



Integrasi Pipeline Big Data NoSQL dan Pemrosesan Bahasa Alami untuk Manajemen Kualitas Berbasis Voice of Customer

Violita Anggraini^{1*}, Muhammad Iqbal Faturohman², Lathiifah Thawafani³

^{1,2}Teknik Industri, Universitas Telkom Purwokerto, Indonesia

³Teknik Industri, Universitas Binus Semarang, Indonesia

Email: violitaanggraini7@gmail.com¹, iqbalfaturohman@telkomuniversity.ac.id²,
lathiifah.thawafani@binus.ac.id³

*Penulis Korespondensi: violitaanggraini7@gmail.com

Abstract. The exponential growth of customer review data on e-commerce platforms generates high volume, variety, and velocity of data, necessitating efficient Big Data Analytics approaches beyond the constraints of traditional relational databases. This study aims to analyze product review sentiments to identify customer satisfaction patterns and inconsistencies between star ratings and textual reviews, while evaluating the efficiency of NoSQL databases in managing such non-relational data structures. The method employed involves text mining and Natural Language Processing (NLP), applying data preprocessing, Term Frequency-Inverse Document Frequency (TF-IDF) feature extraction with bigram-with-punctuation modeling, and text classification using the TextBlob library. The unstructured review data is managed using a NoSQL MongoDB database cluster with a scale-out architecture. The results indicate that integrating symbol-only word filtration significantly improves keyword relevance, and n-gram modeling successfully captures negation contexts like "not good". Word cloud visualizations effectively extract dominant keywords in negative reviews (Score 1), highlighting product quality criticisms ("taste", "product") and origin concerns ("made in"). Furthermore, sentiment distribution inconsistencies were detected, wherein low-rating reviews scores 1 and 2 still exhibited 15.1% and 29.2% positive sentiment, respectively. The implications of this research provide strategic recommendations for e-commerce executive management to integrate real-time sentiment monitoring systems for data-driven decision-making, proactive product development, and customer service optimization.

Keywords: Basis NoSQL; Big Data; Natural Language; Sentimen Pelanggan; Text Mining.

Abstrak. Pertumbuhan eksponensial data ulasan pelanggan pada platform e-commerce menghasilkan volume data yang besar, bervariasi, dan cepat, sehingga memerlukan pendekatan Big Data Analytics yang efisien di luar batasan basis data relasional tradisional. Penelitian ini bertujuan untuk menganalisis sentimen ulasan produk guna mengidentifikasi pola kepuasan pelanggan dan ketidaksesuaian antara rating bintang dengan teks ulasan, sekaligus mengevaluasi efisiensi penggunaan basis data NoSQL dalam mengelola struktur data non-relasional tersebut. Metode yang digunakan adalah text mining dan Natural Language Processing (NLP) dengan menerapkan teknik pembersihan data, ekstraksi fitur berbasis Term Frequency-Inverse Document Frequency (TF-IDF) menggunakan pemodelan frasa dua kata, serta klasifikasi menggunakan library TextBlob. Data ulasan dikelola menggunakan kluster basis data NoSQL MongoDB dengan arsitektur scale-out. Hasil penelitian menunjukkan bahwa integrasi pembersihan kata berbasis simbol meningkatkan relevansi kata kunci sentimental, dan pemodelan n-gram mampu mendeteksi konteks negatif secara akurat seperti frasa "not good". Visualisasi word cloud mengekstrak kata kunci dominan pada ulasan negatif (Skor 1) seperti kritik kualitas produk ("taste", "product") dan asal produk ("made in"). Selain itu, ditemukan inkonsistensi distribusi sentimen di mana ulasan dengan skor rendah 1 dan 2 masih memuat sentimen positif sebesar 15,1% dan 29,2%. Implikasi dari penelitian ini memberikan rekomendasi strategis bagi manajemen puncak e-commerce untuk mengintegrasikan sistem pemantauan sentimen real-time guna pengambilan keputusan berbasis data, peningkatan kualitas layanan, serta perbaikan spesifikasi produk secara proaktif.

Kata kunci: Basis NoSQL; Big Data; Natural Language; Sentimen Pelanggan; Text Mining.

1. LATAR BELAKANG

Dalam era digitalisasi industri yang dinamis, lanskap persaingan sektor e-commerce tidak lagi hanya berfokus pada strategi harga, melainkan bergeser ke arah rekayasa kualitas layanan dan produk secara berkelanjutan. Penjaminan mutu modern mengadopsi prinsip Total

Quality Management (TQM) yang menekankan pentingnya mendengarkan "Suara Konsumen" (Voice of Customer / VoC) untuk mereduksi cacat produk dan mengoptimalkan kinerja rantai pasok (Purba & Aisyah, 2017). Secara tradisional, data VoC dikumpulkan melalui kuesioner terstruktur dengan rating skala Likert (rating bintang), namun rating numerik tersebut sering kali gagal menangkap derajat sentimen emosional dan konteks spesifik keluhan pelanggan secara mendalam (Kim & Lim, 2021). Pertumbuhan eksponensial platform e-commerce memicu ledakan data ulasan berupa teks tidak terstruktur yang memiliki karakteristik Big Data, yaitu volume yang sangat besar (volume), keragaman format data (variety), dan kecepatan pembaruan yang sangat tinggi (velocity) (Mohammadi et al., 2024).

Dalam mengekstrak *value* dari tumpukan data teks pada MongoDB tersebut, diperlukan pendekatan *Big Data Analytics* melalui metode *Natural Language Processing* (NLP) dan analisis sentimen (Chioma Ann Udeh et al., 2024; Goyal & Malviya, 2023; Jamarani et al., 2024; Wamba et al., 2017). Analisis sentimen berbasis teks bertindak sebagai alat kontrol kualitas digital otomatis untuk mengklasifikasikan opini konsumen menjadi kategori positif atau negatif (Lingala, 2025). Melalui ekstraksi fitur berbasis *Term Frequency-Inverse Document Frequency* (TF-IDF) dan pemodelan frasa *n-gram* (khususnya *bigram*), sistem dapat mengenali konteks negasi keluhan mutu seperti frasa "not good", yang kerap kali terlewat jika hanya menggunakan analisis per kata tunggal (*unigram*) (Makkar et al., 2024).

Meskipun penelitian mengenai analisis sentimen telah banyak dilakukan, sebagian besar studi terdahulu hanya berfokus pada akurasi algoritma komputasi tanpa mengaitkannya dengan metrik rekayasa kualitas industri dan anomali perilaku konsumen. Pada kenyataannya, sering kali terjadi inkonsistensi data di mana pelanggan memberikan rating bintang yang sangat rendah (Skor 1 atau 2) namun teks ulasannya justru mengandung elemen sentimen positif, atau sebaliknya. Oleh karena itu, penelitian ini bertujuan untuk merancang sebuah kerangka kerja analisis sentimen terintegrasi yang menggabungkan kapabilitas pemrosesan bahasa alami (NLP) berbasis *bigram* dengan infrastruktur basis data NoSQL MongoDB (Mai et al., 2024; Wong, 2022; Zaki et al., 2022). Melalui pendekatan ini, inkonsistensi hubungan antara rating numerik dengan sentimen tekstual akan dibedah secara mendalam. Hasil akhir dari penelitian ini diharapkan dapat memberikan kontribusi teoritis dan praktis yang signifikan bagi keilmuan Teknik Industri, khususnya dalam memformulasikan strategi *Customer-Driven Quality Management* yang responsif, adaptif, dan berbasis data di industri *e-commerce* modern.

2. KAJIAN TEORITIS

Quality Management (TQM) dan Voice of Customer (VoC) dalam Industri Jasa

Dalam pilar Teknik Industri, penjaminan mutu tidak lagi hanya berfokus pada inspeksi produk akhir di lini manufaktur, melainkan telah bermutasi menjadi pendekatan holistik yang berpusat pada kepuasan pelanggan melalui *Total Quality Management* (TQM). Salah satu instrumen paling krusial dalam TQM adalah penangkapan "Suara Konsumen" (*Voice of Customer* / VoC) yang bertindak sebagai metrik dasar untuk mengevaluasi performa sistem penyampaian jasa dan kualitas produk. Secara tradisional, VoC diukur menggunakan pendekatan konvensional berskala Likert melalui kuesioner (Mittal & Gupta, 2021; Shahin et al., 2024; Torres & Neto, 2018). Namun, dalam ekosistem *e-commerce*, data VoC beralih menjadi ulasan teks tidak terstruktur berkapasitas besar. Menurut (Kim & Lim, 2021), kegagalan menangkap esensi emosional dan keluhan spesifik di balik rating numerik dapat menyebabkan bias dalam pengambilan keputusan pengendalian kualitas.

Infrastruktur Basis Data NoSQL (MongoDB)

Sistem informasi industri konvensional umumnya mengandalkan *Relational Database Management System* (RDBMS) dengan skema tabel terstruktur yang kaku. Ketika dihadapkan pada jutaan baris data ulasan teks tidak terstruktur milik *e-commerce*, RDBMS sering kali mengalami kendala latensi tinggi dan keterbatasan dalam melakukan skalabilitas horizontal (*scale-out*) (Jose & Abraham, 2017; Song et al., 2020). Solusi arsitektur untuk mengatasi batasan ini adalah penggunaan basis data NoSQL (Not Only SQL) berbasis dokumen, seperti MongoDB. MongoDB menyimpan data dalam format semi-terstruktur BSON (*Binary JSON*) yang memberikan fleksibilitas skema tanpa memerlukan definisi relasi yang kaku di awal (Sridhar & Santhosh Kumar, 2025).

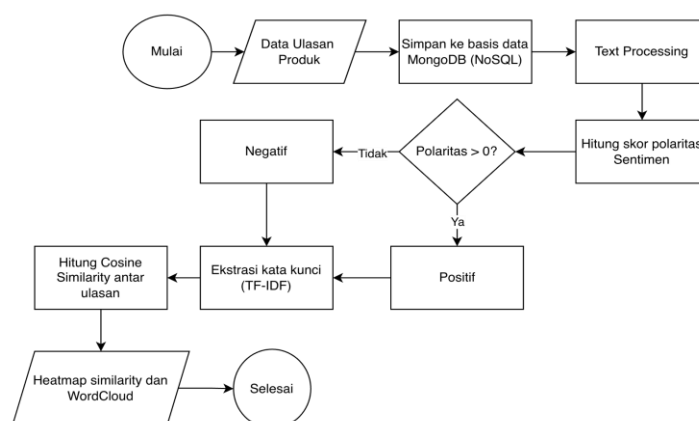
Berdasarkan tinjauan, dapat diidentifikasi adanya research gap yang relevan untuk dijawab. Penelitian-penelitian NLP terdahulu umumnya berfokus pada optimasi algoritmik klasifikasi sentimen tanpa mempertimbangkan skalabilitas arsitektur penyimpanan data, sementara penelitian manajemen kualitas konvensional belum banyak mengintegrasikan analitika teks tidak terstruktur ke dalam kerangka pengambilan keputusan operasional. Penelitian ini berupaya menjembatani kesenjangan tersebut dengan mengintegrasikan basis data NoSQL MongoDB sebagai fondasi penyimpanan data skala besar, teknik NLP untuk ekstraksi kata kunci dan sentimen, serta cosine similarity untuk pemetaan kemiripan ulasan, sebagai satu kesatuan pipeline analitika yang berorientasi pada kebutuhan quality engineering dan customer service di lingkungan industri berbasis data. Perbandingan posisi penelitian ini terhadap studi terdahulu disajikan pada Tabel 1.

Tabel 1. Tinjauan Pustaka.

Penelitian	Sumber Data	Metode Utama	Fokus Analisis
(Apria Ananda Anam et al., 2025)	Ulasan daring terstruktur	Ekstraksi aspek berbasis data	Identifikasi aspek produk dominan
(Sabbeh & Fasihuddin, 2023)	Social Media (Twitter or Facebook)	BiLSTM dan CNN	Klasifikasi sentimen biner
(Atlas et al., 2025)	Ulasan produk daring	BiGRU and RNN	Prediksi sentimen dan permintaan produk
(Kim & Lim, 2021)	Data survei terstruktur, SPC	Statistical Process Control, Six Sigma	Pengendalian kualitas berbasis data numerik
Penelitian ini	Ulasan Amazon berskala besar, teks tidak terstruktur	Pipeline Big Data: NoSQL MongoDB, NLP, bigram, cosine similarity	Kecerdasan pelanggan terintegrasi untuk quality engineering dan customer service

3. METODE PENELITIAN

Penelitian ini menerapkan sebuah data pipeline big data yang terdiri atas empat tahap berurutan yang dapat dilihat pada Gambar xx. Alur metodologi ini dirancang agar data ulasan tidak terstruktur dapat diproses secara sistematis, mulai dari sumber data mentah hingga menghasilkan wawasan (*insight*) yang dapat ditindaklanjuti secara manajerial.



Gambar 1. Alur Pengolahan Data.

Akuisisi Data

Data bersumber dari ulasan produk Amazon, meliputi teks ulasan, ringkasan, rating (skala 1–5), dan identitas produk. Data dikelompokkan berdasarkan kategori produk dan rating untuk analisis komparatif antarkategori.

Penyimpanan Data (MongoDB)

Data disimpan menggunakan MongoDB karena sifatnya yang semi-terstruktur/tidak terstruktur dengan variasi panjang dan format teks yang tinggi. Struktur dokumen BSON memungkinkan penyimpanan tanpa skema tetap, sementara arsitektur scale-out mendukung skalabilitas horizontal seiring pertambahan volume data.

Text Processing Menggunakan NLP

Data teks mentah yang tersimpan pada MongoDB melalui tahap text preprocessing untuk menghilangkan derau (noise) dan menstandarisasi representasi teks. Tahapan ini meliputi tokenization (pemecahan teks menjadi unit kata/token), stopwords removal (penghapusan kata umum yang tidak memberi kontribusi makna signifikan, seperti kata sambung dan kata depan), serta lemmatization (penyederhanaan kata ke bentuk dasarnya untuk mengurangi redundansi variasi morfologis) (Khainur et al., 2025). Selain itu, dilakukan penghapusan tanda baca dan kata yang hanya terdiri atas simbol (symbol-only words), serta penerapan teknik penggabungan kata negasi dengan kata sifat yang mengikutinya, sehingga konteks sentimen negatif yang biasanya hilang pada pemrosesan unigram konvensional dapat dipertahankan secara eksplisit (Chai, 2023).

Analisis Sentimen dan Text Mining

Analisis sentimen berbasis leksikon dilakukan pada ringkasan ulasan: polaritas > 0 dikategorikan positif, ≤ 0 negatif. Hasil klasifikasi dibandingkan dengan rating numerik untuk mengidentifikasi inkonsistensi sentimen tekstual vs. penilaian kuantitatif. Ekstraksi kata kunci menggunakan TF-IDF pada representasi bigram untuk menangkap frasa kontekstual (misal "not great"), divisualisasikan sebagai word cloud per kategori rating. Kemiripan semantik antarulasan pada produk dengan ulasan terbanyak diukur menggunakan cosine similarity terhadap vektor TF-IDF:

$$\cos(\theta) = (A \cdot B) / (\|A\| \|B\|) \quad (1)$$

dengan $A \cdot B$ merupakan hasil kali titik antara vektor bobot TF-IDF dua ulasan, sedangkan $\|A\|$ dan $\|B\|$ masing-masing merupakan norma Euclidean dari vektor A dan B (Sahu & Khandekar, 2020). Nilai cosine similarity berkisar antara 0 hingga 1, dengan nilai 1 menunjukkan kesamaan sempurna dan nilai 0 menunjukkan tidak adanya kemiripan. Hasil perhitungan disusun ke dalam matriks kesamaan (similarity matrix) yang kemudian divisualisasikan dalam bentuk heatmap untuk mengidentifikasi kluster ulasan dengan topik atau sentimen yang serupa (Habibi, 2019).

4. HASIL DAN PEMBAHASAN

Bagian ini menyajikan temuan utama dari penerapan pipeline big data pada data ulasan produk Amazon, meliputi hasil ekstraksi kata kunci berbasis word cloud, hasil analisis kemiripan ulasan menggunakan cosine similarity, hasil analisis sentimen dibandingkan dengan skor rating, serta diskusi implikasi manajerial dari temuan tersebut bagi fungsi manajemen kualitas dan layanan pelanggan.

Ekstraksi Kata Kunci Dominan Berbasis Word Cloud dan Bigram

Hasil visualisasi word cloud untuk setiap kategori rating (1 hingga 5 bintang) menunjukkan pola kata dominan yang berbeda secara jelas antarkategori, sebagaimana ditampilkan pada Gambar 1. Ulasan dengan rating tinggi (4 dan 5 bintang) didominasi oleh kata-kata deskriptif positif yang merepresentasikan kepuasan pelanggan terhadap rasa, kualitas, dan pengalaman penggunaan produk, sedangkan ulasan dengan rating rendah (1 dan 2 bintang) didominasi oleh kata-kata yang mengindikasikan ketidakpuasan, seperti kritik terhadap rasa, tekstur, dan kesesuaian produk dengan deskripsi. Pola gradasi tersebut secara konsisten terlihat dari rating 1 menuju rating 5, yang mengonfirmasi bahwa representasi frekuensi kata pada word cloud mampu menangkap pergeseran sentimen pelanggan secara visual dan intuitif.



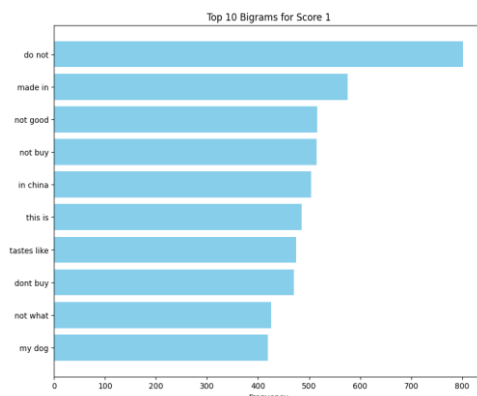
Gambar 2. Word Cloud Kata Kunci Dominan Untuk Setiap Kategori Rating Ulasan (Bintang 1 Sampai Dengan Bintang 5).

Meskipun word cloud pada Gambar 1 telah mampu menunjukkan kata-kata yang paling sering muncul pada setiap kategori rating, pendekatan ekstraksi berbasis unigram tersebut memiliki keterbatasan mendasar, yaitu hilangnya konteks negasi. Kata sifat seperti "good" atau "great" dapat muncul secara dominan pada ulasan ber-rating rendah meskipun sesungguhnya didahului oleh kata negasi "not", sehingga interpretasi sentimen menjadi menyesatkan apabila hanya mengandalkan frekuensi kata tunggal. Untuk mengatasi keterbatasan tersebut, dilakukan penggabungan kata negasi dengan kata sifat yang mengikutinya menjadi satu unit bigram (misalnya "not_good", "not_great") sebelum proses ekstraksi kata kunci dilakukan kembali secara khusus pada ulasan dengan rating ekstrem, yaitu bintang 5 (paling positif) dan bintang 1 (paling negatif), sebagaimana ditunjukkan pada Gambar 2.



Gambar 3. Perbandingan Word Cloud Negation-Aware Pada Ulasan Bintang 5 (Positif) Dan Bintang 1 (Negatif).

Sebagai pelengkap analisis kata kunci, dilakukan pula ekstraksi bigram menggunakan pembobotan TF-IDF disertai penghapusan symbol-only words, yang menghasilkan distribusi frekuensi kata kunci yang lebih bersih pada ulasan bersentimen negatif, sebagaimana ditunjukkan pada Gambar 3.



Gambar 4. Distribusi Frekuensi Kata Kunci Dominan Pada Ulasan Dengan Skor Rating 1 (Sentimen Negatif).

Kata kunci dominan pada kategori rating 1, seperti "don't", "taste", "product", "made in", dan "not good", dapat dikelompokkan ke dalam beberapa tema kritik pelanggan sebagaimana dirangkum pada Tabel 2. Kata-kata negasi secara langsung mengindikasikan ketidakpuasan eksplisit terhadap produk. Kata "taste" dan "product" yang muncul berdampingan dengan sentimen negatif mengindikasikan kritik terhadap atribut kualitas dan karakteristik produk.

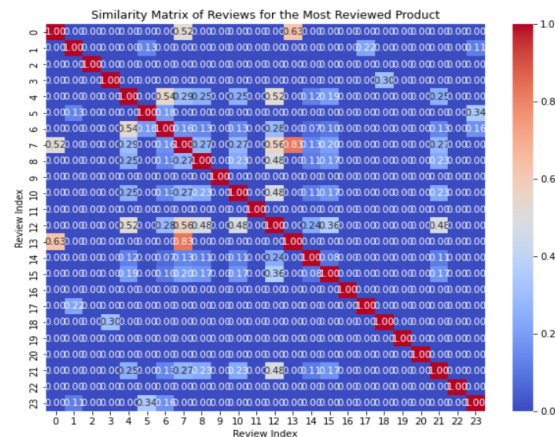
Tabel 2. Ringkasan Tema Kata Kunci Dominan Pada Ulasan Bersentimen Negatif (Rating 1).

Tema Kritik	Kata Kunci Representatif	Implikasi Manajerial
Penolakan/ketidakpuasan eksplisit	don't, not buy, not good, not what	Sinyal langsung bagi customer service untuk prioritas penanganan keluhan
Kritik atribut kualitas produk	taste, product, this is	Input bagi quality engineering untuk investigasi root cause
Kekhawatiran asal produksi	made in, in china	Pertimbangan bagi manajemen rantai pasok dan sourcing
Evaluasi negatif menyeluruh	dont buy, my dog	Masukan bagi tim pemasaran dan pengembangan produk

Kemiripan Ulasan Berbasis Cosine Similarity

Analisis cosine similarity dilakukan terhadap kumpulan ulasan pada produk dengan jumlah ulasan terbanyak. Hasil matriks kesamaan yang divisualisasikan dalam bentuk heatmap pada Gambar 4 menunjukkan bahwa nilai kesamaan pada elemen diagonal matriks selalu bernilai 1, karena merepresentasikan kesamaan setiap ulasan terhadap dirinya sendiri. Warna yang lebih gelap pada sel-sel di luar diagonal mengindikasikan tingkat kemiripan yang lebih

tinggi antara dua ulasan, sehingga area dengan warna gelap yang mengelompok dapat diinterpretasikan sebagai kluster ulasan yang membahas topik atau sentimen yang serupa.



Gambar 5. Heatmap Matriks Cosine Similarity Antarulasan Pada Produk Dengan Jumlah Ulasan Terbanyak.

Identifikasi kluster kemiripan tersebut memiliki nilai praktis yang tinggi bagi organisasi, karena memungkinkan tim customer service mengelompokkan ulasan-ulasan dengan pola keluhan yang serupa ke dalam satu kategori penanganan, alih-alih memproses setiap ulasan secara individual. Sebaliknya, area dengan tingkat kemiripan rendah mengindikasikan keragaman opini atau topik yang lebih luas dalam kumpulan ulasan, yang dapat menjadi sinyal bagi tim quality engineering untuk menelusuri lebih lanjut variasi permasalahan yang dihadapi pelanggan pada produk tersebut.

Konsistensi antara Sentimen dan Skor Rating

Perbandingan antara hasil klasifikasi sentimen berbasis leksikon dengan skor rating numerik pada setiap ulasan menunjukkan adanya proporsi tertentu ulasan yang mengalami inkonsistensi, yaitu ulasan dengan skor rating rendah (1 atau 2) namun memiliki polaritas sentimen tekstual yang positif, maupun sebaliknya. Distribusi sentimen pada setiap kategori rating divisualisasikan dalam bentuk diagram lingkaran sebagaimana ditunjukkan pada Gambar 5.

```
for score, dataframe in dataframes.items():
    if not dataframe.empty:
        # count positive and negative number
        sentiment_counts = dataframe['Sentiment'].value_counts()

        # Define color, ensuring "Negative" and "Positive"
        colors = ['#FF6242', '#6485F6']
        labels = ['Negative', 'Positive']

        if sentiment_counts.index[0] == 'Positive':
            colors = ['#6485F6', '#FF6242']

        # PIE CHART
        plt.figure(figsize=(6, 6))
        plt.pie(sentiment_counts, labels=sentiment_counts.index, autopct='%1.1f%%', startangle=140, colors=colors)
        plt.title(f'Sentiment Distribution for Score {score}')
        plt.show()
```

Gambar 6. Distribusi Proporsi Sentimen Positif Dan Negatif Pada Setiap Kategori Skor Rating.

Temuan inkonsistensi tersebut mengindikasikan bahwa skor rating numerik semata tidak selalu merepresentasikan sentimen aktual pelanggan secara akurat. Fenomena ini dapat disebabkan oleh berbagai faktor, seperti pemberian rating yang dipengaruhi oleh pengalaman non-produk (misalnya keterlambatan pengiriman), penggunaan bahasa sarkastik dalam ulasan, atau keterbatasan pendekatan sentimen berbasis leksikon dalam menangkap nuansa konteks kalimat. Temuan ini menegaskan pentingnya analisis teks ulasan sebagai pelengkap terhadap skor rating dalam proses pengambilan keputusan berbasis data pelanggan.

Dari perspektif manajemen industri, hasil analisis kualitatif yang dihasilkan melalui pipeline big data ini memberikan kontribusi signifikan sebagai input bagi fungsi quality engineering dan customer service. Tema-tema kritik yang teridentifikasi melalui ekstraksi bigram, seperti keluhan terhadap kualitas rasa, atribut produk, dan asal produksi, dapat digunakan sebagai dasar penyusunan kategori akar penyebab (root cause) dalam kerangka perbaikan kualitas produk secara berkelanjutan.

5. KESIMPULAN DAN SARAN

Penelitian ini berhasil merancang pipeline big data yang mengintegrasikan MongoDB dan NLP untuk mengekstraksi customer intelligence dari ulasan produk Amazon. Hasilnya menunjukkan bahwa pendekatan berbasis bigram dan cosine similarity mampu membedakan pola kata kunci ulasan positif-negatif, mengidentifikasi kluster ulasan yang berkaitan secara semantik, serta mendeteksi inkonsistensi antara sentimen tekstual dan skor rating. Temuan ini menjembatani kesenjangan riset antara manajemen kualitas konvensional dan literatur NLP terkait skalabilitas arsitektur data di lingkungan industri. Penelitian ini memiliki tiga keterbatasan: analisis hanya mencakup satu kategori produk sehingga generalisasi perlu validasi lanjutan; pendekatan sentimen berbasis leksikon kurang akurat menangkap konteks kompleks seperti sarkasme dibanding model machine learning; dan perhitungan cosine similarity masih bersifat batch, belum mengakomodasi kebutuhan real-time.

DAFTAR REFERENSI

- Apria Ananda Anam, P., Abdul Fatah, D., & Ali Syakur, M. (2025). Analisis sentimen *review* aplikasi Zalora di *Google Play Store* menggunakan metode *Support Vector Machine*. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 9(2), 2443–2450. <https://doi.org/10.36040/jati.v9i2.13083>
- Atlas, L. G., Arockiam, D., Muthusamy, A., Balusamy, B., Selvarajan, S., Al-Shehari, T., & Alsadhan, N. A. (2025). A modernized approach to sentiment analysis of product reviews using BiGRU and RNN based LSTM deep learning models. *Scientific Reports*, 15(1), 1–24. <https://doi.org/10.1038/s41598-025-01104-0>

- Chai, C. P. (2023). Comparison of text preprocessing methods. *Natural Language Engineering*, 29(3), 509–553. <https://doi.org/10.1017/S1351324922000213>
- Chioma Ann Udeh, Omamode Henry Orieno, Obinna Donald Daraojimba, Ndubuisi Leonard Ndubuisi, & Osato Itohan Oriekhoe. (2024). BIG DATA ANALYTICS: A review of its transformative role in modern business intelligence. *Computer Science & IT Research Journal*, 5(1). <https://doi.org/10.51594/csitrj.v5i1.718>
- Goyal, P., & Malviya, R. (2023). Challenges and opportunities of big data analytics in healthcare. *Health Care Science*, 2(5), 328–338. <https://doi.org/10.1002/hcs2.66>
- Habibi, M. (2019). Implementation of cosine similarity in an automatic classifier for comments. *JISKA (Jurnal Informatika Sunan Kalijaga)*, 3(2), 110. <https://doi.org/10.14421/jiska.2018.32-05>
- Jamarani, A., Haddadi, S., Sarvizadeh, R., Haghi Kashani, M., Akbari, M., & Moradi, S. (2024). Big data and predictive analytics: A systematic review of applications. *Artificial Intelligence Review*, 57(7). <https://doi.org/10.1007/s10462-024-10811-5>
- Jose, B., & Abraham, S. (2017). Exploring the merits of NoSQL: A study based on MongoDB. *2017 International Conference Proceedings*, 266–271. <https://doi.org/10.1109/NETACT.2017.8076778>
- Khainur, A. F. Y., Yuares, T., Fathurrohman, M. H., Widianingsih, & Rozikin, C. (2025). Analisis komparatif efektivitas *pipeline data cleaning* berbasis aturan dan lemmatisasi untuk klasifikasi sentimen. *Jurnal TIMES*, 14(2), 141–149. <https://doi.org/10.51351/jtm.14.2.2025890>
- Kim, J., & Lim, C. (2021). Customer complaints monitoring with customer review data analytics: An integrated method of sentiment and statistical process control analyses. *Advanced Engineering Informatics*, 49. <https://doi.org/10.1016/j.aei.2021.101304>
- Lingala, B. (2025). Strategic implementation of NoSQL technologies in modern enterprise data architectures. *Sarcouncil Journal of Engineering and Computer Sciences*, 4(9), 79–86.
- Mai, X., Zhang, T., Hu, C., & Zhang, Y. (2024). Feature selection algorithm for substation main equipment defect text mining based on natural language processing. *IET Cyber-Physical Systems: Theory and Applications*, 9(3), 238–246. <https://doi.org/10.1049/cps2.12079>
- Makkar, K., Kumar, P., Poriye, M., & Aggarwal, S. (2024). Improving sentiment analysis using negation scope detection and negation handling. *International Journal of Computing and Digital Systems*, 16(1), 239–247. <https://doi.org/10.12785/ijcds/160119>
- Mittal, A., & Gupta, P. (2021). An empirical study on enhancing product quality and customer satisfaction using quality assurance approach in an Indian manufacturing industry. *International Journal of Mathematical, Engineering and Management Sciences*, 6(3), 878–893. <https://doi.org/10.33889/ijmems.2021.6.3.052>
- Mohammadi, A., Ahadi, P., & Fozooni, A. (2024). Analytical evaluation of big data applications in e-commerce: A mixed method approach. *Decision Science Letters*, 8, 429–440. <https://doi.org/10.5267/dsl.2022.11.003>
- Purba, H. H., & Aisyah, S. (2017). *Quality improvement and Lean Six Sigma*. Expert.

- Sabbeh, S. F., & Fasihuddin, H. A. (2023). A comparative analysis of word embedding and deep learning for Arabic sentiment classification. *Electronics*, 12(6). <https://doi.org/10.3390/electronics12061425>
- Sahu, T. P., & Khandekar, S. (2020). A machine learning-based lexicon approach for sentiment analysis. *International Journal of Technology and Human Interaction*, 16(2), 8–22. <https://doi.org/10.4018/IJTHI.2020040102>
- Shahin, M., Chen, F. F., Hosseinzadeh, A., Maghanaki, M., & Eghbalian, A. (2024). A novel approach to voice of customer extraction using GPT-3.5 Turbo: Linking advanced NLP and Lean Six Sigma 4.0. *International Journal of Advanced Manufacturing Technology*, 131(7–8), 3615–3630. <https://doi.org/10.1007/s00170-024-13167-w>
- Song, X., Liu, D., Liu, X., Guo, K., & Mao, F. (2020). Development of database and information management system for data-driven materials design. *Scientia Sinica Technologica*, 50(6), 786–800. <https://doi.org/10.1360/SST-2019-0375>
- Sridhar, K., & Santhosh Kumar, G. (2025). Data science and IoT management system NoSQL databases in real-time systems: Performance, scalability, and use-case analysis. *International Journal of Data Science and IoT Management System*, 2025(4), 32–42.
- Torres, J. G. M., & Neto, P. L. O. C. (2018). World Café method integrated with QFD for obtaining the voice of the customer. *Production*, 28. <https://doi.org/10.1590/0103-6513.20170097>
- Wamba, S. F., Gunasekaran, A., Akter, S., Ren, S. J. F., Dubey, R., & Childe, S. J. (2017). Big data analytics and firm performance: Effects of dynamic capabilities. *Journal of Business Research*, 70, 356–365. <https://doi.org/10.1016/j.jbusres.2016.08.009>
- Wong, A. (2022). The design of an intelligent chatbot with natural language processing capabilities to support learners. In G. G., P. V., & Z. D. (Eds.), *Journal of Physics: Conference Series*, 2251(1), Article 012005. Institute of Physics. <https://doi.org/10.1088/1742-6596/2251/1/012005>
- Zaki, M., Namireddy, S. R., Pittie, T., Bihani, V., Keshri, S. R., Venugopal, V., Gosvami, N. N., & Krishnan, N. M. A. (2022). Natural language processing-guided meta-analysis and structure factor database extraction from glass literature. *Journal of Non-Crystalline Solids: X*, 15. <https://doi.org/10.1016/j.nocx.2022.100103>