ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 11, Nomor 2, Tahun 2022, Halaman 266 - 277
Online di: https://ejournal3.undip.ac.id/index.php/gaussian/

PEMODELAN TOPIK PADA KELUHAN PELANGGAN MENGGUNAKAN ALGORITMA *LATENT DIRICHLET ALLOCATION* DALAM MEDIA SOSIAL *TWITTER*

Diandra Zakeshia Tiara Kannitha[1], Mustafid[2], Puspita Kartikasari [3]
[1, 2, 3] Departemen Statistika, Fakultas Sains dan Matematika, Universitas Diponegoro
***Email**: **dztiarakannitha@gmail.com**

ABSTRACT

Large scale social restrictions (PSBB) is a policy issued by the Government of Indonesia as one of the efforts to reduce the spread of the Covid-19 virus. The impact of the policy is that it requires people to conduct activities online . This makes the internet users in Indonesia in the year 2020 up to 73.7%. Each provider must be able to determine strategies in order to maintain the quality of service and customer loyalty. Good reputation for the company is also important, so customers want to use internet services through their company. One of them is by listening to the complaints of the customers towards the company. In this research, modeling the topic of customer complaints carried out using the Latent Dirichlet Allocation Algorithm. The Latent Dirichlet Allocation Algorithm was chosen because the method has good performance. The topic modelling process is carried out using the gibbs sampling estimation. The topic that is often complained to First Media is that internet was turns off while working, while for IndiHome is that the internet often turns off and disconnect. Based on the results of the interpretation, 70% for First Media and 81,81% for IndiHome that these topics had been in accordance with what is complained by customers through their tweets. From the topic that have been known, it can be used as an evaluation for their company in order to maintain service quality and customer loyalty.

Keywords: First Media, IndiHome, Topic Modeling, Latent Dirichlet Allocation

1. PENDAHULUAN

Jumlah pengguna internet di Indonesia tahun ini naik menjadi 73,7 persen dari populasi atau setara 196,7 juta pengguna (VOI Indonesia, 2020). Hal ini tentu menjadi peluang yang sangat baik bagi para Provider Internet atau yang sering disebut juga *Internet Service Provider* (ISP). Disamping itu, keadaan tersebut juga menjadi tantangan tersendiri bagi para Provider. Masing-masing Provider Internet yang ada di Indonesia harus cermat dalam menentukan strategi agar mampu bersaing satu sama lain. Selain itu, mendengar keluhan maupun masukan dari para pelanggan pun juga menjadi hal yang sangat penting untuk di perhatikan demi mempertahankan loyalitas pelanggan serta reputasi baik terhadap perusahaan.

Salah satu media yang saat ini sering digunakan oleh berbagai perusahaan sebagai sarana dalam melayani pelanggan adalah media sosial. Salah satu media sosial yang sering digunakan oleh masyarakat Indonesia termasuk saat pandemi sekarang ini adalah *Twitter*. Indihome dan First Media merupakan contoh dari Provider Internet yang aktif menggunakan sosial media *Twitter* sebagai kanal pelayanan pelanggan dengan nama akun @IndiHomeCare dan @FirstMediaCares. Kedua Provider Internet tersebut memanfaatkan media sosial *Twitter* sebagai salah satu jembatan untuk para *customer* dalam menyampaikan keluhannya.

Banyaknya jumlah tweet yang diunggah oleh para pelanggan mengenai keluhan yang ingin disampaikan kepada perusahaan, mengakibatkan perusahaan tersebut sulit dalam memahami topik apa saja yang sering menjadi keluhan dan pembicaraan para pelanggan. Perlu dilakukan nya analisis pemodelan topik untuk mengatasi hal tersebut. Pemodelan Topik merupakan salah satu bagian dari Text Mining yang digunakan untuk mengidentifikasi topik yang melekat pada sebuah set data, metode yang sangat popular dan sering digunakan adalah *Latent Dirichlet Allocation* (LDA).

LDA merupakan algoritma dalam text mining yang didasarkan pada model topik statistik Bayesian.Dalam penelitian ini akan dilakukan pemodelan topik menggunakan algoritma *Latent Dirichlet Allocation*. Penelitian akan menggunakan *tweets* dengan kata kunci "@IndiHomeCare" dan "@FirstMediaCares" yang akan diolah menggunakan perangkan lunak RStudio1.2.1335 dan Microsoft Excel 2016.

2. TINJAUAN PUSTAKA

Internet Service Provider (ISP) atau yang biasa disebut Provider Internet merupakan perusahaan kabel atau perusahaan telepon seluler yang menawarkan langganan Internet, selain TV atau layanan komunikasi seluler (Techopedia,2020). Indonesia *Digital Home* (IndiHome) dan First Media merupakan 2 dari sekian Provider Internet yang ada di Indonesia. Indihome merupakan layanan dari PT Telekomunikasi Indonesia yang diluncurkan sebagai pengganti Speedy pada tahun 2015 sebagai salah satu penyelenggara jasa internet di Indonesia. PT First Media Tbk yang sebelumnya bernama PT Broadband Multimedia Tbk merupakan perusahaan publik Indonesia yang terdaftar di Bursa Efek Indonesia. First Media menyediakan jasa layanan internet pita lebar, televisi kabel, dan komunikasi data.

Keluhan pelanggan atau *customer complaint* merupakan ungkapan emosional pelanggan karena adanya sesuatu yang tidak dapat diterimanya, baik yang berkaitan dengan produk yang ditawarkan maupun pelayanan yang diterimanya (Sopiah, et al., 2016). Keluhan pelanggan juga merupakan indikasi atau bahan evaluasi untuk memperbaiki hal-hal yang tidak sesuai dalam perusahaan. Setiap keluhan yang disampaikan oleh para pelanggan harus diidentifikasi, dengan tujuan untuk mencari jalan keluar atau solusi yang paling tepat sehingga kedua belah pihak (Pelanggan dan Perusahaan) tidak merasa dirugikan.

Twitter adalah sebuah situs web yang dimiliki dan dikelola oleh Twitter Inc., yang menawarkan jaringan sosial berupa microblog sehingga memungkinkan pengguna untuk mengirim dan membaca pesan kepada pengguna lainnya (Twitter, 2020). *Crawling* merupakan proses pengambilan data dalam jumlah besar dengan cepat ke dalam suatu tempat penyimpanan lokal dan mencarinya berdasarkan sejumlah kata kunci. Tahapan untuk *crawling data* dari Twitter membutuhkan akun Twitter dan akun pengembang di Twitter.

Text Mining adalah proses penggalian informasi dari sekumpulan dokumen data berupa teks yang mengandung informasi yang tidak terstruktur dengan menggunanakan alat analisis tertentu (Feldman, et al., 2007). Pada dasarnya *text mining* memiliki konsep pengolahan yang hampir sama dengan data *mining*, peredaannya yaitu terdapat pada sumber data yang digunakan. Sumber data *text mining* berupa teks tidak terstruktur sedangkan data *mining* menggunakan data terstruktur. Tahap-tahap *text mining* adalah sebagai berikut:

a. *Text Pre-Processing*

Text Preprocessing dapat didefinisikan sebagai tahap awal dari sistem *text mining* yang mengubah format mentah, tidak terstruktur, dan memiliki format asli menjadi terstruktur dan dapat diolah pada tahapan berikutnya (Feldman, et al., 2007). Tahap-tahap *preprocessing* yang dilakukan antara lain:

1. *Case Folding*, yaitu mengubah semua huruf kapital menjadi huruf kecil pada dokumen.
2. *Remove URL,* yaitu menghapus link URL yang terdapat pada dokumen teks.
3. *Unescape HTML,* yaitu meghapus f*ile HTML.*
4. *Remove Mention*, yaitu menghapus rujukan kepada pengguna akun Twitter lain.
5. *Remove Number*, yaitu menghaous angka.
6. *Remove Punctuation,* yaitu menghapus tanda baca selain alphabet.
7. *Remove Emoticon,*yaitu menghapus simbol *emoticon.*
8. *Strip White Space*, yaitu menghapus spasi yang berlebih.
9. Normalisasi Kata, yaitu mengubah kata yang tidak baku menjadi kata baku.

b. *Feature Selection*

Feature Selection merupakan tahapan untuk mengurangi dimensi dari sebuah data tekstual dengan menghapus kata-kata yang tidak relevan sehingga proses pengelompokkan lebih efektif dan akurat (Feldman, et al., 2007). Proses yang dilakukan pada tahapan ini adalah:

1. *Stopwords Removal*, yaitu proses yang berfungsi untuk menghilangkan kata-kata yang sering muncul dalam suatu dokumen, namun memiliki arti yang tidak deskriptif dan dapat dibuang.
2. *Stemming*, yaitu proses mengubah berbagai kata berimbuhan menjadi kata dasarnya, Pada penelitian ini digunakan *package* "katadasaR" yang ditulus ulang oleh Nurandi dalam bahasa R.
3. *Tokenizing*, yaitu proses pemisahan deretan kata di dalam kalimat menjadi potongan-potongan data tunggal. Tokenisasi juga membuang beberapa karakter tertentu yang dianggap sebagai tanda baca (Susilowati, et al., 2015)

c. *Text Representation*, yaitu tahapan mengubah data tekstual menjadi representasi yang lebih mudah untuk diproses. Salah satu cara untuk *text representation* ini adalah dengan menggunakan *Document Term Matrix*. Dalam pembentukan indeks berdasarkan data dokumen, setiap kata perlu diberi nilai/bobot Dalam penelitian ini menggunakan metode pembobotan *Term-Frequency* (TF). Nilai TF didapatkan dengan persmaan berikut:

$$W_{i,j} = tf_{i,j} \quad (1)$$

dengan:

$W_{i,j}$ adalah Pembobotan TF untuk term ke – i pada dokumen ke – j

tf_{ij} adalah Jumlah kemunculan *term* tj dalam dokumen d_i

Topic modeling merupakan sebuah topik yang terdiri dari kata-kata tertentu yang meyusun topik tersebut, dan dalam satu dokumen memiliki probabilitas masing-masing dari beberapa topik yang dihasilkan. Sehingga secara sederhana *topic modeling* dapat diartikan sebagai algoritma untuk menentukan tema utama dari sebuah kumpulan dokumen yang besar dan tidak terstruktur (Jing, 2014). *Topic modeling* digunakan sebagai alat *text mining* untuk mengklasifikasikan sebuah dokumen berdasarkan hasil kesimpulan topik (Jeong, et al., 2017).

Latent Dirichlet Allocation (LDA) merupakan model *probabilistic generative* untuk sekelompok data diskrit seperti sebuah *corpus*. . LDA mengidentifikasi informasi berupa topik laten (tersembunyi) yang mempresentasikan sebuah dokumen sebagai distribusi probabilitas atas beberapa topik, sementara setiap topik dipresentasikan sebagai distribusi probabilitas atas sejumlah kata (Hong, et al., 2010). Terdapat beberapa istilah yang ada didalam LDA (Blei, et al., 2003):

a. Sebuah kata adalah satuan terkecil dalam data diskrit. Kata didefinisikan sebagai suatu item dari kumpulan kosakata dan diberi indeks $\{1,\ldots,V\}$.
b. Sebuah dokumen adalah urutan dari N kata yang dinotasikan dengan $\mathbf{w} = (w_1, w_2, \ldots, w_N)$.
c. Sebuah *corpus* adalah kumpulan dari M dokumen yang dinotasikan dengan $\boldsymbol{D} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\}$.

LDA direpresentasikan sebagai model grafis probabilistik, dimana terdapat tiga tingkatan pada model LDA. Parameter α digunakan dalam menentuan distribusi topik dalam dokumen. Parameter β digunakan untuk menentukan distribusi kata pada topik. Disitribusi topik dari dokumen (α) mengakibatkan adanya nilai θ sebagai kumpulan campuran topik. Variabel θ merupakan variabel yang berada di tingkat dokumen (M). Variabel θ mempresentasikan distribusi topik untuk dokumen tersebut. variabel z dan w adalah variabel yang berada pada tingkat kata (N). Variabel z mempresentasikan topik dari kata tertentu pada sebuah dokumen, sedangkan variabel w mempresentasikan kata yang berkaitan dengan topik tertentu yang terdapat dalam dokumen. Adapun, proses generatif dari LDA apabila diberikan D korpus yang terdiri dari dokumen sebanyak M, terdiri atas langkah-langkah berikut (Jelodar, et al., 2017):

a. Untuk setiap dokumen **w** ($\mathbf{w} \in \{1, \ldots, M\}$), pilihlah distribusi topik Multinomial ($\boldsymbol{\theta}_w$) dari distribusi Dirichlet dengan parameter α

$$\boldsymbol{\theta}_w \sim Dirichlet\ (\boldsymbol{\alpha})$$

b. Untuk setiap topik $k(k \in \{1, \ldots, K\})$, pilihlah distribusi kata *Multinomial*($\boldsymbol{\varphi}_k$) dari Distribusi Dirichlet dengan parameter β

$$\boldsymbol{\varphi}_k \sim Dirichlet\ (\boldsymbol{\beta})$$

c. Untuk setiap N kata dalam dokumen **w**, $(n \in \{1, \ldots, N_w\})$:

   - Pilih sebuah topik $\boldsymbol{z}_n$ dari $\boldsymbol{\theta}_w$ atau $\boldsymbol{z}_n \sim$ Multinomial ($\boldsymbol{\theta}_w$)

- Pilih sebuah kata $\boldsymbol{w}_n$ dari $\boldsymbol{\varphi}_{z,n}$ atau $\boldsymbol{w}_n \sim$ Multinomial ($\boldsymbol{\varphi}_{z,n}$)

Pembangkitakn model LDA merupakan model yang secara random dibangkitkan melalui data observasi yang didalamnya terdapat variabel laten. Data observasi dalam model ini adalah kumpulan dokumen (*w)* dan variabel laten adalah topik yang ditentukan oleh kata-kata didalamnya. Pembangkitan model LDA membentuk ditribusi probabilitas bersama karena dilakukan melalui pembangkitan data dengan distribusi probabilitas masing-masing parameter. Dari distribusi probabilitas bersama tersebut dapat diestimasi parameter untuk variabel laten yang dicari. Berikut langkah-langkah pemodelan *Latent Dirichlet Allocation*:

1. Memasukkan data yang akan diobservasi. Data observasi yang dibangkitkan dalam model LDA adalah kumpulan dokumen yang berisi kata-kata yang sangat banyak. Didefinisikan *corpus* ***D*** dalam bentuk matriks sebagai berikut:

$$\boldsymbol{D} = \begin{bmatrix} \{w_1, \dots, w_{N1}\} & \text{kata} - \text{kata pada dokumen ke} - 1 \\ & \vdots \\ \{\text{internet}, \text{mati}, \text{jam}, \text{malam}, \text{ganggu}, \text{iya}\} & \end{bmatrix}$$

2. Membangkitkan distribusi probabilitas topik terhadap setiap dokumen. Distribusi tersebut diketahui mempunyai distribusi Dirichlet dengan parameter α dan dapat dinyatakan dengan persamaan:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_{k-1}} \quad (2)$$

Bila dituliskan dalam bentuk matriks adalah sebagai berikut

$$\begin{bmatrix} \theta_{11} & \dots & \theta_{1K} \\ \vdots & \vdots & \vdots \\ \theta_{M1} & \dots & \theta_{MK} \end{bmatrix}$$

dengan θ_{MK} adalah kumpulan campuran topik ke-K pada dokumen ke-M.

3. Membangkitkan distribusi probabilitas kata-kata terhadap topik yang telah ditentukan pada tiap dokumen. Distribusi peluang bersyarat untuk φ juga mengikuti distribusi Dirichlet dengan parameter β.

$$p(\varphi|\beta) = \prod_{k=1}^{K} \frac{\Gamma(\beta_k)}{\prod_{v=1}^{V} \Gamma(\beta_{k,v})} \prod_{v=1}^{V} \varphi_{k,v}^{\beta_{k,v-1}} \quad (3)$$

dengan:
$\varphi_{k,v}$ adalah probabilitas *term v* terhadap topik *k*
V adalah banyaknya kata

Bila dituliskan dalam bentuk matriks adalah sebagai berikut:

$$\begin{bmatrix} \varphi_{11} & \cdots & \varphi_{1v} \\ \vdots & \vdots & \vdots \\ \varphi_{k1} & \cdots & \varphi_{kv} \end{bmatrix}$$

4. Membangkitkan distribusi probabilitas topik terhadap topik-topik yang telah ditentukan pada tiap dokumen berdasarkan kata yang muncul didalamnya. Distribusi tersebut mempunyai distribusi Multinomial. Nilai probabilitasnya dapat dihitung dengan persamaan:

$$p(\mathrm{z}|\theta) = \prod_{w=1}^{M} \prod_{k=1}^{K} \theta_{\mathbf{w},k}^{n_{\mathbf{w},\mathbf{k}}} \tag{4}$$

dengan:
$n_{\mathbf{w},\mathbf{k}}$ adalah jumlah kemunculan topik *k* yang ditentukan dari kata-kata dalam dokumen **w**
Bila dituliskan dalam bentuk matriks adalah sebagai berikut:

$$\begin{bmatrix} z_{11} & \cdots & z_{1N} \\ \vdots & \vdots & \vdots \\ z_{M1} & \cdots & z_{MN} \end{bmatrix}$$

dengan:
z_{MN} adalah topik untuk dokumen ke-M pada kata ke-N, dengan nilai z = {1, …, K}

5. Membangkitkan distribusi probabilitas untuk kata-kata dalam *corpus* terhadap topik yang telah dipilih. Distribusi dari probabilitas tersebut mempunyai distribusi Multinomial dan dapat dihitung menggunakan persamaan:

$$p(w|\mathrm{z},\varphi) = \prod_{k=1}^{K} \prod_{v=1}^{V} \varphi_{k,v}^{n_{k,v}} \tag{5}$$

dengan $n_{k,v}$ adalah jumlah kemunculan topik *k* yang ditentukan dari kata-kata dalam *corpus.*

6. Menghitung distribusi posterior bersama yang merupakan model probabilitas LDA dan dapat dihitung dengan persamaan:

$$p\,(\mathrm{w},\mathrm{z},\theta,\varphi \mid \alpha,\beta) =$$

$$\prod_{\mathrm{w}=1}^{M} p(\theta|\alpha) \prod_{k=1}^{K} p(\varphi|\beta) \prod_{t=1}^{N} p(z|\theta)\, p(\mathrm{w}|\,\varphi, z) \tag{6}$$

Distribusi di atas merupakan model probabilitas *Latent Dirichlet Allocation* (LDA) dimana harus dilakukan estimasi dari variabel laten didalamnya. Distribusi diatas merupakan gabungan dari informasi prior dan informasi sampel yang disebut distribusi posterior bersama.

Model LDA dibentuk melalui *generative process*, dan dalam prakteknya model generatif LDA tidak dapat menemukan variabel yang tersembunyi (laten) secara langsung,sehingga diperlukan estimasi parameter. Estimasi parameter yang digunakan dalam model adalah dengan menggunakan Metode Bayesian. Estimasi parameter dalam model LDA dengan Metode Bayesian dapat dilakukan dengan menggunakan Algoritma Gibbs Sampling. Gibbs Sampling merupakan metode yang digunakan untuk menentukan sebaran terbaik dari LDA (Griffiths, et al., 2004). Rumus dari Gibbs Sampling dapat dinyatakan dengan persamaan berikut (Steyvers, et al., 2006):

$$P(z_i = j \mid z_{-i}, w_i, d_i) \propto \frac{C_{w_i j}^{WT} + \beta}{\sum_{w=1}^{W} C_{w,j}^{WT} + W\beta} \frac{C_{d_i j}^{DT} + \alpha}{\sum_{t=1}^{T} C_{d_i t}^{DT} + T\alpha} \quad (7)$$

dengan:

$(z_i = j)$ adalah penempatan topik untuk kosakata ke-i pada topik ke j

z_{-i} adalah penempatan topik untuk semua token kata

w_i adalah kata yang sedang diobservasi

d_i adalah dokumen yang sedang diobservasi

C^{WT} adalah matriks jumlah kata dalam suatu dokumen

C^{DT} adalah matriks jumlah topik dalam suatu dokumen

$C_{w,j}^{WT}$ adalah jumlah kemunculan kata *w* ditempatkan pada topik *j*

$C_{d_i j}^{DT}$ adalah jumlah kemunculan topik *j* ditempatkan ke sebuah token kata dalam dokumen *d*

α adalah parameter *Dirichlet* atas distribusi topik terhadap dokumen

β adalah parameter *Dirichlet* atas distribusi kata terhadap topik

W adalah jumlah kosakata

T adalah jumlah topik

3. METODE PENELITIAN

Penelitian ini dilakukan terhadap Media Sosial Twitter pada tahun 2021. Data yang digunakan pada penelitian ini merupakan data kualitatif yang diperoleh dari *Twitter Crawling* pada tanggal 23 April 2021 dengan kata kunci "@IndiHomCare" dan "@FirstMediaCares" dalam kategori *tweets* berbahasa Indonesia sebanyak 5.000 *tweets*. Data *tweets* yang diperoleh dihilangkan data duplikatnya hingga tersisa 4.828 *tweets* untuk Indihome dan 4.973 *tweets* untuk First Media. Data hasil deteksi duplikat dieliminasi lagi menjadi 555 dan 700 data *tweets* yang benar-benar merupakan keluhan masyarakat terhadap kedua provider tersebut.

Analisis data pada penelitian ini menggunakan bantuan *software* RStudio 1.2.1335 dan *Microsoft Excel* 2016 Langkah-langkah analisis yang dilakukan yakni sebagai berikut:

1. *Twitter Crawling.*

2. *Data Pre-Processing* (*case folding*, *remove* URL, *unescape* HTML, *remove mention*, *remove number*, *remove punctuation*, *remove emoticon*, *strip white space*, dan normalisasi kata).
3. *Feature selection* (*Stopwords Removal, Stemming*, dan *Tokenizing*).
4. Pembobotan kata dengan metode TF.
5. Melakukan pemodelan topik menggunakan Algoritma LDA
    a. Menentukan jumlah topik.
    b. Membangkitkan distribusi probabilitas topik terhadap setiap dokumen.
    c. Membangkitkan distribusi probabilitas kata-kata terhadap topik.
    d. Membangkitkan distribusi probabilitas topik terhadap topik-topik yang telah ditentukan pada tiap dokumen.
    e. Membangkitkan distribusi probabilitas untuk kata-kata dalam *corpus* terhadap topik yang telah dipilih.
    f. Mendapatkan distribusi posterior bersama yang merupakan model probabilitas LDA.
6. Interpretasi hasil untuk setiap topik pada model LDA.
7. Analisis tren topik untuk keseluruhan dokumen.

4. HASIL DAN PEMBAHASAN

Ekstraksi *tweets* dilakukan pada tanggal 23 April 2021 dengan kata kunci "@IndiHomeCare" dan @FirstMediaCares" sebanyak 5.000 *tweets*. Data *tweets* berkurang menjadi 4.828 *tweets* untuk IndihHome dan 4.973 *tweets* untuk First Media setelah dihilangkan data duplikatnya dengan kurun waktu enam hari yaitu tanggal 17 hingga 22 April 2021. Data hasil deteksi duplikat dieliminasi lagi menjadi 555 *tweets* untuk IndiHome dan 700 *tweets* untuk First Media yang benar-benar merupakan data keluhan pelanggan terhadap kedua provider tersebut. Jumlah data *tweets* tersebut yang digunakan menjadi data yang akan dianalisis.

Proses ini dilakukan untuk mengatasi kecenderungan data yang tidak terstruktur. Dilakukan beberapa proses dalam tahap ini, yaitu mulai dari *Case Folding*, *Remove* URL, *Unescape* HTML, R*emove mention*, *Remove number*, *Remove punctuation*, *Remove emoticon*, *Strip white spaces*, hingga Normalisasi kata. Contoh data yang sudah melewati tahap *pre-processing* diberikan pada Tabel 1.

Tabel 1. Contoh Hasil Proses Text Preprocessing

No. *Tweets*	*Tweets* Hasil *Text Pre-Processing*
1	hai internet saya kembali mati mendadak tapi di lihat firstmedia terlihat normal mohon dibantu untuk
2	admin tolong wifi saya error

Feature Selection terdiri dari 3 proses yaitu *stopwords removal, stemming,* dan *tokenizing.* Pada proses *stopword removal* digunakan kamus *stopword*s Indonesia ditambah beberapa *stopwords* secara manual yang kemudian jika dijumlahkan sebanyak 760 kata. Proses *stemming* menggunakan *package* katadasaR pada Rstudio untuk mengubah kata imbuhan menjadi kata dasar. Serta proses *tokenizing* merupakan proses pemisahan deretan kata di dalam kalimat menjadi potongan data tunggal.

Pada tahap ini, dilakukan perubahan data *tweet* menjadi matriks yang kemudian setiap kata dalam dokumen tersebut diberi bobot. Pada penelitian ini pembobotan kata dihitung menggunakan metode *Term Frequency* (TF). Berikut contoh hasil pembobotan kata menggunakan metode TF yang dapat dilihat pada Tabel 3.

Tabel 2. Pembobotan Kata dengan TF

Tweet	**admin**	**bayar**	**internet**	**…**	**kerja**	**mati**
internet mati luar jadwal	0	0	1	…	0	1
admin iya internet mati	1	0	1	…	0	1

Proses pemodelan topik dengan Algoritma LDA dilakukan dengan bantuan *software* RStudio 1.2.1335 dengan *package* '*topicmodels'*.

Pembentukan model dilakukan dengan estimasi *Gibbs sampling.* Nilai *k* dicari melalui *loglikelihood* distribusi marginal kata dalam dokumen bersyarat topik yaitu *loglikelihood* π ($w|z$). Dengan rentang topik dari 2 hingga 20, didapat nilai *loglikelihood* tertinggi ada pada 10 topik. Sedang kan untuk nilai α diambil sebesar $\frac{50}{k}$ dan nilai β diambil sebesar 0,1. Proses pemodelan topik dilakukan dengan menggunakan *syntax* 'model1 <- LDA(x = dtm.tf, k = kfix$Topik, method = "Gibbs" yang kemudian menghasilkan model seperti pada tabel 3.

Tabel 3. Hasil Pemodelan Topik LDA untuk kata kunci "@FirstMediaCares"

Model LDA Topik 1
0.18*"error" + 0.16*"iya" + 0.09*"banget" + 0.04*" bayar" + 0.03*"wifi" + 0.03*"minggu" + 0.02*"hubung"
Model LDA Topik 2
0.2*"admin" + 0.12*"wifi" + 0.06*"lihat" + 0.04*"mohon" + 0.03*"turun" +0.03*"internet" + 0.02*"halo"

Model LDA Topik 3
0.2*"internet" + 0.14*"langgan" + 0.1*"nomor" + 0.04*"mati" + 0.04*"iya" + 0.03*"rumah" + 0.02*"selamat"
Model LDA Topik 4
0.12*"melulu" + 0.08*"kemarin" + 0.07*"jaring" + 0.07*"banget" + 0.05”ganggu" + 0.05*"lambat" + 0.05*"putus"
Model LDA Topik 5
0.13*"tolong" + 0.1*”pagi" + 0.9*"jam" + 0.06*"bantu" + 0.05*"internet" + 0.03*"sih" +0.03*"pakai"
Model LDA Topik 6
0.13*"ganggu" + 0.08* "wifi" + 0.08*"lambat" + 0.06*"media" + 0.05*"first" + 0.04*"baik" +0.04*"tagih"
Model LDA Topik 7
0.16*"iya" + 0.11*"internet" + 0.1*"koneksi" + 0.9*"malam" + 0.03*"selamat" + 0.03*"teknisi" + 0.03*"kali"
Model LDA Topik 8
0.30*"mati" + 0.05*"daerah"+ 0.05*"nyala" +0.04*"admin"+ 0.03*”kerja"+ 0.03*"internet" +0.03*"turun"
Model LDA Topik 9
0.2*"mati" + 0.04*"siang"+ 0.04*"jam" + 0.04*"deh" + 0.04*"firstmedia"+ 0.03*"internet" +0.03*"modem"
Model LDA Topik 10
0.18*"internet" + 0.13*"halo"+ 0.08*" tolong" + 0.05*"ganggu" + 0.05*"jaring"+ 0.04*"sinyal" +0.02*"kendala"

Dengan langkah yang sama dengan kata kunci “@FirstMediaCares”, nilai *k* yang didapat untuk kata kunci “@IndiHomeCare” adalah sebanyak 11 topik. Untuk nilai α diambil sebesar $\frac{50}{k}$ dan nilai β diambil sebesar 0,1. Proses pemodelan topik dilakukan dengan menggunakan *syntax* ‘model1 <- LDA(x = dtm.tf, k = kfix$Topik, method = "Gibbs” yang kemudian menghasilkan model seperti pada tabel 4.

Model LDA Topik 11
0.22*"wifi"+0.07”iya"+0.06*"minggu"+0.05*"banget" +0.05*"sinyal"+0.04*"kerja"+0.04*"lihat"

Tabel 4. Hasil pemodelan topik LDA untuk kata kunci “@IndiHomeCare”

Model LDA Topik 1
0.12*"internet" +0.10*"koneksi" + 0.05"turun" +0.04*"kemarin"+0.04*"jaring"+0.04*"rusak" +0.03*"banget"
Model LDA Topik 2
0.21*"admin" +0.10*"bantu" + 0.10*"mohon" +0.07*"bayar" + 0.03*"siang"+ 0.02*"error" + 0.02*"kali"

Model LDA Topik 3
0.19*"iya"+ 0.15*"internet" + 0.11*"jaring" +0.04*"restart" + 0.03*"indikator" + 0.03*"melulu" +0.02*"jelek"
Model LDA Topik 4
0.11*"internet"+ 0.1*"admin" + 0.08*"error" +0.05*"pagi" +0.05*"tagih" + 0.05*"baik" +0.04*"kemarin"
Model LDA Topik 5
0.1*"lambat"+0.09*"internet"+0.04*"akses" + 0.04*"banget"+0.04*"kasih"+0.03*"mati"+0.03*"sore"
Model LDA Topik 6
0.15*"tolong"+0.08*"ganggu"+0.07*"mati"+0.06*"los" +0.05*"error"+0.04*"tindak"+0.003*"isolir"
Model LDA Topik 7
0.1*"rusak"+0.1*"internet"+0.09*"jam"+0.07*"rumah" +0.07*"merah"+0.03*"pagi"+ 0.03*"hilang"
Model LDA Topik 8
0.12*"mati"+0.11*"internet"+0.1*"halo"+ 0.04*"kedip"+ 0.04*"kemarin"+0.03*"indihome"+0.02*"putus"
Model LDA Topik 9
0.17*"mati"+0.07*" modem"+0.06*"lampu"+0.05*"malam" + 0.03*"langgan"+0.03*"nyala"+0.03*"teknisi "
Model LDA Topik 10
0.21*"indihome"+0.1*" lambat"+0.05*"kak"+0.05*"bayar" + 0.04*"admin"+0.04*"pakai"+0.02*"loh "

Hasil dari analsis model LDA kemudian diinterpretasikan menjadi suatu kalimat. Hasil dari interpretasi penulis untuk semua topik dari model LDA tersebut kemudian dibandingkan dengan *tweet* yang terkait dengan kedua kata kunci tersebut. Didapatkan persentase sebesar 70% untuk kata kunci "@FirstMediaCares" telah sesuai dengan *tweets* keluhan para pelanggan. Sedangkan untuk kata kunci "@IndiHomeCare" didapatkan angka sebesar 81,81% telah sesuai dengan *tweets* keluhan para pelanggan Sisanya tidak dapat diinterpretasikan.

Tren dari data *tweets* berisi keluhan yang diadukan para pelanggan melalui Twitter dapat dilihat melalui nilai probabilitas topik terhadap seluruh dokumen. Nilai probabilitas tertinggi terdapat pada topik ke -8 untuk kedua kata kunci. Beberapa *term* atau kata yang terdapat pada topik tersebut untuk kata kunci "@FirstMediaCares" adalah "mati","daerah", "admin", "kerja", "internet" hal tersebut dapat diartikan bahwa banyak pelanggan First Media yang menyampaikan keluhannya tentang internet yang mati ketika mereka sedang bekerja. Sedangkan untuk kata kunci "@IndiHomeCare" adalah "mati", "internet", "mati, "halo", "putus" hal tersebut dapat diartikan bahwa banyak pelanggan IndiHome yang menyampaikan keluhannya tentang internet yang suka putus dan mati.

5. KESIMPULAN

Latent Dirichlet Allocation merupakan salah satu model yang dapat diterapkan untuk pemodelan topik dan memiliki performa serta hasil yang cukup baik dengan menggunakan data teks berupa *tweets.*). Pada kata kunci "@FirstMediaCares" dengan menggunakan nilai *loglikelihood*, dihasilkan jumlah topik yang optimum yaitu sebanyak 10 topik, sedangkan untuk kata kunci "@IndiHomeCare" sebanyak 11 topik.

Proses Pemodelan topik dilakukan dengan estimasi *gibbs sampling,* dan berdasarkan hasil analisis tren topik, topik yang sering dikeluhkan para pelanggan kepada First Media adalah internet yang mati ketika mereka sedang bekerja, sedangkan untuk IndiHome yaitu internet yang suka putus dan mati. Berdasarkan hasil interpretasi, didapatkan angka sebesar 70% untuk First Media dan 81,81% untuk IndiHome bahwa topik-topik tersebut telah sesuai dengan apa yang dikeluhkan para pelanggan melalui tweets- tweets nya.

DAFTAR PUSTAKA

Blei, D. M., Ng, A. Y., & Jordan, M. L. (2003). Latent Dirichlet Allocation. (J. Lafferty, Ed.) *Journal of Machine Learning Research 3*, 993-1022.

Destarani, A. R., Slamet , I., & Subanti, S. (2019). Trend Topic Analysis using Latent Dirichlet Allocation (LDA).

Elfira,Tachta.2020. " APJII: Pandemi COVID-19 Buat Pengguna Internet di Indonesia Meningkat Hampir 200 Juta", https://www.kompas.com/skola/read/2021/01/02/191038469/cara-membuat-daftar-pustaka?page=all, diakses pada 30 Maret 2021.

Jelodar, H., & Wang, Y. (2017). Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey

Sopiah, & Sangadji, E. M. (2016). Salesmanship (Kepenjualan). Jakarta: Bumi Aksara.

Qomariyah, S., Iriawan , N., & Fithriasari, K. (2019). Topic modeling Twitter data using Latent Dirichlet Allocation and Latent Semantic Analysis. *AIP Conference Proceeding 2194.*