Estimasi, Vol. 4, No. 1, Januari, 2023, Hal. 53-63
DOI: 10.20956/ejsa.vi.24843

Pemodelan Topik pada Judul Berita *Online* Detikcom Menggunakan *Latent Dirichlet Allocation*

Yayang Matira[1*], Junaidi[2], Iman Setiawan[3]
[123] Departemen Statistika, Fakultas MIPA,
Universitas Tadulako, Palu, 94118, Indonesia
*** Corresponding author, email: yayangmatira@gmail.com**

Abstract

Detikcom is a very popular news portal today. The news on the portal continues to grow time to time, causing the existing news data to pile up. As a result, this is necessary to utilize this large amount of data. One of the ways that can be used is to extract topics from news text data through topic modeling using the Latent dirichlet allocation (LDA) method. This method is very popular because it can perform analysis on very large documents. This research aims to find certain patterns in a document by generating several different topics so that it does not specifically divide documents into a particular topic. This research has three topics obtained, with a coherence score is 0,7586. The first topic discusses conflicts and crises within a country, the second topic discusses issues related to humanitarian, and the third topic discusses the issues of corruption committed by state officials.

Keywords*: Detikcom, Latent Dirichlet Allocation (LDA), Topic Modeling, Text Mining, Coherence Score*

Abstrak

Detikcom merupakan portal berita *online* yang sangat populer saat ini. Berita pada portal tersebut terus bertambah seiring berjalannya waktu sehingga menyebabkan makin menumpuknya data berita yang ada. Oleh karena itu perlu adanya pemanfaatan data yang berjumlah besar. Salah satu cara yang dapat digunakan adalah melakukan ekstraksi topik dari data berita melalui pemodelan topik menggunakan metode *latent dirichlet allocation*. Metode ini merupakan metode pemodelan topik yang sangat populer karena dapat melakukan analisis pada dokumen yang berukuran sangat besar. Penelitian ini bertujuan menemukan pola tertentu pada sebuah dokumen dengan menghasilkan beberapa macam topik yang berbeda, sehingga tidak secara spesifik mengelompokkan dokumen ke dalam sebuah topik tertentu. Dari penelitian diperoleh jumlah topik yang terbentuk sebanyak 3 dengan *coherence score* sebesar 0,7586. Kesimpulan dari setiap topiknya yaitu topik ke-1 membahas konflik dan krisis suatu negara, topik ke-2 membahas isu yang berkaitan dengan kemanusiaan, dan topik ke-3 membahas isu korupsi yang dilakukan oleh pejabat negara.

Kata Kunci: Detikcom, *Latent Dirichlet Allocation* (LDA), Nilai Koheren, Pemodelan Topik, *Text Mining*

1. Pendahuluan

Menurut Hadi (2008) Portal berita *online* merupakan portal yang menyediakan informasi *up to date* (setiap hari) mengenai suatu peristiwa atau kejadian yang menyangkut kehidupan kita sehari-hari seperti pendidikan, olahraga, teknologi, politik dan hidup sehat [1]. Suryawati (2011) mengungkapkan berita *online* memiliki perbedaan

yang cukup signifikan dengan media cetak maupun elektronik meskipun mempunyai tujuan yang sama dalam hal menyampaikan berita. Berita *online* mempermudah bagi khalayak untuk mengakses berita [2].

Nugraha dan Mungaran (2021) mengungkapkan bahwa seiring dengan banyaknya portal berita *online* maka semakin banyak pula berita *online* yang beredar. Berita tersebut terus bertambah seiring dengan berjalanya waktu sehingga menyebabkan semakin menumpuknya data berita yang tersedia. Data dengan jumlah besar tersebut menjadi tantangan tersendiri untuk dapat diolah menjadi bentuk yang lebih bermanfaat. Salah satu bentuk pemanfaatan data berjumlah besar tersebut adalah dengan melakukan ekstraksi topik dari data teks berita dengan pemodelan topik agar data-data tersebut dapat dikategorikan berdasarkan topik pembahasan di dalamnya [3].

Menurut Blei dkk. (2003) pemodelan topik atau *topic modelling* adalah sebuah metode untuk mengekstrak dan merepresentasikan konteks yang digunakan sebagai sebuah arti kata dengan memanfaatkan komputasi statistik untuk sejumlah korpus yang besar dari teks. Tujuan pemodelan topik yaitu menemukan topik dan kata-kata yang terkandung dalam korpus tersebut. Dalam melakukan pemodelan topik, dapat diimplementasikan metode pengelompokan berdasarkan ukuran kedekatan (kemiripan) suatu data, salah satunya dengan menggunakan *Latent Dirichlet Allocation* (LDA). Metode ini merupakan sebuah metode *text mining* untuk menemukan pola tertentu pada sebuah dokumen dengan menghasilkan beberapa macam topik yang berbeda, sehingga tidak secara spesifik mengelompokkan dokumen [4]. Campbell dkk. (2015) mengatakan *Latent Dirichlet Allocation* (LDA) merupakan metode pemodelan topik yang paling populer saat ini. LDA muncul sebagai salah satu metode yang dipilih dalam melakukan analisis pada dokumen yang berukuran sangat besar. LDA dapat digunakan untuk meringkas, melakukan klasterisasi dan menghubungkan maupun memproses data yang sangat besar karena LDA menghasilkan daftar topik yang diberi bobot untuk masing-masing dokumen [5].

Penelitian mengenai analisis topik telah banyak dilakukan sebelumnya, yaitu penelitian yang dilakukan oleh Nugraheni dkk. (2020) menggunakan Metode LDA-*Based Topic modelling* dalam menganalisis *review* pengunjung hotel pada website Agoda dengan studi kasus Ibis Bandung Trans Studio. Hasil yang diperoleh secara keseluruhan adalah ungkapan positif terhadap jasa yang ditawarkan oleh Ibis Bandung Trans Studio Hotel [6]. Penelitian lain dilakukan oleh Nugraha dan Mungaran (2021) dalam memodelkan topik pada portal berita *online* berbahasa Indonesia menggunakan *Latent Dirichlet Allocation* (LDA) dengan hasil terbaik terbentuk lima buah topik dari total sebanyak 68.537 artikel dengan nilai koheren sebesar 0,67 [3]. Berdasarkan beberapa penelitian sebelumnya, maka pada penelitian ini akan dibahas mengenai penerapan metode analisis topik dengan *Latent Dirichlet Allocation* (LDA) dalam menentukan pemodelan topik pada judul berita *online* detikcom kanal detikNews.

2. Material dan Metode

Data yang digunakan dalam penelitian ini adalah data sekunder yaitu data judul berita *online* detikcom kanal detikNews yang diperoleh dari halaman website *https://news.detik.com* periode 16 Desember 2021 hingga 24 Maret 2022. Metode analisis yang digunakan dalam penelitian ini adalah *latent dirichlet allocation* dengan menggunakan *software python* 3.10. Adapun tahapan analisis yang dilakukan dalam penelitian adalah sebagai berikut:

1. Pengambilan data judul berita pada portal detikcom kanal detikNews. Pengumpulan data dilakukan melalui metode *web scraping.*
2. Melakukan input data
3. Melakukan praproses data dengan tujuan menyiapkan data menjadi lebih terstruktur yaitu meliputi *case folding, remove punctuation, stopwords, dan tokenizing*
4. Membentuk model *bi-gram* dan *tri-gram*
5. Melakukan visualisasi data menggunakan *word cloud*
6. Pembobotan kata menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF) digunakan untuk pembobotan kata yang awalnya data teks menjadi data numerik
7. Penentuan jumlah topik menggunakan *topic coherence*
8. Melakukan pemodelan topik menggunakan algoritma *latent dirichlet allocation* (LDA)
9. Melakukan Interpretasi hasil

3. Hasil dan Diskusi

3.1 Praproses Data

Langkah awal sebelum memulai analisis yaitu melakukan tahap praproses data. Tahap ini bertujuan menghilangkan atau membersihkan teks pada data yang tidak diperlukan. Praproses data umumnya bersifat dinamis dan berbeda-beda pada setiap data. Pada penelitian ini, tahapan praproses data dibagi menjadi 4 tahapan, yaitu *case folding, remove punctuation, stopwords* dan *tokenizing*.

Tabel 1. Praproses Data

Tahapan	Sebelum	Sesudah
Case folding	Nelayan Ditangkap Saat Akan Jual 22 Kg Sabu yang Ditemukan Ketika Melaut	nelayan ditangkap saat akan jual 22 kg sabu yang ditemukan ketika melaut
Remove punctuation	nelayan ditangkap saat akan jual 22 kg sabu yang ditemukan ketika melaut	nelayan ditangkap saat akan jual kg sabu yang ditemukan ketika melaut

Stopwords	nelayan ditangkap saat akan jual kg sabu yang ditemukan ketika melaut	nelayan jual kg sabu melaut
Tokenizing	nelayan jual kg sabu melaut	['nelayan', 'jual', 'kg', 'sabu', 'melaut']

Sumber : Hasil Olah Data, 2022

Pada Tabel 1 dapat dilihat bahwa praproses data terdiri atas beberapa tahapan, dimulai pada tahapan *case folding* yang merupakan tahap untuk menyamakan semua karakter dengan cara mengubah huruf kapital menjadi huruf kecil. Selanjutnya, tahap *remove punctuation* yang merupakan tahap untuk menghapus karakter tanda baca seperti angka, tanda tanya, koma, titik dua dan lain sebagainya. Kemudian *stopwords* yang merupakan tahap untuk menghapus kata yang sering muncul, namun tidak memiliki makna. Tahap ini mengambil daftar *stopwords* yang disediakan oleh *Package* Sastrawi di *Python* dan juga yang disediakan berdasarkan teks yang digunakan. Tahap terakhir yaitu *tokenizing* adalah tahap untuk memisahkan deretan kata sehingga diperoleh potongan kata atau token yang akan menjadi entitas yang memiliki nilai dalam penyusunan matriks dokumen pada proses selanjutnya.

3.2 Bi-gram dan Tri-gram

Pada tahap pembentukan *bi-gram dan tri-gram* bertujuan untuk mencari suatu makna kata yang disusun oleh dua kata dan tiga kata. Hal ini dilakukan untuk meminimalisir pemenggalan kata yang dapat menghilangkan makna kata tersebut. Berikut adalah contoh pembentukan *bi-gram* dan *tri-gram*.

Tabel 2. Pembentukan *Bi-gram* dan *Tri-gram*

Sebelum *bi-gram* dan *tri-gram*	**Sesudah *bi-gram***	**Sesudah *tri-gram***
'ridwan', 'kamil', 'lepas', 'ekspor', 'jengkol', 'desa', ‘ciwidey’, ‘dubai’	‘ridwan kamil’, ‘kamil lepas’, ‘lepas ekspor’, ‘ekspor jengkol’, ‘jengkol desa’, ‘desa ciwidey’, ciwidey dubai’	‘ridwan kamil lepas’, ‘kamil lepas ekspor’, ‘lepas ekspor jengkol’, ‘ekspor jengkol desa’, ‘jengkol desa ciwidey’, ‘desa ciwidey dubai’

Sumber : Hasil Olah Data, 2022

Berdasarkan hasil pembentukan *bi-gram* dan *tri-gram* pada Tabel 2 terdapat kata yang bergandengan dua kata membentuk *bi-gram* dan bergandengan tiga kata membentuk *tri-gram*. Jika kemunculan kata hasil proses pembentukan *bi-gram* dan *tri-gram* lebih dari 10 maka hasil akan disimpan dan sebaliknya akan dihapus.

3.3 *Word Cloud*

Word cloud adalah salah satu metode *text mining* dalam menganalisis data berbentuk teks, dimana kata yang memiliki ukuran besar dalam *word cloud* menandakan bahwa kata tersebut memiliki frekuensi kemunculan yang tinggi dalam data, begitu juga sebaliknya. Berikut contoh hasil *word cloud* pada judul berita *online* detikcom.

Gambar 1. *Word cloud*

Dari hasil *word cloud* pada Gambar 1 dapat dilihat beberapa kata dengan frekuensi kemunculan cukup tinggi yang ditandai dengan ukuran kata yang lebih besar dibanding kata lainnya di dalam teks, antara lain Ukraina, Rusia, Jokowi, Banjir dan Gempa, Komisi Pemberantasan Korupsi (KPK).

3.4 Pembobotan TF-IDF

Pembobotan *Term Frequency – Inverse Document Frequency* (TF-IDF) adalah suatu bentuk metode *text mining* yang digunakan untuk tahapan *text transformation* dan *feature selection.* Pada bagian text *transformation*, data yang awalnya berbentuk teks akan diubah menjadi angka karena proses pemodelan topik membutuhkan data yang berbentuk angka. Sementara pada bagian *feature selection*, akan dipilih 1000 kata dengan nilai TF-IDF terbesar untuk dilanjutkan pada tahapan pemodelan topik. Berikut merupakan contoh pembobotan TF-IDF.

Tabel 3. Hasil Pembobotan Kata TF-IDF

D	**Banjir**	**Bentrokan**	⋯	**Ukraina**	⋯	**Walkot**
	⋮	⋮	⋮	⋮	⋮	⋮
12	0	0	0	4,2188	0	0
25	4,7297	0	0	0	0	0
361	0	7,6201	0	0	0	0
731	0	0	0	0	0	6,5214
1500	0	0	0	0	0	0

Sumber : Hasil Olah Data, 2022

Tabel 3 di atas merupakan contoh kata yang menghasilkan nilai TF-IDF dihitung secara otomatis menggunakan software python. Untuk contoh perhitungan manual TF-IDF terlebih dahulu harus menghitung nilai *Term Frequency* (TF).

Tabel 4. Sampel Hasil Nilai TF

D	**Banjir**	**Bentrokan**	⋯	**Ukraina**	⋯	**Walkot**
	⋮	⋮	⋮	⋮	⋮	⋮
12	0	0	0	1	0	0
25	1	0	0	0	0	0
361	0	1	0	0	0	0
731	0	0	0	0	0	1
1500	0	0	0	0	0	0

Sumber : Hasil Olah Data, 2022

Pada perhitungan manual, sebagai contoh kata ‘Banjir’ pada Tabel 4 berada pada dokumen ke 25. Nilai 0 pada tabel TF berarti bahwa dalam satu dokumen kata yang dicari tidak ada dalam dokumen tersebut, sedangkan nilai 1 pada tabel TF yang berarti bahwa dalam satu dokumen kata yang dicari muncul satu kali, dan seterusnya. Selanjutnya menghitung nilai *Document Frequency* (DF) secara manual yaitu jumlah dokumen pada suatu term yang memunculkan kata ‘Banjir’. Nilai DF dari kata ‘Banjir’ dari 1500 dokumen muncul sebanyak 36 kali. Kemudian dihitung nilai IDF atau inverse dari DF, term yang memiliki DF yang rendah memiliki nilai IDF yang tinggi. Kemudian melakukan perhitungan nilai *Inverse Document Frequency* (IDF).

$$IDF = \ln\left(\frac{D}{DF_j}\right) \tag{1}$$

$$IDF = \ln\left(\frac{1500}{36}\right) = 3{,}7297 \tag{2}$$

Dengan,

IDF_j : *Inverse document frequency* pada *term* ke j

D : Jumlah semua dokumen yang ada dalam koleksi

DF_j : Jumlah dokumen yang mengandung *term* (j)

Setelah mendapatkan nilai TF, DF, dan IDF, selanjutnya menghitung nilai pembobotan TF-IDF secara manual, yaitu dengan mengalikan nilai TF dengan nilai IDF.

$$W_{ij} = TF_{ij}\, x\, \left(\ln\left(\frac{D}{DF_j}\right) + 1\right) \tag{3}$$

$$W_{ij} = 1\, x\, (3{,}7297 + 1) = 4{,}7297 \tag{4}$$

Dengan,

i : Dokumen ke-d

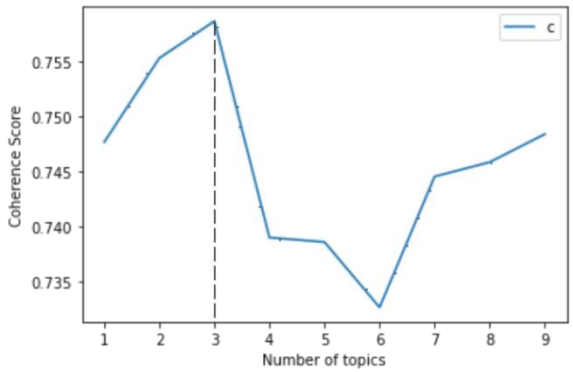
j : Kata ke-t dari kata kunci

W_{ij} : Bobot *term* (*j*) terhadap dokumen (*i*)
TF_{ij} : Jumlah kemunculan *term* (*j*) dalam dokumen (*i*)

Dari hasil pembobotan di atas maka dapat disimpulkan bahwa kata 'Banjir' pada dokumen ke 25 mempunyai pembobotan kata TF-IDF sebesar 4,7297. Dapat diketahui hasil dari pembobotan ($W_{ij} = 4{,}7297$) yang telah dilakukan secara manual menunjukan hasil yang sama dengan hasil yang dilakukan secara otomatis menggunakan *software python* seperti pada Tabel 3.

3.5 *Topic Coherence*

Topic coherence merupakan tahapan untuk menentukan jumlah model topik dalam *topic modelling*. Semakin tinggi nilai *topic coherence* menunjukkan bahwa model yang dihasilkan akan semakin baik. Selanjutnya jumlah topik dengan nilai *topic coherence* tertinggi akan digunakan sebagai topic modeling. Berikut merupakan hasil grafik pada *topic coherence*.



Gambar 2. *Topic Coherence*

Gambar 2 menunjukkan bahwa puncak grafik *topic coherence* berada pada *number of topics* ke-3 atau dengan kata lain jumlah topik terbaik yang terbentuk adalah 3 topik. Guna memudahkan untuk melihat hasil visualisasi dari Gambar 2, maka dapat dilihat menggunakan nilai *coherence score* pada Tabel 5 berikut ini.

Tabel 5. Nilai *Topic coherence*

Num of Topics	***Coherence Score***
1	0,7477
2	0.7553
3	**0,7586**
4	0,7390
6	0,7327
7	0,7445
8	0,7458
9	0,7484

Sumber : Hasil Olah Data, 2022

Hasil *Topic coherence* pada Tabel 5 menunjukkan bahwa nilai dari *Topic coherence* yang paling tinggi yaitu pada *number of topics* ke-3 dengan nilai *coherence score* sebesar 0,7586. Berdasarkan nilai *coherence score*, maka jumlah topik terbaik yang didapatkan akan dijadikan acuan dalam pembuatan model, sehingga dalam penelitian ini digunakan 3 model topik terbaik.

3.6 *Latent Dirichlet Allocation* (LDA)

Model LDA Topik Ke-1

Pada topik ke-1 didapatkan model LDA sebagai berikut:

0.014*"rusia" + 0.012*"ukraina" + 0.010*"gempa" + 0.010*"rusia_ukraina" + 0.008*"minyak_goreng" + 0.007*"kebakaran" + 0.007*"polri" + 0.006*"banjir"+ 0.006*"pemerintah" + 0.006*"perang"'

Kemudian dari model tersebut juga dapat ditampilkan dengan bentuk *word cloud* sebagai berikut:

Gambar 3. *Word cloud* Topik ke-1

Berdasarkan beberapa data yang diperoleh bahwa pada pemodelan LDA topik ke-1 membahas tentang konflik Rusia dan Ukraina, kelangkaan minyak goreng dan bencana alam. Sehingga dapat disimpulkan bahwa bahasan pada topik ke-1 cenderung membahas isu yang berkaitan dengan konflik dan krisis negara.

Model LDA Topik Ke-2

Pada topik ke-2 didapatkan model LDA sebagai berikut:

0.011*"ukraina" + 0.009*"kerangkeng_bupati" + 0.008*"satpol_pp" + 0.008*"bupati" + 0.008*"sungai" + 0.008*"gempa_guncang" + 0.008*"anies" + 0.007*"dunia" + 0.007*"gubernur" + 0.007*"banjir"

Kemudian dari model tersebut juga dapat ditampilkan dengan bentuk *word cloud* sebagai berikut:

Gambar 5. *Word cloud* Topik ke-2

Berdasarkan beberapa data yang diperoleh bahwa pada pemodelan LDA topik ke-2 membahas tentang korban perang ukraina dan kerangkeng manusia Bupati Langkat. Sehingga dapat disimpulkan bahwa bahasan pada topik ke-2 cenderung membahas isu yang berkaitan dengan krisis kemanusiaan.

Model LDA Topik Ke-3

Pada topik ke-3 didapatkan model LDA sebagai berikut

0.010*"tni" + 0.009*"kpk" + 0.008*"jokowi" + 0.008*"polda" + 0.007*"presiden" + 0.007*"ridwan_kamil" + 0.007*"bupati" + 0.006*"dprd" + 0.006*"banjir" + 0.006*"suap"

Kemudian dari model tersebut juga dapat ditampilkan dengan bentuk *word cloud* sebagai berikut:

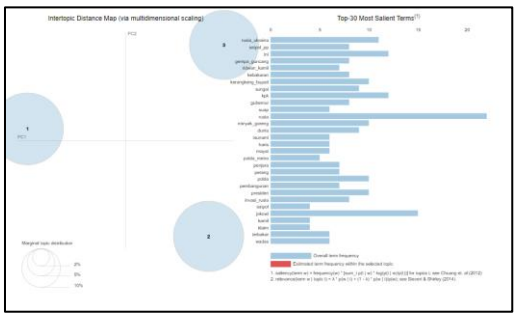
Gambar 6. *Word cloud* Topik Ke-3

Berdasarkan beberapa data yang diperoleh bahwa pada pemodelan LDA topik ke-3 membahas tentang kasus korupsi TNI dan kasus suap anggota DPRD. Sehingga dapat disimpulkan bahwa bahasan pada topik ke-3 cenderung membahas isu yang berkaitan dengan korupsi oleh pejabat negara.

3.7 *Perceptual Map*

Perceptual map atau peta persepsi adalah hubungan antara objek yang dipersepsikan dan dinyatakan sebagai hubungan geometris antara titik-titik di alam ruang yang multidimensional koordinat. Setelah diperoleh model *Latent dirichlet allocation* (LDA) dan *word cloud* maka model tersebut dapat dilihat dengan visualisasi Perceptual Map dan

keterkaitan antar kata yang dihasilkan. Didapatkan hasil visualisasi menggunakan perceptual map sebagai berikut:



Gambar 7. Visualisasi *Perceptual Map*

Berdasarkan Gambar 7 pada panel sisi kanan terdapat 30 buah kata yang paling relevan untuk topik tertentu, sedangkan pada panel sisi kiri merupakan pemetaan jarak antar topik menggunakan multidimensional scaling yang terdapat juga *cluster* topik yang berbentuk lingkaran dengan nomor tertentu pada setiap *cluster* topik. Selanjutnya dapat dilihat bahwa letak setiap topik pada grafik memiliki jarak yang saling berjauhan, sehingga dapat dikatakan bahwa setiap topik tidak memiliki kemiripan dengan topik lainnya.

4. Kesimpulan

Nilai koheren terbaik yang didapatkan pada model *Latent dirichlet allocation* (LDA) sebesar 0,7586 dengan jumlah topik terbaik yaitu sebanyak 3 topik. Berdasarkan hasil analisis dari pemodelan topik dengan metode *Latent dirichlet allocation* (LDA) pada judul berita *online* detikcom diperoleh 3 topik yaitu, model LDA topik ke-1 yaitu membahas konflik dan krisis negara, model LDA topik ke-2 yaitu membahas isu yang berkaitan dengan krisis kemanusiaan, model LDA topik ke-3 yaitu membahas isu korupsi oleh pejabat negara. Adapun saran pada penelitian ini adalah diharapkan pada penelitian selanjutnya menggunakan tahapan *stemming* pada praproses data diharapkan menggunakan portal berita *online* yang lain.

Daftar Pustaka

[1] Hadi, A. F., Bagus, D., & Hasan, M. 2017. *Text mining* pada Media Sosial Twitter Studi Kasus : Masa Tenang Pilkada DKI 2017 Putaran 2. *Seminar Nasional Matematika Dan Aplikasinya*, 21 Oktober 2017 Surabaya, Universitas Airlangga.

[2] Suryawati, I. 2011. *Suatu Pengantar Teori dan Praktek*. Jurnalistik. Ghalia Indonesia. Bogor.

[3] Nugraha, M. A., & Mungaran, L. C. 2021. Pemodelan Topik Berita pada Portal Berita *Online* Berbahasa Indonesia Menggunakan *Latent dirichlet allocation* (LDA). *Jurnal Ilmiah Komputasi*, 20(2).

[4] Blei, D. M., Ng, A. Y., & Jordan, M. I. 2003. *Latent dirichlet allocation*. *Journal of Machine Learning Research*, 3(4–5), 993–1022. https://doi.org/10.1016/b978-0-12-411519-4.00006-9

[5] Campbell, J. C., Hindle, A., & Stroulia, E. 2015. *Latent dirichlet allocation*: extracting topics from software engineering data. *The art and science of analyzing software data (pp. 139–159)*. Elsevier.

[6] Nugraheni, N., Yansari, B. A. R., Listyawan, D. N., Putrayasa, G. E., Sidik, M. H., & Maulana, A. I. 2020. Analisis Review Pengunjung Hotel pada Website Agoda Menggunakan Metode Lda-Based Topic Modeling (Studi Kasus: Ibis Bandung Trans Studio Hotel). *JBMI (Jurnal Bisnis, Manajemen, Dan Informatika)*, 16(3), 252–257. https://doi.org/10.26487/jbmi.v16i3.8484