Novas Métricas para Avaliação da Qualidade de Estratégias de Modelagem de Tópicos Hierárquica

Antônio Pereira , Leonardo Rocha , Felipe Viegas

[1] UFSJ - São João del-Rei, MG , Brasil

antoniopereira@aluno.ufsj.edu.br, lcrocha@ufsj.edu.br

Abstract. *Hierarchical Topic Modeling (HTM) are strategies that aim to automatically extract consistent semantic topics from textual documents, respecting the hierarchy in which the information is structured. Current evaluation metrics for these approaches typically measure the quality of each topic individually. In HTM, other issues need to be considered: (i) Redundancy of topics; (ii) Semantic diversity of constructed topics; and (iii) Topological consistency. In this work, we propose and evaluate three new evaluation metrics that consider these issues, complementing the methodology for evaluating HTM approaches from the perspective of the hierarchical structure in which the topics are constructed.*

Resumo. *Modelagem Hierárquica de Tópicos (MHT) são abordagens que visam extrair automaticamente tópicos semânticos consistentes a partir de documentos textuais, respeitando a hierarquia nas quais as informações se estruturam. As atuais métricas de avaliação dessas abordagens normalmente medem a qualidade de cada tópico individualmente. Em MHT outras questões precisam ser consideradas: (i) Redundância dos tópicos; (ii) Diversidade semântica dos tópicos construídos; (iii) Consistência topológica. O presente trabalho propõem e avalia três novas métricas de avaliação que consideram essas questões, complementando a metodologia de avaliação de abordagens de MHT sob a perspectiva da estrutura hierárquica em que os tópicos são construídos.*

1. Introdução

O grande volume de dados disponível na WEB gerou nos últimos anos um desafiante e intrigante cenário para variadas aplicações: há mais dados que efetivamente pode-se analisar. Representar adequadamente tais informações, sem perdas, bem como desenvolver estratégias efetivas e eficientes para manuseá-las, é uma das tarefas mais desafiadoras em Ciência da Computação. A modelagem de tópicos está entre as abordagens mais exploradas para extrair e organizar informações de grandes quantidades de dados. Essas abordagens visavam encontrar automaticamente tópicos semânticos a partir de documentos textuais [Bicalho et al. 2014], os quais podem ser explorados por outras aplicações, tais como máquinas de busca e sistemas de recomendação, para auxiliar a realização de tarefas específicas.

Além da tradicional modelagem, existe uma segunda vertente que apresenta resultados com um maior grau descritivo, a Modelagem Hierárquica de Tópicos (MHT). MHT é uma tarefa de aprendizado de máquina não supervisionada que visa induzir tópicos latentes de coleções textuais preservando a estrutura hierárquica inerente [Teh et al. 2006]. MHT apresenta seus próprios desafios para garantir sua aplicação: *Coerência de tópicos*

e *Estrutura Semântica*. *Coerência de tópicos* está relacionado à necessidade de aprender tópicos significativos em que as principais palavras que representam um tópico sejam semanticamente consistentes entre si. *Estrutura Semântica* está relacionado à estrutura hierárquica, em que os tópicos próximos da raiz devem ser mais gerais, enquanto os tópicos próximos às folhas são mais específicos e consistentes com seus tópicos pais.

Uma questão importante nesta linha de pesquisa é como avaliar automaticamente a qualidade dos tópicos gerados sem um julgamento humano. As métricas existentes na literatura exploram as t principais palavras para medir a qualidade dos tópicos construídos, com destaque para *Coherence* e NPMI [Nikolenko et al. 2017]. Essas métricas visam capturar a qualidade da interação entre as palavras atribuídas em cada tópico, avaliando a qualidade de cada tópico individualmente. Apesar da efetividade comprovada dessas métricas, estratégias de MT recentes da literatura estão alcançando as pontuações máximas para as mesmas [Viegas et al. 2020]. Do ponto de vista do cenário de MHT, essas métricas convencionais não conseguem avaliar com precisão os modelos de tópicos hierárquicos sob as perspectivas dos desafios previamente mencionados (Coerência de tópicos e Estrutura Semântica). Os tópicos que seguem a mesma topologia devem ter o mesmo contexto, enquanto diferentes topologias devem ter contextos diferentes. Por fim, foi observado no presente trabalho que essas métricas convencionais medem apenas a qualidade dos tópicos, mas não conseguem capturar comportamentos, como duplicidade de tópicos construídos.

Dessa forma, o objetivo deste trabalho é propor uma metodologia de avaliação de estratégias de MHT que capture aspectos distintos e complementares às métricas existentes na literatura. Nesse sentido, foi proposto novas métricas que avaliam eficientemente a qualidade dos tópicos gerados por estratégias de MHT. As métricas propostas conseguem avaliar de forma sintática e semântica a divergência dos tópicos, fornecendo uma avaliação em aspectos ainda não considerados na literatura: (i) Redundância dos tópicos (das principais palavras); (ii) Diversidade semântica dos tópicos construídos; (iii) Consistência das relações topológicas. Apresentamos uma avaliação experimental completa em que foi comparado às duas principais estratégias de MHT existentes na literatura considerando dez coleções de dados distintas, métricas tradicionais propostas na literatura, bem como as novas métricas propostas nesse trabalho. Os resultados deixam claro que as métricas propostas são capazes de complementar as avaliações obtidas por métricas tradicionais, permitindo que outras avaliações sejam feitas sob a perspectiva da estrutura hierárquica com que os tópicos são construídos.

Enfatizamos que a concepção das novas métricas de avaliação de qualidade de abordagens de MHT, bem como todas as implementações e execuções de experimentos foram realizadas pelo aluno Antônio Pereira, sob a orientação do professor Leonardo Rocha. O trabalho faz parte de um projeto maior que visa propor e avaliar novas abordagens de MT e MHT e contou com a colaboração do aluno de pós-graduação Felipe Viegas na concepção do ambiente experimental e nas análises de resultados.

2. Trabalhos Relacionados

Métodos de MHT podem ser agrupados em supervisionados e não-supervisionados. Considerando as estratégias supervisionadas, destacamos o LDA hierárquico supervisionado (HSLDA) [Perotte et al. 2011], extensão para a tradicional LDA, que incorpora a hierarquia de dados com vários rótulos e pré-rótulos em um

único modelo, fornecendo assim recursos de previsão estendidos à tópicos hierárquicos. Considerando estratégias não-supervisionadas de MHT, em [Mimno et al. 2007] é proposto o **hPAM**, uma extensão da técnica de MT conhecida como *Pachinko Allocation* (PAM) [Li and McCallum 2006]. No PAM, os documentos são uma mistura de distribuições em um conjunto de tópicos individuais, usando um gráfico acíclico direcionado para representar as co-ocorrências de tópicos. Em [Viegas et al. 2020] os autores propõem o **CluHTM**, uma nova estratégia de fatoração de matriz não-probabilística que adota a representação denominada *cluwords* em conjunto com uma variante do NMF, especialmente desenvolvida para modelagem de tópicos sobre uma hierarquia de documentos. Em avaliações recentes [Viegas et al. 2020], o CluHTM e o HPAM apresentaram os melhores resultados para as métricas tradicionais, em alguns casos alcançando pontuações próximas à máxima.

2.1. Métricas de Avaliação Tradicionais

Às três principais métricas utilizadas na literatura avaliam a qualidade representativa dos tópicos conforme as ocorrências das palavras mais importantes em cada tópico, sem considerar a correlação entre os tópicos gerados [Nikolenko et al. 2017]. São elas:

- *Coherence:* captura a facilidade de interpretação segundo a coocorrência das palavras sendo definida na Equação 1.

$$c_{\text{tf-idf}}(t, W_t) = \sum_{w_1, w_2 \in W_t} log \frac{\sum_{d: w_1, w_2 \in d} \text{tf-idf}(w_1, d) \times \text{tf-idf}(w_2, d)}{\sum_{d: w_1 \in d} \text{tf-idf}(w_1, d)} \quad (1)$$

  onde a métrica *tf-idf* é dada pela frequência aumentada conforme a Equação 2.

$$\text{tf-idf}(w, d) = \left(\frac{1}{2} + \frac{f(w, d)}{max_{w' \in d} f(w', d)}\right) \times log\left(\frac{|D|}{|\{d \in D : w \in d\}|}\right) \quad (2)$$

  e $f(w, d)$ é o número de ocorrências de uma palavra w no documento d.
- *Normalized Pairwise pointwise Mutual Information (NPMI):* mede quanto uma palavra "ganha" de informação dada à ocorrência de outra palavra, considerando as dependências entre as palavras. Para um determinado conjunto ordenado das palavras mais importantes $W_t = (w_1, ..., w_N)$ de um tópico, a métrica NPMI é calculada segundo a Equação 3.

$$\text{NPMI}_t = \sum_{i<j} \frac{log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}}{-log\, p(w_i, w_j)} \quad (3)$$

3. Métricas Propostas

Esta seção apresenta às três métricas de avaliação de estratégias de MHT desenvolvidas para capturar aspectos distintos e complementares às métricas existentes na literatura, as quais focam nos seguintes aspectos: (i) Redundância dos tópicos (das principais palavras); (ii) Diversidade semântica dos tópicos construídos; (iii) Consistência das relações topológicas. As métricas desenvolvidas exploram dois conceitos comumente utilizados no contexto de algoritmos de agrupamentos: *distância intra grupo* e *distância inter grupo*. Ambos foram adaptados para o contexto de MHT como *distância intra tópico* e *distância inter tópico*. Note que, assim como no cenário de agrupamento, a hierarquia construída por um método deve ser avaliada medindo o nível de correlação entre os tópicos construídos dentro de uma mesma topologia (*distância intra tópico*), bem como, em tópicos construídos em topologias distintas (*distância inter tópico*).

A Equação 4 apresenta a *distância intra tópico* a qual é calculada como a média das distâncias entre os tópicos pertencentes a uma mesma topologia (τ_i).

$$intra = \frac{\sum_{i \in \tau}(distancia\tau_i)}{|\tau|} \tag{4}$$

onde $|\tau|$ corresponde ao número total de topologias criadas.

A Equação 5 refere-se à *distância inter tópico* a qual é calculada pela distância entre os tópicos de uma determinada topologia τ_i para as demais topologias τ_j. Isso é feito para cada uma das topologias e, posteriormente, é calculada a média dessas distâncias.

$$inter = \frac{\sum_{i \in \tau} \sum_{j \in \tau i \neq j} distancia(\tau_i, \tau_j)}{|\tau|} \tag{5}$$

onde $|\tau|$ corresponde ao número total de topologias criadas.

A seguir, será apresentado diferentes maneiras de instanciar estas métricas.

3.1. Avaliação Sintática - Distância de Jaccard

A métrica proposta nesta seção visa avaliar a redundância de tópicos em diferentes níveis da hierarquia e, por consequência, avaliar a consistência das relações topológicas em relação à informação construída pela estratégia. Nesse caso, foi proposto a utilização da distância de Jaccard, sendo a subtração de um pelo coeficiente de Jaccard. O coeficiente de Jaccard mostra a semelhança entre dois conjuntos, sua medida é definida pela razão entre o número de elementos da interseção e o número de elementos da união.

$$J_d(A, B) = 1.0 - \frac{|A \cup B|}{|A \cap B|} \tag{6}$$

Assim, para o cálculo da distância intra tópico, a função $distancia$, apresentada na Equação 4, é instanciada usando a distância de Jaccard J_d, conforme apresentado abaixo:

$$distancia\_intra\_jaccard(\tau_k) = \frac{\sum_{i \in \tau_k} J_d(t_i, t_r)}{|\tau_k|} \tag{7}$$

onde t_r é o tópico do nível imediatamente superior à t_i na mesma topologia analisada.

Para o cálculo da distância inter tópico, a função $distancia$, apresentada na Equação 5, é instanciada usando a distância de Jaccard J_d, conforme apresentado abaixo:

$$distancia\_inter\_jaccard(\tau_k, \tau_{all-k}) = \frac{\sum_{i \in \tau_k} \sum_{j \in \tau_{all-k}} J_d(t_i, t_j)}{|\tau_k| * |\tau|} \tag{8}$$

onde t_j é o tópico do nível imediatamente superior à t_i nas demais topologias τ_{all-k} diferentes de τ_k.

É importante observar que métricas de *distância intra tópico* e *distância inter tópico* possuem referências distintas para medir a distância em termos de Jaccard. Ambas sempre comparam um tópico de referência t_i a tópicos t_j em um nível de hierarquia superior à t_i. Entretanto, enquanto na *distância intra tópico* t_j está mesma topologia de t_i, para o cálculo da *distância inter tópico* t_j está nas outras topologia diferentes da topologia de t_i.

3.2. Avaliação Sintática - Unicidade de tópicos

Essa métrica consiste em avaliar se estratégias de modelagem hierárquica de tópicos conseguem produzir tópicos diversos e sem repetições de palavras. Consideramos a função $unique(t_i, t_j)$ que retorna a quantidade de palavras únicas presentes em ambos os tópicos t_i e t_j. Essa quantidade é dividida pelo total de palavras de ambos os tópicos. Quanto maior o valor dessa razão, mais distintos sintaticamente são os tópicos.

Assim, para o cálculo da distância intra tópico, a função $distancia$, apresentada na Equação 4, é instanciada usando a função $unique(t_i, t_j)$, conforme apresentado abaixo:

$$distancia\_intra\_unicidade(\tau_k) = \frac{\sum_{i \in \tau_k} \sum_{j \in \tau_k i \neq j} \frac{unique(t_i, t_j)}{cont_p(t_i) + cont_p(t_j)}}{|\tau_k|!} \tag{9}$$

onde t_i e t_j são da mesma topologia e mesmo nível hierárquico.

Para o cálculo da distância inter tópico, a função $distancia$, apresentada na Equação 5, é instanciada usando a função $unique(t_i, t_m)$, conforme apresentado abaixo:

$$distancia\_inter\_unicidade(\tau_k, \tau_{all-k}) = \frac{\sum_{i \in \tau_k} \sum_{j \in \tau_{all-k}} \sum_{m \in \tau_j} \frac{unique(t_i, t_m)}{cont_p(t_i) + cont_p(t_m)}}{|\tau_{all}|} \tag{10}$$

onde $\sum_{j \in \tau_{all-k}}$ itera sobre cada uma das topologias diferentes de τ_k e, assim, t_i e t_m são de topologias distintas, mas no mesmo nível hierárquico. Além disso, $|\tau_{all}|$ corresponde ao número total de tópicos presentes em todas as topologias.

3.3. Avaliação Semântica

A terceira métrica avalia a qualidade dos tópicos em relação à semântica. Para capturar a informação semântica dos tópicos, utilizamos as representações vetoriais (*word embedding* [Mikolov et al. 2017]) das palavras mais importantes dos tópicos. O objetivo é avaliar a qualidade em termos da similaridade semântica entre os *word embeddings* das palavras dos tópicos. A métrica explora a Distância de Cosseno entre os *embeddings* para avaliar a diversidade e consistência da hierarquia topológica. Do ponto de vista semântico, quanto mais próximos os *embeddings* dos tópicos em uma mesma topologia (distância intra tópico), mais consistente é a hierarquia (aspecto – iii). Por outro lado, quanto mais distantes os *embeddings* dos tópicos em diferentes topologias (distância inter tópico), mais diversa é a hierarquia criada (aspecto – ii). Para instanciar a função $distancia$, primeiramente vamos definir a distância de cosseno, conforme apresentado pela Equação abaixo:

$$cos_d = 1.0 - \frac{\sqrt{\sum x * y}}{\sqrt{\sum x^2} * \sqrt{\sum y^2}} \tag{11}$$

onde x e y, para nosso cenário, são dois vetores *embeddings*.

Para comparar semanticamente dois tópicos, primeiramente calcula-se o centroide de cada tópico através de uma média vetorial dos *embeddings* das palavras mais representativas de cada tópico, conforme apresentado na equação abaixo:

$$t_i = \frac{\sum_{i=1}^{np} embeddind_{word_i}}{np} \tag{12}$$

onde np é a quantidade de palavras dos tópicos.

Para o cálculo da distância intra tópico, a função $distancia$ (Equação 4), é instanciada usando a distância de cosseno cos_d, conforme apresentado abaixo:

$$distancia\_intra\_cosseno(\tau_k) = \frac{\sum_{i \in \tau_k} cos_d(t_i, t_r)}{|\tau_k|} \quad (13)$$

onde t_r é o tópico do nível imediatamente superior à t_i na mesma topologia analisada.

Para o cálculo da distância inter tópico, a função $distancia$ (Equação 5), é instanciada usando a distância de Jaccard J_d, conforme apresentado abaixo:

$$distancia\_inter\_cosseno(\tau_k, \tau_{all-k}) = \frac{\sum_{i \in \tau_k} \sum_{j \in \tau_{all-k}} cos_d(t_i, t_j)}{|\tau_k| * |\tau|} \quad (14)$$

onde t_j é o tópico do nível imediatamente superior à t_i nas demais topologias τ_{all-k} diferentes de τ_k.

É importante mencionar que as distâncias semânticas intra e inter se diferenciam das distâncias sintáticas. Para ambas as distâncias, sob a perspectiva sintática, é desejável que sejam valores grandes, representando a não ocorrência de repetições entre os termos que compõem os diversos tópicos em uma mesma topologia hierárquica (intra) e em topologias distintas (inter). No entanto, sob a perspectiva semântica, esses valores são antagônicos: a distância semântica intra tópico precisa ser pequena (semanticamente próximos), enquanto distância semântica inter tópico precisa ser grande (semanticamente distantes). Assim como na literatura, o presente trabalho considera uma avaliação da razão entre a distância inter pela distância intra de tal forma que, quanto menor essa razão, mais consistente e mais diversa é toda a hierarquia de tópicos construída.

4. Avaliação Experimental

4.1. Configuração Experimental

Em termos de configuração experimental, foi considerado dez coleções de dados utilizados em trabalhos recentes que avaliam abordagens de MHT [Viegas et al. 2020], nove delas contendo comentários de aplicativos na Google Play Store (i.e. Angrybirds, Dropbox, Evernote, InfoVis-Vas, Pinterest, TripAdvisor, WhatsApp, Facebook e Uber) e uma do Twitter. Os tópicos construídos pelas estratégias de MHT consideradas estado-da-arte pela literatura – CluHTM [Viegas et al. 2020] e o HPAM [Mimno et al. 2007] – foram comparadas em relação ao *Coherence* e NPMI [Nikolenko et al. 2017] (métricas tradicionais - Seção 2.1), bem como as métricas propostas neste trabalho (Seção 3).

A métrica semântica proposta necessita de um modelo de *embedding* para realizar a transformação de palavras e tópicos em vetores. Desta forma, a qualidade do modelo de *embedding* utilizado é fundamental. Diante disso, foi utilizado o modelo FastText[1], treinado a partir de documentos da Wikipedia e documentos de notícias.

Para o cálculo da significância estatística foi utilizado o *Two-way ANOVA test*, o qual avalia o quanto as médias das distâncias para cada tópico variam em ambos os métodos, e por fim, gera-se uma probabilidade das duas ocorrências serem distintas. Foi adotado a significância mínima de 95% de confiança. Os valores marcados em ✘ não possuem significância estatística. Os valores marcados em ✔ revelam uma significância

[1]Em análises preliminares, foi o modelo que apresentou os melhores resultados.

estatística onde o valor de uma métrica é maior a outra. Valores marcados em ✓ possuem uma significância estatística, onde o valor de uma métrica é menor a outra.

4.2. Métricas tradicionais de avaliação de tópicos

Esta seção apresenta os resultados de eficácia em termos de qualidade de tópicos produzidos pelas estratégias CluHTM e HPAM considerando as métricas tradicionais *Coherece* e *NPMI*. A Tabela 1 apresenta os resultados para a métrica *Coherence*. É possível observar que o CluHTM alcança resultados significativos em 3 das 10 bases de dados avaliadas, ocorrendo empate estatístico em outras 5. A Tabela 2 apresenta resultados para a métrica NPMI, mostrando que a estratégia CluHTM conseguiu produzir tópicos, em que palavras compartilham mais informação do que nos tópicos produzidos pelo HPAM. Considerando os resultados das duas métricas de avaliação, a estratégia CluHTM se mostrou superior à estratégia HPAM, conforme apresentado em [Viegas et al. 2020]. Observa-se que para a métrica Coherence há um melhor equilíbrio na qualidade de ambos os métodos. Por outro lado, para a métrica NPMI o CluHTM foi muito melhor em todas as coleções e quase atingiu o valor máximo em alguns casos. As métricas de qualidade propostas na Seção 3 visam complementar as métricas de qualidade de tópicos avaliadas, para verificar se, de fato, outros aspectos podem apoiar os resultados apresentados nesta seção.

Tabela 1. Avaliação da qualidade dos tópicos produzidos em relação a métrica *Coherence*

Coleções	CluHTM	HPAM
ang	**-77.39 ± 41.17 ✘**	**-46.80 ± 16.44**
drop	**-69.56 ± 33.86 ✘**	**-65.52 ± 12.96**
ever	**-54.45 ± 32.81 ✔**	-87.33 ± 8.80
face	**-110.57 ± 51.62 ✘**	**-98.11 ± 12.53**
info	**-4.46 ± 11.36 ✔**	-61.9381 ± 10.1260
pinter	-111.90 ± 48.77	**-58.00 ± 16.46 ✔**
trip	**-55.60 ± 27.54 ✘**	**-68.95 ± 15.60**
tweets	**-93.11 ± 31.38 ✘**	**-92.52 ± 9.71**
uber	**-75.25 ± 39.94 ✔**	-94.72 ± 9.24
wpp	-105.28 ± 38.64	**-48.48 ± 15.55✔**

Tabela 2. Avaliação da qualidade dos tópicos produzidos em relação a métrica NPMI

Coleções	CluHTM	HPAM
ang	**0.8934 ± 0.0514 ✔**	0.3604 ± 0.1005
drop	**0.9002 ± 0.0454 ✔**	0.2529 ± 0.0877
ever	**0.9374 ± 0.0334 ✔**	0.1534 ± 0.0564
face	**0.8686 ± 0.0531 ✔**	0.1517 ± 0.0765
info	**0.9935 ± 0.0190 ✔**	0.1191 ± 0.0533
pinter	**0.8482 ± 0.0535 ✔**	0.3028 ± 0.0988
trip	**0.9265 ± 0.0344 ✔**	0.2745 ± 0.0906
tweets	**0.8950 ± 0.0323 ✔**	0.2130 ± 0.0453
uber	**0.9116 ± 0.0424 ✔**	0.1403 ± 0.0582
wpp	**0.8594 ± 0.0456 ✔**	0.3976 ± 0.0750

4.3. Avaliação Sintática - Distância de Jaccard

A Tabela 3 apresenta os resultados da avaliação *intra* tópicos em relação à distância de Jaccard. Pode-se observar que considerando os ganhos estatísticos, a estratégia HPAM se mostrou superior quando comparado com o CluHTM. Mas, vale ressaltar que esta métrica mede a consistência das relações topológicas em relação à informação construída, e no caso da distância *intra* tópicos, esta métrica está medindo as relações em tópicos de níveis hierárquicos da mesma topologia. É esperado que níveis inferiores da mesma topologia absorvam informações de níveis superiores, visto que o objetivo dos níveis inferiores é uma "especialização" dos níveis superiores da mesma topologia. Pode-se observar este comportamento nas Figuras 1 (a) e (b), na qual o primeiro tópico de ambas Figuras corresponde ao tópico mais generalizado, quando comparado com os demais tópicos, que são os tópicos construídos a partir da informação do respectivo tópico.

A Tabela 4 apresenta os resultados da avaliação *inter* tópicos. Pode-se observar uma leve superioridade do CluHTM que consegue produzir tópicos mais dissimilares, considerando a distância de Jaccard, quando se contrasta diferentes níveis hierárquicos

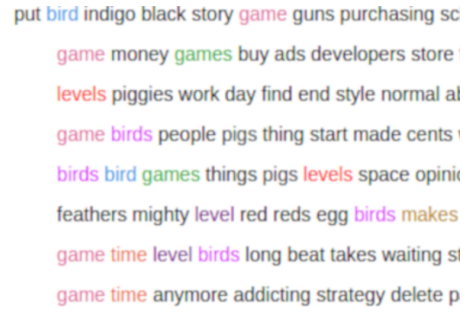
Tabela 3. Avaliação Sintática - Distância de Jaccard – *intra*

Coleções	CluHTM	HPAM
ang	**0.9612 ± 0.0474 ✘**	**0.9947 ± 0.0062**
drop	0.9372 ± 0.0468	**0.9920 ± 0.0049 ✔**
ever	**0.9960 ± 0.0035 ✔**	0.9806 ± 0.0156
face	**0.9747 ± 0.0607 ✘**	**0.9823 ± 0.0088**
info	**0.9816 ± 0.0123 ✘**	**0.9893 ± 0.0107**
pinter	0.9630 ± 0.0301	**0.9921 ± 0.0060 ✔**
trip	**0.9768 ± 0.0266 ✘**	**0.9886 ± 0.0141**
tweets	0.9821 ± 0.0144	**0.9968 ± 0.0048 ✔**
uber	**0.9927 ± 0.0073 ✔**	0.9736 ± 0.0171
wpp	0.9823 ± 0.0109	**0.9958 ± 0.0046 ✔**

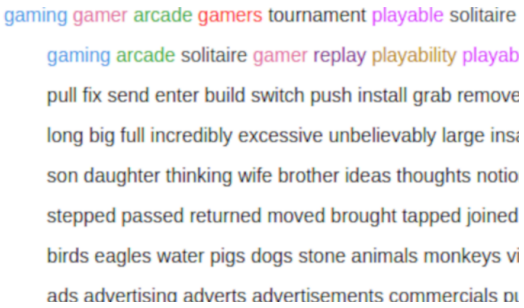
Tabela 4. Avaliação Sintática - Distância de Jaccard – *inter*

Coleções	CluHTM	HPAM
ang	**0.9929 ± 0.0071 ✘**	**0.9946 ± 0.0011**
drop	**0.9966 ± 0.0019 ✔**	0.9939 ± 0.0012
ever	**0.9876 ± 0.0081 ✔**	0.9864 ± 0.0034
face	**0.9949 ± 0.0022 ✘**	**0.9872 ± 0.0026**
info	**0.9924 ± 0.0028 ✘**	**0.9919 ± 0.0021**
pinter	**0.9945 ± 0.0035 ✘**	**0.9943 ± 0.0010**
trip	**0.9898 ± 0.0064 ✘**	**0.9906 ± 0.0017**
tweets	**0.9981 ± 0.0019 ✘**	**0.9978 ± 0.0007**
uber	**0.9925 ± 0.0069 ✔**	0.9773 ± 0.0039
wpp	**0.9937 ± 0.0029 ✘**	**0.9960 ± 0.0014**

em distintas topologias. Contudo, para a maioria das coleções de dados (7 em 10), pode-se observar que ambas estratégias são similares em termos *distância inter tópicos*. Isto pode ser explicado considerando que as coleções de dados avaliadas são de aplicações bem definidas (*review* de aplicativos), e embora cada topologia tenha a sua respectiva informação e "especialização", todos os tópicos compartilham informação central da coleção de dados, sendo de fato a aplicação.



(a) HPAM - Pode-se observar que as palavras *birds*, *levels*, *game* são repetidas em diversos tópicos produzidos pela mesma topologia.



(b) CluHTM - Pode-se observar que as palavras ocorrem em apenas em um tópico da topologia de tópicos.

Figura 1. Exemplo de topologia de tópicos produzidos para coleção Angry birds.

4.4. Avaliação Sintática - Unicidade de tópicos

Tabela 5. Avaliação Sintática - Unicidade dos tópicos – *intra*

Coleções	CluHTM	HPAM
ang	**0.9989 ± 0.0017 ✔**	0.9681 ± 0.0077
drop	**0.9927 ± 0.0080 ✔**	0.9794 ± 0.0056
ever	**1.0000 ± 0.0000 ✔**	0.9519 ± 0.0096
face	**0.9961 ± 0.0104 ✔**	0.9768 ± 0.0089
info	**0.9976 ± 0.0016 ✔**	0.9756 ± 0.0053
pinter	**0.9958 ± 0.0044 ✔**	0.9810 ± 0.0037
trip	**0.9998 ± 0.0003 ✔**	0.9827 ± 0.0041
tweets	**0.9995 ± 0.0016 ✔**	0.9966 ± 0.0020
uber	**1.0000 ± 0.0000 ✔**	0.9554 ± 0.0121
wpp	**0.9983 ± 0.0016 ✔**	0.9921 ± 0.0025

Tabela 6. Avaliação Sintática - Unicidade dos tópicos – *inter*

Coleções	CluHTM	HPAM
ang	**0.9945 ± 0.0009 ✔**	0.9437 ± 0.0036
drop	**0.9954 ± 0.0014 ✔**	0.9548 ± 0.0035
ever	**0.9930 ± 0.0016 ✔**	0.9334 ± 0.0044
face	**0.9963 ± 0.0013 ✔**	0.9552 ± 0.0075
info	**0.9913 ± 0.0017 ✔**	0.9495 ± 0.0039
pinter	**0.9962 ± 0.0006 ✔**	0.9589 ± 0.0028
trip	**0.9950 ± 0.0009 ✔**	0.9571 ± 0.0029
tweets	**0.9985 ± 0.0008 ✔**	0.9716 ± 0.0022
uber	**0.9945 ± 0.0019 ✔**	0.9329 ± 0.0081
wpp	**0.9956 ± 0.0015 ✔**	0.9680 ± 0.0025

A Tabela 5 apresenta os resultados da distância *intra* tópicos em relação à Unicidade dos tópicos. Observa-se que o CluHTM superou o HPAM em todas as coleções de dados. Este comportamento pode ser justificado observando as Figuras 1 (a) e (b) como exemplo. Na Figura 1 (a), note que os tópicos produzidos pelo HPAM repetem palavras (tais como, *games* e *level*) em quase todos os tópicos da mesma topologia e nível

hierárquico. Isto demonstra que o HPAM tem uma capacidade de "especialização" de tópicos inferiores ao CluHTM, ilustrado na Figura 1 (b). A Tabela 6 também apresenta um comportamento similar para a distância *inter* tópicos. O CluHTM se mostrou mais robusta em termos de produzir tópicos únicos para distintas topologias. O HPAM apresentou o mesmo comportamento da Figura 1 (a) em relação à repetição de palavras em diferentes topologias, justificando os ganhos do CluHTM do ponto de vista *inter* tópicos.

4.5. Avaliação Semântica

A Tabela 7 apresenta os resultados da distância *intra* tópicos mostrando que a estratégia HPAM se mostrou superior em todas as coleções de dados, quando comparado com a estratégia CluHTM. Este resultado é interessante e pode ter uma falsa interpretação caso seja analisado individualmente. Na verdade, este resultado é negativo para a estratégia HPAM, quando se analisa este resultado em contraste com os resultados apresentados na seções 4.3 e 4.4. Note que a estratégia HPAM produz tópicos com repetições de palavras, assim, por conta deste viés, é esperado que distância *intra* de tópicos seja superior quando comparado com CluHTM. A Figura 1 (a) novamente pode ilustrar este comportamento quando se observa a ocorrência da palavra *game* em quase todos os tópicos, inclusive no tópico superior da topologia hierárquica. A Tabela 8 apresenta os resultados da distância *inter* tópicos. Pode-se observar que a estratégia CluHTM apresenta os melhores resultados comparado com o HPAM, e contrastando com as métricas das seções 4.3 e 4.4, este resultado reforça que estratégia CluHTM constrói topologias de tópicos dissimilares quando comparado com a métrica HPAM. A Tabela 9 apresenta a razão entre as distâncias *intra* e *inter* tópicos. Esta razão é comumente utilizado no contexto de agrupamento, é uma forma de sumarizar estes resultados, e novamente, a estratégia CluHTM se mostrou superior em contraste com a estratégia HPAM.

Tabela 7. Avaliação Semântica – *intra*

Coleções	CluHTM	HPAM
ang	0.3379 ± 0.0691	**0.2037 ± 0.0271** ✓
drop	0.2961 ± 0.0577	**0.2132 ± 0.0298** ✓
ever	0.3892 ± 0.0281	**0.2094 ± 0.0299** ✓
face	0.3912 ± 0.0537	**0.2227 ± 0.0193** ✓
info	0.2597 ± 0.0207	**0.2096 ± 0.0198** ✓
pinter	0.3257 ± 0.0501	**0.1894 ± 0.0242** ✓
trip	0.3557 ± 0.0331	**0.1940 ± 0.0171** ✓
tweets	0.3319 ± 0.0597	**0.1980 ± 0.0164** ✓
uber	0.3837 ± 0.0576	**0.2023 ± 0.0221** ✓
wpp	0.3207 ± 0.0594	**0.1976 ± 0.0107** ✓

Tabela 8. Avaliação Semântica – *inter*

Coleções	CluHTM	HPAM
ang	**0.4248 ± 0.0110** ✔	0.2047 ± 0.0034
drop	**0.4273 ± 0.0146** ✔	0.2184 ± 0.0032
ever	**0.4560 ± 0.0153** ✔	0.2193 ± 0.0094
face	**0.4915 ± 0.0087** ✔	0.2325 ± 0.0079
info	**0.3033 ± 0.0288** ✔	0.2154 ± 0.0048
pinter	**0.4297 ± 0.0125** ✔	0.1968 ± 0.0058
trip	**0.4413 ± 0.0117** ✔	0.2007 ± 0.0064
tweets	**0.4890 ± 0.0171** ✔	0.1996 ± 0.0046
uber	**0.4661 ± 0.0113** ✔	0.2112 ± 0.0057
wpp	**0.4430 ± 0.0128** ✔	0.1956 ± 0.0026

Tabela 9. Análise Semântica - Razão entre as Distâncias *intra* e *inter*

Coleções	CluHTM	HPAM
ang	**0.7954**✔	0.9951
drop	**0.6929**✔	0.9761
ever	**0.8535**✔	0.9548
face	**0.7959**✔	0.9574
info	**0.8562**✔	0.9730
pinter	**0.7579**✔	0.9623
trip	**0.8060**✔	0.9666
tweets	**0.6787**✔	0.9919
uber	**0.8232**✔	0.9578
wpp	**0.7239**✔	1.0102

5. Conclusão e Trabalhos Futuros

O presente trabalho propõe uma nova metodologia de avaliação de estratégias de MHT que captura aspectos distintos e complementares às métricas existentes na literatura, tais como: (i) redundância; (ii) diversidade semântica; e (iii) consistência topológica dos tópicos construídos. As métricas propostas exploram dois conceitos: *distância intra tópico* e *distância inter tópico*. O primeiro avalia o nível de correlação entre os tópicos construídos dentro de uma mesma topologia e o segundo em topologias distintas. A proposta foi avaliada comparando às duas principais estratégias de modelagem de tópicos hierárquicas existentes na literatura: CluHTM [Viegas et al. 2020] e HPAM [Mimno et al. 2007] considerando dez coleções de dados distintas e amplamente utilizadas na literatura e duas métricas tradicionais. Considerando as métricas tradicionais, observamos que os tópicos gerados pelo CluHTM apresentaram resultados muito melhores do que aqueles construídos pela estratégia HPAM. O CluHTM apresentou resultados acima de 0,90 para NPMI em quase todas as coleções avaliadas, ou seja, próximo do limite superior do NPMI. As métricas de qualidade de tópicos propostas conseguiram capturar comportamentos distintos dos tópicos construídos. Para ambas as visões sintáticas e semânticas, as métricas propostas capturaram duplicidade nos tópicos construídos pelo HPAM. Nossos resultados também mostraram que o CluHTM é capaz de apresentar coerência em termos de apresentar valores de distância menores para tópicos de uma mesma topologia quando comparados às distâncias entre diferentes topologias. No entanto, ainda há espaço significativo para propostas mais robustas do ponto de vista da consistência da estrutura topológica dos tópicos.

Referências

Bicalho, P. V., de Oliveira Cunha, T., Mourão, F. H. J., Pappa, G. L., and Jr., W. M. (2014). Generating cohesive semantic topics from latent factors. In *BRACIS*.

Li, W. and McCallum, A. (2006). Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning*, pages 577–584. ACM.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. (2017). Advances in pre-training distributed word representations. *CoRR*, abs/1712.09405.

Mimno, D., Li, W., and McCallum, A. (2007). Mixtures of hierarchical topics with pachinko allocation. In *Proceedings of the 24th international conference on Machine learning*, pages 633–640.

Nikolenko, S. I., Koltcov, S., and Koltsova, O. (2017). Topic modelling for qualitative studies. *Journal of Information Science*.

Perotte, A. J., Wood, F., Elhadad, N., and Bartlett, N. (2011). Hierarchically supervised latent dirichlet allocation. In *Advances in neural information processing systems*, pages 2609–2617.

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.

Viegas, F., Cunha, W., Gomes, C., Pereira, A., Rocha, L., and Goncalves, M. (2020). Cluhtm-semantic hierarchical topic modeling based on cluwords. In *Proceedings of the 58th ACL*, pages 8138–8150.