Modelagem de tópicos do Twitter em notícias de futebol

Ahmad Fathan Hidayatullah, Elang Cergas Pembrani, Wisnu Kurniawan, Gilang Akbar, Ridwan Pranata
Departamento de Informática Universitas Islam Indonesia, UII
Yogyakarta, Indonésia e-mail: fathan@uii.ac.id, {14523290, 14523264,
14523091, 14523124} @students .uii.ac.id

***Resumo—Junto* com o desenvolvimento das mídias sociais hoje, o Twitter tornou-se uma das mídias sociais que é usada como provedor de informações atuais sobre futebol. O futebol é o esporte mais popular na Indonésia. Pessoas sempre curiosas sobre alguma atualização de notícias de futebol, como previsão de jogos, resultados de jogos, transferências, rumores, etc. Neste artigo, aplicamos modelagem de tópicos para determinar o tópico dos tweets sobre notícias de futebol em Bahasa Indonesia. Os dados usados neste estudo foram retirados de várias contas oficiais do Twitter indonésio que sempre atualizam sobre o futebol e selecionamos antes. *Latent Dirichlet Allocation* (LDA) foi usado como método de modelagem de tópicos para determinar que tipo de tópicos no Twitter. De acordo com a análise de conteúdo, obtivemos vários tópicos perspicazes, como análise pré-jogo, atualização da partida ao vivo, conquistas do clube de futebol, etc. Geralmente, os tópicos postados pela conta do Twitter do provedor de notícias de futebol fornecem informações sobre a competição de futebol em alguns países, como como Indonésia, Inglaterra, Espanha, Itália e Alemanha.**

Palavras-chave-componente; modelagem de tópicos; dirichlet latente alocação; Twitter; notícias de futebol

I. INTRODUÇÃO

A mídia social se tornou o meio e o recurso de informação mais importante para as pessoas em todo o mundo na última década. Há muitas informações sobre as últimas notícias ou eventos atuais, que são postadas a cada segundo nas mídias sociais. Junto com o desenvolvimento das mídias sociais hoje, o Twitter tornou-se uma das mídias sociais que é usada como provedor de informações atuais sobre futebol.

O futebol é o esporte mais popular na Indonésia. Pessoas sempre curiosas sobre alguma atualização de notícias de futebol, como previsão de jogos, resultados de jogos, transferências, rumores, etc.
Além disso, o povo indonésio não está apenas curioso sobre as atualizações de informações da liga nacional de futebol, mas também da liga internacional de futebol, especialmente algumas das principais ligas da Europa. Existem quatro grandes ligas na Europa que as pessoas estão interessadas em receber atualizações, como a Premier League na Inglaterra, a Bundesliga alemã na Alemanha, a La Liga da Espanha na Espanha e a Serie A na Itália. Portanto, é muito importante que os provedores de notícias esportivas compartilhem atualizações de informações dessas grandes ligas via Twitter.

Para os cidadãos, o Twitter é um dos meios de comunicação mais rápidos para receber as últimas notícias sobre futebol.

Esses enormes dados do Twitter fornecem alguns tópicos ocultos e informações importantes. Os tópicos obtidos dos tweets também podem ilustrar e representar as tendências, tópicos quentes, etc. obter o tópico do corpus, o método de modelagem de tópicos pode ser aplicado. Neste artigo, aplicamos a modelagem de tópicos para determinar o tópico dos tweets sobre futebol. Os dados usados neste estudo foram retirados de várias contas oficiais do Twitter indonésio que sempre atualizam sobre o futebol e selecionamos antes. Utilizamos a *Alocação de Dirichlet Latente* (LDA) como método de modelagem de tópicos para determinar que tipo de tópicos no Twitter.

O restante deste artigo está organizado na seguinte estrutura. A seção 2 descreve o trabalho relacionado. Explicamos nossa metodologia de pesquisa na seção 3. Os resultados e discussões são explicados na seção 4. Finalmente, a seção 5 descreve a conclusão de nossa pesquisa.

II. TRABALHO RELACIONADO

A modelagem de tópicos tem sido amplamente aplicada por pesquisadores em vários campos, incluindo área de pesquisa de transporte [1], medicina e saúde [2][3], bioinformática [4], política [5], etc.
A modelagem de tópicos usando dados do Twitter também foi conduzida por alguns pesquisadores antes. A modelagem de tópicos de dados de tweets tem seus próprios desafios em comparação com outros dados de texto devido à sua forma de linguagem não estruturada e tipo de linguagem não padrão [6]. O método LDA foi aplicado para encontrar tópicos no Twitter e houve algumas novas abordagens para melhorar o desempenho do LDA [7][8][9].

Yoon, et al [5] analisaram a opinião pública do Twitter sobre questões políticas na Coréia, identificando os tópicos mais discutidos por meio do modelo de tópicos LDA. Yang e Rim [9]
propôs um novo método de modelagem de tópicos chamado Trend Sensitive-Latent Dirichlet Allocation para extrair tópicos latentes do conteúdo modelando tendências temporais no Twitter ao longo do tempo. Lim, et al [10] propuseram o modelo de tópicos Twitter-Rede para modelar concorrentemente o texto e a rede social de uma forma não paramétrica totalmente Bayesiana.

III. METODOLOGIA

Nesta seção, descrevemos nossa metodologia que realizado em nossa pesquisa.

A. Recuperação de Dados

O Twitter fornece uma API que permite que as pessoas coletem os tweets. Esta pesquisa utiliza a API do Twitter v1.1 e [1] para obter os tweets. O Vantagens da biblioteca GetOldTweets-python desta biblioteca em comparação com o outro Twitter

[1] https://github.com/Jefferson-Henrique/GetOldTweets-python

978-1-5386-6350-9/18/$31,00 ©2018 IEEE

library é coletar dados com base no intervalo de tempo que especificamos conforme desejado, fácil de usar, e os resultados dos dados são organizados ordenadamente no formato csv.

Os dados usados neste estudo foram recuperados de contas confiáveis do Twitter indonésio que postaram notícias de futebol. Essas contas incluem: @bolanet, @detiksport, @goal_id, @panditfootbal, @vivabola. O total de dados obtidos dessas contas do Twitter são 120.639

tweets com um intervalo de tempo de 1º de janeiro de 2017 a 24 de janeiro Dezembro de 2017. A Tabela 1 mostra o conjunto de dados desta pesquisa.

TABELA I. CONJUNTO DE DADOS DO TWITTER

Não	Conta do Twitter	Número de tweets
1	@VIVAbola	31564
2	@panditfootball	15270
3	@GOAL_ID	34204
4	@detiksport	25541
5	@Bolanet	14060
Total		120639

B. Pré -processamento

A etapa de pré-processamento nesta pesquisa é baseada em pesquisas anteriores sobre tarefas de pré-processamento do Twitter [11].
As tarefas de pré-processamento nesta pesquisa são dobragem de maiúsculas, remoção de tags HTML e caracteres Unicode, remoção de símbolos e emoticons, remoção de caracteres não ASCII, remoção de caracteres especiais do Twitter, remoção de URLs, remoção de pontuação, remoção de números e remoção de palavras de parada.

C. Modelagem de tópicos usando LDA A

modelagem de tópicos é um dos métodos mais poderosos em mineração de texto que visa identificar padrões e encontrar relações entre dados de uma coleção de documentos de texto [12]. O método mais popular na modelagem de tópicos é o LDA.
A LDA provou ser uma metodologia de aprendizagem não supervisionada eficaz para encontrar diferentes tópicos em documentos de texto [13]. A modelagem de tópicos LDA é uma técnica não supervisionada em aprendizado de máquina que foi introduzida pela primeira vez por Blei et al [14] como um modelo probabilístico generativo para corpus de texto.

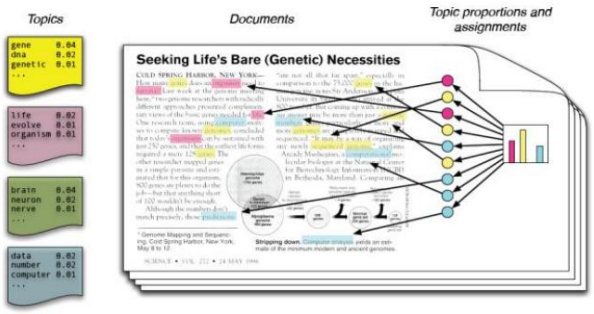

Figura 1. Modelo LDA [15].

O modelo LDA é usado para encontrar a estrutura temática em um documento. O objetivo do método LDA é encontrar a tópicos da coleção de documentos em que cada tópico é uma distribuição sobre palavras ou vocabulário fixo, cada documento é uma mistura de tópicos de todo o corpus e cada termo é retirado de um desses tópicos. O tópico é uma entidade que ilustra a relação entre as palavras, conforme mostrado na Figura 1.

D. Visualização LDA

O resultado do modelo de tópico será visualizado usando a biblioteca Gensim e pyLDAvis em Python. O pyLDAvis é um modelo de visualização de tópicos interativo baseado na web usando LDA que é construído a partir de LDAvis usando uma combinação de R e D3 [16].
Ao usar o pyLDAvis, podemos explorar a relação entre o tópico e os termos para entender o modelo LDA. O PyLDAvis possui dois painéis, o mapa de distribuição de cada tópico e o gráfico de intensidade que representa os termos mais frequentes no corpus.

4. RESULTADO E DISCUSSÃO

A. Análise de Visualização LDA Esta

seção discute o resultado de nossos experimentos. Para realizar a modelagem do tópico LDA, usamos a biblioteca LdaModel fornecida pela biblioteca Gensim em Python. Escolhemos dez tópicos como parâmetro. A visualização do mapa de distância intertópico do nosso modelo é mostrada na figura 2.

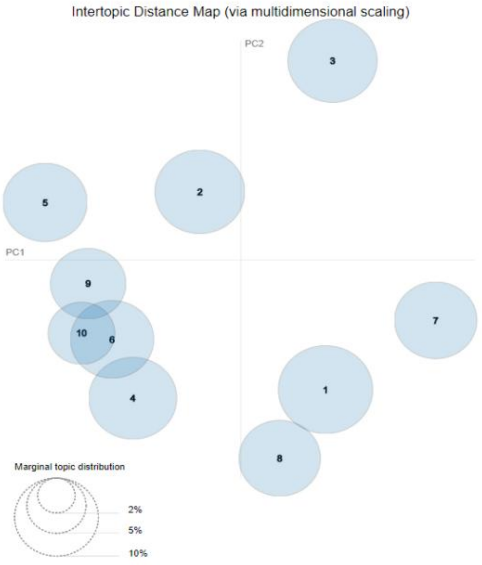

Figura 2. Visualização do Mapa de Distância Intertópico.

De acordo com a figura 2, existem alguns clusters de tópicos que são mutuamente exclusivos, por exemplo, entre os tópicos 6, 9 e 10; tópico número 4 e 6; e entre os tópicos número 1 e 8.
Os grupos de tópicos que se excluem mutuamente indicam que o tópicos têm semelhança. Por outro lado, existem outros tópicos que podem ser agrupados de forma independente, como o agrupamento de tópicos número 2, 3, 5 e 7. Esses agrupamentos abrangeram tópicos específicos que podem ser vistos à distância entre os agrupamentos.
Também indica que a distribuição e frequência da palavra no tópico é muito singular.

A Figura 3 mostra as 30 palavras mais salientes do corpus. Também podemos ver que existem cinco termos que

apareceram com frequência no corpus como “MU”, “Madrid”, “liga”, “Chelsea” e “vs”. Isso indica que Manchester United, Real Madrid e Chelsea são temas bastante discutidos nos noticiários. Além disso, as palavras “liga” e “vs” indicam o tópico sobre partidas de futebol na liga.

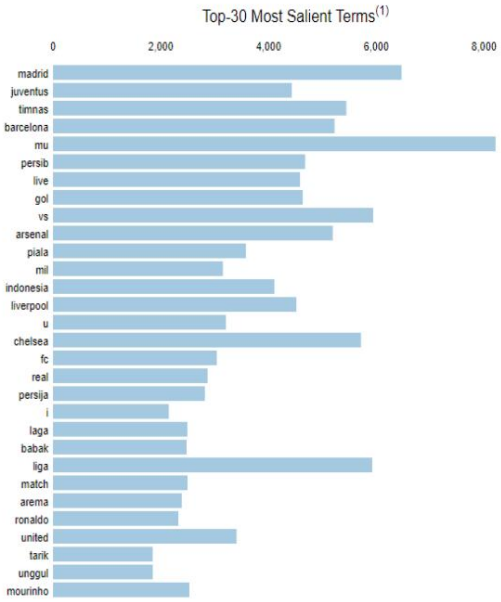

Figura 3. Os 30 Termos Mais Salientes.

B. Análise do Tópico

A visualização de dados de cada tópico é ilustrada usando nuvem de palavras. A nuvem de palavras é uma visualização composta de palavras em dados de texto específicos. A nuvem de palavras exibe a frequência com que as palavras aparecem em uma coleção de texto. O tamanho de cada palavra indica sua importância, o que significa que quanto maior o tamanho da palavra, mais frequente a palavra em um tópico. Além disso, as palavras que dominam a nuvem de palavras provavelmente estão diretamente relacionadas ao tópico da nuvem de palavras.

• Tópico #0: Análise pré-jogo na Premier League

O tópico #0 fala sobre a análise pré-jogo na Premiere League inglesa. Existem algumas palavras dominadas na figura 4 que representam o tópico sobre análise pré-jogo, como *“vs”, “jelang”, “laga”, “rekor”, “fakta”.* As palavras como Chelsea, Tottenham, MU, Arsenal, premier, liga ilustram sobre a Premier League inglesa.

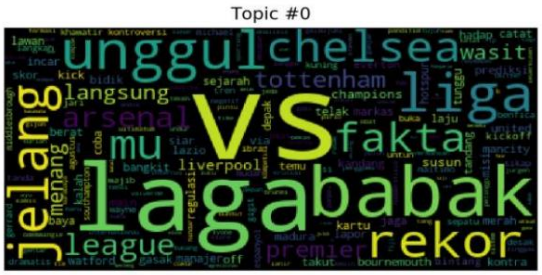

Figura 4. Visualização da Nuvem de Palavras do Tópico #0.

• Tópico #1: Atualização de partidas ao vivo na Liga 1

A Figura 5 mostra a nuvem de palavras do tópico #1. Encontramos a palavra live, match e Persib como o tópico mais dominante. Outras palavras como Arema, Persiba, PSM e Bhayangkara indicam o clube de futebol na Liga Indonésia. A partir dessas palavras dominantes, pode-se concluir que o tópico é sobre partidas ao vivo
na liga de futebol indonésia (Liga 1).

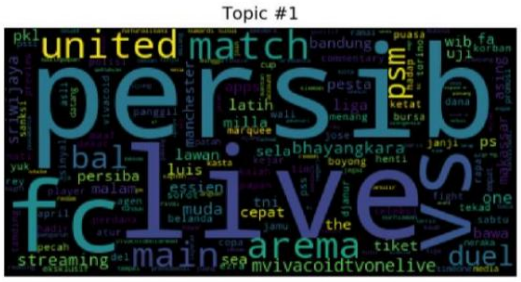

Figura 5. Visualização da Nuvem de Palavras do Tópico 1.

• Tópico #2: Premiere League Inglesa

A visualização da nuvem de palavras do tópico #2 é mostrada na figura 6 abaixo. As palavras Chelsea, Liverpool, City, ManCity, MU estão dominando neste segmento de tópicos. Além disso, também estão listados alguns gerentes da Premier inglesa, como Conte, Guardiola e Klopp.
De acordo com essas palavras, pode-se concluir que o tópico #2 fala sobre a Premier League inglesa.

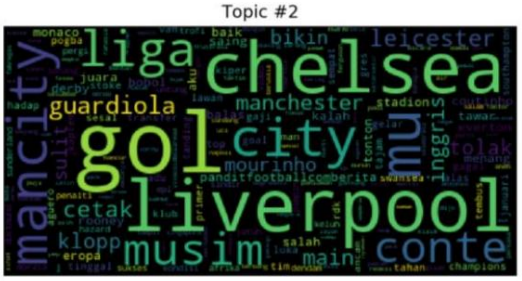

Figura 6. Visualização da Nuvem de Palavras do Tópico 2.

• Tópico #3: Rivalidade entre MU e Arsenal

O tópico discutido no tópico #3 é claramente sobre a rivalidade entre Manchester United e Arsenal. Pode ser visto nos termos Arsenal, MU, Mourinho e Wenger. A nuvem de palavras do tópico #3 pode ser vista na figura 7.

Figura 7. Visualização da Nuvem de Palavras do Tópico 3.

• Tópico #4: Seleção da Indonésia e Indonésia Liga

De acordo com a nuvem de palavras na figura 8, o tópico mais apropriado para o tópico #4 é sobre a seleção da Indonésia e a liga da Indonésia. As palavras timnas e Indonésia indicam o tópico sobre a seleção indonésia. Além disso, existem algumas outras palavras como Persipura, Persija, hasil e klasemen que ilustram o tópico referente à liga indonésia.

Figura 8. Visualização da Nuvem de Palavras do Tópico 4.

• Tópico #5: Série A Itália

A Figura 9 mostra a nuvem de palavras do tópico#5. A partir dos termos da nuvem de palavras, pode-se concluir que o tópico correspondente ao tópico #5 é sobre a Serie A Italia.

Figura 9. Visualização da Nuvem de Palavras do Tópico 5.

• Tópico #6: Copa do Mundo

As palavras do tópico #6 tratam da copa do mundo que pode ser vista a partir de duas palavras dominantes na nuvem de palavras, *"piala"* e *"dunia"*. A nuvem de palavras do tópico #6 é mostrada na figura 10.

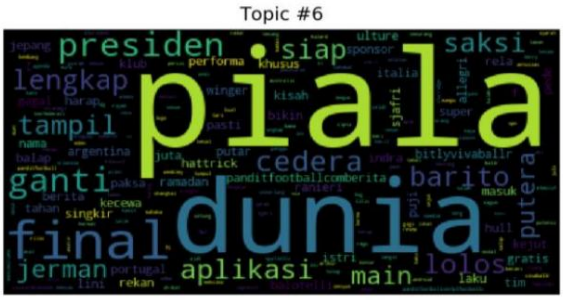

Figura 10. Visualização da Nuvem de Palavras do Tópico 6.

• Tópico #7: El Clássico (Real Madrid x Barcelona)

O tópico #7 refere-se à rivalidade do El Classico entre Real Madrid e FC Barcelona, que pode ser vista em duas palavras dominantes em sua nuvem de palavras, Madrid e Barcelona. Palavras menores também discutem sobre El Clássico, como os nomes de jogadores famosos dos dois clubes, Ronaldo, Messi e Neymar.

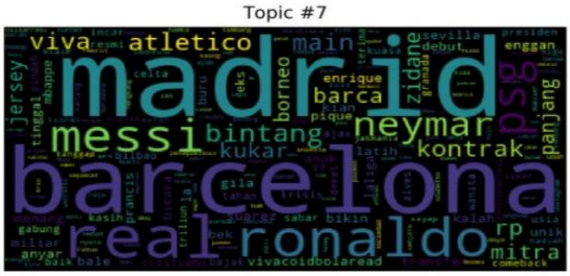

Figura 11. Visualização da Nuvem de Palavras do Tópico 7.

• Tópico #8: Futebol da Indonésia e Semen Padang

O tópico #8 fala sobre um certo clube na Indonésia, Semen Padang. Além disso, este tópico também ilustra sobre o futebol indonésio.

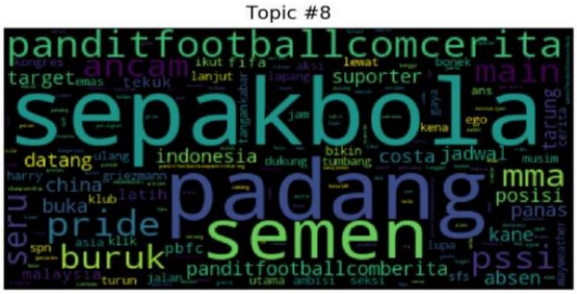

Figura 12. Visualização da Nuvem de Palavras do Tópico 8.

• Tópico #9: A conquista de alguns grandes clubes de futebol da Europa

O tópico #9 refere-se à conquista de alguns grandes clubes da Europa. Isso pode ser visto nas palavras dominantes em sua nuvem de palavras, como os nomes de clubes como Bayern, AC (significa AC Milan), Mil (possivelmente a origem da palavra Milan afetada pelo processo de derivação), Inter, Chelsea.

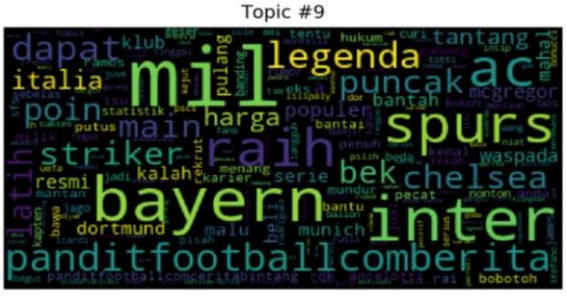

Figura 13. Visualização da Nuvem de Palavras do Tópico nº 9.

V. CONCLUSÃO

Este artigo explorou o uso de modelos tópicos a serem aplicados a mensagens do Twitter que falam sobre futebol usando o método *Latent Dirichlet Allocation* . De acordo com a análise de conteúdo, obtivemos vários tópicos perspicazes, como análise pré-jogo, atualização do jogo ao vivo, conquistas do clube de futebol, etc. Geralmente, os tópicos postados pela conta do Twitter do provedor de notícias de futebol fornecem informações sobre a competição de futebol em alguns países, como Indonésia, Inglaterra, Espanha, Itália e Alemanha.

REFERÊNCIAS

[1] L. Sun e Y. Yin, “Descobrindo temas e tendências na pesquisa de transporte usando modelagem de tópicos”, *Transp. Res. Parte* C, v. 77, pág. 49–66, 2017.

[2] XP Zhang, XZ Zhou, HK Huang, Q. Feng, SB Chen e BY Liu, “Modelo de tópico para diagnóstico de medicina chinesa e análise de regularidades de prescrição: Caso sobre diabetes”, *Chin. J. Integr. Med.,* vol. 17, não. 4, pp. 307-313, 2011.

[3] S. Wang, MJ Paul e M. Dredze, “Exploring Health Topics in Chinese Social Media: An Analysis of Sina Weibo”, em *Workshops na Vigésima Oitava Conferência AAAI sobre Inteligência Artificial,* 2014, pp. 20– 23.

[4] L. Liu, L. Tang, W. Dong, S. Yao e W. Zhou, “Uma visão geral da modelagem de tópicos e suas aplicações atuais em bioinformática”, *Springerplus,* vol. 5, não. 1, pág. 1608, 2016.

[5] HG Yoon, H. Kim, CO Kim e M. Song, “Detecção de polaridade de opinião em dados do Twitter combinando regressão de encolhimento e modelagem de tópicos”, *J. Informetr.,* vol. 10, não. 2, pp. 634-644, 2016.

[6] AO Steinskog, JF Therkelsen e B. Gambäck, “Twitter Topic Modeling by Tweet Aggregation”, em *Proceedings of the 21st Nordic Conference of Computational Linguistics,* 2017, no. Maio, pp. 77-86.

[7] G. Lansley e PA Longley, "Computadores, Ambiente e Sistemas Urbanos A geografia dos tópicos do Twitter em Londres", *Comput. Ambiente. Sistema Urbano,* vol. 58, pp. 85-96, 2016.

[8] K. Sasaki, T. Yoshikawa e T. Furuhashi, “Modelo de tópico online para Twitter Considerando a dinâmica dos interesses do usuário e tendências de tópicos”, em *Anais da Conferência de 2014 sobre Métodos Empíricos em Processamento de Linguagem Natural (EMNLP),* 2014 , pp. 1977-1985.

[9] MC Yang e HC Rim, “Identificando conteúdos interessantes do Twitter usando análise tópica,” *Sistemas Especialistas com Aplicações,* vol. 41, nº. 9, Elsevier Ltd, pp. 4330-4336, 2014.

[10] KW Lim, C. Chen e W. Buntine, “Modelo de Tópico de Rede do Twitter: Um Tratamento Bayesiano Completo para Rede Social e Modelagem de Texto”, pp. 1–6, 2016.

[11] AF Hidayatullah e MR Ma'arif, “Pre-processing Tasks in Indonesian Twitter Messages,” in *IOP Conf. Série: Journal of Physics,* 2017, vol. 801.

[12] H. Jelodar, Y. Wang, C. Yuan e X. Feng, “Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey”, 2017.

[13] L. Bolelli, ÿ. Ertekin e CL Giles, “Detecção de tópicos e tendências em coleções de texto usando alocação de dirichlet latente”, *Lect. Computação de Notas. Sci. (incluindo Subser. Lect. Notas Artif. Intell. Lect. Notas Bioinformática),* vol. 5478 LNCS, pp. 776–780, 2009.

[14] DM Blei, AY Ng e MI Jordan, “Latent Dirichlet Allocation,” *J. Mach. Aprender. Res.,* vol. 3, pp. 993-1022, 2003.

[15] DM Blei, “Modelos de tópicos probabilísticos”, *Commun. ACM,* vol. 55, não. 4, pp. 77-84, 2012.

[16] C. Sievert e K. Shirley, “LDAvis: Um método para visualizar e interpretar tópicos”, em *Anais do Workshop sobre Aprendizagem de Linguagem Interativa, Visualização e Interfaces,* 2014, pp. 63–70.