ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 11, Nomor 4, Tahun 2022, Halaman 554 - 561
Online di: https://ejournal3.undip.ac.id/index.php/gaussian/

# PEMODELAN TOPIK ULASAN APLIKASI NETFLIX PADA GOOGLE PLAY STORE MENGGUNAKAN *LATENT DIRICHLET ALLOCATION*

**Gina Rosalinda[1*], Rukun Santoso[2], Puspita Kartikasari[3]**
[1,2,3] Departemen Statistika, Fakultas Sains dan Matematika, Universitas Diponegoro
***e-mail: gina.rosalinda@gmail.com**

**DOI: 10.14710/j.gauss.11.4.554-561**

**Article Info:**
Received: 2022-06-08
Accepted: 2022-12-12
Available Online: 2023-02-25

**Keywords:**
*Topic Modeling; Latent Dirichlet Allocation; Topic Coherence; Netflix; Google Play Store.*

**Abstract:** The vast amount of review data available on the Google Play Store can be utilized to extract hidden essential information. These reviews have an unstructured format that requiring particular methods to automatically collect and analyze the review data. Topic modeling is an extension of text analysis that can find main themes or trends hidden in large sets of unstructured documents. This study applies topic modeling with the Latent Dirichlet Allocation (LDA) method to Netflix application review data sourced from the Google Play Store web. The Latent Dirichlet Allocation (LDA) method is a generative probabilistic model from textual data that can explain the hidden semantic themes in the review document. This research aims to analyze hidden topics that application users discuss. These hidden topics contain essential valuable information for Netflix users and the company. Users can use this information to decide before using Netflix services. Meanwhile, Netflix can use this information to improve the quality of its services. This research use data from a web scraping Netflix review on the Google Play Store from January 2021–August 2021. The results of topic modeling show that of the twelve topics generated, the most discussed topic by users is payment methods.

## 1. PENDAHULUAN

Perkembangan teknologi dan internet yang mendorong adanya transformasi digital telah memberikan alternatif pada cara seseorang menikmati konten hiburan. Salah satunya melalui layanan berbasis langganan atau biasa disebut *Subscription Video-on-Demand* (SVOD). Pengguna harus membayar biaya berlangganan untuk dapat menikmati konten yang tersedia. Konten tersebut dapat diakses dengan bebas kapanpun dan di manapun selama pengguna memiliki koneksi internet, tanpa adanya jadwal penyiaran tertentu (Wayne, 2018). Menurut survei yang dilakukan oleh DailySocial.id, Netflix termasuk salah satu aplikasi layanan *Video-on-Demand* yang memiliki banyak pengguna di Indonesia. Perangkat yang paling sering digunakan untuk mengakses Netflix adalah *smartphone* (Dailysocial.id, 2020). Netflix merupakan layanan *streaming* berlangganan yang membebaskan penggunanya menonton film, acara televisi, dokumenter, dan Netflix Original tanpa iklan pada perangkat yang terkoneksi ke internet (Netflix, 2021).

Pengguna aplikasi Netflix dapat memberikan ulasan dan rating publik di Google Play Store. Data ulasan pengguna aplikasi Netflix yang sangat besar dapat dimanfaatkan untuk mengekstraksi tren topik yang ada. Ulasan tersebut memiliki format yang tidak terstruktur sehingga memerlukan metode khusus untuk mengumpulkan dan menganalisis data ulasan agar didapatkan informasi yang bermanfaat. Hal ini karena sangat tidak memungkinkan untuk mengolahnya dengan proses manual.

Terdapat banyak metode yang digunakan untuk mengekstraksi topik, salah satunya pemodelan topik menggunakan *Latent Dirichlet Allocation* (LDA). *Latent Dirichlet Allocation* (LDA) merupakan model probabilistik generatif dari kumpulan teks yang disebut

korpus. Metode ini memiliki ide dasar bahwa setiap dokumen mewakili campuran topik yang acak sekaligus tersembunyi dan karakter setiap topiknya ditentukan oleh distribusi kata per topik (Blei *et al*., 2003). Beberapa penelitian yang menjadi acuan mengenai LDA untuk mengidentifikasi topik diantaranya yaitu penelitian yang dilakukan oleh Annisa *et al*. (2019), penelitian ini menerapkan LDA *topic modeling* pada 1.187 ulasan pengunjung hotel di Mandalika yang terdapat pada Traveloka. Ulasan yang digunakan merupakan ulasan dalam bahasa Indonesia. Penelitian tersebut berhasil mengidentifikasi delapan topik dari kata-kata yang terdapat pada setiap topik. Penelitian lainnya dilakukan oleh Agustina (2017) yang menganalisis 4.400 suara pelanggan PT. Petrokimia Gresik pada aplikasi Pusat Layanan Pelanggan. Analisis menggunakan metode *Latent Dirichlet Allocation*. Hasil penelitian berhasil mengidentifikasi 35 topik yang dikelompokkan ke dalam tujuh kategori. Nilai akurasinya sebesar 83.7% yaitu 190 dokumen dari 227 dokumen. Penelitian ini menganalisis topik-topik terkait layanan aplikasi Netflix yang sedang dibahas oleh pengguna pada Google Play Store sehingga didapatkan informasi bermanfaat berdasarkan hasil interpretasi yang diperoleh.

## 2. TINJAUAN PUSTAKA

*Text mining* dapat diartikan sebagai proses mengekstraksi informasi bermanfaat dari kumpulan dokumen dengan cara mengidentifikasi dan mengeksplorasi pola yang menarik menggunakan seperangkat alat analisis. Sumber data *text mining* merupakan data tekstual yang tidak terstruktur dan tidak berformat (Feldman dan Sanger, 2007). Tahapan *text mining* umumnya meliputi praproses teks (*text preprocessing)* dan seleksi fitur (*feature selection)* (Berry dan Kogan, 2010; Feldman dan Sanger, 2007). *Text preprocessing* merupakan tahapan pertama dari *text mining*. Tahapan ini meliputi segala proses, aktivitas, serta metode untuk mempersiapkan data sebelum digunakan untuk penemuan pengetahuan sistem *text mining*. *Preprocessing* data pada penelitian ini terdiri atas: *case folding, remove number, remove punctuatiom, spelling normalization, tokenizing, stopword removing,* dan *stemming*.

Dokumen teks tidak dapat langsung diproses dalam bentuk aslinya. Oleh karena itu dokumen perlu direpresentaskan sebagai vektor fitur agar lebih mudah dikelola (Feldman dan Sanger, 2007). Ekstraksi fitur dilakukan menggunakan *Document Term Matrix* (DTM). Setiap baris dalam DTM mewakili dokumen dan setiap kolom dalam DTM mewakili *term*, sementara nilai entrinya adalah jumlah frekuensi kemunculan *term* dalam dokumen. Nilai nol menunjukkan bahwa *term* tersebut tidak muncul di dalam dokumen.

Menurut Blei (2012) *topic modeling* atau pemodelan topik adalah serangkaian algoritma yang memiliki tujuan untuk menemukan topik-topik tersembunyi pada sekumpulan besar dokumen tidak terstruktur. Algoritma pemodelan topik tidak memerlukan pelabelan dokumen sebelumnya karena topik muncul dari hasil analisis teks asli. Pemodelan topik terdiri atas entitas-entitas yaitu (Blei *et al*., 2003):

1. Kata merupakan unit dasar dari data diskrit dalam dokumen yang didefinisikan sebagai item dari kumpulan kosakata berindeks {1, ..., *V*} untuk setiap kata unik pada dokumen.
2. Dokumen adalah susunan *N* kata-kata yang dinotasikan dengan $\boldsymbol{w} = (w_1, w_2, \dots, w_N)$, dengan $w_n$ merupakan kata ke-n dalam suatu barisan kata.
3. Sebuah korpus adalah kumpulan *M* dokumen yang dinotasikan dengan $\boldsymbol{D} = (\boldsymbol{w_1}, \boldsymbol{w_2}, \dots, \boldsymbol{w_M})$.

Sedangkan topik merupakan distribusi kosakata-kosakata yang bersifat tetap. Sederhananya, setiap dokumen pada suatu korpus memiliki proporsi tersendiri dari topik-topik yang dibicarakan sesuai dengan kata-kata yang terkandung di dalamnya.

Metode pemodelan topik yang paling populer adalah *Latent Dirichlet Allocation* (LDA). LDA merupakan salah satu metode yang dapat digunakan untuk menganalisis kumpulan

dokumen teks dalam ukuran yang sangat besar. LDA dapat meringkas, mengklasterisasi, menautkan, dan memproses data yang sangat besar karena LDA menghasilkan daftar topik yang memiliki bobot untuk setiap dokumen (Campbell *et al*., 2014). Distribusi Dirichlet digunakan untuk mendapatkan distribusi topik terhadap dokumen. Proses generatif hasil dari Dirichlet kemudian digunakan untuk mengalokasikan kata-kata pada dokumen untuk topik yang berbeda. Objek yang dapat diamati dalam LDA adalah dokumen, sementara distribusi kata terhadap topik, distribusi topik terhadap dokumen, dan penggolongan setiap kata pada topik tertentu merupakan struktur tersembunyi (*latent*). Oleh karena itu, algoritma ini dinamakan *Latent Dirichlet Allocation* (LDA) (Blei, 2012).

LDA mengikuti proses generatif berikut untuk setiap dokumen **w** pada sebuah korpus **D** yang berisi M dokumen (Campbell *et al*., 2014):

1. Untuk setiap topik ke-*k*, $k \in \{1, \dots, K\}$, memilih $\varphi_k \sim Dir(\beta)$.
2. Untuk setiap dokumen ke-*m*, $m \in \{1, \dots, M\}$:
   a. Memilih $\theta_m \sim Dir(\alpha)$.
   b. Untuk setiap kata ke-*n*, $n \in \{1, \dots, N_m\}$ dalam dokumen $m$:
      i. Memilih topik $z_{m,n} \sim Multinomial(\theta_m)$.
      ii. Memilih kata $w_{m,n} \sim Multinomial(\varphi_{z_{m,n}})$.

Distribusi bersama (posterior) dari semua variabel yang diketahui ($w$) dan tersembunyi ($z, \theta, \varphi$) diberikan parameter $\alpha$ dan $\beta$ seperti pada persamaan (1) (Heinrich, 2005).

$$P(w, z, \theta, \varphi|\alpha, \beta) = P(\theta|\alpha)P(z|\theta)P(\varphi|\beta)P(w|z, \varphi) \quad (1)$$

Persamaan (1) memperlihatkan bahwa terdapat empat komponen pada ruas sebelah kanan persamaan tersebut. Pertama, distribusi topik terhadap dokumen ($\theta$)dibangkitkan mengikuti distribusi Dirichlet dengan parameter $\alpha$.Nilai probabilitasnya dihitung dengan rumus pada persamaan (2).

$$P(\theta|\alpha) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \quad , \alpha_k > 0 \quad (2)$$

Kedua, distribusi probabilitas suatu kata terhadap dokumen untuk masing-masing topik ($z$) bergantung pada distribusi $\theta$ tersebut di atas. Setiap kata $w_n$ dalam dokumen dari N kata diberi nilai dari 1,…,K. Distribusi peluang untuk $z$ yang diamati dengan $\theta$ yang diamati mengikuti distribusi Multinomial yang dihitung menggunakan persamaan (3).

$$P(z|\theta) = \prod_{m=1}^{M} \prod_{k=1}^{K} \theta_{m,k}^{n_{m,k}} \quad (3)$$

dengan:

$\theta_{m,k}$ = distribusi topik *k* terhadap dokumen *m*

$n_{m,k}$ = jumlah kemunculan topik *k* yang ditentukan berdasarkan kata-kata dalam dokumen *m*

Ketiga, distribusi kata terhadap topik ($\varphi$) juga dibangkitkan mengikuti distribusi Dirichlet dengan parameter $\beta$. Probabilitas $\varphi$ untuk seluruh topik dan seluruh kata dalam kosakata diformulasikan pada persamaan (4).

$$P(\varphi|\beta) = \prod_{k=1}^{K} \frac{\Gamma(\beta_k)}{\prod_{v=1}^{V} \Gamma(\beta_{k,v})} \prod_{v=1}^{V} \varphi_{k,v}^{\beta_{k,v} - 1} \quad (4)$$

dengan:

$\varphi_{k,v}$ = probabilitas *term v* terhadap topik *k* dan $\beta_k > 0$

Terakhir, distribusi peluang kata-kata ($w$) dalam korpus terhadap topik tertentu ($z$)bergantung pada distribusi kata terhadap topik ($\varphi$). Distribusi tersebut mengikuti distribusi Multinomial yang didapatkan menggunakan persamaan (5).

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

dengan:

$\varphi_{k,v}$ = probabilitas *term v* terhadap topik *k*

$n_{k,v}$ = jumlah kemunculan topik *k* yang ditentukan berdasarkan *term v* dalam korpus

*Gibbs Sampling* adalah kasus khusus dari *Markov-chain Monte Carlo* (MCMC) dan sering kali menghasilkan algoritma yang relatif sederhana untuk memperkirakan inferensi dari model berdimensi tinggi seperti LDA (Heinrich, 2005). Estimasi parameter metode LDA pada kenyataannya sangat kompleks sehingga sulit untuk diterapkan secara langsung. Hal ini dikarenakan distribusi posterior dari variabel $(z, \theta, \varphi)$ yang tersembunyi atau tidak diketahui dan satu kata dapat mengandung dua topik atau lebih. Distibusi yang menjadi target untuk diestimasi diformulasikan pada persamaan (6) (Steyvers dan Griffiths, 2006):

$$P(z_i = j|z_{-i}, w_i, d_i) \propto \frac{C_{w,j}^{WT}+\beta}{\sum_{w=1}^{W} C_{w,j}^{WT}+W\beta} \frac{C_{d,j}^{DT}+\alpha}{\sum_{t=1}^{T} C_{d,j}^{DT}+T\alpha} \quad (6)$$

*Topic coherence* mengukur nilai setiap topik dengan cara mengukur tingkat kesamaan semantik antarkata dengan nilai tinggi dalam topik. Pengukuran ini dapat digunakan untuk membedakan antara topik hasil temuan inferensi statistik dengan topik yang dapat diinterpretasi secara semantik (Stevens *et al*., 2012). Model yang baik akan menghasilkan topik yang koheren, yaitu topik dengan skor koherensi yang tinggi. Salah satu indikator bahwa topik yang dihasilkan baik atau bermakna adalah kemudahan sesorang dalam memberikan label pendek untuk menggambarkan topik tersebut (Newman *et al*., 2010). Jika suatu topik tidak koheren, sangat mungkin terdapat kata-kata dalam topik yang tidak terkait secara semantik dengan kata lain (Mimno *et al*., 2011). Nilai koherensi topik dapat dihitung dengan persamaan (7) (Stevens *et al*., 2012).

$$coherence\ (V) = \sum_{(v_i,v_j)\in V} score(v_i, v_j, \epsilon) \quad (7)$$

*Word cloud* merupakan metode visualisasi untuk menganalisis dokumen teks. Visualisasi dalam ruang dua dimensi dihasilkan dengan memplot kata-kata dengan frekuensi kemunculan tinggi pada suatu dokumen. Frekuensi kemunculan kata ditunjukkan dengan besar kecilnya ukuran huruf kata tersebut. Semakin besar frekuensi kata tersebut muncul dalam dokumen maka ukuran kata juga semakin besar (Castellà dan Sutton, 2014).

## 3. METODE PENELITIAN

Penelitian ini menggunakan data sekunder berupa data teks. Data teks yang digunakan adalah ulasan aplikasi Netflix berbahasa Indonesia dari situs web Google Play Store. Proses pengambilan data menggunakan teknik *web scraping* dengan alat ekstensi bawaan Google Chrome yaitu Data Miner 5.2.90. Jumlah ulasan yang diperoleh dari hasil pengambilan data selama rentang waktu Januari 2021–Agustus 2021 sebanyak 3028 ulasan. Variabel dalam penelitian ini meliputi: nama (nama pengguna aplikasi Netflix), tanggal (tanggal pengguna aplikasi dalam membuat ulasan), dan ulasan (isi ulasan pengguna aplikasi Netflix).

Analisis data pada penelitian ini menggunakan bantuan perangkat lunak Rstudio 1.3.1093 dan ekstensi Data Miner 5.2.90 yang ada di Google Chrome. Tahapan analisis dalam penelitian ini yaitu:

1. *Web scraping* data ulasan menggunakan ekstensi pada Google Chrome yaitu Data Miner.
2. *Pre-processing* data (*case folding, remove number, remove punctuation, spelling normalization, tokenizing, stopword removing, stemming).*
3. Pembentukan *Document Term Matrix.*
4. Memodelkan topik menggunakan *Latent Dirichlet Allocation* (LDA) dengan insisiasi jumlah topik *k* dan maksimum iterasi *i.*
5. Hasil pemodelan berupa kata-kata untuk setiap topik.
6. Visualisasi kata-kata dengan *word cloud* untuk setiap topik.
7. Evaluasi hasil pemodelan topik dengan menghitung nilai *topic coherence.*

8. Melakukan interpretasi topik berdasarkan hasil kata-kata pada masing-masing topik.

## 4. HASIL DAN PEMBAHASAN

*Preprocessing* data dilakukan untuk membersihkan data agar dapat diolah untuk tahap selanjutnya. Tahapan yang dilakukan meliputi: *case folding*, *remove number*, *remove punctuatiom, spelling normalization, tokenizing, stopword removing* dan *stemming*. Setelah melalui tahap *text preprocessing*, *term* yang terlalu umum dan terlalu spefisik akan dihapus. *Term* yang muncul di lebih dari sama dengan 15% dari total dokumen dan *term* yang muncul pada kurang dari sama dengan lima dokumen akan dihapus. Hal ini dilakukan agar dalam pemodelan topik didapatkan kumpulan istilah yang cukup umum dan digunakan bersama oleh beberapa dokumen untuk menunjukkan topik tersembunyi tetapi juga cukup unik sehingga tidak dimiliki oleh semua dokumen. Setelah mendapatkan *term* yang tidak terlalu umum dan tidak terlalu spesifik maka kumpulan *term* tersebut direpresentasikan dalam bentuk *Document Term Matrix*. Proses ini mengubah data ulasan menjadi bentuk matriks yang isinya merupakan frekuensi kemunculan *term* pada setiap dokumen.

Jumlah topik paling optimal perlu diestimasi terlebih dahulu agar model yang dihasilkan dapat mengektraksi topik tren yang ada secara maksimal. Jika jumlah topik yang dipilih terlalu kecil maka makna pada setiap topik akan terlalu luas sehingga sulit untuk diinterpretasikan. Namun, jika jumlah topik yang dipilih terlalu besar maka akan menyebabkan pengelompokan data yang berlebihan. Hal ini mengakibatkan topik yang dihasilakan tidak berguna karena topik terlalu banyak memiliki kesamaan satu sama lain. Nilai rata-rata koherensi topik tertinggi terdapat pada jumlah topik sebanyak dua belas topik dengan nilai rata-rata koherensi topiknya sebesar 0,10616. Oleh karena itu, didapatkan jumlah topik sebanyak dua belas untuk analisis pemodelan topik pada penelitian ini.

Pembentukan model pada penelitian ini menggunakan *package* "textmineR" pada perangkat lunak RStudio. Estimasi parameter menggunakan algoritma *Gibbs Sampling* dengan jumlah topik (k) sebanyak dua belas. Sedangkan *default* untuk nilai $\alpha$ sebesar 0,1 dan $\beta$ sebesar 0,05. Langkah selanjutnya yaitu membangkitkan distribusi probabilitas topik terhadap setiap dokumen. Distribusi tersebut mengikuti distribusi Dirichlet dengan parameter $\alpha$ sebesar 0,1. Proses pembangkitan secara keseluruhan dilakukan dengan menggunakan bantuan perangkat lunak RStudio dan mengacu pada persamaan (2).

Nilai theta dalam pemodelan topik menggunakan LDA merupakan sebuah matriks yang elemen-elemennya merupakan nilai probabilitas suatu topik terhadap sebuah dokumen. Hasil komputasi dari probabilitas topik terhadap dokumen berdasarkan output `netflix_lda$theta`. Setelah mendapatkan distribusi probabilitas topik terhadap dokumen, selanjutnya membangkitkan distribusi probabilitas kata-kata terhadap topik yang telah ditentukan pada setiap dokumen. Distribusi dari kata-kata terhadap setiap topik mengikuti distribusi Dirichlet dengan parameter $\beta$ sebesar 0,05. Proses pembangkitan data dilakukan menggunakan bantuan perangkat lunak RStudio dengan mengacu pada persamaan (4). Hasil komputasi dari probabilitas topik terhadap dokumen berdasarkan output `netflix_lda$phi`.

Setelah selesai membangkitkan distribusi probabilitas kata terhadap topik, langkah selanjutnya yaitu membangkitkan probabilitas topik terhadap topik-topik yang telah ditentukan pada setiap dokumen berdasarkan kata yang muncul didalamnya. Distribusi tersebut mengikuti distribusi Multinomial. Proses pembangkitan secara keseluruhan mengacu pada persamaan (3).

Langkah terakhir adalah membangkitkan distribusi probabilitas untuk kata-kata dalam korpus terhadap topik yang telah terpilih. Distribusi dari probabilitas tersebut mengikuti ditribusi Multinomial. Proses pembangkitan dilakukan dengan megacu pada


Gambar 1. *Word Cloud* Setiap Topik

persamaan (5). Proses pembangkitan empat distribusi di atas menghasilkan ditribusi posterior bersama yang merupakan model probabilitas LDA. Proses tersebut juga menghasilkan probabilitas kata-kata yang muncul sebagai hasil akhir pembentukan model. Total probabilitas sebuah dokumen berdasarkan grafik model LDA dapat dihitung menggunakan persamaan (1). Tabel 1 menyajikan sepuluh kata dengan probabilitas tertinggi untuk setiap topik.

Tabel 1. Sepuluh Kata dengan Probabilitas Tertinggi untuk Setiap Topik

| **Topik 1** | **Topik 2** | **Topik 3** | **Topik 4** | **Topik 5** | **Topik 6** |
|---|---|---|---|---|---|
| gopay | kartu | buka | kartu | film | daftar |
| mudah | kredit | handphone | kredit | bagus | suruh |
| metode | tolong | download | debit | indonesia | masuk |
| ovo | metode | coba | daftar | tonton | ribet |
| coba | orang | mohon | ribet | lengkap | langsung |
| langgan | pulsa | tulis | susah | tolong | download |
| dana | langgan | tolong | masuk | banyak | lihat |
| tolong | tambah | hapus | orang | langgan | langgan |
| via | debit | maaf | langsung | update | buka |
| ribet | gopay | bantu | pilih | suka | tonton |
| **Topik 7** | **Topik 8** | **Topik 9** | **Topik 10** | **Topik 11** | **Topik 12** |
| beli | kartu | masuk | download | film | kasih |
| tonton | kredit | daftar | gratis | tonton | bintang |
| paket | tolong | login | tonton | lengkap | bagus |
| langgan | debit | email | mending | bagus | tolong |
| gratis | orang | akun | sesal | ulas | film |
| akun | pulsa | kali | kuota | suka | tonton |
| mahal | metode | susah | duit | anime | terima |
| harga | mudah | salah | unduh | mantap | maaf |
| mending | gopay | sandi | orang | cari | coba |
| premium | langgan | tidak bisa | suruh | tayang | sayang |

Gambar 1 menunjukan bahwa kata yang memiliki probabilitas paling tinggi memiliki font paling besar. Sebagi contoh, pada topik satu kata “gopay”, “mudah”, dan “metode”

merupakan kata dengan font paling besar. Hal ini berarti ketiga kata tersebut merupakan tiga kata yang memiliki probabilitas paling tinggi pada topik satu.

Topik dengan nilai koherensi paling tinggi adalah topik delapan dengan nilai 0,18715. Kata-kata pada Topik delapan memiliki keterkaitan yang lebih besar jika dibandingan dengan kata-kata pada topik lainnya. Topik dengan nilai koherensi yang lebih tinggi menunjukkan bahwa topik tersebut lebih mudah diinterpretasikan berdasarkan kata-kata didalamnya.

Tabel 1 menunjukkan sepuluh kata yang terdapat dalam setiap topik, dengan banyaknya topik adalah dua belas. Kata-kata tersebut kemudian diinterpretasikan secara manual menggunakan intuisi dan penilaian peneliti menjadi label tertentu yang mewakili topik tersebut. Beberapa topik mengandung kata yang sama karena metode LDA memperbolehkan suatu kata masuk dalam beberapa topik. Sebagai contoh, pada topik dua, topik empat, dan topik delapan muncul kata "kartu", dan "kredit". Hal ini mengakibatkan ketiga topik tersebut memiliki kemiripan saat diinterpretasikan. Tabel 2 menyajikan hasil interpretasi topik.

Tabel **2**. Hasil Ekstraksi Topik

| **Topik ke-** | **Ektraksi Topik** |
|---|---|
| 1 | Pembayaran via e-wallet |
| 2 | Penambahan metode pembayaran |
| 3 | Sulit diinterpretasikan |
| 4 | Kartu kredit |
| 5 | Film berkualitas |
| 6 | Mendaftar dan masuk |
| 7 | Pembelian paket langganan |
| 8 | Penambahan metode pembayaran |
| 9 | Salah sandi |
| 10 | *Download* gratis |
| 11 | Koleksi film |
| 12 | Rating aplikasi |

## 5. KESIMPULAN

Hasil penelitian menunjukkan bahwa topik yang sering dibahas pelanggan Netflix adalah metode pembayaran yang kurang variatif sehingga sulit dijangkau seluruh kalangan. Pembayaran melalui kartu kredit dirasa rumit sehingga pelanggan menginginkan adanya metode pembayaran lain seperti e-wallet dan pulsa. Nilai koherensi setiap topik dieksplorasi dan didapatkan bahwa topik delapan merupakan topik yang paling mudah diinterpretasikan dengan kemampuan manusia dengan nilai koherensi sebesar 0,18715. Hasil interpretasi menunjukkan bahwa dari dua belas topik yang didapatkan, hanya satu topik yang sulit diinterpretasikan yaitu topik tiga dengan nilai koherensi 0,04190. Hal ini membuktikan bahwa semakin tinggi nilai koherensinya maka semakin mudah topik tersebut diinterpretasikan. Selain itu, hal ini juga menunjukkan bahwa pemodelan topik menggunakan metode *Latent Dirichlet Allocation* berhasil mengungkapkan tren topik yang tersembunyi dalam korpus yang besar.

## DAFTAR PUSTAKA


Agustina, A. (2017). *Analisis dan Visualisasi Suara Pelanggan pada Pusat Layanan Pelanggan dengan Pemodelan Topik Menggunakan Latent Dirichlet Allocation (LDA)*. December.

Annisa, R., Surjandari, I., & Zulkarnain. (2019). *Opinion Mining on Mandalika Hotel Reviews Using Latent Dirichlet Allocation*. Procedia Computer Science, 161, 739–746.

Berry, M. W. & Kogan, J. (2010). *Text Mining Applications and Theory*. United Kingdom: WILEY.

Blei, D. M. (2012). *Probabilistic Topic Models. Communications of the ACM, 55(4), 77–84.*

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). *Latent Dirichlet Allocation*. Journal of Machine Learning Research, 3, 993–1022.

Campbell, J. C., Hindle, A., & Stroulia, E. (2014). *Latent Dirichlet Allocation: Extracting Topics from Software Engineering Data*. The Art and Science of Analyzing Software Data, 139–159.

Castellà, Q. & Sutton, C. (2014). *Word Storms: Multiples of Word Clouds for Visual Comparison of Documents*. WWW 2014 - Proceedings of the 23rd International Conference on World Wide Web, 665–675.

Dailysocial.id. (2020). *Menengok Sederet Aplikasi Hiburan Terpopuler Selama Pandemi*. Dailysocial. https://dailysocial.id/post/menengok-sederet-aplikasi-hiburan-terpopuler-selama-pandemi. Diakses: 6 Januari 2021

Feldman, R. & Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. New York: Cambridge University Press.

Heinrich, G. (2005). *Parameter Estimation for Text Analysis*. Bernoulli, 35, 1–31.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). *Optimizing Semantic Coherence in Topic Models*. EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of The Conference, 2, 262–272.

Netflix, A. (2021). *Tentang Netflix*. https://media.netflix.com/id/about-netflix. Diakses : 6 Januari 2021

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). *Automatic Evaluation of Topic Coherence*. NAACL HLT 2010 - Human Language Technologies: The 2010 Annual Conference of The North American Chapter of The Association for Computational Linguistics, Proceedings of The Main Conference, June, 100–108.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). *Exploring Topic Coherence Over Many Models and Many Topics*. EMNLP-CoNLL 2012 - 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Proceedings of the Conference, July, 952–961.

Steyvers, M. & Griffiths, T. (2006). *Probalistic Topic Models. Latent Semantic Analysis: A Road To Meaning*, 3(3), 993–1022.

Wayne, M. L. (2018). *Netflix, Amazon, and Branded Television Content in Subscription Video On-Demand Portals*. Media, Culture and Society, 40(5), 725–741.