

Web Scraping Data Ulasan Pelanggan untuk Kemajuan Bisnis E-Commerce pada Official Store dan Non-Official Store dengan Pendekatan Natural Language Processing

Muhamad Rafif Al Hafizh^{1)*}, Aldi Daim Fauzan²⁾, Woro Isti Rahayu³⁾, Kiki Mustaqim⁴⁾,
Rahma Hanum⁵⁾

^{1,2,3,4,5)} Program Studi Sains Data, Universitas Logistik dan Bisnis Internasional

Received: 3 April 2025

Accepted: 27 May 2025

Published: 1 July 2025



*rafif.alhafizh37@gmail.com

Kata Kunci: *Sentiment Analysis, Web Scraping, Tokopedia, Official Store, Non-Official Store, Customer Satisfaction*

DSI: Jurnal Data Science Indonesia is licensed under a Creative Commons Attribution-Non Commercial 4.0 International (CC BY-NC 4.0).

Abstrak : Penelitian ini bertujuan mengaplikasikan teknik *Web Scraping* untuk mendapatkan data ulasan *official store* dan *non-official store* dari Tokopedia untuk dibandingkan dan dilakukan *Sentiment Analysis* untuk menghapus prasangka bahwa *official store* lebih unggul daripada *non-official store*. *Web Scraping* adalah teknik ekstraksi data secara otomatis dari web melalui HTML yang kemudian di-*framing* menjadi *dataframe* agar mudah dibaca. Data yang sudah diubah akan melalui *Preprocessing Data*, yang mencakup *Data Cleaning*, Normalisasi, *Stopword Removal*, *Tokenize*, *Stemming*, dan *Translate*. Proses ini menghasilkan kata yang siap diolah ke *Labelling*, dan diberi sentimen positif, negatif, dan netral. Hasil perbandingannya adalah Faith Industries mendapatkan 80.9% positif, 4.4% negatif, dan 14.6% netral, sedangkan Julclothing mendapatkan 69.1% positif, 18.0% negatif, dan 13.0% netral. Perbandingan tersebut tidak signifikan, dapat disimpulkan bahwa *non-official store* dapat bersaing dengan *official store*. Perbandingan selanjutnya terdapat di frekuensi sentimen terbanyak. Kedua toko tersebut memiliki kata 'fit' sebagai sentimen positif dan 'measure' sebagai sentimen netral, sedangkan sentimen negatif Faith Industries mendapatkan kata 'long', Julclothing mendapatkan kata 'small'. Klasifikasi negatif tersebut disimpulkan bahwa permasalahan utama pelanggan tidak puas karena ukuran baju. Solusi permasalahan utama bagi toko dan pelanggan, yaitu toko menyediakan *chart size* yang sesuai dan pelanggan mengukur baju sesuai *chart size* sebelum membeli produk dari toko tersebut.

PENDAHULUAN

Tokopedia, salah satu *platform e-commerce* terbesar di Indonesia, menyediakan *official store* dengan lisensi resmi dan *non-official store* yang beroperasi secara mandiri. Penelitian ini bertujuan untuk memahami perbedaan persepsi pelanggan terhadap kedua jenis toko tersebut, mengingat kecenderungan pelanggan lebih mempercayai *official store* meskipun harga di *non-official store* lebih bersaing. Persepsi citra merek secara signifikan dipengaruhi oleh pengalaman mereka pada toko resmi [1]. Dalam beberapa kasus, terlihat bahwa pelanggan cenderung mempercayai *official store* dan menghindari *non-official store*. *Web Scraping* merupakan proses pengambilan dokumen semi-terstruktur dari *web*, biasanya berupa *web* yang menggunakan bahasa *markup* seperti HTML atau XHTML. Dokumen tersebut kemudian dianalisis untuk mengambil data yang dibutuhkan [2]. Dengan *Web Scraping*, ulasan pelanggan dikumpulkan secara otomatis, lalu dianalisis menggunakan *Sentiment Analysis* untuk mengidentifikasi kepuasan dan ketidakpuasan pelanggan. Hasil penelitian diharapkan dapat mengungkap persepsi pelanggan terhadap *official* dan *non-official store* di Tokopedia, serta membantu *non-official store* bersaing dalam meningkatkan

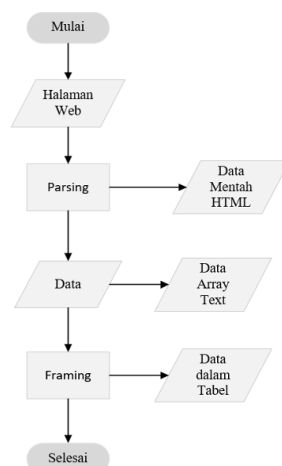
TINJAUAN LITERATUR

Beberapa studi sebelumnya telah melakukan proses *web scraping* dan *sentiment analysis* menggunakan bahasa pemrograman python dan mengimplementasikannya dalam berbagai macam bidang. Penelitian oleh Faris Arkan Ans (2023) dengan judul "Implementasi *Web Scraping* untuk Pencarian *Freelance Job*" [3] menunjukkan bahwa penggunaan metode *web scraping* sebagai cara untuk mengumpulkan data dalam jumlah besar dari *website* adalah teknik yang efektif. Penelitian selanjutnya oleh Anggi Vivi Febrianti (2020) dengan judul "Analisis Sentimen Data Ulasan Pengunjung Objek Wisata Lawang Sewu Kota Semarang pada Situs Tripadvisor"[4] menunjukkan bahwa penggunaan analisis sentimen merupakan metode yang efektif untuk melihat *review* positif, negatif, ataupun netral dari suatu objek. Hasil analisis sentimen tersebut dapat dipakai untuk pengambilan keputusan bagi pemangku kepentingan. Penelitian selanjutnya oleh Triyuliana Anisam dkk (2021) dengan judul "*Website Quality for E-Commerce using Sentiment Analysis (Case Studies in Tokopedia and Shopee)*"[5] menunjukkan bahwa mengetahui sentimen atau *feedback* yang diberikan *user* terhadap kualitas *website* Tokopedia dan Shopee dapat memberikan informasi tentang preferensi *user*, dan alasan-alasan mengapa *user* tersebut bisa memilih satu aplikasi dibandingkan dengan aplikasi lainnya.

Penelitian terakhir oleh Rita Apriani (2021) dengan judul Analisis Sentimen Menggunakan Algoritma Naive Bayes Terhadap Komentar Aplikasi Tokopedia" [6] menunjukkan bahwa penggunaan Naive Bayes untuk menentukan analisis sentimen terbukti sangat akurat, dan mendapatkan hasil yang memuaskan. Penelitian ini menggunakan teori kepuasan pelanggan untuk melihat apakah pelanggan tersebut merasakan puas berbelanja di suatu toko atau tidak. Kepuasan pelanggan merujuk pada perasaan senang atau kekecewaan yang timbul setelah membandingkan kinerja suatu produk dengan harapan yang dimiliki. Jika kinerja produk berada di bawah harapan, pelanggan akan merasa tidak puas. Namun, jika kinerja produk melebihi harapan, pelanggan akan merasa senang [7]. Ada 4 cara mengukur kepuasan pelanggan, diantaranya yaitu *Grievance and Suggestion System*, *Patron Delight Surveys*, *Ghost Shopping*, *Lost Purchaser Analysis* [8].

METODE PENELITIAN

Studi Literatur menjadi salah satu metode penelitian yang digunakan untuk penelitian ini. Metode studi literatur adalah serangkaian kegiatan yang berkenaan dengan metode pengumpulan data pustaka, membaca, dan mencatat, serta mengelola bahan penelitian [9]. Data dari penelitian ini menggunakan data sekunder, data yang sudah ada dari halaman di *review website* Tokopedia, di toko Faith Industries dan Julclothing. Data diperoleh menggunakan proses *web scraping* menggunakan *library* BeautifulSoup untuk melakukan parsing HTML dan XML dalam Python [3], dan Selenium untuk membantu proses *testing* otomatisasi dari *platform web* [10].



Gambar 1 Alur Pengumpulan Data

Penelitian ini melalui tahap *preprocessing data*, yaitu proses mengubah data mentah ke bentuk yang lebih mudah dipahami. *Preprocessing Data* untuk menghindari data mentah yang memiliki format berantakan dan tidak teratur. Berikut adalah alur tahapan melakukan *preprocessing data*.



Gambar 2 Alur *Preprocessing Data*

Selanjutnya, dilakukan proses *labelling*. *Labelling* dalam *Sentiment Analysis* adalah langkah di mana data mentah di beri label untuk menemukan konteks yang dituju. Dalam konteks *sentiment analysis*, *labelling* melibatkan penandaan teks dengan label seperti positif, negatif, atau netral. Proses *labelling* menggunakan *library* NLTK dan Textblob. NLTK (*Natural Language Toolkit*) adalah modul dalam Python yang menyediakan banyak tipe data NLP (*Natural Language Processing*) [11]. Sedangkan Textblob adalah *text processing package* untuk Python. *Package* ini menyediakan API dasar untuk melakukan NLP seperti *part-of-speech tagging*, *noun phrase extraction*, *sentiment analysis*, *classification* dan *translation* [12].

Untuk memperjelas hasil penelitian, dilakukan Visualisasi agar hasil data yang telah dianalisis dapat diubah menjadi representasi grafis. Visualisasi membantu mengidentifikasi data-data dengan sentimen positif, negatif, atau netral dengan mudah. *Library* yang digunakan adalah Matplotlib, Seaborn, dan WordCloud. Matplotlib adalah *library* Python yang digunakan untuk membuat visualisasi data seperti grafik, diagram, dan plot [13]. Sama seperti Matplotlib, Seaborn pada dasarnya adalah sebuah *library* untuk *plotting*. Wordcloud adalah *library* visualisasi yang terkhususkan untuk *word cloud* (gumpalan kata) dari teks,

HASIL PENELITIAN

Penelitian ini bertujuan untuk membandingkan kepuasan pelanggan antara *official store* dan *non-official store* di Tokopedia, khususnya toko baju Faith Industries sebagai *official store* dan Julclothing sebagai *non-official store*. Proses penelitian ini diawali dengan *web scraping* dengan menggunakan *library* dalam Python.

1) Data Uji Coba

Web Scraping menghasilkan data uji coba dalam format CSV, berisikan data mentah yang perlu diolah dalam Python untuk analisis sentimen. Data dari toko Faith Industries mencakup 362 baris dengan 5 kolom yaitu kolom Produk, Nama, Waktu, Review, dan Bintang. Sementara untuk data toko Julclothing terdiri dari 362 baris dengan 5 kolom yang sama. Kedua data tersebut masing-masing diperoleh pada 5 Juni 2024.

2) *Preprocessing Data*

Preprocessing bertujuan untuk mengurangi ukuran, data, menemukan relasi antar data, normalisasi data, dan mengekstrak fitur yang ada pada data [14]. Langkah-langkah pada penelitian ini meliputi sebagai berikut.

a. *Data Cleaning*

Data Cleaning digunakan untuk merujuk pada semua jenis tugas dan aktivitas untuk mendeteksi dan memperbaiki kesalahan dalam data [15].

b. Normalisasi

Dengan melakukan normalisasi, kata-kata yang disingkat akan diperjelas. Contohnya 'tdk' menjadi 'tidak'.

c. *Stopword*

Stopword adalah kata yang paling sering muncul dan mengandung sedikit informasi yang biasanya tidak diperlukan [16]. Contohnya "yang, dan, di" dsb.

d. *Tokenize*

Tokenize adalah proses memecah konten tekstual menjadi kata-kata atau elemen bermakna lainnya yang disebut token [17]. Contohnya "ini teks" menjadi ["ini", "teks"].

e. *Stemming*

Stemming adalah proses bentuk kata varian diptakan ke bentuk dasarnya [10]. Contohnya "berlari" menjadi "lari".

f. *Translate*

Kolom *review* akan diubah sepenuhnya dari Bahasa Indonesia menjadi Bahasa Inggris.

3) *Labelling*

Proses ini akan memberikan kolom tambahan yaitu label yang isinya menunjukkan angka polaritas. Polaritas adalah angka *float* di kisara -1.0 dan 1.0. Nilai mendekati 1 menunjukkan positif, dan nilai mendekat -1 menunjukkan negatif, sedangkan jika 0 menunjukkan netral. Hasil *labelling* dari kedua toko tersebut adalah sebagai berikut.

Tabel 1 Hasil Analisis Data Faith Industries

Hasil Analisis Data Faith Industries	
Klasifikasi Positif	293
Klasifikasi Negatif	16
Klasifikasi Netral	53
Total Data	362

Tabel 2 Hasil Analisis Data Julclothing

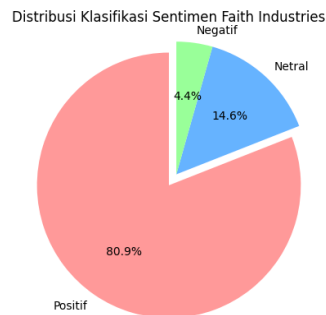
Hasil Analisis Data Julclothing	
Klasifikasi Positif	250
Klasifikasi Negatif	65
Klasifikasi Netral	47
Total Data	362

4) *Visualisasi*

Teknik visualisasi adalah konversi data ke dalam format visual atau tabel sehingga karakteristik dari data dan relasi di antara *item* data atau atribut dapat di analisis dan di laporkan [18]. Penelitian ini menggunakan *Pie Chart*, *Bar Chart*, dan *Word Cloud*.

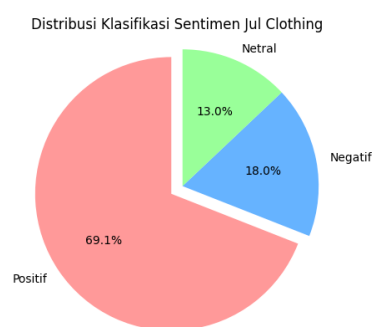
a. *Pie Chart*

Pie Chart adalah diagram yang menunjukkan perbandingan menyeluruh antar data dengan cara membagi lingkaran dengan sudut pusat yang sesuai dengan perbandingan tersebut [19].



Gambar 3 *Pie Chart* Faith Industries

Gambar tersebut menunjukkan sentimen dominan positif sebesar 80.9%, negatif 4.4%, dan netral 13% pada Faith Industries.

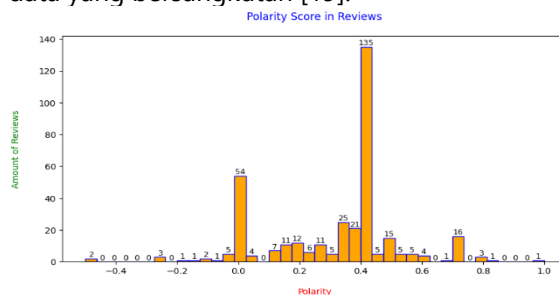


Gambar 4 *Pie Chart* Julclothing

Gambar tersebut menunjukkan sentimen dominan positif sebesar 69.1%, negatif 13.0%, dan netral 18.0% pada Julclothing.

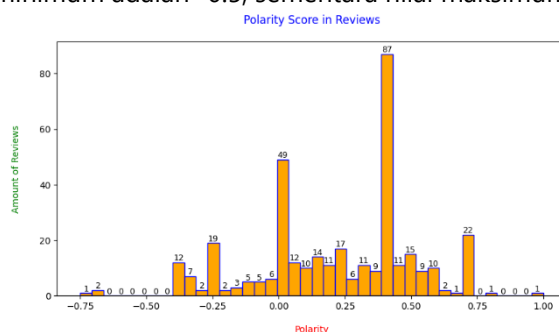
b. *Bar Chart Polarity Score*

Bar Chart adalah grafik data yang berbentuk persegi panjang dan dilengkapi dengan skala atau ukuran sesuai data yang bersangkutan [19].



Gambar 5 *Bar Chart Polarity Score* Faith Industries

Gambar menunjukkan skor polaritas pada Faith Industries memiliki skor tertinggi sebesar 0.4, dan 0.0. Nilai minimum adalah -0.5, sementara nilai maksimum adalah 1.



Gambar 6 *Bar Chart Polarity Score* Julclothing

Gambar menunjukkan skor polaritas pada Julclothing, dengan skor tertinggi sebesar 0.4, dan 0.0 netral. Nilai minimum adalah -0.75, sedangkan nilai maksimum adalah 1.

c. *Word Cloud*

Word Cloud adalah kumpulan kata yang merepresentasikan atau mewakili makna dari *cluster* [20].



Gambar 7 *Word Cloud* Faith Industries

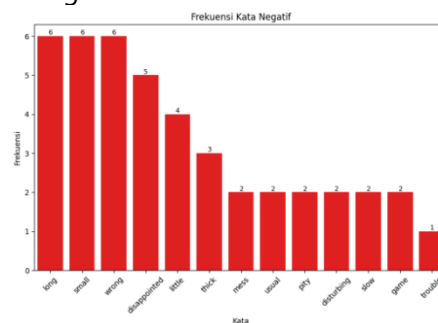
Word Cloud tersebut menampilkan kata positif yang sering muncul dalam *review* dari Faith Industries seperti "good, fit, cool, comfortable, fast".



Gambar 8 *Word Cloud* Julclothing

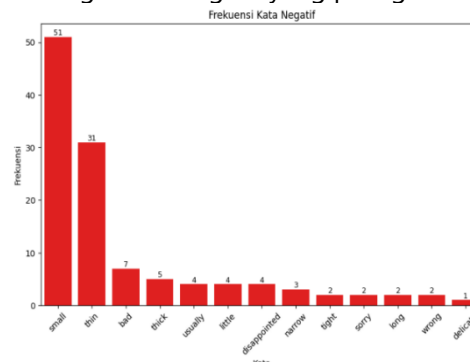
Word Cloud tersebut menampilkan kata positif yang sering muncul dalam *review* dari Julclothing seperti "good, comfortable, fit, fast, great".

d. *Bar Chart* Frekuensi Kata Negatif



Gambar 9 *Bar Chart* Frekuensi Kata Negatif Faith Industries

Faith Industries memiliki kata "long" sebagai kata negatif yang sering muncul pada data, dan juga kata "trouble" sebagai kata negatif yang paling sedikit.



Gambar 10 *Bar Chart* Frekuensi Kata Negatif Julclothing

Julclothing memiliki kata "small" sebagai kata negatif yang sering muncul pada data, dan kata "delicate" sebagai kata negatif yang paling sedikit.

PEMBAHASAN

Faith Industries memiliki sentimen positif 80.9%, negatif 4.4%, dan netral 14.6%, sedangkan Julclothing memiliki 69.1%, negatif 18.0%, dan netral 13.0%. Visualisasi menunjukkan bahwa kata "fit" dominan dalam ulasan positif kedua toko, sementara kata "long, small, measure" mendominasi sentimen negatif dan netral. Penelitian ini menggunakan visualisasi yang sudah dibuat yang kemudian diubah menjadi tabel untuk membantu pengambilan keputusan guna peningkatan layanan.

Tabel 3 Kata Negatif

Kata	Frekuensi
<i>Small</i>	51
<i>Thin</i>	31
<i>Thick</i>	7
<i>Bad</i>	5
<i>Usually</i>	4
<i>Little</i>	4

Tabel di atas menunjukkan kata negatif yang sering muncul pada kedua toko, yang mengindikasikan keluhan mengenai ukuran dan ketebalan produk. Langkah selanjutnya adalah mencari ulasan yang mengandung kata-kata tersebut untuk membantu pemilik toko mengidentifikasi area perbaikan. Berikut adalah ringkasan ulasan negatif terbanyak dari kedua toko.

Tabel 4 Hasil Ulasan

Review	Klasifikasi
bahan tipis ukur l nya ukur m kecil sempit	Negatif
ukur xl nya kecil bahan tipis sesuai harga lah d	Negatif
ukur pas ukur terlalu besar tinggi badan 171-180cm ukur terlalu kecil tinggi badan 180cm berat badan 80kg	Negatif
baju nya keren salah ukur min kan sen xl datengnya xxl tapu gapapa untung gak kirim ukur kecil	Negatif

Kepuasan pelanggan merujuk pada perasaan senang atau kekecewaan yang timbul setelah membandingkan kinerja suatu produk dengan harapan yang dimiliki [7]. Pada penelitian ini, pengukuran kepuasan pelanggan dapat dilihat dari keluhan dan saran pelanggan. Berdasarkan analisis yang telah dilakukan, dapat ditarik kesimpulan bagi pihak toko dan pelanggan yaitu sebagai berikut.

Tabel 5 Saran bagi Pelanggan dan Toko

Toko	Pelanggan
- Menyediakan ukuran baju yang lebih besar. - Menyediakan tabel ukuran untuk pengukuran yang akurat. - Meningkatkan kualitas bahan agar sesuai standar. - Menambahkan deskripsi bahan produk secara rinci. - Melakukan pemeriksaan ulang pesanan pelanggan untuk menghindari kesalahan pengiriman. - Mengadakan kebijakan refund untuk menangani kekecewaan pelanggan.	- Lebih teliti dalam memilih ukuran baju - Mengukur badan sesuai chart size agar mendapat ukuran yang sesuai dengan ekspektasi - Melihat deskripsi produk dan/atau melihat ulasan dari pelanggan lain tentang ketebalan bahan - Mencari toko yang memiliki kebijakan refund

KESIMPULAN

Penelitian berjudul "Web Scraping Data Ulasan Pelanggan untuk Kemajuan Bisnis E-Commerce pada Official Store dan Non-Official Store dengan Pendekatan Natural Language Processing" menghasilkan beberapa temuan kunci. Data diperoleh menggunakan teknik *Web Scraping* dengan total 362 ulasan untuk masing-masing toko. Hasil analisis menunjukkan bahwa Faith Industries memiliki 293 sentimen positif, 16 negatif, dan 53 netral, sedangkan Julclothing mencatat 250 positif, 65 negatif, 47 netral. Masalah utama yang diidentifikasi adalah kinerja produk yang tidak memenuhi harapan pelanggan dan kesalahan pengiriman. Untuk meningkatkan kepuasan pelanggan, toko perlu menyediakan ukuran lebih besar, menampilkan tabel *chart size*, dan meningkatkan kualitas bahan. Kebijakan *refund* juga perlu diterapkan. Sementara itu, pelanggan disarankan untuk lebih teliti dalam memilih ukuran dan memeriksa deskripsi serta ulasan produk.

REFERENCES

- [1] L. D. Hollebeek and T. Chen, "Exploring positively- versus negatively-valenced brand engagement: a conceptual model," *Journal of Product & Brand Management*, vol. 23, no. 1, pp. 62–74, Mar. 2014, doi: 10.1108/JPBM-06-2013-0332.
- [2] M. Turland, *PHP|ARCHITECT'S GUIDE TO WEB SCRAPING*. 2010.
- [3] F. Ans, "IMPLEMENTASI WEB SCRAPING UNTUK PENCARIAN FREELANCE JOB," *Jurnal Teknologi Pintar*, vol. 3, no. 9, Dec. 2023.
- [4] A. Febrianti, "ANALISIS SENTIMEN DATA ULASAN PENGUNJUNG OBJEK WISATA LAWANG SEWU KOTA SEMARANG PADA SITUS TRIPADVISOR," *Repository UNNES*, Dec. 2020.
- [5] T. Annisa, K. Kusumahadi, and E. Nurhazizah, "Website Quality for E-Commerce using Sentiment Analysis (Case Studies in Tokopedia and Shopee)," *European International Conference on Industrial Engineering and Operations Management Rome*, vol. 4, Aug. 2021.
- [6] R. Apriani and D. Gustian, "ANALISIS SENTIMEN DENGAN NAÏVE BAYES TERHADAP KOMENTAR APLIKASI TOKOPEDIA," *Jurnal Rekayasa Teknologi Nusa Putra*, vol. 6, no. 1, pp. 54–62, Sep. 2019, doi: 10.52005/rekayasa.v6i1.86.
- [7] F. Sayedi, R. Ilato, R. Hafid, U. Moonti, and S. Sudirman, "Pengaruh Kualitas Produk Terhadap Kepuasan Konsumen Toko Amira Meubel.," *JOURNAL of ECONOMIC and BUSINESS EDUCATION*, vol. 1, no. 2, pp. 3–3, May 2023, Accessed: Jun. 04, 2024. [Online]. Available: <https://ejurnal.ung.ac.id/index.php/JEBE/article/download/18651/6516>
- [8] A. Zikri and M. I. Harahap, "Analisis Kualitas Pelayanan Pengiriman Barang terhadap Kepuasan Konsumen pada PT Pos Indonesia (Persero) Regional I Sumatera."
- [9] M. Zed, *Metode Penelitian Kepustakaan*, vol. 3. Jakarta: Yayasan Pustaka Obor Indonesia, 2014.
- [10] J. Singh and V. Gupta, "Text Stemming," *ACM Comput Surv*, vol. 49, no. 3, pp. 1–46, Sep. 2017, doi: 10.1145/2975608.
- [11] S. Bird, "NLTK: The Natural Language Toolkit," *COLING/ACL*, pp. 1–1, 2006.
- [12] O. Abiola, A. Abayomi-Alli, O. A. Tale, S. Misra, and O. Abayomi-Alli, "Sentiment analysis of COVID-19 tweets from selected hashtags in Nigeria using VADER and Text Blob analyser," *Journal of Electrical Systems and Information Technology*, vol. 10, no. 1, p. 5, Jan. 2023, doi: 10.1186/s43067-023-00070-9.
- [13] R. Guntara, "Visualisasi Data Laporan Penjualan Toko Online Melalui Pendekatan Data Science Menggunakan Google Colab," *Jurnal Ilmiah Multidisiplin*, vol. 2, no. 6, Apr. 2023.
- [14] S. A. Alasadi and W. S. Bhaya, "Review of Data Preprocessing Techniques in Data Mining," *Journal of Engineering and Applied Sciences*, vol. 12, no. 16, pp. 1–1, 2017.
- [15] I. F. Ilyas and X. Chu, *Data Cleaning*, 1st ed. New York, NY, USA: ACM, 2019. doi: 10.1145/3310205.
- [16] J. Kaur and P. K. Buttar, "A Systematic Review on Stopword Removal Algorithms," *International Journal on Future Revolution in Computer Science & Communication Engineering*, vol. 4, no. 4, Apr. 2018.
- [17] S. Vijayarani and R. Janni, "TEXT MINING: OPEN SOURCE TOKENIZATION TOOLS – AN ANALYSIS," *Advanced Computational Intelligence: An International Journal (ACII)*, vol. 3, no. 1, Jan. 2016.

- [18] D. Fernando, "Visualisasi Data Menggunakan Google Data Studio," *Prosiding Seminar Nasional Rekayasa Teknologi Informasi | SNARTISI*, vol. 1, Nov. 2018.
- [19] Febina Sembiring, Noni Simbolon, Refina Br Tarigan, and Anastasya Br Sinuraya, "Gambaran Pemahaman Mahasiswa MIK STIKes Santa Elisabeth Medan Mengenai Penyajian Data dalam Bentuk Diagram, Tabular dan Grafikal," *INSOLOGI: Jurnal Sains dan Teknologi*, vol. 3, no. 4, pp. 481–487, Aug. 2024, doi: 10.55123/insologi.v3i4.4092.
- [20] K. Septiani, "Perbandingan Analisis Sentimen Terhadap Pembayaran Digital 'Go-Pay' Dan 'Ovo' Di Media Sosial Twitter Menggunakan Metode Naive Bayes Dan Word Cloud," *Jurnal Ilmu Data*, vol. 2, no. 10, pp. 3–3, Jun. 2022.