



**PROGRAM STUDI S1 MAJOR ILMU KOMPUTER  
DEPARTEMEN ILMU KOMPUTER, FMIPA-IPB**  
Jl. Meranti, Kampus IPB Darmaga, Wing 20 Level V,  
Darmaga 16680 Bogor  
Telp/Fax (0251) 8625584, e-mail: ilkom@ipb.ac.id

**FORMULIR KESIAPAN UJIAN TUGAS AKHIR (KOM499)  
TAHUN AKADEMIK 2021/2022**

Yang bertanda tangan di bawah ini, Pembimbing & Penguji Tugas Akhir mahasiswa Mayor Ilmu Komputer menyatakan bahwa:

Nama Mahasiswa : Amin Elhan  
NIM : G64170109  
Bidang Kajian : Kecerdasan Komputasional dan Optimasi (CIO)

telah SIAP DIUJI pada:

Hari, Tanggal : Senin, 19 September 2022  
P u k u l : 10.00 - 12.00 WIB  
T e m p a t : Google Meeting Zoom  
dengan Dosen Penguji sebagai berikut:

Penguji 1:

Nama : Sony Hartono Wijaya, S.Kom, M.Kom, Ph.D  
NIP : 19810809 200812 1 002  
Tanda tangan :



Demikian surat keterangan ini dibuat untuk dipergunakan sebagaimana mestinya.

Bogor, 15 September 2022  
Komisi Pembimbing,



Medria K. D. Hardhienata, S.Komp. Ph.D  
NIP. 19860822 202012 2 001



**PROGRAM STUDI S1 MAYOR ILMU KOMPUTER  
DEPARTEMEN ILMU KOMPUTER, FMIPA-IPB**

Jl. Meranti, Kampus IPB Darmaga, Wing 20 Level V,  
Darmaga 16680 Bogor  
Telp/Fax (0251) 8625584, e-mail: ilkom@ipb.ac.id

**KELENGKAPAN UJIAN TUGAS AKHIR (KOM499)  
TAHUN AKADEMIK 2021/2022**

Nama : Amin Elhan  
NIM : G64170109

No	Persyaratan Ujian	Keterangan*)
1.	Telah selesai melaksanakan kegiatan Praktik Kerja Lapang (PKL) (salinan daftar nilai PKL terlampir)	✓
2.	Telah melaksanakan seminar tugas akhir (salinan nilai seminar terlampir)	✓
3.	Telah menyelesaikan sekurang-kurangnya 142 SKS. IPK minimum 2.00 dan tanpa nilai E (salinan transkrip seluruh semester terlampir)	✓
4.	Telah menyerahkan formulir kesiapan ujian tugas akhir yang telah ditanda-tangani oleh Ketua Komisi Pembimbing	✓
5.	Menyerahkan 1 (satu) eksemplar laporan yang siap diuji, ditanda-tangani oleh Ketua Komisi Pembimbing	✓
6.	Telah melunasi SPP semester terakhir (salinan bukti pembayaran SPP terlampir)	✓

Bogor, 15 September 2022

Komisi Pendidikan Program S1,



Hari Agung Ardianto S.Kom, M.Si, Ph.D  
NIP: 19760917 200501 1 001

Keterangan:

- \*) Beri tanda ✓ pada kolom keterangan jika mahasiswa yang bersangkutan telah melengkapi persyaratan ujian tugas akhir.

## Nilai KKN

1	FMP400	Pengantar Bioinformatika	Interdepartemen	<span style="color: green;">A</span>	3 (2 - 1)	4	12
2	IKK334	Manajemen Sumberdaya Keluarga	SC	<span style="color: yellow;">C</span>	3 (3 - 0)	2	6
3	KOM302	Etika Komputasi	Mayor	<span style="color: green;">AB</span>	2 (2 - 0)	3.5	7
4	KOM311	Sistem Operasi *)	Mayor	<span style="color: green;">B</span>	3 (2 - 1)	3	9
5	KOM330	Manajemen Proyek Perangkat Lunak	Mayor	<span style="color: green;">B</span>	3 (2 - 1)	3	9
6	KOM398	Metode Penelitian dan Telaah Pustaka	Mayor	<span style="color: green;">AB</span>	2 (2 - 0)	3.5	7
7	KOM401	Analisis Algoritme	Mayor	<span style="color: yellow;">BC</span>	3 (2 - 1)	2.5	7.5
8	MAN111	Pengantar Manajemen	SC	<span style="color: yellow;">BC</span>	3 (3 - 0)	2.5	7.5
9	IPB400	Kuliah Kerja Nyata Tematik (SP)	Interdepartemen	<span style="color: green;">A</span>	4 (1 - 9)	4	16
<b>IP Semester : 2.95</b> <b>IP Kumulatif : 2.98</b>					<b>SKS Kumulatif : 140</b>	<b>22</b>	<b>65</b>
<b>Kelanjutan Studi: Tanpa Syarat</b>							

## Nilai Seminar

2021/2022 Genap							
1	KOM497	Kolokium	Mayor	<span style="color: green;">A</span>	1 (0 - 1)	4	4
2	KOM498	Seminar	Mayor	<span style="color: green;">A</span>	1 (0 - 1)	4	4
3	KOM499	Tugas Akhir	Mayor	<span style="color: grey;">BL</span>	4 (0 - 4)	0	0
<b>IP Semester : 4</b> <b>IP Kumulatif : 3.01</b>					<b>SKS Kumulatif : 148</b>	<b>2</b>	<b>8</b>
<b>Kelanjutan Studi: Tanpa Syarat</b>							



**KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI**  
**INSTITUT PERTANIAN BOGOR**  
**DAFTAR PRESTASI AKADEMIK MAHASISWA**

Nama : Amin Elhan  
NIM : G64170109  
Semester : Semester 1 (2017/2018 Ganjil)  
Program Studi : Ilmu Komputer

No.	Kode Mata Kuliah	Nama Mata Kuliah	sks	Huruf Mutu	Angka Mutu	Nilai Mutu
1	AGB100	Pengantar Kewirausahaan	1	A	4.00	4.00
2	BIO101	Biologi Umum	2	AB	3.50	7.00
3	FIS100	Fisika	3	AB	3.50	10.50
4	IPB100	Pendidikan Agama Islam	3	A	4.00	12.00
5	IPB107	Pengantar Ilmu Pertanian	2	A	4.00	8.00
6	IPB111	Pendidikan Pancasila	2	AB	3.50	7.00
7	KPM130	Sosiologi Umum	3	BC	2.50	7.50
8	MAT100	Pengantar Matematika	3	A	4.00	12.00
Total sks			19			
sks Kumulatif			19			
Indeks Prestasi (IP)			3.58			
IP Kumulatif			3.58			
Status Kelanjutan Studi			Tanpa Syarat			

Keterangan :

A : 4.00                    BL : Belum Lengkap  
AB : 3.50                 \* : Mengulang  
B : 3.00  
BC : 2.50  
C : 2.00  
D : 1.00  
E : 0.00

Bogor, 04 Agustus 2022

Direktur Administrasi Pendidikan  
dan Penerimaan Mahasiswa Baru

  
digitally signed  
dsign.ipb.ac.id

Dr. Ir. Nurhayati, M.Sc.  
196201211986012001



**KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI  
INSTITUT PERTANIAN BOGOR  
DAFTAR PRESTASI AKADEMIK MAHASISWA**

Nama : Amin Elhan  
NIM : G64170109  
Semester : Semester 2 (2017/2018 Genap)  
Program Studi : Ilmu Komputer

No.	Kode Mata Kuliah	Nama Mata Kuliah	skls	Huruf Mutu	Angka Mutu	Nilai Mutu
1	EKO100	Ekonomi Umum	3	A	4.00	12.00
2	IPB106	Bahasa Indonesia	2	A	4.00	8.00
3	IPB108	Bahasa Inggris	3	B	3.00	9.00
4	IPB112	Olahraga dan Seni	1	A	4.00	4.00
5	KIM100	Kimia Umum	2	AB	3.50	7.00
6	KOM101	Algoritme	3	B	3.00	9.00
7	KOM201	Penerapan Komputer	3	A	4.00	12.00
8	MAT103	Kalkulus	3	AB	3.50	10.50
Total sks			20			
sks Kumulatif			39			
Indeks Prestasi (IP)			3.58			
IP Kumulatif			3.58			
Status Kelanjutan Studi			Tanpa Syarat			

Keterangan :

A : 4.00      BL : Belum Lengkap  
AB : 3.50     \* : Mengulang  
B : 3.00  
BC : 2.50  
C : 2.00  
D : 1.00  
E : 0.00

Bogor, 12 Agustus 2022

Direktur Administrasi Pendidikan  
dan Penerimaan Mahasiswa Baru

Dr. Ir. Nurhayati, M.Sc.  
196201211986012001



**KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI  
INSTITUT PERTANIAN BOGOR  
DAFTAR PRESTASI AKADEMIK MAHASISWA**

Nama : Amin Elhan  
NIM : G64170109  
Semester : Semester 3 (2018/2019 Ganjil)  
Program Studi : Ilmu Komputer

No.	Kode Mata Kuliah	Nama Mata Kuliah	sks	Huruf Mutu	Angka Mutu	Nilai Mutu
1	KOM200	Dasar Pemrograman	3	C	2.00	6.00
2	KOM203	Rangkaian Digital	3	B	3.00	9.00
3	KOM209	Struktur Diskret	3	BC	2.50	7.50
4	KOM220	Pengantar Matematika Komputasi	3	B	3.00	9.00
5	MAT219	Aljabar Linier	3	C	2.00	6.00
6	STK202	Pengantar Hitung Peluang	3	A	4.00	12.00
7	STK211	Metode Statistika	3	C	2.00	6.00
Total sks			21			
sks Kumulatif			60			
Indeks Prestasi (IP)			2.64			
IP Kumulatif			3.25			
Status Kelanjutan Studi			Tanpa Syarat			

Keterangan :

A : 4.00      BL : Belum Lengkap  
AB : 3.50     \* : Mengulang  
B : 3.00  
BC : 2.50  
C : 2.00  
D : 1.00  
E : 0.00

Bogor, 12 Agustus 2022

Direktur Administrasi Pendidikan  
dan Penerimaan Mahasiswa Baru

Dr. Ir. Nurhayati, M.Sc.  
196201211986012001



**KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI  
INSTITUT PERTANIAN BOGOR  
DAFTAR PRESTASI AKADEMIK MAHASISWA**

Nama : Amin Elhan  
NIM : G64170109  
Semester : Semester 4 (2018/2019 Genap)  
Program Studi : Ilmu Komputer

No.	Kode Mata Kuliah	Nama Mata Kuliah	sk	Huruf Mutu	Angka Mutu	Nilai Mutu
1	KOM204	Bahasa Pemrograman	3	C	2.00	6.00
2	KOM205	Basis Data	3	AB	3.50	10.50
3	KOM206	Organisasi dan Arsitektur Komputer	3	C	2.00	6.00
4	KOM207	Struktur Data	3	BC	2.50	7.50
5	KOM322	Metode Kuantitatif	3	BC	2.50	7.50
6	KOM331	Rekayasa Perangkat Lunak	3	A	4.00	12.00
7	MAT234	Graf Algoritmik	3	D	1.00	3.00
Total sks			21			
sks Kumulatif			81			
Indeks Prestasi (IP)			2.50			
IP Kumulatif			3.06			
Status Kelanjutan Studi			Tanpa Syarat			

Keterangan :

A : 4.00      BL : Belum Lengkap  
AB : 3.50     \* : Mengulang  
B : 3.00  
BC : 2.50  
C : 2.00  
D : 1.00  
E : 0.00

Bogor, 12 Agustus 2022

Direktur Administrasi Pendidikan  
dan Penerimaan Mahasiswa Baru

Dr. Ir. Nurhayati, M.Sc.  
196201211986012001



**KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI  
INSTITUT PERTANIAN BOGOR  
DAFTAR PRESTASI AKADEMIK MAHASISWA**

Nama : Amin Elhan  
NIM : G64170109  
Semester : Semester 5 (2019/2020 Ganjil)  
Program Studi : Ilmu Komputer

No.	Kode Mata Kuliah	Nama Mata Kuliah	sks	Huruf Mutu	Angka Mutu	Nilai Mutu
1	KOM300	Grafika Komputer	3	C	2.00	6.00
2	KOM311	Sistem Operasi	3	E	0.00	0.00
3	KOM321	Kecerdasan Buatan	3	C	2.00	6.00
4	KOM325	Komputasi Numerik	3	C	2.00	6.00
5	KOM333	Interaksi Manusia dan Komputer	3	B	3.00	9.00
6	KOM335	Sistem Informasi	3	B	3.00	9.00
Total sks			18			
sks Kumulatif			99			
Indeks Prestasi (IP)			2.00			
IP Kumulatif			2.86			
Status Kelanjutan Studi			Tanpa Syarat			

Keterangan :

A : 4.00      BL : Belum Lengkap  
AB : 3.50     \* : Mengulang  
B : 3.00  
BC : 2.50  
C : 2.00  
D : 1.00  
E : 0.00

Bogor, 12 Agustus 2022

Direktur Administrasi Pendidikan  
dan Penerimaan Mahasiswa Baru

Dr. Ir. Nurhayati, M.Sc.  
196201211986012001



**KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI  
INSTITUT PERTANIAN BOGOR  
DAFTAR PRESTASI AKADEMIK MAHASISWA**

Nama : Amin Elhan  
NIM : G64170109  
Semester : Semester 6 (2019/2020 Genap)  
Program Studi : Ilmu Komputer

No.	Kode Mata Kuliah	Nama Mata Kuliah	sks	Huruf Mutu	Angka Mutu	Nilai Mutu
1	GFM221	Klimatologi	3	AB	3.50	10.50
2	IKK233	Perilaku Konsumen	3	B	3.00	9.00
3	KOM320	Sistem Cerdas	3	B	3.00	9.00
4	KOM324	Pengolahan Citra Digital	3	D	1.00	3.00
5	KOM332	Data Mining	3	B	3.00	9.00
6	KOM334	Pengembangan Sistem Berorientasi Objek	3	A	4.00	12.00
Total sks			18			
sks Kumulatif			117			
Indeks Prestasi (IP)			2.92			
IP Kumulatif			2.87			
Status Kelanjutan Studi			Tanpa Syarat			

Keterangan :

A : 4.00      BL : Belum Lengkap  
AB : 3.50      \* : Mengulang  
B : 3.00  
BC : 2.50  
C : 2.00  
D : 1.00  
E : 0.00

Bogor, 12 Agustus 2022

Direktur Administrasi Pendidikan  
dan Penerimaan Mahasiswa Baru

Dr. Ir. Nurhayati, M.Sc.  
196201211986012001



**KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI**  
**INSTITUT PERTANIAN BOGOR**  
**DAFTAR PRESTASI AKADEMIK MAHASISWA**

Nama : Amin Elhan  
NIM : G64170109  
Semester : Semester 7 (2020/2021 Ganjil)  
Program Studi : Ilmu Komputer

No.	Kode Mata Kuliah	Nama Mata Kuliah	sks	Huruf Mutu	Angka Mutu	Nilai Mutu
1	FMP400	Pengantar Bioinformatika	3	A	4.00	12.00
2	IKK334	Manajemen Sumberdaya Keluarga	3	C	2.00	6.00
3	IPB400	Kuliah Kerja Nyata Tematik	4	A	4.00	16.00
4	KOM302	Etika Komputasi	2	AB	3.50	7.00
5	KOM311	Sistem Operasi *)	3	B	3.00	9.00
6	KOM330	Manajemen Proyek Perangkat Lunak	3	B	3.00	9.00
7	KOM398	Metode Penelitian dan Telaah Pustaka	2	AB	3.50	7.00
8	KOM401	Analisis Algoritme	3	BC	2.50	7.50
9	MAN111	Pengantar Manajemen	3	BC	2.50	7.50
Total sks			26			
sks Kumulatif			140			
Indeks Prestasi (IP)			2.95			
IP Kumulatif			2.98			
Status Kelanjutan Studi			Tanpa Syarat			

Keterangan :

A : 4.00      BL : Belum Lengkap  
AB : 3.50     \* : Mengulang  
B : 3.00  
BC : 2.50  
C : 2.00  
D : 1.00  
E : 0.00

Bogor, 12 Agustus 2022

Direktur Administrasi Pendidikan  
dan Penerimaan Mahasiswa Baru

Dr. Ir. Nurhayati, M.Sc.  
196201211986012001



**KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI  
INSTITUT PERTANIAN BOGOR  
DAFTAR PRESTASI AKADEMIK MAHASISWA**

Nama : Amin Elhan  
NIM : G64170109  
Semester : Semester 8 (2020/2021 Genap)  
Program Studi : Ilmu Komputer

No.	Kode Mata Kuliah	Nama Mata Kuliah	sks	Huruf Mutu	Angka Mutu	Nilai Mutu
1	KOM312	Komunikasi Data dan Jaringan Komputer	3	C	2.00	6.00
2	KOM324	Pengolahan Citra Digital *)	3	C	2.00	6.00
3	KOM497	Kolokium	1	BL	0.00	0.00
4	KOM498	Seminar	1	BL	0.00	0.00
5	KOM499	Tugas Akhir	4	BL	0.00	0.00
6	MAT234	Graf Algoritmik *)	3	D	1.00	3.00
Total sks			15			
sks Kumulatif			143			
Indeks Prestasi (IP)			1.67			
IP Kumulatif			2.98			
Status Kelanjutan Studi			Tanpa Syarat			

Keterangan :

A : 4.00      BL : Belum Lengkap  
AB : 3.50      \* : Mengulang  
B : 3.00  
BC : 2.50  
C : 2.00  
D : 1.00  
E : 0.00

Bogor, 12 Agustus 2022

Direktur Administrasi Pendidikan  
dan Penerimaan Mahasiswa Baru

Dr. Ir. Nurhayati, M.Sc.  
196201211986012001



**KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI  
INSTITUT PERTANIAN BOGOR  
DAFTAR PRESTASI AKADEMIK MAHASISWA**

Nama : Amin Elhan  
NIM : G64170109  
Semester : Semester 9 (2021/2022 Ganjil)  
Program Studi : Ilmu Komputer

No.	Kode Mata Kuliah	Nama Mata Kuliah	sks	Huruf Mutu	Angka Mutu	Nilai Mutu
1	IPB308	Kepemimpinan Inklusif & Inovatif	3	A	4.00	12.00
2	KOM497	Kolokium	1	BL	0.00	0.00
3	KOM498	Seminar	1	BL	0.00	0.00
4	KOM499	Tugas Akhir	4	BL	0.00	0.00
Total sks			9			
sks Kumulatif			146			
Indeks Prestasi (IP)			4.00			
IP Kumulatif			3.00			
Status Kelanjutan Studi			Tanpa Syarat			

Keterangan :

A : 4.00      BL : Belum Lengkap  
AB : 3.50      \* : Mengulang  
B : 3.00  
BC : 2.50  
C : 2.00  
D : 1.00  
E : 0.00

Bogor, 12 Agustus 2022

Direktur Administrasi Pendidikan  
dan Penerimaan Mahasiswa Baru

Dr. Ir. Nurhayati, M.Sc.  
196201211986012001



**KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI  
INSTITUT PERTANIAN BOGOR  
DAFTAR PRESTASI AKADEMIK MAHASISWA**

Nama : Amin Elhan  
NIM : G64170109  
Semester : Semester 10 (2021/2022 Genap)  
Program Studi : Ilmu Komputer

No.	Kode Mata Kuliah	Nama Mata Kuliah	sks	Huruf Mutu	Angka Mutu	Nilai Mutu
1	KOM497	Kolokium	1	A	4.00	4.00
2	KOM498	Seminar	1	A	4.00	4.00
3	KOM499	Tugas Akhir	4	BL	0.00	0.00
		Total sks	6			
		sks Kumulatif	148			
		Indeks Prestasi (IP)	4.00			
		IP Kumulatif	3.01			
		Status Kelanjutan Studi	Tanpa Syarat			

Keterangan :

A : 4.00      BL : Belum Lengkap  
AB : 3.50     \* : Mengulang  
B : 3.00  
BC : 2.50  
C : 2.00  
D : 1.00  
E : 0.00

Bogor, 12 Agustus 2022

Direktur Administrasi Pendidikan  
dan Penerimaan Mahasiswa Baru

Dr. Ir. Nurhayati, M.Sc.  
196201211986012001



**KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI  
INSTITUT PERTANIAN BOGOR  
DAFTAR PRESTASI AKADEMIK MAHASISWA**

Nama : Amin Elhan  
NIM : G64170109  
Semester : Semester 11 (2022/2023 Ganjil)  
Program Studi : Ilmu Komputer

No.	Kode Mata Kuliah	Nama Mata Kuliah	sks	Huruf Mutu	Angka Mutu	Nilai Mutu
1	KOM499	Tugas Akhir	4	BL	0.00	0.00
		Total sks	4			
		sks Kumulatif	148			
		Indeks Prestasi (IP)	0.00			
		IP Kumulatif	3.01			
		Status Kelanjutan Studi				

Keterangan :

A : 4.00      BL : Belum Lengkap  
AB : 3.50     \* : Mengulang  
B : 3.00  
BC : 2.50  
C : 2.00  
D : 1.00  
E : 0.00

Bogor, 15 September 2022

Direktur Administrasi Pendidikan  
dan Penerimaan Mahasiswa Baru

Dr. Ir. Nurhayati, M.Sc.  
196201211986012001



**ANALISIS TOPIK DAN SENTIMEN PENGGUNA TWITTER  
TERHADAP VAKSINASI COVID-19 DI INDONESIA MENGGUNAKAN  
LATENT DIRICHLET ALLOCATION DAN RANDOM FOREST**

**AMIN ELHAN**



**DEPARTEMEN ILMU KOMPUTER  
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM  
INSTITUT PERTANIAN BOGOR  
BOGOR  
2022**



## **PERNYATAAN MENGENAI SKRIPSI DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA**

Dengan ini saya menyatakan bahwa skripsi dengan judul “Analisis Topik dan Sentimen Pengguna Twitter Terhadap Vaksinasi COVID-19 di Indonesia Menggunakan *Latent Dirichlet Allocation* dan *Random Forest* adalah karya saya dengan arahan dari dosen pembimbing dan belum diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir skripsi ini.

Dengan ini saya melimpahkan hak cipta dari karya tulis saya kepada Institut Pertanian Bogor.

Bogor, September 2022

Amin Elhan  
G64170109

## ABSTRAK

AMIN ELHAN. Analisis Topik dan Sentimen Pengguna Twitter terhadap Vaksinasi COVID-19 di Indonesia Menggunakan *Latent Dirichlet Allocation* dan *Random Forest*. Dibimbing oleh MEDRIA KUSUMA DEWI HARDHIENATA dan YENI HERDIYENI

Pandemi Covid-19 mendorong banyak pihak agar mampu beradaptasi dengan kondisi terkini. Salah satu program yang diluncurkan pemerintah agar dapat mengatasi penyebaran Covid-19 adalah dengan menjalankan program vaksinasi. Agar dapat mengetahui animo masyarakat terkait program vaksinasi Covid-19 yang diluncurkan, perlu dilakukan analisis topik dan analisis sentimen. Analisis sentimen pada umumnya dilakukan untuk mendapatkan informasi terkini dari korpus yang besar. Tujuan penelitian ini adalah mengetahui topik-topik terkait vaksin Covid-19 yang dibicarakan masyarakat di Twitter dan melakukan analisis sentimen pengguna Twitter terhadap vaksinasi corona. Untuk mendapatkan topik-topik pembicaraan terkait vaksin Covid-19 digunakan metode *Latent Dirichlet Allocation* (LDA), sedangkan untuk melakukan analisis sentimen digunakan algoritme *Random Forest*. Metode penelitian yang dilakukan meliputi praproses data, pelabelan sentimen, penentuan jumlah, pemodelan topik, dan analisis topik. Hasil dari penelitian yang dilakukan adalah berupa topik-topik terkait vaksinasi Covid-19 yang sedang diperbincangkan di media Twitter di Indonesia. Selain itu hasil analisis sentimen pengguna Twitter terhadap vaksinasi Covid-19 di Indonesia menggunakan algoritme *Random Forest* menghasilkan akurasi 85% dan F1-score sebesar 81%.

Kata Kunci : analisis topik, analisis sentimen, Covid-19, vaksinasi, LDA, Random Forest, Twitter

## ABSTRACT

AMIN ELHAN. Sentiment and Topic Analysis of Twitter User About COVID-19 Vaccination in Indonesia Using *Latent Dirichlet Allocation* and *Random Forest*. Supervised by MEDRIA KUSUMA DEWI HARDHIENATA and YENI HERDIYENI.

The Covid-19 pandemic has pushed many stakeholders to be able to adapt to the current conditions. One of the programs launched by the government in order to overcome the spread of Covid-19 is to run a vaccination program. In order to know the public's interest regarding the Covid-19 vaccination program that was launched, it is necessary to conduct a topic analysis and sentiment analysis. Sentiment analysis is generally carried out to obtain up-to-date information from a large corpus. The purpose of this study was to find out the topics related to the Covid-19 vaccine that were discussed by the public on Twitter and to analyze the sentiments of Twitter users towards corona vaccination. To get topics of discussion related to the Covid-19 vaccine, *Latent Dirichlet Allocation (LDA)* method is used, while the *Random Forest* algorithm is used to carry out sentiment analysis. The research methods include data preprocessing, sentiment labeling, number determination, topic modeling, and topic analysis. The results of the research carried out are in the form of topics related to Covid-19 vaccination which are being discussed on Twitter media in Indonesia. In addition, the results of the analysis of Twitter users' sentiment on Covid-19 vaccination in Indonesia using the Random Forest algorithm yielded an accuracy of 85% and an F1-score of 81%.

Keywords : topic analysis, sentiment analysis Covid-19, vaccination, LDA, Random Forest, Twitter

© Hak Cipta milik IPB, tahun 2022  
Hak Cipta dilindungi Undang-Undang

*Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan atau menyebutkan sumbernya. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik, atau tinjauan suatu masalah, dan pengutipan tersebut tidak merugikan kepentingan IPB.*

*Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apa pun tanpa izin IPB.*

**ANALISIS SENTIMEN PENGGUNA TWITTER TERHADAP  
VAKSINASI COVID-19 DI INDONESIA MENGGUNAKAN  
PEMODELAN TOPIK LATENT DIRICHLET ALLOCATION**

**AMIN ELHAN**

Skripsi  
sebagai salah satu syarat untuk memperoleh gelar  
Sarjana pada  
Program Studi Ilmu Komputer

**DEPARTEMEN ILMU KOMPUTER  
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM  
INSTITUT PERTANIAN BOGOR  
BOGOR  
2022**

Tim Penguji pada Ujian Skripsi:

1 Sony Hartono Wijaya, S.Kom, M.Kom, Ph.D



Judul Skripsi : Analisis Topik dan Sentimen Pengguna Twitter terhadap  
Vaksinasi COVID-19 di Indonesia Menggunakan *Latent Dirichlet  
Allocation* dan *Random Forest*

Nama : Amin Elhan  
NIM : G64170109

Disetujui oleh

Pembimbing 1:  
Medria K. D. Hardhienata, S.Komp. Ph.D

Pembimbing 2:  
Dr. Yeni Herdiyeni, S.Si, M.Komp

Diketahui oleh

Ketua Ketua Departemen Ilmu Komputer :  
Dr. Sony Hartono Wijaya, S.Komp., M.Kom  
NIP. 19810809 200812 1 002

Tanggal Ujian:  
19 September 2022

Tanggal Lulus:  
(tanggal penandatanganan oleh Dekan  
Fakultas/Sekolah ...)

## PRAKATA

Puji dan syukur penulis panjatkan kepada Allah subhanaahu wa ta'ala atas segala karunia-Nya sehingga karya ilmiah ini berhasil diselesaikan. Tema yang dipilih dalam penelitian yang dilaksanakan sejak bulan September 2020 sampai bulan September 2022 ini ialah analisis Twitter, dengan judul “Analisis Topik dan Sentimen Pengguna Twitter terhadap Vaksinasi COVID-19 di Indonesia Menggunakan *Latent Dirichlet Allocation* dan *Random Forest*”.

Terima kasih penulis ucapan kepada para pembimbing, Ibu Medria K.D. Hardhienata, S.Komp. Ph.D dan Dr. Yeni Herdiyeni, S.Si., M.Kom. yang telah membimbing dan banyak memberi saran. Ucapan terima kasih juga disampaikan kepada pembimbing akademik, moderator seminar, dan penguji luar komisi pembimbing. Ungkapan terima kasih juga disampaikan kepada kedua orang tua, Bapak Hansastri dan Ibu Elimarlina, kedua adik, serta seluruh keluarga yang telah memberikan dukungan, doa, dan kasih sayangnya sehingga penelitian ini dapat diselesaikan. Terima kasih tak lupa penulis sampaikan kepada rekan-rekan S1 IPB Angkatan 54 dan 55 yang membantu dalam proses pelabelan data, rekan-rekan BEM FMIPA IPB, dan sahabat lainnya yang tak bisa disebut satu per satu, yang telah menyertai dan menemani setiap Langkah penulis dalam menempuh studi.

Semoga karya ilmiah ini bermanfaat bagi pihak yang membutuhkan dan bagi kemajuan ilmu pengetahuan..

Bogor, September 2022

*Amin Elhan*

## DAFTAR ISI

<b>I</b>	<b>PENDAHULUAN</b>	<b>1</b>
1.1	Latar Belakang	1
1.2	Rumusan Masalah	2
1.3	Tujuan	2
1.4	Manfaat	2
1.5	Ruang Lingkup	3
<b>II</b>	<b>TINJAUAN PUSTAKA</b>	<b>4</b>
2.1	Analisis Topik Covid-19	4
2.2	Analisis Sentimen Covid-19	4
2.3	<i>Latent Dirichlet Allocation</i>	4
2.4	Pemodelan Topik	6
2.5	<i>Online Variational Inference for Latent Dirichlet Allocation</i>	7
2.6	Ukuran Koherensi Topik	7
2.7	Visualisasi Pemodelan Topik	7
2.8	<i>Term Frequency – Inverse Document Frequency (TF-IDF)</i>	8
2.9	<i>Random Forest</i>	9
2.10	<i>Bidirectional Encoder Representation from Transformer (BERT)</i>	9
2.11	K-Fold Cross Validation	10
2.12	<i>Confusion Matrix</i>	10
2.13	<i>Algoritme Synthetic Minority Oversampling Technique)</i>	11
<b>III</b>	<b>METODE</b>	<b>11</b>
3.1	Tahapan Penelitian	11
3.2	Pengumpulan Data	12
3.3	Pelabelan Data	12
3.4	Eksplorasi Data	13
3.5	Praproses Data	13
3.6	<i>Bag of Words (BOW)</i>	14
3.7	Mengubah Nilai BOW Menjadi TF-IDF	14
3.8	Mengukur Koherensi Topik	15
3.9	Analisis Topik dengan <i>Latent Dirichlet Allocation</i>	15
3.10	Analisis Sentimen dengan <i>Random Forest</i>	16
3.11	Analisis Sentimen dengan BERT	16
3.12	Hyperparameter Tuning	16
3.13	Evaluasi Model Klasifikasi	16
3.14	Lingkungan Pengembangan	17
<b>IV</b>	<b>HASIL DAN PEMBAHASAN</b>	<b>18</b>
4.1	Eksplorasi Data	18
4.2	Koherensi Topik	18
4.3	Pemodelan Topik	19
4.4	Pembagian Data	20
4.5	<i>Hyperparameter Tuning</i>	20
4.6	Evaluasi Model Klasifikasi dengan Data Asli	21
4.7	Evaluasi Model Klasifikasi dengan <i>Random Under Sampling</i>	21
4.8	Evaluasi Model Klasifikasi dengan <i>Random Over Sampling</i>	22

4.9	Evaluasi Model Klasifikasi dengan <i>Over Sampling</i> SMOTE	22
4.10	Perbandingan Evaluasi Model Random Forest <i>Imbalance Data</i>	23
4.11	Perbandingan Klasifikasi dan Waktu <i>RF</i> dengan BERT	23
V	SIMPULAN DAN SARAN	25
5.1	Simpulan	25
5.2	Saran	25
	DAFTAR PUSTAKA	26
	RIWAYAT HIDUP	28

## DAFTAR TABEL

1	Penelitian terdahulu mengenai LDA	6
2	<i>Confusion Matrix</i>	10
3	Contoh tweet setelah pelabelan	13
4	Topik yang dihasilkan LDA	18
5	Tabel jumlah pembagian data	18
6	Parameter terbaik hasil hyperparameter tuning	18
7	<i>Classification Report</i> dari data asli	19
8	<i>Confusion Matrix</i> dari data asli	19
9	<i>Classification Report</i> dari data Random Under Sampling (RUS)	20
10	<i>Confusion Matrix</i> dari data Random Under Sampling (RUS)	20
11	<i>Classification Report</i> dari data Random Over Sampling (ROS)	20
12	<i>Confusion Matrix</i> dari data data Random Over Sampling (ROS)	20
13	<i>Classification Report</i> dari data Over Sampling dengan SMOTE	21
14	<i>Confusion Matrix</i> dari data Over Sampling dengan SMOTE	21
15	Perbandingan berdasarkan perlakuan data	21
16	<i>Classification Report</i> klasifikasi BERT dengan data asli	23
17	<i>Confusion Matrix</i> klasifikasi BERT dengan data asli	23
18	<i>Classification Report</i> klasifikasi BERT dengan data ROS	23
19	<i>Confusion Matrix</i> klasifikasi BERT dengan data ROS	23
20	Perhitungan waktu komputasi	24

## DAFTAR GAMBAR

1	<i>Plate Notation</i> LDA	5
2	<i>Representasi Random Forest</i>	9
3	Ilustrasi <i>pre-training</i> dan <i>fine-tuning</i> BERT	10
4	Ilustrasi SMOTE	10
5	Tahapan penelitian	12
6	Perbandingan jumlah data tweet tiap kelas	16
7	Perbandingan panjang kata tiap kelas	16
8	Grafik koherensi topik	17
9	Visualisasi pyLDAviz	17

## I PENDAHULUAN

### 1.1 Latar Belakang

Tipe baru *coronavirus* dengan tipe SARS-CoV-2 yang menyebabkan epidemi penyakit pernafasan akut muncul pada bulan Desember 2019 di kota Wuhan, China (Bavel *et al.* 2020). The World Health Organization (WHO) mendeklarasikan pandemi Covid-19 sejak 11 Maret 2020, dengan bukti *outbreak* penyakit Covid-19 yang terjadi di beberapa negara seperti Korea Utara, Iran, Amerika Serikat dan Eropa (Rockett *et al.* 2020). Kebutuhan informasi saat ini terkait Covid-19 perlu didukung dengan teknologi canggih yang mampu diakses secara luas. Salah satu platform sosial media, yaitu Twitter mengalami peningkatan aktivitas yang cukup signifikan saat pandemi Covid-19 (Rosenberg *et al.* 2020). Selama pandemi Covid-19, pemerintah di berbagai negara menggunakan Twitter sebagai kanal komunikasi untuk mengabarkan kebijakan terbaru dan berita terkait Covid-19 ke ranah publik (Xue *et al.* 2020b).

Media *microblogging* Twitter merupakan alat komunikasi yang dapat membagikan mengenai opini seseorang, dengan format bebas dan akses yang sangat mudah (Pak dan Paroubek, 2020). Jumlah pengguna Twitter saat ini sudah mencapai lebih dari 3,8 miliar orang (Cuello-Garcia *et al.* 2020). Penggunaan Twitter sebagai data analisis untuk melakukan prediksi politik sudah pernah dipakai sebelumnya dalam pemilu Jerman, Singapura, Irlandia, Amerika Serikat, dan Perancis (Qamar *et al.* 2014).

Berbagai bentuk analisis dengan data Twitter dilakukan oleh para peneliti di seluruh dunia untuk memahami perubahan aktivitas selama pandemi Covid-19. Analisis sentimen data Twitter terkait *lockdown* di India memperlihatkan sentimen positif masyarakat dan negara tersebut mulai memahami cara untuk menurunkan jumlah penderita Covid-19 (Barkur *et al.* 2020). Analisis pemodelan topik dan sentimen dari diskursus publik yang dibicarakan dilakukan dengan menggunakan data Twitter menggunakan LDA (Xue *et al.* 2020a).

Selain beberapa manfaat dari penggunaan analisis topik dan analisis sentimen menggunakan data Twitter yang telah dijelaskan diatas, diskusi mengenai vaksinasi Covid-19 di Twitter menjadi pembicaraan terhangat saat Rusia mengizinkan vaksin Covid-19 pertama di dunia (Lyu *et al.* 2021). Diskusi vaksinasi Covid-19 di Amerika Serikat dan Brazil terlihat didominasi oleh emosi yang negatif, hal ini terkait dengan penanganan kasus yang meningkat, laporan kasus, dan data statistik (Garcia dan Berton 2021). Vaksinasi pertama Covid-19 di Indonesia dimulai pada 13 Januari 2021 yang diawali oleh Presiden Indonesia Joko Widodo (Sastramidjaja dan Rosli 2021). Hal ini tentunya menimbulkan berbagai reaksi dari masyarakat karena perbedaan pandangan terhadap vaksinasi.

Beberapa analisis topik untuk mendapatkan tema diskusi mengenai vaksinasi Covid-19 sudah dilakukan di Indonesia. Analisis topik menggunakan LDA dengan 5583 *tweet* memperlihatkan diskusi pengguna Twitter mengandung topik mengenai kinerja pemerintah, harapan masyarakat, dan keamanan vaksin (Rachman dan Pramana 2020). Diskusi mengenai korupsi bantuan sosial dan ajakan memakai masker juga menjadi topik utama yang dihasilkan metode LDA pada rentang waktu November 2020 hingga Februari 2021 (Hakim *et al.* 2021).

Topik yang dihasilkan LDA sangat membantu untuk membuat kesimpulan dari banyaknya data *tweet* yang beredar.

Berbagai bentuk analisis sentimen berupa klasifikasi *tweet* sudah dilakukan, seperti analisis dengan keyword vaksin merah putih dan *sinovac* dengan 845 *tweet* dengan *classifier SVM* menghasilkan 84% dan *classifier Naïve Bayes* sebesar 85% (Laurensz *et al.* 2021). Analisis menggunakan data Twitter yang sama dengan data pada penelitian ini sudah pernah dilakukan dengan klasifikasi *Support Vector Machine (SVM)* menghasilkan nilai akurasi sebesar 90% (Chairunnisa *et al.* 2022). Meski demikian, klasifikasi sentimen menggunakan *Random Forest* mengenai vaksinasi Covid-19 di Indonesia masih sedikit ditemukan. Penelitian teks bahasa *malayalam* (bahasa lokal di India) menggunakan klasifikasi sentimen *Random Forest* dengan data Twitter yang tidak terkait dengan Covid-19 mendapatkan nilai akurasi sebesar 95% (S. dan K.V. 2020). Nilai akurasi besar menjadikan *Random Forest* menjadi pilihan model klasifikasi dengan data *tweet* vaksinasi pada penelitian ini. Hasil dari algoritme *Random Forest* akan dibandingkan algoritme *BERT* untuk mendapatkan pembanding dari hasil klasifikasi dan waktu komputasi. Oleh karena itu, penelitian ini bertujuan untuk melakukan analisis topik dan sentimen pengguna Twitter terhadap vaksinasi Covid-19 dengan pemodelan topik *Latent Dirichlet Allocation* dan *Random Forest*.

## 1.2 Rumusan Masalah

Perumusan masalah penelitian ini adalah sebagai berikut:

1. Belum banyak penelitian terkait analisis topik dan analisis sentimen yang menggunakan data vaksinasi Covid-19 di Indonesia
2. Apa saja topik yang dibicarakan yang terkait dengan vaksinasi Covid-19?
3. Bagaimana performa analisis sentimen pengguna Twitter menggunakan *Random Forest* dengan data vaksinasi Covid-19?

## 1.3 Tujuan

Tujuan penelitian ini adalah melakukan analisis topik pengguna Twitter terhadap vaksinasi Covid-19 dengan pemodelan topik *Latent Dirichlet Allocation* (LDA) dan analisis sentimen dengan model klasifikasi *Random Forest*.

## 1.4 Manfaat

Manfaat penelitian ini adalah untuk mendapatkan analisis topik terkait topik-topik yang diperbincangkan mengenai vaksinasi Covid-19 dan mengetahui performa dari model klasifikasi *Random Forest* pada data vaksinasi Covid-19. Hasil dari penelitian diharapkan dapat membantu banyak pihak untuk menjawab pertanyaan publik yang sedang berkembang.

## 1.5 Ruang Lingkup

Lingkup penelitian ini adalah :

1. Data *tweet* diambil pada rentang 8 September 2020 sampai 1 Juni 2021
2. Data *tweet* yang digunakan adalah tweet dengan kata kunci “Vaksin corona” dengan bahasa Indonesia yang hanya mengambil fitur *tweet*, *retweetCount*, *likeCount*, dan tanggal.

## II TINJAUAN PUSTAKA

### **2.1 Analisis Topik Covid-19**

Analisis topik bertujuan untuk mendapatkan topik-topik pembicaraan dari korpus yang besar. Topik yang dihasilkan LDA sangat membantu untuk membuat kesimpulan dari banyaknya data *tweet* yang beredar. Penelitian terkait untuk mendapatkan tema diskusi mengenai vaksinasi Covid-19 sudah dilakukan di Indonesia. Analisis topik menggunakan LDA dengan 5583 *tweet* memperlihatkan diskusi pengguna Twitter mengandung topik mengenai kinerja pemerintah, harapan masyarakat, dan keamanan vaksin (Rachman dan Pramana 2020). Diskusi mengenai korupsi bantuan sosial dan ajakan memakai masker juga menjadi topik utama yang dihasilkan metode LDA pada rentang waktu November 2020 hingga Februari 2021 (Hakim *et al.* 2021).

### **2.2 Analisis Sentimen Covid-19**

Analisis sentimen bertujuan mendapatkan ekstraksi informasi subjektif untuk membuat strukturisasi dan penerapan pengetahuan (Pozzi *et al.* 2017). Penelitian terkait analisis sentimen mengenai vaksin Covid sudah pernah dilakukan sebelumnya menggunakan data Twitter yang sama, dengan menggunakan *classifier* SVM dihasilkan nilai akurasi 90% (Chairunnisa *et al.* 2022). Penelitian lainnya yang terkait data Twitter dengan kata kunci "vaksin sinovac" dan "vaksin merah putih" menggunakan *classifier* Naïve Bayes dan SVM menghasilkan akurasi 84% dan 85% (Prakosa *et al.* 2021).

### **2.3 Latent Dirichlet Allocation**

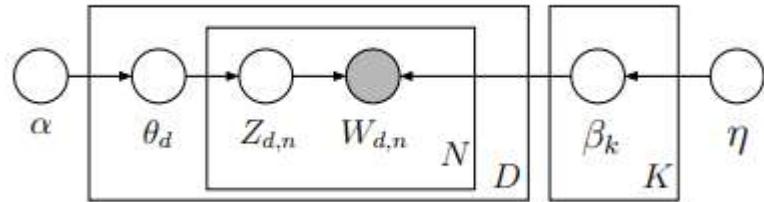
*Latent Dirichlet Allocation* (LDA) merupakan suatu cara untuk mendapatkan *insight* dari data yang masif untuk melihat struktur laten dalam suatu data. LDA dipakai untuk melakukan identifikasi pola, tema, struktur dan menjelaskan bagaimana tema tersebut terhubung (Xue *et al.* 2020a). Metode ini mengasumsikan setiap dokumen mengandung campuran berbagai topik, dan setiap topik dapat terlihat dari campuran kata-kata (Jamison *et al.* 2020). LDA sering digunakan diberbagai bidang. Pada bidang genetika, LDA digunakan untuk melakukan karakterisasi pola genetika dari suatu populasi dan identifikasi ekspresi dari setiap individual. LDA juga digunakan pada bidang *computer vision*, untuk klasifikasi gambar, menghubungkan gambar dengan *caption*, dan membuat hierarki dari gambar (Blei 2012).

LDA juga merupakan bagian dari bidang *probabilistic modeling*. Pada pemodelan probabilistik generatif, data diperlakukan dengan proses generatif yang melibatkan variabel laten. Proses generatif menggunakan distribusi *joint probability* dengan variabel tampak dan laten. Analisis data dari distribusi *joint probability* untuk menghitung *conditional distribution* dari variable tersembunyi dengan menggunakan variabel yang terlihat. *Conditional distribution* juga disebut dengan *posterior distribution*. Proses generatif untuk LDA akan dijalankan mengikuti *joint distribution* dari variabel laten dan tampak. Persamaan untuk menghitung LDA dirumuskan sebagai berikut: (Blei 2012)

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^K p(\beta_i) \prod_{d=1}^D p(\theta_d) \left( \prod_{n=1}^N p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:K}, z_{d,n}) \right), \quad (1)$$

dengan  $\beta_K$  merupakan distribusi dari pembendaharaan ke- $k$ ,  $\theta_d$  proporsi topik untuk dokumen ke- $d$ ,  $z_d$  merupakan pilihan topik untuk dokumen ke- $d$ , dan  $w_d$  merupakan kata yang terlihat pada dokumen ke- $d$ .

LDA direpresentasikan pada Gambar 1 untuk menjelaskan asumsi probabilistik dan kelompok distribusi probabilistik. Parameter *alpha* ( $\alpha$ ) merupakan parameter untuk mengontrol *prior distribution* kandungan topik tiap dokumen, dan parameter *eta* ( $\eta$ ) merupakan parameter untuk mengontrol *prior distribution* kandungan kata pada tiap topik.



Gambar 1 *Plate Notation LDA*

Permasalahan komputasi yang dihadapi untuk mendapatkan struktur topik adalah menghitung *posterior distribution*, yaitu *conditional distribution* dari variabel laten yang berada pada dokumen. Perhitungan *posterior* dijabarkan sebagai berikut (Blei 2012):

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D} | w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})} \quad (2)$$

dengan  $\beta_{1:K}$  merupakan distribusi dari pembendaharaan untuk semua  $K$ ,  $\theta_{1:D}$  proporsi topik untuk semua dokumen  $D$ ,  $z_d$  merupakan pilihan topik untuk semua dokumen  $D$ , dan  $w_d$  merupakan kata yang terlihat untuk semua dokumen  $D$ .

Pembilang merupakan *joint distribution* dari semua variabel random yang masih bisa dihitung, namun pembagi merupakan *marginal probability* dari observasi, yang mana merupakan probabilitas dari korpus yang terlihat dari topik. Secara teori, pembagi dapat dihitung dengan menjumlahkan semua *joint distribution* dari setiap struktur topik laten. Namun, jumlah dari semua struktur topik sangat besar, sehingga sangat sulit untuk dihitung. Algoritme pemodelan topik bisa melakukan pendekatan dengan membuat distribusi alternatif terhadap struktur topik yang laten yang dapat mendekati *true posterior* (Blei 2012).

LDA digunakan pada berbagai penelitian analisis sentimen, seperti interaksi saat kejadian bencana alam dan kesehatan saat epidemi berlangsung (Shurrah *et al.* 2021). Tabel 1 merupakan beberapa penerapan LDA dalam melakukan analisis topik.

Tabel 1 Penelitian terdahulu mengenai LDA untuk menganalisis topik

Penulis	Judul Paper	Dataset	Hasil
IR Putri dan R Kusumaningrum (2017)	Latent Dirichlet Allocation (LDA) for Sentiment Analysis Toward Tourism Review in Indonesia	Trip Advisor	<i>shows the best accuracy is about 60% as an average accuracy of all folds and the best accuracy is about 80%</i>
Xue J, Chen J, Chen C, Zheng C, Li S, dan Zhu T (2020)	Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter	Tweet from Twitter API	-Outbreak potensial terjadi di New York -Ketakutan menjadi emosi yang dominan.
Jamison, A., Broniatowski, D.A., Smith, M.C., Parikh, K.S., Malik, A., Dredze, M. dan Quinn (2020)	Adapting and extending a typology to identify vaccine misinformation on Twitter	Tweet from Twitter API	-Melakukan pemetaan tipologi dengan diskusi yang berkembang -Mendapatkan fakta dari narasi yang digunakan oleh pendukung dan penolak vaksin

## 2.4 Pendekatan Inferensi

Pendekatan inferensi untuk mendapatkan nilai probabilitas biasanya digunakan dengan dua pendekatan, yaitu algoritme berbasis *sampling* dan *variational*. Algoritme berbasis *sampling* dapat mengoleksi contoh dari *posterior* untuk mendekati distribusi empiris. Algoritme sampling yang biasa digunakan yaitu *Gibbs Sampling*, dimana *Markov Chain* membuat rangkaian dari variabel random yang saling bergantung yang akan membuat batas distribusi dari *posterior*. *Markov Chain* dibentuk dari variabel topik laten dari korpus yang ditentukan, dan algoritme menjalankan *chain* dalam waktu yang lama, mengambil sampel dari batas distribusi, dan menyimpulkan nilai distribusi dari sampel yang sudah didapatkan (Blei 2012).

Metode *Variational Inference* (VI) merupakan alternatif dari algoritme berbasis sampling. Berbeda dengan mengambil sample untuk menyimpulkan *posterior*, metode *variational* menentukan kelompok parameter distribusi dari struktur laten dan mencari kelompok yang dekat dengan *posterior*. Permasalahan inferensi berubah menjadi permasalahan optimisasi. Metode *Variational Bayes* (VB) memberikan konvergensi lebih cepat dibandingkan *Gibbs Sampling* (Asuncion *et al.* 2009). *Online Learning for Latent Dirichlet Allocation* merupakan salah satu metode yang dikembangkan dari *Batch Varional Bayes* yang dapat memproses data masif (Hoffman *et al.* 2010).

## 2.5 Online Variational Inference for Latent Dirichlet Allocation

Metode inferensi dengan menggunakan *Online Variational Inference for Latent Dirichlet Allocation* merupakan ekstensi dari algoritme *Variational Bayes* (VB) yang berbasis *online stochastic optimization* (Hoffman *et al.* 2010). Metode ini menunjukkan dapat menghasilkan parameter yang bagus dan lebih cepat dari VB sebelumnya untuk memproses dataset yang besar. Pada inferensi VB, *true posterior* diperkirakan dengan distribusi  $q(z, \theta, \beta)$  yang terindex oleh sekumpulan parameter bebas. VB akan mencari parameter yang membuat aproksimasi terdekat dengan *posterior distribution*. Parameter ini dioptimisasi untuk memaksimalisasi *Evidence Lower Bound (ELBO)*. Memaksimalisasi nilai ELBO sama saja dengan meminimalkan divergensi Kullback–Leibler antara  $q(z, \theta, \beta)$  dan posterior  $p(z, \theta, \beta|w, \alpha, \eta)$ . Berikut rumus untuk menghitung nilai ELBO (Hoffman *et al.* 2010) :

$$L(n, \lambda) \triangleq \sum d l(n_d, \gamma_d, \varphi_d, \lambda) \quad (3)$$

dimana  $n_d$  merupakan nomor dari dokumen,  $\gamma_d$   $\varphi_d$  merupakan parameter variational dan  $\lambda$  merupakan topik.

## 2.6 Ukuran Koherensi Topik

Terdapat empat dimensi pengukuran tingkat koherensi yang bersifat saling bebas sehingga dapat digabungkan. Koherensi adalah sekumpulan kata yang mengukur keterikatan dan kesesuaian kata-kata tunggal atau subset dari kumpulan kata.

Dimensi pertama yaitu jenis segmentasi yang digunakan untuk membagi kumpulan data menjadi bagian-bagian yang lebih kecil. Potongan ini dibandingkan satu dengan yang lain, sehingga segmentasi tersebut bisa menjadi pasangan kata. Dimensi kedua adalah estimasi peluang, dimana nilai probabilitas dihitung dari sumber data. Penghitungan peluang menggunakan *boolean document* yang menghitung probabilitas kata yang muncul dibagi dengan jumlah dokumen, *boolean paragraph* dimana penghitungan memperhatikan paragraf atau kalimat, dan *boolean sliding window* dimana kata dihitung menggunakan metode *sliding window*. Dimensi ketiga adalah ukuran konfirmasi yang menilai kesepakatan dari pasangan tertentu, misalnya NPMI dari dua kata. Himpunan ukuran konfirmasi adalah M. Himpunan metode untuk memperkirakan peluang adalah P, yang membentuk dimensi ketiga dari ruang konfigurasi. Terakhir, menggabungkan nilai skalar yang dihitung oleh ukuran konfirmasi membentuk dimensi keempat. Himpunan fungsi agregasi adalah  $\Sigma$ . Secara visual, metode agregasi ukuran koherensi dapat dilihat pada gambar dibawah ini. Ukuran koherensi merupakan produk silang dari  $C = S \times P \times M \times \Sigma$  (Röder *et al.* 2015).

## 2.7 Visualisasi Pemodelan Topik

Hasil dari pemodelan topik dapat disajikan dengan berbagai visualisasi, mulai dari menampilkan sebaran probabilitas dari tiap kata pada suatu topik, membuat *wordcloud* kata sesuai probabilitas tiap kata pada topik, dan membuat visualisasi dengan menampilkan relevansi dan jarak antar topik.

LDAvis merupakan alat yang berupa *web-based* untuk membantu menginterpretasikan topik yang sudah dihasilkan oleh metode LDA. LDAvis membantu user untuk menjawab apa makna dari setiap topik, bagaimana kelaziman dari topik yang muncul, dan apakah tiap topik mempunyai hubungan. Untuk menjawab pertanyaan tersebut, diperlukan beberapa perhitungan yaitu relevansi dan *saliency*. Relevansi dapat dihitung sebagai berikut (Sievert dan Shirley 2014) :

$$r(w, k | \lambda) = \lambda \log(\varphi_{kw}) + (1 - \lambda) \log \left( \frac{\varphi_{kw}}{p_w} \right) \quad (7)$$

dengan  $w$  merupakan kata yang muncul,  $k$  merupakan nomor topik,  $\lambda$  menentukan kandungan yang diberikan antara probabilitas kata dan topik,  $\varphi_{kw}$  merupakan probabilitas kata tiap topik, dan  $p_w$  merupakan distribusi empiris dari korpus. *Saliency* dapat memberikan urutan kata yang koheren membangun sebuah topik. *Saliency* dan *distinctiveness* dapat dihitung sebagai berikut (Chuang 2012) :

$$\begin{aligned} \text{distinctiveness}(w) &= \sum_T P(T|w) \log \frac{P(T|w)}{P(T)} \\ \text{saliency}(w) &= P(w) \times \text{distinctiveness}(w) \end{aligned} \quad (8)$$

dengan  $w$  sebagai kata,  $P(T|w)$  sebagai *conditional probability* yang menjadi *likelihood* kata  $w$  yang dihasilkan dari topik laten  $T$ ,  $P(T)$  yang merupakan *likelihood* dari kata acar yang dihasilkan topik  $T$ . Selanjutnya akan dihitung *distinctiveness* kata  $w$  menggunakan divergensi Kullback-Leibler antara  $P(T|w)$  dan  $P(T)$ , dan *saliency* yang merupakan perkalian antara *distinctiveness* dan  $P(w)$ .

## 2.8 Term Frequency – Inverse Document Frequency (TF-IDF)

*Term Frequency – Inverse Document Frequency* (TF-IDF) pertama diperkenalkan oleh Salton dan Buckley pada 1988 (Salton dan Buckley 1988). Metode ini memberikan pembobotan otomatis terhadap suatu kata. TF (*Term Frequency*) menghitung seberapa sering istilah muncul dalam suatu dokumen atau korpus. Perhitungan nilai TF dihitung seperti berikut (Rajaraman dan Ullman 2011) :

$$TF_{ij} = \frac{f_{ij}}{\max_k f_{kj}} \quad (9)$$

dengan  $f_{ij}$  sebagai frekuensi dari kemunculan kata  $i$  pada dokumen  $j$  dan notasi  $\max_k f_{kj}$  menunjukkan total term yang muncul pada dokumen. IDF (*Inverse Document Frequency*) akan mengukur seberapa unik kata tersebut pada korpus. IDF akan dihitung sebagai berikut (Rajaraman dan Ullman 2011) :

$$IDF_i = \log_2 \left( \frac{N}{n_i} \right) \quad (10)$$

dengan  $n_i$  merupakan nomor dokumen dimana kata  $I$  muncul, dan  $N$  merupakan jumlah dokumen pada koleksi. TF-IDF merupakan hasil dari perkalian nilai  $TF_{ij}$

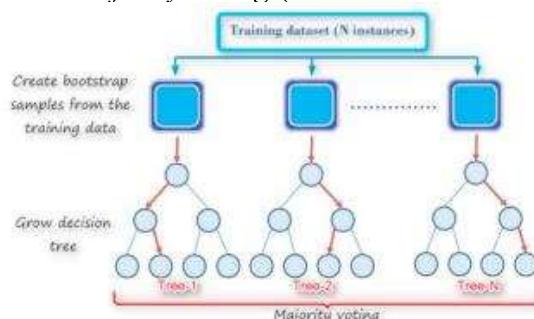
dan  $IDF_i$ . Kata dengan nilai TF-IDF tertinggi biasanya merupakan kata terbaik untuk merepresentasikan topik pada dokumen tersebut.

## **2.9 Random Forest**

*Random forest* merupakan kombinasi dari pohon prediksi dimana setiap pohon bergantung pada nilai vektor acak (Breiman 2001). Algoritme ini menggabungkan konsep dari *random subspaces* dan *bootstrap aggregating (bagging)*. Algoritme *Random Forest* dapat dijabarkan sebagai berikut. Pertama, melakukan *bootstrap sampling* dengan penggantian dari data *train*. Kedua, untuk setiap *bootstrap* pada tahap pertama, pohon akan berkembang dengan *sampling* acak dari variabel input dan pilih pembagian terbaik dari variabel tersebut. Kemudian tahap pertama dan kedua diulang hingga iterasi ke- $k$  hingga pohon terbentuk. Setelah dilakukan training, dalam melakukan prediksi dapat dilakukan dengan mengambil nilai rata-rata dari semua pohon regresi. Persamaan regresi dijabarkan sebagai berikut (al Amrani *et al.* 2018) :

$$F_{rf}^B = \frac{1}{B} \sum_{b=1}^B T_b(x) \quad (3)$$

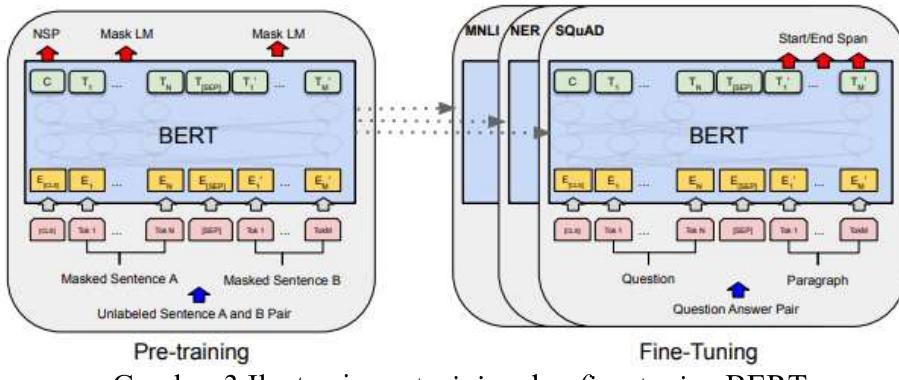
dengan  $B$  merupakan jumlah ulangan *bagging*, dan  $Tb$  merupakan nilai dari pohon regresi. Untuk klasifikasi, prediksi diambil dari *majority voting* dari setiap pohon klasifikasi. Gambar 2 menjelaskan bagaimana metode Random Forest mengambil kesimpulan menggunakan *majority voting* (al Amrani *et al.* 2018).



Gambar 2 Representasi *Random Forest*

## 2.10 Bidirectional Encoder Representation from Transformer (BERT)

BERT merupakan singkatan dari *Bidirectional Encoder Representation from Transformer* (Devlin *et al.* 2018). BERT didesain untuk melakukan pra-training dengan representasi *deep bidirectional* dari teks tidak berlabel (Devlin *et al.* 2018). Encoder pada BERT membaca semua sekuen dari kata yang masuk secara menyeluruh. Model dibuat untuk dapat mengerti makna yang terkandung sebelum dan sesudah kata. Karena kemampuan BERT untuk memahami makna, BERT dapat meningkatkan performa pada berbagai kegiatan pada bidang NLP seperti *Natural Language Inference* (NLI) dan *Question Answering* (QA) (Azhar dan Khodra 2020).

Gambar 3 Ilustrasi *pre-training* dan *fine-tuning* BERT

Gambar 3 menunjukkan prosedur *pre-training* dan *fine-tuning* yang dilakukan oleh BERT. Arsitektur yang sama digunakan pada *pre-training* dan *fine-tuning*, perbedaan terletak pada *output layer*. *Pre-trained* model digunakan untuk menginisiasi model untuk pekerjaan *downstream* yang berbeda. Saat *fine-tuning*, semua parameter sudah disetting dengan baik (Devlin *et al.* 2018).

## 2.11 K-Fold Cross Validation

*K-fold cross validation* merupakan salah satu cara untuk membagi gugus data menjadi data latih dan data uji. Metode ini digunakan untuk mengurangi bias dalam pengambilan contoh. K merupakan nilai untuk menentukan banyaknya pembagian gugus data, dimana Sebagian dari gugus data akan dijadikan data uji pada pemodelan. Penggunaan cross validation berulang direkomendasikan, selama masih bisa dikomputasikan. Penggunaan  $k=5$  dan  $k=10$  lebih baik, karena bias akan berkurang jika dibandingkan dengan  $k=2$  (Rodríguez *et al.* 2010).

## 2.12 Confusion Matrix

Klasifikasi menentukan prediksi akan memberikan empat hasil, diantaranya *confusion matrix*, *precision*, *accuracy*, *recall*, dan *F1-score*. Tabel Confusion Matrix untuk data tiga kelas dijelaskan sebagai berikut (Sholehah 2018):

Tabel 2 *Confusion Matrix*

Fakta	Prediksi		
	Positif	Negatif	Netral
Positif	TP	FN <sub>1</sub>	FN <sub>t1</sub>
Negatif	FP <sub>1</sub>	T <sub>Ng</sub>	FN <sub>t2</sub>
Netral	FP <sub>2</sub>	FN <sub>Ng2</sub>	T <sub>Ng</sub>

dengan TP adalah jumlah prediksi yang benar untuk data aktual positif, FP<sub>1</sub> dan FP<sub>2</sub> adalah jumlah prediksi yang salah untuk data aktual positif, TN<sub>t</sub> adalah jumlah prediksi yang benar untuk data aktual netral, FN<sub>t1</sub> dan FN<sub>t2</sub> adalah jumlah prediksi yang salah untuk data netral, T<sub>Ng</sub> adalah jumlah prediksi yang benar untuk data aktual negatif, FN<sub>Ng1</sub> dan FN<sub>Ng2</sub> adalah jumlah prediksi yang salah untuk data aktual negatif. Nilai dari *precision*, *accuracy*, *recall*, dan *F1-score* akan ditampilkan dibawah ini (Sholehah 2018):

$$Akurasi = \frac{TP + TN}{P + N} \quad (11)$$

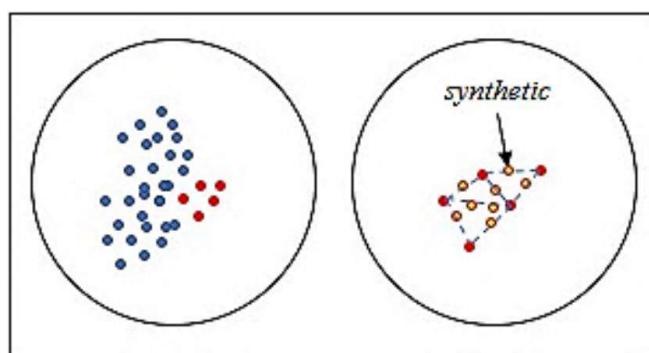
$$Presisi = \frac{TP}{TP + FP1 + FP2} \quad (12)$$

$$Recall = \frac{TP}{TP + FN_{g1} + FN_{t1}} \quad (13)$$

$$F - measure = \frac{2}{1/presisi + 1/recall} \quad (14)$$

### 2.13 Algoritme Synthetic Minority Oversampling Technique)

SMOTE adalah penanganan dataset imbalance dengan melakukan sintesis data minoritas berdasarkan *k-nearest neighbor* antar kelas minoritas (Bunkhumpornpat *et al.* 2009). Proses SMOTE melewati beberapa proses yaitu identifikasi fitur vektor dan tentangga terdekat, lalu ambil perbedaan jarak antar kedua titik, penggandaan perbedaan dengan jumlah acak antara 0 dan 1, kemudian identifikasi titik baru pada garis segmen dengan menambahkan nomor acak ke vektor (Bisri dan Rachmatika 2019). Ilustrasi SMOTE dapat dilihat pada Gambar 4 (Bisri dan Rachmatika 2019).

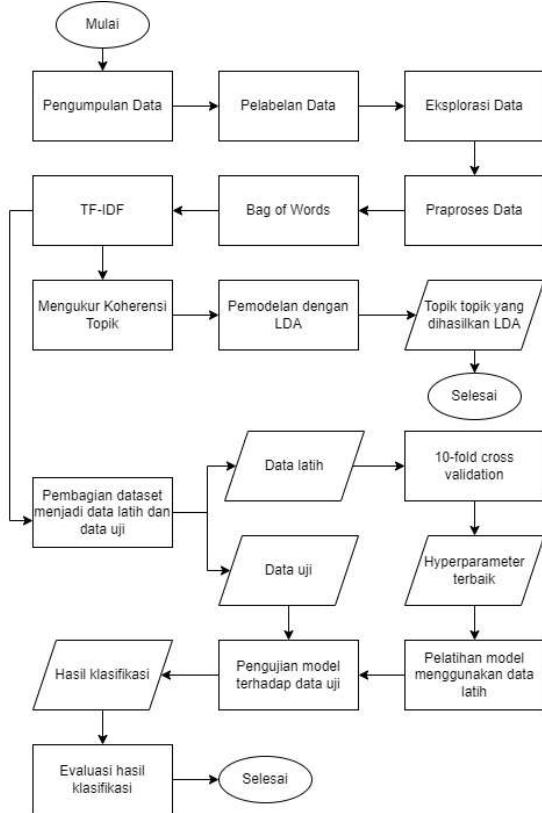


Gambar 4 Ilustrasi SMOTE

## III METODE

### 3.1 Tahapan Penelitian

Tahapan penelitian dapat dilihat pada gambar dibawah ini. Tahapan ini terdiri dari beberapa tahapan. yaitu pengumpulan data, pelabelan data, praproses data, mengukur koherensi topik, *bag of words*, pemodelan topik LDA, pemetaan probabilitas pada *tweet*, pemodelan klasifikasi, *hyperparameter tuning* dan evaluasi model klasifikasi.



Gambar 5 Tahapan penelitian

### 3.2 Pengumpulan Data

Data *tweet* diambil menggunakan *package snscreape*. Data yang diambil sebanyak 47.140 *tweet* yang terdiri dari 4 atribut. Rentang tanggal tweet yang diambil mulai dari 1 September 2020 hingga 30 Juni 2021. Data *tweet* duplikat dibersihkan sehingga tersisa 18.805 *tweet*.

### 3.3 Pelabelan Data

Pelabelan dibagi menjadi tiga kelas, yaitu positif, negatif, dan netral. Pemilihan sentimen diatur dengan konsensus. Sentimen positif menunjukkan setuju atau dukungan terhadap vaksin dan kebijakan pemerintah. Sentimen negatif berisi pernyataan kontra terhadap vaksin dan kebijakan pemerintah. Sentimen netral berisi berita atau informasi yang tidak mengarah kepada dukungan maupun kontra. Pelabelan manual dilakukan oleh 13 orang mahasiswa IPB University dari berbagai program studi. Hasil pelabelan manual menunjukkan jumlah sentimen positif sebesar 5841 tweet, sentimen negatif sebesar 3256 tweet, dan sentimen netral sebesar 9708 tweet. Berikut contoh *tweet* yang sudah dilabeli dengan sentimen dapat dilihat pada Tabel 3.

Tabel 3 Contoh tweet setelah pelabelan

Sentimen	Tweet
Positif	Saya pikir ini konsep yg sangat bagus kebijakan pemerintah terkait akan memberikan vaksin Corona Virus Disease atau Covid-19 secara gratis diutamakan bagi masyarakat tidak mampu menjadi bukti kehadiran negara. #VaksinGratis
Netral	Tim riset uji klinis vaksin Covid-19 buatan Sinovac membenarkan adanya relawan yang positif terpapar virus corona. <a href="https://t.co/4nhnxET37H">https://t.co/4nhnxET37H</a> #iNewsid #Finance
Negatif	Nah ini, berarti sebenarnya dilakuin vaksinasi juga percuma karena dia ga nyembuhin. Dan pemerintah mengembanggemborkan seolah vaksin ini obat dan corona bakal selesai dengan ini vaksin. Agak misleading ga sih?

### 3.4 Eksplorasi Data

Data yang sudah dilabeli akan dilihat gambaran besarnya melalui eksplorasi data yang akan melakukan visualisasi panjang data tiap kelas dan jumlah data yang terlabeli dalam grafik.

### 3.5 Praproses Data

Data awal perlu dibersihkan agar data yang diproses menjadi ringan. Tujuan dari proses ini untuk standarisasi bentuk kata, memperbaiki kata yang cacat, menghilangkan URL twitter yang masih terkandung dalam *cell*, dan objek lainnya (Angiani *et al.* 2016). Berikut pra-proses yang dilakukan dalam penelitian ini :

1. *Case Folding*  
Merupakan proses untuk menyamakan seluruh bentuk kata. Huruf kapital diproses menjadi huruf kecil. Proses ini bertujuan untuk mengurangi indeks berlebih pada kata yang sama.
2. *Data Cleaning*  
Proses pembersihan bertujuan untuk mengurangi *noise* pada data. Proses ini menghapus simbol, tanda baca, angka, link URL, emotikon, *mention*, *hashtag*.
3. Normalisasi Bahasa  
Mengembalikan bentuk penulisan yang tidak baku menjadi penulisan yang sesuai KBBI seperti gk menjadi tidak, dan kata-kata singkatan lainnya.
4. Tokenisasi  
Proses pemecahan kalimat menjadi potongan atau *token* untuk diproses lebih lanjut dengan program. Berikut contoh dari bentuk tokenisasi :

[‘ndak’, ‘usah’, ‘panikan’, ‘pak’, ‘tetap’, ‘patuhi’, ‘prokes’]

### 5. Penghapusan *Stopword*

Kata-kata yang sering muncul namun tidak mempunyai pemaknaan dalam proses ini. Contoh *stopword* yang dimaksud yaitu : “yang”, “dan”, “di”.

### 6. Stemming

Proses ini mengubah kata menjadi bentuk dasarnya, sehingga tidak ada pemaknaan lebih dari bentuk kata yang sama. Contohnya “memahami” akan diganti dengan kata “paham”. Package *sastrawi* digunakan dalam proses ini untuk membantu proses *stemming* dengan Bahasa Indonesia.

Beberapa data *tweet* yang sudah diproses akan terlihat seperti :

[‘saya’, ‘pikir’, ‘ini’, ‘konsep’, ‘sangat’, ‘bagus’, ‘kebijakan’, ‘pemerintah’, ‘terkait’, ‘akan’, ‘memberikan’, ‘vaksincorona’, ‘virus’, ‘desease’, ‘covid’, ‘secara’, ‘gratis’, ‘diutamakan’, ‘bagi’, ‘masyarakat’, ‘tidak’, ‘mampu’, ‘menjadi’, ‘bukti’, ‘kehadiran’, ‘negara’]

[‘tim’, ‘riset’, ‘uji’, ‘klinis’, ‘vaksin’, ‘covid’, ‘buatan’, ‘sinovac’, ‘membenarkan’, ‘adanya’, ‘relawan’, ‘positif’, ‘terpapar’, ‘virus’, ‘corona’]

### 3.6 *Bag of Words (BOW)*

Kata dibobotkan dengan beberapa metode yaitu *Bag of Words*. Pada *Bag of Words* kata dimasukkan pada vektor dan nilainya bertambah sesuai kata-kata yang di proses. *Bag of Words* merupakan representasi teks dalam vektor yang mudah diimplementasikan, yang langsung korespondensi dengan kata dalam setiap pengaturannya. Kata atau *n-gram* digunakan untuk representasi disebut *terms*. Semua kemunculan term yang sama akan diperlakukan dengan cara yang sama, tidak tergantung posisi atau kata yang berkaitan. (Cichosz 2018). Nilai *Bag of Words* akan terlihat seperti berikut :

$\{[(31, 1), (68, 1), (98, 1), (151, 1), (282, 1), (399, 1), (405, 1), (576, 1)]\}$

### 3.7 Mengubah Nilai BOW Menjadi TF-IDF

Model representasi data *bag of words* diubah menjadi model TF-IDF. Nilai *bag of words* yang sebelumnya hanya menampilkan jumlah kemunculan data pada suatu dokumen berubah menjadi nilai TF-IDF. Nilai BOW yang sudah diubah menjadi TF-IDF akan terlihat sebagai berikut :

$\{[(31, 0.19045678217087234), (68, 0.2298831617658562)]\}$

### 3.8 Mengukur Koherensi Topik

Nilai ukuran koherensi akan dicari dari data yang sudah dilakukan praproses. Mekanisme pengukuran nilai koherensi terdiri dari empat tahap. Pertama, proses segmentasi menggunakan *s-one-set* membandingkan kata terhadap *word set W* menggunakan *words context vector*. Kedua, penghitungan probabilitas menggunakan metode  $Psw(110)$ . Ketiga, ukuran konfirmasi menggunakan pengukuran secara *indirect*. Terakhir, proses agregasi untuk memberikan nilai tengah dari ukuran konfirmasi.

### 3.9 Analisis Topik dengan *Latent Dirichlet Allocation*

Proses ini bertujuan untuk mendapatkan nilai dari distribusi kata yang membentuk suatu topik menggunakan *Latent Dirichlet Allocation* (LDA). Nilai Bag-of-word (BOW) dengan TF-IDF dimodelkan menjadi *document topic matrix* yang berisi keterkaitan dokumen dengan topik yang dipilih secara acak dan *topic word matrix* yang berisi keterkaitan topik dengan kata yang masuk. Untuk mendapatkan nilai yang akurat, dilakukan optimisasi dengan proses iterasi untuk semua dokumen dan semua kata. Tiap dokumen dimodelkan mempunyai beberapa kata sebagai berikut

$$\begin{aligned} D1 &= w1, w2, w3, w4, w5, w6, w7 \\ D2 &= w1, w2, w3, w4, w5, w6, w7, w8, w9 \\ D3 &= w1, w2, w3, w4, w5, w6, w7, w8, w9, w10 \\ D4 &= w1, w2, w3, w4, w5, w6, w7, w8, w9, w10, w11 \end{aligned}$$

dimana D merupakan dokumen, dan w merupakan kata yang terkandung pada dokumen. Setiap kata pada dokumen akan dibobotkan dengan topik yang dipilih secara acak, sehingga model dari dokumen akan terlihat seperti berikut

$$\begin{aligned} D1 &= w1(k5), w2(k3), w3(k1), w4(k2) \\ D2 &= w1(k5), w2(k1), w3(k5), w4(k3), w5(k1) \\ D3 &= w1(k3), w2(k5), w3(k2), w4(k2), w5(k3), w6(k2) \\ D4 &= w1(k2), w2(k3), w3(k5), w4(k2), w5(k3), w6(k5), w7(k2) \end{aligned}$$

dimana nilai k merupakan topik yang diinisiasi diawal. *Mixture* dari topik untuk tiap dokumen akan terlihat sebagai berikut

$$\begin{aligned} D1 &= k5 + k3 + k1 + k2 \\ D2 &= k5 + k1 + k5 + k3 + k1 \\ D3 &= k3 + k5 + k2 + k2 + k3 + k2 \\ D4 &= k2 + k3 + k5 + k2 + k3 + k5 + k2 \end{aligned}$$

dimana tiap dokumen sudah mengandung nilai topik yang dikandung secara acak. *Mixture* dari setiap kata pada topik juga dimodelkan sebagai berikut

$$\begin{aligned} k1 &= w3 + w5 \\ k2 &= w4 + w3 + w6 + w1 + w7 \end{aligned}$$

$$\begin{aligned} k3 &= w2 + w4 + w1 + w5 + w2 \\ k4 &= w4 \end{aligned}$$

dimana setiap topik merupakan kumpulan dari campuran tiap kata. LDA akan melakukan penyesuaian peletakan topik pada kata yang sesuai dengan menghitung dua probabilitas, yaitu proporsi topik pada dokumen dokumen  $D$  dan proporsi kata pada topik  $k$ . Nilai proporsi tersebut dikalikan dan akan menjadi produk untuk mengidentifikasi topik baru yang lebih relevan dengan kata sebelumnya.

Package python yang dipakai yaitu *ldamodel* dari *Gensim*, dan *pyLDAvis*. Pemodelan ini memberikan nilai dari kata-kata yang menjadi bagian dari topik. Package *pyLDAvis* menghasilkan infografis yang memudahkan pemahaman pengguna terhadap hasil yang terbentuk.

### 3.10 Analisis Sentimen dengan *Random Forest*

Analisis sentimen menggunakan algoritme *Random Forest* dilakukan untuk mendapatkan performa klasifikasi model dalam memilih kategori label dari data yang sudah di latih. Penelitian terkait perbandingan data *train* dan *test* menyarankan rasio 80:20 sebagai pilihan untuk menjadi rasio pemisahan data, hal ini untuk bertujuan memberikan data training yang cukup untuk klasifikasi multi kelas (Rácz *et al.* 2021). *Random Forest* akan melakukan proses *bootstrap* untuk menarik contoh acak untuk dijadikan data latih. Pohon klasifikasi akan dibuat dengan menggunakan *random feature selection*. Peubah penjelas dipilih yang terbaik untuk dijadikan sebagai penyekat dan dilanjutkan dengan pemisahan menjadi dua simpul baru. Proses ini terus berlanjut hingga ukuran minimum amatan dalam simpul tercapai. Pembuatan pohon terus dilakukan hingga diperoleh pohon klasifikasi. Tiap pohon klasifikasi menghasilkan satu suara, dan penentuan klasifikasi didasarkan pada suara terbanyak (*majority vote*).

Package python yang digunakan adalah *RandomForestClassifier* dari *Sklearn*. Algoritme *Random Forest* dilatih dengan rasio data latih sebesar 80% dan data uji sebesar 20%.

### 3.11 Analisis Sentimen dengan BERT

Analisis sentimen menggunakan algoritme *BERT* dipilih untuk menjadi pembanding klasifikasi dan waktu komputasi dari algoritme *Random Forest*. Algoritme *Random Forest* dilatih dengan rasio data latih sebesar 80% dan data uji sebesar 20%.

### 3.12 Hyperparameter Tuning

Hyperparameter tuning pada *Random Forest* digunakan dengan cara grid search. Hyperparameter Tuning dilakukan dengan menggunakan metode *grid search cross validation* dengan *10-fold cross validation*.

### 3.13 Evaluasi Model Klasifikasi

Evaluasi akan menjabarkan hasil dari klasifikasi. Nilai ini akan digunakan juga untuk menghitung akurasi, presisi, *recall*, dan *f-measure*. Metode

penanganan data *imbalance* antar kelas diterapkan untuk mengevaluasi model. Metode yang digunakan yaitu *Random Under Sampling* (RUS) dan *Over Sampling* dengan menggunakan SMOTE. Metode *Random Over Sampling* (ROS) juga digunakan untuk membandingkan performa antara kedua model klasifikasi, yaitu algoritme *Random Forest* dan *BERT*.

### 3.14 Lingkungan Pengembangan

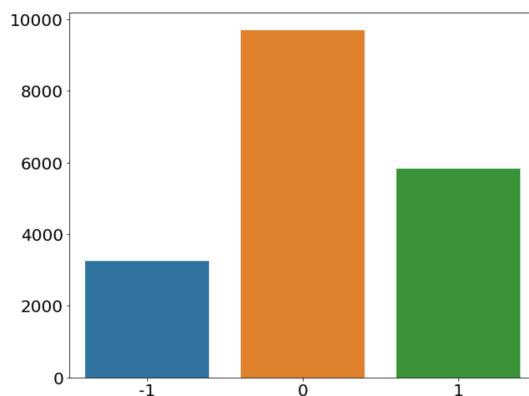
Spesifikasi perangkat keras dan perangkat lunak yang digunakan dalam penelitian ini adalah sebagai berikut :

- Perangkat Keras
  - Prosesor AMD Ryzen 3
  - RAM 8GB
- Perangkat Lunak
  - Windows 10 Home
  - Bahasa pemrograman Python 3.7
  - Anaconda Interpreter
  - Package Python terkait

## IV HASIL DAN PEMBAHASAN

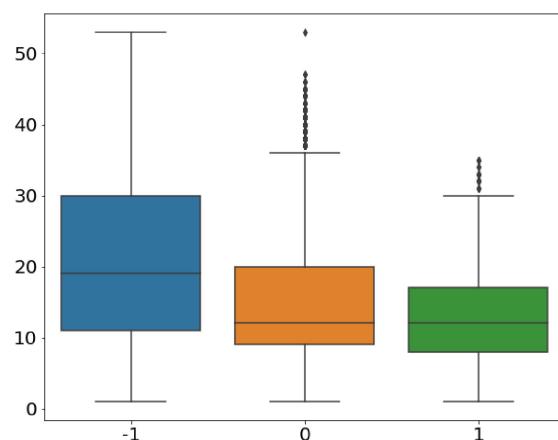
### 4.1 Eksplorasi Data

Data yang dihimpun berjumlah 18.805 *tweet* dengan rincian sentimen positif sebesar 5841 *tweet*, sentimen negatif sebesar 3256 *tweet*, dan sentimen netral sebesar 9708 *tweet*.



Gambar 6 Perbandingan jumlah data tweet tiap kelas

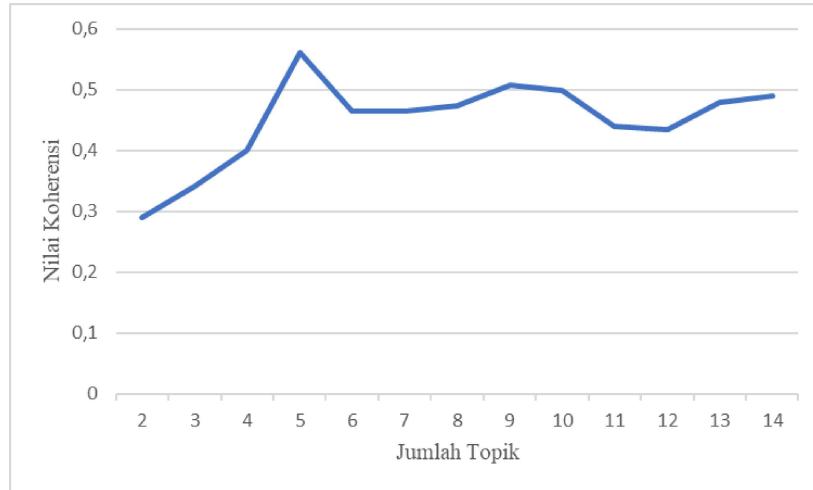
Jika dilihat dari panjangnya teks, sentimen negatif terlihat mempunyai panjang teks yang paling panjang diantara dua sentimen lainnya. Sentimen positif mempunyai rata-rata 19 kata, sentimen netral dengan rata-rata 12 kata, dan sentimen positif dengan rata-rata 12 kata.



Gambar 7 Perbandingan panjang kata tiap kelas

### 4.2 Koherensi Topik

Data berikut menjelaskan nilai koherensi pada setiap topik. Jumlah topik yang dipilih menjadi kandidat yaitu 2 topik hingga 14 topik. Nilai koherensi tertinggi terdapat pada jumlah topik 5 dengan nilai koherensi sebesar 0.56.

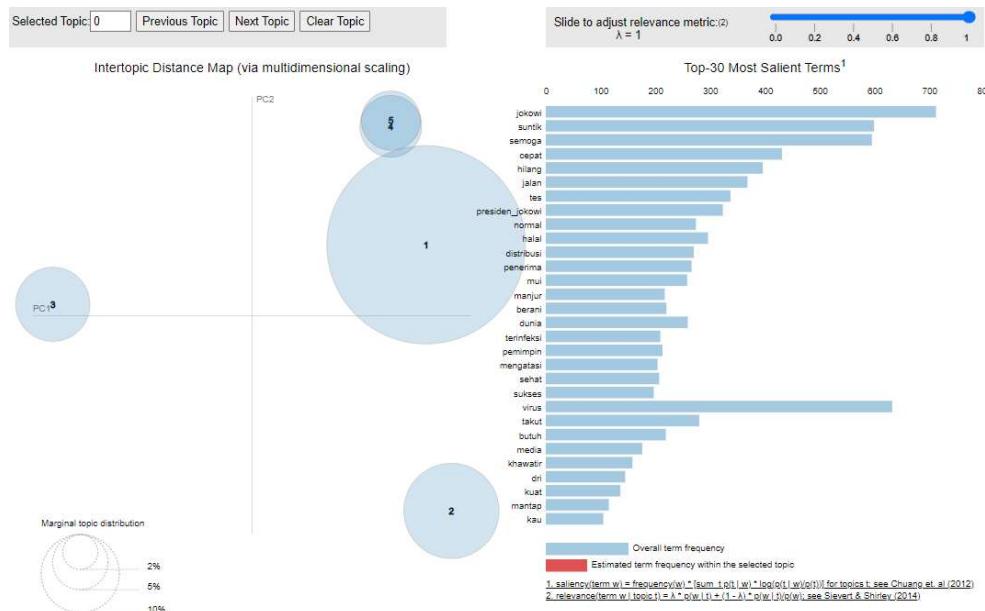


Gambar 8 Grafik Koherensi Topik

Topik 5 diprediksi dapat memberikan interpretasi topik yang mempunyai pemaknaan yang tinggi. Nilai koherensi diagregasi dari nilai kata-kata yang saling mendukung Selanjutnya jumlah topik 5 akan digunakan pada proses pemodelan topik.

### 4.3 Pemodelan Topik

Pemodelan topik menggunakan LDA menghasilkan topik yang mengandung probabilitas kata-kata. Berikut merupakan visualisasi dengan pyLDAviz mengenai hubungan antar topik dan kata yang paling umum ada pada dataset.



Gambar 9 Visualisasi pyLDAviz

Berikut 5 topik dengan 15 kata yang mempunyai nilai probabilitas tertinggi pada setiap topik :

Tabel 4 Topik yang dihasilkan LDA

No	Topik	Kata yang mewakili
1	Varian Vaksinasi	'tidak', 'covid', 'virus', 'indonesia', 'pemerintah', 'sinovac', 'gratis', 'china', 'disuntik', 'warga', 'negara', 'saya', 'masyarakat', 'pfizer', 'rakyat'
2	Ketakutan terhadap Covid-19 dan ajakan menggunakan protokol kesehatan	'takut', 'virus', 'dokter', 'lawan', 'tubuh', 'pakai', 'hidup', 'ayo', 'protokol_kesehatan', 'mutasi', 'mencegah', 'mari', 'kasih'
3	Kehalalan dari vaksin yang didistribusikan pemerintah	'jokowi', 'suntik', 'halal', 'distribusi', 'penerima', 'mui', 'haram', 'pelaksanaan'
4	Keinginan dan harapan untuk bisa pulih dari pandemi	'jalan', 'tes', 'normal', 'berani', 'pimpin', 'mengatasi', 'sukses', 'khawatir', 'semoga', 'lekas'
5	Harapan dan doa untuk semua orang	'semoga', 'cepat', 'hilang', 'manjur', 'terinfeksi', 'sehat', 'kuat', 'pergi', 'aamiin', 'selesai'

#### 4.4 Pembagian Data

Data dibagi menjadi data training dan data testing dengan proporsi 80% dan 20%. Jumlah pembagian data latih dan data uji dapat dilihat pada Tabel 5.

Tabel 5 Jumlah pembagian data

Sentimen	Data Latih	Data Uji
Positif	4685	648
Netral	7751	1957
Negatif	2608	1156
Total	15044	3761

#### 4.5 Hyperparameter Tuning

*Hyperparameter Tuning* dilakukan dengan menggunakan metode grid search cross validation dengan *10-fold cross validation*. Ruang pencarian untuk tiap *hyperparameter* adalah *minimum sample leaf* 1-5, *minimum sample split* 2-5 dan *n estimator* 100, 200, 300 dan 1000. Nilai *tuning* yang didapatkan dari setiap penerapan penanganan asumsi data *imbalance* dapat dilihat pada Tabel 6.

Tabel 6 Parameter terbaik hasil *hyperparameter tuning*

Penanganan Data	Parameter
Data asli	min_samples_split=4, n_estimators=1000
<i>Random Under Sampling</i> (RUS)	min_samples_split=5, n_estimators=1000
<i>Random Over Sampling</i> (ROS)	n_estimators=300
Over Sampling dengan SMOTE	n_estimators=300

#### 4.6 Evaluasi Model Klasifikasi dengan Data Asli

Nilai akurasi dan *F1-score* dari model ini terlihat menghasilkan sebesar 81% dan 70% dengan data asli. Terlihat nilai recall dan *F1-score* pada sentimen negatif terlihat sangat kecil yaitu 23% dan 36%. *Classification Report* dari data asli dapat dilihat pada Tabel 7.

Tabel 7 *Classification Report* dari data asli

	Precision	Recall	F1-score	Support
-1	0.81	0.23	0.36	648
0	0.78	0.93	0.85	1957
1	0.87	0.93	0.90	1156
Accuracy			0.81	3761
Macro Avg	0.82	0.70	0.70	3761

Berdasarkan data yang didapat dari *confusion matrix* yang merupakan data uji, terdapat 148 tweet negatif, 448 tweet netral, dan 52 tweet positif yang diklasifikasikan sebagai negatif. Terdapat 35 tweet negatif, 1819 tweet netral, dan 103 tweet positif yang diklasifikasikan sebagai netral. Terdapat 0 tweet negatif, 76 tweet netral dan 1080 tweet positif yang diklasifikasikan sebagai positif. *Confusion Matrix* dari data asli dapat dilihat pada Tabel 8.

Tabel 8 *Confusion Matrix* dari data asli

	-1	0	1
-1	148	448	52
0	35	1819	103
1	0	76	1080

Jika data tweet diasumsikan mempunyai data yang tidak seimbang karena berbanding dengan rasio 3:5:9, selanjutnya data akan diperlakukan dengan menggunakan metode *Random Under Sampling* (RUS), *Random Over Sampling* (ROS) dan *Over Sampling* menggunakan SMOTE.

#### 4.7 Evaluasi Model Klasifikasi dengan *Random Under Sampling* (RUS)

*Random Under Sampling* (RUS) menggunakan data penyetaraan jumlah data dengan data terkecil dari kelas yang ada. Pada kasus ini, data kelas negatif mempunyai jumlah data latih terkecil yaitu 2605 tweet. Data tiap kelas akan disetarakan menjadi 2605 tweet dengan pilihan acak. Setelah dilakukan klasifikasi, nilai akurasi turun dari 81% menjadi 77%. Nilai *F1-score* model terlihat naik dari 70% menjadi 75%. Nilai *recall* dan nilai *F1-score* kelas negatif naik menjadi 83% dan 62%. *Classification Report* dan *Confusion Matrix* dari data RUS dapat dilihat pada Tabel 9 dan Tabel 10.

Tabel 9 *Classification Report* dari data *Random Under Sampling* (RUS)

	Precision	Recall	F1-score	Support
-1	0.49	0.83	0.62	651
0	0.93	0.62	0.74	1941
1	0.83	0.97	0.90	1168
Accuracy			0.77	3760
Macro Avg	0.81	0.81	0.75	3760

Tabel 10 *Confusion Matrix* dari data *Random Under Sampling* (RUS)

	-1	0	1
-1	553	69	49
0	557	1192	192
1	0	31	1137

#### 4.8 Evaluasi Model Klasifikasi dengan *Random Over Sampling* (ROS)

*Random Over Sampling* (ROS) menggunakan data penyetaraan jumlah data latih dengan data terbesar dari kelas yang ada. Pada kasus ini, data kelas netral mempunyai jumlah data terbesar yaitu 7751 tweet. Data tiap kelas akan disetarakan menjadi 7751 tweet dengan pilihan acak. Nilai akurasi mempunyai nilai yang sama jika dibandingkan data asli. *F1-score* model terlihat naik dibanding data asli sebesar 70% menjadi 74%. *Classification Report* dan *Confusion Matrix* dari data RUS dapat dilihat pada Tabel 11 dan Tabel 12.

Tabel 11 *Classification Report* dari data *Random Under Sampling* (RUS)

	Precision	Recall	F1-score	Support
-1	0.62	0.41	0.50	651
0	0.81	0.85	0.83	1941
1	0.87	0.94	0.90	1168
Accuracy			0.81	3760
Macro Avg	0.76	0.74	0.74	3760

Tabel 12 *Confusion Matrix* dari data *Random Under Sampling* (RUS)

	-1	0	1
-1	270	331	50
0	163	1659	119
1	2	68	1098

#### 4.9 Evaluasi Model Klasifikasi dengan *Over Sampling SMOTE*

*Over Sampling* dengan SMOTE menggunakan data penyetaraan dengan data terbesar dari kelas yang ada. Pada kasus ini, data kelas netral mempunyai jumlah data terkecil yaitu 7751 tweet. Data tiap kelas akan disetarakan menjadi 7751 tweet dengan pilihan acak. Kekurangan data pada kelas positif dan negatif dilengkapi dengan sintesis data oleh SMOTE. Terlihat nilai akurasi dan *F1-score*

meningkat menjadi 85%. *Classification Report* dan *Confusion Matrix* dari data *Over Sampling* dengan SMOTE dapat dilihat pada Tabel 13 dan Tabel 14.

Tabel 13 *Classification Report* dari data *Over Sampling* dengan SMOTE

	Precision	Recall	F1-score	Support
-1	0.71	0.58	0.64	648
0	0.87	0.87	0.87	1957
1	0.88	0.97	0.92	1156
Accuracy			0.85	3761
Macro Avg	0.82	0.81	0.81	3761

Tabel 14 *Confusion Matrix* dari data *Over Sampling* dengan SMOTE

	-1	0	1
-1	379	222	47
0	153	1693	111
1	2	36	1118

#### 4.10 Perbandingan Evaluasi Model Random Forest *Imbalance Data*

Perbandingan nilai akurasi dan *F1-score* dari data asli, data dengan perlakuan *Random Under Sampling* (RUS), *Random Over Sampling* (ROS) dan *Over Sampling* menggunakan SMOTE dapat dilihat pada Tabel 15.

Tabel 15 Perbandingan berdasarkan perlakuan data

	Akurasi	F1-score
Data Asli	81%	70%
<i>Random Under Sampling</i> (RUS)	77%	75%
<i>Random Over Sampling</i> (ROS)	81%	74%
<i>Over Sampling</i> menggunakan SMOTE	85%	81%

#### 4.11 Perbandingan Klasifikasi dan Waktu Random Forest dengan BERT

Sebagai perbandingan hasil *Random Forest* dengan metode lain, metode *neural network* dengan algoritme BERT digunakan sebagai pembanding untuk melihat perbedaan performa pada dataset yang sama. Data training diperlakukan dengan metode data asli dan *Random Over Sampling* (ROS). Klasifikasi dengan algoritme BERT menunjukkan hasil yang lebih besar dibandingkan dengan algoritme *random forest* dengan data yang sama. Terlihat nilai akurasi naik dari 79% menjadi 82%. *Classification Report* dan *Confusion Matrix* dari klasifikasi BERT dari data asli dapat dilihat pada Tabel 16 dan Tabel 17.

Tabel 16 *Classification Report* klasifikasi BERT dengan data asli

	Precision	Recall	F1-score	Support
-1	0.61	0.57	0.59	651
0	0.82	0.86	0.84	1941
1	0.93	0.88	0.91	1168
Accuracy			0.82	3760
Macro Avg	0.78	0.77	0.78	3760

Tabel 17 *Confusion Matrix* klasifikasi BERT data asli

	-1	0	1
-1	371	254	26
0	225	1669	47
1	17	124	1027

Klasifikasi pemodelan BERT dengan data *Random Over Sampling* (ROS) dilakukan, nilai akurasi dan *F1-Score* terlihat meningkat dibandingkan data ROS dengan model klasifikasi *Random Forest*. *Classification Report* dan *Confusion Matrix* dengan data ROS dapat dilihat pada tabel 18 dan Tabel 19.

Tabel 18 *Classification Report* klasifikasi BERT dengan data ROS

	Precision	Recall	F1-score	Support
-1	0.60	0.60	0.60	651
0	0.85	0.83	0.84	1941
1	0.91	0.93	0.92	1168
Accuracy			0.82	3760
Macro Avg	0.78	0.79	0.79	3760

Tabel 19 *Confusion Matrix* klasifikasi BERT data ROS

	-1	0	1
-1	391	226	34
0	246	1620	75
1	16	71	1081

Algoritme BERT yang menggunakan konsep *neural network* dapat menghasilkan hasil dari akurasi dan *F1-score* yang lebih baik. Namun disisi lain, biaya komputasi menggunakan algoritme BERT memerlukan *computational cost* yang cukup besar. Jika *computational cost* diukur dari waktu proses, maka algoritme BERT memerlukan waktu yang lebih banyak. Perbandingan penggunaan waktu untuk mendapatkan hasil dari algoritme *Random Forest* dan BERT dapat dilihat pada Tabel 20.

Tabel 20 Perhitungan waktu komputasi

Algoritme	Waktu Komputasi
<i>Random Forest</i> data asli	100 menit
<i>Random Forest</i> data RUS	18 menit
<i>Random Forest</i> data ROS	140 menit
<i>Random Forest</i> data Over Sampling SMOTE	134 menit
BERT data asli	454 menit
BERT data ROS	759 menit

## V SIMPULAN DAN SARAN

### 5.1 Simpulan

Metode LDA menghasilkan 5 topik yaitu varian vaksinasi Covid-19 oleh pemerintah dan diskusi terkait rakyat disuntik vaksin, ketakutan terhadap virus Covid-19 dan ajakan menggunakan protokol kesehatan untuk mencegah mutasi dari virus yang berkembang, kehalalan dari vaksin yang didistribusikan oleh pemerintah, keinginan untuk pemimpin dapat sukses mengatasi kondisi saat ini dan harapan untuk bisa kembali pulih, harapan dan doa untuk orang yang terinfeksi bisa segera sehat. Pemodelan dengan klasifikasi *Random Forest* dengan data asli menghasilkan akurasi 81% dan *F1-score* sebesar 70%. Pengklasifikasian dengan data dengan metode *Random Under Sampling* (RUS) menghasilkan akurasi 77% dan *F1-score* sebesar 75%. Pengklasifikasian dengan data metode Random Over Sampling (ROS) menghasilkan akurasi 81% dan *F1-score* 74%. Pengklasifikasian dengan data *Oversampling* dengan SMOTE menghasilkan akurasi 85% dan *F1-Score* sebesar 81%.

### 5.2 Saran

Pelabelan secara manual memungkinkan subyektifitas dalam pemilihan label, yang akan mempengaruhi *machine learning* dalam mempelajari data. Konsensus dan kepakaran diperlukan saat pelabelan *tweet* yang mempunyai beberapa makna. Penggunaan *classifier* yang tepat dapat meningkatkan kredibilitas model. Model klasifikasi berbasis *neural network* bisa menjadi arah penelitian kedepan mempunyai performa dan kapabilitas yang bisa ditingkatkan dengan berbagai pengembangan.

## DAFTAR PUSTAKA

- al Amrani Y, Lazaar M, el Kadirp KE. 2018. Random forest and support vector machine based hybrid approach to sentiment analysis. Di dalam: *Procedia Computer Science*. Volume ke-127. Elsevier B.V. hlm 511–520.
- Angiani G, Ferrari L, Fontanini T, Fornacciari P, Iotti E, Magliani F, Manicardi S. 2016. A Comparison between Preprocessing Techniques for Sentiment Analysis in Twitter. <http://alt.qcri.org/semeval2016/task4/>.
- Asuncion A, Welling M, Smyth P, Teh YW. 2009. On Smoothing and Inference for Topic Models.
- Azhar AN, Khodra ML. 2020. *2020 7th International Conference on Advance Informatics: Concepts, Theory and Applications (ICAICTA) : September 8-9, 2020, online conference*.
- Barkur G, Vibha, Kamath GB. 2020. Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from India. *Asian Journal of Psychiatry*. 51. doi:10.1016/j.ajp.2020.102089.
- Bavel JJV, Baicker K, Boggio PS, Capraro V, Cichocka A, Cikara M, Crockett MJ, Crum AJ, Douglas KM, Druckman JN, et al. 2020. Using social and behavioural science to support COVID-19 pandemic response. *Nature Human Behaviour*. 4(5):460–471. doi:10.1038/s41562-020-0884-z.
- Bisri A, Rachmatika R. 2019. Integrasi Gradient Boosted Trees dengan SMOTE dan Bagging untuk Deteksi Kelulusan Mahasiswa. Volume ke-8. <https://forlap.ristekdikti.go.id>.
- Blei DM. 2012. Introduction to Probabilistic Topic Models.
- Breiman L. 2001. Random Forests. Volume ke-45.
- Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C. 2009. Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem. Volume ke-5476.
- Chairunnisa QA, Herdiyeni Y, Kusuma M, Hardhienata D, Adisantoso J. 2022. Analisis Sentimen Pengguna Twitter Terhadap Program Vaksinasi Covid-19 di Indonesia Menggunakan Algoritme Support Vector Machine. <http://journal.ipb.ac.id/index.php/jika>.
- Chuang J. 2012. *Proceedings of the International Working Conference on Advanced Visual Interfaces*.
- Cuello-Garcia C, Pérez-Gaxiola G, van Amelsvoort L. 2020. Social media can have an impact on how we manage and investigate the COVID-19 pandemic. *Journal of Clinical Epidemiology*. 127:198–201. doi:10.1016/j.jclinepi.2020.06.028.
- Devlin J, Chang M-W, Lee K, Toutanova K. 2018 Okt 10. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. <http://arxiv.org/abs/1810.04805>.
- Garcia K, Berton L. 2021. Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. *Applied Soft Computing*. 101. doi:10.1016/j.asoc.2020.107057.
- Hakim KF, Silvianti P, Soleh AM. 2021. LATENT DIRICHLET ALLOCATION DALAM IDENTIFIKASI RESPON MASYARAKAT INDONESIA

- TERHADAP COVID-19 TAHUN 2020-2021. *Xplore: Journal of Statistics*. 10(3):248–257. doi:10.29244/xplore.v10i3.836.
- Hoffman MD, Blei DM, Bach F. 2010. Online Learning for Latent Dirichlet Allocation.
- Jamison A, Broniatowski DA, Smith MC, Parikh KS, Malik A, Dredze M, Quinn SC. 2020. Adapting and extending a typology to identify vaccine misinformation on twitter. *American Journal of Public Health*. 110:S331–S339. doi:10.2105/AJPH.2020.305940.
- Laurensz B, Sentimen A, Sediyo E. 2021. Analisis Sentimen Masyarakat terhadap Tindakan Vaksinasi dalam Upaya Mengatasi Pandemi Covid-19 (Analysis of Public Sentiment on Vaccination in Efforts to Overcome the Covid-19 Pandemic). Volume ke-10.
- Lyu JC, Han E le, Luli GK. 2021. Covid-19 vaccine-related discussion on twitter: Topic modeling and sentiment analysis. *Journal of Medical Internet Research*. 23(6). doi:10.2196/24435.
- Pozzi FA, Fersini E, Messina E, Liu B. 2017. *Sentiment analysis in social networks*.
- Prakosa HA, Riyanto AB, Nasiroh DS. 2021. Analisis sentimen dan pemodelan topik pandemi Covid-19 pada media sosial Twitter menggunakan Naïve Bayes Classifier dan Latent Dirichlet Allocation.
- Qamar AM, Razzaq MA, Angel M, Pardo A, Syed H, Bilal M. 2014. Prediction and analysis of Pakistan election 2013 based on sentiment analysis Related papers A review on political analysis and social media Prediction and Analysis of Pakistan Election 2013 based on Sentiment Analysis. ASONAM.
- Rachman FF, Pramana S. 2020. Analisis Sentimen Pro dan Kontra Masyarakat Indonesia tentang Vaksin COVID-19 pada Media Sosial Twitter. Volume ke-8.
- Rácz A, Bajusz D, Héberger K. 2021. Effect of dataset size and train/test split ratios in qsar/qspr multiclass classification. *Molecules*. 26(4). doi:10.3390/molecules26041111.
- Rajaraman A, Ullman J. 2011. Mining of Massive Datasets. <http://en.wikipedia.org/wiki/1854>.
- Rockett RJ, Arnott A, Lam C, Sadsad R, Timms V, Gray KA, Eden JS, Chang S, Gall M, Draper J, et al. 2020. Revealing COVID-19 transmission in Australia by SARS-CoV-2 genome sequencing and agent-based modeling. *Nature Medicine*. 26(9):1398–1404. doi:10.1038/s41591-020-1000-7.
- Röder M, Both A, Hinneburg A. 2015. Exploring the space of topic coherence measures. Di dalam: *WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining*. Association for Computing Machinery. hlm 399–408.
- Rodríguez JD, Pérez A, Lozano JA. 2010. Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 32(3):569–575. doi:10.1109/TPAMI.2009.187.
- Rosenberg H, Syed S, Rezaie S. 2020. The Twitter pandemic: The critical role of Twitter in the dissemination of medical information and misinformation during the COVID-19 pandemic. *Canadian Journal of Emergency Medicine*. 22(4):418–421. doi:10.1017/cem.2020.361.

- S. S, K.V. P. 2020. Sentiment analysis of malayalam tweets using machine learning techniques. *ICT Express.* 6(4):300–305. doi:10.1016/j.icte.2020.04.003.
- Salton G, Buckley C. 1988. TERM-WEIGHTING APPROACHES IN AUTOMATIC TEXT RETRIEVAL.
- Sastramidjaja Y, Rosli AA. 2021. Tracking the Swelling COVID-19 Vaccine Chatter on TikTok in Indonesia. <https://www.youtube.com/watch?v=7iVMVG7bgYY>.
- Sholehah NA. 2018. ANALISIS SENTIMEN MENGGUNAKAN NAIVE BAYES PADA DATA TWITTER BAHASA INDONESIA.
- Shurrah S, Shannak Y, Almshnanah A, Khazaleh H, Najadat H. 2021. Attitudes Evaluation Toward COVID-19 Pandemic: An Application of Twitter Sentiment Analysis and Latent Dirichlet Allocation. Di dalam: *2021 12th International Conference on Information and Communication Systems, ICICS 2021*. Institute of Electrical and Electronics Engineers Inc. hlm 265–272.
- Sievert C, Shirley KE. 2014. LDAvis: A method for visualizing and interpreting topics.
- Xue J, Chen J, Chen C, Zheng C, Li S, Zhu T. 2020a. Public discourse and sentiment during the COVID 19 pandemic: Using latent dirichlet allocation for topic modeling on twitter. *PLoS ONE.* 15 9 September. doi:10.1371/journal.pone.0239441.
- Xue J, Chen J, Hu R, Chen C, Zheng C, Su Y, Zhu T. 2020b. Twitter discussions and emotions about the COVID-19 pandemic: Machine learning approach. *Journal of Medical Internet Research.* 22(11). doi:10.2196/20550.

## RIWAYAT HIDUP

Penulis dilahirkan di kota Bengkulu pada 24 Juli 1999 sebagai anak ke pertama dari pasangan bapak Hansastri dan ibu Elimarlina. Pendidikan sekolah menengah atas (SMA) ditempuh di SMAIT Nurul Fikri , dan lulus pada tahun 2017 Pada tahun 2017 penulis diterima sebagai mahasiswa program sarjana (S-1) di Program Studi Ilmu Komputer di IPB.

Selama mengikuti program S-1, penulis aktif menjadi anggota Biro Internal dan Eksternal BEM FMIPA IPB pada tahun kedua, Kepala Biro Internal dan Pengembangan BEM FMIPA IPB pada tahun ketiga, dan Ketua G-Family dan FMIPA Family Day (Masa Pengenalan Fakultas FMIPA) .

## LAPORAN PEMBAYARAN SPP

NAMA : Amin Elhan  
NIM : G64170109  
PROGRAM STUDI : Ilmu Komputer  
TAHUN MASUK : 2017  
STRATA : S1  
TANGGAL CETAK : 15/09/2022 11:46

SEMESTER	TAGIHAN (IDR)	PEMBAYARAN (IDR)	SELISIH (IDR)
1	Rp 29.900.000	Rp 29.900.000	Rp 0
2	Rp 11.000.000	Rp 11.000.000	Rp 0
3	Rp 11.000.000	Rp 11.000.000	Rp 0
4	Rp 11.000.000	Rp 11.000.000	Rp 0
5	Rp 11.000.000	Rp 11.000.000	Rp 0
6	Rp 11.000.000	Rp 11.000.000	Rp 0
7	Rp 11.000.000	Rp 11.000.000	Rp 0
8	Rp 11.000.000	Rp 11.000.000	Rp 0
9	Rp 3.850.000	Rp 3.850.000	Rp 0
10	Rp 5.500.000	Rp 5.500.000	Rp 0
11	Rp 5.500.000	Rp 5.500.000	Rp 0
<b>Total</b>	<b>Rp 121.750.000</b>	<b>Rp 121.750.000</b>	<b>Rp 0</b>

<b>Total Tagihan</b>	<b>Rp 121.750.000</b>
<b>Total Pembayaran</b>	<b>Rp 121.750.000</b>
<b>Selisih</b>	<b>Rp 0</b>

Status Pembayaran : LUNAS

Informasi ini hasil cetakan komputer dan tidak memerlukan tanda tangan