

**PEMODELAN TOPIK CUITAN MEDIA ONLINE PADA TWITTER
MENGUNAKAN LATENT DIRICHLET ALLOCATION**

SKRIPSI

Diajukan untuk memenuhi salah satu syarat memperoleh gelar Sarjana
Komputer pada Program Studi S1 Informatika



Disusun Oleh
Anlsykurli Faza Ramadhanl
18.11.2438

**PROGRAM SARJANA
PROGRAM STUDI INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2023**

**PEMODELAN TOPIK CUITAN MEDIA ONLINE PADA TWITTER
MENGUNAKAN LATENT DIRICHLET ALLOCATION**

SKRIPSI

Diajukan untuk memenuhi salah satu syarat memperoleh gelar Sarjana
Komputer pada Program Studi S1 Informatika



Disusun Oleh
Anlsykurll Faza Ramadhani
18.11.2438

**PROGRAM SARJANA
PROGRAM STUDI INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2023**

HALAMAN PERSETUJUAN

SKRIPSI

**PEMODELAN TOPIK CUITAN MEDIA ONLINE PADA TWITTER
MENGUNAKAN LATENT DIRICHLET ALLOCATION**

yang disusun dan diajukan oleh
Anisykurli Faza Ramadhani
18.11.2438

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 7 Juni 2023

Dosen Pembimbing,



Anggit Dwi Hartanto, M.Kom
NIK. 190302163

HALAMAN PENGESAHAN

SKRIPSI

PEMODELAN TOPIK CUITAN MEDIA ONLINE PADA TWITTER
MENGUNAKAN LATENT DIRICHLET ALLOCATION

yang disusun dan diajukan oleh
Anisykurli Faza Ramadhani
18.11.2438

Telah dipertahankan di depan Dewan Penguji
pada tanggal 26 Juni 2023

Susunan Dewan Penguji

Nama Penguji

Arif Dwi Laksito, M.kom
NIK. 190302150

Andi Sunyoto, M.Kom., Dr.
NIK. 190302052

Anggit Dwi Hartanto, M.Kom
NIK. 190302163

Tanda Tangan



Skrripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 26 Juni 2023

DEKAN FAKULTAS ILMU KOMPUTER



Hanif Al Fatta, S.Kom., M.Kom.
NIK. 190302096

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama Mahasiswa : Anisykurli Faza Ramadhani
NIM : 18.11.2438

Menyatakan bahwa skripsi dengan judul berikut:

PEMODELAN TOPIK CUITAN MEDIA ONLINE PADA TWITTER MENGUNAKAN LATENT DIRICHLET ALLOCATION

Dosen Pembimbing : Anggit Dwi Hartanto, M.Kom

1. Karya tulis ini adalah benar-benar **ASLI** dan **BELUM PERNAH** diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian **SAYA** sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab **SAYA**, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
5. Pernyataan ini **SAYA** buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka **SAYA** bersedia menerima **SANKSI AKADEMIK** dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 26 Juni 2023

Yang Menyatakan,



Anisykurli Faza Ramadhani

HALAMAN PERSEMBAHAN

Dengan telah diselesaikannya skripsi ini, penulis mempersembahkan karya tulis ini kepada:

1. Kedua orang tua saya yang selalu memberi motivasi dan doa untuk saya.
2. Bapak Anggit Dwi Hartanto, M.Kom. yang telah membimbing saya dalam mengerjakan skripsi.
3. Teman-teman terdekat saya yang selalu memberikan semangat dalam mengerjakan skripsi.
4. Teman-teman kelas 18-IF09 yang menemani saya dari awal hingga akhir perkuliahan di Universitas Amikom Yogyakarta.



KATA PENGANTAR

Puji dan syukur penulis ucapkan kehadirat Allah SWT. Yang telah memberikan rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan skripsi ini dengan judul **“PEMODELAN TOPIK CUITAN MEDIA ONLINE PADA TWITTER MENGGUNAKAN LATENT DIRICHLET ALLOCATION”** yang menjadi salah satu syarat untuk menyelesaikan masa studi program sarjana di Universitas AMIKOM Yogyakarta.

Dalam menyusun skripsi ini, penulis banyak mendapat dukungan, bimbingan dan kemudahan dari berbagai pihak. Pada kesempatan ini penulis ingin mengucapkan terima kasih kepada:

1. Bapak Prof. Dr. M. Suyanto, M.M. selaku Rektor Universitas AMIKOM Yogyakarta.
2. Bapak Hanif Al Fatta, M.Kom. selaku Dekan Fakultas Ilmu Komputer Universitas AMIKOM Yogyakarta.
3. Ibu Windha Mega Pradnya D., M.Kom. selaku Ketua Program Studi SI Informatika Universitas AMIKOM Yogyakarta.
4. Bapak Anggit Dwi Hartanto, M.Kom. selaku Dosen Pembimbing yang selalu memberikan bimbingan, saran dan masukan dalam penulisan skripsi.

Penulis menyadari bahwa skripsi ini masih memiliki banyak kekurangan, maka dari itu kritik dan saran yang membangun sangat dibutuhkan penulis untuk perbaikan karya selanjutnya. Akhir kata **semoga penelitian ini dapat bermanfaat bagi pembaca dan menambah wawasan khususnya dalam bidang *natural language processing*.**

DAFTAR ISI

HALAMAN JUDUL.....	i
HALAMAN PERSETUJUAN.....	ii
HALAMAN PENGESAHAN.....	iii
HALAMAN PERNYATAAN KEASLIAN SKRIPSI.....	iv
HALAMAN PERSEMBAHAN.....	v
KATA PENGANTAR.....	vi
DAFTAR ISI.....	vii
DAFTAR TABEL.....	xi
DAFTAR GAMBAR.....	xii
DAFTAR KODE.....	xiv
INTISARI.....	xv
ABSTRACT.....	xvi
BAB I.....	1
PENDAHULUAN.....	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah.....	2
1.3 Batasan Masalah.....	2
1.4 Maksud dan Tujuan Penelitian.....	3
1.5 Manfaat Penelitian.....	3
1.6 Metode Penelitian.....	3
1.6.1 Pengumpulan data.....	4
1.6.2 Preprocessing.....	4
1.6.3 Analisis data.....	4
1.6.4 Pengembangan.....	4
1.6.5 Analisis hasil.....	4
1.7 Sistematika Penulisan.....	4
BAB II.....	6
TINJAUAN PUSTAKA.....	6

2.1 Studi Literatur.....	6
2.2 Dasar Teori.....	9
2.2.1 Media online.....	9
2.2.2 Twitter application programming interface (API).....	10
2.2.3 Text mining.....	10
2.2.4 Preprocessing.....	11
2.2.4.1 Tokenization.....	11
2.2.4.2 Standardization dan cleaning.....	12
2.2.4.3 Stemming.....	12
2.2.4.4 Stopword removal.....	13
2.2.5 Pemodelan topik.....	13
2.2.6 Latent dirichlet allocation.....	14
2.2.7 Topic coherence.....	17
BAB III.....	19
METODE PENELITIAN.....	19
3.1 Alur Penelitian.....	19
3.1.1 Studi literatur.....	20
3.1.2 Pengambilan dan analisis data.....	20
3.1.3 Preprocessing.....	21
3.1.4 Pemodelan topik dengan LDA.....	21
3.1.5 Analisis hasil.....	22
3.2 Alat dan Bahan Penelitian.....	22
3.2.1 Data penelitian.....	22
3.2.2 Alat penelitian.....	23
3.2.2.1 Perangkat keras (hardware).....	24
3.2.2.2 Perangkat lunak (software).....	24
BAB IV.....	26
HASIL DAN PEMBAHASAN.....	26
4.1 Pengambilan dan Analisis Data.....	26

4.2 Preprocessing.....	31
4.2.1 Standardization dan cleaning.....	31
4.2.1.1 Lowercasing.....	31
4.2.1.2 Remove newline.....	32
4.2.1.3 Remove hashtag.....	33
4.2.1.4 Remove URL.....	34
4.2.1.5 Remove punctuation.....	35
4.2.1.6 Remove digits.....	35
4.2.1.7 Remove emoji.....	36
4.2.2 Tokenization.....	37
4.2.3 Stopword removal.....	38
4.2.4 Stemming.....	39
4.2.5 Visualisasi frekuensi kata hasil preprocessing.....	40
4.3 Pemodelan Topik.....	43
4.3.1 Import dataset.....	43
4.3.2 Membuat dictionary dan corpus.....	43
4.3.3 Evaluasi topic coherence.....	45
4.3.4 Pembuatan model optimal dengan LDA.....	48
4.3.5 Visualisasi hasil LDA.....	54
4.4 Analisis Hasil.....	56
4.4.1 Topik ke-1.....	56
4.4.2 Topik ke-2.....	58
4.4.3 Topik ke-3.....	59
4.4.4 Topik ke-4.....	61
4.4.5 Topik ke-5.....	63
4.4.6 Topik ke-6.....	64
4.4.7 Topik ke-7.....	66
4.4.8 Topik ke-8.....	67
4.4.9 Topik ke-9.....	69

4.4.10 Topik ke-10.....	70
4.4.11 Topik ke-11.....	72
4.4.12 Topik ke-12.....	73
BAB V.....	76
PENUTUP.....	76
5.1 Kesimpulan.....	76
5.2 Saran.....	76
DAFTAR PUSTAKA.....	79
LAMPIRAN.....	82



DAFTAR TABEL

Tabel 2.1 Perbandingan penelitian.....	8
Tabel 3.1 Kebutuhan perangkat keras.....	24
Tabel 3.2 Kebutuhan perangkat lunak.....	24
Tabel 3.2 Kebutuhan perangkat lunak (lanjutan)	25
Tabel 4.1 Atribut dataset.....	26
Tabel 4.2 Timeline data.....	27
Tabel 4.3 Jumlah data pada setiap akun.....	27
Tabel 4.3 Jumlah data pada setiap akun (lanjutan).....	28
Tabel 4.4 Sampel dataset csv.....	30
Tabel 4.5 Contoh lowercasing.....	32
Tabel 4.6 Contoh remove newline.....	33
Tabel 4.7 Contoh remove hashtag.....	34
Tabel 4.8 Contoh remove url.....	34
Tabel 4.9 Contoh remove punctuation.....	35
Tabel 4.10 Contoh remove digits.....	36
Tabel 4.11 Contoh remove emoji.....	37
Tabel 4.12 Contoh tokenization.....	38
Tabel 4.13 Contoh stopword removal.....	39
Tabel 4.14 Contoh stemming.....	40
Tabel 4.15 Sampel data dictionary.....	44
Tabel 4.16 Sampel data corpus.....	45
Tabel 4.17 Coherence score pada setiap topik.....	47
Tabel 4.18 Persebaran kata pada setiap topik.....	48
Tabel 4.18 Persebaran kata pada setiap topik (lanjutan).....	49
Tabel 4.19 Sampel data distribusi topik pada setiap dokumen.....	50
Tabel 4.20 Sampel data topik dominan pada setiap dokumen.....	51
Tabel 4.21 Dokumen dengan probabilitas tertinggi pada setiap topik.....	52
Tabel 4.21 Dokumen dengan probabilitas tertinggi pada setiap topik (lanjutan)	53

DAFTAR GAMBAR

Gambar 2.1 Alur text mining.....	11
Gambar 2.2 Proses tokenization.....	12
Gambar 2.3 Konsep topic modeling	14
Gambar 2.4 Diagram LDA	16
Gambar 3.1 Alur penelitian	19
Gambar 3.2 Contoh tweet	20
Gambar 3.3 Alur pemodelan topik LDA.....	22
Gambar 3.4 Dataset mentah	23
Gambar 4.1 Frekuensi kata pada dataset.....	42
Gambar 4.2 Grafik nilai evaluasi topic coherence	46
Gambar 4.3 Hasil pemodelan topik dengan pyldavis	55
Gambar 4.4 Frekuensi kata seluruh topik dengan wordcloud.....	55
Gambar 4.5 Distribusi jumlah dokumen pada setiap topik	56
Gambar 4.6 Hasil pyldavis topik ke-1.....	57
Gambar 4.7 Hasil wordcloud topik ke-1.....	57
Gambar 4.8 Mind map pembahasan topik ke-1.....	58
Gambar 4.9 Hasil pyldavis topik ke-2.....	58
Gambar 4.10 Hasil wordcloud topik ke-2.....	59
Gambar 4.11 Mind map pembahasan topik ke-2.....	59
Gambar 4.12 Hasil pyldavis topik ke-3.....	60
Gambar 4.13 Hasil wordcloud topik ke-3.....	60
Gambar 4.14 Mind map pembahasan topik ke-3.....	61
Gambar 4.15 Hasil pyldavis topik ke-4.....	62
Gambar 4.16 Hasil wordcloud topik ke-4.....	62
Gambar 4.17 Mind map pembahasan topik ke-4.....	63
Gambar 4.18 Hasil pyldavis topik ke-5.....	63
Gambar 4.19 Hasil wordcloud topik ke-5	64
Gambar 4.20 Mind map pembahasan topik ke-5.....	64

Gambar 4.21 Hasil pyldavis topik ke-6	65
Gambar 4.22 Hasil wordcloud topik ke-6.....	65
Gambar 4.23 Mind map pembahasan topik ke-6	66
Gambar 4.24 Hasil pyldavis topik ke-7.....	66
Gambar 4.25 Hasil wordcloud topik ke-7	67
Gambar 4.26 Mind map pembahasan topik ke-7.....	67
Gambar 4.27 Hasil pyldavis topik ke-8	68
Gambar 4.28 Hasil wordcloud topik ke-8.....	68
Gambar 4.29 Mind map pembahasan topik ke-8	69
Gambar 4.30 Hasil pyldavis topik ke-9	69
Gambar 4.31 Hasil wordcloud topik ke-9.....	70
Gambar 4.32 Mind map pembahasan topik ke-9	70
Gambar 4.33 Hasil pyldavis topik ke-10	71
Gambar 4.34 Hasil wordcloud topik ke-10.....	71
Gambar 4.35 Mind map pembahasan topik ke-10	72
Gambar 4.36 Hasil pyldavis topik ke-11	72
Gambar 4.37 Hasil wordcloud topik ke-11.....	73
Gambar 4.38 Mind map pembahasan topik ke-11	73
Gambar 4.39 Hasil pyldavis topik ke-12.....	74
Gambar 4.40 Hasil wordcloud topik ke-12	74
Gambar 4.41 Mind map pembahasan topik ke-12.....	75

DAFTAR KODE

Kode 3.1 Query pengambilan tweet.....	23
Kode 4.1 Fungsi lowercasing.....	31
Kode 4.2 Fungsi remove newline.....	32
Kode 4.3 Fungsi remove hashtag.....	33
Kode 4.4 Fungsi remove url.....	34
Kode 4.5 Fungsi remove punctuation.....	35
Kode 4.6 Fungsi remove digits.....	36
Kode 4.7 Fungsi remove emoji.....	36
Kode 4.8 Tokenization.....	37
Kode 4.9 Stopword removal.....	38
Kode 4.9 Stopword removal (lanjutan).....	39
Kode 4.10 Stemming.....	40
Kode 4.11 Import dataset.....	43
Kode 4.12 Membuat dictionary.....	44
Kode 4.13 Membuat corpus.....	44
Kode 4.14 Evaluasi topic coherence.....	45
Kode 4.15 Pembuatan model optimal LDA.....	48
Kode 4.16 Visualisasi hasil LDA menggunakan pyLDAvis.....	54

INTISARI

Media sosial twitter merupakan salah satu *platform* yang wajib digunakan oleh media-media online di Indonesia untuk menyebarkan berita dalam bentuk *tweet*. Banyaknya *tweet* yang dibuat oleh media online menyebabkan sulit untuk menganalisis topik-topik apa saja yang dibahas dalam pemberitaan khususnya berita tentang covid-19. Tujuan dari penelitian ini adalah untuk menemukan topik-topik tersembunyi dari kumpulan *tweet* yang dibuat oleh beberapa media online berbahasa Indonesia dengan teknik pemodelan topik.

Metode LDA merupakan metode pemodelan topik populer yang banyak digunakan dan juga merupakan metode penyempurnaan dari metode-metode sebelumnya seperti *Latent Semantic Analysis* (LSA) dan *Probabilistic Latent Semantic Analysis* (PLSA). Untuk mendapatkan model yang maksimal, dilakukan tahapan *preprocessing* pada data teks yang akan digunakan dan juga evaluasi model menggunakan *topic coherence* untuk menentukan banyak topik yang akan dihasilkan, kemudian topik yang dihasilkan akan dilakukan analisis untuk setiap topiknya menggunakan visualisasi hasil, sehingga akan diketahui makna atau pembahasan apa yang terkandung pada masing-masing topik.

Pemodelan topik yang dilakukan menghasilkan 12 topik dengan *coherence score* sebesar 0,403044. Kata-kata yang muncul pada setiap topik dapat diinterpretasi dengan baik sehingga setiap topik memiliki makna yang dapat dipahami. Hasil menunjukkan bahwa topik tentang *update* kasus harian covid-19 muncul paling banyak di antara topik lainnya, sedangkan topik tentang perkembangan kasus positif di Jawa-Bali menjadi topik yang paling sedikit dibahas oleh media online.

Kata kunci: *Latent Dirichlet Allocation*, Pemodelan Topik, Berita, Media Online

ABSTRACT

Twitter social media is one of the platforms used by online media in Indonesia to spread news in the form of tweets. The large number of tweets made by online media makes it difficult to analyze what topics are discussed in the news, especially news about covid-19. The purpose of this research is to find hidden topics from a collection of tweets made by several Indonesian-language online media with topic modeling techniques.

The LDA method is a popular topic modeling method that is widely used and is also a refinement of previous methods such as Latent Semantic Analysis (LSA) and Probabilistic Latent Semantic Analysis (PLSA). To get the maximum model, the preprocessing stage is carried out on the text data to be used and also evaluates the model using topic coherence to determine the number of topics that will be generated, then the resulting topics will be analyzed for each topic using visualization of the results, so that it will be known what meaning or discussion is contained in each topic.

Topic modeling resulted in 12 topics with a coherence score of 0.403044. The words that appear in each topic can be interpreted well so that each topic has a meaning that can be understood. The results show that the topic of daily covid-19 case updates appears the most among other topics, while the topic of positive case developments in Java-Bali is the topic least discussed by online media.

Keywords: *Latent Dirichlet Allocation, Topic Modeling, News, Online Media*