Modelagem Probabilística de Tópicos: Uma Comparação Empírica

Leonardo H. Rocha[1], **Daniel Welter**[1], **Denio Duarte**[1]

[1]Universidade Federal da Fronteira Sul (UFFS)
Campus Chapecó
Chapecó – SC – Brazil

leoheiro@hotmail.com, danielluiswelter@gmail.com, duarte@uffs.edu.br

Abstract. *Extracting the main topics that represent the subjects covered by a text collection can give valuable insights. To this end, approaches for topic modeling have been used to tackle such problems as information discovery and topic extraction with thematic information. In this context, this work presents a comparasion of four topic modeling approaches: Hierarchical Dirichlet Process (HDP), Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF), and Dirichlet-multinomial Regression (DMR). As input collection, we use news from five websites: Breibart, Business Insider, The Atlantic, CNN e New York Times. The collection contains 50,000 articles. We evaluate the discovered topics using a coherence metrics called C_v. The results have shown that DMR and LDA are the best models to extract topics from the proposed document collection.*

Resumo. *Abordagens probabilísticas de tópicos são ferramentas para descobrir e explorar estruturas temáticas escondidas em coleções de textos. Dada uma coleção de documentos, a tarefa de extrair os tópicos consiste em criar um vocabulário a partir da coleção, verificar a probabilidade de cada palavra pertencer a um documento da coleção. Em seguida, baseado no número de tópicos desejado, a probabilidade de cada palavra estar associada a um determinado tópico é contabilizada. Assim, um tópico é um conjunto de palavras ordenadas pela probabilidade de estar associada ao tópico. Várias abordagens são encontradas na literatura para criação de modelos de tópicos, e.g., Hierarchical Dirichlet Process (HDP), Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF) e Dirichlet-multinomial Regression (DMR). Este trabalho procura identificar a qualidade dos tópicos construídos pelas quatro abordagens citadas. A Qualidade será medida por métricas de coerência e todas as abordagens terão a mesma coleção de documentos como entrada: notícias de websites dos jornais Breibart, Business Insider, The Atlantic, CNN e New York Times contendo 50.000 artigos. Os resultados mostram que DMR e LDA são os melhores modelos para extrair tópicos da coleção utilizada.*

1. Introdução

Modelagem de tópicos vem sendo utilizada ha décadas para extrair estruturas latentes de coleção de documentos [Blei 2012]. Dada uma coleção de documentos de entrada,

um vocabulário é gerado e as palavras desse vocabulário são contadas nos documentos para identificar a probabilidade de cada uma estar associada ao documento. Após a contagem, as palavras são associadas ao número de tópicos configurado (K), baseado nas probabilidades de aparecerem nos documentos. No fim do processo, todas as palavras do vocabulário (também chamado de dicionário) estarão associadas aos K tópicos definidos. Geralmente, escolhe-se as top-10 palavras para caracterizar um tópico, ou seja, as 10 palavras com maior chance de estar associadas ao tópico. Os documentos, por sua vez, também serão associados aos tópicos por probabilidade. Um documento pode ter mais de um tópico associado, porém a soma das probabilidades de tais tópicos deverá ser igual a um [Steyvers and Griffiths 2007].

Existem várias abordagens que utilizam diferentes estratégias para extrair tópicos das coleções de documentos. Dentre as abordagens, este trabalho selecionou as seguintes para realizar uma comparação empírica a partir dos tópicos gerados: *Latent Dirichlet Allocation* (LDA), *A Hierarchical Dirichlet Process* (HDP), *Dirichlet-Multinomial Regression* (DMR)e *Non-negative Matrix Factorization* (NMF). Para todas as abordagens, a coleção de entrada é formada por notícias e está disponível no sítio da *Kaggle*[1]. Esta coleção é composta por 50.000 notícias dos veículos: *Breibart*, *Business Insider*, *The Atlantic*, *CNN* e *New York Times*.

Os tópicos gerados pelas abordagens selecionadas são comparados utilizando métricas de coerência propostas em [Röder et al. 2015]. Das métricas propostas, a métrica C_v foi a escolhida por ter maior relação com o julgamento humano. Assim, a contribuição deste trabalho é apresentar uma comparação das abordagens selecionadas, classificando pela qualidade dos tópicos baseado na métrica C_v. Os experimentos apontam que DMR teve o melhor desempenho entre os outros modelos gerados. Porém, LDA se mostrou bastante similar ao resultado do DMR. Por outro lado, NMF obteve o pior desempenho.

O restante deste trabalho está organizado da seguinte forma. A próxima seção apresenta alguns conceitos importantes para este trabalho. A Seção 3 apresenta alguns trabalhos similares ao apresentado aqui. Em seguida, são apresentados os resultados da comparação realizada entre as quatro abordagens. Finalmente, a Seção 5 apresenta as conclusões deste trabalho, bem como algumas direções futuras.

2. Fundamentos Teóricos

Esta seção apresenta brevemente alguns conceitos que auxiliam o entendimento deste trabalho. Inicialmente, é apresentado o conceito de modelagem de tópicos e, em seguida, as abordagens consideradas neste trabalho e as métricas de avaliação.

2.1. Modelagem de Tópicos

Modelagem de tópicos se refere a um conjunto de algoritmos cujo objetivo é extrair, dada uma coleção de documentos, os principais tópicos que representam os assuntos cobertos pela coleção [Blei 2012, Steyvers and Griffiths 2007]. Em aprendizado de máquina, modelagem de tópicos pertencem à classe de algoritmos não supervisionados em que os dados de entrada não possuem rótulos para categorizar cada exemplo [Duarte and Ståhl 2019]. O fato de ser não supervisionado traz alguns desafios para

[1]https://www.kaggle.com/snapcrack/all-the-news

Tópico 247

word	prob.
drugs	.069
drug	.060
medicine	.027
effects	.026
body	.023
medicines	.019
pain	.016
person	.016
marijuana	.014
label	.012
alcohol	.012
dangerous	.011
abuse	.009
effect	.009
known	.008
pills	.008

Tópico 5

word	prob.
red	.202
blue	.099
green	.096
yellow	.073
white	.048
color	.048
bright	.030
colors	.029
orange	.027
brown	.027
pink	.017
look	.017
black	.016
purple	.015
cross	.011
colored	.009

Tópico 43

word	prob.
mind	.081
thought	.066
remember	.064
memory	.037
thinking	.030
professor	.028
felt	.025
remembered	.022
thoughts	.020
forgotten	.020
moment	.020
think	.019
thing	.016
wonder	.014
forget	.012
recall	.012

Tópico 56

word	prob.
doctor	.074
dr.	.063
patient	.061
hospital	.049
care	.046
medical	.042
nurse	.031
patients	.029
doctors	.028
health	.025
medicine	.017
nursing	.017
dental	.015
nurses	.013
physician	.012
hospitals	.011

Tabela 1. Uma ilustração de quatro (de 300) tópicos extraídos da coleção de documentos da TASA [Steyvers and Griffiths 2007].

a criação e avaliação dos modelos construídos: o melhor número de tópicos para uma determinada coleção, avaliar a qualidade dos tópicos e a melhor métrica para fazer essa avaliação [Chang et al. 2009, Lau et al. 2014, Röder et al. 2015].

2.2. Tópicos

Tópicos são derivados da distribuição probabilística das palavras nos documentos da coleção de entrada. O conjunto de palavras que por relação de ordem, frequência e semântica representam certos assuntos (temas). Assim, por meio desses relacionamentos, é possível definir um tema como um tópico, ou seja, a distribuição probabilística das palavras com a frequência e semântica que faz sentido no contexto do tópico.

A Tabela 1 apresenta um exemplo com quatro de trezentos tópicos construídos a partir da coleção de documentos extraída do corpus Touchstone Applied Science Associates (TASA), composta por de mais de 37 mil documentos sobre materiais educacionais (*e.g.*, linguagem e artes, estudos sociais, saúde, ciências) [Steyvers and Griffiths 2007]. Essa tabela mostra as dezesseis palavras com maior probabilidade em cada tópico. Perceba que como não existem rótulos, um especialista no assunto deve definir a semântica de cada tópico. No caso do exemplo da Tabela 1, pode-se inferir que os tópicos estão relacionadas ao uso de drogas (247), cores (5), mente e memória (43), e consultas médicas (56).

2.3. Documentos

A modelagem de tópicos é baseada na ideia que documentos são misturas de tópicos, ou seja, documentos estão associados a múltiplos tópicos [Steyvers and Griffiths 2007, Blei 2012]. Assim, documentos podem ser gerados a partir de diferentes distribuições de tópicos. Um documento pode ser definido como uma sequência de palavras **w**=(w_1, w_2,

$\ldots, w_n)$, em que n é o número de palavras em **w**. Similarmente, um *corpus* (ou coleção) é um conjunto de m documentos D=$\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m\}$. Além disso, um documento pode possuir qualquer conteúdo, *e.g.*, uma notícia ou um comentário em uma rede social.

Em modelagem de tópicos, muitas abordagens consideram o documento como um saco de palavras, isto é, a ordem das palavras no documento não importa. Também, um pré-processamento deve ser executado na coleção para prepará-la para a extração dos tópicos. A fase de pré-processamento pode ser composta pelos seguintes passos [Steyvers and Griffiths 2007]: (i) remoção das *stop-words*, *i.e.*, remoção das palavras espúrias (*e.g.*, artigos, pontuação e pronomes pessoais) da coleção, (ii) tokenização, *i.e.*, transformar a coleção em um lista de palavras, (iii) stemização, *i.e.*, reduzir as palavras para as suas formas raízes (*e.g.*, *studying*, *studies* e *study* se tornam *stud*), e (iv) lematização, *i.e.*, agrupa as palavras nas suas formas inflexionadas (*e.g.*, *studying*, *studies* e *study* se tornam *study*). É preferível utilizar a lematização ao invés da stemização, pois a segunda pode gerar palavras não existentes na língua da coleção. Porém, a lematização é mais demorada em processamento computacional, pois é necessário consultar um dicionário com as flexões.

2.4. Abordagens para Modelagem de Tópicos

As abordagens de criação de modelos de tópicos diferem, principalmente, pela função de distribuição de probabilidade. *Latent Dirichlet Allocation* (LDA) utiliza a distribuição de Dirichlet para iniciar o processo de alocação de probabilidade às palavras em relação aos documentos e tópicos [Blei et al. 2003]. *A Hierarchical Dirichlet Process* (HDP) é baseado na LDA porém o número de tópicos é encontrado automaticamente, dito não paramétrico [Teh et al. 2006]. Utilizando também a distribuição de Dirichlet, *Dirichlet-Multinomial Regression* (DMR) utiliza os atributos extraídos dos documentos (*e.g*, TF-IDF) durante o processo de distribuição das probabilidades [Mimno and McCallum 2008]. *Non-negative Matrix Factorization* (NMF) é uma família de algoritmos baseada na álgebra linear para identificar estrutura latentes nos dados representados como matrizes cujos valores são não negativos.

2.5. Métricas

Métricas são utilizadas para medir e avaliar modelos em aprendizado de máquina. Na modelagem de tópicos, como em qualquer modelo não supervisionado, avaliar os modelos é um desafio porque os conjuntos de dados não possuem rótulos para verificar a consistência dos resultados automaticamente. Avaliações podem ser feitas por humanos, porém, é uma tarefa complexa e onerosa [Röder et al. 2015]. Nesse mesmo trabalho, os autores apresentam um estudo comparando várias métricas de coerência para modelagem de tópicos. O estudo foi realizado para identificar qual métrica é a mais próxima da avaliação humana. Como resultado, a métrica C_v foi a mais próxima da percepção humana. C_v é baseada na versão normalizada da *PMI* (Pointwise Mutual Information) (veja Equação 1) a qual o resultado é fixado na faixa de $[-1, 1]$, em que 1 indica completa coocorrência entre as palavras (w_i and w_j) e -1 sem coocorrência. Também, uma janela deslizante é utilizada para calcular a coocorrência entre as palavras. Por exemplo, dado o conjunto de palavras $C = \{w_1, w_2, w_3, w_4, w_5, w_6\}$, uma janela deslizante de tamanho três poderia conter a janela $W_1 = \{w_2, w_3, w_4\}$ e deslizar para $W_2 = \{w_3, w_4, w_5\}$.

$$PMI(w_i, w_j) = \log\left(\frac{P(w_i, w_j) + \epsilon}{P(w_i) \cdot P(w_j)}\right) \qquad NPMI(w_i, w_j) = \frac{PMI(w_i, w_j)}{-\log\left(P(w_i, w_j)\right)} \tag{1}$$

Devido ao excelente desempenho de C_v em relação à percepção humana, ela será utilizada neste trabalho para encontrar os melhores hiperparâmetros (*e.g.* número de tópicos ideal) e para identificar a qualidade dos tópicos gerados pelas abordagens consideradas.

3. Trabalhos Relacionados

Após investigar trabalhos correlatos em modelagem de tópicos foram encontrados diversos artigos, porém, são apresentados aqueles mais relevantes e recentes para a comparação proposta. Por exemplo, a comparação conduzida por [Agade and Balpande 2020] utiliza duas abordagens, LDA e NMF, em que usam um conjunto de notícias sobre Covid-19 e avaliam os impactos não relacionados a saúde. As métricas utilizadas para avaliar os modelos são C_v e Perplexidade. Os resultados do trabalho apontam que para uma melhor coerência entre os tópicos, LDA foi a melhor abordagem. O trabalho de [Mifrah and Benlahmar 2020] também tem como tema Covid-19 e os modelos LDA e NMF, e se apoiam apenas na métrica C_v. Os autores concluem que para alguns tópicos o NMF teve melhores resultados, porém, avaliando os tópicos globalmente, o LDA gera tópicos mais coerentes e concisos.

[Williams and Betak 2018] realizam uma comparação entre LDA e LSA usando como tema uma base de dados sobre a causa de acidentes ferroviários. Neste trabalho não é usado uma métrica para avaliar os resultados. Foram gerados os tópicos de cada modelo, os mesmos foram avaliados visualmente e chegaram a conclusão que para esses modelos e para o tema escolhido, os modelos se complementam. A abordagem usada por [Chehal et al. 2021] utiliza os modelos LSA, PLSA, LDA e CTM, e uma base de dados de avaliações de usuários de celulares Moto E5 do site Amazon.com. A ideia dos autores é de gerar um sistema de recomendação personalizadas em sites de e-commerce por meio de modelagem de tópicos. Usando a métrica C_v novamente o modelo gerado pelo LDA teve o melhor desempenho.

Após uma análise sobre os 4 trabalhos apresentados, percebe-se que LDA é a abordagem com melhor desempenho nas análises propostas. A comparação aqui apresentada se apoia em algumas das abordagens utilizadas pelos trabalhos relacionados (*i.e.*, LDA e NMF) e acrescenta HDP e DMR.

4. Experimentos

Para realização dos experimentos foi coletado um conjunto dados do sítio *Kaggle* com 50.000 notícias. Foram usadas notícias de janeiro de 2016 a junho de 2017, dado que esse período somava mais do que 92% do conteúdo da coleção. As notícias são das seguintes fontes: Breibart (23.781), Business Insider (6.757), The Atlantic (171), CNN (11.488) e New York Times (7.803). O conjunto possui temas como: política, policial, saúde, esporte, entretenimento, estilo de vida e acontecimentos mundiais. Os experimentos foram realizados utilizando a linguagem *Python* e as bibliotecas *tomotopy*, *sklearn*, *gensim* e *spacy*.

4.1. Projeto

A etapa de pré-processamento foi dividida em três passos. No primeiro passo, foram usados métodos da biblioteca do *gensim* para retirar caracteres especiais e palavras com menos de três caracteres. No segundo passo, foi usado recursos da biblioteca *spacy* para lematizar e remover *stopwords* do conjunto de notícias, mantendo palavras cujo *part of speech* (POS) se classifica em NOUN e PROPN (substantivos e nome próprios). Ainda na segunda etapa, foram gerados bigramas usando os recursos da classe *Phrases* do *gensim*. Palavras que aparecem juntas com uma determinada frequência são unidas por _, por exemplo, "Super Bowl" resulta em "super_bowl". Notícias que resultaram em menos do que 10 palavras após o pré-processamento foram excluídas do conjunto, dado que muitas dessas notícias resultaram em palavras sem semântica. Com o conjunto pré-processado, foi gerado um dicionário mantendo 10.000 palavras, retirando palavras muito frequentes que apareciam em mais do que 80% dos documentos e também removendo palavras pouco frequentes que ocorriam em menos do que 20 documentos. Por fim, a coleção resultante passou pelo processo de tokenização. O resultado do pré-processamento pode ser visto na Tabela 2. Percebe-se que o documento pós-processado tem um tamanho menor, sem os verbos e com alguns bigramas (*e.g.*, hit_song). Os bigramas não influenciam no resultado da C_v, pois são propagados para a coleção utilizada na avaliação.

Original
The reaction on social media to the news that Beyoncé is pregnant with twins was by turns swift, sweet and strange. While Beyonce's Super Bowl performance of the hit song "Formation" was seen as a pointed political statement, addressing issues like black pride and police brutality, it and the video drew anger from some in law enforcement who said they were an attack on their profession. Then there were others like the police officer in Virginia who recreated her formidable choreography to inspire high school students. A phone call to the Atlanta department seeking comment was not immediately returned on Wednesday.
Pre-processado
reaction medium news beyonce twin beyonce super_bowl performance hit_song formation statement issue pride police_brutality video anger law_enforcement attack profession police_officer virginia choreography school_student phone atlanta department comment wednesday

Tabela 2. comparação de pré-processamento do texto original e pré-processado.

A mesma coleção de notícias pré-processadas foi usada para executar todos os testes. As implementações LDA, HDP e DMR são da biblioteca *tomotopy*, já NMF é da biblioteca *gensim*. Para escolher os melhores hiperparâmetros dos modelos testados, foram geradas combinações segundo a documentação de cada modelo. Testes anteriores usando NMF e LDA resultaram em 40 tópicos como tendo o melhor resultado conforme a métrica C_v. Com isso, foi mantido número de tópicos fixo em 40 para todos os modelos. Para os demais hiperparâmetros foi mantida a combinação que obteve o maior valor na métrica C_v. A Tabela 3 apresenta as combinações utilizadas para cada abordagem e, em negrito, o valor selecionado. NMF é apresentada separadamente na tabela, pois possui um conjunto de hiperâmetros diferentes das demais. As combinações foram escolhidas empiricamente e para as abordagens que não possuíam determinado hiperparâmetro, a célula da tabela foi preenchida com –.

4.2. Resultados

As Tabelas 4 a 6 apresentam as dez palavras mais frequentes dos cinco melhores tópicos dos modelos LDA, DMR, NMF e HDP, respectivamente. Os valores da métrica C_v

Modelo	Term Weight	Alpha	Gamma	ETA	Sigma	Epsilon	Iters/ Passes
LDA	PMI **IDF** ONE	1/40 0.1 **0.01** 0.05	–	**0.1** 0.05 0.01 0.08	–	–	300 400 500 **800**
DMR	PMI **IDF** ONE	1/40 0.1 **0.01**	–	0.01 **0.1** 0.5	**0.85** 1 1.1	**1e-10**	300 400 500 **800**
HDP	PMI **IDF** ONE	**0.01** 0.05	0.5 0.7 **1**	**1** 0.5	–	–	**1500**

	chunk size	passes	kappa	min prob	max iter (w)	(w) stop condition	max iter (h)	(h) stop condition	eval every
NMF	3% **5%** 10%	500 **1000** 1500	**0.1** 0.5 1	**0.01** 0.1 0.5	100 200 **300**	0.001 **0.0001** 0.01	50 **100** 200	**0.001** 0.0001 0.01	5 **10** 15

Tabela 3. Tabela com os hiperparâmetros por abordagem.

são apresentados entre parênteses ao lado de cada tópico, lembrando que quanto mais próximo de um, mais coerente é o tópico. As execuções foram realizadas com a melhor combinação de hiperparâmetros obtida nos testes. Todos os tópicos podem ser encontrados em `bit.ly/3vip3ky`.

Analisando os melhores cinco tópicos, é notável a semelhança entre os resultados do LDA (Tabela 4) com o DMR (Tabela 5). Por exemplo, o tópico 18 do LDA e o 15 do DMR que podem ser rotulados como *futebol americano*, e o 18 do DMR e 15 do LDA que podem ser rotulados como *empresas de tecnologia*.

Topico 18 (0.732)	**Topico 1 (0.710)**	**Topico 39 (0.661)**	**Topico 15 (0.657)**	**Topico 35 (0.642)**
game	athlete	syria	company	china
player	sport	isis	apple	north_korea
team	player	force	google	russia
espn	game	iraq	facebook	putin
nfl	team	assad	user	tillerson
season	olympics	aleppo	uber	japan
kaepernick	match	war	app	south_korea
sport	tournament	military	technology	missile
league	coach	troop	internet	beijing
super_bowl	games	islamic_state	microsoft	united_states

Tabela 4. cinco melhores tópicos do modelo LDA de acordo com a métrica C_v.

Os tópicos com melhores resultados na métrica C_v tendem a mostrar uma melhor relação semântica entre as palavras do tópico, dado que é possível interpretar o assunto que está sendo abordado. Por exemplo, no tópico 19 do modelo NMF (Tabela 6) é possível perceber palavras relacionadas a corrida eleitoral americana em 2016. Assim como no tópico 28 do modelo HDP (Tabela 7) é possível perceber palavras relacionadas a desas-

Topico 15 (0.738)	Topico 7 (0.691)	Topico 18 (0.649)	Topico 1 (0.6291)	Topico 29 (0.628)
game	attack	company	migrant	car
player	isis	apple	europe	tesla
team	islamic_state	facebook	germany	company
season	muslims	google	france	uber
espn	islam	user	european_union	vehicle
sport	terrorist	app	britain	ford
nfl	terrorism	internet	refugee	model
league	group	technology	merkel	driver
kaepernick	mosque	microsoft	italy	technology
coach	syria	device	party	automaker

Tabela 5. cinco melhores tópicos do modelo DMR de acordo com a métrica C_v.

Topico 13 (0.657)	Topico 11 (0.649)	Topico 20 (0.643)	Topico 19 (0.621)	Topico 24 (0.620)
china	game	child	campaign	rubio
north_korea	team	family	candidate	february
beijing	player	school	voter	sanders
trade	season	parent	donald_trump	people
policy	point	mother	poll	south_carolina
united_states	sport	life	hillary_clinton	kasich
sea	fan	father	race	carson
japan	league	kid	vote	hillary
missile	week	son	supporter	nevada
south_korea	coach	daughter	mrs_clinton	america

Tabela 6. cinco melhores tópicos do modelo NMF de acordo com a métrica C_v.

Topico 31 (0.837)	Topico 38 (0.755)	Topico 26 (0.718)	Topico 39 (0.705)	Topico 28 (0.694)
cosby	art	animal	study	storm
christie	city	game	brain	water
train	work	dog	child	earthquake
constand	artist	pokemon	health	quake
new_jersey	building	rhino	patient	flooding
governor	museum	specie	researcher	resident
prosecutor	life	player	research	hurricane
trial	painting	dinosaur	diet	rain
bill_cosby	fashion	bird	woman	county
transit	design	gorilla	doctor	weather

Tabela 7. cinco melhores tópicos do modelo HDP de acordo com a métrica C_v.

tres naturais, podendo estar ligadas a onda de furacões que atingem a América do Norte todos os anos. Além disso, o tópico 39 do mesmo modelo mostra palavras sobre saúde e pesquisas na área. Tópicos relacionados a esportes também são notados nos modelos LDA, DMR e NMF, podendo-se observar palavras relacionadas ao campeonato anual de futebol americano, como "super_bowl" e "nfl" nos casos do tópico 18 do LDA e tópico 15 do DMR. Juntamente, os dois modelos também trazem tópicos que fazem menção a empresas de tecnologia muito conhecidas no mercado, como o tópico 15 do LDA e o tópico 18 do DMR.

Finalmente, a Tabela 8 apresenta o desempenho dos modelos considerando a média de métrica C_v utilizando todos os 40 tópicos. Apesar de o modelo HDP obter um melhor resultado para os top-5 tópicos (*i.e*, 0.84, 0.76, 0.72 e 0.70), no cálculo geral,

obteve a terceiro melhor desempenho. O modelo com o melhor desempenho geral foi o DMR, seguido pelo LDA. Os trabalhos relacionados apontavam o LDA como a melhor abordagem, o que não foi confirmado na comparação aqui realizada. Porém, a diferença entre DMR e LDA é pequena (0.003) e pode-se considerar como desempenho similar entre os dois modelos. Já o modelo NMF teve os piores desempenhos tanto no top-5 tópicos como no geral.

Modelo	**Total** C_v
DMR	0.702
LDA	0.699
HDP	0.678
NMF	0.584

Tabela 8. Desempenho geral da métrica C_v para cada modelo.

Além da métrica escolhida para calcular a qualidade do tópico, outra maneira de avaliar o tópico é de forma visual ou seja, utilizando a percepção humana, comparando se as palavras do tópico fazem sentido juntas (veja [Röder et al. 2015] para maiores detalhes). Com isso, é notável que o valor total resultante da métrica escolhida reflete na qualidade dos tópicos de cada modelo.

A qualidade dos tópicos também é resultado de outros fatores essenciais: as etapas de pré-processamento e a escolha de hiperparâmetros. Um bom pré-processamento vai garantir uma maior qualidade no resultado dos tópicos, facilitando a compreensão do conjunto de palavras. Além de que a escolha dos hiperparâmetros ideais ajuda a extrair bons resultados. Porém, as combinações de hiperparâmetros podem levar muito tempo para serem executadas até chegar no melhor resultado utilizando uma métrica para calcular a qualidade dos tópicos de cada combinação.

5. Conclusão

Este trabalho teve como objetivo comparar quatro abordagens de modelagem de tópicos utilizando a métrica C_v. Entre os resultados obtidos, o modelo DMR se mostrou mais eficiente gerando tópicos mais coerentes. Entretanto, o LDA também teve um bom desempenho gerando bons tópicos, com um resultado muito semelhante ao DMR. O modelo HDP, nesta pesquisa, foi forçado a gerar 40 tópicos deixando uma de suas características de lado que é encontrar o número ideal de tópicos automaticamente, tal fator pode ter prejudicado o desempenho do modelo. Entretanto, se mostrou eficiente na abordagem escolhida quando comparado com os outros modelos, com um bom resultado. Por outro lado, o modelo NMF quando comparado com os demais modelos não apresentou um bom resultado, mostrando não ser uma boa opção essa categoria de abordagem em específico.

Como trabalhos futuros, pode-se citar: rotular os tópicos para identificar o assunto recorrente extraído por cada modelo, incluir pessoas na avaliação dos tópicos gerados, utilizar outras métricas de coerência propostas por [Röder et al. 2015] e avaliar a escalabilidade das abordagens aumentando a quantidade de documentos nas coleções.

Agradecimentos

Leonardo H. Rocha é parcialmente financiado pela Universidade Federal da Fronteira Sul. Projeto PES-2020-0078.

Referências

Agade, A. and Balpande, S. (2020). Exploring the non-medical impacts of covid-19 using natural language processing. In *Preprints 2020110056*. Preprints Plataform.

Blei, D. M. (2012). Probabilistic topic models. *Commun. ACM*, 55(4):77–84.

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

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., and Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Proceedings of the Twenty-third Advances in neural information processing systems*, pages 288–296.

Chehal, D., Gupta, P., and Gulati, P. (2021). Implementation and comparison of topic modeling techniques based on user reviews in e-commerce recommendations. *Journal of Ambient Intelligence and Humanized Computing*, 12(5):5055–5070.

Duarte, D. and Ståhl, N. (2019). Machine learning: a concise overview. In Said, A. and Torra, V., editors, *Data Science in Practice*, pages 27–58. Springer.

Lau, J. H., Newman, D., and Baldwin, T. (2014). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539.

Mifrah, S. and Benlahmar, E. (2020). Topic modeling coherence: A comparative study between lda and nmf models using covid'19 corpus. *Int. J. Adv. Trends Comput. Sci. Eng*, 9:5756–5761.

Mimno, D. and McCallum, A. (2008). Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *UAI*, UAI'08, page 411–418, Arlington, Virginia, USA. AUAI Press.

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, USA. Association for Computing Machinery.

Steyvers, M. and Griffiths, T. (2007). Probabilistic topic models. In Landauer, T. K., McNamara, D. S., Dennis, S., and Kintsch, W., editors, *Handbook of latent semantic analysis*, chapter 21, pages 424–440. Laurence Erlbaum Associates.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476):1566–1581.

Williams, T. and Betak, J. (2018). A comparison of lsa and lda for the analysis of railroad accident text. *Procedia computer science*, 130:98–102.