KLASIFIKASI JUDUL BERITA COVID-19 MENGGUNAKAN METODE LATENT DIRICHLET ALLOCATION

TUGAS AKHIR

Diajukan Sebagai Salah Satu Syarat untuk Memperoleh Gelar Sarjana Teknik
pada Program Studi Teknik Informatika Fakultas Sains dan Teknologi

UIN SUSKA RIAU

Oleh:

NUR PRATIWI NURID

11751201978

PROGRAM STUDI TEKNIK INFORMATIKA

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS ISLAM NEGERI SULTAN SYARIF KASIM RIAU

PEKANBARU

2022

LEMBAR PERSETUJUAN

KLASIFIKASI JUDUL BERITA COVID-19 MENGGUNAKAN METODE LATENT DIRICHLET ALLOCATION

TUGAS AKHIR

Oleh

NUR PRATIWI NURID

NIM. 11751201978

Telah diperiksa dan disetujui sebagai Laporan Tugas Akhir
di Pekanbaru, pada tanggal 4 Februari 2022

Pembimbing I,

RESKI MAI CANDRA, S.T., M.Sc

NIP. 198605052015031006

LEMBAR PENGESAHAN

KLASIFIKASI JUDUL BERITA COVID-19 MENGGUNAKAN METODE LATENT DIRICHLET ALLOCATION

Oleh

NUR PRATIWI NURID

NIM. 11751201978

Telah dipertahankan di depan sidang dewan penguji
sebagai salah satu syarat untuk memperoleh gelar Sarjana Teknik
pada Universitas Islam Negeri Sultan Syarif Kasim Riau
di Pekanbaru, pada tanggal 4 Februari 2022

Pekanbaru, 4 Februari 2022
Mengesahkan,

Dekan,	Ketua Jurusan,
Dr. Hartono, M.Pd	**Iwan Iskandar, M.T**
NIP. 19640301 199203 1 003	**NIP. 19821216 201503 1 003**

DEWAN PENGUJI

Ketua : Iwan Iskandar, S.T., M.T
Pembimbing I : Reski Mai Candra, S.T., M.Sc
Penguji I : Suwanto Sanjaya, S.T., M.Kom
Penguji II : Fitri Insani, S.T., M.Kom

SURAT PERNYATAAN

Saya yang bertanda tangan dibawah ini :

Nama	:	NUR PRATIWI NURID
NIM	:	11751201978
Tempat/Tgl. Lahir	:	Pekanbaru, 18 Agustus 1999
Fakultas/Pascasarjana	:	Fakultas Sains dan Teknologi
Judul Skripsi	:	

KLASIFIKASI JUDUL BERITA COVID-19 MENGGUNAKAN METODE LATENT DIRICHLET ALLOCATION

Menyatakan dengan sebenar-benarnya bahwa :

1. Penulisan Skripsi lainnya dengan judul sebagaimana tersebut diatas adalah hasil pemikiran dan penelitian saya sendiri.
2. Semua kutipan pada karya tulis saya ini sudah disebutkan sumbernya.
3. Oleh karena itu Skripsi lainnya saya ini, saya nyatakan bebas dari plagiat.
4. Apa bila dikemudian hari terbukti terdapat plagiat dalam penulisan Skripsi saya tersebut, maka saya bersedia menerima sanksi sesuai peraturan perundang-undangan.

Demikianlah Surat pernyataan ini saya buat dengan penuh kesadaran dan tanpa paksaan dari pihak manapun juga.

Pekanbaru, 22 Juli 2022
Yang Membuat Pernyataan

METERAI TEMPEL
FC7B9AJX974730569

NUR PRATIWI NURID
NIM. 11751201978

SENTIMETER (Seminar Nasional Teknologi Informasi, Mekatronika dan Ilmu Komputer)
Universitas Nusa Putra, 8 Januari 2022

KLASIFIKASI JUDUL BERITA COVID-19 MENGGUNAKAN METODE LATENT DIRICHLET ALLOCATION

Nur Pratiwi Nurid [1)], dan Reski Mai Candra [2)]
[1,2] Fakultas Sains dan Teknologi, Universitas Islam Negeri Sultan Syarif Kasim Riau, Jalan HR. Soebrantas Pekanbaru
e-mail: 11751201978@students.uin-suska.ac.id[1)], reski.candra@uin-suska.ac.id [2)]

ABSTRAK

Arus globalisasi yang menyebar di masyarakat berupa penyampaian informasi semakin berkembang dari tahun ke tahun seperti televisi dan radio. Lalu dengan adanya internet penyampaian informasi semakin berkembang yaitu berupa teks seperti berita online. Penyebaran coronavirus disease 2019 atau covid-19 di Indonesia merupakan informasi berita yang paling banyak dicari dikarenakan banyaknya kekhawatiran dari seluruh masyarakat, maka dari itu pada masa pandemi untuk mengetahui informasi yang tepat dan terpercaya mengenai covid-19 di Indonesia dapat dilihat pada situs resmi covid-19.go.id. Jenis berita yang terdapat pada website tersebut berupa penanganan kesehatan, vaksinasi covid-19 dan pemulihan ekonomi. Dengan banyaknya berita yang disajikan, penulis telah melakukan penelitian mengenai klasifikasi judul dengan menggunakan metode latent dirichlet allocation. Tujuan dari penelitian ini adalah untuk mengklasifikasikan judul berita pada situs covid-19 dan mengetahui topik berita pada keseluruhan berita di situs covid-19.go.id. Penelitian ini dilakukan melalui beberapa tahapan yaitu pengumpulan judul berita berbahasa Indonesia pada situs covid19.go.id dari bulan Januari 2021 hingga November 2021. Setelah melakukan pengumpulan judul berita, data tersebut dilakukan proses preprocessing. Tahapan pembobotan yang mengubah kata pada data menjadi bentuk numeric dengan melakukan pembobotan TF-IDF. Penerapan metode Latent Dirichlet Allocation untuk memperoleh topik dari keseluruhan data judul berita. Berdasarkan hasil penelitian, terdapat 3 topik dengan jumlah probabilitas yang tertinggi dan berkaitan dari keseluruhan data judul berita pada website covid19.go.id yaitu topik update dengan nilai probabilitas 0,047363, topik vaksinasi dengan nilai probabilitas 0,064872 dan topik sembuh dengan nilai probabilitas 0,054829.

Kata Kunci: covid-19, judul berita, klasifikasi, *latent dirichlet allocation*

ABSTRACT

The wave of globalization that spreads in society in the form of delivering information is growing from year to year through television and radio. In addition, with the existence of internet, the delivery of information is also growing in the form of text such as online news. The spread of coronavirus disease 2019 or covid-19 in Indonesia is the most searched news information due to many concerns from the whole society, therefore to find out the right and reliable information about covid-19 in Indonesia during this pandemic can be accesed on the official website namely covid19.go.id. The types of news contained on the website are about the management of health, covid-19 vaccination, and economic recovery. With these many presented, the author has done research about the classification of titles using latent dirichlet allocation method. This research is carried out through several stages, namely (1) collecting Indonesian news titles on the covid19.go.id site from January 2021 to November 2021. (2) After collecting news titles, the data was preprocessed. (3) The weighting stage changes the words in the data into numeric form by weighting the TF-IDF. (4) Application of the Latent Dirichlet Allocation method to obtain topics from the entire news headline data. Based on the results of the study, there are 3 topics with the highest number of probabilities and are related to the overall news headline data on the covid19.go.id website, namely (1) data with a probability value of 0.097838, (2) a level with a probability value of 0.048895 and (3) vaccines with a probability value of 0.048895 probability value 0.031822.

Keywords: *covid-19, news title, classification, latent dirichlet allocation*

I. PENDAHULUAN

Arus globalisasi yang menyebar di masyarakat berupa penyampaian informasi berkembang menggunakan gelombang elektronik seperti Televisi dan Radio. Tahun 1990 hingga tahun 2000 memasuki era digital

dengan munculnya internet. Internet sebagai alat informasi dan komunikasi yang tak dapat terabaikan [1]Maka semakin berkembang pula dalam penyampaian informasi berupa teks seperti berita *online*. Berita yang diunggah di internet sudah sangat banyak dengan rentang waktu yang cepat.

Coronavirus disease 2019 atau Covid-19 merupakan penyakit yang teridentifikasi pertama kali terdeteksi di China pada akhir tahun 2019 dan telah menyebar hingga berujung pada pandemi di seluruh dunia. Sementara Covid-19 di Indonesia pertama kali dideteksi pada tanggal 02 Maret 2020 [2], ketika dua orang terkonfirmasi tertular dari seorang warga negara Jepang. Pandemi menyebar dengan cepat keseluruh provinsi di Indonesia pada tanggal 09 April 2020.

Pada masa pandemi, informasi yang tepat dan terpercaya mengenai covid-19 di Indonesia dapat dilihat pada situs resmi covid19.go.id. Pada situs tersebut terdapat berita mengenai vaksinasi covid-19, penanganan kesehatan (3M dan 3T) dan pemulihan ekonomi. Dengan banyaknya berita, diperlukan waktu untuk menentukan topik apa saja yang paling sering muncul pada berita vaksinasi covid-19, berita penanganan kesehatan (3M dan 3T), dan berita pemulihan ekonomi. Maka dari itu, penelitian ini akan memberikan solusi dalam melakukan klasifikasi trend topik pada judul berita yang diterbitkan oleh situs covid-19.go.id. Penggunaan *topic modeling* untuk mengetahui *trend* topik yang menjadi perhatian pada judul berita, sehingga mendapatkan informasi yang lebih ringkas, yang lebih luas untuk kebutuhan informasi.

Berdasarkan yang telah dijelaskan, maka didapatkan rumusan permasalahan dari penelitian ini yaitu apakah metode *latent dirichlet allocation* dapat mengklasifikasi judul berita dan mengetahui topik dari keseluruhan data.

II. STUDI PUSTAKA

A. Pengambilan Data

Tahap pertama pada penelitian ini adalah pengambilan data dari situs covid19.go.id. Yang diambil berupa judul berita dari penanganan kesehatan, vaksinasi covid-19 dan pemulihan ekonomi. Contoh salah satu data judul berita terdapat pada Tabel I.

Judul Berita
Pemerintah Percepat Bantuan Ekonomi di Masa PPKM Darurat

Tabel I Contoh Judul Berita

B. *Text PreProcessing*

Proses pada teks yang akan digunakan dipersiapkan dahulu disebut dengan proses *text preprocessing*. Tahapan *text preprocessing* mencakup dengan semua tahapan dengan mempersiapkan data yang akan digunakan [3]

1.) *Tokenization*

Tokenization atau tokenisasi merupakan proses pemisahan atau pemecahan suatu kata [4]. Tabel II merupakan contoh dari data judul berita yang telah di tokenisasi.

Pemerintah	**Percepat**	**Bantuan**	**Ekonomi**	**Di**
Masa	**PPKM**	**Darurat**		

Tabel II Contoh Tokenization

2.) *Case Folding*

Case Folding merupakan tahapan dengan mengubah seluruh huruf atau kata pada data menjadi huruf kecil semua. Tabel III merupakan contoh dari data judul berita yang telah dilakukan *case folding*.

pemerintah	**percepat**	**bantuan**	**ekonomi**	**di**
masa	**ppkm**	**darurat**		

Tabel III Contoh Case Folding

3.) *Cleaning*

Cleaning merupakan tahapan pembersihan data dari kata yang tidak diperlukan. Kata yang akan dibersihkan atau dihilangkan pada data biasanya berupa karakter simbol, angka dan link url.

4.) *Filtering*

Stopword Removal atau *filtering* merupakan proses dalam pemilihan kata-kata. Tabel IV merupakan contoh dari data judul berita yang telah dilakukan filtering dengan membuang kata-kata yang cukup umum namun tidak mempunyai pengaruh yang signifikan [5].

Pemerintah	**percepat**	**bantuan**	**ekonomi**	**masa**
Ppkm	**darurat**			

Tabel IV Contoh Filtering

5.) *Stemming*

Stemming merupakan proses dengan mencari akar dari kata yang dihasilkan dari tahapan *filtering*. Tabel V merupakan contoh dari data judul berita yang telah dilakukan proses stemming untuk mengurangi kumpulan-kumpulan teks yang dihasilkan, dengan mengubah kata imbuhan menjadi kata dasar [6].

Pemerintah	**cepat**	**bantu**	**ekonomi**	**masa**
Ppkm	**darurat**			

Tabel V Contoh Stemming

C. Pembobotan TF-IDF

Pada tahapan ini, yang digunakan untuk memberikan bobot pada suatu *term* atau kata pada data [7].. Untuk menghitung TF-IDF dapat menggunakan rumus :

$$W_{i,j} = tf_{i,j} \times ((\log \frac{N}{n}) + 1$$

Dimana,

$W_{i,j}$ bobot kata t_j terhadap dokumen d_i

tf_{ij} = jumlah kemunculan kata t_j dalam d_i

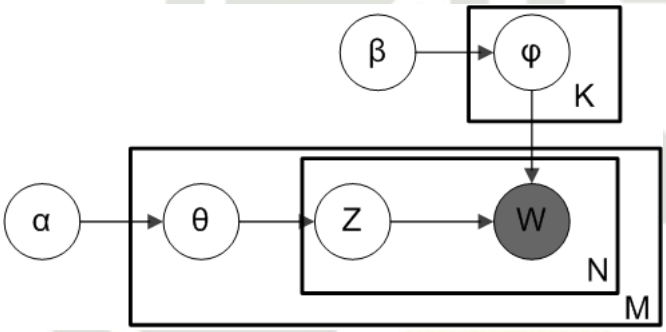
N = jumlah semua dokumen yang ada

n = jumlah dokumen yang mengandung kata t_j

D. Metode *Latent Dirichlet Allocation*

Pada tahapan ini menggunakan metode *latent dirichlet allocation* dari keseluruhan data untuk mengklasifikasi dan mengetahui topik yang sering muncul pada judul berita di website covid19.go.id. Mekanisme kerja LDA terbagi menjadi dua bagian yaitu penalaran dan implementasi. Blei [8] menjelaskan intuisi dari LDA adalah setiap dokumen terdapat berbagai topik. Dokumen pada LDA merupakan objek yang bisa diamati namun topik, penggolongan kata untuk topik per-dokumen adalah tersembunyi dan belum diketahui [9].

Menurut [8] metode LDA adalah *probabilistic model* seperti pada gambar I, parameter α dan ß merupakan parameter distribusi topik yang terdapat pada kumpulan data. Untuk parameter α digunakan untuk menentukan distribusi topik data, apabila nilai α semakin besar pada data, maka semakin banyak topik yang dibahas. Dan apabila nilai ß semakin tinggi, maka semakin banyak kata pada topik, jika semakin kecil maka semakin sedikit pada topik.



Gambar I Metode Latent Dirichlet Allocation

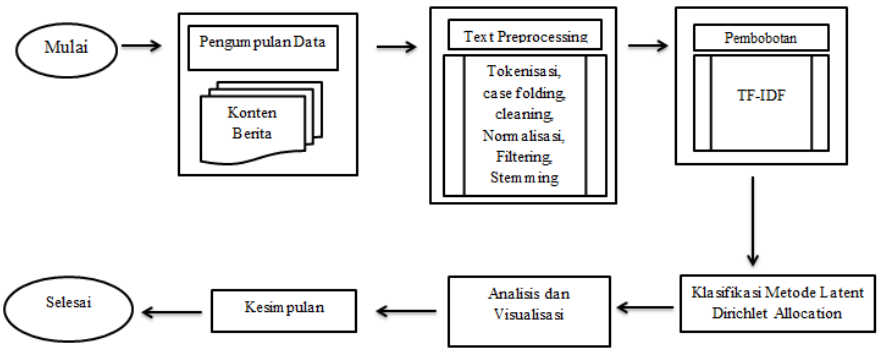
III. METODE PENELITAN

A. Pengumpulan Data

Data yang digunakan pada penelitian ini adalah sekumpulan judul berita covid-19 sejumlah 1660 judul berita. Kumpulan berita tersebut dalam bahasa Indonesia yang diambil pada situs covid19.go.id. Data judul berita penelitian ini berupa dokumen excel. Data penelitian ini akan dilakukan tahapan berupa *text preprocessing*, pembobotan *term*, penggunaan metode *latent dirichlet allocation* dan visualisasi dalam bentuk pyLDAvis.

B. Tahapan Penelitian

Tahapan penelitian merupakan tahapan-tahapan yang dilakukan pada sebuah penelitian. Tahapan penelitian menjelaskan uraian analisis-analisis mengenai permasalahan yang diselesaikan sesuai tahapan yang dilakukan. Berikut tahapan-tahapan dalam penelitian :



Gambar II Tahapan Penelitian

IV. HASIL DAN PEMBAHASAN

A. Analisa Data

Analisa kebutuhan data digunakan untuk mengetahui data yang akan digunakan pada proses klasifikasi. Penelitian ini menggunakan data berupa judul berita pada website resmi covid19.go.id dengan 3 jenis berita yang diambil secara keseluruhan dari bulan Januari 2021 hingga November 2021.

B. Latent Dirichlet Allocation

Penerapan LDA bertujuan untuk memperoleh persebaran data yang menghasilkan topik. Prinsip dasar pada metode ini adalah setiap dokumen dengan campuran topik yang tersembunyi dan belum diketahui, dan setiap topik terdiri dari banyak kata. Berikut cara kerja metode LDA pada penelitian ini

a. Membuat kumpulan-kumpulan berita *online*
b. Melakukan inisialisasi parameter seperti jumlah berita, jumlah topik, jumlah iterasi, *random state*, nilai α, nilai ß dan sebagainya.
c. Menentukan topik berdasarkan *latent dirichlet allocation*
d. Menghitung nilai probabilitas kata terhadap setiap topik
e. Memasukkan kata kedalam topik yang memiliki nilai tertinggi
f. Melakukan tahap b sampai e hingga seluruh berita dalam *corpus* telah diproses

Probabilitas suatu kata pada dokumen berita dihitung dengan melihat jumlah kata pada suatu topik dengan menggunakan persamaan berikut :

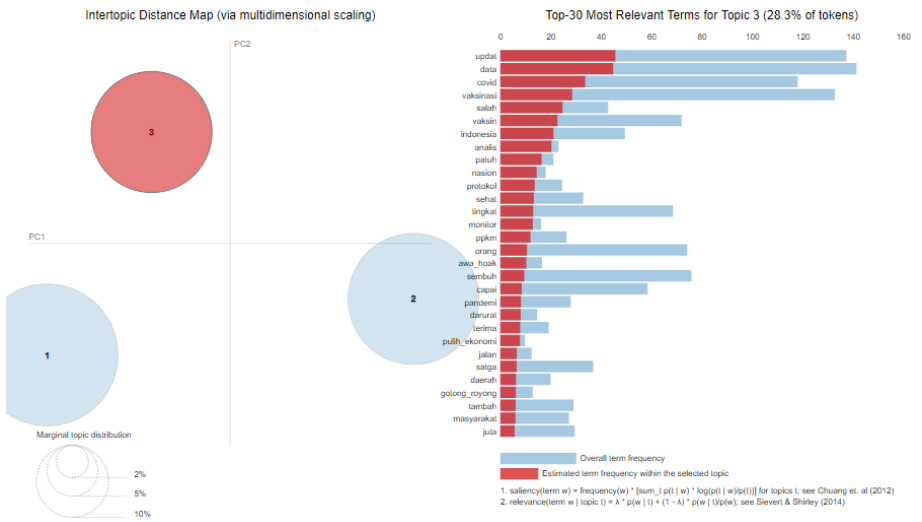
$$P(j|w_i, d_i) = \frac{C_{wij}^{WT}}{\sum_{w=1}^{W} C_{wj}^{WT}} \cdot \frac{C_{dij}^{DT}}{\sum_{t=1}^{T} C_{dit}^{DT}}$$

j = topik yang sedang di kalkulasi
w_i = kata yang sedang di kalkulasi
d_i = dokumen yang sedang di kalkulasi
C^{WT} = matriks jumlah kata dalam suatu dokumen
C^{DT} = matriks jumlah topik dalam suatu dokumen
C_{wij}^{WT} = jumlah kata w_i dalam topik j
C_{wj}^{WT} = jumlah kata w dalam topik j
C_{dij}^{DT} = jumlah kata dalam d_i termasuk topik j
C_{dit}^{DT} = jumlah kata dalam d_i termasuk topik t

C. Hasil

Data yang dikumpulkan sebanyak 1660 data judul berita dilakukan *preprocessing*. Hasil *prerprocessing* dari data akan dilakukan proses tf-idf untuk pembobotan *term* pada data. Lalu metode *latent dirichlet allocation* untuk memperoleh data yang akan menghasilkan topik dari nilai probabilitas tertinggi. Berdasarkan metode LDA yang diperoleh pada topik ke-1 terdapat pada Tabel VI dengan visualisasi *PyLDAvis* pada Gambar IV maka dapat dilihat bahwa topik yang muncul berkaitan satu sama lain mengenai update.

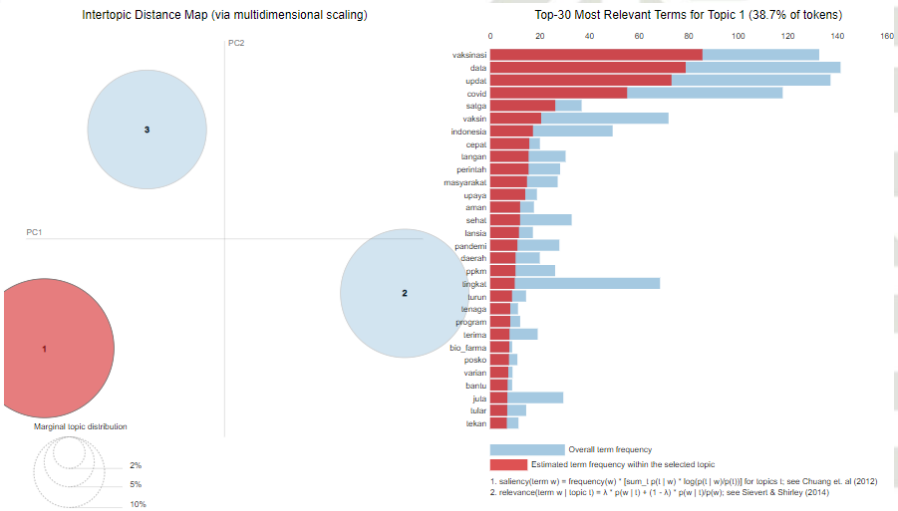
Gambar III Word Cloud Topik-1



Gambar IV Visualisasi Topik-1

Berdasarkan metode LDA yang diperoleh pada topik ke-2 terdapat pada Tabel VI dengan visualisasi *PyLDAvis* pada Gambar VI maka dapat dilihat bahwa topik yang muncul berkaitan satu sama lain mengenai vaksinasi.

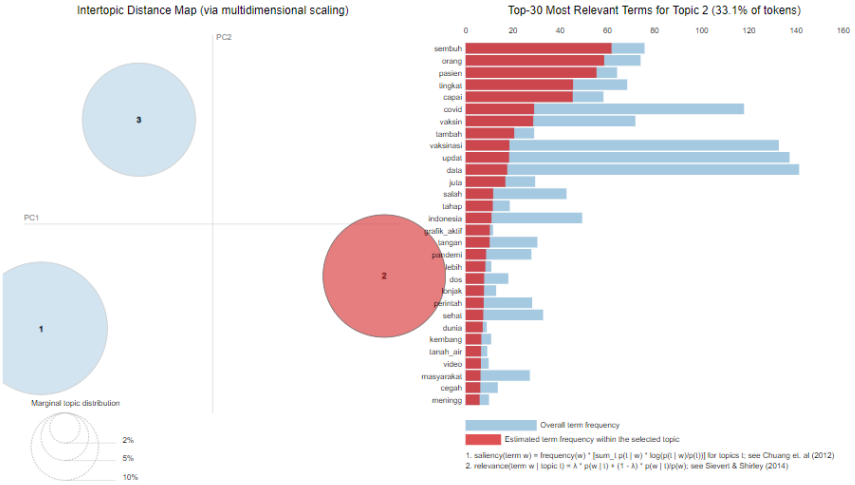
Gambar V Word Cloud Topik-2



Gambar VI Visualisasi Topik-2

Berdasarkan metode LDA yang diperoleh pada topik ke-3 terdapat pada Tabel VI dengan visualisasi *PyLDAvis* pada Gambar VIII maka dapat dilihat bahwa topik yang muncul berkaitan satu sama lain mengenai sembuh.

Gambar VII Word Cloud Topik 3



Gambar VIII Visualisasi Topik-3

Berdasarkan hasil dari penelitian, didapatkan 3 topik dengan nilai probabilitas tertinggi dari keseluruhan data dapat dilihat pada Tabel VI.

Topik	Kata	Probabilitas
1	Update	0,047363
2	Vaksinasi	0,064872
3	Sembuh	0,054829

Tabel VI Hasil Penelitian

Lalu dilakukan pengujian data berupa judul berita "BPOM Terbitkan Izin Vaksin Sinovac Untuk 6-11 Tahun" didapatkan nilai akurasi nya sebesar 84.62% dari topik Pasien,, 8.42% dari topik Data dan 6.95% dari topik Vaksin dapat dilihat pada Tabel VII.

Score	Kata	Probabilitas
85,58%	Vaksinasi	0,065
7,74%	Update	0,047
6,69%	Sembuh	0,055

Tabel VII Pengujian Data

V. KESIMPULAN

Berdasarkan tahapan-tahapan yang telah dilakukan, Metode *Latent Dirichlet Allocation* dapat diterapkan untuk klasifikasi judul berita covid-19. Berdasarkanasil dari pemodelan topik menggunakan *Latent Dirichlet Allocation* pada judul berita covid-19 diperoleh 3 topik sebagai berikut Model LDA topik ke-1 update dengan nilai probabilitas 0,047363, Model LDA topik ke-2 vaksinasi dengan nilai probabilitas 0,064872 dan Model LDA topik ke-3 sembuh dengan nilai probabilitas 0,054829. Hasil Pengujian pada salah satu judul didapatkan dengan nilai akurasi pada topik vaksinasi sebesar 85,58%, topik Update sebesar 7,74% dan topik Sembuh sebesar 6,69%.

DAFTAR PUSTAKA

[1] Ardianto, K. Lukiati and Elvinaro, Komunikasi Massa Suatu Pengantar, Bandung: Simbiosa Rekatama Media, 2004.

[2] "bbccom," 25 Juni 2021. [Online]. Available: https://www.bbc.com/indonesia/dunia-57590872.

[3] R. Feldman and J. Sanger, "The text mining handbook," *Advanced aproaches in analyzing unstructured data*, 2007.

[4] A. T. Jaka H, "Preprocessing Text untuk Meminimalisir Kata Yang Tidak Berarti dalam Proses Text Mining," 2015.

[5] H. Maarifah, A. P. Wibawa and M. I. Akbar, "Klasifikasi artikel ilmiah dengan berbagai skenario preprocessing," *Jurnal Sakti FKTI*, pp. 70-78, 2020.

[6] F. Z. Tala, A study of Stemming Efects on Information Retrieval in Bahasa Indonesia, Amsterdam: Institute for Logic, Language and Computation, 2003.

[7] C. Kamayasa and I. B. Mahendra, "Implementasi Vector Space Model dan Beberapa Notasi Metode Term Frequency Inverse Document Frequency (TF-IDF) pada Sistem Temu Kembali Informasi," *Jurnal Elektronik Ilmu Komputer Udayana*, 2012.

[8] D. M. Blei and M. I. Jordan, "Latent Dirichlet Allocation," pp. 993-1022, 2012.

[9] Campbell, H. J. C and E. Stroulia, Latent Dirichlet Allocation: Extracting Topics from Software Engineering Data, 2014.