Xplore (2021), Vol. 10, No. 03, 249-258
DOI: 10.29244/xplore.v10i3.836

Xplore: Journal of Statistics
Available at https://journal.stats.id/index.php/xplore
e-ISSN: 2655-2744

Latent Dirichlet Allocation dalam Identifikasi Respon Masyarakat Indonesia terhadap Covid-19 Tahun 2020-2021*

Karel Fauzan Hakim[1], Pika Silvianti[2‡], Agus Mohamad Soleh[3]

[1,2,3]Department of Statistics, IPB University, Indonesia
‡corresponding author: pikasilvianti@apps.ipb.ac.id

Abstract

Covid-19 is a very troubling disease in Indonesia. Therefore, understanding public opinion is required to find solutions and evaluate the government performance in handling the pandemic. Twitter can be helpful to identify the public opinion of significant events. Twitter's tweet is a large dimension text-based big data. It requires text sampling and text mining to be processed efficiently and effectively. Stratified random sampling with 20 repetitions applied to assume days as strata followed by topic modeling with latent Dirichlet allocation (LDA). This research aims to find out public opinion regarding Covid-19 and its growth over time. Other than that, this research also aims to find out sampling effects on tweet data using stratified random sampling. Therefore, the extracted topics will be transformed into time-series data and considering the variety of the pattern made. Afterward, the transformation results will be explored and interpreted. This research suggests that discussions related to Covid-19 are divided into four topics by the first model, namely: "Vaccine", "Positive or affected people", "Health protocol", and "Indonesia" then nine topics by the second model, namely: "Vaccine", "Prayer", "Health protocol", "Social aid and corruption", "Affected people", "Indonesian economy", "Work", "Persuading to wear mask", and "Willing to watch". Furthermore, some topics peak whenever a significant event occurs in Indonesia. Afterward, this research suggests that 20 repetitions of stratified random sampling could provide good results.

Keywords: Covid, Latent Dirichlet Allocation, text mining, text sampling, Twitter.

* Received: July 2021; Reviewed: Aug 2021; Published: Sep 2021

1. Pendahuluan

Covid-19 atau *Corona virus desease* 2019 merupakan sebuah penyakit yang sangat meresahkan bagi masyarakat Indonesia. Hal ini disebabkan oleh tingginya tingkat penyebaran penyakit tersebut (Settersten Jr, 2020). Memahami pendapat masyarakat Indonesia sangat dibutuhkan untuk mengevaluasi kinerja pemerintah dalam penanganan Covid-19. Dengan memahami pendapat masyarakat, pemerintah juga dapat mencari solusi terbaik untuk menghadapi pandemi ini. Oleh karena itu, perlu dilakukan sebuah analisis untuk mengetahui pendapat masyarakat Indonesia yang berkembang seiring peningkatan kasus harian Covid-19.

Data *tweet* dari Twitter dapat memberikan informasi berupa opini masyarakat terhadap sebuah fenomena (Bollen *et al.* 2011, Cody *et al.* 2015, Gurajala *et al.* 2019). Bollen *et al.* (2011) menganalisa opini Twitter untuk memprediksi pasar saham. Cody *et al.* (2015) menggunakan data *tweet* untuk mengetahui sentimen publik terhadap perubahan iklim. Gurajala *et al.* (2019) menggunakan data *tweet* untuk mengetahui respon publik terhadap kualitas udara di New Delhi. Data *tweet* dari Twitter merupakan *big data* yang tidak terstruktur dalam bentuk teks. Oleh karena itu, penarikan contoh dan *topic modeling* akan diaplikasikan untuk menganalisis data tersebut.

Menurut Webb dan Wang (2013), penarikan contoh penting dilakukan untuk mempelajari sebuah media sosial yang memiliki jumlah data yang sangat besar dan terus bertambah. Penarikan contoh merupakan memilih dan mengobservasi bagian dari sebuah populasi untuk mengestimasi sesuatu tentang populasi tersebut (Thompson 2012 dalam Webb dan Wang 2013). Penelitian ini bertujuan untuk mengetahui perkembangan pendapat seiring waktu. Oleh karena itu, penarikan contoh dilakukan berdasarkan periode seperti yang dilakukan oleh Bollen *et al.* (2011) dalam Webb dan Wang (2013).

Menurut Blei *et al.* (2003) dalam Xie dan Xing (2013), *topic modeling* adalah model yang dibentuk untuk memodelkan teks dan mendeteksi topik yang belum diketahui dari sebuah dokumen. Pada penelitian ini, *latent* Dirichlet *allocation* (LDA) digunakan untuk melakukan *topic modeling*. LDA adalah model yang dibentuk untuk mencari topik di dalam dokumen dan peluang kemunculan kata dalam sebuah topik. Model ini menghasilkan keluaran berupa campuran topik di dalam sebuah dokumen yang menyebar Dirichlet. Penelitian sebelumnya yang menggunakan LDA untuk memodelkan topik dari media sosial adalah Gurajala *et al.* (2019) dan Han *et al.* (2020). Gurajala *et al.* (2019) memodelkan LDA dengan data Twitter untuk mengidentifikasi respon masyarakat New Delhi terhadap polusi udara sedangkan Han *et al.* (2020) memodelkan LDA dengan data Weibo untuk mengidentifikasi respon masyarakat China terhadap Covid-19.

Penelitian sebelumnya mengemukakan bahwa beberapa topik yang memiliki perkembangan dari waktu ke waktu bersangkutan pada kejadian nyata yang terjadi. Gurajala *et al.* (2019) menyatakan bahwa topik "*air policy*" dan "*health*" memiliki perkembangan yang paling serupa dengan perkembangan kadar polusi udara dari waktu ke waktu. Han *et al.* (2020) menyatakan bahwa topik berkembang tergantung stadium Covid-19 yang terjadi. Data deret waktu ini merupakan data jumlah peluang sebuah topik per harinya sehingga jumlah data deret waktu yang terbentuk adalah sejumlah topik yang terbentuk. Data deret waktu ini kemudian akan dieksplorasi

berdasarkan pola yang terbentuk dan dianalisis berdasarkan kejadian yang terjadi seiring berjalannya waktu.

2. Metodologi

2.1 Data

Data yang digunakan pada penelitian ini merupakan data yang diperoleh dengan *scraping* Twitter dengan *package* “Rtweet” serta *keyword* “covid”, “pandemi”, atau “corona” pada periode 25 November 2020 hingga 5 Februari 2020 sejumlah 1.716.920 *tweet* dari 582.959 akun. “Rtweet” melakukan *scraping* berdasarkan bahasa yang digunakan dalam *tweet* yaitu bahasa Indonesia. Namun, penggunaan fitur tersebut juga menyebabkan beberapa kata asing yang serupa dengan bahasa Indonesia tertarik. Akan tetapi, permasalahan ini dapat ditangani oleh LDA karena metode ini dapat mengelompokkan bahasa asing menjadi sebuah topik lain. Selain itu, *tweet* yang didapatkan tidak mencakup re-*tweet* serta tidak menghiraukan jumlah *re-tweet*, *quote*, dan *like*. Data Twitter dirincikan pada Tabel 1.

Tabel 1 Rincian peubah data pada Twitter.

Nama peubah	Keterangan
created_at	Waktu unggah *tweet*
screen_name	Username Twitter
text	Isi *tweet*

2.2 Prosedur Analisis Data

Analisis data dilakukan dengan menggunakan perangkat lunak R Studio dengan pemrograman R versi 3.6.3 menggunakan paket “tm”, “textclean”, “textmineR”, “katadasaR”, “tokenizers”, “wordcloud”, “dplyr”, “dygraph”, dan “ggplot2”. Tahapan analisis pada penelitian ini diilustrasikan dan dijelaskan sebagai berikut:

1. Data yang diperoleh dari *scraping* Twitter kemudian diseleksi dengan menghapus akun berita, akun *buzzer*, dan akun bot.
2. Melakukan *stratified random sampling* terhadap data yang terseleksi dengan asumsi data heterogen antar periode/hari sebanyak 20 kali. Jumlah *tweet* pada setiap contoh ditentukan dengan rumus Slovin.
3. Melakukan pra proses pada masing-masing contoh dengan menghapus url, mengubah huruf besar menjadi huruf kecil, normalisasi kata yang tidak baku, menghapus karakter selain alfabet, dan menghapus *stopwords*.
4. Memodelkan data dengan LDA pada masing-masing contoh menggunakan dua pemodelan dengan penentuan jumlah topik yang berbeda. Langkah-langkah pemodelan LDA dijelaskan sebagai berikut:
    a. Membentuk *Bag of Words* (BoW) (Deepu *et al.* 2016) per kata dari data *tweet*. Kata diasumsikan sebagai satu kosakata atau dua kosakata (frasa).
    b. Memodelkan LDA dengan Gibbs *sampling* pada setiap kata dalam data. Algoritma Gibbs *sampling* dijelaskan sebagai berikut:

i. Labelkan setiap kata di setiap *tweet* dengan sebuah topik secara acak.
ii. Pada kata ke-*i* dalam *tweet* ke-*k*, keluarkan kata tersebut dari model BoW lalu hitung: (Griffiths dan Steyvers 2004)

$$P(z_i = j, x_i = k | w_i = m, z_{-i}, x_{-i}) = \frac{C_{mj}^{WT} + \beta}{\sum_{n=1}^{V} C_{nj}^{WT} + V\beta} \frac{C_{kj}^{DT} + \alpha}{\sum_{n=1}^{T} C_{kn}^{DT} + T\alpha}$$

Dengan:

$(z_i = j, x_i = k)$	:	penempatan kata ke-*i* pada topik ke-*j* dan *tweet* ke-*k*
$w_i = m$	:	observasi dimana kata ke-*i* adalah kata ke-*m* di dalam daftar kata
z_{-i}, x_{-i}	:	seluruh penempatan topik dan *tweet* tanpa termasuk kata ke-*i*
C_{mj}^{WT}	:	jumlah kemunculan kata ke-*m* ditempatkan pada topik ke-*j* tanpa termasuk kata ke-*m*
C_{kj}^{DT}	:	jumlah kemunculan topik ke-*j* pada *tweet* ke-*k* tanpa termasuk kata ke-*m*
α	:	parameter sebaran Dirichlet distribusi topik terhadap *tweet*
β	:	parameter sebaran Dirichlet distribusi kata terhadap topik
V	:	jumlah kata
T	:	jumlah topik

iii. Tempatkan kata ke-*i* ke topik yang memiliki nilai $P(z_i = j, x_i = k | w_i = m, z_{-i}, x_{-i})$ terbesar.
iv. Ulangi langkah ii dan iii pada setiap kata di seluruh *tweet*.
v. Ulangi langkah iv sebanyak iterasi yang diinginkan.

c. Untuk pemodelan pertama, hitung maksimum lokal *topic coherence* dengan inisiasi 5 topik. Nilai *coherence* dirumuskan sebagi berikut: (Jones 2019)

$$C\left(t; V^{(t)}\right) = \frac{\sum_{m=2}^{M} \sum_{l=1}^{m-1} P\left(v_l^{(t)} \middle| v_m^{(t)}\right) - P(v_m^{(t)})}{\frac{1}{2} M(M-1)}$$

dengan:

$V^{(t)} = (v_1^{(t)}, \ldots, v_M^{(t)})$	:	sebuah vektor *M* kata terbaik di dalam topik *t*.
$P\left(v_l^{(t)} \middle| v_m^{(t)}\right)$	:	peluang kemunculan kata ke-l di seluruh kemunculan kata ke-m pada seluruh dokumen
$P(v_m^{(t)})$	:	peluang kemunculan kata ke-m pada seluruh dokumen

Sedangkan *topic coherence* didapatkan dari nilai tengah dari seluruh $C\left(t; V^{(t)}\right)$.

d. Untuk pemodelan kedua, lakukan langkah b terhadap seluruh contoh dengan jumlah topik 20.

5. Memilah sebaran topik berdasarkan kemiripan sehingga didapatkan kumpulan topik yang bermakna serupa dari setiap contoh pada kedua model. Pemilahan topik dilakukan dengan menyaring sebuah topik dengan makna tertentu dari setiap contohnya dan mengumpulkannya.
6. Mentransformasikan kumpulan topik menjadi data deret waktu. Hal ini dilakukan dengan menjumlahkan peluang setiap topik pada setiap *tweet* per harinya sehingga didapatkan data deret waktu kemunculan topik setiap harinya.
7. Mengeksplorasi dan menginterpretasikan data deret waktu dengan melihat pola, nilai tengah, median, dan keragaman dari jumlah *tweet* per harinya pada setiap contoh.

3. Hasil dan Pembahasan

3.1 Seleksi *Tweet*

Data *tweet* yang didapatkan masih mengandung *tweet* yang bukan merupakan opini. Oleh karena itu perlu dilakukan penyeleksian agar tidak menimbulkan bias pada hasil analisis. Penyeleksian dilakukan dengan membuang akun yang telah mengunggah *tweet* dengan *keyword* covid, pandemi, atau corona lebih dari 20 kali selama periode yang ditetapkan dan seluruh *tweet* tersebut lebih dari 50% mengandung *url* atau lebih dari 50% mengandung *hashtag* atau lebih dari 70% mengandung kata “khofifah”. Kriteria ini ditetapkan berdasarkan hasil pengamatan terhadap akun yang dipercaya merupakan akun bot, *buzzer*, dan berita. Selain itu, akun yang mengandung kata “covid”, “corona”, “berita”, “fm”, “idntimes”, “news”, “humas” juga dibuang. Penyeleksian tersebut menyisakan 1.245.638 *tweet* dari 575.462 akun.

3.2 Stratified Random Sampling

Data *tweet* sejumlah 1.245.638 masih terlalu berat untuk dilakukan komputasi. Oleh karena itu, akan diaplikasikan *stratified random sampling* (Scheaffer *et al.* 2011) dengan mengasumsikan data *tweet* heterogen antar periode atau hari. Data *tweet* diasumsikan heterogen antar hari karena perbincangan masyarakat terhadap sebuah akan berbeda-beda pada setiap waktunya.

Jumlah contoh yang ditarik pada setiap strata adalah 28.144. Angka ini ditentukan berdasarkan rumus Slovin (Israel 1992) dengan *margin of error* = 0.05 dan N = jumlah *tweet* per hari dengan algoritma sebagai berikut:
Untuk setiap hari (i) dalam tanggal 25 November 2020 hingga 5 Februari 2021, hitung:

$$n_i = \frac{N_i}{1 + N_i(0.05)^2}$$

dengan:
n_i : jumlah contoh hari ke-i
N_i : jumlah *tweet* hari ke-i
Kemudian hitung:

$$Jumlah\ contoh = \sum n_{(i)}$$

Namun, satu contoh tersebut belum tentu dapat mewakilkan seluruh *tweet*. Oleh karena itu, akan diulang proses penarikan contoh sebanyak 20 kali sehingga didapatkan contoh sebanyak 28.144 × 20 *tweet*. Seluruh contoh tersebut kemudian akan dibandingkan untuk mendapatkan hasil yang lebih akurat.

3.3 Pra Proses

Melakukan pra proses adalah hal yang penting dalam pengolahan data teks. Hal ini disebabkan teks yang tidak ter-pra proses dapat mengurangi efisiensi komputasi serta menghasilkan hasil yang tidak bermakna (Vijayarani *et al.* 2015). Oleh karena itu, pra proses dilakukan pada setiap contoh dengan menghapus *url*, mengubah huruf besar menjadi huruf kecil, normalisasi kata menjadi bentuk baku, menghapus karakter selain alfabet, dan menghapus *stopwords*.

3.4 *Latent* Dirichlet *Allocation*

LDA diterapkan pada setiap contoh yang terbentuk pada proses sebelumnya. LDA dimulai dengan mentransformasikan setiap *tweet* ke dalam model BoW untuk setiap satu kosakata dan dua kosakata. Model BoW tersebut kemudian dimodelkan ke dalam LDA dengan Gibbs *sampling*. Griffiths dan Steyvers (2004) menawarkan $\alpha = 50 \,/\, (\text{Jumlah topik})$ dan $\beta = 200 \,/\, (\text{Jumlah kosakata})$. Namun, kedua parameter tersebut digunakan untuk penelitiannya yang menggunakan jumlah topik lebih dari 50 dan lebih dari puluhan ribu kata sehingga parameter tersebut perlu disesuaikan (Allen dan Xiong 2012). Penelitian ini menggunakan jumlah topik kurang dari atau sama dengan 20 dan memiliki puluhan ribu kata. Putri dan Kusumaningrum (2017) menyatakan bahwa $\alpha = 0.1$ merupakan pilihan yang baik selain $\alpha = 50 \,/\, (\text{Jumlah topik})$. Oleh karena itu penelitian ini menggunakan parameter $\alpha = 0.1$ dan $\beta = 200 \,/\, (\text{Jumlah kosakata})$ pada setiap contohnya.

Untuk mengeksplorasi hasil yang lebih luas, LDA dilakukan dengan dua pemodelan yang berbeda. Kedua pemodelan tersebut menggunakan parameter yang sama di setiap contohnya tetapi menggunakan penentuan jumlah topik berbeda, yaitu berdasarkan maksimum lokal pertama pada grafik *topic coherence* setiap contohnya pada pemodelan pertama dan jumlah topik sejumlah 20 pada pemodelan kedua. Perhitungan *topic coherence* yang dilakukan menggunakan M = 5 karena *package* "textmineR" menetapkan M = 5 sebagai default dan tidak dapat dirubah. Kemudian, tidak semua topik yang terbentuk memiliki makna yang relevan. Oleh karena itu, topik-topik yang memiliki makna yang relevan dan terdapat di setiap contoh akan dipilih. Pemilahan topik dilakukan secara subjektif dengan cara mengamati topik yang terbentuk pada setiap contohnya lalu menyaring dan mengelompokkan topik yang memiliki makna serupa dari seluruh contoh.

Pemodelan pertama menginisiasi perhitungan jumlah topik dari 5 topik. Hal ini disebabkan pembagian menjadi 4 topik atau kurang masih terlalu umum untuk dideskripsikan. Topik yang umum tersebut dibuktikan dengan sebaran kata yang merata atau tidak mengerucut pada sebuah kata. Pada pemodelan ini, setiap contoh memiliki jumlah topik yang beragam. Contoh-contoh tersebut memiliki jumlah topik dengan kisaran lima hingga sembilan topik. Pemodelan tersebut mengidentifikasi

empat topik yang relevan dan terdapat di setiap contoh, yaitu: “Vaksin”, “Orang positif atau terkena covid”, “Protokol kesehatan”, dan “Indonesia”.

Pemodelan Kedua menggunakan 20 topik seperti yang dilakukan Han *et al.* (2020). Mereka mengobservasi topik-topik yang terbentuk dan membuang topik-topik tidak relevan. Pemodelan kedua mengidentifikasi 9 topik yang relevan, yaitu: “Vaksin”, “Doa”, “Protokol Kesehatan”, “Bantuan sosial dan korupsi”, “Orang terkena Covid”, “Ekonomi Indonesia”, “Kerja”, “Ajakan memakai masker”, dan “Keinginan untuk menonton”.

3.5 Eksplorasi Deret Waktu

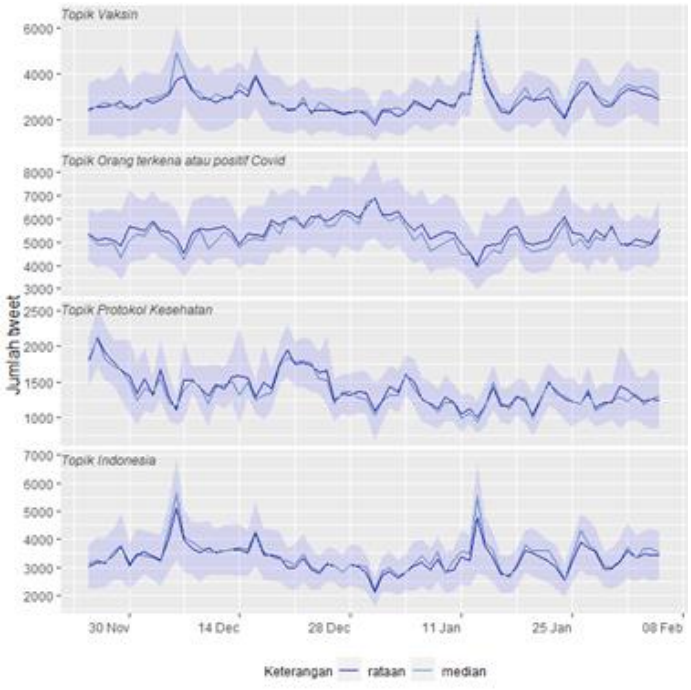
Kedua pemodelan LDA meringkas *tweet* ke dalam bentuk distribusi topik. Penelitian ini bertujuan untuk melihat perkembangan dari setiap topik dari waktu ke waktu dengan menjumlahkan distribusi peluang topik pada seluruh *tweet* per harinya dan mentransformasikan ke dalam bentuk deret waktu sehingga didapatkan 20 data deret waktu per topiknya. Proses transformasi dijelaskan sebagai berikut:
Untuk seluruh *tweet* (*i*) dalam hari ke-*j* hitung:

$$T(T_k; H_j; S_l) = \sum_i P(T_k; H_j; S_l; D_i)$$

dengan:

$T(T_k; H_j; S_l)$: Total peluang topik ke-k, hari ke-j, dan contoh ke-l
$P(T_k; H_j; S_l; D_i)$: Peluang topik ke-k, hari ke-j, contoh ke-l, dan *tweet* ke-i.

Sehingga didapatkan 20 data deret waktu untuk setiap topiknya. Namun, menampilkan 20 deret waktu sekaligus akan menyulitkan proses eksplorasi. Oleh karena itu, akan dihitung rata-rata, median, dan ragam dari $T(T_k; H_j; S_l)$.



Gambar 1 Perkembangan topik pada pemodelan pertama

Berdasarkan Gambar 1, terdapat *peak* pada tanggal 6 Desember 2020 di topik “Indonesia”. *Peak* lain terjadi pada tanggal 13 Januari 2021 di topik “Vaksin” dan “Indonesia”. Topik “Orang terkena atau positif Covid” dan “Protokol Kesehatan” tidak memiliki fluktuasi yang signifikan. Selain itu, topik tersebut tampak memiliki ragam yang besar diikuti oleh topik “Vaksin” dan “Indonesia”.

Gambar 2 Perkembangan topik pada pemodelan kedua

Berdasarkan Gambar 2, terdapat *peak* kecil pada tanggal 1 Desember 2020 di topik “Doa”. Kemudian, *peak* besar muncul pada tanggal 6 Desember 2020 di topik “Bantuan sosial dan korupsi”. Setelah itu, *peak* kecil lain terlihat pada tanggal 7 Desember 2020 dan 16 Desember 2020 di topik “Vaksin”. Pada tanggal 21 Desember 2020 topik “Doa” perlahan naik, mengalami penurunan sedikit pada tanggal 26 Desember 2020, hingga akhirnya memuncak pada 1 Januari 2021. Kemudian, *peak* besar lain muncul pada tanggal 13 Januari 2021 pada topik “Vaksin”. *Peak* besar tersebut diikuti oleh *peak* kecil lainnya pada tanggal 27 Januari 2021 dan 2 Februari 2021 di topik yang sama. Topik-topik lain seperti “Protokol Kesehatan”, “Orang positif Covid”, “Ekonomi Indonesia”, “Kerja”, “Ajakan memakai masker”, dan “Keinginan selama pandemi” tidak memiliki fluktuasi yang signifikan. Selain itu, topik “Orang positif covid”, “Kerja”, dan “Keinginan selama pandemi” tampak memiliki keragaman yang besar.

3.6 Pembahasan

Pengulangan *stratified random sampling* memberikan hasil yang cukup baik. Hal ini dibuktikan dengan didapatkannya beberapa topik yang memiliki keragaman pola minimum dalam pemodelan kedua. Keragaman minimum pada topik menunjukan bahwa pola perkembangan topik tersebut konsisten dalam 20 kali pengulangan. Pengulangan penarikan contoh berpengaruh pada jumlah topik yang terpilah pada

proses pemilahan. Pemilahan topik pada penelitian ini mengumpulkan topik yang memiliki kemunculan sebanyak 20 atau muncul di seluruh pengulangan. Oleh karena itu, topik-topik yang memiliki peluang kemunculan kecil pada populasi berpeluang lebih besar untuk tidak terpilah dan tidak dilanjutkan dalam proses analisa dengan memperbanyak pengulangan. Namun, topik-topik yang berpeluang besar untuk muncul dalam populasi akan berpeluang besar untuk muncul pada setiap pengulangannya. Untuk membuktikan hal tersebut, penelitian selanjutnya sangat diharapkan untuk menguji jumlah pengulangan penarikan contoh pada pemodelan topik.

Perbincangan terhadap topik terkait Covid akan memuncak apabila terdapat suatu kejadian besar di Indonesia. Hal ini didukung dengan beberapa topik yang memuncak pada tanggal tertentu. Sebagai contoh, perbincangan terhadap topik "Vaksin" pada kedua pemodelan memuncak pada 13 Januari 2021. Pada tanggal tersebut Presiden Joko Widodo divaksin untuk pertama kalinya dan hal tersebut sempat menjadi pembicaraan hangat di Indonesia. Selain itu, topik "Doa" juga perlahan memuncak pada akhir Desember 2020 hingga tahun baru 2021. Hal ini disebabkan masyarakat yang menuliskan doa dan harapan mereka terkait Covid-19.

4. Simpulan dan Saran

4.1 Simpulan

Stratified random sampling yang diulang 20 kali dapat menunjukkan hasil yang cukup baik. Perbincangan terkait Covid-19 di Indonesia dapat dibagi menjadi empat topik menurut pemodelan pertama dan sembilan topik menurut pemodelan kedua. Empat topik pemodelan pertama meliputi: "Vaksin", "Orang positif atau terkena covid", "Protokol kesehatan", dan "Indonesia". Sembilan topik pemodelan kedua meliputi: "Vaksin", "Doa", "Protokol Kesehatan", "Bantuan sosial dan korupsi", "Orang terkena Covid", "Ekonomi Indonesia", "Kerja", "Ajakan memakai masker", dan "Keinginan untuk menonton". Eksplorasi terhadap deret waktu mengemukakan kuantitas perbincangan tentang topik terkait Covid-19 di Indonesia mengalami peningkatan yang signifikan pada setiap kejadian besar di Indonesia.

4.2 Saran

Berdasarkan hasil penelitian, ada beberapa hal yang dapat dijadikan saran yaitu:

1. Penelitian selanjutnya diharapkan menguji pengaruh pengulangan penarikan contoh terhadap hasil pemodelan topik.
2. Penelitian selanjutnya diharapkan menguji pengaruh parameter LDA ($\alpha, \beta,$ iterasi, dan jumlah topik) terhadap pola serta keragaman deret waktu yang terbentuk.

Daftar Pustaka

Allen TT, Xiong H. 2012. Pareto charting using multifield freestyle text data applied to toyota camry user reviews. Applied Stochastic Models in Business and Industry. 28: 152-163. DOI: 10.1002/asmb.947

Bollen J, Mao H, Zeng XJ. 2011. Twitter mood predicts the stock market. Journal of Computational Science. 2(1): 1–8. DOI:10.1016/j. jocs.2010.12.007.

Cody EM, Reagan AJ, Mitchell L, Dodds PS, Danforth CM. 2015. Climate change sentiment on twitter: an unsolicited public opinion poll. Plos One. 10(8). DOI: 10.1371/e0136092

Deepu S, Raj P, Rajaraajeswari S. 2016. A framework for text analytics using the bag of words model for prediction. International Journal Advanced Networking & Applications (IJANA).

Griffiths TL, Steyvers M. 2004. Finding scientific topics. Proceeding of the National Academy of Science. 101:5228-5235.

Gurajala S, Dhaniyala S, Matthews JN. 2019. Understanding public response to air quality using *tweet* analysis. Social Media + Society. 1-14. DOI: 10.1177/2056305119867656.

Han X, Wang J, Zhang M, Wang X. 2020. Using social media to mine and analyze public opinion related to COVID-19 in China. International Journal of Enviromental Research and Public Health. 17(8): 2788. DOI: 10.3390/ijerph17082788

Israel, GD. 1992. Determining sample size. University of Florida Cooperative Extension Service, Institute of Food and Agriculture Sciences, EDIS, Florida.

Jones TW. 2019. Topic Modelling. Cran R Project. Diunduh 2021 Jun 25. https://cran.r-project.org/web/*packages*/textmineR/vignettes/c_topic_modeling.html

Putri IR, Kusumaningrum R. 2017. Latent dirichlet allocation (LDA) for sentiment analysis toward tourism review in indonesia. Journal of Physics: Conference Series. 801. DOI:10.1088/1742-6596/801/1/012073

Scheaffer, RL. Mendenhall III W, Ott RL, Gerow KG. Elementary survey sampling. Cengage Learning, 2011.

Settersten Jr RA, Bernardi L, Harkonen J, Antonucci TC, Dykstra PA, Heckhausen J, Kuh D, Mayer KU, Moen P, Mortimer JT, Mulder CH, Smeeding TM, Lippe TVD, Hagestad GO, Kohli M, Levy R, Schoon I, Thomson E. 2020. Understanding the effects of Covid-19 through a life course lens. Elsevier Public Health Emergency Collection. 46: 100360. DOI: 10.1016/j.alcr.2020.100360.

Steyvers M, Griffiths T. 2007. Probabilistic topic models. Handbook of Latent Semantic Analysis. 427(7): 424-440.

Vijayarani S, Ilamathi MJ, Nithya M. 2015. Preprocessing techniques for text mining-an overview. International Journal of Computer Science & Communication Networks. 5(1): 7-16.

Webb LM, Wang Y. 2014. Techniques for sampling online text-based data sets. IGI Global. 95-114.

Xie P, Xing EP. 2013. Integrating document clustering and topic modeling. Proceedings of the 30th Conference on Uncertainly in Artificial Intelligence.