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

Syahril Akbar[1)], Muhammad Faisal*[2)], Rizki Yusliana Bakti[3)], Muhammad Syafaat[4)], Andi Makbul Syamsuri[5)], Muhyiddin AM Hayat[6)], Andi Lukman Anas[7)]

1. Informatika, Universitas Muhammadiyah Makassar
email: syahrilakbar63@gmail.com
2. Informatika, Universitas Muhammadiyah Makassar
email: muhfaisal@unismuh.ac.id
3. Informatika, Universitas Muhammadiyah Makassar
email: rizkiyusliana@unismuh.ac.id
4. Teknik Pengairan, Universitas Muhammadiyah Makassar
email: syafaat_skuba@unismuh.ac.id
5. Teknik Pengairan, Universitas Muhammadiyah Makassar
email: amakbulsyamsuri@unismuh.ac.id
6. Informatika, Universitas Muhammadiyah Makassar
email: muhyiddin@unismuh.ac.id
7. Informatika, Universitas Muhammadiyah Makassar
email: lukmananas@unismuh.ac.id

Abstract

Manual analysis of large-scale and unstructured textual feedback data is often inefficient and subjective, thereby hindering data-driven decision-making. This study aims to design and implement an integrated analytical workflow to automatically filter, cluster, and classify feedback data consisting of criticisms and suggestions. The research employs a hybrid approach that begins with TF-IDF-based data filtering, followed by dimensionality reduction using Latent Semantic Analysis (LSA), and topic clustering through K-Means clustering optimized with the Silhouette Score. The resulting cluster labels are then used as training data to build a Multinomial Naive Bayes classification model. The results show that this workflow successfully identified two main thematic clusters, namely "Criticism and Expectations" and "Suggestions and Compliments", and the classification model achieved an overall accuracy of 91%. Although class imbalance affected the recall of the minority class (47%), the model demonstrated high precision (95%) for that class. It is concluded that this hybrid approach effectively transforms raw data into structured insights, and utilizing clustering results as training data is an efficient strategy for automating feedback categorization, providing a reliable tool for institutional analysis.

***Kata Kunci** : Analisis Sentimen, TF-IDF, K-Means, Latent Semantic Analysis, NLP.*

A. PENDAHULUAN

Analisis sentimen telah menjadi pendekatan krusial untuk mengekstraksi wawasan bermakna dari data tekstual bervolume besar di era digital [1]. Bagi instansi pemerintah seperti Balai Besar Pengembangan SDM dan Penelitian (BBPSDMP) Kominfo Makassar, umpan balik dari kegiatan Monitoring dan Evaluasi (Monev) merupakan fondasi penting untuk perbaikan layanan. Namun, tingginya volume komentar yang tidak terstruktur, bersifat umum, dan sering kali ditulis dalam bahasa informal menjadi tantangan signifikan [2][3]. Proses analisis manual tidak hanya memakan waktu dan sumber daya, tetapi juga rentan terhadap bias subjektivitas, sehingga menghambat pengambilan keputusan berbasis data yang efektif.

Berdasarkan kajian yang dilakukan, penelitian ini bertujuan untuk merancang dan mengimplementasikan sebuah sistem otomatis yang mampu menyaring dan mengolah data kritik dan saran secara lebih efektif. Secara spesifik, penelitian ini berupaya menjawab pertanyaan-pertanyaan berikut: (1) Bagaimana menerapkan proses penyaringan data untuk memisahkan komentar yang relevan dan substantif dari yang tidak relevan berdasarkan bobot TF-IDF dan jumlah kata?; (2) Bagaimana menerapkan pendekatan hibrida topic modeling menggunakan *Latent Semantic Analysis* (LSA) dan *K-Means clustering* untuk mengelompokkan data 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?.

Sejalan dengan rumusan masalah, tujuan penelitian ini adalah: pertama, mengimplementasikan metode penyaringan data berbasis TF-IDF untuk memperoleh data komentar yang berkualitas. Kedua, menerapkan pendekatan hibrida LSA dan *K-Means* untuk mengelompokkan data kritik dan saran ke dalam klaster tematik yang relevan. Ketiga, membangun sebuah model klasifikasi berbasis *Multinomial Naive Bayes* yang dilatih menggunakan label hasil *clustering* untuk memprediksi kategori topik dari data baru secara otomatis.

Untuk mencapai tujuan tersebut, penelitian ini menerapkan pendekatan hibrida yang menggabungkan beberapa metode *machine learning* [4]. Tahap awal adalah penyaringan data menggunakan kriteria bobot *Term Frequency-Inverse Document Frequency* (TF-IDF) dan jumlah kata [5]. Selanjutnya, data yang relevan diolah dengan pendekatan *unsupervised learning* yang mengombinasikan *Latent Semantic Analysis* (LSA) untuk reduksi dimensi [6] dan *K-Means Clustering* untuk pengelompokan topik [7]. Hasil dari pengelompokan ini kemudian dimanfaatkan sebagai data latih berlabel untuk membangun model klasifikasi *supervised learning* menggunakan algoritma *Multinomial Naive Bayes*, yang terbukti efektif untuk klasifikasi teks [8][9].

Research gap yang diisi oleh studi ini adalah penerapan alur kerja analitis yang komprehensif dari hulu ke hilir pada data umpan balik institusional yang spesifik di Indonesia. Berbeda dengan penelitian sebelumnya yang sering kali cenderung berfokus pada (misalnya, hanya *clustering* atau hanya klasifikasi), penelitian ini mengintegrasikan penyaringan kuantitatif, *clustering* hibrida, dan klasifikasi otomatis dalam satu prototipe sistem. Kontribusi ilmiahnya terletak pada validasi empiris bahwa pemanfaatan hasil *unsupervised learning* sebagai data latih merupakan strategi yang efektif untuk

membangun model klasifikasi yang akurat tanpa memerlukan pelabelan manual yang intensif [10]. Dengan akurasi model mencapai 91%, penelitian ini memberikan bukti kuat akan relevansi dan utilitas praktis dari sistem yang diusulkan untuk mendukung peningkatan kualitas layanan publik.

B. METODE PENELITIAN

Penelitian ini menerapkan pendekatan kuantitatif dengan merancang alur kerja analitis terintegrasi untuk membangun sebuah sistem prototipe. Objek penelitian adalah data tekstual sekunder berupa 2.601 komentar kritik dan saran yang dikumpulkan melalui teknik dokumentasi dari basis data Monitoring dan Evaluasi (Monev) milik BBPSDMP Kominfo Makassar. Keseluruhan alur kerja penelitian ini, mulai dari pemuatan data hingga evaluasi model, divisualisasikan secara rinci melalui diagram alir pada Gambar 1.

Berikut adalah penjelasan rinci untuk setiap tahapan yang digambarkan dalam diagram alir tersebut:

1. Pemuatan dan Pra-pemrosesan Data

Tahap awal dimulai dengan memuat *dataset* dari dokumen bernama kritik_saran (format .xlsx) menggunakan pustaka Pandas. Data teks mentah ini kemudian melalui serangkaian proses pra-pemrosesan yang bertujuan untuk membersihkan dan mengubahnya menjadi format terstruktur yang siap diolah. Proses ini meliputi: (a) Konversi ke *String*, untuk memastikan konsistensi tipe data; (b) *Stemming*, menggunakan pustaka Sastrawi untuk mereduksi setiap kata ke bentuk dasarnya dan mengurangi variasi morfologis; (c) Penghapusan.

Gambar 1. Flowchart Perancangan Sistem

2. Penyaringan Data dan Analisis Klastering

Untuk memastikan kualitas analisis, data yang telah divektorisasi kemudian disaring berdasarkan kriteria relevansi kuantitatif. Sebuah komentar dianggap relevan jika memiliki skor total TF-IDF lebih dari 1 dan terdiri dari lebih dari 2 kata dipilih berdasarkan observasi awal untuk menyingkirkan komentar formalitas seperti “bagus”, “baik” atau “mantap” yang berdiri sendiri. Data yang tidak memenuhi kriteria ini dipisahkan dan tidak diikutsertakan dalam tahap analisis selanjutnya.

Data relevan yang telah tersaring kemudian dikelompokkan (*clustering*) menggunakan algoritma *K-Means* untuk

menemukan struktur atau topik laten. Sebelum klastering, reduksi dimensi diterapkan menggunakan *Latent Semantic Analysis* (LSA) yang diimplementasikan dengan algoritma *TruncatedSVD* untuk mengatasi dimensionalitas tinggi dari data TF-IDF. Kombinasi parameter terbaik, yaitu jumlah klaster (K) dan jumlah komponen SVD, ditentukan dengan mencari nilai *Silhouette Score* tertinggi. Berdasarkan pengujian, konfigurasi optimal yang diperoleh adalah 2 klaster dengan 50 komponen SVD.

3. Interpretasi, Pelabelan, dan Pelatihan Model

Makna dari setiap klaster yang terbentuk diinterpretasikan secara kualitatif dengan menganalisis kata-kata kunci (*top words*) yang paling sering muncul di dalamnya. Berdasarkan interpretasi, kedua klaster diberi label "Kritik dan Harapan" dan "Saran dan Pujian". Data yang telah dilabeli ini kemudian digunakan untuk melatih model klasifikasi *Multinomial Naive Bayes*. *Dataset* dibagi menjadi 80% data latih dan 20% data uji untuk memastikan evaluasi yang objektif.

4. Evaluasi dan Validasi Model

Kinerja model klasifikasi yang telah dilatih dievaluasi pada data uji menggunakan *Confusion Matrix*. Metrik performa utama seperti akurasi, presisi, *recall*, dan *F1-Score* dihitung untuk mengukur kemampuan model dalam mengklasifikasikan topik secara akurat. Sebagai validasi fungsional, model yang telah jadi diuji coba pada beberapa kalimat baru untuk mensimulasikan penggunaannya dalam aplikasi dunia nyata. Terakhir, model Naive Bayes dan *vectorizer* TF-IDF yang telah terlatih disimpan untuk penggunaan di masa depan tanpa perlu melakukan pelatihan ulang.

C. HASIL DAN PEMBAHASAN

Penelitian ini berhasil mengimplementasikan alur kerja analitis dari hulu ke hilir untuk mengubah data umpan balik mentah menjadi wawasan terstruktur. Hasil dari setiap tahapan, mulai dari penemuan topik hingga evaluasi model klasifikasi, disajikan dan dibahas di bawah ini.

1. Penemuan Topik Otomatis melalui Klastering Hibrida

Tahap awal analisis berfokus pada peningkatan kualitas data. Dari 2.601 komentar awal, proses penyaringan kuantitatif berbasis skor TF-IDF dan jumlah kata berhasil mengidentifikasi 2.307 (88,7%) komentar sebagai data yang relevan dan substantif. Langkah ini krusial karena memastikan bahwa analisis selanjutnya hanya berfokus pada masukan yang memiliki nilai informasional, sehingga mengurangi derau (*noise*) dari data yang bersifat terlalu umum atau formalitas.

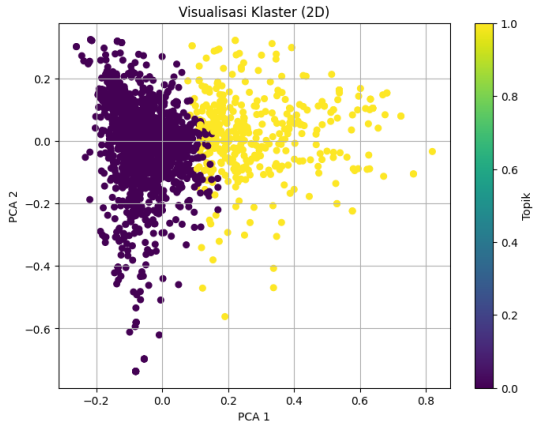
Selanjutnya, pendekatan *unsupervised learning* diterapkan pada data yang telah disaring. Penggunaan *Latent Semantic Analysis* (LSA) untuk reduksi dimensi sebelum pengelompokan dengan *K-Means* terbukti efektif. Validasi menggunakan *Silhouette Score* menunjukkan bahwa jumlah klaster optimal adalah dua (K=2), dengan skor tertinggi 0.1212.

Tabel 1. 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 1, nilai K=2 dipilih sebagai jumlah klaster optimal karena memberikan skor *Silhouette* tertinggi, meskipun nilai 0.1212 adalah yang tertinggi, skor ini mengindikasikan adanya tumpang tindih yang signifikan antar klaster, yang wajar terjadi pada data teks yang kompleks dan bernuansa, yang mengindikasikan bahwa data secara alami paling baik terbagi menjadi dua kelompok tematik.



Gambar 2. Visualisasi Hasil Klastering K-Means

Gambar 2 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.

Analisis kualitatif terhadap kata-kata kunci dengan bobot TF-IDF tertinggi pada setiap klaster berhasil mengungkap dua tema utama yang koheren dan dapat diinterpretasikan secara intuitif:

Tabel 2. Interpretasi dan Pelabelan Klaster

	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.
Label Final	**Kritik dan Harapan**	**Saran dan Pujian**

Temuan ini menunjukkan bahwa pendekatan hibrida LSA dan *K-Means* mampu melampaui pencocokan kata kunci sederhana untuk menangkap struktur tematik laten dalam data. Keberhasilan ini menjadi fondasi penting, karena memungkinkan pelabelan data secara otomatis dan mengubah masalah dari yang sepenuhnya tidak terstruktur menjadi masalah klasifikasi yang terstruktur.

2. Kinerja dan Implikasi Model Klasifikasi

Dengan memanfaatkan label yang dihasilkan dari tahap *clustering*, sebuah model klasifikasi *Multinomial Naive Bayes* dibangun dan dievaluasi. Model ini mencapai tingkat akurasi global yang tinggi sebesar 91% pada data uji,

menunjukkan kemampuannya yang kuat dalam menggeneralisasi pola dari data latih. Namun, analisis yang lebih mendalam pada *Confusion Matrix* mengungkap dampak signifikan dari ketidakseimbangan kelas dalam *dataset* (rasio sekitar 5:1 antara kelas "Kritik dan Harapan" dan "Saran dan Pujian").

Tabel 3. 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

Tabel 4. 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**

Dampak ini paling terlihat pada metrik *recall*. Model menunjukkan *recall* yang sangat tinggi (99%) untuk kelas mayoritas ("Kritik dan Harapan"), namun *recall* yang jauh lebih rendah (47%) untuk kelas minoritas ("Saran dan Pujian"). Ini berarti model sangat efektif dalam mengidentifikasi hampir semua kritik, tetapi melewatkan lebih dari separuh saran atau pujian. Fenomena ini merupakan konsekuensi klasik dari pelatihan pada data yang tidak seimbang, di mana model menjadi lebih bias terhadap kelas yang lebih sering muncul.

Meskipun demikian, temuan yang paling signifikan secara praktis adalah nilai presisi yang sangat tinggi (95%) untuk kelas minoritas. Ini mengimplikasikan bahwa meskipun model tidak menangkap semua saran atau pujian, setiap komentar yang berhasil diklasifikasikan sebagai "Saran dan Pujian" memiliki kemungkinan 95% benar. Dari perspektif institusional, hal ini sangat berharga. Tim dapat secara langsung menindaklanjuti umpan balik yang teridentifikasi ini dengan keyakinan tinggi tanpa memerlukan verifikasi manual, sehingga menciptakan sistem penyaringan yang efisien untuk masukan yang bersifat konstruktif dan positif. Dengan demikian, penelitian ini tidak hanya membuktikan efektivitas alur kerja yang diusulkan tetapi juga memberikan wawasan tentang bagaimana menafsirkan kinerja model dalam kondisi data dunia nyata yang tidak ideal.

D. KESIMPULAN DAN SARAN

Penelitian ini berhasil mendemonstrasikan sebuah alur kerja analitis terintegrasi yang secara efektif mengubah data umpan balik mentah dan tidak terstruktur menjadi wawasan yang dapat ditindaklanjuti. Dengan menerapkan penyaringan data berbasis TF-IDF dan pengelompokan topik hibrida menggunakan *Latent Semantic Analysis* (LSA) dan K-Means, sistem mampu mengidentifikasi dua klaster tematik yang koheren: "Kritik dan Harapan" dan "Saran dan Pujian". Pembangunan model klasifikasi *Multinomial Naive Bayes* yang dilatih menggunakan label hasil *clustering* ini mencapai akurasi global sebesar 91%. Temuan ini secara langsung menjawab

rumusan masalah dengan membuktikan bahwa pendekatan hibrida ini mampu menemukan struktur tematik dan bahwa hasil *unsupervised learning* dapat dimanfaatkan secara efisien sebagai data latih untuk membangun model klasifikasi yang andal, yang ditunjukkan oleh presisi tinggi (95%) pada kelas minoritas.

Meskipun berhasil, penelitian ini memiliki keterbatasan, terutama terkait dampak ketidakseimbangan kelas pada metrik *recall* (47% untuk kelas minoritas) dan ketergantungan pada representasi fitur *bag-of-words* (TF-IDF). Oleh karena itu, penelitian selanjutnya disarankan untuk fokus pada penanganan ketidakseimbangan kelas melalui teknik *resampling* seperti SMOTE. Selain itu, eksplorasi model yang lebih canggih seperti LSTM atau *Transformer* (misalnya, IndoBERT) yang mampu menangkap konteks sekuensial dapat meningkatkan kinerja klasifikasi secara signifikan. Mengembangkan sistem lebih lanjut untuk melakukan analisis sentimen *granular* (positif, negatif, netral) di dalam setiap topik juga akan memberikan wawasan yang lebih kaya dan mendalam.

E. REFERENSI

[1] 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

[2] Egger, R., & Yu, J. (2022). A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. Frontiers in Sociology, 7, 886498. https://doi.org/10.3389/fsoc.2022.886498

[3] Khomsah, S., & Aribowo, A. S. (2020). Model text-preprocessing komentar YouTube dalam bahasa Indonesia. Masa Berlaku Mulai, 1(3), 648–654.

[4] Hanan, D. A., Husodo, A. Y., & Rassy, R. P. (2025). Sentiment study of ChatGPT on Twitter data with hybrid K-Means and LSTM. Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer, 24(2). https://doi.org/10.30812/matrik.v24i2.4791

[5] Zikrina, S. A., & Fitriyani. (2025). Advancing hate speech detection in Indonesian language using graph neural networks and TF-IDF. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 9(1), 137–145. https://doi.org/10.29207/resti.v9i1.6179

[6] García-Jurado, A., Pérez-Barea, J. J., & Nova, R. (2021). A new approach to social entrepreneurship: A systematic review and meta-analysis. Sustainability, 13(5), 2754. https://doi.org/10.3390/su13052754

[7] Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. Electronics, 9(8), 1295. https://doi.org/10.3390/electronics9081295

[8] Fakhrezi, M. F., Rochim, A. F., & Nugraheni, D. M. K. (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

[9] Killamsetty, K., Sivasubramanian, D., Ramakrishnan, G., & Iyer, R. (2021). GLISTER: Generalization based data subset selection for efficient and robust learning. ArXiv preprint arXiv:2102.10423. Retrieved from https://medium.com/syncedreview/the-staggering-cost-of-

[10] 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