IMPLEMENTASI K-MEANS DAN ANALISIS SENTIMEN KRITIK SARAN BERBASIS NLP PADA DATA MONEV BBPSDMP KOMINFO MAKASSAR

SKRIPSI

Diajukan Sebagai Salah Satu Syarat untuk Mendapat Gelar Sarjana Komputer
(S.Kom) Program Studi Informatika

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

SYAHRIL AKBAR
105841103821

PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNIK
UNIVERSITAS MUHAMMADIYAH MAKASSAR
2025

IMPLEMENTASI K-MEANS DAN ANALISIS SENTIMEN KRITIK SARAN BERBASIS NLP PADA DATA MONEV BBPSDMP KOMINFO MAKASSAR

Diajukan Sebagai Salah Satu Syarat untuk Mendapat Gelar Sarjana Komputer
(S.Kom) Program Studi Informatika

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

Disusun dan Diajukan Oleh :

SYAHRIL AKBAR
105841103821

PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNIK
UNIVERSITAS MUHAMMADIYAH MAKASSAR
2025

MAJELIS PENDIDIKAN TINGGI PIMPINAN PUSAT MUHAMMADIYAH
UNIVERSITAS MUHAMMADIYAH MAKASSAR
FAKULTAS TEKNIK

PENGESAHAN

Skripsi atas nama SYAHRIL AKBAR dengan nomor induk Mahasiswa 105841103821, dinyatakan diterima dan disahkan oleh Panitia Ujian Tugas Akhir/Skripsi sesuai dengan Surat Keputusan Dekan Fakultas Teknik Universitas Muhammadiyah Makassar Nomor : 0004/SK-Y/55202/091004/2025, sebagai salah satu syarat guna memperoleh gelar Sarjana Komputer pada Program Studi Informatika Fakultas Teknik Universitas Muhammadiyah Makassar pada hari Sabtu, 30 Agustus 2025.

Panitia Ujian :

Makassar, 17 Rabi'ul Awal 1447 H
10 September 2025 M

1. Pengawas Umum
   a. Rektor Universitas Muhammadiyah Makassar
      Dr. Ir. H. Abd. Rakhim Nanda, ST.,MT.,IPU. : ....
   b. Dekan Fakultas Teknik Universitas Hasanuddin
      Prof. Dr. Eng. Muhammad Isran Ramli, S.T.,M.T.,ASEAN.,Eng. : ....
2. Penguji
   a. Ketua : Dr. Ir. Zahir Zainuddin, M.Sc. : ....
   b. Sekertaris : Lukman, S.Kom., M.T. : ....
3. Anggota : 1. Fahrim Irhamna Rachman, S.Kom., M.T. : ....
   2. Ir. Ida, S.Kom., M.T. : ....
   3. Muhyiddin A.M. Hayat, S.Kom., M.T. : ....

Mengetahui :

Pembimbing I	Pembimbing II
Ir. Muhammad Faisal, S.SI., M.T., Ph.D., IPM.	**Rizki Yusliana Bakti, S.T., M.T.**

Dekan

Ir. Muh. Syafaat S. Kuba, S.T., M.T.
NBM : 795 288

Gedung Menara Iqra Lantai 3
Jl. Sultan Alauddin No. 259 Telp. (0411) 866 972 Fax (0411) 865 588 Makassar 90221
Web: https://teknik.unismuh.ac.id/, e-mail: teknik@unismuh.ac.id

MAJELIS PENDIDIKAN TINGGI PIMPINAN PUSAT MUHAMMADIYAH
UNIVERSITAS MUHAMMADIYAH MAKASSAR
FAKULTAS TEKNIK

UNGGUL

HALAMAN PENGESAHAN

Tugas Akhir ini diajukan untuk memenuhi syarat ujian guna memperoleh gelar Sarjana Komputer (S.Kom) Program Studi Informatika Fakultas Teknik Universitas Muhammadiyah Makassar.

Judul Skripsi : **IMPLEMENTASI K-MEANS DAN ANALISIS SENTIMEN KRITIK SARAN BERBASIS NLP PADA DATA MONEV BBPSDMP KOMINFO MAKASSAR**

Nama : SYAHRIL AKBAR
Stambuk : 105 84 11038 21

Makassar, 10 September 2025

Telah Diperiksa dan Disetujui
Oleh Dosen Pembimbing;

Pembimbing I	Pembimbing II
Ir. Muhammad Faisal, S.SI., M.T., Ph.D., IPM.	**Rizki Yusliana Bakti, S.T., M.T.**

Mengetahui,
Ketua Prodi Informatika

Rizki Yusliana Bakti, S.T., M.T.
NBM : 1307 284

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

MOTTO DAN PERSEMBAHAN

Motto

"Jika ingin mendapatkan performa yang terbaik, jadikan keraguan orang lain sebagai pemicu. Ketika rasa ingin membuktikan itu diubah menjadi motivasi, ia dapat menjadi bahan bakar untuk melampaui batas kemampuanmu."

Persembahan

Tiada lembar skripsi yang paling indah dalam laporan skripsi ini kecuali lembar persembahan. Bismillahirrahmanirrahim, dengan penuh rasa syukur skripsi ini saya persembahkan pertama-tama kepada Allah SWT, atas segala rahmat, hidayah, dan karunia-Nya sehingga karya ini dapat terselesaikan. Selanjutnya, persembahan tulus ini tertuju kepada Ayahanda Kaswan dan Ibunda Reny yang tercinta, sebagai pilar kekuatan atas doa yang tak pernah putus, kasih sayang, serta dukungan moril maupun materiel yang tak ternilai harganya. Persembahan ini juga menjadi pengingat bagi diri sendiri akan sebuah perjuangan dalam membagi waktu antara dunia akademik dan tanggung jawab di jalanan, di mana setiap kilometer di jalanan dan setiap baris kode adalah langkahku menuju wisuda. Tentu, keberhasilan ini tidak terlepas dari bimbingan, kesabaran, dan ilmu tak ternilai dari Dosen Pembimbing I, Bapak Ir. Muhammad Faisal, S.SI., M.T., Ph.D., IPM, dan Dosen Pembimbing II, Ibu Rizki Yusliana Bakti, S.T., M.T.. Secara khusus, terima kasih penulis haturkan untuk seseorang yang istimewa, Rahma Sari, yang kehadirannya menjadi alasan terkuat untuk terus melangkah dan sumber semangat yang tak pernah padam dalam setiap lelah. Tidak lupa, terima kasih untuk sahabat-sahabat seperjuangan: Muh Ulil Amri, Risal, Nur Alam, Sony Achmad Djalil, dan Muhammad Adil Syaputra, serta seluruh rekan Angkatan 2021 khususnya kelas B, atas segala tawa dan kebersamaannya. Akhir kata, karya ini saya persembahkan untuk almamater tercinta, Universitas Muhammadiyah Makassar.

ABSTRAK

SYAHRIL AKBAR, Implementasi K-Means dan Analisis Sentimen Kritik Saran Berbasis NLP pada Data Monev BBPSDMP Kominfo Makassar. (dibimbing oleh Ir. Muhammad Faisal, S.SI., M.T., Ph.D., IPM, dan Rizki Yusliana Bakti, S.T., M.T) .

Balai Besar Pengembangan SDM dan Penelitian (BBPSDMP) Kominfo Makassar menghadapi tantangan dalam menganalisis data kritik dan saran yang bervolume besar dan tidak terstruktur. yang menghambat pengambilan keputusan berbasis data. Penelitian ini bertujuan untuk merancang dan mengimplementasikan sebuah sistem terintegrasi untuk menyaring, mengelompokkan, dan mengklasifikasikan umpan balik secara otomatis. Menggunakan pendekatan *Natural Language Processing* (NLP) hibrida, penelitian ini menerapkan filterisasi data kuantitatif berbasis TF-IDF dan jumlah kata, dilanjutkan dengan pengelompokan topik menggunakan *Latent Semantic Analysis* (LSA) dan *K-Means clustering* yang dioptimalkan dengan *Silhouette Score*, dan diakhiri dengan pembangunan model klasifikasi *Multinomial Naive Bayes* yang dilatih menggunakan label hasil *clustering*. Temuan utama menunjukkan bahwa sistem berhasil menyaring 2.307 data relevan dan mengidentifikasi dua klaster topik yang koheren, yaitu "Kritik dan Harapan" serta "Saran dan Pujian". Model klasifikasi yang dihasilkan mampu mengkategorikan data baru dengan akurasi global mencapai 91%, meskipun menunjukkan sensitivitas yang lebih rendah pada kelas minoritas akibat ketidakseimbangan data. Penelitian ini menyimpulkan bahwa alur kerja yang diusulkan efektif dalam mengubah data mentah menjadi wawasan terstruktur, menghasilkan prototipe sistem yang dapat meningkatkan efisiensi analisis umpan balik untuk peningkatan kualitas layanan publik.

Kata Kunci: Analisis Sentimen, *K-Means Clustering*, *Natural Language Processing* (NLP), Pemodelan Topik, Klasifikasi Teks

ABSTRACT

SYAHRIL AKBAR, *Implementation of K-Means and Sentiment Analysis of Criticism and Suggestions Based on NLP on Monev Data at BBPSDMP Kominfo Makassar. (supervised by Ir. Muhammad Faisal, S.SI., M.T., Ph.D., IPM, and Rizki Yusliana Bakti, S.T., M.T).*

The Center for Human Resources Development and Research (BBPSDMP) of Kominfo Makassar faces a challenge in analyzing large volumes of unstructured criticism and suggestion data, which hinders data-driven decision-making. This research aims to design and implement an integrated system to automatically filter, cluster, and classify feedback. Employing a hybrid Natural Language Processing (NLP) approach, this study implements quantitative data filtering based on TF-IDF scores and word count, followed by topic modeling using Latent Semantic Analysis (LSA) and K-Means clustering optimized with the Silhouette Score, and concludes with the development of a Multinomial Naive Bayes classification model trained using the labels generated from the clustering process. The main findings indicate that the system successfully filtered 2,307 relevant data entries and identified two coherent topic clusters: "Kritik dan Harapan" (Criticism and Hopes) and "Saran dan Pujian" (Suggestions and Praise). The resulting classification model is capable of categorizing new data with a global accuracy reaching 91%, although it exhibits lower sensitivity for the minority class due to data imbalance. This study concludes that the proposed workflow is effective in transforming raw data into structured insights, yielding a system prototype that can enhance the efficiency of feedback analysis for improving public service quality.

Keywords: Sentiment Analysis, K-Means Clustering, Natural Language Processing (NLP), Topic Modeling, Text Classification

KATA PENGANTAR

Syukur Alhamdulillah, segala puji penulis panjatkan kehadirat Allah SWT, atas rahmat dan hidayah-Nya sehingga penulis dapat menyelesaikan karya akademik yang merupakan salah satu persyaratan dalam menempuh Program Studi Informatika di Fakultas Teknik, Universitas Muhammadiyah Makassar. Adapun judul yang diangkat dalam tugas akhir ini adalah "***Implementasi K-Means dan Analisis Sentimen Kritik Saran Berbasis NLP pada Data Monev BBPSDMP Kominfo Makassar***".

Penulis menyadari sepenuhnya bahwa penyusunan skripsi ini masih jauh dari kesempurnaan, baik dari aspek isi, analisis, maupun tata bahasa, yang disebabkan oleh keterbatasan pengetahuan dan pengalaman penulis. Oleh karena itu, dengan segala kerendahan hati, penulis mengharapkan kritik dan saran yang membangun demi perbaikan di masa mendatang.

Terselesaikannya skripsi ini tidak terlepas dari bantuan, bimbingan, dan dukungan dari berbagai pihak. Untuk itu, dengan tulus ikhlas, penulis menghaturkan ucapan terima kasih kepada:

1. Bapak Dr. Ir. H. Abd. Rakhim Nanda, S.T., M.T., IPU, selaku Rektor Universitas Muhammadiyah Makassar.
2. Bapak Muh. Syafaat S. Kuba, ST., M.T, selaku Dekan Fakultas Teknik Universitas Muhammadiyah Makassar.
3. Ibu Rizki Yusliana Bakti, S.T., M.T, selaku Ketua Program Studi Informatika Fakultas Teknik Universitas Muhammadiyah Makassar.
4. Bapak Ir. Muhammad Faisal, S.SI., M.T., Ph.D., IPM, selaku Dosen Pembimbing I yang telah memberikan arahan dan bimbingan dalam penyusunan skripsi ini.
5. Ibu Rizki Yusliana Bakti, S.T., M.T, selaku Dosen Pembimbing II yang turut memberikan masukan berharga.

6. Segenap Dosen dan Staf di lingkungan Fakultas Teknik Universitas Muhammadiyah Makassar atas ilmu dan kemudahan administrasi yang diberikan.
7. Pegawai Tata Usaha Fakultas Teknik Universitas Muhammadiyah Makassar yang telah membantu dalam pengurusan berkas dan administrasi sehingga proses penyusunan skripsi ini berjalan dengan baik.
8. Bapak/Ibu Pegawai BBPSDMP Kominfo Makassar, yang telah memberikan bantuan, izin pengumpulan data, serta informasi yang sangat berarti dalam kelancaran penelitian ini.
9. Kedua orang tua yang tercinta, Ayahanda Kaswan dan Ibunda Reny, atas segala doa, kasih sayang, serta dukungan moral dan material yang tak ternilai harganya.
10. Rahma Sari, mahasiswa Matematika yang telah banyak membantu penulis memahami rumus dan logika perhitungan dalam penyusunan skripsi ini. Terima kasih atas dukungan dan semangatnya.
11. Teman-teman, Angkatan 2021 khususnya kelas B Program Studi Informatika, Fakultas Teknik, Universitas Muhammadiyah Makassar, atas dukungan, doa, dan semangat yang telah diberikan.
12. Kepada semua pihak yang sudah membantu, Penulis mengucapkan banyak terima kasih yang sebesar-besarnya.

Akhir kata, penulis berharap skripsi ini dapat memberikan sumbangsih bagi pengembangan ilmu pengetahuan dan bermanfaat bagi semua pihak yang memerlukan.

Makassar, 26 Agustus 2025

Syahril Akbar

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

DAFTAR ISI

UNIVERSITAS MUHAMMADIYAH MAKASSAR
PERPUSTAKAAN DAN PENERBITAN

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

DAFTAR GAMBAR

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

DAFTAR TABEL

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

DAFTAR LAMPIRAN

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

DAFTAR ISTILAH

Accuracy	Metrik evaluasi yang mengukur proporsi total prediksi yang benar dari keseluruhan data uji.
Baseline Model	Model dasar yang kinerjanya digunakan sebagai titik acuan untuk membandingkan model lain yang lebih kompleks.
Case Folding	Tahap pra-pemrosesan untuk menyeragamkan penggunaan huruf, biasanya dengan mengubah semua karakter menjadi huruf kecil.
Centroid	Titik pusat dari sebuah klaster dalam algoritma K-Means, dihitung sebagai rata-rata dari semua titik data dalam klaster tersebut.
Class Imbalance	Kondisi di mana jumlah sampel untuk setiap kelas dalam dataset tidak terdistribusi secara merata.
Clustering	Teknik unsupervised learning untuk mengelompokkan data ke dalam beberapa grup (klaster) berdasarkan kemiripan karakteristiknya.
Confusion Matrix	Tabel yang memvisualisasikan kinerja model klasifikasi dengan membandingkan kelas aktual dengan kelas yang diprediksi.
Corpus	Kumpulan besar dokumen atau teks yang digunakan untuk analisis linguistik atau melatih model NLP.

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

Dataset	Kumpulan data terstruktur yang digunakan untuk melatih dan menguji model machine learning.
F1-Score	Rata-rata harmonik dari Precision dan Recall, berguna untuk mengevaluasi model pada dataset yang tidak seimbang.
Feature	Atribut atau properti individual yang dapat diukur dari sebuah data, seperti kata atau bobot TF-IDF.
Jarak Euclidean	Metrik jarak garis lurus antara dua titik dalam ruang multidimensi, digunakan dalam K-Means.
Klasifikasi	Tugas dalam supervised learning untuk memprediksi label kategori diskrit (kelas) dari sebuah data masukan.
K-Means	Algoritma clustering yang mempartisi data ke dalam K jumlah klaster berdasarkan kedekatan data dengan centroid.
Latent Semantic Analysis (LSA)	Teknik statistik untuk menganalisis hubungan semantik antara dokumen dan istilah dengan cara mereduksi dimensi.
Machine Learning	Cabang kecerdasan buatan yang memungkinkan sistem untuk belajar dari data tanpa diprogram secara eksplisit.
Monev	Singkatan dari Monitoring dan Evaluasi, proses sistematis untuk memantau dan menilai pelaksanaan program.

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

Multinomial Naive Bayes	Varian algoritma Naive Bayes yang cocok untuk klasifikasi teks berdasarkan frekuensi kemunculan kata.
Natural Language Processing (NLP)	Bidang ilmu yang berfokus pada interaksi antara komputer dengan bahasa manusia, termasuk pemrosesan teks.
Noise	Elemen dalam teks yang tidak relevan atau mengganggu proses analisis, seperti tanda baca atau simbol.
Overfitting	Kondisi di mana model terlalu "menghafal" data latih sehingga kinerjanya buruk pada data baru.
Pendekatan Hibrida	Metode yang mengombinasikan dua atau lebih teknik berbeda untuk mencapai hasil yang lebih baik.
Polaritas	Arah dari sebuah sentimen, yang umumnya dikategorikan menjadi positif, negatif, atau netral.
Precision	Metrik yang mengukur proporsi prediksi positif yang benar dari total prediksi yang diklasifikasikan sebagai positif.
Preprocessing	Serangkaian langkah untuk membersihkan dan mentransformasi data mentah menjadi format yang siap diolah.
Prototipe	Model awal atau implementasi dasar dari sebuah sistem yang dibangun untuk menguji konsep atau alur kerja.

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

Recall	Metrik yang mengukur kemampuan model untuk menemukan kembali semua sampel positif yang relevan dari data.
Reduksi Dimensi	Proses mengurangi jumlah fitur dalam dataset dengan tetap mempertahankan informasi yang paling penting.
Resampling	Teknik statistik untuk menyeimbangkan distribusi kelas dalam dataset, seperti oversampling atau undersampling.
Ruang Semantik	Representasi matematis dari kata atau dokumen di mana jarak antar vektor menunjukkan hubungan makna.
Sentiment Analysis	Proses komputasional untuk mengidentifikasi dan menguantifikasi opini subjektif (positif, negatif, netral) dalam data teks.
Silhouette Score	Metrik untuk mengevaluasi kualitas hasil clustering dengan mengukur kepadatan dan pemisahan antar klaster.
Singular Value Decomposition (SVD)	Teknik dekomposisi matriks yang menjadi dasar matematis dari LSA.
SMOTE	Singkatan dari Synthetic Minority Over-sampling Technique, metode untuk mengatasi ketidakseimbangan kelas.
Sparse Data	Matriks data di mana sebagian besar elemennya bernilai nol, sering dihasilkan oleh representasi TF-IDF.

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

Stemming	Proses mereduksi kata ke bentuk dasarnya (root word) dengan menghilangkan imbuhan.
Stopword	Kata-kata umum yang sering muncul namun memiliki sedikit makna informasional (misalnya: "dan", "di", "yang").
Supervised Learning	Paradigma machine learning di mana model dilatih menggunakan data yang telah diberi label.
TF-IDF	Singkatan dari Term Frequency-Inverse Document Frequency, teknik pembobotan statistik untuk mengukur signifikansi kata.
Tokenizing	Proses memecah rangkaian teks menjadi unit-unit yang lebih kecil yang disebut token (biasanya kata).
Unsupervised Learning	Paradigma machine learning di mana model dilatih menggunakan data yang tidak memiliki label.
Validasi	Proses untuk mengevaluasi dan memastikan bahwa model yang dibangun memiliki kinerja yang andal.
Vektorisasi	Proses mengubah data teks menjadi representasi numerik (vektor) agar dapat diproses oleh algoritma.

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

BAB I
PENDAHULUAN

A. Latar Belakang

Di era digital saat ini, analisis sentimen telah menjadi pendekatan krusial untuk mengekstraksi wawasan bermakna dari data tekstual bervolume besar (Jain et al., 2021). Salah satu sumber data opini publik yang paling kaya berasal dari platform interaktif, termasuk kolom umpan balik pada layanan pemerintah, di mana pengguna secara aktif menyampaikan pandangan mereka terhadap berbagai isu dan program (Daniel Păvăloaia, 2024). Namun, tingginya volume komentar yang bersifat umum dan minim substansi sering kali menjadi tantangan signifikan dalam proses pengambilan keputusan berbasis data.

Konteks tersebut sangat relevan bagi instansi seperti Balai Besar Pengembangan SDM dan Penelitian (BBPSDMP) Kominfo Makassar, yang secara berkala melaksanakan kegiatan Monitoring dan Evaluasi (Monev) untuk menilai kualitas layanannya. Melalui kolom kritik dan saran, para peserta program pelatihan memberikan umpan balik yang menjadi fondasi bagi perbaikan. Akan tetapi, mayoritas tanggapan yang diterima bersifat sangat umum, seperti komentar "bagus", "baik", atau "mantap" yang tidak memberikan masukan konkret. Jumlah entri yang dapat mencapai ribuan, dengan variasi tinggi dari sisi format maupun konten (Jain et al., 2021), menjadikan analisis manual tidak efisien. Selain memakan waktu dan sumber daya yang besar, pendekatan manual juga rentan terhadap bias subjektivitas serta risiko terlewatnya informasi penting (Gandhi et al., 2023).

Berdasarkan tantangan tersebut, diperlukan sebuah sistem otomatis yang mampu menyaring dan mengolah data kritik dan saran secara lebih efektif. Penelitian ini mengusulkan sebuah alur kerja analitis yang komprehensif, dimulai dengan penyaringan data kuantitatif untuk memisahkan komentar yang relevan dari yang tidak relevan berdasarkan kriteria bobot *Term Frequency-Inverse Document Frequency* (TF-IDF) dan jumlah kata. Pendekatan ini memastikan bahwa hanya data berkualitas yang akan diproses lebih lanjut.

Untuk mengelompokkan data yang relevan, penelitian ini menerapkan pendekatan hibrida yang menggabungkan *Latent Semantic Analysis* (LSA) dan *K-Means Clustering*. LSA digunakan terlebih dahulu untuk mereduksi dimensi dan menangkap makna semantik laten di dalam teks, sehingga pengelompokan dengan *K-Means* tidak hanya didasarkan pada kemiripan kata kunci, tetapi juga pada kedekatan makna konseptual (Qureshi & Sabih, 2021). Proses ini, yang dioptimalkan menggunakan *Silhouette Score*, bertujuan untuk menghasilkan klaster-klaster topik yang koheren, seperti "Kritik dan Harapan" serta "Saran dan Pujian".

Rangkaian proses analisis ini diakhiri dengan pemanfaatan hasil *unsupervised learning* (*clustering*) sebagai data latih berlabel untuk membangun model *supervised learning*. Secara spesifik, sebuah model klasifikasi *Multinomial Naive Bayes* dilatih untuk mengotomatisasi proses kategorisasi data kritik dan saran baru. Keseluruhan alur kerja ini, mulai dari text *preprocessing* (*stemming*, *stopword removal*), ekstraksi fitur TF-IDF, penyaringan data, *clustering* hibrida LSA dan *K-Means*, hingga klasifikasi *Naive Bayes* dirancang untuk mengubah data mentah menjadi wawasan terstruktur yang dapat ditindaklanjuti.

Penelitian ini diharapkan mampu menghasilkan sebuah prototipe sistem yang dapat memberikan kontribusi nyata dalam mendukung pengambilan keputusan berbasis data serta memperkuat upaya peningkatan kualitas layanan publik di lingkungan BBPSDMP Kominfo Makassar.

B. Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, rumusan masalah dalam penelitian ini sebagai berikut:

1. Bagaimana menerapkan proses penyaringan data untuk memisahkan komentar yang relevan dan substantif dari komentar yang tidak relevan berdasarkan bobot TF-IDF dan jumlah kata?
2. Bagaimana menerapkan pendekatan hibrida *topic modeling Latent Semantic Analysis (LSA)* dan *K-Means clustering* untuk mengelompokkan data kritik dan saran secara efektif ke dalam kategori-kategori yang bermakna?

3. Bagaimana membangun model klasifikasi *Multinomial Naive Bayes* dengan memanfaatkan data berlabel hasil dari proses *clustering* untuk mengotomatisasi kategorisasi masukan baru secara akurat dan efisien?

C. Tujuan Penelitian

Tujuan Penelitian ini sebagai berikut:

1. Mengimplementasikan metode penyaringan data berbasis TF-IDF dan jumlah kata untuk memperoleh data komentar yang berkualitas.
2. Mengimplementasikan pendekatan hibrida yang menggabungkan *topic modeling Latent Semantic Analysis (LSA)* dan *K-Means clustering* untuk mengelompokkan data kritik dan saran ke dalam klaster yang relevan secara tematik.
3. Membangun sebuah model klasifikasi berbasis *Multinomial Naive Bayes* yang dilatih menggunakan label hasil *clustering* untuk dapat memprediksi kategori topik dari data kritik dan saran baru secara otomatis.

D. Manfaat Penelitian

Penelitian ini diharapkan memberikan manfaat sebagai berikut:

1. Manfaat Akademis
    a. Memberikan kontribusi pada literatur di bidang *Natural Language Processing* (NLP) dan *Machine Learning*, khususnya dalam penerapan teknik hibrida *topic modeling* dengan *Latent Semantic Analysis* (LSA) dan *clustering* dengan *K-Means* pada data teks berbahasa Indonesia.
    b. Menyajikan studi kasus empiris mengenai pemanfaatan hasil *unsupervised learning (clustering)* sebagai data latih berlabel untuk model *supervised learning* (klasifikasi Naive Bayes), yang dapat menjadi referensi metodologis bagi penelitian sejenis.
    c. Menjadi bahan rujukan bagi peneliti lain yang tertarik untuk mengembangkan sistem analisis sentimen dan pengelompokan topik pada data internal lembaga, khususnya dalam konteks evaluasi layanan publik.

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

2. Manfaat Praktis
   a. Menghasilkan sebuah prototipe sistem yang dapat membantu BBPSDMP Kominfo Makassar dalam mengotomatisasi proses analisis data kritik dan saran, sehingga meningkatkan efisiensi dan mengurangi beban kerja manual.
   b. Mempermudah identifikasi masukan yang substantif dan konstruktif dari ribuan komentar melalui metode penyaringan otomatis berbasis TF-IDF, sehingga fokus perbaikan layanan menjadi lebih terarah.
   c. Mendukung pengambilan keputusan berbasis data (*data-driven decision making*) bagi manajemen BBPSDMP Kominfo Makassar dengan menyediakan wawasan terstruktur mengenai area-area yang paling sering menjadi sorotan peserta.
3. Manfaat terhadap Peneliti
   a. Menambah wawasan dan keterampilan praktis dalam mengimplementasikan seluruh alur kerja analisis data teks, mulai dari *preprocessing*, ekstraksi fitur, *clustering*, hingga klasifikasi dan evaluasi model.
   b. Memberikan pengalaman berharga dalam menerapkan pengetahuan teoritis untuk menyelesaikan permasalahan nyata di sebuah institusi.
   c. Memenuhi salah satu syarat kelulusan untuk memperoleh gelar Sarjana Komputer (S.Kom) pada Program Studi Informatika, Fakultas Teknik, Universitas Muhammadiyah Makassar.

E. Ruang Lingkup Penelitian

Untuk menjaga agar penelitian ini tetap terarah dan fokus pada tujuan yang telah ditetapkan, maka perlu dirumuskan batasan-batasan atau ruang lingkupnya. Adapun ruang lingkup dalam penelitian ini adalah sebagai berikut:

1. Sumber Data: Data yang digunakan dalam penelitian ini merupakan data sekunder, yaitu kumpulan teks kritik dan saran yang berasal dari basis data Monitoring dan Evaluasi (Monev) milik BBPSDMP Kominfo Makassar. Data yang dianalisis tidak mencakup sumber eksternal lain seperti media sosial atau

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

platform ulasan publik, sehingga fokus penelitian terbatas pada data internal lembaga tersebut.

2. Pemrosesan Teks (*Preprocessing*): Proses *preprocessing* data teks terbatas pada teknik *case folding*, *stemming* menggunakan pustaka Sastrawi, dan penghapusan *stopword*s bahasa Indonesia. Penelitian ini tidak menangani koreksi ejaan (tipografi) atau normalisasi kata non-standar (*slang*) secara mendalam.
3. Penyaringan dan Ekstraksi Fitur: Penyaringan data untuk memisahkan komentar relevan dan tidak relevan dilakukan berdasarkan kriteria bobot TF-IDF dan jumlah kata. Representasi fitur teks menggunakan model *Term Frequency-Inverse Document Frequency* (TF-IDF) dan tidak membandingkannya dengan metode ekstraksi fitur lain seperti *Word2Vec* atau BERT.
4. Metode Pengelompokan Topik: Pengelompokan topik atau *clustering* menerapkan pendekatan hibrida yang menggabungkan *Latent Semantic Analysis* (LSA) melalui *TruncatedSVD* untuk reduksi dimensi dan algoritma *K-Means*. Jumlah klaster yang dihasilkan dibatasi pada dua kategori utama ("Kritik dan Harapan" serta "Saran dan Pujian") sesuai hasil optimasi menggunakan *Silhouette Score*.
5. Model Klasifikasi: Model klasifikasi yang dibangun untuk prediksi topik secara otomatis adalah *Multinomial Naive Bayes*. Penelitian ini tidak melakukan perbandingan performa dengan algoritma klasifikasi lain seperti *Support Vector Machine* (SVM) atau *Decision Tree*.
6. Luaran Penelitian: Fokus penelitian adalah pada implementasi dan evaluasi prototipe sistem untuk analisis data internal, bukan untuk membangun sebuah aplikasi analisis sentimen yang bersifat umum atau komersial.

F. Sistematika Penulisan

Skripsi ini disusun dalam lima bab dengan struktur sebagai berikut:

BAB I: PENDAHULUAN

Bab ini menyajikan pendahuluan yang mencakup latar belakang, rumusan masalah, tujuan, dan manfaat penelitian. Bagian ini juga menetapkan ruang lingkup dan menguraikan sistematika penulisan skripsi.

BAB II: TINJAUAN PUSTAKA

Bab ini mengulas landasan teori yang mencakup Analisis Sentimen, *Natural Language Processing* (NLP), algoritma *K-Means*, *Latent Semantic Analysis* (LSA), *Multinomial Naive Bayes*, serta metode evaluasi yang relevan. Tinjauan terhadap penelitian terdahulu juga disajikan pada bab ini.

BAB III: METODE PENELITIAN

Bab ini merinci metodologi penelitian secara sistematis, mulai dari teknik pengumpulan data, tahapan *preprocessing*, implementasi *clustering* dengan pendekatan hibrida LSA dan *K-Means*, hingga perancangan dan pengujian model klasifikasi *Multinomial Naive Bayes*.

BAB IV: HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil implementasi dan pembahasan secara mendalam. Cakupannya meliputi hasil penyaringan data, interpretasi klaster yang terbentuk, serta evaluasi performa model klasifikasi yang telah dibangun.

BAB V: PENUTUP

Bab ini merupakan bagian penutup yang merangkum kesimpulan dari seluruh rangkaian penelitian. Berdasarkan hasil yang diperoleh, dirumuskan pula saran untuk pengembangan penelitian selanjutnya.

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

BAB II

TINJAUAN PUSTAKA

A. Landasan Teori

1. Analisis Sentimen

Analisis Sentimen merupakan proses sistematis untuk mengidentifikasi opini subjektif dalam data teks guna menentukan polaritasnya, positif, negatif, atau netral (Jain et al., 2021). Analisis ini menjadi penting dalam memahami persepsi publik terhadap layanan, produk, atau isu tertentu, termasuk dalam konteks umpan balik peserta layanan publik (Dang et al., 2020; Li et al., 2021).

Analisis sentimen secara luas memanfaatkan teknik *machine learning* dan NLP, dimulai dari *text preprocessing* hingga tahap klasifikasi. Terdapat tiga pendekatan umum dalam analisis sentimen, antara lain: berbasis leksikon, berbasis pembelajaran mesin (termasuk *deep learning*), dan pendekatan hibrida (Gandhi et al., 2023).

Penelitian ini menerapkan pendekatan *machine learning* hibrida yang diawali dengan tahap *unsupervised learning* untuk menemukan struktur tematik pada data yang tidak berlabel (Brunton et al., 2020).

Secara spesifik, algoritma *clustering K-Means* digunakan untuk mengelompokkan data kritik dan saran ke dalam beberapa klaster berdasarkan kemiripan kontennya (Ahmed et al., 2020). Hasil dari *clustering* ini, yang merupakan kategori-kategori bermakna, kemudian diinterpretasi dan diberi label secara manual. Data yang telah berlabel tersebut selanjutnya dimanfaatkan sebagai data latih untuk membangun model klasifikasi *supervised learning*. Model *Multinomial Naive Bayes* dipilih untuk melatih sistem agar mampu mengotomatisasi proses kategorisasi masukan baru secara akurat dan efisien (Fakhrezi et al., 2023).

2. *Pre-processing* Text

Tahap *preprocessing* teks memegang peranan fundamental dalam domain analisis tekstual, dengan tujuan utama mentransformasi data teks dalam bentuknya yang asli menjadi format yang lebih terorganisir, bersih, serta siap untuk diolah lebih lanjut menggunakan algoritma komputasional (Khomsah & Sasmito Aribowo, 2020). Signifikansi tahapan ini terletak pada dampaknya terhadap kualitas luaran, di mana hasil *preprocessing* yang optimal berkorelasi positif dengan tingkat efektivitas dan akurasi model analisis data tekstual yang dikembangkan (Zikrina dan Fitriyani, 2025). Khususnya dalam analisis data yang bersumber dari kritik dan saran berbahasa Indonesia, terdapat beberapa teknik *preprocessing* yang lazim diimplementasikan, meliputi:

a. *Case Folding.*

Teknik ini bertujuan untuk menyeragamkan kasus huruf pada keseluruhan teks, lazimnya dengan mengonversi semua karakter menjadi huruf kecil. Upaya ini esensial guna menjamin konsistensi dalam proses analisis, sehingga variasi kapitalisasi pada kata yang identik tidak menyebabkan kata tersebut diperlakukan sebagai entitas leksikal yang berbeda (Adhe et al., 2020).

b. *Text Cleaning.*

Proses ini mencakup eliminasi berbagai karakter atau elemen yang tidak relevan dan berpotensi menjadi derau (*noise*) dalam analisis semantik teks. Elemen-elemen tersebut meliputi tanda baca, digit angka, simbol, serta karakter non-alfabetik lainnya. Tujuan dari pembersihan ini adalah untuk mereduksi kompleksitas data sekaligus memusatkan fokus analisis pada substansi linguistik yang signifikan (Zikrina dan Fitriyani, 2025).

c. *Tokenizing.*

Menyusul tahap pembersihan, teks kemudian melalui proses tokenisasi. Tahapan ini bertujuan untuk melakukan segmentasi terhadap rangkaian teks atau kalimat menjadi unit-unit diskrit yang lebih kecil, yang dikenal sebagai token. Lazimnya, token tersebut merepresentasikan kata-kata individual yang akan menjadi unit dasar dalam analisis selanjutnya (Khomsah & Sasmito Aribowo, 2020).

d. *Stopword Removal.*

Tahapan ini dikonsentrasikan pada identifikasi dan penghilangan kata-kata frekuensi tinggi yang umum dijumpai dalam suatu bahasa, namun memiliki kontribusi informasional yang rendah terhadap pemahaman konten spesifik sebuah dokumen (misalnya, "yang", "dan", "di", "adalah"). Implementasi proses ini mengacu pada daftar *stopword* yang telah disesuaikan secara kontekstual untuk bahasa Indonesia (Adhe et al., 2020).

e. *Word Normalization.*

Proses normalisasi ditujukan untuk mengonversi kosakata yang tidak sesuai dengan kaidah standar termasuk di dalamnya varian ejaan, bentuk singkatan non-formal, atau terminologi *slang* ke dalam bentuk leksikal baku yang merujuk pada kamus standar. Langkah ini memiliki urgensi tinggi, khususnya dalam penanganan data tekstual yang berasal dari interaksi pengguna, seperti komentar pada platform media sosial, di mana prevalensi penggunaan bahasa informal cukup tinggi dan memerlukan standardisasi demi konsistensi analisis (Baskara Nugraha, 2020).

f. *Stemming.*

Stemming merupakan prosedur yang bertujuan untuk mereduksi kata-kata terinfleksi atau terderivasi ke bentuk kata dasar atau akarnya *(root word)*. Hal ini dicapai melalui eliminasi seluruh afiks (imbuhan), yang mencakup prefiks (awalan), sufiks (akhiran), infiks (sisipan), serta kombinasi antar afiks. Dalam konteks morfologi bahasa Indonesia yang kompleks, algoritma *stemming* seperti Nazief-Adriani kerap diaplikasikan untuk tujuan ini (Adhe et al., 2020).

Implementasi cermat dan komprehensif atas keseluruhan tahapan *preprocessing* yang telah dipaparkan akan menghasilkan representasi data tekstual yang optimal. Representasi data dengan kualitas optimal demikian dapat meningkatkan mutu data masukan untuk proses analisis selanjutnya, termasuk pengelompokan (*clustering*) dan analisis sentimen.

3. Ekstraksi Fitur dengan TF-IDF

Setelah melalui tahap *preprocessing*, data teks yang masih bersifat kualitatif perlu ditransformasikan ke dalam format numerik terstruktur agar dapat diolah oleh algoritma *machine learning*. Proses ini dikenal sebagai ekstraksi fitur (*feature extraction*). Salah satu metode yang paling fundamental dan efektif dalam domain analisis teks adalah *Term Frequency-Inverse Document Frequency* (TF-IDF). TF-IDF merupakan sebuah teknik pembobotan statistik yang mampu mengukur signifikansi sebuah kata (*term*) dalam suatu dokumen relatif terhadap keseluruhan kumpulan dokumen (Priambodo & Zuliarso, 2024).

Tujuan utama dari TF-IDF adalah untuk memberikan bobot yang tinggi pada *term-term* yang memiliki nilai informasional penting dan menurunkan bobot dari *term-term* yang umum dijumpai di seluruh dokumen (Khomsah & Sasmito Aribowo, 2020). Perhitungan bobot TF-IDF didasarkan pada perkalian dua komponen utama: *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF).

a. *Term Frequency* (TF)

Term Frequency mengukur frekuensi kemunculan sebuah *term* t dalam sebuah dokumen d. Logikanya, semakin sering sebuah *term* muncul dalam suatu dokumen, semakin penting *term* tersebut untuk merepresentasikan isi dari dokumen tersebut. Rumus matematis untuk TF dapat dinyatakan sebagai berikut (Hanan et al., 2022):

$$\mathrm{TF}(t, d) = \text{Frekuensi kemunculan term } t \text{ dalam dokumen } d \qquad (1)$$

b. Inverse Document Frequency (IDF)

Inverse Document Frequency adalah metrik yang mengukur tingkat kelangkaan atau keunikan sebuah *term* di seluruh korpus. Komponen ini berfungsi untuk mengurangi bobot *term* yang sering muncul di banyak dokumen (misalnya kata "produk" dalam ulasan e-commerce) dan meningkatkan bobot *term* yang spesifik untuk dokumen tertentu. IDF dihitung dengan menggunakan logaritma dari rasio jumlah total dokumen (N) terhadap jumlah dokumen yang mengandung *term* t (df(t)) (Cendikiawan et al., 2023). Rumusnya adalah sebagai berikut:

$$\mathrm{IDF}(t) = \log\left(\frac{N}{\mathrm{df}(t)}\right) \qquad (2)$$

Dengan menggabungkan kedua komponen tersebut, bobot TF-IDF untuk setiap *term* dalam setiap dokumen dihitung menggunakan rumus:

$$\text{TF-IDF}(t,d) = \text{TF}(t,d) \times \text{IDF}(t) \tag{3}$$

Hasil dari proses ini adalah sebuah matriks numerik di mana setiap baris merepresentasikan sebuah dokumen dan setiap kolom merepresentasikan sebuah *term* unik dalam korpus. Nilai dalam setiap sel matriks adalah bobot TF-IDF dari *term* tersebut. Matriks inilah yang kemudian menjadi masukan (input) untuk algoritma *machine learning* selanjutnya, seperti *K-Means clustering* dan klasifikasi *Multinomial Naive Bayes* (Fakhrezi et al., 2023). Penggunaan TF-IDF terbukti efektif dalam berbagai penelitian analisis sentimen karena kemampuannya untuk menonjolkan kata-kata kunci yang paling representatif dari sebuah opini (Zikrina & Fitriyani, 2025).

4. *Natural Language Processing* (NLP)

Natural Language Processing (NLP) adalah cabang ilmu interdisipliner yang menggabungkan ilmu komputer, kecerdasan buatan, dan linguistik, dengan tujuan memungkinkan komputer memahami dan memproses bahasa manusia secara alami (S. Singh & Mahmood, 2021). Perkembangan NLP modern sangat dipengaruhi oleh model bahasa berskala besar (*Large Language Models* / LLMs) yang dilatih melalui *pre-training* pada korpus teks masif, kemudian disesuaikan (*fine-tuned*) untuk tugas spesifik (Bender et al., 2021).

Model berbasis arsitektur *Transformer* seperti BERT, GPT, dan T5 telah menjadi fondasi dari banyak aplikasi NLP mutakhir. Gu et al. (2022) menunjukkan bahwa *domain-specific pre-training* yaitu pelatihan model pada data dengan konteks khusus dapat meningkatkan kinerja dibandingkan model umum. Dalam konteks analisis sentimen, NLP berperan penting dalam tahapan *preprocessing* (normalisasi, tokenisasi, *stemming*), serta dalam transformasi data teks ke bentuk numerik seperti TF-IDF agar dapat diproses oleh algoritma pembelajaran mesin (Dang et al., 2020; Gandhi et al., 2023).

5. *Machine Learning*

Machine Learning (ML) merupakan sekumpulan metode komputasional yang memungkinkan sistem untuk belajar langsung dari data guna mengekstraksi pola serta informasi yang signifikan (Brunton et al., 2020). Sebagai bidang yang terus berkembang, ML berperan penting dalam menyelesaikan berbagai tugas kompleks dengan menekankan pemahaman teoretis terkait cara model melakukan pembelajaran dan generalisasi (Cohen et al., 2021). Pendekatan ini sering kali mengombinasikan data observasional dengan model domain yang sudah ada untuk meningkatkan performa, terutama dalam kondisi data yang tidak sempurna atau terbatas (Karniadakis et al., 2021).

Inti dari kapabilitas ML terletak pada pembentukan model yang mampu melakukan prediksi atau pengambilan keputusan berbasis data, suatu aspek yang sangat krusial di tengah ledakan informasi digital (Brunton et al., 2020; Karniadakis et al., 2021). Secara umum, ML terbagi ke dalam dua paradigma utama, yakni *supervised learning* dan *unsupervised learning*.

a. *Supervised Learning*

Supervised learning adalah pendekatan pembelajaran mesin di mana model dilatih menggunakan data berlabel, artinya setiap data masukan disertai dengan keluaran yang telah diketahui (Brunton et al., 2020). Tujuan utama pendekatan ini adalah untuk mempelajari relasi atau pemetaan antara input dan output, sehingga model dapat melakukan prediksi terhadap data baru secara akurat (Karniadakis et al., 2021). Proses pelatihan melibatkan iterasi penyesuaian parameter untuk meminimalkan kesalahan prediksi terhadap label sebenarnya.

Pendekatan ini umum digunakan dalam klasifikasi sentimen, di mana teks dikategorikan sebagai positif, negatif, atau netral berdasarkan data latih berlabel (Gandhi et al., 2023). Algoritma yang sering digunakan dalam *supervised learning* antara lain: *Regresi Linear*, *Regresi Logistik*, *Support Vector Machine* (SVM), *Decision Tree*, *Random Forest*, dan berbagai arsitektur *Neural Network*.

b. *Unsupervised Learning*

Berbeda dari pendekatan sebelumnya, *unsupervised learning* bekerja pada data yang tidak memiliki label (Brunton et al., 2020). Tujuan utamanya adalah mengidentifikasi struktur atau pola tersembunyi dalam data tanpa panduan eksplisit mengenai output yang diharapkan (Kamiadakis et al., 2021). Algoritma dalam paradigma ini mengelompokkan data berdasarkan kemiripan atribut, atau melakukan reduksi dimensi untuk mempermudah visualisasi dan analisis.

Salah satu metode yang paling banyak digunakan dalam *unsupervised learning* adalah *clustering*. Algoritma *K-Means* merupakan metode *clustering* yang populer karena kesederhanaannya dan efisiensinya dalam menemukan segmentasi alami pada data (Ahmed et al., 2020; Ghazal et al., 2021). Aplikasi *K-Means* mencakup eksplorasi awal data, pengelompokan dokumen, segmentasi pelanggan, serta tahap *preprocessing* dalam pipeline *machine learning* (Moradi Fard et al., 2020).

6. Reduksi Dimensi dengan *Latent Semantic Analysis* (LSA)

LSA adalah sebuah metode *text mining* yang bertujuan untuk mengubah data teks tidak terstruktur menjadi objek data terstruktur dengan cara mengekstrak dan menguraikan faktor-faktor laten kunci yang ada di dalamnya (García-Jurado et al., 2021). Menurut Qureshi & Sabih (2021), LSA mampu mengubah informasi berdimensi tinggi menjadi sebuah "ruang semantik" berdimensi lebih rendah dengan mengidentifikasi sinonimi (kata-kata berbeda dengan makna serupa) dan polisemi (kata yang sama dengan banyak makna). Dengan demikian, LSA tidak hanya mengurangi jumlah fitur, tetapi juga menangkap hubungan konseptual yang lebih dalam antar kata dan dokumen.

Secara teknis, LSA diimplementasikan melalui dekomposisi matriks, khususnya Singular *Value Decomposition* (SVD), yang diterapkan pada matriks TF-IDF. Proses ini menguraikan matriks awal menjadi tiga matriks terpisah yang merepresentasikan hubungan antara *term* dengan konsep laten, pentingnya setiap konsep, dan hubungan antara dokumen dengan konsep laten tersebut (García-Jurado et al., 2021). Dalam penelitian ini, varian SVD yang digunakan adalah *TruncatedSVD*, yang sangat efisien untuk data renggang seperti hasil dari TF-IDF.

Penerapan LSA menghasilkan representasi data yang lebih padat (*dense*) dan kaya secara semantik. Representasi baru ini menjadi masukan yang jauh lebih optimal untuk algoritma *clustering* seperti *K-Means*, karena pengelompokan tidak lagi didasarkan pada kemiripan kata secara harfiah, melainkan pada kedekatan makna konseptual. Hal ini terbukti dapat meningkatkan performa klasifikasi dan membantu menemukan klaster yang lebih koheren dan bermakna (Qureshi & Sabih, 2021; Dash et al., 2023).

7. Algoritma *K-Means*

Algoritma *K-Means* digunakan untuk membagi data ke dalam sejumlah klaster tertentu (K) berdasarkan kedekatan tiap data terhadap titik pusat klaster (*centroid*) (Ahmed et al., 2020; Ghazal et al., 2021). Sebagai algoritma berbasis *centroid*, *K-Means* berupaya meminimalkan varians intra-klaster melalui iterasi langkah berikut (Priambodo dan Zuliarso, 2024):

a. Inisialisasi: Menentukan jumlah klaster (K) secara acak sebagai *centroid* awal.

b. *Assignment*: Menghitung jarak antara setiap titik data dengan setiap *centroid* menggunakan metrik jarak. Dalam penelitian ini, sebagaimana umum digunakan dan disebutkan dalam penelitian Priambodo dan Zuliarso (2024), jarak *Euclidean* digunakan. Rumus jarak *Euclidean* antara dua titik χ dan γ dalam ruang η −dimensi adalah:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{4}$$

c. Setiap titik data kemudian di-*assign* ke klaster dengan *centroid* terdekat.

d. *Update Centroid*: Setelah semua titik data di-*assign* ke klaster, posisi *centroid* untuk setiap klaster dihitung ulang dengan mengambil rata-rata dari semua titik data yang termasuk dalam klaster tersebut. Rumus untuk *update centroid* C adalah (Priambodo & Zuliarso, 2024):

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

$$C_j = \frac{1}{|S_j|} \sum_{x_i \in S_j} x_i \tag{5}$$

Di mana:

1) C_j adalah *centroid* untuk klaster ke-j.
2) $|S_j|$ adalah jumlah data poin (atau anggota) dalam klaster ke-j (yang direpresentasikan oleh himpunan S_j.
3) x_i adalah data poin individual yang termasuk dalam klaster ke-j (anggota dari himpunan S_j).

Langkah 2 dan 3 diulang hingga posisi *centroid* tidak lagi berubah secara signifikan atau hingga jumlah iterasi maksimum tercapai (Ahmed et al., 2020; Ghazal et al., 2021). Meskipun *K-Means* dikenal sederhana dan efisien, algoritma ini menghadapi tantangan seperti kesulitan dalam menentukan jumlah klaster (K) yang optimal dan sensitivitas terhadap inisialisasi *centroid* awal (Moradi Fard et al., 2020; Ran et al., 2021).

Dalam Kendati sederhana, *K-Means* memiliki keterbatasan seperti sensitivitas terhadap inisialisasi awal dan kesulitan dalam menentukan nilai K yang optimal (Moradi Fard et al., 2020; Ran et al., 2021). Namun demikian, *K-Means* tetap menjadi pilihan utama dalam penelitian ini karena efisiensinya dalam mengolah data representasi numerik seperti TF-IDF, serta kompatibilitasnya dengan pipeline *Natural Language Processing* (Priambodo & Zuliarso, 2024).

Selain algoritma berbasis partisi seperti *K-Means*, pendekatan lain yang umum digunakan untuk menemukan struktur tematik dalam data teks adalah model topik probabilistik, di mana *Latent Dirichlet Allocation Latent Semantic Analysis (LSA)* menjadi salah satu metode yang paling fundamental.

8. Model Klasifikasi *Multinomial Naive Bayes*

Setelah data diberi label melalui proses *clustering*, sebuah model klasifikasi *supervised* dapat dilatih untuk mengotomatisasi kategorisasi data baru. Penelitian ini menggunakan algoritma *Multinomial Naive Bayes*, sebuah model klasifikasi probabilistik yang didasarkan pada teorema Bayes dengan asumsi independensi ("naif") antar fitur (Killamsetty et al., 2021). Asumsi ini menyatakan bahwa keberadaan suatu fitur dalam sebuah kelas tidak terkait dengan keberadaan fitur lainnya, yang menyederhanakan komputasi secara signifikan tanpa mengorbankan efektivitas secara drastis.

Dari beberapa varian yang ada, model *Multinomial Naive Bayes* (MNB) secara khusus dirancang untuk menangani fitur-fitur diskrit, seperti data yang merepresentasikan frekuensi kemunculan kata dalam sebuah dokumen (Saura et al., 2023). Hal ini menjadikannya sangat cocok untuk tugas klasifikasi teks, di mana data sering kali direpresentasikan dalam format *Term Frequency* (TF) atau bobot TF-IDF. Model ini bekerja dengan menghitung probabilitas sebuah dokumen masuk ke dalam suatu kategori berdasarkan frekuensi kata-kata yang terkandung di dalamnya.

Penerapan *Multinomial Naive Bayes* telah terbukti efektif dalam berbagai domain. Dalam konteks analisis sentimen pada media sosial, Saura et al. (2023) mengimplementasikan MNB sebagai salah satu *classifier* standar untuk membagi data teks ke dalam kategori positif, negatif, dan netral. Selain itu, *Naive Bayes* juga sering diposisikan sebagai model dasar (*baseline*) yang kuat untuk perbandingan kinerja. Sebagai contoh, V. Singh et al. (2022) menggunakannya sebagai salah satu pembanding untuk mengevaluasi performa model *deep neural network* dalam diagnosis penyakit. Killamsetty et al. (2021) juga merujuk pada *Naive Bayes* sebagai salah satu model klasifikasi fundamental dalam kerangka kerja seleksi subset data.

Berdasarkan landasan tersebut, pemilihan *Multinomial Naive Bayes* dalam penelitian ini didasarkan pada efisiensi komputasionalnya, kemampuannya yang terbukti dalam menangani data teks berbasis frekuensi seperti TF-IDF, serta relevansinya yang kuat sebagai metode standar dalam analisis sentimen dan klasifikasi dokumen.

9. Evaluasi Confusion Matrix

Evaluasi performa model klasifikasi dalam analisis data umumnya menggunakan *Confusion Matrix*, yakni suatu tabel kontingensi yang menggambarkan kinerja model berdasarkan hasil klasifikasi aktual dan prediksi (Dwita et al., 2023). Matriks ini menyediakan informasi mendetail tentang jumlah data yang diklasifikasikan dengan benar maupun yang salah dalam setiap kategori kelas, sehingga memungkinkan analisis terhadap metrik evaluasi penting seperti *accuracy*, *precision*, *recall*, dan *F1-score*.

Secara matematis, *Confusion Matrix* terdiri atas empat komponen utama: *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Berdasarkan komponen tersebut, sejumlah metrik evaluasi dapat dihitung sebagai berikut:

a. *Accuracy* (akurasi) mengukur proporsi prediksi yang benar terhadap keseluruhan data:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

b. *Precision* (presisi) menggambarkan proporsi prediksi positif yang benar dari seluruh prediksi positif:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

c. *Recall* (sensitivitas) mengukur kemampuan model dalam mengidentifikasi seluruh data aktual yang positif:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{8}$$

d. *F1-Score* adalah rata-rata harmonik dari *precision* dan *recall*, berguna saat diperlukan keseimbangan antara keduanya:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

Penggunaan *Confusion Matrix* menjadi penting ketika klasifikasi melibatkan distribusi kelas yang tidak seimbang, sebab nilai akurasi semata tidak cukup merepresentasikan kualitas prediksi model secara menyeluruh (Fakhrezi et al., 2023). Dalam konteks analisis sentimen, misalnya, ketidakseimbangan antara jumlah komentar positif, negatif, dan netral dapat menyebabkan ilusi akurasi tinggi apabila model hanya memprediksi kelas mayoritas.

Penelitian terdahulu oleh Dwita et al. (2023) menunjukkan bahwa evaluasi menggunakan *Confusion Matrix* dapat membedakan efektivitas berbagai algoritma klasifikasi. Pada penelitian tersebut, algoritma C4.5 memperoleh akurasi sebesar 97%, sedangkan *Naive Bayes* mencapai akurasi 95%, dengan nilai *precision* dan *recall* masing-masing dihitung berdasarkan distribusi TP, FP, FN, dan TN pada dataset kinerja pegawai. Demikian pula, Eser dan Şahin (2024) dalam penelitian mengenai analisis sentimen ulasan aplikasi, menggunakan metrik *Confusion Matrix* untuk mengevaluasi model berbasis *transformer* seperti *BERT* dan *RoBERTa*, dengan menekankan bahwa nilai *F1-score* lebih representatif dalam mengukur kinerja model dibanding akurasi semata.

Dengan demikian, evaluasi menggunakan *Confusion Matrix* tidak hanya memberikan gambaran menyeluruh terhadap performa prediktif suatu model, tetapi juga memungkinkan analisis mendalam terhadap kesalahan klasifikasi, terutama pada sistem berbasis teks dan analisis sentimen, seperti yang diimplementasikan dalam penelitian ini.

B. Penelitian Terkait

Penelitian ini merangkum penelitian terdahulu yang relevan, sebagaimana ditampilkan pada Tabel 1.

Tabel 1. Penelitian Terkait

No	Penelitian	Kontribusi
1	(Adhe et al., 2020), Implementasi Text Mining Pengelompokkan	Penelitian ini memberikan kontribusi dalam penerapan algoritma *K-Means* untuk mengelompokkan dokumen skripsi berbahasa

	Dokumen Skripsi Menggunakan Metode K-Means Clustering.	Indonesia. Kontribusi utamanya adalah menunjukkan efektivitas *K-Means* dalam menemukan pola tematik pada data tekstual akademik dan validasi jumlah klaster optimal menggunakan *Silhouette Coefficient*, yang menghasilkan dua klaster dominan.
2	(Fakhrezi et al., 2023), Comparison of Sentiment Analysis Methods Based on Accuracy Value Case Study: Twitter Mentions of Academic Article.	Penelitian ini menyajikan studi komparatif untuk mengevaluasi performa beberapa algoritma klasifikasi, termasuk *Multinomial Naive Bayes*, dalam analisis sentimen pada data Twitter. Kontribusi utamanya adalah pembuktian empiris bahwa Naive Bayes mencapai akurasi tertinggi (95,45%) dibandingkan metode lain, menjadikannya pilihan yang solid untuk klasifikasi sentimen pada teks media sosial.
3	(Priambodo & Zuliarso, 2024), Combination K-Means and LSTM for Social Media Black Campaign Detection of Indonesia Presidential Candidates 2024.	Penelitian ini berkontribusi dengan mengimplementasikan pendekatan hibrida yang menggabungkan *unsupervised learning* (*K-Means*) dan *supervised learning* (LSTM). Kontribusi utamanya adalah pemanfaatan hasil *clustering K-Means* sebagai data latih berlabel untuk mendeteksi kampanye hitam, menunjukkan bahwa pengelompokan awal dapat meningkatkan efektivitas model klasifikasi sekuensial.
4	(Khomsah & Sasmito Aribowo, 2020), Model Text-Preprocessing	Penelitian ini memberikan kontribusi signifikan pada tahapan *preprocessing* untuk teks berbahasa Indonesia yang bersifat

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

	Komentar Youtube Dalam Bahasa Indonesia.	informal (*noise*). Kontribusi utamanya adalah menunjukkan secara empiris bahwa kombinasi *preprocessing* standar, penghapusan *stopwords*, dan konversi kata *slang* ke dalam kosakata baku (KBBI) mampu meningkatkan akurasi model analisis sentimen secara signifikan.
5	(Afryzal Hanan et al., 2025), Sentiment Study of ChatGPT on Twitter Data with Hybrid K-Means and LSTM.	Penelitian ini mengaplikasikan metode hibrida *K-Means* dan LSTM untuk menganalisis sentimen publik terhadap topik teknologi terkini (ChatGPT). Kontribusinya terletak pada penggunaan ekstraksi fitur gabungan (TF-IDF dan *Word2Vec*) sebagai input untuk *clustering*, yang kemudian hasilnya dievaluasi menggunakan model sekuensial, serta menunjukkan efektivitas pendekatan ini dalam menangani data dinamis dari media sosial.

Meskipun penelitian penelitian sebelumnya telah berhasil menerapkan metode seperti *K-Means* untuk pengelompokan dokumen (Adhe et al., 2020) dan membuktikan efektivitas *Multinomial Naive Bayes* untuk klasifikasi sentimen (Fakhrezi et al., 2023), masih terdapat celah penelitian dalam penerapan sebuah alur kerja yang terintegrasi secara penuh, mulai dari penyaringan data mentah hingga pembangunan model klasifikasi otomatis pada data umpan balik institusional yang spesifik. Sebagian besar penelitian cenderung berfokus pada salah satu tahap, baik *clustering* maupun klasifikasi tanpa mengimplementasikan sebuah metode penyaringan data kuantitatif di tahap awal untuk memisahkan masukan yang substantif dari yang tidak relevan.

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

Penelitian ini menawarkan kebaruan dengan mengimplementasikan sebuah alur kerja analitis yang komprehensif dan terstruktur. Kebaruan utama terletak pada penerapan pendekatan hibrida yang menggabungkan beberapa teknik secara berurutan:

1. Proses penyaringan data kuantitatif berbasis bobot TF-IDF dan jumlah kata untuk memastikan hanya data yang relevan dan berkualitas yang diolah lebih lanjut.
2. Penggunaan *Latent Semantic Analysis* (LSA) untuk reduksi dimensi sebelum *clustering*, yang bertujuan menangkap makna semantik laten dan menghasilkan klaster yang lebih koheren secara tematik.
3. Pemanfaatan hasil *unsupervised clustering* (*K-Means*) sebagai data latih berlabel untuk membangun model klasifikasi *supervised* (*Multinomial Naive Bayes*) secara otomatis.

Dengan demikian, penelitian ini tidak hanya menguji setiap algoritma secara terpisah, tetapi juga membangun dan mengevaluasi sebuah prototipe sistem dari hulu ke hilir yang mampu mengubah data kritik dan saran mentah menjadi wawasan terstruktur yang dapat ditindaklanjuti secara efisien.

C. Kerangka Pikir

Kerangka pikir penelitian ini menggambarkan alur logis dalam menjawab permasalahan yang telah dirumuskan, dimulai dari identifikasi masalah, solusi yang diusulkan, hingga tahapan-tahapan yang akan dilakukan untuk mencapai tujuan penelitian. Alur kerja penelitian ini secara visual disajikan pada Gambar 1.

MASALAH

Banyaknya data kritik dan saran pada data Monev BBPSDMP Kominfo Makassar yang bersifat umum dan tidak terstruktur menyulitkan proses penyaringan dan pemanfaatan manual untuk pengambilan keputusan berbasis data.

SOLUSI

Membangun sebuah sistem terintegrasi yang menerapkan pendekatan hibrida: unsupervised learning (LSA dan K-Means) untuk pengelompokan topik secara otomatis, yang hasilnya kemudian digunakan untuk melatih model supervised learning (Multinomial Naive Bayes) untuk klasifikasi data baru.

JUDUL

Implementasi K-Means dan Analisis Sentimen Kritik Saran Berbasis NLP pada Data Monev BBPSDMP Kominfo Makassar.

HASIL

Output sistem adalah sebuah prototipe yang mampu menyaring, mengelompokkan, dan mengklasifikasikan data kritik dan saran secara otomatis. Sistem ini membantu BBPSDMP Kominfo Makassar dalam menganalisis umpan balik secara lebih efisien dan terstruktur untuk pengambilan keputusan yang lebih baik.

Gambar 1. Kerangka Pikir

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

BAB III
METODE PENELITIAN

A. Tempat dan Waktu Penelitian

1. Tempat Penelitian

Penelitian ini dilaksanakan di BBPSDMP Kominfo Makassar yang beralamat di Jl. Prof. Abdurrahman Basalamah II No.25, Karampuang, Kec. Panakkukang, Kota Makassar, Sulawesi Selatan 90231.

2. Waktu Penelitian

Adapun pelaksanaan penelitian ini dilakukan selama bulan Mei – Juli 2025.

B. Alat dan Bahan

1. Kebutuhan Hardware (perangkat keras)
   a. Laptop TUF Gaming FX505DD
2. Kebutuhan Software (perangkat lunak)
   a. Visual Studio Code

      Sebagai *code editor* untuk pengembangan berbagai aplikasi perangkat lunak (Zhang et al., 2023).

   b. Excel

      Perangkat lunak lembar kerja untuk analisis data, perhitungan, dan visualisasi informasi (Zhao et al., 2024).

   c. Python

      Bahasa pemrograman untuk pengembangan aplikasi dan komputasi ilmiah (Ariamajd et al., 2025).

   d. Jupyter Notebook

      Platform interaktif untuk dokumen yang mengintegrasikan kode, visualisasi, dan narasi (Huang et al., 2025).

   e. Pandas

      Pustaka Python untuk manipulasi dan analisis data terstruktur secara efisien (B. P. Singh et al., 2025).

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

f. *Scikit-learn*

Pustaka Python berisi algoritma pembelajaran mesin untuk pemodelan prediktif data (Tran et al., 2025).

g. NLTK dan/atau Sastrawi

Pustaka Python untuk beragam tugas pemrosesan bahasa alami (Hossain et al., 2024).

h. Matplotlib

Pustaka Python untuk pembuatan beragam grafik dan visualisasi data(Goswami et al., 2025).

C. Perancangan Sistem

Perancangan sistem ini menjelaskan rangkaian proses otomatis untuk mengelompokkan dan menganalisis komentar dari hasil survei Monitoring dan Evaluasi. Sistem ini terdiri dari beberapa tahapan utama yang masing-masing memiliki tujuan dan mekanisme spesifik. Keseluruhan alur kerja tersebut divisualisasikan secara rinci melalui *flowchart* pada Gambar 2 untuk memperjelas setiap langkah yang dilakukan.

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

Gambar 2. Flowchart Perancangan Sistem

Berikut adalah uraian setiap proses perancangan sistem berdasarkan *flowchart* diatas:

1. Teknik Pengumpulan Data

   Pengumpulan data merupakan langkah penting dalam penelitian dan dapat dilakukan melalui berbagai metode seperti observasi, wawancara, kuesioner, eksperimen, dan dokumentasi. Teknik yang digunakan bergantung pada jenis data dan tujuan penelitian.

   Dalam penelitian ini, digunakan teknik dokumentasi, yaitu dengan memanfaatkan data sekunder dari hasil survei Monitoring dan Evaluasi (Monev) BBPSDMP Kominfo Makassar. Data berupa komentar pada kolom "Kritik dan Saran" yang di isi oleh peserta pelatihan dari berbagai daerah, dan diperoleh dalam format *spreadsheet* (.xlsx).

   Data dikumpulkan melalui sistem survei digital internal lembaga, digunakan tanpa manipulasi isi, dan dikelola sesuai etika penelitian, termasuk menjaga kerahasiaan responden.

2. Memuat Dataset

   Tahap awal dimulai dengan memuat dataset yang berisi kritik dan saran dari file kritik_saran.xlsx ke dalam program menggunakan library Pandas.

3. *Preprocessing* Teks

   Tahap ini bertujuan untuk membersihkan dan mengubah data teks mentah menjadi format yang terstruktur dan siap diolah. Prosesnya meliputi:

   a. Konversi ke String: Memastikan semua data dalam kolom ulasan berformat string untuk menghindari error saat pemrosesan.

   b. *Stemming*: Mengubah setiap kata ke bentuk dasarnya menggunakan library Sastrawi.

   c. Hapus *Stopwords*: Menghilangkan kata-kata umum yang tidak memiliki makna signifikan (misalnya, "dan", "di", "yang").

   d. Vektorisasi TF-IDF: Mengubah teks menjadi representasi numerik menggunakan metode *Term Frequency-Inverse Document Frequency* (TF-IDF). TF-IDF memberikan bobot pada kata berdasarkan frekuensi dan keunikannya di seluruh kumpulan data.

4. Filter Relevansi Data

Untuk memastikan kualitas analisis, ulasan disaring berdasarkan kriteria relevansi. Data yang relevan adalah yang memiliki skor TF-IDF total di atas 1 dan jumlah kata lebih dari 2. Data yang tidak memenuhi kriteria ini dianggap tidak relevan.

Data disaring dengan menghitung TF-IDF Score dan jumlah kata untuk setiap ulasan. Hasilnya, data dibagi menjadi ulasan relevan dan ulasan tidak relevan.

5. Analisis Klastering (*K-Means*)

Data relevan dikelompokkan ke dalam beberapa klaster atau topik menggunakan algoritma *K-Means*. Tujuannya adalah menemukan kelompok ulasan yang memiliki kemiripan konten.

a. Mencari Parameter Terbaik: Sebelum klastering, dimensi data direduksi menggunakan *TruncatedSVD*. Kombinasi jumlah komponen SVD dan jumlah klaster (K) terbaik dicari menggunakan *Silhouette Score*. Hasil optimal yang didapat adalah 2 klaster dengan 50 komponen SVD.

b. Model *K-Means* diterapkan pada data yang telah direduksi dimensinya untuk membagi ulasan ke dalam 2 klaster.

6. Interpretasi & Pelabelan Klaster

Setiap klaster yang terbentuk diinterpretasikan maknanya dengan menganalisis kata-kata yang paling sering muncul di dalamnya. Berdasarkan interpretasi ini, setiap klaster diberi label.

Setelah menganalisis kata-kata kunci, dua klaster yang terbentuk diberi label "Kritik Dan Harapan" untuk klaster 0 dan "Saran dan Pujian" untuk klaster 1. Data yang telah dilabeli kemudian disimpan ke file hasil_klaster_teks_asli.xlsx.

7. Pelatihan Model Klasifikasi (*Naive Bayes*)

Data yang sudah memiliki label hasil klastering digunakan untuk melatih model klasifikasi. Tujuannya adalah agar model dapat memprediksi topik dari ulasan baru secara otomatis.

   a. Pembagian Data: Dataset dibagi menjadi data latih (80%) dan data uji (20%).
   b. Pelatihan Model: Model *Multinomial Naive Bayes* dilatih menggunakan data latih yang telah divektorisasi dengan TF-IDF. Algoritma ini dipilih karena efektivitasnya dalam klasifikasi teks.

8. Evaluasi Model

   Kinerja model yang telah dilatih dievaluasi menggunakan data uji untuk mengukur seberapa baik kemampuannya dalam mengklasifikasikan topik. Evaluasi dilakukan menggunakan *Classification Report* dan *Confusion Matrix*.

9. Uji Coba & Penyimpanan Model

   Model diuji coba pada beberapa kalimat baru untuk mensimulasikan penggunaannya dalam aplikasi nyata. Selanjutnya, model Naive Bayes dan TF-IDF *Vectorizer* yang telah dilatih disimpan ke dalam file agar dapat digunakan kembali tanpa perlu melatih ulang.

10. Selesai

   Proses berakhir setelah semua tahapan selesai dieksekusi, menghasilkan model yang siap digunakan dan laporan analisis data.

D. Teknik Pengujian Sistem

Pengujian sistem dalam penelitian ini dilakukan untuk mengevaluasi kinerja dan kualitas dari dua komponen utama, yaitu model *clustering K-Means* dan model klasifikasi *Multinomial Naive Bayes*. Pendekatan pengujian berfokus pada analisis kuantitatif dan kualitatif untuk memastikan bahwa sistem mampu mencapai tujuan yang telah ditetapkan.

1. Pengujian Kualitas *Clustering* dengan *Silhouette Score*

   Pengujian ini bertujuan untuk mengukur seberapa baik model *K-Means* dapat mengelompokkan data kritik dan saran yang relevan ke dalam klaster-klaster yang koheren secara tematik.

   a. Parameter Pengukuran: Kualitas klaster diukur menggunakan *Silhouette Score*. Metrik ini mengevaluasi seberapa mirip sebuah data dengan klasternya sendiri dibandingkan dengan klaster lain. Nilai *silhouette score*

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

berkisar antara -1 hingga 1, di mana nilai yang mendekati 1 menunjukkan bahwa data telah dikelompokkan dengan sangat baik.

b. Langkah-langkah Pengujian:

   1) Data ulasan relevan yang telah melalui *preprocessing* dan ekstraksi fitur TF-IDF direduksi dimensinya menggunakan *TruncatedSVD*.
   2) Model *K-Means* diuji dengan beberapa kombinasi parameter: jumlah klaster (K) dari 2 hingga 8 dan jumlah komponen SVD dari 50 hingga 250.
   3) *Silhouette Score* dihitung untuk setiap kombinasi parameter.
   4) Kombinasi parameter yang menghasilkan *silhouette* score tertinggi dipilih sebagai model *clustering* terbaik. Berdasarkan hasil pengujian, kombinasi terbaik adalah 2 klaster dengan 50 komponen SVD.

2. Pengujian Performa Model Klasifikasi dengan *Confusion Matrix*

Pengujian ini bertujuan untuk mengukur akurasi dan efektivitas model *Multinomial Naive Bayes* dalam mengklasifikasikan data kritik dan saran baru berdasarkan label yang dihasilkan dari tahap *clustering*.

a. Parameter Pengukuran: Performa model diukur menggunakan metrik yang diturunkan dari *Confusion Matrix*, yaitu:

   1) *Accuracy*: Mengukur proporsi total prediksi yang benar.
   2) *Precision*: Mengukur tingkat ketepatan prediksi positif.
   3) *Recall*: Mengukur kemampuan model untuk menemukan kembali semua data positif yang sebenarnya.
   4) *F1-Score*: Rata-rata harmonik dari *precision* dan *recall*, memberikan gambaran performa yang seimbang.

b. Langkah-langkah Pengujian:

   1) Dataset yang telah dilabeli ("Kritik Dan Harapan" dan "Saran dan Pujian") dibagi menjadi 80% data latih dan 20% data uji.
   2) Model *Multinomial Naive Bayes* dilatih menggunakan data latih.
   3) Model yang telah dilatih digunakan untuk memprediksi label pada data uji.

4) Hasil prediksi dibandingkan dengan label sebenarnya untuk membangun *Confusion Matrix* dan menghitung metrik performa.

3. Uji Coba Kualitatif pada Data Baru

Pengujian ini bersifat fungsional dan bertujuan untuk memvalidasi kemampuan sistem dalam mengklasifikasikan kalimat baru yang belum pernah dilihat sebelumnya, sebagai simulasi penggunaan di dunia nyata.

a. Cara Pengujian: Beberapa kalimat contoh yang mewakili setiap kategori dimasukkan ke dalam sistem.

b. Langkah-langkah Pengujian:

1) Kalimat baru seperti "materi sangat mudah dipahami" dan "kursi keras dan ruang terlalu panas" diproses oleh *vectorizer* TF-IDF yang telah disimpan.

2) Model *Naive Bayes* yang telah dilatih digunakan untuk memprediksi kategori dari kalimat-kalimat tersebut.

3) Hasil prediksi diamati untuk memastikan kesesuaian dengan makna kalimat. Hasilnya menunjukkan prediksi yang akurat sesuai dengan konten kalimat.

E. Teknik Analisis Data

Analisis data akan dilakukan secara langsung untuk menjawab setiap rumusan masalah penelitian:

1. Analisis Efektivitas Penyaringan Data

Analisis akan dilakukan dengan mengevaluasi hasil dari metode penyaringan data. Analisis kuantitatif akan membandingkan jumlah data sebelum dan sesudah proses filterisasi berdasarkan kriteria bobot TF-IDF dan jumlah kata. Analisis ini bertujuan untuk membuktikan bahwa metode yang diterapkan mampu memisahkan komentar yang relevan dan substantif dari yang tidak relevan, sehingga menghasilkan dataset yang lebih berkualitas untuk tahap analisis selanjutnya.

2. Analisis Hasil Pengelompokan Topik

Analisis hasil *clustering* akan dilakukan melalui dua pendekatan. Pertama, analisis kuantitatif akan menggunakan metrik *Silhouette Score* untuk mengevaluasi dan menentukan jumlah klaster (K) yang paling optimal. Kedua, analisis kualitatif akan dilakukan dengan mengidentifikasi kata-kata kunci yang memiliki bobot tertinggi di setiap klaster yang terbentuk. Interpretasi terhadap kata-kata kunci ini bertujuan untuk memvalidasi bahwa pendekatan hibrida LSA dan *K-Means* mampu mengelompokkan data secara efektif ke dalam kategori yang bermakna, seperti "Kritik dan Harapan" serta "Saran dan Pujian".

3. Analisis Kinerja Model Klasifikasi

kinerja model klasifikasi *Multinomial Naive Bayes* akan dianalisis secara kuantitatif. Analisis akan didasarkan pada *Confusion Matrix* yang dihasilkan dari prediksi model terhadap data uji. Dari matriks tersebut, akan dihitung beberapa metrik evaluasi utama, yaitu *Accuracy*, *Precision*, *Recall*, dan *F1-Score*. Analisis ini bertujuan untuk mengukur tingkat akurasi dan keandalan model dalam mengotomatisasi kategorisasi masukan baru, sekaligus membuktikan efektivitas pemanfaatan data berlabel hasil *clustering* untuk melatih model supervised *learning*.

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

BAB IV
HASIL DAN PEMBAHASAN

A. Deskripsi Data Penelitian

1. Karakteristik Dataset Penelitian

Dataset yang dianalisis dalam penelitian ini merupakan himpunan data tekstual yang diekstraksi dari file kritik_saran.xlsx. Sesuai dengan ruang lingkup yang ditetapkan, data ini mencakup respons dari peserta program *Vocational School Graduate Academy* (VSGA) yang berasal dari berbagai wilayah di Indonesia.

Berdasarkan analisis yang dijalankan oleh program, total data yang berhasil diolah adalah 2.601 *record* individual. Setiap *record* merepresentasikan satu unit umpan balik unik dari peserta. Struktur data mentah yang menjadi fokus analisis disajikan pada Tabel 2.

Tabel 2. Struktur Data Mentah yang Dianalisis

Nama Kolom	Tipe Data	Deskripsi
kritik dan saran	Teks	Konten tekstual asli dari umpan balik yang diberikan peserta.

Variabel sentral dalam analisis ini adalah kolom kritik dan saran, yang berisi data linguistik tidak terstruktur (*unstructured linguistic data*) dalam Bahasa Indonesia. Data ini menunjukkan tingkat variabilitas yang tinggi dan mencakup beberapa tantangan komputasional, di antaranya:

a. Variasi Leksikal: Penggunaan kosakata yang beragam, mulai dari gaya bahasa formal hingga non-formal (bahasa gaul).

b. *Inkonsistensi Ortografis*: Adanya singkatan (contoh: "yg", "tdk"), kesalahan ketik (*typographical errors*), dan penggunaan tanda baca yang tidak standar.

c. Komentar Bernilai Informasional Rendah: Sebagaimana diidentifikasi dalam latar belakang masalah, terdapat proporsi data yang signifikan yang bersifat umum dan hanya merupakan formalitas (contoh: "bagus", "mantap"), sehingga tidak memberikan masukan yang substantif.

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

B. Tahap Pra-pemrosesan dan Filterisasi Data

1. Hasil Pra-pemrosesan Teks

Setiap ulasan dalam dataset mentah melewati serangkaian proses pembersihan dan normalisasi. Proses ini, seperti yang dieksekusi pada cell 3 dari notebook sentiment_analysis.ipynb, meliputi:

a. *Stemming*: Proses ini bertujuan untuk mereduksi setiap kata ke bentuk dasarnya (kata dasar) menggunakan library Sastrawi untuk Bahasa Indonesia. Tujuannya adalah untuk mengonsolidasikan variasi kata yang memiliki akar makna yang sama, sehingga mengurangi kompleksitas kosakata.

b. *Stopword Removal*: Proses ini menghilangkan kata-kata umum yang memiliki frekuensi tinggi namun tidak membawa makna signifikan (misalnya, "yang", "di", "dan"). Hal ini memungkinkan model untuk fokus pada kata-kata yang lebih esensial dalam menentukan topik.

Tabel 3 di bawah ini menyajikan contoh transformasi sebuah ulasan sebelum dan sesudah melalui tahap pra-pemrosesan.

Tabel 3. Hasil Pra-pemrosesan Teks

Tahap	Contoh Teks
Teks Asli	"Cara penyampaian materinya sangat mudah untuk dipahami"
Hasil Akhir	"cara sampai materi sangat mudah paham"

2. Hasil Vektorisasi TF-IDF

Setelah teks dibersihkan, tahap selanjutnya adalah mengubahnya menjadi representasi numerik melalui vektorisasi. Metode yang digunakan adalah TF-IDF (*Term Frequency-Inverse Document Frequency*). TF-IDF memberikan bobot pada setiap kata dalam sebuah ulasan, yang merefleksikan seberapa penting kata tersebut dalam konteks ulasan itu dan keseluruhan dataset.

a. Perhitungan Manual TF-IDF

Perhitungan skor TF-IDF untuk sebuah kata (*term* t) dalam sebuah dokumen (d) adalah hasil perkalian antara *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF).

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t, d)$$

scikit-learn menggunakan formula yang sedikit dimodifikasi untuk meningkatkan akurasi:

1) *Term Frequency* (TF): Frekuensi kata dalam dokumen.
2) *Inverse Document Frequency* (IDF): $\log\left[\frac{n_{\text{docs}}+1}{\text{df}(t)+1}\right] + 1$ di mana n_docs adalah total dokumen dan df(t) adalah jumlah dokumen yang mengandung kata t. Penambahan +1 (*smoothing*) mencegah pembagian dengan nol.
3) Normalisasi L2: Hasil akhir untuk setiap dokumen dinormalisasi agar memiliki panjang vektor 1.

Contoh Perhitungan Berdasarkan Data Aktual: Kita akan menghitung skor TF-IDF untuk kata kunci dari salah satu ulasan aktual dalam dataset.

1) Ulasan (Dokumen) yang Dipilih: "acnya kurang dingin"
2) Hasil Pra-pemrosesan: "acnya kurang dingin"
3) Kata yang Dihitung: "acnya"

Langkah 1: Hitung *Term Frequency* (TF) Kata 'acnya' muncul 1 kali dalam ulasan yang memiliki total 3 kata (setelah pra-pemrosesan).

$$\text{TF('acnya')} = \frac{1}{3} = 0.3333$$

Langkah 2: Hitung *Inverse Document Frequency* (IDF) Dari total 2601 ulasan dalam dataset, kata 'acnya' muncul di 9 ulasan berbeda.

$$\text{IDF('acnya')} = \log\left(\frac{2601+1}{9+1}\right) + 1 = 6.5615$$

Langkah 3: Hitung Skor TF-IDF (Termasuk Normalisasi L2) Skor mentah (TF x IDF) kemudian dinormalisasi bersama dengan semua skor kata lain dalam dokumen yang sama. Skor akhir yang dihasilkan oleh *TfidfVectorizer* untuk kata ini adalah:

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

Skor TF-IDF Final untuk 'acnya': 0.7752

Nilai ini secara akurat merepresentasikan bobot atau pentingnya kata tersebut dalam konteks ulasan spesifik ini dan kelangkaannya di seluruh kumpulan data, yang menjadi dasar untuk filterisasi dan analisis selanjutnya.

3. Hasil Filterisasi Data

Tidak semua ulasan memiliki bobot informatif yang cukup untuk dianalisis. Oleh karena itu, dilakukan filterisasi berdasarkan dua kriteria yang diperoleh dari tahap sebelumnya:

a. Skor Total TF-IDF: Agregasi skor TF-IDF dari seluruh kata dalam ulasan harus lebih besar dari 1.
b. Jumlah Kata: Ulasan harus memiliki lebih dari 2 kata.

Hasil dari proses filterisasi ini, seperti yang ditunjukkan pada output cell 4 di notebook, merangkum pembagian data sebagai berikut:

Tabel 4. Hasil Filterisasi Data Ulasan

Kategori	Jumlah Ulasan	Keterangan	Berkas Output
Data Relevan	2.307	Ulasan yang memenuhi kriteria dan digunakan untuk analisis selanjutnya.	ulasan_relevan.xlsx
Data Tidak Relevan	134	Ulasan yang disaring karena dianggap tidak informatif.	ulasan_tidak_relevan.xlsx
Total Data Awal	2.441	Jumlah keseluruhan ulasan sebelum filterisasi.	kritik_saran.xlsx

Dengan demikian, seluruh tahapan analisis selanjutnya, mulai dari klastering hingga klasifikasi, secara eksklusif menggunakan 2.307 ulasan relevan yang telah teridentifikasi.

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

C. Analisis Klastering untuk Penemuan Topik (Metode *K-Means*)

1. Penentuan Parameter Optimal Menggunakan *Silhouette Score*

Tantangan utama dalam *K-Means* adalah menentukan jumlah klaster (K) yang paling representatif untuk data. Selain itu, karena data TF-IDF bersifat sangat dimensional (*high-dimensional*), diperlukan teknik reduksi dimensi untuk meningkatkan performa klastering. Dalam penelitian ini, digunakan *TruncatedSVD* untuk reduksi dimensi.

Untuk menemukan kombinasi parameter terbaik (jumlah komponen SVD dan jumlah klaster), dilakukan pengujian sistematis seperti yang terlihat pada cell 7 di notebook. Kualitas setiap kombinasi diukur menggunakan *Silhouette Score*, sebuah metrik yang mengevaluasi seberapa baik sebuah objek (ulasan) cocok dengan klasternya sendiri dibandingkan dengan klaster lain. Skor berkisar dari -1 hingga 1, di mana nilai yang lebih tinggi menunjukkan klaster yang lebih padat dan terdefinisi dengan baik.

Hasil pengujian menunjukkan bahwa kombinasi 50 komponen SVD dan 2 klaster (K=2) menghasilkan skor Silhouette tertinggi. Tabel 5 merangkum hasil skor untuk berbagai jumlah klaster pada konfigurasi komponen SVD terbaik.

Tabel 5. Hasil Silhouette Score untuk Penentuan Jumlah Klaster (K)

Jumlah Klaster (K)	Silhouette Score
2	0.1212
3	0.0854
4	0.0571
5	0.0622
6	0.0703
7	0.0793
8	0.0813

Berdasarkan hasil pada Tabel 5, nilai K=2 dipilih sebagai jumlah klaster optimal karena memberikan skor *Silhouette* tertinggi, yang mengindikasikan bahwa data secara alami paling baik terbagi menjadi dua kelompok tematik.

a. Perhitungan Manual *Silhouette Score*

Skor Silhouette untuk satu titik data (i) dihitung dengan rumus:

$$s(i) = \frac{b(i) - a(i)}{\max\big(a(i), b(i)\big)}$$

Di mana:

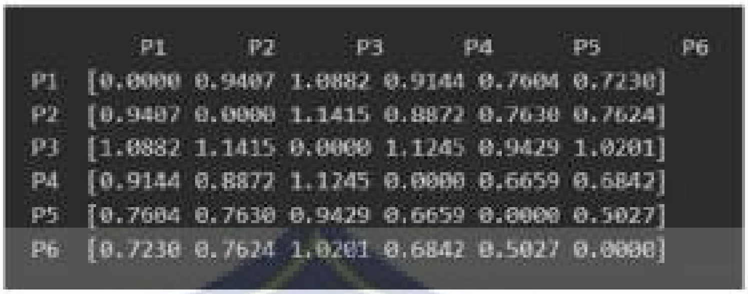
1) a(i): Jarak rata-rata dari titik i ke semua titik lain di dalam klaster yang sama (ukuran kohesi).
2) b(i): Jarak rata-rata dari titik i ke semua titik di klaster terdekat berikutnya (ukuran separasi).

Contoh Perhitungan Berdasarkan Data Aktual: Untuk demonstrasi, kita akan mengambil 3 sampel ulasan dari masing-masing klaster yang dihasilkan oleh model. Kita akan menghitung skor silhouette untuk P1.

Sampel Data:

1) P1: Ulasan "perlu perhati proyektor nya agar tidak buram sehingga mudah ..." (Klaster 0)
2) P2: Ulasan "untuk proyektor mungkin bisa di siap lebih baik lagi agar se..." (Klaster 0)
3) P3: Ulasan "untuk fasilitas latih beri wifi agar mudah dalam dapat mater..." (Klaster 0)
4) P4: Ulasan "acnya kurang dingin..." (Klaster 1)
5) P5: Ulasan "moga latih di luas lagi lingkup di kota parepare..." (Klaster 1)
6) P6: Ulasan "lumayan dinginn dan lumayan bersih..." (Klaster 1)

Matriks Jarak (*Euclidean*) Antar Sampel:



```
        P1      P2      P3      P4      P5      P6
P1  [0.0000 0.9407 1.0882 0.9144 0.7604 0.7230]
P2  [0.9407 0.0000 1.1415 0.8872 0.7630 0.7624]
P3  [1.0882 1.1415 0.0000 1.1245 0.9429 1.0201]
P4  [0.9144 0.8872 1.1245 0.0000 0.6659 0.6842]
P5  [0.7604 0.7630 0.9429 0.6659 0.0000 0.5027]
P6  [0.7230 0.7624 1.0201 0.6842 0.5027 0.0000]
```

Langkah-langkah Perhitungan untuk P1 (dari Klaster 0):

1) Hitung a(P1) (Kohesi): Jarak rata-rata dari P1 ke titik lain di Klaster 0 (P2 dan P3).
   a) Jarak(P1, P2) = 0.9407
   b) Jarak(P1, P3) = 1.0882
   c) $a(P1) = \frac{0.9407+1.0882}{2} = 1.0145$
2) Hitung b(P1) (Separasi): Jarak rata-rata dari P1 ke semua titik di klaster terdekat (Klaster 1), yaitu P4, P5, dan P6.
   a) Jarak(P1, P4) = 0.9144
   b) Jarak(P1, P5) = 0.7604
   c) Jarak(P1, P6) = 0.7230
   d) $b(P1) = \frac{0.9144+0.7604+0.7230}{3} = 0.7993$
3) Hitung Skor Silhouette $s(P1) = \frac{b(P1)-a(P1)}{\max(a(P1),b(P1))}$

$$= \frac{0.7993 - 1.0145}{\max(1.0145,0.7993)}$$

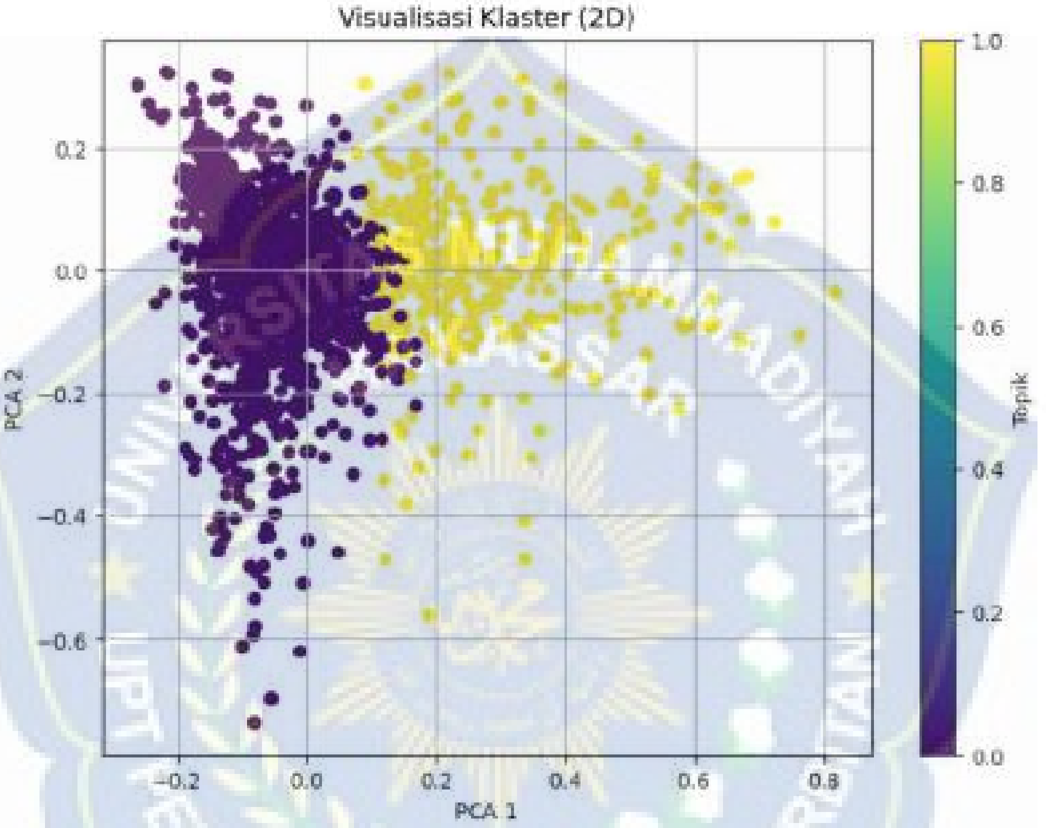
$$= \frac{-0.2152}{1.0145} = -0.2121$$

Skor Silhouette keseluruhan (0.1212) adalah rata-rata dari skor s(i) yang dihitung untuk semua 2307 titik data dalam himpunan data relevan.

UNIVERSITAS MUHAMMADIYAH MAKASSAR
PERPUSTAKAAN DAN PENERBITAN

2. Visualisasi Hasil Klastering

Untuk memvisualisasikan hasil pengelompokan, data TF-IDF yang telah direduksi dimensinya dipetakan ke dalam ruang 2D menggunakan PCA (Principal Component Analysis). Hasilnya disajikan pada Gambar 3.



UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

Gambar 3. Visualisasi Hasil Klastering K-Means

Gambar 3 secara visual mengkonfirmasi hasil kuantitatif. Setiap titik mewakili satu ulasan, diwarnai berdasarkan keanggotaan klasternya. Terlihat dua pengelompokan utama: Klaster 0 (ungu) dan Klaster 1 (kuning). Meskipun ada beberapa tumpang tindih yang wajar untuk data teks yang kompleks, kedua klaster secara umum menempati wilayah yang berbeda pada plot, yang mendukung kesimpulan bahwa terdapat dua tema yang berbeda secara signifikan dalam data.

3. Interpretasi dan Pelabelan Klaster

Makna dari setiap klaster diinterpretasikan dengan menganalisis kata-kata kunci (*top words*) yang paling representatif (memiliki skor TF-IDF rata-rata tertinggi) dan divalidasi dengan memeriksa sampel ulasan aktual dari setiap klaster yang tercatat dalam hasil_klaster_teks_asli.xlsx.

	Klaster 0	**Klaster 1**
Kata Kunci Utama	lebih, baik, kurang, ruang, makan, latih, nya, moga, sangat, fasilitas	materi, ajar, paham, jelas, mudah, sangat, baik, cepat, cara, terlalu
Interpretasi	Berisi umpan balik evaluatif mengenai kondisi, fasilitas, dan harapan perbaikan.	Berisi umpan balik mengenai proses pembelajaran, materi, dan metode pengajaran.
Contoh Ulasan	"acnya kurang dingin", "cahaya ruang kurang", "suara berisik mesin dari belakang kelas"	"ajar jelas dengan sangat baik sehingga materi materi yang ajar mudah paham", "materi yang sampai jelas dan mudah paham"
Label Final	**Kritik dan Harapan**	**Saran dan Pujian**

Analisis gabungan antara kata kunci dan data ulasan aktual memberikan dasar yang kuat untuk pelabelan akhir, yang akan digunakan sebagai kelas target (*target class*) pada tahap klasifikasi selanjutnya.

D. Hasil Klasifikasi Topik (Metode Naive Bayes)

1. Metodologi Klasifikasi

a. Pemilihan Algoritma: *Multinomial Naive Bayes*

Algoritma yang dipilih untuk tugas klasifikasi ini adalah *Multinomial Naive Bayes*. Algoritma ini sangat cocok untuk klasifikasi teks dengan representasi fitur TF-IDF karena beberapa alasan:

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

1) Probabilistik: Model ini menghitung probabilitas sebuah ulasan masuk ke dalam kelas tertentu berdasarkan kata-kata yang dikandungnya.
2) Efisien: Naive Bayes memiliki komputasi yang cepat, baik saat pelatihan maupun prediksi.
3) Asumsi "Naive": Model ini mengasumsikan bahwa setiap kata dalam ulasan bersifat independen (tidak saling memengaruhi), sebuah penyederhanaan yang bekerja dengan sangat baik dalam praktik untuk klasifikasi dokumen.

Secara konseptual, model ini menerapkan Teorema Bayes untuk menghitung probabilitas *posterior* P(Kelas|Ulasan), yaitu probabilitas sebuah ulasan termasuk dalam suatu kelas (misalnya, "Kritik dan Harapan") berdasarkan bukti (kata-kata dalam ulasan tersebut).

b. Protokol Pengujian

Untuk mengevaluasi kinerja model secara objektif, dataset berlabel (2.307 ulasan) dibagi menjadi dua himpunan, seperti yang dieksekusi pada cell 9 di notebook:

1) Himpunan Data Latih (80%): 1.845 ulasan yang digunakan untuk melatih parameter model.
2) Himpunan Data Uji (20%): 462 ulasan yang tidak pernah dilihat oleh model selama pelatihan dan digunakan untuk mengukur kemampuan generalisasinya.

Pembagian ini dilakukan menggunakan *stratified sampling* untuk memastikan proporsi kedua kelas topik tetap terjaga di kedua himpunan data.

2. Hasil Kinerja Model

Kinerja model pada himpunan data uji dievaluasi secara kuantitatif menggunakan *Confusion Matrix* dan metrik-metrik turunannya.

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

a. Analisis *Confusion Matrix*

Confusion Matrix pada Tabel 6 menyajikan dekomposisi hasil prediksi model dibandingkan dengan label aktualnya.

Tabel 6. Confusion Matrix Hasil Klasifikasi

	Prediksi: Kritik & Harapan	**Prediksi: Saran & Pujian**	**Total Aktual**
Aktual: Kritik & Harapan	383 (TP)	2 (FN)	385
Aktual: Saran & Pujian	41 (FP)	36 (TN)	77
Total Prediksi	424	38	462

Keterangan: Untuk tujuan perhitungan, kelas "Kritik & Harapan" dianggap sebagai kelas Positif.

1) *True Positive* (TP): 383 ulasan "Kritik & Harapan" diprediksi dengan benar.
2) *False Negative* (FN): 2 ulasan "Kritik & Harapan" salah diprediksi sebagai "Saran & Pujian".
3) *False Positive* (FP): 41 ulasan "Saran & Pujian" salah diprediksi sebagai "Kritik & Harapan".
4) *True Negative* (TN): 36 ulasan "Saran & Pujian" diprediksi dengan benar.

b. Laporan Metrik Klasifikasi

Metrik kinerja yang lebih detail, yang dihitung dari *Confusion Matrix*, disajikan pada Tabel 7.

Tabel 7. Laporan Metrik Klasifikasi Model

Kelas	**Presisi**	**Recall**	**F1-Score**	**Jumlah (Support)**
Kritik & Harapan	0.90	0.99	0.95	385
Saran & Pujian	0.95	0.47	0.63	77
Akurasi Global			0.91	462

UNIVERSITAS MUHAMMADIYAH MAKASSAR
PERPUSTAKAAN DAN PENERBITAN

3. Perhitungan Manual dan Interpretasi Metrik

Berikut adalah rincian perhitungan manual untuk setiap metrik utama yang diringkas pada Tabel 7.

a. Akurasi (*Accuracy*) Mengukur proporsi total prediksi yang benar di semua kelas.

1) Rumus: $\frac{TP+TN}{TP+TN+FP+FN}$
2) Perhitungan: $\frac{383+36}{383+36+41+2} = \frac{419}{462}$
3) Hasil: 0.9069 atau 90.7% . Ini menunjukkan bahwa model memiliki tingkat keandalan prediksi yang sangat tinggi secara keseluruhan.

b. Presisi (*Precision*) Mengukur tingkat ketepatan dari prediksi positif. Dari semua yang diprediksi sebagai suatu kelas, berapa persen yang benar?

1) Kelas "Kritik & Harapan":
   a) Rumus: $\frac{TP}{TP+FP}$
   b) Perhitungan: $\frac{383}{383+41} = \frac{383}{424} = 0.9033 \quad (90.3\%)$
2) Kelas "Saran & Pujian"
   a) Rumus: $\frac{TN}{TN+FN}$
   b) Perhitungan: $\frac{36}{36+2} = \frac{36}{38} = 0.9473 \quad (94.7\%)$
3) Interpretasi: Presisi yang tinggi di kedua kelas (terutama 'Saran & Pujian") menunjukkan bahwa ketika model membuat prediksi, prediksi tersebut sangat dapat dipercaya.

c. *Recall* (*Sensitivity*) Mengukur tingkat kelengkapan atau sensitivitas model. Dari semua data yang seharusnya masuk ke suatu kelas, berapa persen yang berhasil ditemukan oleh model?

1) Kelas "Kritik & Harapan":
   a) Rumus: $\frac{TP}{TP+FN}$
   b) Perhitungan: $\frac{383}{383+2} = \frac{383}{385} = 0.9948 \quad (99.5\%)$

UNIVERSITAS MUHAMMADIYAH MAKASSAR UPT PERPUSTAKAAN DAN PENERBITAN

2) Kelas "Saran & Pujian":

   a) Rumus: $\frac{\text{TN}}{\text{TN+FP}}$

   b) Perhitungan: $\frac{36}{36+41} = \frac{36}{77} = 0.4675 \quad (46.8\%)$

3) Interpretasi: *Recall* yang sangat tinggi untuk "Kritik & Harapan" menunjukkan model sangat efektif menangkap kelas ini. Sebaliknya, *recall* yang rendah untuk "Saran & Pujian" adalah indikator utama dari dampak *class imbalance*, di mana model kurang sensitif dalam mendeteksi kelas minoritas.

d. *F1-Score* Rata-rata harmonik dari Presisi dan *Recall*, memberikan satu metrik tunggal yang menyeimbangkan kedua aspek tersebut.

1) Kelas "Kritik & Harapan":

   a) Rumus: $2 \times \frac{\text{Presisi} \times \text{Recall}}{\text{Presisi} + \text{Recall}}$

   b) Perhitungan: $\frac{2 \times (0.9033 \times 0.9948)}{0.9033 + 0.9948} = 0.9468 \quad (94.7\%)$

2) Interpretasi: Skor F1 yang sangat tinggi ini mengkonfirmasi performa model yang sangat baik dan seimbang untuk kelas mayoritas. Skor F1 yang lebih rendah untuk kelas minoritas (0.63) secara langsung merefleksikan tarikan ke bawah dari nilai *recall* yang rendah.

E. Uji Coba dan Validasi Model pada Data Baru

1. Metodologi Uji Coba

Untuk melakukan pengujian ini, model *Naive Bayes* (Model_Naive_bayes.pkl) dan *Vectorizer* TF-IDF (Model_TF-IDF.pkl) yang telah dilatih dan disimpan pada tahap sebelumnya dimuat kembali. Proses ini mensimulasikan penerapan model dalam lingkungan produksi, di mana model yang sudah jadi digunakan untuk mengklasifikasikan data baru secara *real-time*.

Beberapa contoh kalimat yang merepresentasikan kedua topik yang telah diidentifikasi diumpankan ke dalam model. Model kemudian melakukan serangkaian langkah: pra-pemrosesan (*stemming*, *stopword removal*), vektorisasi TF-IDF, dan akhirnya prediksi probabilitas untuk setiap kelas.

2. Hasil Uji Coba

Hasil dari pengujian pada dua kalimat baru, seperti yang dieksekusi pada cell 11 di notebook, disajikan pada Tabel 8 di bawah ini.

Tabel 8. Hasil Uji Coba Model pada Kalimat Baru

Kasus Uji	Input Teks	Hasil Prediksi Model	Analisis Validasi
1	"materi sangat mudah dipahami"	Saran dan Pujian	Akurat. Kalimat ini mengandung kata kunci "materi", "mudah", dan "paham" yang memiliki asosiasi kuat dengan kelas "Saran dan Pujian". Model berhasil mengidentifikasi sentimen positif dan fokus pada aspek pembelajaran.
2	"kursi keras dan ruang terlalu panas"	Kritik dan Harapan	Akurat. Kalimat ini mengandung kata-kata seperti "keras" dan "panas" yang jelas merupakan keluhan terhadap fasilitas. Model dengan tepat mengklasifikasikannya ke dalam kelas "Kritik dan Harapan".

3. Implikasi Hasil Uji Coba

Keberhasilan model dalam mengklasifikasikan kedua kasus uji dengan benar menunjukkan beberapa poin penting:

a. Kemampuan Generalisasi: Model tidak mengalami *overfitting* (terlalu menghafal data latih), melainkan telah berhasil mempelajari pola leksikal yang mendasari setiap topik.

b. Utilitas Praktis: Hasil ini mengkonfirmasi bahwa alur kerja yang dibangun—mulai dari pra-pemrosesan hingga prediksi berfungsi secara *end-to-end* dan siap untuk diterapkan pada kasus penggunaan nyata.

Dengan demikian, tahap uji coba ini menjadi validasi akhir yang menegaskan bahwa model klasifikasi yang dihasilkan tidak hanya akurat secara statistik pada data historis, tetapi juga fungsional dan andal untuk aplikasi di masa depan.

F. Pembahasan Hasil Penelitian

1. Analisis Alur Kerja: Dari Data Mentah ke Prediksi Sentimen

Alur kerja penelitian ini dirancang secara sistematis untuk mengubah data ulasan yang tidak terstruktur menjadi wawasan yang dapat ditindaklanjuti. Setiap tahapan memiliki peran krusial dalam validitas hasil akhir.

a. Filterisasi Data: Memfokuskan Analisis pada Ulasan Relevan

Tahap awal penelitian adalah filterisasi data. Berdasarkan analisis pada sentiment_analysis.ipynb, proses ini bertujuan untuk menyaring ulasan yang paling informatif. Kriteria yang digunakan adalah:

1) Skor TF-IDF Total: Ulasan dengan skor di atas 1 dianggap memiliki bobot kata yang signifikan.
2) Jumlah Kata: Ulasan harus memiliki lebih dari 2 kata untuk dianggap substantif.

Proses ini berhasil memisahkan 2.307 ulasan relevan dari total keseluruhan data. Langkah ini sangat penting karena:

1) Meningkatkan Kualitas Data: Dengan membuang ulasan yang tidak informatif (misalnya, "ok", "bagus"), model dapat belajar dari contoh-contoh yang lebih kaya makna.
2) Mendefinisikan Fokus Analisis: Mengarahkan analisis pada ulasan yang benar-benar mengandung kritik atau saran.
3) Efisiensi Komputasi: Mengurangi volume data membuat proses pelatihan model lebih cepat.

Keberhasilan tahap ini menjadi fondasi yang kuat, memastikan bahwa topik yang ditemukan dan model yang dibangun benar-benar relevan dengan masalah yang ingin dipecahkan.

b. Klastering *K-Means*: Penemuan Topik Otomatis

Setelah data difilter, pendekatan *unsupervised learning* dengan *K-Means Clustering* diterapkan pada 2.307 ulasan relevan untuk menemukan struktur atau tema laten. Validasi menggunakan *Silhouette Score* menunjukkan bahwa data secara alami paling baik terbagi menjadi dua klaster (k=2).

Analisis kualitatif terhadap kata-kata dengan bobot TF-IDF tertinggi di setiap klaster sesuai output notebook mengarah pada penamaan yang intuitif:

1) Klaster 0: "Kritik dan Harapan": Berisi kata-kata representatif seperti lebih, baik, kurang, ruang, fasilitas, dan moga (semoga). Ini mengindikasikan ulasan yang fokus pada fasilitas fisik, proses pelatihan, dan harapan untuk perbaikan.
2) Klaster 1: "Saran dan Pujian": Berisi kata-kata representatif seperti materi, ajar, paham, jelas, mudah, dan cepat. Ini jelas merujuk pada ulasan tentang kualitas penyampaian materi, metode mengajar, dan kemudahan pemahaman.

Keberhasilan ini sangat krusial karena memungkinkan pelabelan data secara otomatis, yang mengubah masalah dari *unsupervised* menjadi *supervised* dan menjadi fondasi untuk tahap klasifikasi.

2. Evaluasi Kinerja Model Klasifikasi Naive Bayes

Dengan 2.307 data yang telah dilabeli, data dibagi menjadi 80% untuk pelatihan (1.845 ulasan) dan 20% untuk pengujian (462 ulasan). Model *Multinomial Naive Bayes* kemudian dilatih dan dievaluasi.

a. Analisis *Confusion Matrix*

Confusion Matrix dari sentiment_analysis.ipynb memberikan gambaran rinci tentang 462 prediksi pada data uji.

Tabel 9. Hasil Klasifikasi Confusion Matrix

	Prediksi: Kritik & Harapan	**Prediksi: Saran & Pujian**	**Total Aktual**
Aktual: Kritik & Harapan	TP = 383	FN = 2	385
Aktual: Saran & Pujian	FP = 41	TN = 36	77
Total Prediksi	424	38	462

1) *True Positive* (TP): 383 ulasan "Kritik & Harapan" diprediksi dengan benar.
2) *False Negative* (FN): Hanya 2 ulasan "Kritik & Harapan" yang salah diprediksi.
3) *False Positive* (FP): 41 ulasan "Saran & Pujian" salah diprediksi sebagai "Kritik & Harapan".
4) *True Negative* (TN): 36 ulasan "Saran & Pujian" diprediksi dengan benar.

Matriks ini menunjukkan model sangat andal dalam mengenali "Kritik & Harapan", namun cenderung salah mengklasifikasikan "Saran & Pujian" sebagai "Kritik & Harapan".

b. Analisis Metrik Kinerja Agregat

Berdasarkan *classification report* di notebook, metrik kinerja utama disajikan pada tabel berikut.

Tabel 10. Perbandingan Kinerja Model per Kelas

Metrik	**Kelas Mayoritas ("Kritik & Harapan")**	**Kelas Minoritas ("Saran & Pujian")**	**Akurasi Global**
Recall	0.99 (99%)	0.47 (47%)	91%
Precision	0.90 (90%)	0.95 (95%)	
F1-Score	0.95	0.63	

1) Akurasi Global (91%): Menunjukkan bahwa 91% dari total prediksi pada data uji adalah benar.

2) *Recall*: Model sangat sensitif dalam mendeteksi kelas mayoritas (99%), namun hanya mampu mendeteksi 47% dari kelas minoritas. Ini berarti model berhasil menemukan hampir semua "Kritik & Harapan", tetapi melewatkan lebih dari separuh "Saran & Pujian".

3) *Precision*: Tingkat kepercayaan prediksi sangat tinggi untuk kedua kelas, terutama untuk kelas minoritas (95%).

3. Interpretasi Hasil dan Implikasi

a. Mengupas Fenomena Ketidakseimbangan Kelas (*Class Imbalance*)

Penyebab utama perbedaan nilai *recall* adalah ketidakseimbangan kelas. Data latih didominasi oleh kelas "Kritik dan Harapan" (1.540 sampel) dibandingkan "Saran dan Pujian" (305 sampel), dengan rasio sekitar 5:1. Akibatnya, model lebih banyak "belajar" tentang ciri-ciri kelas mayoritas, membuatnya sangat ahli dalam mengidentifikasi kelas tersebut (*Recall* 99%). Namun, ini juga membuatnya cenderung lebih sering memilih kelas mayoritas ketika menghadapi kasus yang ambigu, yang menyebabkan banyak ulasan "Saran & Pujian" salah diklasifikasikan (FP = 41).

b. Implikasi Praktis: Keandalan Prediksi vs. Sensitivitas Deteksi

Meskipun *recall* untuk "Saran dan Pujian" rendah, presisi yang sangat tinggi (95%) memiliki implikasi praktis yang kuat. Ini berarti, meskipun sistem mungkin tidak menangkap semua ulasan saran/pujian, setiap ulasan yang berhasil ditandai sebagai "Saran dan Pujian" memiliki kemungkinan 95% benar. Dalam konteks bisnis, ini sangat berharga: tim dapat langsung menindaklanjuti umpan balik ini dengan keyakinan tinggi tanpa perlu verifikasi manual.

BAB V
KESIMPULAN DAN SARAN

A. Kesimpulan

Berdasarkan analisis dan pembahasan yang telah diuraikan, ditarik kesimpulan yang menjawab secara tuntas ketiga rumusan masalah penelitian sebagai berikut:

1. Mengenai Penerapan Proses Penyaringan Data: Penelitian ini berhasil menerapkan proses penyaringan data untuk memisahkan komentar yang relevan dari yang tidak relevan. Dengan menggunakan kriteria gabungan antara bobot TF-IDF (>1) dan jumlah kata (>2), sistem mampu menyaring 2.307 ulasan substantif dari total data awal. Metode ini terbukti efektif dalam meningkatkan kualitas dataset secara signifikan dengan mengeliminasi ulasan yang tidak informatif, sehingga analisis selanjutnya menjadi lebih fokus dan akurat.
2. Mengenai Penerapan Pendekatan Hibrida untuk Pengelompokan Topik: Pendekatan hibrida yang menggabungkan *Latent Semantic Analysis* (LSA) untuk reduksi dimensi dan *K-Means Clustering* berhasil mengelompokkan data kritik dan saran secara efektif. Validasi dengan *Silhouette Score* mengonfirmasi bahwa jumlah klaster optimal adalah dua. Analisis kualitatif terhadap kata-kata kunci menghasilkan dua kategori yang bermakna: "Kritik dan Harapan" (terkait fasilitas, ruang, dan harapan perbaikan) dan "Saran dan Pujian" (terkait materi, metode ajar, dan pemahaman). Ini membuktikan bahwa pendekatan hibrida mampu menemukan struktur tematik yang koheren dari data teks tidak terstruktur.
3. Mengenai Pembangunan dan Kinerja Model Klasifikasi: Sebuah model klasifikasi *Multinomial Naive Bayes* berhasil dibangun dengan memanfaatkan data berlabel hasil dari proses *clustering*. Model ini mampu mengotomatisasi kategorisasi masukan baru dengan tingkat akurasi global mencapai 91% pada data uji. Meskipun dipengaruhi oleh ketidakseimbangan kelas (*recall* 47% untuk kelas minoritas), nilai presisi yang tinggi (95% untuk kelas minoritas) menunjukkan bahwa prediksi yang dibuat sangat dapat diandalkan. Dengan

demikian, pemanfaatan hasil *clustering* sebagai data latih terbukti menjadi strategi yang efisien dan efektif untuk membangun model klasifikasi yang akurat.

B. Keterbatasan Penelitian

Penelitian yang telah dilakukan memiliki beberapa keterbatasan yang perlu diakui untuk memberikan konteks pada temuan yang ada:

1. Ketidakseimbangan Kelas: Kinerja model, khususnya metrik *recall* untuk kelas "Saran dan Pujian" (47%), secara signifikan dipengaruhi oleh distribusi data yang tidak seimbang (rasio sekitar 5:1).
2. Representasi Fitur: Analisis terbatas pada fitur *bag-of-words* dari TF-IDF dan belum mempertimbangkan informasi kontekstual atau urutan kata yang lebih kompleks yang dapat ditangkap oleh model bahasa yang lebih canggih.
3. Generalisasi Model: Model yang dibangun spesifik untuk dataset internal BBPSDMP Kominfo Makassar. Kinerjanya mungkin tidak dapat digeneralisasi secara langsung ke domain umpan balik lain tanpa adanya *fine-tuning* atau pelatihan ulang.

C. Saran

Berdasarkan kesimpulan dan keterbatasan penelitian, berikut adalah beberapa saran strategis untuk pengembangan di masa mendatang:

1. Penanganan *Class Imbalance*: Sangat disarankan untuk menerapkan teknik resampling pada penelitian selanjutnya. Metode *oversampling* seperti SMOTE *(Synthetic Minority Over-sampling Technique)* dapat digunakan untuk membuat sampel sintetis bagi kelas minoritas, yang berpotensi besar meningkatkan sensitivitas (*recall*) model terhadap topik "Saran dan Pujian".
2. Eksplorasi Model Alternatif: Melakukan studi komparatif dengan arsitektur model yang lebih canggih, seperti LSTM (*Long Short-Term Memory*) atau model berbasis *Transformer* (misalnya, IndoBERT). Model-model ini mampu memahami konteks sekuensial dan semantik yang lebih dalam, sehingga berpotensi memberikan peningkatan kinerja klasifikasi.

3. Analisis Sentimen Granular: Mengembangkan sistem lebih lanjut untuk tidak hanya mengklasifikasikan topik, tetapi juga menganalisis polaritas sentimen (positif, negatif, netral) di dalam setiap topik. Ini akan memberikan wawasan yang lebih kaya, misalnya, mampu secara otomatis membedakan antara "kritik" (negatif) dan "harapan" (netral/positif) di dalam klaster pertama.

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

DAFTAR PUSTAKA

Adhe, D., Rachman, C., Goejantoro, R., Deny, F., & Amijaya, T. (2020). Implementasi Text Mining Pengelompokkan Dokumen Skripsi Menggunakan Metode K-Means Clustering. *Jurnal EKSPONENSIAL*, *11*(2).

Afryzal Hanan, D., Yudo Husodo, A., & Pasca Rassy, R. (2025). *Sentiment Study of ChatGPT on Twitter Data with Hybrid K-Means and LSTM*. *24, No. 2*. https://doi.org/10.30812/matrik.v24i2.4791

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. In *Electronics (Switzerland)* (Vol. 9, Issue 8, pp. 1–12). MDPI AG. https://doi.org/10.3390/electronics9081295

Ariamajd, A., de Castro, R. L.-R., & Volkamer, A. (2025). *PyPackIT: Automated Research Software Engineering for Scientific Python Applications on GitHub*.

Baskara Nugraha, G. B. (2020). *Normalisasi Kata Tidak Baku yang Tidak Disingkat dengan Jarak Perubahan*.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Brunton, S. L., Noack, B. R., & Koumoutsakos, P. (2020). Machine Learning for Fluid Mechanics. *Annual Review of Fluid Mechanics*, *52*(1), 477–508. https://doi.org/10.1146/annurev-fluid-010719-060214

Cohen, O., Malka, O., & Ringel, Z. (2021). Learning curves for overparametrized deep neural networks: A field theory perspective. *Physical Review Research*, *3*(2). https://doi.org/10.1103/PhysRevResearch.3.023034

Dang, N. C., Moreno-García, M. N., & De la Prieta, F. (2020). Sentiment analysis based on deep learning: A comparative study. *Electronics (Switzerland)*, *9*(3). https://doi.org/10.3390/electronics9030483

Daniel Păvăloaia, V. (2024). Clustering Algorithms in Sentiment Analysis Techniques in Social Media-A Rapid Literature Review. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 15, Issue 3). www.ijacsa.thesai.org

Dash, G., Sharma, C., & Sharma, S. (2023). Sustainable Marketing and the Role of Social Media: An Experimental Study Using Natural Language Processing (NLP). In *Sustainability (Switzerland)* (Vol. 15, Issue 6). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/su15065443

Dwita, I., Tarigan, S., Habibi, R., Nuraini, R., & Fatonah, S. (2023). Evaluasi Algoritma Klasifikasi Machine Learning Kategori Nilai Akhir Tunjangan Kinerja Pegawai. In *Jurnal Sistem Cerdas* (Vol. 6, Issue 3).

Fakhrezi, M. F., Adian Fatchur Rochim, & Dinar Mutiara Kusomo Nugraheni. (2023). Comparison of Sentiment Analysis Methods Based on Accuracy Value Case Study: Twitter Mentions of Academic Article. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *7*(1), 161–167. https://doi.org/10.29207/resti.v7i1.4767

Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., & Hussain, A. (2023). Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, *91*, 424–444. https://doi.org/10.1016/j.inffus.2022.09.025

García-Jurado, A., Pérez-Barea, J. J., & Nova, R. (2021). A new approach to social entrepreneurship: A systematic review and meta-analysis. *Sustainability (Switzerland)*, *13*(5), 1–16. https://doi.org/10.3390/su13052754

Ghazal, T. M., Hussain, M. Z., Said, R. A., Nadeem, A., Hasan, M. K., Ahmad, M., Khan, M. A., & Naseem, M. T. (2021). Performances of k-means clustering algorithm with different distance metrics. *Intelligent Automation and Soft Computing*, *30*(2), 735–742. https://doi.org/10.32604/iasc.2021.019067

Goswami, K., Mathur, P., Rossi, R., & Dernoncourt, F. (2025). *PlotGen: Multi-Agent LLM-based Scientific Data Visualization via Multimodal Feedback*.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2022). Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, *3*(1). https://doi.org/10.1145/3458754

Hossain, M. S., Islam, M. R., Riskhan, B., Hasan, M. M., & Islam, R. (2024). Political sentiment analysis using natural language processing on social media. *International Journal of Applied Methods in Electronics and Computers*. https://doi.org/10.58190/ijamec.2024.108

Huang, R., Ravi, S., He, M., Tian, B., Lerner, S., & Coblenz, M. (2025). *How Scientists Use Jupyter Notebooks: Goals, Quality Attributes, and Opportunities*.

Jain, P. K., Pamula, R., & Srivastava, G. (2021). A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews. *Computer Science Review*, *41*, 100413. https://doi.org/10.1016/j.cosrev.2021.100413

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. In *Nature Reviews Physics* (Vol. 3, Issue 6, pp. 422–440). Springer Nature. https://doi.org/10.1038/s42254-021-00314-5

Khomsah, S., & Sasmito Aribowo, A. (2020). Model Text-Preprocessing Komentar Youtube Dalam Bahasa Indonesia. *Masa Berlaku Mulai*, *1*(3), 648–654.

Killamsetty, K., Sivasubramanian, D., Ramakrishnan, G., & Iyer, R. (2021). *GLISTER: Generalization based Data Subset Selection for Efficient and Robust Learning*. https://medium.com/syncedreview/the-staggering-cost-of-

Li, R., Chen, H., Feng, F., Ma, Z., Wang, X., & Hovy, E. (2021). Dual Graph Convolutional Networks for Aspect-based Sentiment Analysis. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 6319–6329. https://doi.org/10.18653/v1/2021.acl-long.494

Moradi Fard, M., Thonet, T., & Gaussier, E. (2020). Deep k-Means: Jointly clustering with k-Means and learning representations. *Pattern Recognition Letters*, *138*, 185–192. https://doi.org/10.1016/j.patrec.2020.07.028

Priambodo, W., & Zuliarso, E. (2024). Kombinasi K-Means dan LSTM untuk Deteksi Black Campaign di Media Sosial pada Calon Presiden Indonesia 2024. *Jurnal Teknik Informatika (JUTIF)*, *5*(2), 539–550. https://doi.org/10.52436/1.jutif.2024.5.2.1635

Qureshi, K. A., & Sabih, M. (2021). Un-Compromised Credibility: Social Media Based Multi-Class Hate Speech Classification for Text. *IEEE Access*, *9*, 109465–109477. https://doi.org/10.1109/ACCESS.2021.3101977

Ran, X., Zhou, X., Lei, M., Tepsan, W., & Deng, W. (2021). A novel K-means clustering algorithm with a noise algorithm for capturing urban hotspots. *Applied Sciences (Switzerland)*, *11*(23). https://doi.org/10.3390/app112311202

Saura, J. R., Palacios-Marqués, D., & Ribeiro-Soriano, D. (2023). Exploring the boundaries of open innovation: Evidence from social media mining. *Technovation*, *119*. https://doi.org/10.1016/j.technovation.2021.102447

Singh, B. P., Kumar, P., Bhattacharya, C., & Sudarshan, S. (2025). *Efficient Dataframe Systems: Lazy Fat Pandas on a Diet*.

Singh, S., & Mahmood, A. (2021). The NLP Cookbook: Modern Recipes for Transformer Based Deep Learning Architectures. *IEEE Access*, *9*, 68675–68702. https://doi.org/10.1109/ACCESS.2021.3077350

Singh, V., Asari, V. K., & Rajasekaran, R. (2022). A Deep Neural Network for Early Detection and Prediction of Chronic Kidney Disease. *Diagnostics*, *12*(1). https://doi.org/10.3390/diagnostics12010116

Tran, K. A., Pearson, J. V., & Waddell, N. (2025). *xML-workFlow: an end-to-end explainable scikit-learn workflow for rapid biomedical experimentation*.

Zhang, B., Liang, P., Zhou, X., Ahmad, A., & Waseem, M. (2023). *Practices and Challenges of Using GitHub Copilot: An Empirical Study*. https://doi.org/10.18293/SEKE2023-077

Zhao, W., Hou, Z., Wu, S., Gao, Y., Dong, H., Wan, Y., Zhang, H., Sui, Y., & Zhang, H. (2024). *NL2Formula: Generating Spreadsheet Formulas from Natural Language Queries*.

Zikrina, S. A., & Fitriyani. (2025). Advancing Hate Speech Detection in Indonesian Language Using Graph Neural Networks and TF-IDF. *Jurnal RESTI*, *9*(1), 137–145. https://doi.org/10.29207/resti.v9i1.6179

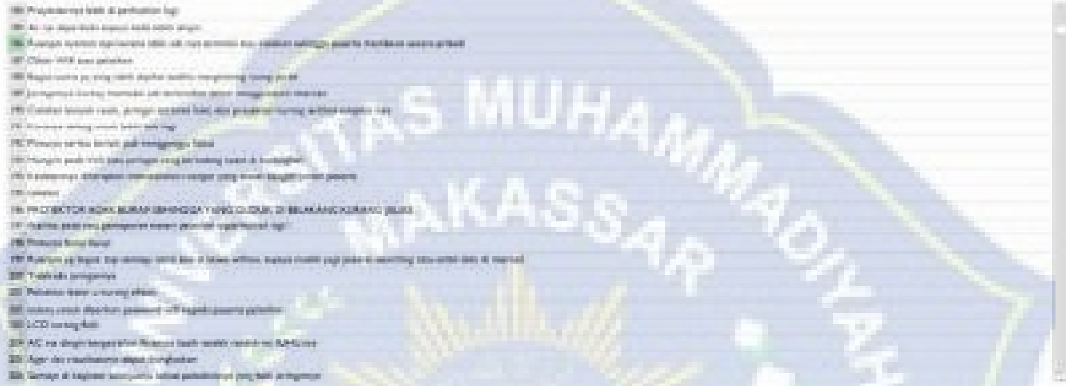
UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

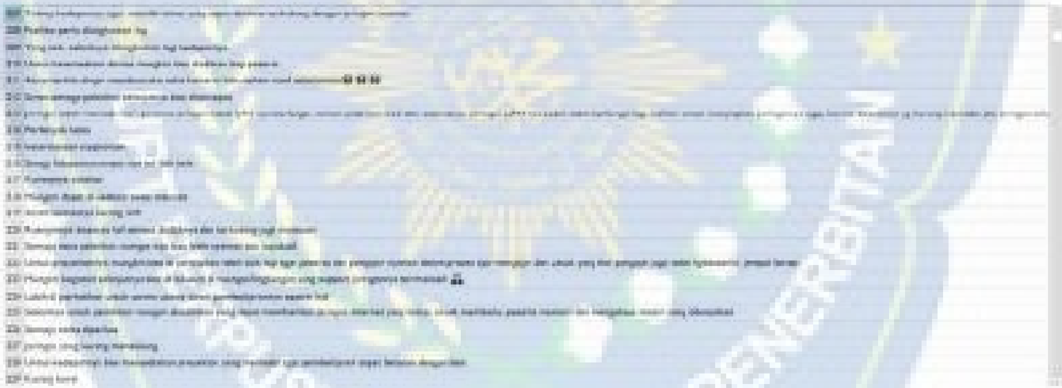
LAMPIRAN

Lampiran 1. Dataset

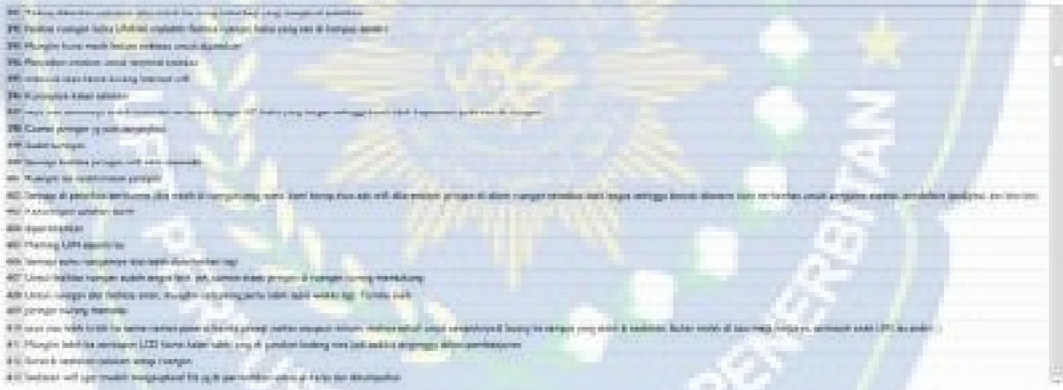
UNIVERSITAS MUHAMMADIYAH MAKASSAR
PERPUSTAKAAN DAN PENERBITAN

UNIVERSITAS MUHAMMADIYAH MAKASSAR
PERPUSTAKAAN DAN PENERBITAN

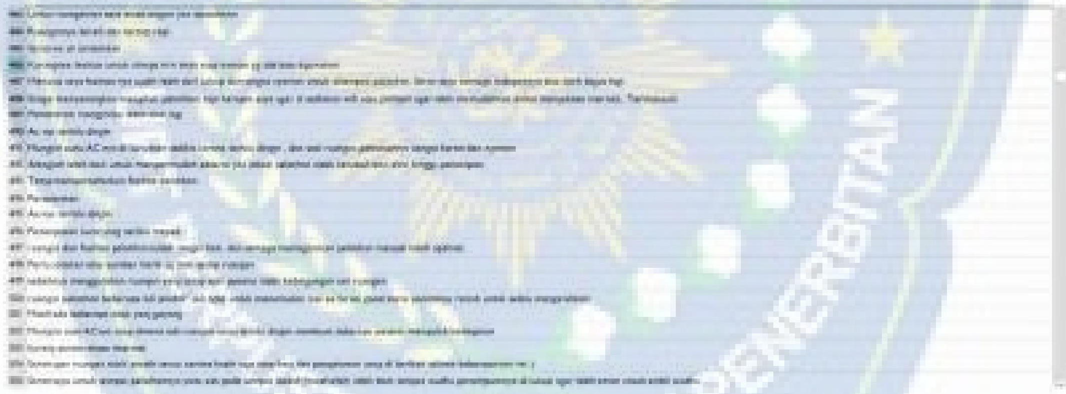




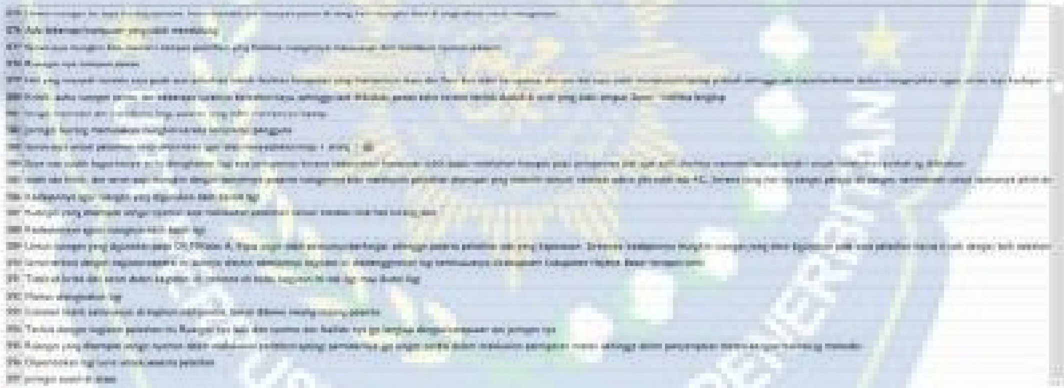
UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN



UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN



UNIVERSITAS MUHAMMADIYAH MAKASSAR
PERPUSTAKAAN DAN PENERBITAN

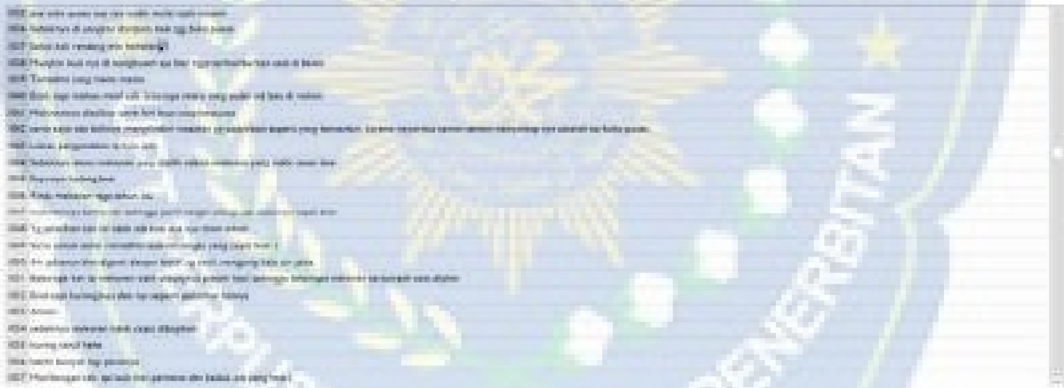


UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

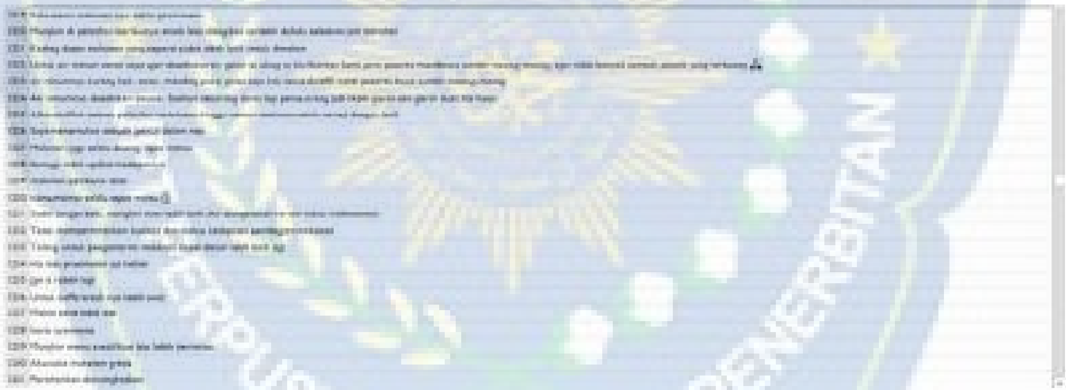
UNIVERSITAS MUHAMMADIYAH MAKASSAR
PERPUSTAKAAN DAN PENERBITAN

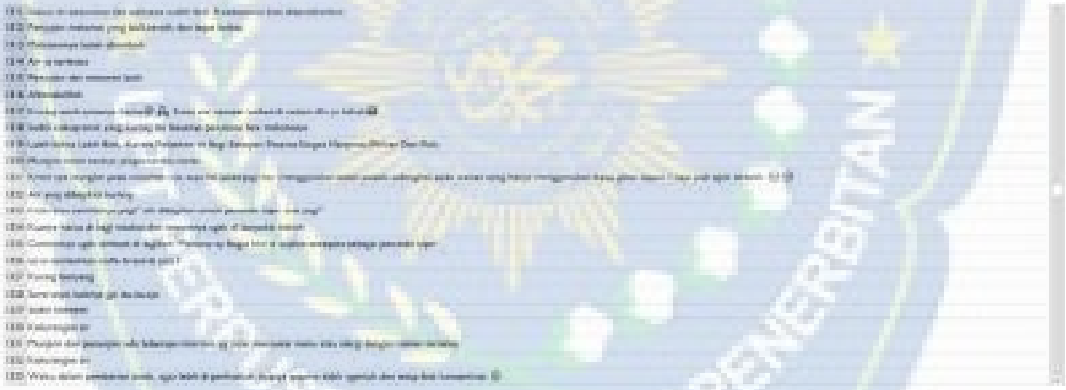
UNIVERSITAS MUHAMMADIYAH MAKASSAR
PERPUSTAKAAN DAN PENERBITAN

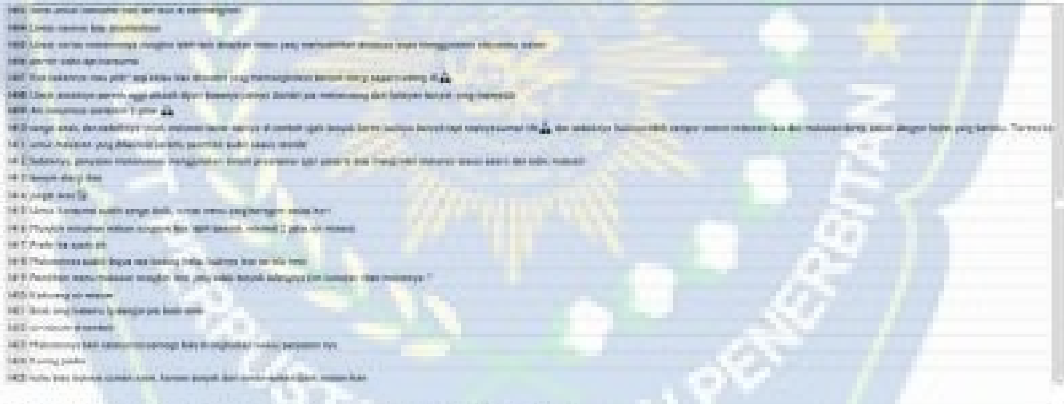


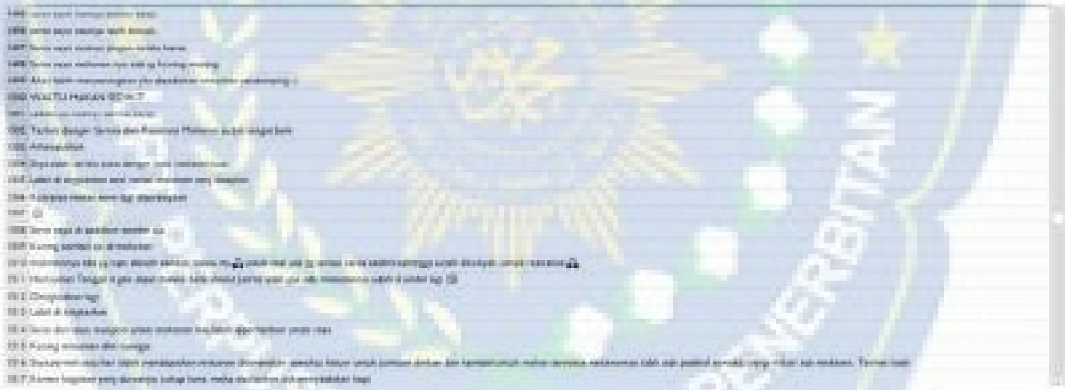
UNIVERSITAS MUHAMMADIYAH MAKASSAR
PERPUSTAKAAN DAN PENERBITAN

UNIVERSITAS MUHAMMADIYAH MAKASSAR
PERPUSTAKAAN DAN PENERBITAN

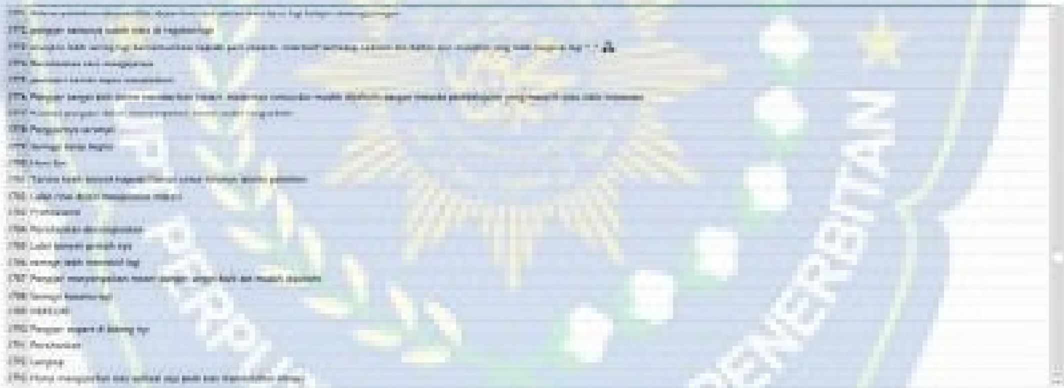




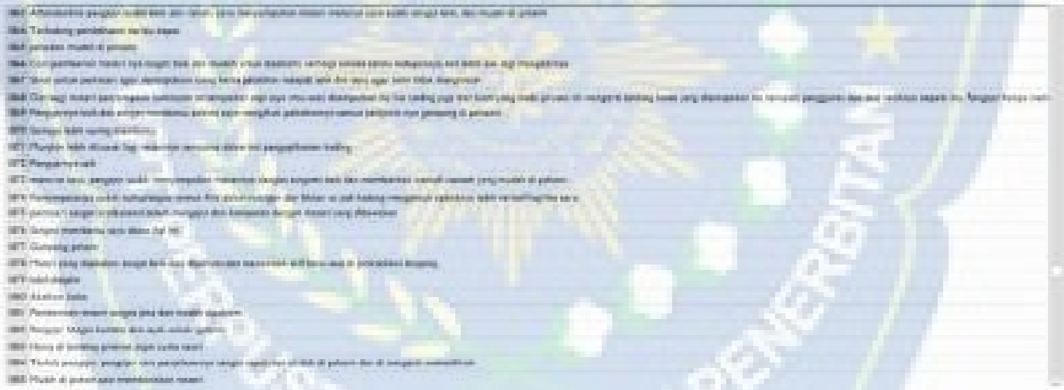




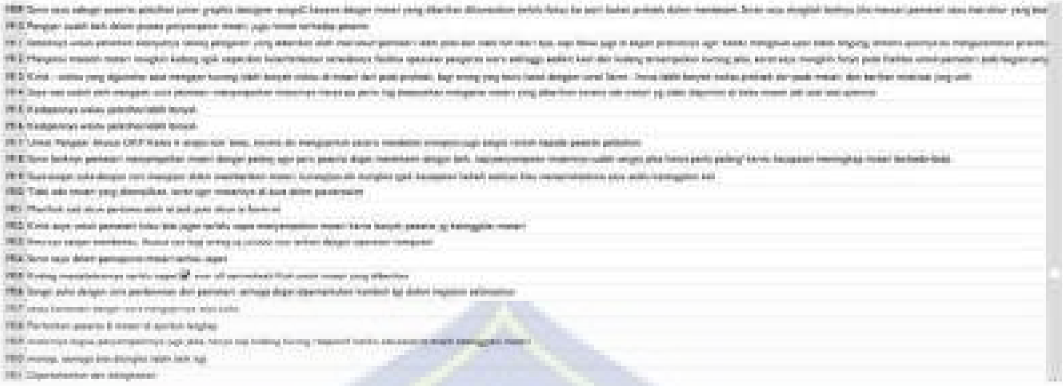
UNIVERSITAS MUHAMMADIYAH MAKASSAR
PERPUSTAKAAN DAN PENERBITAN

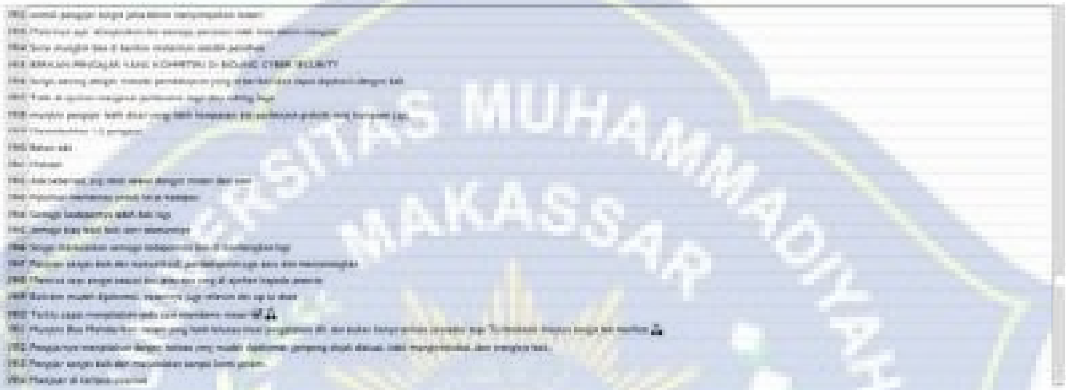


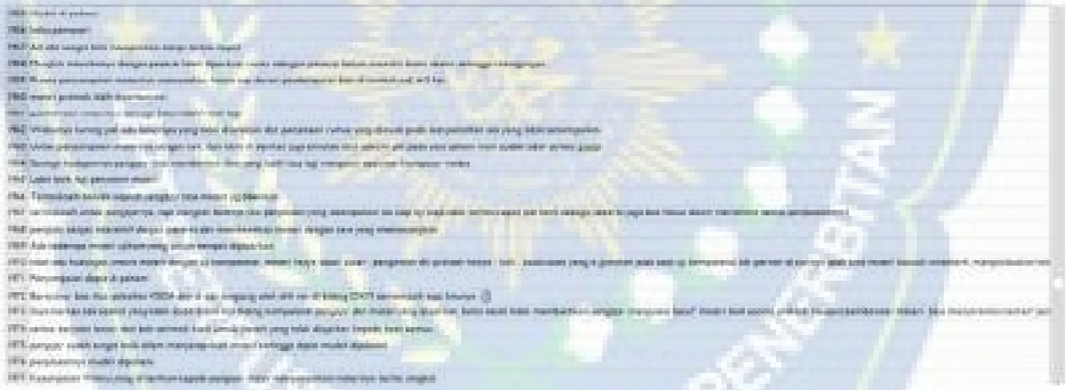
UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN



UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

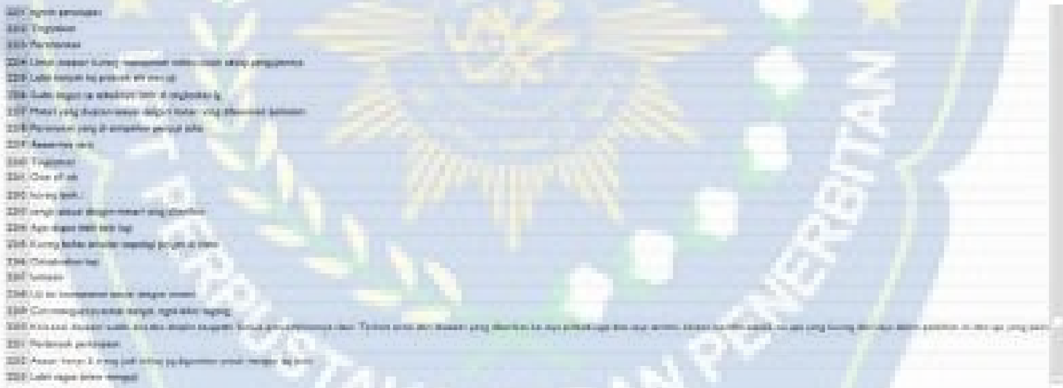


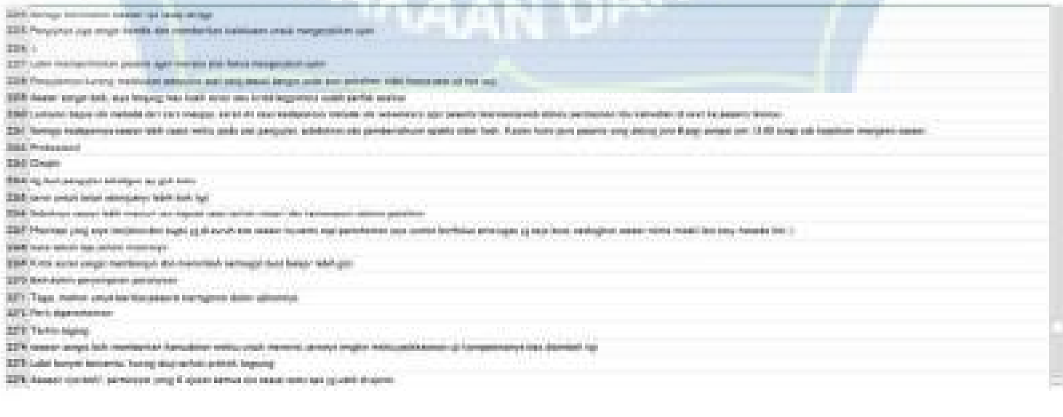




UNIVERSITAS MUHAMMADIYAH MAKASSAR
PERPUSTAKAAN DAN PENERBITAN

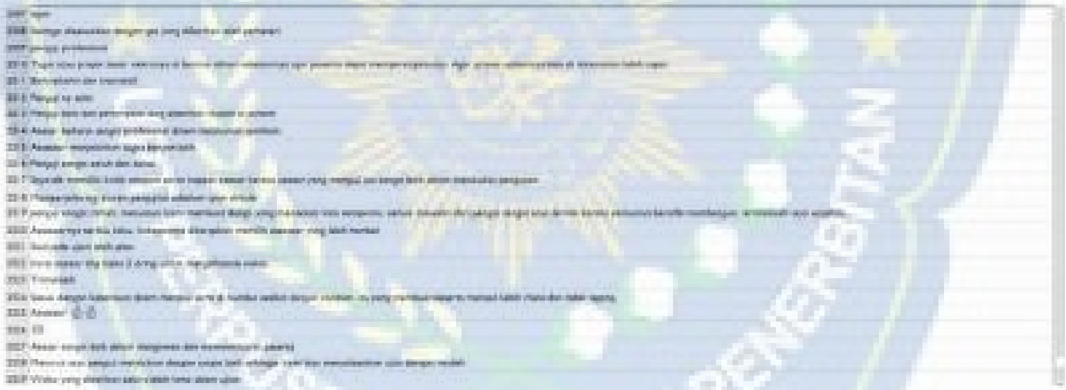
UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN



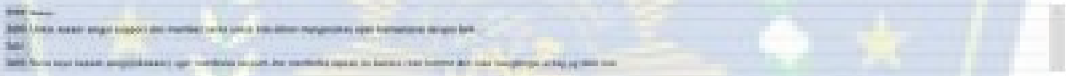


UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

UNIVERSITAS MUHAMMADIYAH MAKASSAR
PERPUSTAKAAN DAN PENERBITAN



UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN



UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

Lampiran 2. Source Code

```
# %% [markdown]
# ## 1. Inisialisasi & Import Library
# Langkah pertama adalah mengimpor semua library Python yang diperlukan untuk analisis,
termasuk `pandas` untuk manipulasi data, `numpy` untuk operasi numerik, dan `scikit-learn`
untuk machine learning.

# %%
import pandas as pd
import numpy as np
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score

# %% [markdown]
# ## 2. Memuat Dataset Awal
# Memuat data dari file Excel yang berisi kritik dan saran

# %%
df = pd.read_excel('kritik_saran.xlsx', sheet_name="Sheet1")

# %% [markdown]
# ## 3. Preprocessing Teks
# Tahap ini adalah inti dari persiapan data teks. Proses ini meliputi:
# 1.  **Mengubah tipe data** kolom ulasan menjadi string.
# 2.  **Stemming**: Mengubah kata ke bentuk dasarnya menggunakan library Sastrawi.
# 3.  **Stopword Removal**: Menghapus kata-kata umum yang tidak memiliki makna
signifikan (misalnya: 'dan', 'di', 'yang').
# 4.  **Vektorisasi TF-IDF**: Mengubah teks menjadi representasi numerik yang dapat
diolah oleh model machine learning. TF-IDF (Term Frequency-Inverse Document Frequency)
memberikan bobot pada kata-kata berdasarkan frekuensinya dalam dokumen dan
kelangkaannya di seluruh korpus.

# %%
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
from Sastrawi.StopWordRemover.StopWordRemoverFactory import
StopWordRemoverFactory
from sklearn.feature_extraction.text import TfidfVectorizer

# Mengubah tipe data kolom 'ULASAN' menjadi string untuk menghindari error
df['ULASAN'] = df['ULASAN'].astype(str)

# Inisialisasi Stemmer
stemmer_factory = StemmerFactory()
```

UNIVERSITAS MUHAMMADIYAH MAKASSAR
PERPUSTAKAAN DAN PENERBITAN

```
stemmer = stemmer_factory.create_stemmer()

# Inisialisasi StopWordRemover
stopword_factory = StopWordRemoverFactory()
indonesian_stopwords = stopword_factory.get_stop_words()

# Terapkan stemming pada kolom 'ULASAN'
df['ULASAN'] = df['ULASAN'].apply(lambda x: stemmer.stem(x))

# Buat TF-IDF Vectorizer dengan stopwords bahasa Indonesia
vectorizer = TfidfVectorizer(stop_words=indonesian_stopwords)
X_tfidf = vectorizer.fit_transform(df['ULASAN'])

# %% [markdown]
# ## 4. Filterisasi Data Berdasarkan Relevansi
# Untuk memastikan kualitas analisis, ulasan yang tidak relevan akan disaring. Kriteria relevansi ditentukan berdasarkan:
# - **Skor TF-IDF**: Ulasan dengan skor total di atas ambang batas dianggap lebih signifikan.
# - **Jumlah Kata**: Ulasan yang terlalu pendek (misalnya, kurang dari 3 kata) dianggap kurang informatif.
#
# Data kemudian dipisahkan menjadi dua kategori: relevan dan tidak relevan.

# %%
# Hitung total TF-IDF score untuk setiap dokumen (ulasan)
tfidf_scores = X_tfidf.sum(axis=1) # hasilnya matrix

# Ubah ke array 1D
df['TFIDF_Score'] = np.asarray(tfidf_scores).flatten()

# Hitung jumlah kata di setiap kalimat
df['Jumlah_Kata'] = df['ULASAN'].apply(lambda x: len(str(x).split()))

# Tandai relevan dan tidak relevan sesuai kriteria
df['Relevan'] = (df['TFIDF_Score'] > 1) & (df['Jumlah_Kata'] > 2)
df['Tidak_Relevan'] =(df['TFIDF_Score'] <= 1) & (df['Jumlah_Kata'] <= 2)

# Pisahkan data relevan dan tidak relevan
df_relevan = df[df['Relevan']]
df_tidak_relevan = df[df['Tidak_Relevan']]

# Cek hasilnya (opsional)
print(f"Jumlah data relevan: {len(df_relevan)}")
print(f"Jumlah data tidak relevan: {len(df_tidak_relevan)}")
```

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

```
# %% [markdown]
# ## 5. Menyimpan Hasil Filterisasi
# Data yang telah difilter disimpan ke dalam file Excel terpisah untuk dokumentasi dan
analisis lebih lanjut.

# %%
df_relevan[['ULASAN', 'TFIDF_Score']].to_excel('ulasan_relevan.xlsx', index=False)
df_tidak_relevan[['ULASAN', 'TFIDF_Score']].to_excel('ulasan_tidak_relevan.xlsx',
index=False)

# %% [markdown]
# ## 6. Analisis Klastering (Topic Modeling)
# Pada tahap ini, kita akan mengelompokkan ulasan yang relevan ke dalam beberapa topik
atau klaster menggunakan algoritma K-Means.

# %% [markdown]
# ### 6.1. Memuat Data Relevan
# Memuat kembali data ulasan yang sudah difilter sebagai dasar untuk klastering.

# %%
df = pd.read_excel('ulasan_relevan.xlsx', sheet_name="Sheet1")

# %% [markdown]
# ### 6.2. Menentukan Kombinasi Parameter Terbaik
# Sebelum melakukan klastering, kita perlu mengurangi dimensi data TF-IDF menggunakan
`TruncatedSVD` (mirip dengan PCA untuk data sparse). Selanjutnya, kita mencari kombinasi
jumlah komponen SVD (`n_components`) dan jumlah klaster (`n_clusters`) yang optimal
dengan menggunakan **Silhouette Score**. Skor ini mengukur seberapa baik sebuah objek
ditempatkan dalam klasternya dibandingkan dengan klaster lain. Nilai yang lebih tinggi
menunjukkan klaster yang lebih padat dan terpisah dengan baik.

# %%
from Sastrawi.StopWordRemover.StopWordRemoverFactory import
StopWordRemoverFactory
from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer
from sklearn.decomposition import LatentDirichletAllocation, TruncatedSVD, PCA
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# Stopwords bahasa Indonesia
factory = StopWordRemoverFactory()
```

UNIVERSITAS MUHAMMADIYAH MAKASSAR
PERPUSTAKAAN DAN PENERBITAN

```
stopwords_id = factory.get_stop_words()

# TF-IDF vectorizer
vectorizer_tfidf = TfidfVectorizer(stop_words=stopwords_id)
X_tfidf = vectorizer_tfidf.fit_transform(df['ULASAN'])

# CountVectorizer dan LDA
vectorizer_count = CountVectorizer(max_df=0.9, min_df=5, stop_words=stopwords_id)
X_counts = vectorizer_count.fit_transform(df['ULASAN'])
lda = LatentDirichletAllocation(n_components=2, random_state=42)
lda.fit(X_counts)

# Topik LDA ke df
topic_distribution = lda.transform(X_counts)
dominant_topic = np.argmax(topic_distribution, axis=1)
df['Topik_LDA'] = dominant_topic

# Cari kombinasi terbaik n_components dan n_clusters
n_components_list = [50, 100, 150, 200, 250]
n_clusters_list = range(2, 9)

best_score = -1
best_config = (None, None)

for n_components in n_components_list:
    svd = TruncatedSVD(n_components=n_components, random_state=0)
    X_reduced = svd.fit_transform(X_tfidf)

    for k in n_clusters_list:
        kmeans = KMeans(n_clusters=k, random_state=0)
        labels = kmeans.fit_predict(X_reduced)
        score = silhouette_score(X_reduced, labels)
        print(f"n_components={n_components}, n_clusters={k}, Silhouette={score:.4f}")

        if score > best_score:
            best_score = score
            best_config = (n_components, k)

print(f"\nBest combo: n_components={best_config[0]}, n_clusters={best_config[1]} with
Silhouette={best_score:.4f}")

# Buat model dengan parameter terbaik
best_n_components, best_n_clusters = best_config
svd = TruncatedSVD(n_components=best_n_components, random_state=0)
X_reduced = svd.fit_transform(X_tfidf)
```

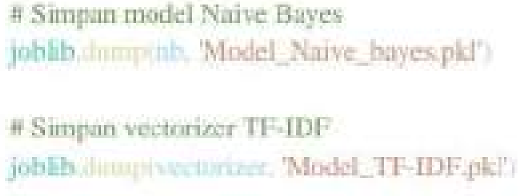
UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

```
kmeans = KMeans(n_clusters=best_n_clusters, random_state=0)
df['Topik'] = kmeans.fit_predict(X_reduced)

# Fungsi top tfidf kata per cluster
def top_tfidf_words(docs, top_n=10):
    vec = TfidfVectorizer(stop_words=stopwords_id)
    X = vec.fit_transform(docs)
    mean_tfidf = np.asarray(X.mean(axis=0)).ravel()
    top_indices = mean_tfidf.argsort()[::-1][:top_n]
    return [vec.get_feature_names_out()[i] for i in top_indices]

for i in range(best_n_clusters):
    cluster_docs = df[df['Topik'] == i]['ULASAN']
    print(f"\nTop kata di Klaster {i}: {top_tfidf_words(cluster_docs)}")

# Visualisasi hasil klaster 2D
pca = PCA(n_components=2, random_state=0)
X_2d = pca.fit_transform(X_reduced)

plt.figure(figsize=(8,6))
plt.scatter(X_2d[:, 0], X_2d[:, 1], c=df['Topik'], cmap='viridis', s=30)
plt.title("Visualisasi Klaster (2D)")
plt.xlabel("PCA 1")
plt.ylabel("PCA 2")
plt.colorbar(label='Topik')
plt.grid(True)
plt.show()

# %% [markdown]
# ## 7. Interpretasi dan Pelabelan Klaster
# Setelah klaster terbentuk, langkah selanjutnya adalah menginterpretasi makna dari setiap
klaster. Ini dilakukan dengan melihat kata-kata yang paling representatif dari setiap klaster.
Berdasarkan interpretasi tersebut, setiap klaster diberi label yang sesuai (misalnya, 'Kritik &
Harapan', 'Saran & Pujian'). Hasilnya kemudian disimpan dalam file Excel baru.

# %%
# 1. Buat label yang bermakna untuk tiap topik
topic_labels = {
    0: "Kritik Dan Harapan",
    1: "Saran dan Pujian"
}

# 2. Map ke DataFrame
df['Topik_Label'] = df['Topik'].map(topic_labels)
```

```
# 3. Tampilkan hasil lengkap

pd.set_option('display.max_colwidth', None)
print(df[['ULASAN', 'Topik', 'Topik_Label']])
# Lihat contoh ulasan dari tiap topik
print(df[df['Topik'] == 0]['ULASAN'].sample(5))
print(df[df['Topik'] == 1]['ULASAN'].sample(5))


df[['ULASAN', 'Topik', 'Topik_Label']].to_excel("hasil_klaster_teks_asli.xlsx", index=False)

# %% [markdown]
# ## 8. Pelatihan Model Klasifikasi (Naive Bayes)
# Setelah data memiliki label topik (hasil dari klastering), kita dapat menggunakan data ini untuk melatih model klasifikasi. Model ini nantinya dapat memprediksi topik dari ulasan baru secara otomatis.
#
# Prosesnya adalah sebagai berikut:
# 1. **Pembagian Data**: Data dibagi menjadi set pelatihan (80%) dan set pengujian (20%).
# 2. **Vektorisasi TF-IDF**: Teks kembali diubah menjadi vektor numerik.
# 3. **Pelatihan Model**: Model `Multinomial Naive Bayes` dilatih menggunakan data pelatihan. Algoritma ini sangat cocok untuk klasifikasi teks.

# %%
df['ULASAN'] = df['ULASAN'].fillna("").astype(str)
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer

# Fitur dan label
X = df['ULASAN']
y = df['Topik_Label']

# Split data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42, stratify=y)

# TF-IDF
vectorizer = TfidfVectorizer(stop_words=stopwords_id)
X_train_tfidf = vectorizer.fit_transform(X_train)
X_test_tfidf = vectorizer.transform(X_test)

from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import classification_report, confusion_matrix
```

```
# Model
nb = MultinomialNB()
nb.fit(X_train_tfidf, y_train)

# Prediksi
y_pred = nb.predict(X_test_tfidf)

# %% [markdown]
# ## 9. Evaluasi Model
# Performa model dievaluasi menggunakan data pengujian. Metrik yang digunakan adalah:
# - **Classification Report**: Menampilkan `precision`, `recall`, dan `f1-score` untuk setiap kelas. Ini memberikan gambaran lengkap tentang seberapa baik model mengklasifikasikan setiap topik.
# - **Confusion Matrix**: Menunjukkan jumlah prediksi yang benar dan salah untuk setiap kelas. Ini membantu untuk melihat di mana model sering membuat kesalahan.

# %%
print("\n📊 Classification Report:")
print(classification_report(y_test, y_pred))

print("\n📉 Confusion Matrix:")
print(confusion_matrix(y_test, y_pred))

# %% [markdown]
# ## 10. Uji Coba Model pada Data Baru
# Menggunakan model yang telah dilatih untuk memprediksi topik dari beberapa contoh kalimat baru. Ini menunjukkan bagaimana model akan bekerja dalam aplikasi nyata.

# %%
new_texts = ["materi sangat mudah dipahami", "kursi keras dan ruang terlalu panas"]
new_vec = vectorizer.transform(new_texts)
predictions = nb.predict(new_vec)

for txt, label in zip(new_texts, predictions):
    print(f"\n📝 \"{txt}\" → 🏷 Predicted Topik: {label}")

# %% [markdown]
# ## 11. Menyimpan Model dan Vectorizer
# Langkah terakhir adalah menyimpan model Naive Bayes dan TF-IDF Vectorizer yang telah dilatih ke dalam file menggunakan `joblib`. Ini memungkinkan kita untuk memuat kembali model dan vectorizer di lain waktu tanpa harus melatih ulang dari awal.

# %%
import joblib
```

UNIVERSITAS MUHAMMADIYAH MAKASSAR
PERPUSTAKAAN DAN PENERBITAN



```
# Simpan model Naive Bayes
joblib.dump(nb, 'Model_Naive_bayes.pkl')

# Simpan vectorizer TF-IDF
joblib.dump(vectorizer, 'Model_TF-IDF.pkl')
```

UNIVERSITAS MUHAMMADIYAH
MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

Lampiran 3. Surat Permohonan Data Penelitian dari Program Studi

MAJELIS PENDIDIKAN TINGGI PIMPINAN PUSAT MUHAMMADIYAH
UNIVERSITAS MUHAMMADIYAH MAKASSAR
FAKULTAS TEKNIK
PROGRAM STUDI INFORMATIKA

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

Makassar, 13 Muharram 1447 H
8 Juli 2025 M

Nomor : 050/05/IF/C.4-VI/VII/47/2025
Lamp. : -
Hal : **Permohonan Data Penelitian**

Kepada yang Terhormat,
Ketua LP3M Unismuh Makassar
Di –
Tempat

Assalamu 'Alaikum Warahmatullahi Wabarakatuh

Dengan Rahmat Allah SWT. Semoga aktivitas kita bernilai ibadah di Sisi–Nya. Dalam rangka penyelesaian Tugas Akhir pada Program Studi Informatika dengan judul "Implementasi K-Means dan Analisis Sentimen Kritik Saran Berbasis NLP pada Data Monev BBPSDMP Kominfo Kota Makassar". Bersama ini kami sampaikan mahasiswa:

Stambuk	Nama
105 84 11038 21	Syahril Akbar

Sehubungan dengan hal tersebut, maka kami memohon dibuatkan surat pengantar pada instansi di bawah ini:

Nama Instansi : Balai Besar Pengembangan SDM dan Penelitian Komunikasi dan Informatika Makassar
Alamat : Jl. Prof. Abdurrahman Basalamah IINo.25, Karampuang, Kec. Panakkukang, Kota Makassar

Demikian surat kami atas perhatian dan kerja samanya kami haturkan banyak terima kasih.
Jazakumullah Khaeran Katsiran
Wassalamu 'Alaikum warahmatullah Wabarakatuh

Ketua Program Studi
Informatika

Muhyiddin A M Hayat, S.Kom., M.T.
NBM : [illegible]

Tembusan:
1. Dekan Fakultas Teknik
2. Arsip

Gedung Menara Iqra Lantai 3
Jl. Sultan Alauddin No. 259 Telp. (0411) 866 972 Fax (0411) 865 588 Makassar 90221
Web: https://teknik.unismuh.ac.id/, e-mail: teknik@unismuh.ac.id

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

Lampiran 4. Surat Permohonan Izin Pelaksanaan Penelitian dari LP3M

UNIVERSITAS MUHAMMADIYAH MAKASSAR
LEMBAGA PENELITIAN PENGEMBANGAN DAN PENGABDIAN KEPADA MASYARAKAT
Jl. Sultan Alauddin No. 259 Telp. 866972 Fax. (0411) 865588 Makassar 90221 e-mail lp3m@unismuh.ac.id

بســـــــم الله الرحمن الرحيم

Nomor : 89/LP3M/05/C.4-VIII/VII/1447/2025
Lampiran : 1 (satu) rangkap proposal
Hal : Permohonan Izin Pelaksanaan Penelitian

Kepada Yth:
Bapak Gubernur Provinsi Sulawesi Selatan
Cq. Kepala Dinas Penanaman Modal & PTSP Provinsi Sulawesi Selatan
di-
Makassar

Assalamu Alaikum Wr. Wb

Berdasarkan surat Dekan Fakultas Program Study Informatika, nomor: 050 tanggal: 10 Juli 2025, menerangkan bahwa mahasiswa dengan data sebagai berikut:

Nama : SYAHRIL AKBAR
Nim : 105841103821
Fakultas : Teknik
Prodi : Informatika

Bermaksud melaksanakan penelitian/pengumpulan data dalam rangka penulisan laporan tugas akhir Skripsi dengan judul:

"Implementasi K-Means Dan Analisis Sentimen Kritik Saran Berbasis NLP Pada Data Monev BBPSDMP Kominfo Makassar"

Yang akan dilaksanakan dari tanggal 16 Juli 2025 s/d 16 September 2025.

Sehubungan dengan maksud di atas, kiranya Mahasiswa tersebut diberikan izin untuk melakukan penelitian sesuai ketentuan yang berlaku.

Demikian, atas perhatian dan kerjasamanya diucapkan jazakumullahu khaeran katsiraa.
Billahi Fii Sabilil Haq, Fastabiqul Khairat.
Wassalamu Alaikum Wr. Wb.

Makassar, 15 Muharram 1447
11 Juli 2025

Ketua LP3M Unismuh Makassar,

Dr. Muh. Arief Muhsin, M.Pd.
NBM. 112 7761

Lampiran 5. Surat Izin Penelitian dari Dinas Penanaman Modal dan PTSP Provinsi Sulawesi Selatan

PEMERINTAH PROVINSI SULAWESI SELATAN
DINAS PENANAMAN MODAL DAN PELAYANAN TERPADU SATU PINTU
Jl.Bougenvile No.5 Telp. (0411) 441077 Fax. (0411) 448936
Website : http://simap-new.sulselprov.go.id Email : ptsp@sulselprov.go.id
Makassar 90231

Nomor : **15648/S.01/PTSP/2025**
Lampiran : -
Perihal : **Izin penelitian**

Kepada Yth.
Kepala Balai Besar Pengembangan Sumber Daya Manusia dan Penelitian Kominfo Makassar

di-
Tempat

Berdasarkan surat Ketua LP3M Unismuh Makassar Nomor : 89/LP3M/05/C.4-VIII/VII/1447/2025 tanggal 11 Juli 2025 perihal tersebut diatas, mahasiswa/peneliti dibawah ini:

N a m a : **SYAHRIL AKBAR**
Nomor Pokok : 105841103821
Program Studi : Teknik Informatika
Pekerjaan/Lembaga : Mahasiswa (S1)
Alamat : Jl. Slt Alauddin No. 259, Makassar

Bermaksud untuk melakukan penelitian di daerah/kantor saudara dalam rangka menyusun SKRIPSI, dengan judul :

" IMPLEMENTASI K-MEANS DAN ANALISIS SENTIMEN KRITIK SARAN BERBASIS NLP PADA DATA MONEV BBPSDMP KOMINFO MAKASSAR "

Yang akan dilaksanakan dari : Tgl. ***16 Juli s/d 16 September 2025***

Sehubungan dengan hal tersebut diatas, pada prinsipnya kami ***menyetujui*** kegiatan dimaksud dengan ketentuan yang tertera di belakang surat izin penelitian.

Demikian Surat Keterangan ini diberikan agar dipergunakan sebagaimana mestinya.

Diterbitkan di Makassar
Pada Tanggal 16 Juli 2025

KEPALA DINAS PENANAMAN MODAL DAN PELAYANAN TERPADU SATU PINTU PROVINSI SULAWESI SELATAN

ASRUL SANI, S.H., M.Si.
Pangkat : PEMBINA UTAMA MUDA (IV/c)
Nip : 19750321 200312 1 008

Tembusan Yth
1. Ketua LP3M Unismuh Makassar di Makassar;
2. *Pertinggal.*

Lampiran 6. Surat Balasan Izin Penelitian dari BBPSDMP Kominfo Makassar

KEMENTERIAN KOMUNIKASI DAN INFORMATIKA RI
BADAN PENGEMBANGAN SUMBER DAYA MANUSIA KOMUNIKASI DAN INFORMATIKA
BALAI BESAR PENGEMBANGAN SUMBER DAYA MANUSIA DAN PENELITIAN
KOMUNIKASI DAN INFORMATIKA MAKASSAR
Indonesia Terkoneksi: Makin Digital, Makin Maju
KOMINFO
Jl. Prof. Abdurrahman Basalamah II No 25 Makassar 90231 Telp. (0411) 460084 || www.kominfo.go.id

Nomor : B-389/BBPSDMP.73/UM.01.01/08/2025
Sifat : Biasa
Hal : Balasan Surat Izin Penelitian

Makassar, 1 Agustus 2025

Kepada Yth.
Kepala Dinas Penanaman Modal dan Pelayanan Terpadu
Satu Pintu Provinsi Sulawesi Selatan
di Tempat

Sehubungan dengan Surat Permohonan Izin Penelitian No. 15648/S.01/PTSP/2025 yang diajukan kepada kami oleh mahasiswa Bapak/Ibu atas nama :

Nama : Syahril Akbar
Nomo Pokok : 105841103821
Program Studi : Teknik Informatika

Bersama surat ini kami memberikan izin kepada mahasiswa tersebut untuk melakukan Penelitian Skripsi dengan judul Implementasi K-Means dan Analisis Sentimen Kritik Saran Berbasis NLP Pada Data Monev BBPSDMP Kominfo Makassar. Kontak Person yang bisa dihubungi atas nama Mardiana (085919646455 9) atau Dhilah (085395808110).

Selanjutnya dalam upaya mendukung Zona Integritas Wilayah Bebas dari Korupsi, Apabila dalam Pelayanan kami terdapat penyimpangan atau pelanggaran kode etik oleh pegawai kami, mohon perkenan Bapak/Ibu melaporkan dengan disertai bukti otentik (identitas pelapor akan dijamin kerahasiaannya).

Demikian surat balasan ini dibuat agar dapat dipergunakan sebagaimana mestinya.

Kepala BBPSDMP Makassar

Baso Saleh

Lampiran 7. Surat Keterangan Bebas Plagiat

MAJELIS PENDIDIKAN TINGGI PIMPINAN PUSAT MUHAMMADIYAH
UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN
Alamat kantor: Jl.Sultan Alauddin NO.259 Makassar 90221 Tlp.(0411) 866972,881593, Fax.(0411) 865588

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

SURAT KETERANGAN BEBAS PLAGIAT

UPT Perpustakaan dan Penerbitan Universitas Muhammadiyah Makassar, Menerangkan bahwa mahasiswa yang tersebut namanya di bawah ini:

Nama : Syahril Akbar
Nim : 105841103421
Program Studi : Pendidikan Guru (Pendidikan Anak Usia Dini)

Dengan nilai:

No	Bab	Nilai	Ambang Batas
1	Bab 1	2%	10 %
2	Bab 2	12%	25 %
3	Bab 3	4%	10 %
4	Bab 4	7%	10 %
5	Bab 5	5%	5 %

Dinyatakan telah lulus cek plagiat yang diadakan oleh UPT- Perpustakaan dan Penerbitan Universitas Muhammadiyah Makassar Menggunakan Aplikasi Turnitin.

Demikian surat keterangan ini diberikan kepada yang bersangkutan untuk dipergunakan seperlunya.

Makassar, 25 Agustus 2025
Mengetahui,

Kepala UPT- Perpustakaan dan Pernerbitan,

Nur[illegible] S.[illegible] M.I.P
NBM. 964 591

Jl. Sultan Alauddin no 259 makassar 90222
Telepon (0411)866972,881 593,fax (0411)865 588
Website: www.library.unismuh.ac.id
E-mail : perpustakaan@unismuh.ac.id

Lampiran 8. Hasil Turnitin

BAB I Syahril Akbar 105841103821

by Tahap Tutup

Submission date: 25-Aug-2025 08:41AM (UTC+0700)
Submission ID: 2734624745
File name: Skripsi_BAB_I_Syahril_Akbar_105841103821_Informatika.docx (169.57K)
Word count: 1328
Character count: 9008

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

BAB I Syahril Akbar 105841103821

ORIGINALITY REPORT

LULUS

2%	2%	3%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	kc.umn.ac.id Internet Source	2%

Exclude quotes	Off	Exclude matches	[illegible]
Exclude bibliography	Off		

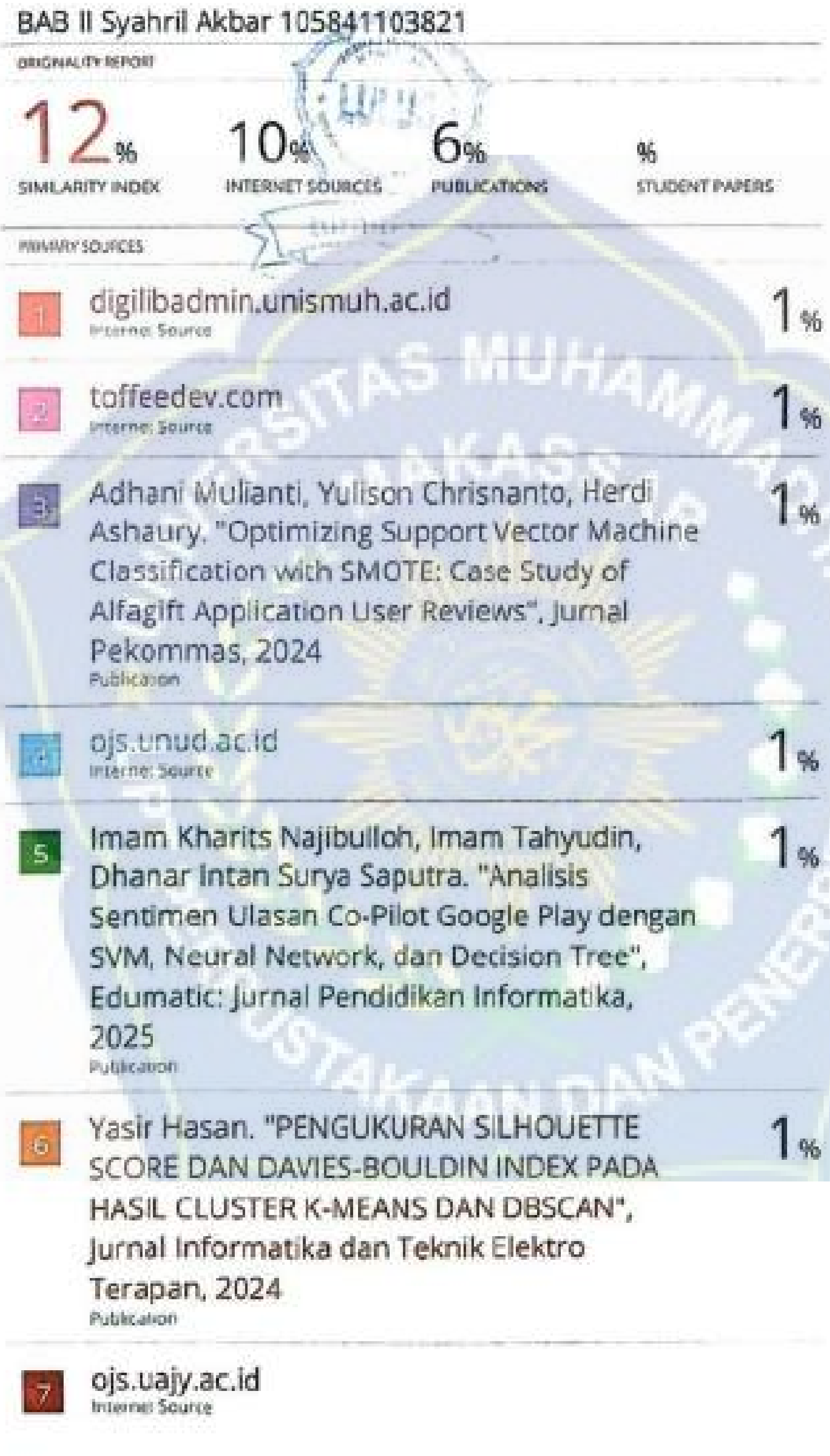
UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

BAB II Syahril Akbar 105841103821

by Tahap Tutup

Submission date: 25-Aug-2025 08:42AM (UTC+0700)
Submission ID: 2734625151
File name: Skripsi_BAB_II_Syahril_Akbar_105841103821_Informatika.docx (409.77K)
Word count: 3382
Character count: 22776

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN



BAB II Syahril Akbar 105841103821

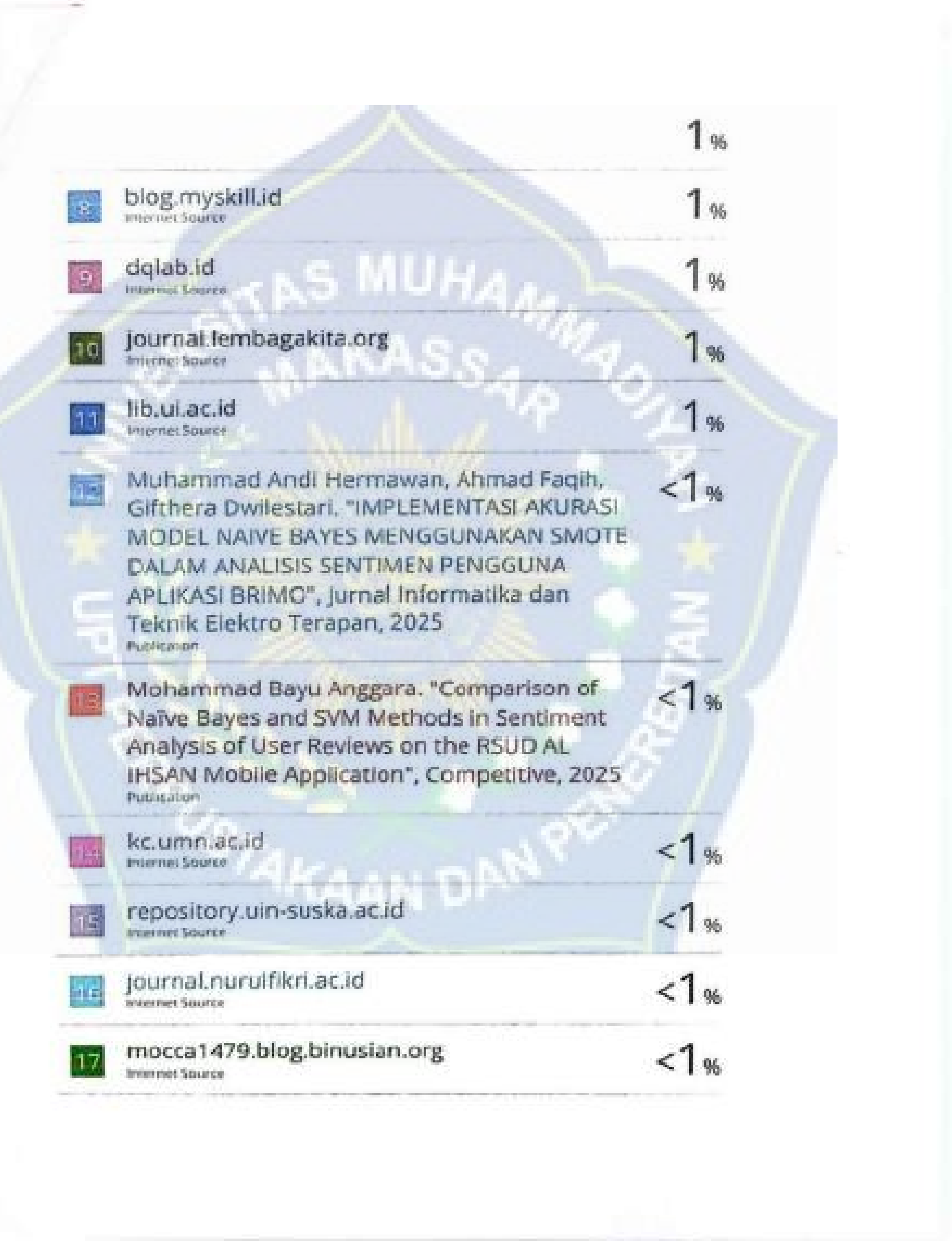
ORIGINALITY REPORT

12%	10%	6%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

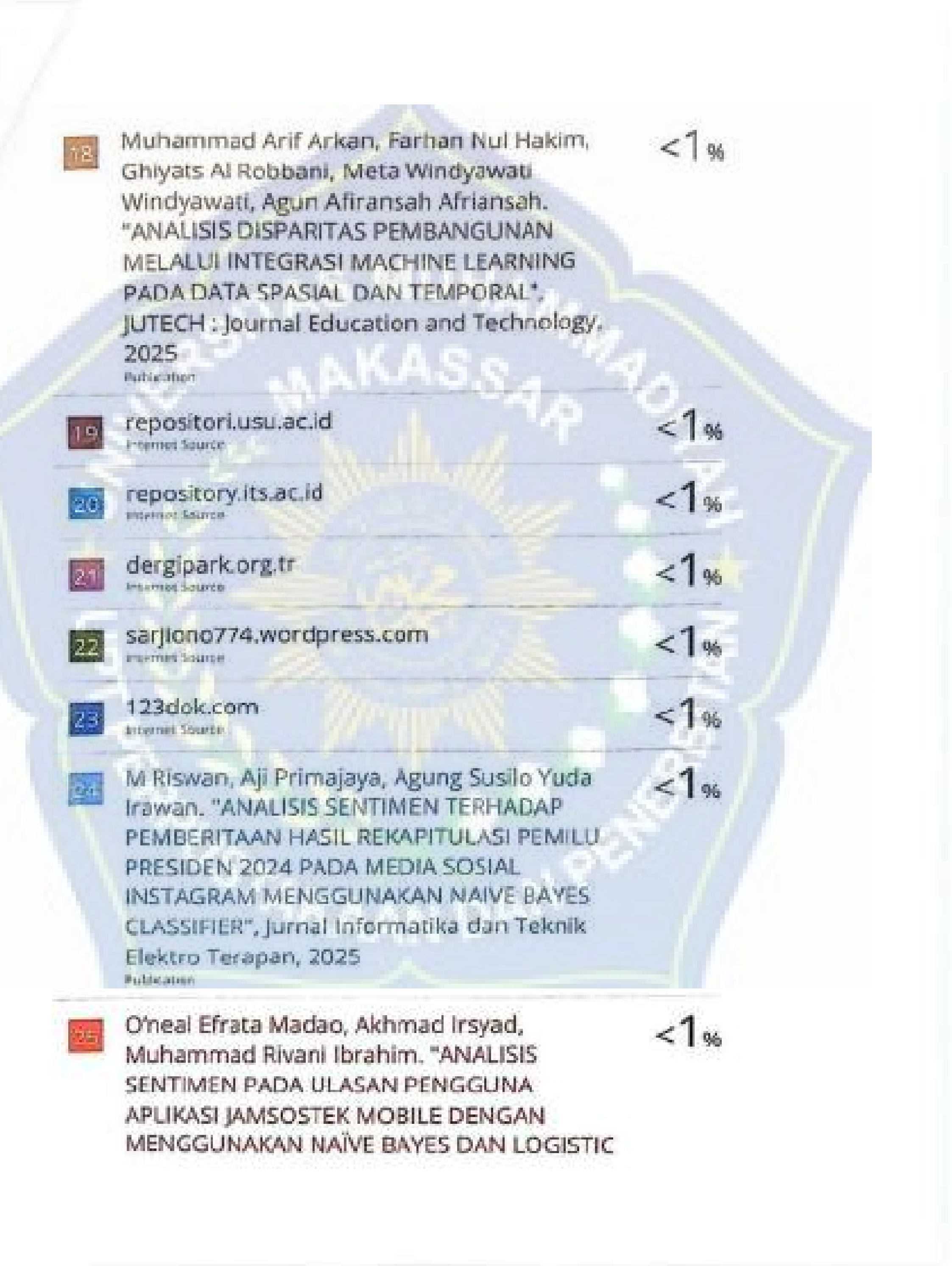
1	digilibadmin.unismuh.ac.id Internet Source	1%
2	toffeedev.com Internet Source	1%
3	Adhani Mulianti, Yulison Chrisnanto, Herdi Ashaury. "Optimizing Support Vector Machine Classification with SMOTE: Case Study of Alfagift Application User Reviews", Jurnal Pekommas, 2024 Publication	1%
4	ojs.unud.ac.id Internet Source	1%
5	Imam Kharits Najibulloh, Imam Tahyudin, Dhanar Intan Surya Saputra. "Analisis Sentimen Ulasan Co-Pilot Google Play dengan SVM, Neural Network, dan Decision Tree", Edumatic: Jurnal Pendidikan Informatika, 2025 Publication	1%
6	Yasir Hasan. "PENGUKURAN SILHOUETTE SCORE DAN DAVIES-BOULDIN INDEX PADA HASIL CLUSTER K-MEANS DAN DBSCAN", Jurnal Informatika dan Teknik Elektro Terapan, 2024 Publication	1%
7	ojs.uajy.ac.id Internet Source	

UNIVERSITAS MUHAMMADIYAH MAKASSAR
PERPUSTAKAAN DAN PENERBITAN



		1%
8	blog.myskill.id Internet Source	1%
9	dqlab.id Internet Source	1%
10	journal.lembagakita.org Internet Source	1%
11	lib.ui.ac.id Internet Source	1%
12	Muhammad Andi Hermawan, Ahmad Faqih, Gifthera Dwilestari. "IMPLEMENTASI AKURASI MODEL NAIVE BAYES MENGGUNAKAN SMOTE DALAM ANALISIS SENTIMEN PENGGUNA APLIKASI BRIMO", Jurnal Informatika dan Teknik Elektro Terapan, 2025 Publication	<1%
13	Mohammad Bayu Anggara. "Comparison of Naïve Bayes and SVM Methods in Sentiment Analysis of User Reviews on the RSUD AL IHSAN Mobile Application", Competitive, 2025 Publication	<1%
14	kc.umn.ac.id Internet Source	<1%
15	repository.uin-suska.ac.id Internet Source	<1%
16	journal.nurulfikri.ac.id Internet Source	<1%
17	mocca1479.blog.binusian.org Internet Source	<1%

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN



18 Muhammad Arif Arkan, Farhan Nul Hakim, Ghiyats Al Robbani, Meta Windyawati Windyawati, Agun Afiransah Afriansah. "ANALISIS DISPARITAS PEMBANGUNAN MELALUI INTEGRASI MACHINE LEARNING PADA DATA SPASIAL DAN TEMPORAL", JUTECH : Journal Education and Technology, 2025 <1%
Publication

19 repositori.usu.ac.id <1%
Internet Source

20 repository.its.ac.id <1%
Internet Source

21 dergipark.org.tr <1%
Internet Source

22 sarjiono774.wordpress.com <1%
Internet Source

23 123dok.com <1%
Internet Source

24 M Riswan, Aji Primajaya, Agung Susilo Yuda Irawan. "ANALISIS SENTIMEN TERHADAP PEMBERITAAN HASIL REKAPITULASI PEMILU PRESIDEN 2024 PADA MEDIA SOSIAL INSTAGRAM MENGGUNAKAN NAIVE BAYES CLASSIFIER", Jurnal Informatika dan Teknik Elektro Terapan, 2025 <1%
Publication

25 O'neal Efrata Madao, Akhmad Irsyad, Muhammad Rivani Ibrahim. "ANALISIS SENTIMEN PADA ULASAN PENGGUNA APLIKASI JAMSOSTEK MOBILE DENGAN MENGGUNAKAN NAÏVE BAYES DAN LOGISTIC <1%

	REGRETION", Djtechno: Jurnal Teknologi Informasi, 2025 Publication	
26	Relin Pramudiya, Aldo Kadafi, Daniel Udjulawa. "Analisis Sentimen Opini Publik terhadap Kasus Korupsi Timah di Youtube Menggunakan Metode Oversampling dan Algoritma Decision Tree", Arcitech: Journal of Computer Science and Artificial Intelligence, 2024 Publication	<1%
27	caraguna.com Internet Source	<1%
28	course-net.com Internet Source	<1%
29	docplayer.info Internet Source	<1%
30	eprints.uny.ac.id Internet Source	<1%
31	es.scribd.com Internet Source	<1%
32	fr.scribd.com Internet Source	<1%
33	Rizky Ichwan Purnama, Lutfan Hakim, Aryansyah Al Giffari, Faizal Andre Febrianto, Farizi Ilham. "Analisis Klaster Kekayaan Tokoh Terkaya Indonesia Tahun 2007 Menggunakan Algoritma K-Means", RIGGS: Journal of Artificial Intelligence and Digital Business, 2025 Publication	<1%

UNIVERSITAS MUHAMMADIYAH MAKASSAR
PERPUSTAKAAN DAN PENERBITAN

BAB III Syahril Akbar 105841103821

by Tahap Tutup

Submission date: 25-Aug-2025 08:42AM (UTC+0700)
Submission ID: 2734625454
File name: Skripsi_BAB_III_Syahril_Akbar_105841103821_Informatika.docx (397.78K)
Word count: 1515
Character count: 9899

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

BAB III Syahril Akbar 105841103821

ORIGINALITY REPORT

4%	4%	2%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	M Riswan, Aji Primajaya, Agung Susilo Yuda Irawan. "ANALISIS SENTIMEN TERHADAP PEMBERITAAN HASIL REKAPITULASI PEMILU PRESIDEN 2024 PADA MEDIA SOSIAL INSTAGRAM MENGGUNAKAN NAIVE BAYES CLASSIFIER", Jurnal Informatika dan Teknik Elektro Terapan, 2025 Publication	2%
2	docplayer.info Internet Source	2%

Exclude quotes Off
Exclude bibliography Off
Exclude matches < 2%

UPT PERPUSTAKAAN DAN PENERBITAN
MUHAMMADIYAH

BAB IV Syahril Akbar 105841103821

ORIGINALITY REPORT

7%	5%	6%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	kc.umn.ac.id Internet Source	1%
2	Dedi Irawan, Yayu Sri Rahayu, Mohamad Farozi. "KLASIFIKASI SENTIMEN PELANGGAN RESTORAN KOKI SUNDA MENGGUNAKAN ALGORITMA NAIVE BAYES", Jurnal Manajemen Informatika dan Sistem Informasi, 2025 Publication	1%
3	jpti.journals.id Internet Source	1%
4	jurnal.kdi.or.id Internet Source	1%
5	oda.oslomet.no Internet Source	<1%
6	rudyekoprasetya.wordpress.com Internet Source	<1%
7	repository.ar-raniry.ac.id Internet Source	<1%
8	Rizqia Lestika Atimi, Enda Esyudha Pratama. "Implementasi Model Klasifikasi Sentimen Pada Review Produk Lazada Indonesia", Jurnal Sains dan Informatika, 2022 Publication	<1%
9	download.garuda.ristekdikti.go.id Internet Source	<1%

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

10	Fauzan Baehaqi, Nuri Cahyono. "Analisis Sentimen Terhadap Cyberbullying Pada Komentar Di Instagram Menggunakan Algoritma Naïve Bayes", Indonesian Journal of Computer Science, 2024 Publication	<1%
11	repository.teknokrat.ac.id Internet Source	<1%
12	Shazifa Azhari, Nining Rahaningsih, Raditya Danar Dana, Mulyawan .. "PENINGKATAN AKURASI ANALISIS SENTIMEN PADA APLIKASI LOKLOK DENGAN METODE NAÏVE BAYES", Jurnal Informatika dan Teknik Elektro Terapan, 2025 Publication	<1%
13	Dimas Rahma Samudra, Jusuf Herlambang Herdiyana, Muhammad Fauzan, Aqiel Maulidan Zuhdi et al. "Klasterisasi Data Ekonomi dan Demografi Menggunakan K-Means: Studi Kasus Saham, Neraca Pembayaran, GDP, dan Kependudukan", RIGGS: Journal of Artificial Intelligence and Digital Business, 2025 Publication	<1%
14	Gregorius Guntur Sunardi Putra, Windra Swastika, Paulus Lucky Tirma Irawan. "Perbandingan Particle Swarm Optimization dengan Genetic Algorithm dalam Feature Selection untuk Analisis Sentimen pada Permendikbudristek PPKS-LPT", Jurnal Edukasi dan Penelitian Informatika (JEPIN), 2022 Publication	<1%

UNIVERSITAS MUHAMMADIYAH MAKASSAR
PERPUSTAKAAN DAN PENERBITAN

15	Indra Indra, Nahya Nur, Muh. Iqram, Nurul Inayah. "Perbandingan K-Means dan Hierarchical Clustering dalam Pengelompokan Daerah Beresiko Stunting", INOVTEK Polbeng - Seri Informatika, 2023 Publication	<1%
16	Ivan Aditya Nugraha, Irving Vitra Paputungan. "Analisis Sentimen Video Review Starlink Indonesia Menggunakan Pendekatan Lexicon Dictionary Bahasa Indonesia Dan Algoritma Random Forest", Innovative: Journal Of Social Science Research, 2025 Publication	<1%
17	Rizky Ichwan Purnama, Lutfan Hakim, Aryansyah Al Giffari, Faizal Andre Febrianto, Farizi Ilham. "Analisis Klaster Kekayaan Tokoh Terkaya Indonesia Tahun 2007 Menggunakan Algoritma K-Means", RIGGS: Journal of Artificial Intelligence and Digital Business, 2025 Publication	<1%
18	jurnal.untan.ac.id Internet Source	<1%
19	Rezqiwati Ishak, Nurmawanti Nurmawanti, Amiruddin Bengnga. "Optimization of K-Means in Disease Clustering of Pregnant Women Using Random Forest", Jambura Journal of Electrical and Electronics Engineering, 2025 Publication	<1%

UNIVERSITAS MUHAMMADIYAH MAKASSAR
PERPUSTAKAAN DAN PENERBITAN

BAB V Syahril Akbar 105841103821

by Tahap Tutup

Submission date: 25-Aug-2025 08:43AM (UTC+0700)
Submission ID: 2734626327
File name: Skripsi_BAB_V_Syahril_Akbar_105841103821_Informatika.docx (152.7K)
Word count: 520
Character count: 3565

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN

BAB V Syahril Akbar 105841103821

ORIGINALITY REPORT

LULUS

5%	5%	2%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	text-id.123dok.com Internet Source	3%
2	eprints.walisongo.ac.id Internet Source	2%

Exclude quotes Off
Exclude bibliography Off

Exclude matches Off

UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN