com o BERTopic, realizamos a escolha de hiperparâmetros do último modelo, o NMF. A escolha consistiu em buscar o número de tópicos que obtivessem a maior coerência de tópicos C_V para cada uma das três amostras. A implementação do NMF utilizada foi a *nmf*, presente na biblioteca Gensim [Rehůřek and Sojka 2010].

[8]https://bitermplus.readthedocs.io/en/stable/index.html

Table I. Médias totais de coerência C_V e distância de Jaccard para os três tamanhos diferentes de amostra.

Métrica	25 mil tweets	50 mil tweets	100 mil tweets
Coerência C_V	0,275	0,299	0,235
Distância de Jaccard	0,837	0,884	0,842
C_V/d_J	0,329	0,338	0,279

Table II. Valores máximos de coerência de tópico C_V para cada amostra em cada modelo.

Amostra	LDA	BTM	BERTopic	NMF
1	0,594	0,624	0,780	0,651
2	0,540	0,592	0,783	0,641
3	0,561	0,610	0,772	0,662
Média	0,565	0,609	0,778	0,651

4. RESULTADOS

Na tabela I, observa-se um resumo sobre os experimentos com distintos tamanhos de amostra, ou seja, com variação da quantidade de *tweets* coletados. Optamos por trabalhar com 50 mil tweets nas amostras tendo em vista um compromisso entre melhor desempenho na razão C_V/d_J e um médio custo computacional, em comparação com as quantidades de 100 mil e 25 mil tweets.

Na tabela II, é apresentado um resumo dos desempenhos alcançados nos experimentos de escolha dos hiperparâmetros de cada modelo. De acordo com os experimentos, o BERTopic obteve melhor desempenho, seguido do NMF, em terceiro o BTM e por último o LDA. Isso demonstra que esse modelo neural de modelagem de tópicos apresenta vantagem em relação aos modelos mais antigos, LDA e BTM, que se baseiam mais estritamente em ferramentas probabilísticas e NMF, que se baseia em técnicas de álgebra linear.

O BERTopic poderia ter usado os dados sem realizar pré-processamento por se tratar de um modelo que consegue considerar o contexto de um texto. Ele tem a capacidade de ponderar a ordem das palavras, as palavras de parada (*stopwords*) e pontuações. Provavelmente ele obtivesse um desempenho ainda maior em relação aos outros modelos. Entretanto, optamos em manter a mesma metodologia para todos os modelos para realizar a comparação.

5. CONCLUSÃO

Neste trabalho, apresentamos uma metodologia para realizar análise de mídias sociais com base em modelagem de tópicos. Primeiramente, definimos uma forma de especificar os dados para o Brasil tendo por base o idioma. Em segundo lugar, apresentamos um procedimento para a coleta e o pré-processamento. Em terceiro lugar, definiu-se um método de análise de tamanho de amostra para definir a quantidade de coletas que pode ser representativa para um estudo. E por fim, foram indicados elementos para melhorar o desempenho dos modelos por meio de seus hiperparâmetros.

Por meio dos resultados, pudemos realizar uma análise comparativa e conseguimos observar que há um melhor desempenho por parte modelo mais recente, que é caracterizado como uma estratégia neural de aprendizado profundo na categoria dos *transformers* pré-treinados, o BERTopic. Também se verificou a vantagem de modelos que não são baseados em *bag of words* [Shadrova 2021], como é o caso do BERTopic, que cria representações de textos no espaço vetorial multi-dimensional levando em consideração todo o contexto e por meio disso realiza o agrupamento dos textos em tópicos.

Possíveis perspectivas e trabalhos futuros estão na análise os tópicos produzidos pelos modelos juntamente com um especialista sobre o tema, principalmente aqueles gerados pelo BERTopic, que obteve maior desempenho. Além disso, o BERTopic tem um potencial, visto os resultados e a discussão de Shadrova [2021]. Seria interessante usá-lo para analisar outros temas sociais utilizando a metodologia desse trabalho.

REFERÊNCIAS

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., and Kochut, K. A brief survey of text mining: Classification, clustering and extraction techniques, 2017.

Angelov, D. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*, 2020.

Bird, S., Klein, E., and Loper, E. *Natural language processing with Python*. O'Reilly Media, 2009.

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

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

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41 (6): 391–407, 1990.

Gallagher, R. J., Reing, K., Kale, D., and Ver Steeg, G. Anchored correlation explanation: Topic modeling with minimal domain knowledge. *Transactions of the Association for Computational Linguistics* vol. 5, pp. 529–542, 2017.

Grootendorst, M. Bertopic. `https://github.com/MaartenGr/BERTopic`, 2022a.

Grootendorst, M. Bertopic: Neural topic modeling with a class-based tf-idf procedure, 2022b.

Halbwachs, M. La mémoire collective [la memoria colectiva]. *Paris, Francia: Presses Universitaires de France*, 1950.

Hofmann, T. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '99. Association for Computing Machinery, New York, NY, USA, pp. 50–57, 1999.

Honnibal, M. and Montani, I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, 2017. To appear.

Jónsson, E. An evaluation of topic modelling techniques for twitter, 2016.

Krippendorff, K. *Content analysis*. SAGE Publications, Thousand Oaks, CA, 2018.

Machado, M. G. and Colevati, J. Anticomunismo e Gramscismo Cultural no Brasil. *Revista Aurora* 14 (Edição Especial): 23–34, July, 2021. Number: Edição Especial.

Nisha and Kumar R, D. A. Implementation on text classification using bag of words model. *SSRN Electron. J.*, 2019.

Paatero, P. and Tapper, U. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* 5 (2): 111–126, 1994.

Panichella, A. A systematic comparison of search-based approaches for LDA hyperparameter tuning. *Information and Software Technology* vol. 130, pp. 106411, 2021.

Rehůřek, R. and Sojka, P. Software framework for topic modelling with large corpora. pp. 45–50, 2010.

Roberts, M. E., Stewart, B. M., Tingley, D., Airoldi, E. M., et al. The structural topic model and applied social science. In *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation*. Vol. 4. Harrahs and Harveys, Lake Tahoe, pp. 1–20, 2013.

Robila, M. and Robila, S. A. Applications of artificial intelligence methodologies to behavioral and social sciences. *Journal of Child and Family Studies* 29 (10): 2954–2966, Oct., 2020.

Röder, M., Both, A., and Hinneburg, A. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. WSDM '15. Association for Computing Machinery, New York, NY, USA, pp. 399–408, 2015.

Shadrova, A. Topic models do not model topics: epistemological remarks and steps towards best practices. *Journal of Data Mining & Digital Humanities* vol. 2021, 2021.

Sridhar, V. K. R. Unsupervised topic modeling for short texts using distributed representations of words. In *Proceedings of the 1st workshop on vector space modeling for natural language processing*. pp. 192–200, 2015.

Sumikawa, Y., Jatowt, A., and Düring, M. Digital history meets microblogging: Analyzing collective memories in twitter. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*. JCDL '18. Association for Computing Machinery, New York, NY, USA, pp. 213–222, 2018.

Teh, Y., Jordan, M., Beal, M., and Blei, D. Sharing clusters among related groups: Hierarchical dirichlet processes. *Advances in neural information processing systems* vol. 17, 2004.

Wang, Y.-X. and Zhang, Y.-J. Nonnegative matrix factorization: A comprehensive review. *IEEE Transactions on Knowledge and Data Engineering* 25 (6): 1336–1353, 2013.

Yan, X., Guo, J., Lan, Y., and Cheng, X. A biterm topic model for short texts. WWW '13. Association for Computing Machinery, New York, NY, USA, pp. 1445–1456, 2013.