

## Sentiment Analysis of the Top 5 E-commerce Platforms in Indonesia using Text Mining and Natural Language Processing (NLP)

R. A. E. Virgana Targa Sapanji<sup>1</sup>, Dani Hamdani<sup>2</sup>, Parlindungan Harahap<sup>3</sup>

\* Sistem Informasi, Universitas Widyatama Bandung

[rae.virgana@widyatama.ac.id](mailto:rae.virgana@widyatama.ac.id)<sup>1</sup>, [dani.hamdani@widyatama.ac.id](mailto:dani.hamdani@widyatama.ac.id)<sup>2</sup>, [parlindungan.mt@widyatama.ac.id](mailto:parlindungan.mt@widyatama.ac.id)<sup>3</sup>

### Article Info

#### Article history:

Received 2023-09-09

Revised 2023-09-19

Accepted 2023-09-21

#### Keyword:

*E-commerce,*  
*Sentiment Analysis,*  
*NLP,*  
*Text Mining,*  
*R Programming.*

### ABSTRACT

This research attempts to depict a sentiment comparison of the top 5 E-commerce platforms in Indonesia by gathering the emotional tone behind sentence contents related to customer sentiments, customer experiences, and the brand reputation of E-commerce. Data were collected using Python 3.11.4 with the google-play-scraper library, extracted from user reviews/comments on each play store page of the top 5 E-commerce platforms in Indonesia. A sampling of 10,000 records was taken to form a long document term matrix (DTM) of 59,981,785 due to the limitation of CPU capacity for data matrix size. R Programming version 4.3.1 was employed for sentiment analysis in this study. It can be concluded that user comments or reviews on the top five (5) E-commerce platforms in Indonesia show positive sentences indicating user satisfaction (3664 sentences), neutral sentences indicating average user appreciation (2282 sentences), and negative sentences indicating user dissatisfaction (4054 sentences). At least with more positive and neutral sentences, it is indicated that 59.64% of E-commerce users in Indonesia express a positive opinion on the performance of the top 5 E-commerce platforms in the country.



This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.

### I. PENDAHULUAN

Penelitian Analisis sentimen 5 *E-commerce* terbesar di Indonesia menggunakan *Text Mining* dan *Natural Language Processing* (NLP), *E-commerce* telah berkembang di Indonesia dengan nilai pasar bruto mencapai US\$ 32 miliar, perkiraan Google, Bain, dan Temasek mencapai US\$ 83 miliar pada tahun 2025, sudah 138,1 juta orang yang telah berbelanja online di Indonesia [1].

Tetapi sentimen negatif, positif dan netral terhadap *E-commerce* tersebut tidak bisa dihindari, mulai dari ketidakpuasan layanan pelanggan, aplikasi yang lambat, sulit diakses, data yang hilang, penjual yang tidak amanah, komentar-komentar pelanggan ini tercurah dalam sentimen tulisan yang negatif, positif maupun netral.

Penelitian ini mencoba menggambarkan komparasi sentimen terhadap 5 *E-commerce* tersebut, sehingga dihasilkan analisis sentimen membantu mengumpulkan nada emosional di balik isi teks tentang sentimen pelanggan, pengalaman pelanggan dan reputasi merek *E-commerce*.

Penelitian ini menjadi sangat penting untuk para *E-commerce* tersebut, karena ada masalah bersifat gunung es dengan para pelanggan yang tidak puas, baik terhadap layanan yang disediakan oleh *E-commerce* itu sendiri seperti aplikasi, *helpdesk*, *call center*, maupun dengan pihak-pihak lain seperti para penjual, jasa pengiriman, bahkan para penipu di *E-commerce*, dan akhirnya semua permasalahan tersebut berujung pada sentimen buruk terhadap para *E-commerce* tersebut.

Analisis sentimen 5 *E-commerce* terbesar di Indonesia (Tokopedia, Shopee, Bukalapak, Lazada, Blibli) menggunakan *Text Mining* dan *Natural Language Processing* (NLP), merupakan penambangan opini dengan pendekatan pemrosesan bahasa alami (NLP) untuk mengidentifikasi nada emosional di balik isi teks komentar/review para pelanggan *E-commerce* tersebut, melibatkan penggunaan penambangan data, pembelajaran mesin, kecerdasan buatan, dan linguistik komputasional untuk menambang teks sentimen dan informasi subyektif untuk mengungkapkan perasaan positif, negatif, atau netral.

Akhirnya dengan penelitian ini bisa tergambarkan dari 5 *E-commerce* tersebut, mana yang paling baik dalam memahami pelanggan, penjual, jasa pengiriman, siapapun yang terkait dengan industri *E-commerce* ini.

Analisis sentimen berfokus pada polaritas teks (positif, negatif, netral) tetapi juga melampaui polaritas untuk mendeteksi perasaan dan emosi tertentu (marah, senang, sedih, dll), urgensi (mendesak, tidak mendesak) dan bahkan niat (tertarik vs tidak tertarik) [2].

Analisis sentimen dari sejumlah besar ulasan pengguna pada platform e-commerce dapat secara efektif meningkatkan kepuasan pengguna. Makalah ini mengusulkan model analisis sentimen baru-SLCABG, yang didasarkan pada leksikon sentimen dan menggabungkan Convolutional Neural Network (CNN) dan Bidirectional Gated Recurrent Unit (BiGRU) [3].

Serangkaian percobaan dilakukan pada kumpulan data skala besar yang melibatkan lebih dari 500 ribu ulasan produk. Hasilnya menunjukkan bahwa model yang diusulkan mengungguli model pembelajaran mendalam lainnya, termasuk RNN, BiGRU, dan Bert-BiLSTM, yang dapat mencapai akurasi lebih dari 95,5% [4].

Ulasan konsumen dalam sistem E-commerce biasanya diperlakukan sebagai sumber daya penting yang mencerminkan pengalaman, perasaan, dan keinginan pengguna untuk membeli barang. Semua informasi ini dapat melibatkan pandangan konsumen terhadap hal-hal yang dapat mengungkapkan minat, sentimen, dan opini. Banyak jenis penelitian telah menunjukkan bahwa orang lebih cenderung mempercayai satu sama lain dengan sikap yang sama terhadap hal yang serupa [5].

Penambangan data dapat digunakan untuk analisis data pengguna media sosial yang mengunjungi E-Commerce. Penelitian ini menggunakan teknik data mining yang bertujuan untuk membandingkan klasifikasi dalam analisis sentimen dari pandangan pelanggan E-Commerce yang telah dituliskan di Twitter [6].

Opinion Mining adalah nama lain untuk analisis sentimen. Berbagai macam data teks dihasilkan dalam bentuk saran, umpan balik, tweet, dan komentar. Portal E-Commerce menghasilkan banyak data setiap hari dalam bentuk ulasan pelanggan. Menganalisis data E-Commerce akan membantu pengecer online untuk memahami harapan pelanggan, memberikan pengalaman berbelanja yang lebih baik, dan meningkatkan penjualan [7].

Teknik Machine learning yang disebut Support Vector Machine (SVM) digunakan untuk merancang model dan model ini telah diimplementasikan pada aplikasi E-commerce. Data yang digunakan dalam penelitian ini adalah ulasan produk online yang dikumpulkan dari Amazon.com. Eksperimen analisis sentimen dilakukan untuk dua level kategorisasi: level review dan level kalimat [8].

Analisis data tersegmentasi untuk sentimen berdasarkan masing-masing kelompok usia dan jenis kelamin. Terakhir, analisis sentimen dilakukan dengan menggunakan berbagai pendekatan Pembelajaran Mesin (ML) termasuk entropi

maksimum, mesin vektor pendukung, jaringan saraf konvolusional, dan memori jangka pendek panjang untuk mempelajari dampak usia dan jenis kelamin pada ulasan pengguna [9].

Analisis berita yang memengaruhi harga produk, dan menetapkan model baru untuk prediksi harga. Hasil penelitian menunjukkan bahwa peristiwa berita yang signifikan berdampak pada harga jual produk elektronik, dan dapat meningkatkan akurasi peramalan harga [10].

Ulasan konsumen adalah informasi penting yang mencerminkan kualitas barang dan layanan E-commerce dan masalah yang ada setelah berbelanja. Karena kemungkinan perbedaan dalam pengalaman konsumen dengan kualitas barang dan jasa, ulasan konsumen dapat melibatkan ekspresi emosi atau pendapat multi-aspek [11].

Sistem analisis sentimen membantu organisasi mengumpulkan wawasan tentang sentimen pelanggan waktu nyata, pengalaman pelanggan, dan reputasi merek. Umumnya, alat ini menggunakan analitik teks untuk menganalisis sumber online seperti email, postingan blog, ulasan online, tiket dukungan pelanggan, artikel berita, tanggapan survei, studi kasus, obrolan web, tweet, forum, dan komentar. Algoritma digunakan untuk mengimplementasikan metode penilaian berbasis aturan, otomatis, atau gabungan apakah pelanggan mengungkapkan kata-kata positif, kata-kata negatif, atau kata-kata netral. Selain mengidentifikasi sentimen, analisis sentimen dapat mengekstraksi polaritas atau jumlah positif dan negatif, subjek dan pemegang opini dalam teks. Pendekatan ini digunakan untuk menganalisis berbagai bagian teks, seperti dokumen lengkap atau paragraf, kalimat atau subkalimat [12].

Analisis sentimen adalah penambangan teks kontekstual yang mengidentifikasi dan mengekstrak informasi subyektif dalam materi sumber, dan membantu bisnis memahami sentimen sosial merek, produk, atau layanan mereka sambil memantau percakapan online. Dengan kemajuan terkini dalam pembelajaran mendalam, kemampuan algoritme untuk menganalisis teks telah meningkat pesat. Penggunaan kreatif teknik kecerdasan buatan tingkat lanjut dapat menjadi alat yang efektif untuk melakukan penelitian mendalam [13].

Bisnis menggunakan analisis sentimen untuk memperoleh kecerdasan dan menyusun rencana yang dapat ditindaklanjuti di berbagai bidang [14].

Metode penambangan teks memungkinkan menyorot kata kunci yang paling sering digunakan dalam paragraf teks. Seseorang dapat membuat cloud kata, juga disebut cloud teks atau cloud tag, yang merupakan representasi visual dari data teks. Prosedur membuat cloud kata sangat sederhana di R. Paket penambangan teks (tm) dan paket pembuat cloud kata (wordcloud) tersedia di R untuk membantu menganalisis teks dan dengan cepat memvisualisasikan kata kunci sebagai cloud kata [15, p. 1].

Lima langkah utama untuk membuat word cloud di R [15, p. 2]

- Langkah 1: Buat file teks
- Langkah 2 : Instal dan muat paket yang diperlukan

- Langkah 3: Penambahan teks
- Langkah 4: Bangun matriks term-dokumen
- Langkah 5: Hasilkan cloud Word

Analisis Sentimen R Notebook, R Markdown Ini adalah dokumen R Markdown. Markdown adalah sintaks pemformatan sederhana untuk membuat dokumen HTML, PDF, dan MS Word. Untuk detail lebih lanjut tentang penggunaan R Markdown. Saat mengklik tombol Knit, sebuah dokumen akan dihasilkan yang menyertakan konten serta keluaran dari setiap potongan kode R yang disematkan di dalam dokumen [16].

Menggunakan paket TM, topik tutorial penambahan teks seperti hubungan kata, analisis sentimen, dan frekuensi dokumen terbalik dibahas secara mendetail. Sejalan dengan topik-topik ini ada paket lain yang berguna yang memungkinkan peneliti untuk mendapatkan wawasan yang berguna dari teks. Paket tm dibuat oleh Ingo Feinerer dan memungkinkan peneliti pemula (seperti saya) untuk memanfaatkan kekuatan R tanpa pemahaman mendalam tentang bahasa pemrograman. Dengan pemahaman ini, mari jelajahi beberapa aplikasi praktis dari paket tm [17].

Salah satu hal pertama yang penting untuk dipelajari tentang analisis teks kuantitatif adalah bahwa sebagian besar program komputer, teks atau string juga memiliki basis numerik yang disebut pengkodean karakter. Pengkodean karakter adalah gaya penulisan teks dalam kode komputer yang membantu program seperti browser web mengetahui cara menampilkan teks. Saat ini ada lusinan jenis pengkodean karakter berbeda yang dihasilkan tidak hanya dari kemajuan teknologi komputasi—dan pengembangan gaya yang berbeda untuk sistem operasi yang berbeda—tetapi juga untuk bahasa yang berbeda (dan bahkan bahasa baru seperti emoji) [18].

Alur kerja yang biasa menggunakan paket penambahan teks, yaitu, TM dan wordcloud, dan kerangka kerja penambahan teks akan menganalisis frekuensi kata dari file teks yang berbeda dan akhirnya membuat cloud kata yang bagus dari kata yang dibagikan di seluruh dokumen dan memvisualisasikan distribusi kata yang sering digunakan, Corpus struktur utama yang tm digunakan untuk menyimpan dan memanipulasi dokumen teks. Ada dua jenis VCorpus (Volatile Corpus) dan PCorpus (Permanent Corpus), perbedaan utama antara kedua implementasi ini adalah bahwa yang pertama menyimpan dokumen sebagai objek R di memori sedangkan yang terakhir berurusan dengan dokumen yang disimpan di luar lingkungan R [19].

Filter data dengan beberapa kondisi di R menggunakan Dplyr, memfilter dataframe dengan berbagai kondisi dalam bahasa pemrograman R menggunakan paket dplyr, Fungsi filter() digunakan untuk menghasilkan subset dari bingkai data, mempertahankan semua baris yang memenuhi kondisi yang ditentukan. Metode filter() dalam bahasa pemrograman R dapat diterapkan pada data yang dikelompokkan dan tidak dikelompokkan [20].

Sentiment-Analysis-with-R, analisis sentimen dari review atau ulasan pelanggan, untuk mendapatkan hasil analisis

kesimpulan dari analisis ini bisa dilihat pada gambar sentiment negatif, netral dan positif. secara umum ulasan negatif banyak membahas tentang delay pada maskapai lion air. sementara yang positif berkaitan dengan "good" dan "time" untuk melihat apa yang berkaitan dengan 2 kata tersebut bisa dilihat dari asosiasi kata pada ulasan positif [21].

Analisis Sentimen pada Pesawat Udara Indonesia, memodelkan analisis sentimen untuk salah satu Pesawat tertua di Indonesia, Sriwijaya Airlines, analisis Sentimen di pesawat [22].

WebScraping, pemanenan web, atau ekstraksi data web adalah pengikisan data yang digunakan untuk mengekstraksi data dari situs web. Perangkat lunak pengikis web dapat langsung mengakses World Wide Web menggunakan Hypertext Transfer Protocol atau browser web. Meskipun pengikisan web dapat dilakukan secara manual oleh pengguna perangkat lunak, istilah tersebut biasanya mengacu pada proses otomatis yang diterapkan menggunakan bot atau perayap web. Ini adalah bentuk penyalinan di mana data tertentu dikumpulkan dan disalin dari web, biasanya ke database atau spreadsheet lokal pusat, untuk pengambilan atau analisis nanti [23].

Corpus adalah kumpulan dokumen teks di mana kami akan menerapkan penambahan teks atau rutinitas pemrosesan bahasa alami untuk mendapatkan kesimpulan. Dalam paket tm R, ada fungsi untuk membuat korpus dari file, vektor, dll. Ada ekstensi tm untuk membuat korpus dari email juga [24], pengantar korpus menunjukkan fungsionalitas yang disediakan oleh paket corpus R [25].

Pra-pemrosesan teks: Hentikan penghapusan kata menggunakan perpustakaan yang berbeda, adalah proses mempersiapkan data teks sehingga mesin dapat menggunakan yang sama untuk melakukan tugas-tugas seperti analisis, prediksi, dll. Ada banyak langkah berbeda dalam pra-pemrosesan teks tetapi dalam artikel ini, kita hanya akan mengenal berhenti kata-kata, mengapa kami menghapusnya, dan perpustakaan berbeda yang dapat digunakan untuk menghapusnya [26].

Stop words untuk Bahasa Indonesia, stop words adalah kata umum (common words) yang biasanya muncul dalam jumlah besar dan dianggap tidak memiliki makna. Stop words umumnya dimanfaatkan dalam task information retrieval, termasuk oleh Google (penjelasannya di sini). Contoh stop words untuk bahasa Inggris diantaranya "of", "the". Sedangkan untuk bahasa Indonesia diantaranya "yang", "di", "ke" [27].

Matriks dokumen istilah adalah cara merepresentasikan kata-kata dalam teks sebagai tabel (atau matriks) angka. Baris matriks mewakili teks tanggapan yang akan dianalisis, dan kolom matriks mewakili kata-kata dari teks yang akan digunakan dalam analisis [28]. Document Term Matrix merupakan algoritma – Metode perhitungan yang sering kita temui dalam text mining. Document Term Matrix merupakan sebuah representasi numerik dari dokumen dalam corpus. Corpus hanyalah kumpulan dokumen. Dengan kumpulan kata yang "lebih besar" ini, Melalui Document

Term Matrix, kita dapat melakukan analisis yang lebih menarik. Mudah untuk menentukan jumlah kata individual untuk setiap dokumen atau untuk semua dokumen. Misalkan untuk menghitung agregat dan statistik dasar seperti jumlah istilah rata-rata, mean, median, mode, varians, dan deviasi standar dari panjang dokumen, serta dapat mengetahui istilah mana yang lebih sering dalam kumpulan dokumen dan dapat menggunakan informasi tersebut untuk menentukan istilah mana yang lebih mungkin mewakili dokumen tersebut [29].

## II. METODE PENELITIAN

### A. Metode Analisis

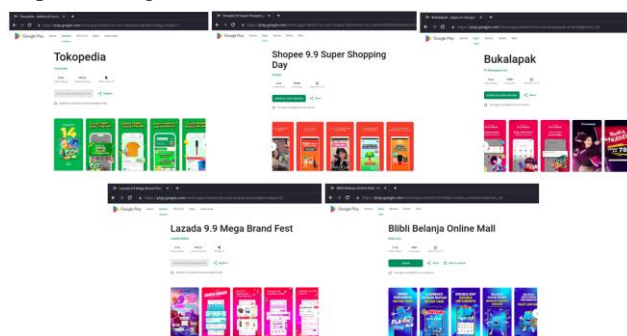
Metode analisis sentimen menggunakan model pembelajaran mesin untuk melakukan analisis teks bahasa manusia. Metrik yang digunakan dirancang untuk mendeteksi apakah keseluruhan sentimen sepotong teks positif, negatif, atau netral, metode analisis sentimen umumnya mengikuti langkah-langkah berikut [12]:

- Mengumpulkan data. Teks yang dianalisis diidentifikasi dan dikumpulkan. Ini melibatkan penggunaan bot pengikis web atau antarmuka pemrograman aplikasi pengikis.
- Bersihkan datanya. Data diproses dan dibersihkan untuk menghilangkan noise dan part of speech yang tidak memiliki makna yang relevan dengan sentimen teks. Ini termasuk kontraksi, seperti I'm, dan kata-kata yang memiliki sedikit informasi seperti is, artikel seperti, tanda baca, URL, karakter khusus, dan huruf besar. Ini disebut sebagai standarisasi.
- Ekstrak fitur. Algoritme pembelajaran mesin secara otomatis mengekstrak fitur teks untuk mengidentifikasi sentimen negatif atau positif. Pendekatan ML yang digunakan meliputi teknik bag-of-words yang melacak kemunculan kata-kata dalam sebuah teks dan teknik penyematan kata yang lebih bernuansa yang menggunakan jaringan saraf untuk menganalisis kata-kata dengan makna yang mirip.
- Pilih model ML. Alat analisis sentimen menilai teks menggunakan model ML berbasis aturan, otomatis, atau hibrid. Sistem berbasis aturan melakukan analisis sentimen berdasarkan aturan berbasis leksikon yang telah ditentukan sebelumnya dan sering digunakan dalam domain seperti hukum dan kedokteran di mana diperlukan presisi tingkat tinggi dan kontrol manusia. Sistem otomatis menggunakan ML dan teknik pembelajaran mendalam untuk belajar dari kumpulan data. Model hibrid menggabungkan kedua pendekatan dan umumnya dianggap sebagai model yang paling akurat. Model ini menawarkan pendekatan berbeda untuk menetapkan skor sentimen ke potongan teks.
- Klasifikasi sentimen. Setelah model dipilih dan digunakan untuk menganalisis sepotong teks, itu memberikan skor sentimen ke teks termasuk positif, negatif atau netral. Organisasi juga dapat memutuskan

untuk melihat hasil analisis mereka pada tingkat yang berbeda, termasuk tingkat dokumen, yang sebagian besar berkaitan dengan tinjauan dan cakupan profesional; tingkat kalimat untuk komentar dan ulasan pelanggan; dan tingkat sub-kalimat, yang mengidentifikasi frase atau klausa dalam kalimat.

### B. Sumber Data

Data diambil/web scraping dari review/komentar pengguna pada setiap halaman playstore 5 E-commerce terbesar di Indonesia, yang digunakan dalam penelitian ini yaitu : Tokopedia, Shopee, Bukalapak, Lazada, Blibli, dengan tampilan sebagai berikut :



Gambar 1. Sumber Data Penelitian, Review User di 5 E-commerce Terbesar di Indonesia

### C. Teknik Pengambilan Data

Data diambil dengan teknik *web scraping* dari *reviews user* di Playstore 5 E-commerce terbesar di Indonesia, *web scraping* menggunakan bahasa pemrograman python 3.11.4 dengan library google-play-scraper.

Data diambil secara *sampling* sebanyak 10.000 record data, yang akan membentuk panjang *document term matrix* (dtm) sebesar 59.981.785, seperti tampak dari gambar berikut :

```
m matrix 53981785 413.1 MB Large matrix (53981785 elements, 433.1 MB)
```

Gambar 2. Besaran *Document Term Matrix* yang terbentuk penelitian ini

Dikarenakan besarnya matrix yang terbentuk sehingga *data sampling* dibatasi dikarenakan keterbatasan kekuatan CPU yang digunakan.

Jumlah data yang digunakan dalam penelitian ini sebanyak 10.000 record data, yang terdiri dari :

- Jumlah data yang digunakan dalam penelitian ini sebanyak 10.000 record data, yang terdiri dari :
- Data *reviews users playstore* Tokopedia : 2000 record data, tanggal data antara June/9/2023 15:22 sampai dengan tanggal July/31/2023 23:45
- Data *reviews users playstore* Shopee : 2000 record data, tanggal data antara July/19/2023 0:13 sampai dengan tanggal July/31/2023 23:51
- Data *reviews users playstore* Bukalapak: 2000 record data, tanggal data antara June/4/2021 15:15 sampai dengan tanggal July/31/2023 23:57

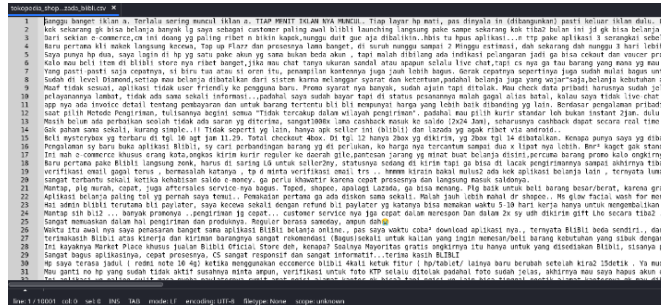
- Data reviews users playstore Lazada : 2000 record data, tanggal data antara July/31/2023 23:48 sampai dengan tanggal July/14/2023 19:40
- Data reviews users playstore Blibli : 2000 record data, tanggal data antara Nov/28/2022 11:01 sampai dengan tanggal July/31/2023 19:39

### III. HASIL DAN PEMBAHASAN

#### A. Persiapan Tools

Bahasa pemrograman yang digunakan dalam penelitian ini adalah R Programming versi 4.3.1 digunakan untuk proses *sentiment analysis* dan Python versi 3.11.4 digunakan untuk *web scraping*. IDE (*Integrated Development Environments*) R Programming yang digunakan adalah R Studio 2023.06.2 Build 561. Pustaka pemrograman R yang digunakan dalam penelitian ini sebagai berikut : ggplot2, NLP, tm, SnowballC, RColorBrewer, wordcloud, plyr, dplyr, stringr.

#### B. Unggah Data ke Corpus



Gambar 3. Raw Data dalam bentuk csv

Raw Data 10.000 record dalam bentuk tipe file csv, diberi nama "tokopedia\_shopee\_bukalapak\_lazada\_blibli.csv", di loading ke variabel filepath dan text.

#### C. Transformasi dan Pembersihan Teks

Transformasi dilakukan menggunakan fungsi `tm_map()` :

- Mengganti, beberapa kata dari teks, mengganti “/”, “@” dan “” dengan spasi
- Konversi teks ke huruf kecil semuanya
- Menghapus nama-nama bulan Indonesia
- Menghapus nama-nama bulan English
- Menghapus beberapa kata khusus
- Menghapus angka
- Menghapus punctuations
- Menghapus spasi berlebih

#### D. Stop Words

Stopwords adalah setiap kata dalam stop list (atau stoplist atau kamus negatif) yang disaring (yaitu dihentikan) sebelum atau setelah pemrosesan data bahasa alami (teks) [27].

Stop words adalah kata umum (common words) yang biasanya muncul dalam jumlah besar dan dianggap tidak memiliki makna [26].

Karena studi kasus ini kita pakai data bahasa Indonesia maka koleksi kata stopwords yang digunakan yang bahasa Indonesia, sebagai berikut :

Menghapus stopwords dari corpus :  
`myStopwords = readLines("id.stopwords.02.01.2016.txt")`  
`docs <- tm_map(docs, removeWords, myStopwords)`  
 Menghapus secara manual kata-kata yang tidak diinginkan untuk diproses :  
`docs <- tm_map(docs, removeWords, c("tokopedia", "bukalapak", "blibli", "lazada", "shopee", "aplika si", "app", "just", "cache", "even", "internet", "use", "like", "time"))`.

#### E. Membuat Term Document Matrix

Document Term Matrix merupakan sebuah representasi numerik dari dokumen dalam corpus. Corpus hanyalah kumpulan dokumen. Dengan kumpulan kata yang “lebih besar” ini, melalui Document Term Matrix, dapat dilakukan analisis yang lebih menarik [29].

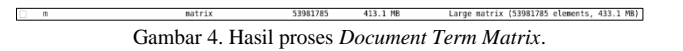
Membuat term document matrix adalah adalah cara untuk merepresentasikan kata-kata dalam teks sebagai tabel atau matriks angka. Baris-baris matriks mewakili tanggapan teks yang akan dianalisis, dan kolom-kolom matriks mewakili kata-kata dari teks yang akan digunakan dalam analisis [28].

Document Term Matrix merupakan algoritma – Metode perhitungan yang sering ditemui dalam text minning. Mudah untuk menentukan jumlah kata individual untuk setiap dokumen atau untuk semua dokumen. Misalkan untuk menghitung agregat dan statistik dasar seperti jumlah istilah rata-rata, mean, median, mode, varians, dan deviasi standar dari panjang dokumen, serta dapat mengetahui istilah mana yang lebih sering dalam kumpulan dokumen dan dapat menggunakan informasi tersebut untuk menentukan istilah mana yang lebih mungkin mewakili dokumen tersebut.

Dengan proses document-term-matrix sebagai berikut :

```
dtm <- TermDocumentMatrix(docs)
m <- as.matrix(dtm)
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)
```

Hasil proses document-term-matrix :  
 Dari 10.000 record data yg digunakan membentuk panjang *document term matrix* (dtm) sebesar 59.981.785, seperti tampak dari gambar berikut :



Gambar 4. Hasil proses Document Term Matrix.

#### F. Menampilkan 50 Kata sering muncul

Setelah proses document-term-matrix selesai, coba ditampilan 50 data pertama yang paling sering muncul, word | freq : nya 2677, barang 1988, gak 1713, ongkir 1210, bagus 1066, aja 1060, pengiriman 890, udah 878, gratis 852, tolong 815, banget 813, biaya 766, toko 685, beli 662, mudah 640, cepat 624, kasih 618, transaksi 605, pake 596, buka 587, pakai 567, bayar 547, kalo 545, pembayaran 544, promo 542, sesuai 521, akun 516, pesanan 504, membantu 482, apk 477, kecewa 467, suka 461, kurir 447, update 441, murah 433, voucher 433, bintang 432,

paylater 432, masuk 431, harga 422, mahal 411, lemot 410, kali 403, online 403, dana 400, produk 392, semoga 373, susah 361, penjual 359.

G. Membentuk Word Cloud pertama

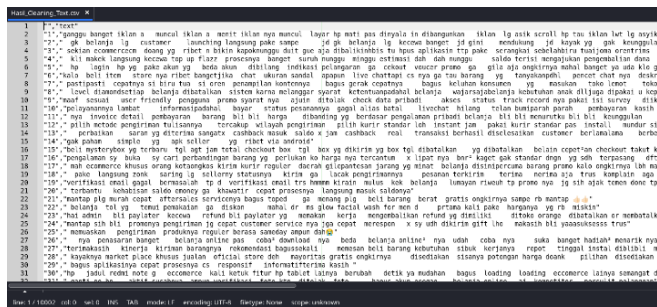
Sampai sini data sudah dibersihkan dan dikonversi, tapi belum dilakukan klasifikasi sentiment analysis, data hasil term-document-matrix akan dibentuk visualisasi word cloud :



Gambar 5. Visualisasi word cloud pertama sebelum proses klasifikasi sentiment analysis

H. Menyimpan Data hasil Pembersihan

Simpan hasil cleaning text, disimpan dahulu hasil proses pembersihan dan konversi :



Gambar 6. Data teks yang sudah dibersihkan

I. Frequent Term - Kata yang muncul 150 kali

Sebelum proses sentiment analysis coba menjelajahi kata yang sering digunakan dan asosiasinya, anda dapat melihat frequent term dalam term-document matrix sebagai berikut.

Dalam contoh ini mencari kata yang muncul setidaknya kata yang sama muncul 150 kali : findFreqTerms(dtm, lowfreq = 150)

Table showing the output of findFreqTerms(dtm, lowfreq = 150) with columns for term frequency and the terms themselves.

Gambar 7. Seratus lima puluh (150) Kata yang paling sering muncul dalam penelitian ini

J. Proses Skoring Sentiment Analysis

Setelah proses pembersihan dan konversi diatas kita mulai proses sentiment analysis, akan memilah kata atau kalimat yang masuk sebagai kata atau kalimat : positif atau negatif atau netral, dalam proses ini dibutuhkan file kata positif dan file kata negatif dalam bahasa Indonesia.

Hasil proses skoring sentiment analysis sebagai berikut, terdiri dari 2 kolom score dan text (teks sentimen), score dalam range angka nilai positif, 0, dan negatif.

Tampak gambar 8, contoh hasil skoring penelitian ini:

Table showing the results of sentiment scoring for 10,000 records, with columns for 'score' and 'text'.

Gambar 8. Hasil Skoring 10.000 record data teks sentimen

Proses selanjutnya klasifikasi konversi nilai score menjadi kata negatif, netral dan positif, sebagai berikut :

Table showing the conversion of numerical scores into sentiment labels (Netral, Negatif, Positif) for the same 10,000 records.

Gambar 9. Konversi Nilai Score ke Klasifikasi Positif, Netral, Negatif

Proses selanjutnya menghapus kolom komentar yang kosong kemudian di export menjadi suatu file dengan nama "Hasil\_sentiment\_pos\_net\_neg.csv", dengan proses dan hasil sebagai berikut :







Word	Count	Word	Count	Word	Count	Word	Count
(1) "jaja"	49	(17) "kecewa"	48	(33) "pengiriman"	47	(49) "tanggul"	46
(9) "ongki"	46	(25) "gratis"	45	(41) "menarik"	44	(57) "apka"	43
(17) "kecewa"	45	(33) "pengiriman"	44	(49) "tanggul"	43	(65) "laya"	42
(25) "gratis"	44	(41) "menarik"	43	(57) "apka"	42	(73) "neat"	41
(33) "pengiriman"	43	(49) "tanggul"	42	(65) "laya"	41	(81) "pembelian"	40
(41) "menarik"	42	(57) "apka"	41	(73) "neat"	40	(89) "masa"	39
(49) "tanggul"	41	(65) "laya"	40	(81) "pembelian"	39	(97) "dla"	38
(57) "apka"	40	(73) "neat"	39	(89) "masa"	38	(105) "budang"	37
(65) "laya"	39	(81) "pembelian"	38	(97) "dla"	37	(113) "bagi"	36
(73) "neat"	38	(89) "masa"	37	(105) "budang"	36	(121) "pengembalian"	35
(81) "pembelian"	37	(97) "dla"	36	(113) "bagi"	35	(129) "dapat"	34
(89) "masa"	36	(105) "budang"	35	(121) "pengembalian"	34	(137) "respon"	33
(97) "dla"	35	(113) "bagi"	34	(129) "dapat"	33	(145) "tarik"	32
(105) "budang"	34	(121) "pengembalian"	33	(137) "respon"	32	(153) "pengirimannya"	31
(113) "bagi"	33	(129) "dapat"	32	(145) "tarik"	31	(161) "alasan"	30
(121) "pengembalian"	32	(137) "respon"	31	(153) "pengirimannya"	30	(169) "alamat"	29
(129) "dapat"	31	(145) "tarik"	30	(161) "alasan"	29	(177) "jua"	28
(137) "respon"	30	(153) "pengirimannya"	29	(169) "alamat"	28	(185) "email"	27
(145) "tarik"	29	(161) "alasan"	28	(177) "jua"	27	(193) "paket"	26
(153) "pengirimannya"	28	(169) "alamat"	27	(185) "email"	26	(201) "ratur"	25
(161) "alasan"	27	(177) "jua"	26	(193) "paket"	25	(209) "loadng"	24
(169) "alamat"	26	(185) "email"	25	(201) "ratur"	24	(217) "status"	23
(177) "jua"	25	(193) "paket"	24	(209) "loadng"	23	(225) "membeli"	22
(185) "email"	24	(201) "ratur"	23	(217) "status"	22	(233) "dipake"	21
(193) "paket"	23	(209) "loadng"	22	(225) "membeli"	21	(241) "bonus"	20
(201) "ratur"	22	(217) "status"	21	(233) "dipake"	20	(249) "ksh"	19
(209) "loadng"	21	(225) "membeli"	20	(241) "bonus"	19	(257) "terimakasih"	18
(217) "status"	20	(233) "dipake"	19	(249) "ksh"	18	(265) "pindah"	17
(225) "membeli"	19	(241) "bonus"	18	(257) "terimakasih"	17	(273) "lambat"	16
(233) "dipake"	18	(249) "ksh"	17	(265) "pindah"	16		
(241) "bonus"	17	(257) "terimakasih"	16	(273) "lambat"	15		
(249) "ksh"	16						
(257) "terimakasih"	15						
(265) "pindah"	14						
(273) "lambat"	13						

Gambar 23. Kata yang muncul di ulang sebanyak 40 kali terkonotasi positif

Penelitian Analisis sentimen 5 *E-commerce* terbesar di Indonesia (Tokopedia, Shopee, Bukalapak, Lazada, Blibli) menggunakan *Text Mining* dan *Natural Language Processing* (NLP), *E-commerce* telah berkembang di Indonesia dengan nilai pasar bruto mencapai US\$ 32 miliar, perkiraan Google, Bain, dan Temasek mencapai US\$ 83 miliar pada tahun 2025, sudah 138,1 juta orang yang telah berbelanja online di Indonesia [1].

Tetapi sentimen negatif, positif dan netral terhadap *E-commerce* tersebut tidak bisa dihindari, mulai dari ketidakpuasan layanan pelanggan, aplikasi yang lambat, sulit diakses, data yang hilang, penjual yang tidak amanah, komentar-komentar pelanggan ini tercurah dalam sentimen tulisan yang negatif, positif maupun netral. Penelitian ini mencoba menggambarkan komparasi sentimen terhadap 5 *E-commerce* tersebut, sehingga dihasilkan analisis sentimen membantu mengumpulkan nada emosional di balik isi teks tentang sentimen pelanggan, pengalaman pelanggan dan reputasi merek *E-commerce*.

Data diambil secara *web scraping* menggunakan bahasa pemrograman python 3.11.4 dengan library google-play-scraper diambil dari review/komentar pengguna pada setiap halaman playstore 5 *E-commerce* terbesar di Indonesia, yang digunakan dalam penelitian ini yaitu : Tokopedia, Shopee, Bukalapak, Lazada, Blibli.

Data diambil secara *sampling* sebanyak 10.000 record data, yang akan membentuk panjang *document term matrix* (dtm) sebesar 167.136.712, dikarenakan besarnya matrix yang terbentuk sehingga *data sampling* dibatasi dikarenakan keterbatasan kekuatan CPU yang digunakan.

Bahasa pemrograman yang digunakan dalam penelitian ini adalah R Programming versi 4.3.1 digunakan untuk proses *sentiment analysis* dan Python versi 3.11.4 digunakan untuk *web scraping*, dengan IDE (*Integrated Development Environments*) R Programming yang digunakan adalah R Studio 2023.06.2 Build 561.

#### IV. KESIMPULAN

Bisa disimpulkan komentar pengguna atau *reviews user* lima (5) *E-commerce* terbesar di Indonesia (Tokopedia, Shopee, Bukalapak, Lazada, Blibli) menunjukkan kalimat

positif menunjukkan apresiasi kepuasan pengguna (3664 kalimat), kalimat netral menunjukkan apresiasi biasa-biasa pengguna (2282 kalimat) dan negatif menunjukkan ketidakpuasan pengguna (4054 kalimat), setidaknya dengan kalimat positif dan netral lebih banyak, menunjukkan 59.46% pengguna *E-commerce* di Indonesia menyatakan baik kinerja 5 *E-commerce* terbesar di Indonesia.

#### UCAPAN TERIMA KASIH

Ucapan terima kasih ditujukan kepada Yayasan Widyatama, Rektorat Universitas Widyatama, Lembaga Penelitian, Pengabdian kepada Masyarakat & Modal Intelektual Universitas Widyatama (LP2M Utama) dukungan penelitian ini.

#### DAFTAR PUSTAKA

- [1] Medium, "The Top 10 Marketplace E-Commerce in Indonesia in 2021 | by 9cv9 official | Medium." [Online]. Available: <https://medium.com/@9cv9official/the-top-10-marketplace-e-commerce-in-indonesia-in-2021-6846d699345b>. [Accessed: 16-Mar-2023].
- [2] MonkeyLearn, "Sentiment Analysis Guide." [Online]. Available: <https://monkeylearn.com/sentiment-analysis/>. [Accessed: 16-Mar-2023].
- [3] L. Yang, Y. Li, J. Wang, and R. S. Sherratt, "Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning," *IEEE Access*, vol. 8, pp. 23522–23530, 2020.
- [4] Y. Liu, J. Lu, J. Yang, and F. Mao, "Sentiment analysis for e-commerce product reviews by deep learning model of Bert-BiGRU-Softmax," *Math. Biosci. Eng.*, vol. 17, no. 6, pp. 7819–7837, Nov. 2020.
- [5] S. Zhang and H. Zhong, "Mining Users Trust From E-Commerce Reviews Based on Sentiment Similarity Analysis," *IEEE Access*, vol. 7, pp. 13523–13535, 2019.
- [6] A. Bayhaqy, S. Sfenrianto, K. Nainggolan, and E. R. Kaburuan, "Sentiment Analysis about E-Commerce from Tweets Using Decision Tree, K-Nearest Neighbor, and Naïve Bayes," *2018 Int. Conf. Orange Technol. ICOT 2018*, Jul. 2018.
- [7] S. Vanaja and M. Belwal, "Aspect-Level Sentiment Analysis on E-Commerce Data," *Proc. Int. Conf. Inven. Res. Comput. ICIRCA 2018*, pp. 1275–1279, Dec. 2018.
- [8] J. Jabbar, I. Urooj, W. Junsheng, and N. Azeem, "Real-time sentiment analysis on E-Commerce application," *Proc. 2019 IEEE 16th Int. Conf. Networking, Sens. Control. ICNSC 2019*, pp. 391–396, May 2019.
- [9] S. Kumar, M. Gahalawat, P. P. Roy, D. P. Dogra, and B. G. Kim, "Exploring impact of age and gender on sentiment analysis using machine learning," *Electron.*, vol. 9, no. 2, Feb. 2020.
- [10] K. K. Tseng, R. F. Y. Lin, H. Zhou, K. J. Kurniajaya, and Q. Li, "Price prediction of e-commerce products through Internet sentiment analysis," *Electron. Commer. Res.*, vol. 18, no. 1, pp. 65–88, Mar. 2018.
- [11] S. Zhang, D. Zhang, H. Zhong, and G. Wang, "A multiclassification model of sentiment for e-commerce reviews," *IEEE Access*, vol. 8, pp. 189513–189526, 2020.
- [12] N. Barney, "What Is Sentiment Analysis (Opinion Mining)? | Definition from TechTarget." [Online]. Available: <https://www.techtarget.com/searchbusinessanalytics/definition/opinion-mining-sentiment-mining>. [Accessed: 16-Mar-2023].
- [13] S. Gupta, "Sentiment Analysis: Concept, Analysis and Applications | by Shashank Gupta | Towards Data Science." [Online]. Available: <https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17>. [Accessed: 16-Mar-2023].

- [14] A. AWS, "Apa itu Analisis Sentimen? - Penjelasan tentang Analisis Sentimen - AWS." [Online]. Available: <https://aws.amazon.com/id/what-is/sentiment-analysis/>. [Accessed: 16-Mar-2023].
- [15] STHDA, "Text mining and word cloud fundamentals in R : 5 simple steps you should know - Easy Guides - Wiki - STHDA." [Online]. Available: <http://www.sthda.com/english/wiki/text-mining-and-word-cloud-fundamentals-in-r-5-simple-steps-you-should-know>. [Accessed: 16-Mar-2023].
- [16] A. Salcedo, "RPubs - Sentiment Analysis R." [Online]. Available: <https://rpubs.com/aneudissalcedo/week5-homework>. [Accessed: 16-Mar-2023].
- [17] T. S. Holliger, "RPubs - A Practical Application of the TM Package." [Online]. Available: <https://rpubs.com/tsholliger/301914>. [Accessed: 16-Mar-2023].
- [18] P. Chris Bail and D. University, "Basic Text Analysis in R." [Online]. Available: [https://sicss.io/2020/materials/day3-text-analysis/basic-text-analysis/rmarkdown/Basic\\_Text\\_Analysis\\_in\\_R.html](https://sicss.io/2020/materials/day3-text-analysis/basic-text-analysis/rmarkdown/Basic_Text_Analysis_in_R.html). [Accessed: 16-Mar-2023].
- [19] Afxwilhelm, "Introduction to Text Mining Package (TM)." [Online]. Available: <http://afxwilhelm.github.io/statsWithR/tutorials/textMiningIntro.html>. [Accessed: 16-Mar-2023].
- [20] Geeksforgeeks, "Filter data by multiple conditions in R using Dplyr - GeeksforGeeks." [Online]. Available: <https://www.geeksforgeeks.org/filter-data-by-multiple-conditions-in-r-using-dplyr/>. [Accessed: 16-Mar-2023].
- [21] Supriyd, "GitHub - supriyd/Sentiment-Analysis-with-R: Analisis sentimen kali ini data yang digunakan adalah ulasa maskapai lion air yang telah diunduh pada postingan sebelumnya. Analisis dilakukan dengan menggunakan program R. semoga membantu terimakasih." [Online]. Available: <https://github.com/supriyd/Sentiment-Analysis-with-R>. [Accessed: 16-Mar-2023].
- [22] L. Florence, "RPubs - Sentiment Analysis Sriwijaya Air." [Online]. Available: <https://rpubs.com/LauraEflor/sasriwijaya>. [Accessed: 16-Mar-2023].
- [23] S. R. A. E. Virgana Targa, "Web scraping with Chrome Extensions - Web Scraper - Free Web Scraping." [Online]. Available: [https://www.youtube.com/watch?v=a-Nqe\\_GDoGU](https://www.youtube.com/watch?v=a-Nqe_GDoGU). [Accessed: 16-Mar-2023].
- [24] A. Chandramohan, "What is corpus in R? - Quora." [Online]. Available: <https://www.quora.com/What-is-corpus-in-R>. [Accessed: 16-Mar-2023].
- [25] C. Project, "Introduction to corpus." [Online]. Available: <https://cran.r-project.org/web/packages/corpus/vignettes/corpus.html>. [Accessed: 16-Mar-2023].
- [26] C. Khanna, "Text preprocessing: Stop words removal | Chetna | Towards Data Science." [Online]. Available: <https://towardsdatascience.com/text-pre-processing-stop-words-removal-using-different-libraries-f20bac19929a>. [Accessed: 16-Mar-2023].
- [27] Y. Wibisono, "Stop words untuk Bahasa Indonesia." [Online]. Available: <https://yudiwbs.wordpress.com/2008/07/23/stop-words-untuk-bahasa-indonesia/>. [Accessed: 16-Mar-2023].
- [28] C. Facer, "Text Analysis: Hooking up Your Term Document Matrix to Custom R Code - Displayr." [Online]. Available: <https://www.displayr.com/text-analysis-hooking-up-your-term-document-matrix-to-custom-r-code/>. [Accessed: 16-Mar-2023].
- [29] Softscients, "Membuat Document Term Matrix - Softscients." [Online]. Available: <https://softscients.com/2021/02/16/membuat-document-term-matrix/>. [Accessed: 16-Mar-2023]