PENERAPAN *LEXICON-BASED* DAN *LATENT DIRICHLET ALLOCATION* PADA *"CLASH OF CHAMPIONS" SEASON* 1

ALBERTO CRISTOFER LAUREN
NIMH1091211059

SKRIPSI

PROGRAM STUDI STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS TANJUNGPURA
PONTIANAK
2025

PENERAPAN *LEXICON-BASED* DAN *LATENT DIRICHLET ALLOCATION* PADA *"CLASH OF CHAMPIONS" SEASON* 1

ALBERTO CRISTOFER LAUREN
NIMH1091211059

Skripsi
Sebagai salah satu syarat untuk memperoleh gelar
Sarjana Statistika pada Program Studi Statistika

PROGRAM STUDI STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS TANJUNGPURA
PONTIANAK
2025

LEMBAR PENGESAHAN

Judul Tugas Akhir	: Penerapan *Lexicon-Based* Dan *Latent Dirichlet Allocation* Pada *"Clash Of Champions" Season* 1
Nama Mahasiswa	: Alberto Cristofer Lauren
NIM	: H1091211059
Jurusan/Program Studi	: Matematika/Statistika
Tanggal Lulus	: 31 Juli 2025
SK Pembimbing	: No.2633/UN22.8/TD.06/2024/Tanggal 2 September 2024
SK Penguji	: No.2454/UN22.8/TD.06/2025/Tanggal 25 Juli 2025

Dosen Pembimbing

Pembimbing I	Pembimbing II
Nurfitri Imro'ah, M.Si	Dr. Yundari, M.Sc
NIP 198907182019032021	NIP 198310202008012012

Dosen Penguji

Ketua Penguji	Anggota Penguji
Shantika Martha, M.Si	Meliana Pasaribu, M.Sc
NIP 198403082008122003	NIP 198710192019032006

Pimpinan Sidang (merangkap anggota penguji)	Sekretaris Sidang (merangkap anggota penguji)
Nurfitri Imro'ah, M.Si	Dr. Yundari, M.Sc
NIP 198907182019032021	NIP 198310202008012012

Mengesahkan
Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Tanjungpura

Prof. Dr. Gusrizal, S.Si., M.Si.
NIP 197108122000031001

PERNYATAAN INTEGRITAS AKADEMIK

Saya yang bertanda tangan di bawah ini:

Nama : Alberto Cristofer Lauren

NIM : H1091211059

Program Studi/Jurusan : Matematika/Statistika

Fakultas : Matematika dan Ilmu Pengetahuan Alam

dengan ini menyatakan bahwa dokumen ilmiah Tugas Akhir yang disajikan ini tidak mengandung unsur pelanggaran integritas akademik sesuai Peraturan Menteri Pendidikan, Kebudayaan, Riset, Dan Teknologi Republik Indonesia Nomor 39 Tahun 2021. Apabila di kemudian hari dokumen ilmiah Tugas Akhir ini mengandung unsur pelanggaran integritas akademik sesuai ketentuan perundangan tersebut, maka saya bersedia menerima sanksi akademik dan/atau sanksi hukum yang berlaku.

Demikian pernyataan ini untuk digunakan sebagaimana mestinya.

Pontianak, 25 Juli 2025



Alberto Cristofer Lauren

NIM H1091211059

Penerapan *Lexicon-Based* dan *Latent Dirichlet Allocation* Pada *"Clash of Champions" Season* 1

Abstrak

Clash of Champions (CoC) merupakan kompetisi berbasis pendidikan yang diproduksi oleh Ruangguru, salah satu industri pendidikan terbesar di Indonesia. Tayangan ini berhasil menarik perhatian publik, yang ditunjukan melalui berbagai respon dan tanggapan penonton pada kolom komentar *YouTube*. Penelitian ini bertujuan untuk menganalisis sentimen penonton menggunakan metode *Lexicon-Based* serta mengidentifikasi topik pembahasan dalam kolom komentar menggunakan pemodelan topik *Latent Dirichlet Allocation* (LDA). Data dikumpulkan melalui teknik *data extraction* dengan bantuan API *YouTube* selama periode penayangan, yaitu dari 29 Juni 2024 hingga 17 Agustus 2024. Sebanyak 35.047 komentar dari seluruh episode telah melalui proses *text preprocessing* sebelum dianalisis. Setiap episode memiliki jumlah komentar yang bervariasi, seperti pada episode pertama memiliki 7.807 komentar dan episode sebelas memiliki 3.639 komentar. Hasil analisis sentimen menunjukkan bahwa mayoritas komentar bersifat positif di setiap episode. Sementara itu, pemodelan topik menggunakan LDA yang dievaluasi dengan *Cv Coherence* untuk menghasilkan topik-topik yang relevan. Topik pada episode satu, dua, empat, lima, dan sebelas mencerminkan ekspresi kekaguman, kebanggaan, serta harapan terhadap peserta maupun pihak penyelenggara. Pada episode tiga, enam, sembilan, dan sepuluh, topik yang muncul berkaitan dengan sosok individu tertentu seperti Seong Beom, Agas, Sandy, dan Kadit. Adapun episode tujuh dan delapan menampilkan topik yang unik, di mana perhatian penonton terpusat pada elemen visual, khususnya warna latar belakang peserta yang menimbulkan interpretasi simbolik terhadap hasil pertandingan.

Kata Kunci: Analisis Sentimen, Pemodelan Topik, Ruangguru, *SentiWordNet, Cv Coherence.*

***Lexicon-Based and Latent Dirichlet Allocation on "Clash of Champions" Season* 1**

Abstract

Clash of Champions (CoC) is an educational competition program produced by Ruangguru, one of Indonesia's leading education technology companies. The show has garnered significant public attention, as evidenced by the wide range of viewer responses and reactions expressed in the YouTube comment section. This study aims to examine public sentiment using a Lexicon-Based approach and to uncover key discussion topics through Latent Dirichlet Allocation (LDA) topic modeling. Data were collected via YouTube API extraction throughout the airing period, from June 29 to August 17, 2024. A total of 35,047 comments from all episodes were preprocessed prior to analysis. The number of comments varied by episode for instance, the first episode received 7,807 comments, while the eleventh episode garnered 3,639. Sentiment analysis revealed that most comments across episodes were predominantly positive. Meanwhile, LDA topic modeling, evaluated using the Cv Coherence score, generated thematically relevant topics. Topics in episodes one, two, four, five, and eleven predominantly reflected admiration, pride, and hope directed at the participants and organizers. In contrast, episodes three, six, nine, and ten were centered around individual figures such as Seong Beom, Agas, Sandy, and Kadit. Interestingly, episodes seven and eight surfaced distinctive topics where viewers focused on visual elements particularly the background colors used which sparked symbolic interpretations of the competition outcomes.

***Key Words*:** *Sentiment Analysis*, *Topic Modeling*, *Ruangguru*, *SentiWordNet*, *Cv Coherence*.

PRAKATA

Puji syukur kehadirat Tuhan yang Maha Esa atas rahmat dan karunia-Nya penulis dapat menyelesaikan penulisan Tugas Akhir yang berjudul **"Penerapan *Lexicon-Based* dan *Latent Dirichlet Allocation* Pada "*Clash of Champions*" *Season* 1"**. Adapun penelitian ini dibuat untuk memenuhi syarat penulis untuk memperoleh gelar Sarjana Statistika pada Program Studi Statistika di Fakultas Matematika dan Ilmu Pengetahuan Alam di Universitas Tanjungpura. Penulisan ini dapat dilakukan dan dapat diselesaikan tidak terlepas dari bimbingan, dukungan, saran, dan dukungan dari berbagai pihak. Oleh karena itu, penulis berterima kasih kepada:

1. Kedua orang tua, kakak Perempuan, dan ketiga keponakan yang saya cintai, serta keluarga besar saya yang telah memberikan saya semangat, dorongan, dan motivasi dalam menyelesaikan skripsi ini.
2. Bapak Hendra Perdana, S.Si., M.Sc. selaku Dosen Pembimbing Akademik yang telah banyak membantu dan membimbing saya selama proses perkuliahan.
3. Ibu Nurfitri Imro'ah, M.Si. selaku dosen pembimbing pertama yang telah banyak memberikan bimbingan dan arahan selama penulisan skripsi ini.
4. Ibu Dr Yundari, M.Sc. selaku dosen pembimbing kedua yang telah banyak memberikan bimbingan, arahan, serta masukkan selama penulisan skripsi ini.
5. Tegas Raja Priyatna, Surya Adinda Fitria Pramesti, Anisa Shafarianti, David Jordy Dhandio, Della Zaria, dan Sy Farini Nurhaliza yang selalu membentu penulis selama proses menyelesaikan skripsi ini serta mewarnai kehidupan perkuliahan penulis.
6. Muhammad Silmi Hadi, Irvan Meilandra, Juwan Prioabil Dwi Wicaksono, Ledy Claudia Margareta, Raja Mahendra, Novi, Angela Nadya Putri Ditya, Farrel Pahlevi, dan Novelya Amanda yang selalu memberikan motivasi dan mendukung penulis dalam menyelesaikan skripsi serta memberikan penulis banyak warna selama masa perkuliahan.

7. Teman-teman statistika terutama Angkatan 2021 “Expos” yang telah menemani perjalanan perkuliahan penulis.
8. Semua pihak yang penulis kenal yang memberikan dukungan, motivasi, nasehat, dan doa yang tidak bisa penulis sebutkan satu-satu.

Penulis berharap skripsi ini dapat menjadi bekal untuk kemajuan ilmu pengetahuan. Penulis menyadari bahwa skripsi ini belum sempurna. Oleh karena itu, kritik dan saran yang membangun dapat membantu penulis untuk menjadikan skripsi ini menjadi lebih baik.

Pontianak, 25 Juli 2025

Alberto Cristofer Lauren

DAFTAR ISI

Abstrak ..... **iv**
PRAKATA ..... **vi**
DAFTAR ISI ..... **viii**
DAFTAR GAMBAR ..... **x**
DAFTAR TABEL ..... **xi**
DAFTAR LAMPIRAN ..... **xii**
DAFTAR SIMBOL ..... **xiii**
BAB I PENDAHULUAN **1**
1.1 Latar Belakang 1
1.2 Rumusan Masalah 3
1.3 Tujuan Penelitian 3
1.4 Batasan Masalah 4
1.5 Tinjauan Pustaka 4
1.6 Metodologi Penelitian 6
BAB II LANDASAN TEORI **10**
2.1 Data Mining 10
2.2 Text Mining 10
2.3 Ruangguru 11
2.4 Analisis Sentimen 12
2.5 Pemodelan Topik 13
2.6 *Bag of Words* (BoW) 14
2.7 Probabilitas 16
2.8 Distribusi Multinomial 17
2.9 Distribusi Dirichlet 17
BAB III *LEXICON BASED & LATENT DIRICHLET ALLOCATION* **19**
3.1 *Data Extraction* 19
3.2 *Lexicon Based* 19
3.3 *LibraySentiWordNet* 20
3.4 *Latent Dirichlet Allocation* (LDA) 20

3.5 *Normalized Pointwise Mutual Information* (NPMI) 26
3.6 *Cv Coherence* 30
BAB IV HASIL DAN PEMBAHASAN..... 33
4.1 Pengambilan Data Menggunakan Teknik *Data Extraction*.....33
4.2 *Translate* Data Ulasan Komentar Video *Youtube* Ruangguru 34
4.3 *Cleaning* Data Ulasan Komentar Video *Youtube* Ruangguru.....36
4.4 *Case Folding* Data Ulasan Komentar Video *Youtube* Ruangguru.....36
4.5 *Tokenization* Data Ulasan Komentar Video *Youtube* Ruangguru 37
4.6 *Stopword Removal* Data Ulasan Komentar Video *Youtube* Ruangguru .38
4.7 *Lemmatization* Data Ulasan Komentar Video *Youtube* Ruangguru.....38
4.8 *Reconstruction* Data Ulasan Komentar Video *Youtube* Ruangguru.....39
4.9 Hasil *Preprocessing*.....40
4.10 Pelabelan Sentimen Menggunakan *Library SentiWordNet*.....41
4.11 Interpretasi Hasil Pelabelan Sentimen.....42
4.12 Membuat Representasi BoW.....43
4.13 Membuat Model LDA.....45
4.14 Menghitung *Coherence Score* Tiap Model 45
4.15 Menghitung *Coherence Score* Tiap Topik.....47
4.16 Interpretasi Hasil Pemodelan Topik.....47
BAB V KESIMPULAN DAN SARAN 51
5.1 Kesimpulan.....51
5.2 Saran.....52
DAFTAR PUSTAKA 53
LAMPIRAN..... 58

DAFTAR GAMBAR

Gambar 1. 1 *Flowchart Preprocessing Data* 8
Gambar 1. 2 *Flowchart* Analisis Sentimen 8
Gambar 1. 3 *Flowchart* Pemodelan Topik 9
Gambar 4. 1 Grafik Jumlah Ulasan 40
Gambar 4. 2 Grafik Jumlah Ulasan Sentimen 42

DAFTAR TABEL

Tabel 4.1 *Data Extraction* 34
Tabel 4.2 *Translate Data* 35
Tabel 4.3 *Cleaning Data* 36
Tabel 4.4 *Case Folding* 37
Tabel 4.5 *Tokenization* 37
Tabel 4.6 *Stopwords Removal* 38
Tabel 4.7 *Lemmatization* 39
Tabel 4.8 *Reconstruction* 39
Tabel 4.9 Klasifikasi Kelas Sentimen 41
Tabel 4.10 Representasi BoW pada Ulasan Episode 1 43
Tabel 4.11 *ID* Ulasan Episode 1 44
Tabel 4.12 Representasi BoW Episode 1 44
Tabel 4.13 Jumlah Topik Dan Nilai *Coherence Score* Tiap Episode 45
Tabel 4.14 Hasil Pemodelan Topik 48

DAFTAR LAMPIRAN

Lampiran 1 *Syntax Data Extraction* 58
Lampiran 2 *Syntax Text Preprocessing* 60
Lampiran 3 *Syntax Lexicon Based SentiWordNet* 62
Lampiran 4 *Syntax* LDA dan Cv Coherence 63

DAFTAR SIMBOL

$P(E)$	: Peluang kejadian E
$n(E)$	: Banyaknya kejadian E
n	: Jumlah percobaan
E_i	: Kejadian ke-$i, i = 1, 2, \dots, n$
S	: Ruang sampel
k_i	: Banyak kejadian untuk kategori ke-i
p_i	: Probabilitas kategori ke-i
r	: Jumlah kategori
X	: Vektor probabilitas
a_q	: Parameter konsentrasi untuk komponen ke-$q, q = 1, 2, \dots, m$
m	: Jumlah komponen dalam vektor
$\Gamma(x)$	: Fungsi gamma euler
D	: Seluruh dokumen
M	: Jumlah dokumen
N_d	: Jumlah kata dalam dokumen
θ_d	: Distribusi topik dalam dokumen $d, d = 1, 2, \dots, M$
φ_t	: Distribusi kata dalam topik t
α, β	: Hyperparameter Dirichlet
z_{d_l}	: Topik untuk kata ke-$l, l = 1, 2, \dots, N$ dalam dokumen d
w_{d_l}	: Kata ke-l dalam dokumen d
T	: Himpunan topik
t	: Sebuah topik
w_u, w_v	: Pasangan kata dalam topik
w_u	: Sebuah kata
$p(w_u, w_v)$	: Probabilitas kemunculan bersama kata w_u dan kata w_v

BAB I
PENDAHULUAN

1.1 Latar Belakang

Pendidikan memegang peranan penting dalam membentuk generasi penerus yang berkualitas. Pendidikan yang dibutuhkan adalah pendidikan yang tidak hanya mentransfer pengetahuan, tetapi juga mampu mencetak generasi unggul, yakni individu-individu yang memiliki kemampuan untuk membangun, berinovasi, dan bersaing secara kompetitif di berbagai bidang (Bakiyah, 2022). Di Indonesia, perhatian terhadap sektor pendidikan semakin meningkat seiring dengan tuntutan globalisasi dan pesatnya perkembangan teknologi, Indonesia terus mengupayakan untuk memperkuat digitalisasi di berbagai sektor sebagai langkah untuk meningkatkan daya saing di berbagai bidang, khususnya dalam bidang pendidikan (Yandri, 2023). Upaya untuk memperbaiki kualitas pendidikan terus dilakukan melalui kebijakan pemerintah, seperti inovasi kurikulum, pelatihan tenaga pendidik, hingga pemanfaatan teknologi digital dalam pembelajaran. Meskipun demikian, tantangan dalam sektor pendidikan di Indonesia masih sangat beragam dan sulit untuk diatasi, menurut (Patandung & Panggua, 2022) pendidikan di Indonesia masih menghadapi berbagai permasalahan, yaitu pemerataan akses pendidikan dan kualitas pendidikan, efisiensi dalam pelaksanaan pendidikan, serta relevansi pendidikan dengan kebutuhan masyarakat, selain itu, pendidikan di Indonesia juga dihadapkan pada sejumlah tantangan seperti perkembangan ilmu pengetahuan dan teknologi, pertumbuhan populasi yang cepat, dan kondisi geografis Indonesia yang beragam. Oleh karena itu, diperlukan langkah-langkah yang inovatif dan kreatif serta partisipasi aktif dari semua pihak untuk mewujudkan sistem pendidikan yang adil, dan berkelanjutan.

Kompetisi berbasis pendidikan semakin mendapat perhatian sebagai salah satu strategi untuk mendorong siswa maupun mahasiswa agar berprestasi, berani bersaing, serta mengembangkan potensi dan keterampilan yang dimiliki. Beragam bentuk kompetisi, mulai dari ajang akademik seperti Olimpiade Sains Nasional

hingga kompetisi kreatif seperti debat dan inovasi teknologi, diharapkan dapat meningkatkan motivasi belajar peserta. Kompetisi dalam ranah pendidikan dinilai memiliki berbagai manfaat, antara lain memfasilitasi pengembangan diri, menumbuhkan semangat belajar, serta membangun keterampilan sosial dan mental yang penting dalam menghadapi tantangan masa depan (IELC Creative Team, 2023). Meskipun demikian, perlu disadari bahwa tidak semua kompetisi memperoleh respons positif yang merata dari kalangan pelajar maupun masyarakat. Terdapat sejumlah faktor yang memengaruhi tingkat partisipasi dan minat peserta, seperti aksesibilitas, kesesuaian materi lomba dengan kurikulum yang digunakan, serta bentuk penghargaan yang diberikan kepada peserta.

Industri Pendidikan di Indonesia sudah banyak menyelenggarakan kompetisi berbasis pendidikan yang beragam demi mendukung semangat serta motivasi pelajar Indonesia terhadap pendidikan melalui berbagai macam *platform* seperti media digital dan *platform* daring lainnya. Salah satu kompetisi edukatif yang berhasil mencuri perhatian adalah "*Clash of Champions*" dari Ruangguru, sebuah platform pembelajaran daring ternama di Indonesia. Kompetisi ini menghadirkan 50 mahasiswa terpilih dari berbagai universitas, baik dalam maupun luar negeri, yang lolos melalui proses seleksi ketat. Disajikan dalam format game show, acara ini terdiri dari 11 episode yang ditayangkan secara berkala setiap 6 hingga 7 hari sekali, mulai dari 29 Juni 2024 hingga 17 Agustus 2024. Setiap episodenya menampilkan tantangan berbeda yang menguji kemampuan literasi, numerik, berpikir kritis, komunikasi, kolaborasi, serta pemecahan masalah. Perpaduan antara unsur edukasi dan hiburan inilah yang membuat "*Clash of Champions*" tampil unik dan menarik dibandingkan kompetisi pendidikan lainnya.

Penelitian ini bertujuan untuk menganalisis pandangan penonton terhadap *game show "Clash of Champions"* menggunakan perpaduan antara *Data Mining* dan *Text Mining* terhadap ulasan penonton di kolom komentar *youtube* ruangguru. Data ulasan penonton yang diperoleh dianalisis menggunakan metode analisis sentimen *Lexicon-Based*, dengan tujuan untuk mengetahui respon postitif, negatif, dan netral dari penonton terhadap program ini. Analisis dilakukan dalam rentang waktu 6 hingga 7 hari per episode setelah masa tayang video di media sosial *youtube*

sebanyak 11 episode. Selanjutnya, akan dihitung proporsi sentimen yang didapat dari setiap episode untuk melihat perubahan sentimen penonton.

Perubahan sentimen yang telah terdeteksi dapat dijelaskan alasannya dengan menggunakan analisis *Latent Dirichlet Allocation* (LDA), yang kemudian dilakukan evaluasi model terhadap data dengan menggunakan *Coherence score* untuk melihat tingkat efektivitas model dalam mengklasifikasikan kata. Analisis ini bertujuan untuk mengeksplorasi topik-topik yang muncul pada setiap episode sehingga memungkinkan penelusuran topik-topik yang relevan dengan video dalam ulasan penonton di kolom komentar *youtube*. Melalui kombinasi metode pada penelitian ini, dapat melihat bagaimana reaksi dan tanggapan penonton terhadap lomba berbasis pendidikan, serta dapat mengetahui dimana fokus penonton pada video tersebut.

1.2 Rumusan Masalah

Rumusan masalah dari penelitian ini adalah:

1. Bagaimana cara kerja dan penerapan metode *Lexicon Based* dan *Latent Dirichlet Allocation* serta bagaimana evaluasi kualitas model topik dilakukan menggunakan *Coherence Score* dalam konteks penelitian ini?
2. Bagaimana pandangan penonton terhadap *game show "Clash of Champions"* yang diselenggarakan oleh ruangguru, berdasarkan analisis ulasan penonton pada setiap episode?
3. Topik apa yang paling sering dibahas dalam ulasan penonton terhadap *game show "Clash of Champions"* pada setiap episode, berdasarkan hasil analisis LDA?

1.3 Tujuan Penelitian

Adapun tujuan penelitian ini adalah:

1. Mengkaji metode *Lexicon Based* dan *Latent Dirichlet Allocation* beserta *Coherence Score*.
2. Menganalisis pandangan penonton terhadap *game show "Clash of Champions"* berdasarkan ulasan penonton untuk mengetahui respon dari *game show* tersebut pada setiap episode.

3. Mengeksplorasi topik-topik yang muncul dalam ulasan penonton melalui analisis *Latent Dirichlet Allocation* (LDA) untuk mengidentifikasi topik apa saja yang menjadi perhatian utama dalam ulasan tiap episode.

1.4 Batasan Masalah

Batasan masalah yang digunakan pada penelitian ini adalah:

1. Ulasan penonton yang digunakan pada penelitian ini diperoleh dari kolom komentar *youtube* ruangguru pada setiap episode *"Clash of Champions"* selama periode masa tayang yaitu 29 Juni 2024 hingga 17 Agustus 2024.
2. Nilai parameter untuk α dan β adalah 0,5
3. Nilai ϵ adalah 10^{-12}

1.5 Tinjauan Pustaka

Yogi et al., (2024) melakukan penelitian analisis sentimen dengan menggunakan metode *Lexicon Based* terhadap ulasan pelanggan pada hotel Maestro Pontianak. Ulasan yang digunakan dalam penelitian ini diperoleh melalui *web scraping* dari situs *Traveloka.* Setelah data ulasan terkumpul, dilakukan *pre-processing* terlebih dahulu sebelum analisis lebih lanjut, yang mencakup menerjemahkan semua ulasan ke dalam bahasa inggris, *cleaning data, case folding, tokenizing,* dan *stopword removal.* Penelitian ini juga menggunakan pembobotan kata dengan bantuan *libray* VADER *sentiment* untuk memperjelas hasil sentimen. Berdasarkan analisis sentimen yang dilakukan, *score polarity* ≤ -0.05 menunjukkan sentimen negatif, $-0.05 < coumpound\ score < 0.05$ menunjukkan sentimen netral, dan ≥ 0.05 menunjukkan sentiment positif. Diperoleh tingkat akurasi presentase dari ulasan tersebut adalah sebanyak 65.9% untuk ulasan sentimen positif, 23.61% untuk ulasan sentimen negatif, dan 10.49% untuk ulasan sentimen netral.

Pardede & Waskita, (2023) melakukan penelitian terhadap topik ulasan tentang aplikasi PeduliLindungi menggunakan metode *Latent Dirichlet Allocation* (LDA) dan melakukan evaluasi pada topik yang didapatkan menggunakan metode *topic coherence*. Data yang digunakan berupa ulasan pengguna PeduliLindungi yang diambil dari *Google Play Store* sebanyak 20.000 ulasan. Data yang telah diperoleh

harus diproses terlebih dahulu melalui tahapan *pre-processing* yang meliputi *case folding, filtering, tokenizing, normalization, stemming,* dan penghapusan *stopword.* Setelah tahapa *pre-processing* selesai, langkah berikutnya adalah menerapkan LDA untuk mengidentifikasikan topik yang terdapat di dalam ulasan, yang menghasilkan sebanyak 5 topik. Jumlah topik yang diperoleh digunakan sebagai dasar untuk mengevaluasi topik menggunakan metode evaluasi *topic coherence.* Nilai *coherence score* untuk 5 topik tersebut adalah sebesar 0.3963 dan memiliki nilai rata-rata sebesar 0.3643, nilai ini menunjukkan bahwa 5 topik tersebut menghasilkan distribusi topik yang cukup stabil. Presentase pada topik 1 sebesar 32%, topik 2 sebesar 26.9%, topik 3 sebesar 16.1%, topik 4 sebesar 14.1%, dan topik 5 sebesar 10.9%, hal ini menunjukkan bahwa topik 1 merupakan topik yang paling sering dibahas dan memiliki perhatian atau daya tarik lebih besar dibandingkan dengan topik-topik lainnya dalam ulasan tersebut.

Anggai et al., (2024) melakukan penelitian ekstraksi topik dalam dataset menggunakan teknik pemodelan topik *Latent Dirichlet Allocation* (LDA) terhadap topik-topik utama yang sering dibahas dalam pidato, dan publikasi media terkait Presiden Joko Widodo. LDA digunakan untuk menganalisis, mengidentifikasi, serta memprediksi perubahan topik dari 20 Oktober 2014 hingga 2 Maret 2024, yang didapatkan dari situs resmi *web* Sekretariat Kabinet Presiden Republik Indonesia dengan menggunakan *web scrapping* sebanyak 5988 naskah pidato Presiden. Setelah data didapatkan, perlu dilakukan tahapan *preprocessing text* yang meliputi *tokenization, stopwords removal,* dan *stemming/lemmatization,* dan selanjutnya dibuat *corpus* dalam bentuk *words-vector*, serta membuat *dictionary-term* untuk memetakan kata unik ke dalam numerik ID. Data yang telah diubah akan dilakukan *topic modeling* menggunakan pendekatan LDA dan dilakukan evaluasi topik menggunakan *coherence* dan *perplexity*. Model topik yang digunakan sebanyak 16, karena memiliki nilai *coherence* yang tertinggi yaitu sebesar 0.554 dan menunjukkan konsistensi serta relevansi pada topik yang dihasilkan. Sebaliknya, model dengan topik sebanyak 21 menghasilkan nilai *perplexity* terbaik yaitu sebesar -13.130, yang menunjukkan kemampuan model dalam menangani ketidakpastian dalam data.

1.6 Metodologi Penelitian

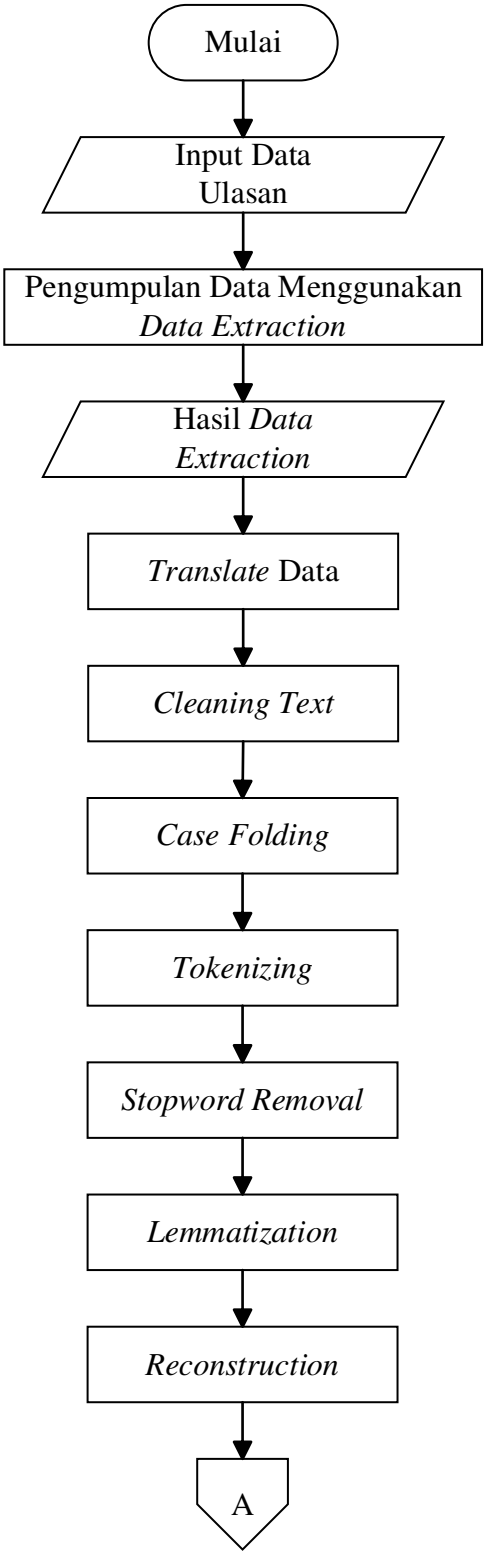
Metodologi yang dilakukan pada penelitian ini merupakan studi literatur dengan mencari teori-teori pendukung mengenai *Lexicon Based* dan *Latent Dirichlet Allocation*. Data yang digunakan pada penelitian ini merupakan data sekunder yang berupa ulasan komentar penonton yang diperoleh dari kolom komentar video *youtube "Clash of Champions" by* Ruangguru. Pengambilan data dilakukan dengan menggunakan teknik *data extraction* melalui *Application Programming Interface* (API) resmi *youtube* dengan bantuan *visual code python* dan bahasa pemrograman *python.*

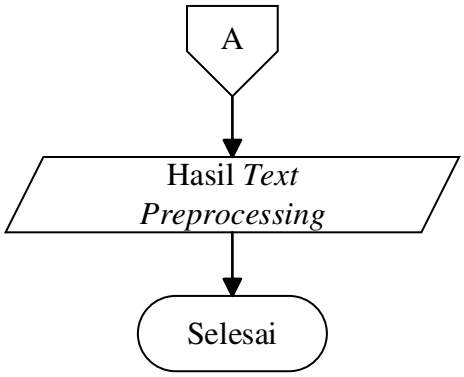
Penelitian ini dilakukan dengan menggunakan beberapa tahapan yang sistematis. Tahapan pertama diawali dengan *data extraction*, yaitu pengumpulan data ulasan dari kolom komentar video *youtube* Ruangguru. Proses ini memerlukan *ID* API pribadi yang diperoleh melalui *Google Cloud Console* setelah pendaftaran akun, lalu digunakan bersama dengan skrip *python* di *Visual Studio Code* untuk mengekstrak ulasan secara otomatis.

Tahap berikutnya adalah *text preprocessing* yang digunakan untuk menyiapkan teks mentah menjadi data yang bersih dan terstruktur. Proses *preprocessing* ini mencakup penerjemahan ulasan ke dalam bahasa Inggris, *cleaning text, case folding, tokenizing, stopword removal, lemmatization,* dan *recronstruction*. Setelah data bersih diperoleh, dilakukan analisis menggunakan metode *Lexicon Based* untuk mengidentifikasi polaritas sentimen dalam ulasan. Sentimen yang dihasilkan diklasifikasikan ke dalam kategori positif, negatif, atau netral, sebagai dasar interpretasi opini publik terhadap tiap episode.

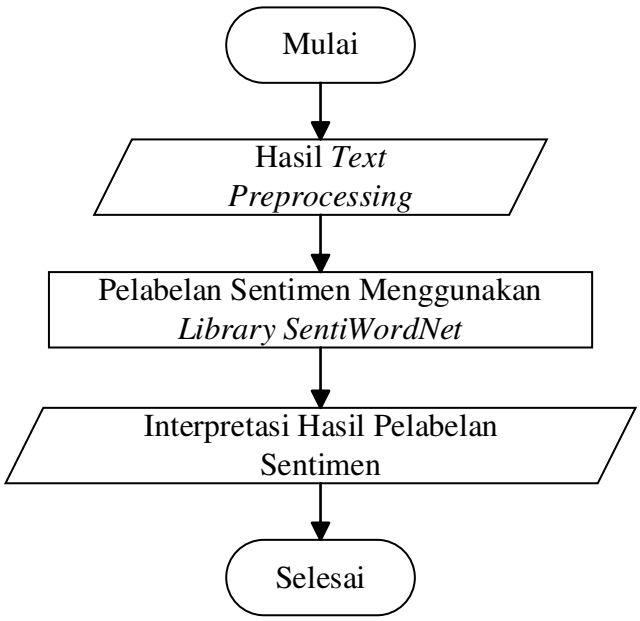
Tahapan selanjutnya adalah pemodelan topik menggunakan metode *Latent Dirichlet Allocation* (LDA) untuk mengidentifikasi tema-tema dominan dalam kumpulan ulasan. Teknik ini membantu dalam memperjelas konteks dari sentimen yang telah dianalisis sebelumnya, serta memberikan gambaran tentang fokus perhatian penonton. Evaluasi terhadap topik dilakukan dengan menggunakan metode *Topic Coherence* menggunakan metrik C_v *Coherence* untuk memastikan bahwa setiap topik yang terbentuk memiliki keterkaitan semantik yang kuat antar kata penyusunnya.

Alur untuk melakukan penelitian ini dapat dilihat pada gambar berikut.

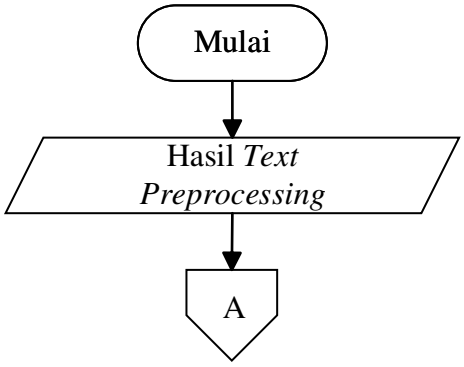


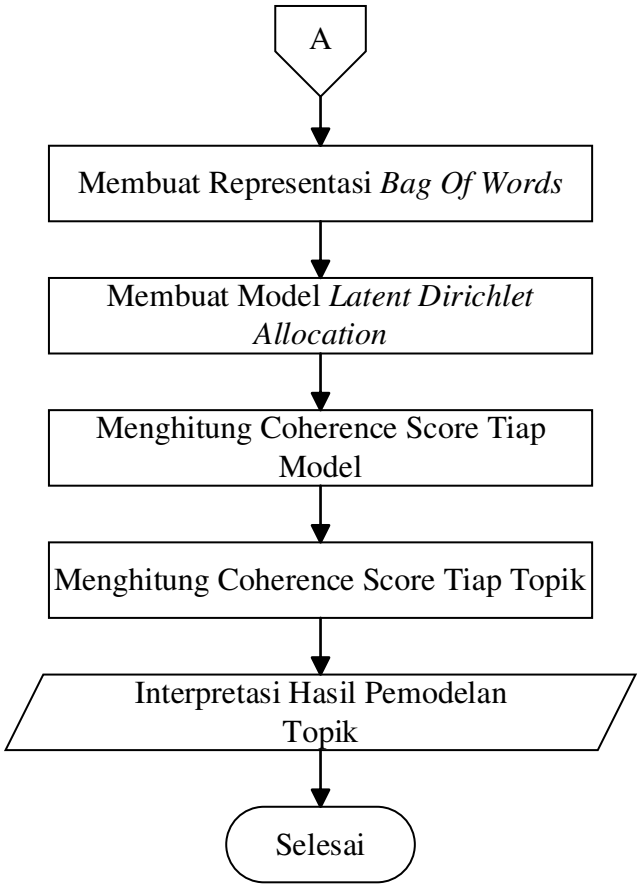


Gambar 1. 1 *Flowchart Preprocessing Data*



Gambar 1. 2 *Flowchart* Analisis Sentimen





Gambar 1. 3 *Flowchart* Pemodelan Topik