2018 3rd International Conference on Computer and Communication Systems

# Modelagem de Tópicos no Twitter no Football News

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* - Com o desenvolvimento da mídia social nos dias de hoje, o Twitter se tornou a mídia social que é utilizada como um fornecedor de informações atuais sobre futebol. O futebol é o esporte mais popular na Indonésia. As pessoas sempre curiosas sobre algumas atualizações de notícias sobre futebol, tais como previsão de jogos, resultados de jogos, transferências, boatos, etc. Neste artigo, aplicamos modelos de tópicos para determinar o tópico dos tweets sobre notícias de futebol em Bahasa Indonésia. Os dados utilizados neste estudo foram retirados de várias contas oficiais indonésias no Twitter que sempre atualizam sobre o futebol e nós já selecionamos antes. A *Latent Dirichlet Allocation* (LDA) foi utilizada como método de modelagem de tópicos para determinar que tipo de tópicos no Twitter. De acordo com a análise de conteúdo, obtivemos vários tópicos de visão, tais como análise pré-jogo, atualização de jogos ao vivo, realizações de clubes de futebol, etc. Geralmente, os tópicos postados pela conta no Twitter do provedor de notícias de futebol dão informações sobre competições de futebol em alguns países como Indonésia, Inglaterra, Espanha, Itália e Alemanha.**

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

## I. INTRODUÇÃO

A mídia social tornou-se 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 publicadas a cada segundo nas mídias sociais. Junto com o desenvolvimento da mídia social hoje, o Twitter se tornou a mídia social que é usada como um fornecedor de informações atuais sobre futebol.

O futebol é o esporte mais popular na Indonésia. As pessoas estão sempre curiosas sobre algumas notícias de futebol atualizadas, como previsão de partidas, resultados de jogos, transferências, boatos, 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 principais da Europa. Existem quatro grandes ligas principais na Europa que as pessoas estão interessadas em obter atualizações, como a Premier League na Inglaterra, a Bundesliga alemã na Alemanha, a La Liga espanhola na Espanha e a Serie A na Itália. Portanto, é muito importante que os fornecedores de notícias esportivas compartilhem as atualizações de informações dessas ligas principais via Twitter. Para os cidadãos, o Twitter é a mídia mais rápida 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, temas quentes, etc. Para

978-1-5386-6350-9/18/$31.00 ©2018 IEEE

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 utilizados neste estudo foram retirados de várias contas oficiais indonésias no Twitter que sempre atualizam sobre o futebol e nós já selecionamos antes. Utilizamos a *Latent Dirichlet Allocation* (LDA) como método de modelagem de tópicos para determinar que tipo de tópicos no Twitter.

O restante deste documento 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 temática tem sido amplamente aplicada por pesquisadores em vários campos, incluindo a área de pesquisa em transporte [1], médica e saúde [2][3], bioinformática [4], política [5], etc. A modelagem de tópicos usando dados do Twitter também já foi conduzida por alguns pesquisadores antes. A modelagem tópica de dados de tweet tem seus próprios desafios em comparação com outros dados de texto devido a sua forma de linguagem não estruturada e tipo de linguagem não-padrão [6]. O método LDA tem sido 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 através do modelo de tópicos da LDA. Yang e Rim [9] propuseram um novo método para modelagem de tópicos chamado Trend Sensitive-Latent Dirichlet Allocation para extrair tópicos latentes do conteúdo, modelando as tendências temporais no Twitter ao longo do tempo. Lim, et al [10] propuseram o modelo temático Twitter-Network para modelar simultaneamente o texto e a rede social de uma forma totalmente Bayesiana e não paramétrica.

## III. METODOLOGIA

Nesta seção, descrevemos nossa metodologia que foi conduzida em nossa pesquisa.

### *A. Recuperação de dados*

O Twitter fornece uma API que permite que as pessoas reúnam os tweets. Esta pesquisa utiliza a API v1.1 do Twitter e a biblioteca GetOldTweets-python[1] para obter os tweets. As vantagens desta biblioteca em comparação com as outras do Twitter

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

A biblioteca é para coletar dados com base no intervalo de tempo que especificamos como desejado, fácil de usar, e os resultados dos dados são ordenados em formato csv.

Os dados utilizados neste estudo foram recuperados de contas confiáveis no Twitter indonésio que postaram sobre notícias de futebol. Essas contas incluem: @bolanet, @detiksport, @goal_id, @panditfootbal, @vivabola. Os dados totais obtidos dessas contas do Twitter são 120.639 tweets com um período de tempo de 1st de janeiro de 2017 a 24th de dezembro de 2017. A tabela 1 mostra o conjunto de dados desta pesquisa.

TABELA I. DATASET DO TWITTER

| Não | Conta no 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 na pesquisa anterior sobre tarefas de pré-processamento do Twitter [11]. As tarefas de pré-processamento nesta pesquisa são: dobrar casos, remover tags HTML e caracteres Unicode, remover símbolos e emoticons, remover caracteres não ASCII, remover caracteres especiais do Twitter, remover URLs, remover pontuações, remover números e remover palavras de parada.

### *C. Modelagem de tópicos usando LDA*

A modelagem tópica é um dos métodos mais poderosos na mineração de textos que visa identificar padrões e encontrar relação 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 eficaz de aprendizagem sem supervisão para encontrar diferentes tópicos em documentos de texto [13]. A modelagem de tópicos LDA é uma técnica não supervisionada na aprendizagem mecânica que foi introduzida pela primeira vez por Blei, et al [14] como um modelo probabilístico generativo para corpus textual.

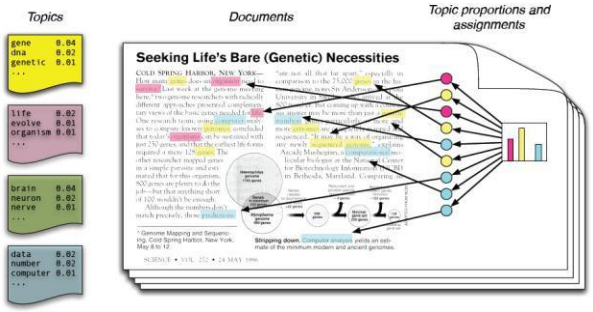


Figura 1. Modelo LDA [15].

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 onde cada tópico é uma distribuição sobre palavras ou vocabulário fixo, cada documento é uma mistura de tópicos de todo o corpo, e cada termo é retirado de um desses tópicos. O tópico é uma entidade que ilustra a relação entre palavras, como mostrado na Figura 1.

### *D. Visualização LDA*

O resultado do modelo tópico será visualizado usando a biblioteca Gensim e pyLDAvis em Python. A pyLDAvis é uma visualização de modelo tópico interativa baseada na web usando LDAvis que é construída a partir de LDAvis usando uma combinação de R e D3 [16]. Usando a pyLDAvis, temos permissão para explorar a relação entre tópico e termos para entender o modelo LDA. PyLDAvis tem dois painéis, o mapa de distribuição de cada tópico e o gráfico de intensidade que representa os termos mais freqüentes no corpus.

## IV. RESULTADO E DISCUSSÃO

### *A. Análise de visualização LDA*

Esta seção discute o resultado de nossas experiências. Para conduzir a modelagem temática da LDA, usamos a biblioteca de modelos LdaModel fornecida pela biblioteca Gensim em Python. Nós escolhemos dez tópicos como parâmetro. A visualização do mapa de distância intertópico de nosso modelo é mostrada na figura 2.

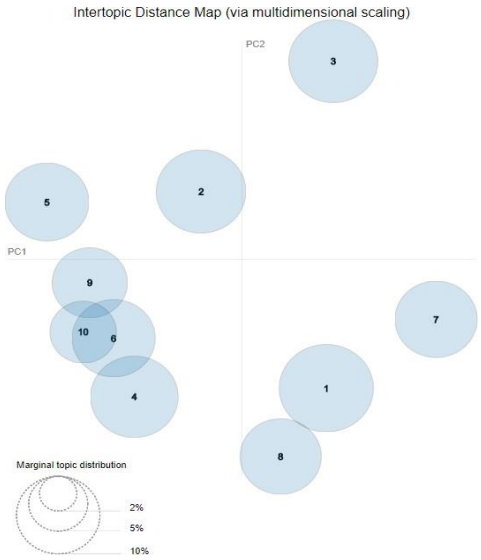


Figura 2. Visualização do Mapa Intertópico de Distância.

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

A figura 3 mostra as 30 palavras mais salientes do corpus. Podemos ver também que há cinco termos que

freqüentemente apareceram no corpus como "MU", "Madrid", "liga", "Chelsea" e "vs". Isto indica que o Manchester United, o Real Madrid e o Chelsea são tópicos a serem discutidos com bastante freqüência nas notícias. Além disso, as palavras "liga" e "vs" indicam o tópico sobre jogos de futebol na liga.

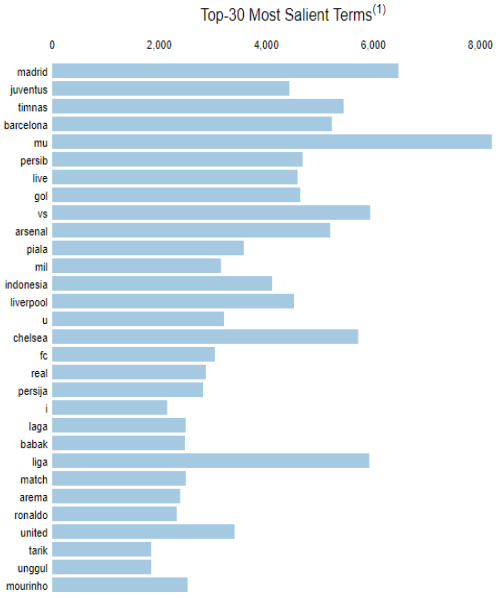


Figura 3. Top-30 Termos Mais Salientes.

## *B. Análise Tópica*

A visualização de dados de cada tópico é ilustrada usando a palavra nuvem. A nuvem de palavras é uma visualização composta de palavras em um dado de texto particular. A nuvem de palavras mostra a freqüência com que as palavras aparecem em uma coleção de textos. O tamanho de cada palavra indica sua importância, o que significa que quanto maior o tamanho da palavra, mais freqüente a palavra em um tópico. Além disso, as palavras que dominam a nuvem de palavras provavelmente estão diretamente relacionadas com o 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. Há 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.

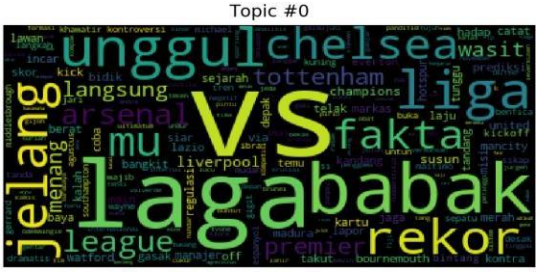


Figura 4. Visualização da Nuvem de Palavras do Tema #0.

- Tópico #1: Atualização do jogo ao vivo na Liga 1

  A figura 5 mostra a palavra nuvem do tópico nº 1. Encontramos a palavra "ao vivo", "fósforo" e "persas" como o tópico mais dominante. Outras palavras como Arema, Persiba, PSM, e Bhayangkara indicam o clube de futebol da Liga Indonésia. A partir dessas palavras dominantes, pode-se concluir que o tópico é sobre jogos ao vivo na liga de futebol indonésia (Liga 1).

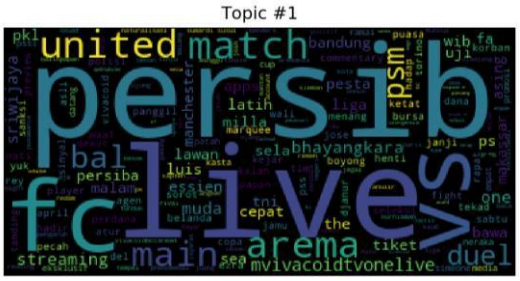


Figura 5. Visualização da nuvem de palavras do tópico nº 1.

- Tópico #2: Premiere League Inglês

  A palavra visualização de nuvem do tópico nº 2 é mostrada na figura 6 abaixo. As palavras Chelsea, Liverpool, City, ManCity, MU são dominantes neste segmento do tópico. Além disso, há também alguns gerentes Premier em inglês como Conte, Guardiola, e Klopp. De acordo com essas palavras, pode-se concluir que o tópico nº 2 fala sobre a Premier League inglesa.

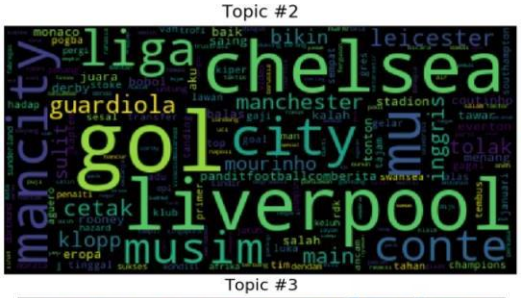


Figura 6. Visualização da nuvem de palavras do tópico nº 2.

- Tema #3: Rivalidade entre a MU e o Arsenal

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


Figura 7. Visualização da nuvem de palavras do tópico nº 3.

- Tema #4: A equipe nacional da Indonésia e a Liga Indonésia

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

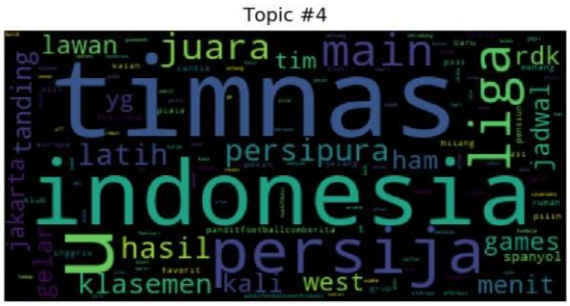


Figura 8. Visualização da Nuvem de Palavras do tópico #4.

- Tema #5: Serie A Italia

  A figura 9 mostra a palavra nuvem do tópico#5. A partir dos termos da palavra nuvem, pode-se concluir que o tópico correspondente do tópico #5 é sobre a Série A Itália.


Figura 9. Visualização da Nuvem de Palavras do Tema #5.

- Tópico #6: Copa do Mundo

  As palavras no tópico #6 tratam da copa do mundo que pode ser vista a partir de duas palavras dominantes na palavra nuvem, *"piala"* e *"dunia"*. A palavra nuvem do tópico #6 é mostrada pela figura 10.

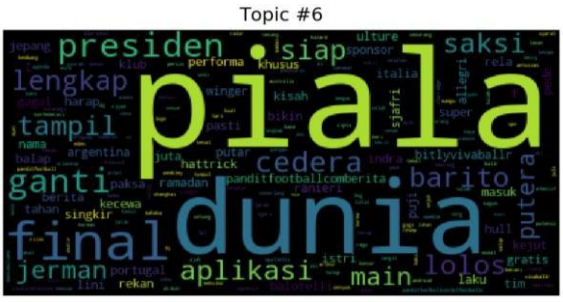


Figura 10. Visualização da nuvem de palavras do tópico #6.

- Tema #7: El Classico (Real Madrid vs Barcelona)

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

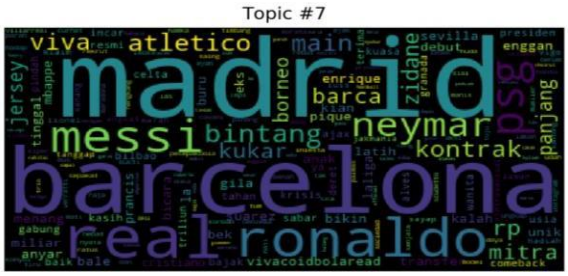


Figura 11. Visualização da nuvem de palavras do tópico #7.

- Tema #8: Indonésia Futebol e Sêmen Padang

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

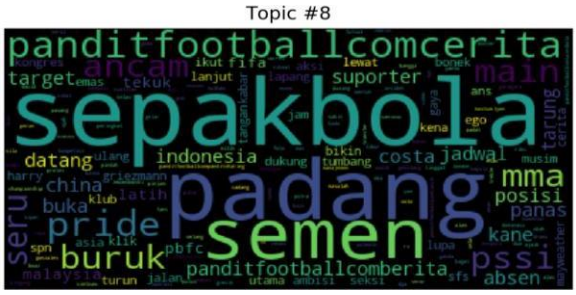


Figura 12. Visualização da nuvem de palavras do tópico nº 8.

- Tema #9: A conquista de alguns grandes clubes de futebol na Europa

  O tópico nº 9 diz respeito à realização de alguns grandes clubes na Europa. Pode ser visto pelas palavras dominantes em sua palavra nuvem, como os nomes dos clubes como Bayern, AC (significa AC Milan), Mil (possivelmente a origem da palavra Milan afetada pelo processo de criação), Inter, Chelsea.

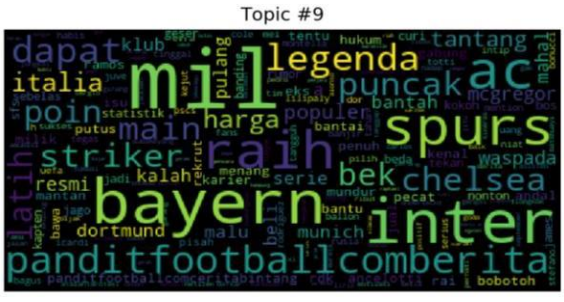


Figura 13. Visualização da nuvem de palavras do tópico #9.

## V. CONCLUSÃO

Este artigo explorou o uso de modelos atuais a serem aplicados às mensagens do Twitter que falam sobre futebol usando o método de *alocação Latent Dirichlet.* De acordo com a análise de conteúdo, obtivemos vários tópicos esclarecedores, tais como análise pré-jogo, atualização de jogos ao vivo, realizações de clubes de futebol, etc. Geralmente, os tópicos postados pela conta no Twitter do provedor de notícias de futebol dão informações sobre a competição futebolística em alguns países como Indonésia, Inglaterra, Espanha, Itália e Alemanha.

## REFERÊNCIAS


[1] L. Sun e Y. Yin, "Descobrindo temas e tendências na pesquisa de transportes usando a modelagem de tópicos", *Transp. Res. Parte C*, vol. 77, pp. 49- 66, 2017.

[2] X. P. Zhang, X. Z. Zhou, H. K. Huang, Q. Feng, S. B. Chen, e B. Y. Liu, "Modelo tópico para diagnóstico da medicina chinesa e análise das regularidades de prescrição": Case on diabetes", *Chin. J. Integr. Med.* , vol. 17, no. 4, pp. 307-313, 2011.

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

[4] L. Liu, L. Tang, W. Dong, S. Yao, e W. Zhou, "An overview of topic modeling and its current applications in bioinformatics", *Springerplus*, vol. 5, no. 1, p. 1608, 2016.

[5] H. G. Yoon, H. Kim, C. O. Kim, e M. Song, "Opinion polarity detection in Twitter data combining shrinkage regression and topic modeling", *J. Informetr.* , vol. 10, no. 2, pp. 634-644, 2016.

[6] A. O. Steinskog, J. F. 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 P. A. Longley, "Computers , Environment and Urban Systems The geography of Twitter topics in London", *Comput. Ambiente. Urban Syst.* , vol. 58, pp. 85-96, 2016.

[8] K. Sasaki, T. Yoshikawa, e T. Furuhashi, "Online Topic Model for Twitter Considering Dynamics of User Interests and Topic Trends", em *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1977-1985.

[9] M. C. Yang e H. C. Rim, "Identifying interesting Twitter contents using topical analysis", *Expert Systems with Applications*, vol. 41, no. 9, Elsevier Ltd, pp. 4330-4336, 2014.

[10] K. W. Lim, C. Chen, e W. Buntine, "Twitter-Network Topic Model": Um tratamento Bayesiano completo para redes sociais e modelagem de texto".
pp. 1–6, 2016.

[11] A. F. Hidayatullah e M. R. Ma'arif, "Tarefas de pré-processamento em Mensagens do Twitter Indonésio", no *OIP Conf. Série IOP: 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, Ş. Ertekin, e C. L. Giles, "Topic and trend detection in text collections using latent dirichlet allocation", *Lect. Notas Informática. Sci. (incluindo Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5478 LNCS, pp. 776-780, 2009.

[14] D. M. Blei, A. Y. Ng, e M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Aprenda. Res.* , vol. 3, pp. 993-1022, 2003.

[15] D. M. Blei, "Probabilistic topic models", *Commun. ACM*, vol. 55, no. 4, pp. 77-84, 2012.

[16] C. Sievert e K. Shirley, "LDAvis: A Method for Visualizing and Interpreting topics", em *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 2014, pp. 63-70.