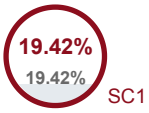
# Similarity Report

## Metadata

Name of the organization
**Universitas Muhammadiyah Sidoarjo**
Title
**Analisis Sentimen Pengguna Aplikasi Shopee**
Author(s) Coordinator
**perpustakaan umsidaprist**
Organizational unit
**Perpustakaan**

## Record of similarities

SCs indicate the percentage of the number of words found in other texts compared to the total number of words in the analysed document. Please note that high coefficient values do not automatically mean plagiarism. The report must be analyzed by an authorized person.




**3971**
Length in words

**29568**
Length in characters

## Alerts

In this section, you can find information regarding text modifications that may aim at temper with the analysis results. Invisible to the person evaluating the content of the document on a printout or in a file, they influence the phrases compared during text analysis (by causing intended misspellings) to conceal borrowings as well as to falsify values in the Similarity Report. It should be assessed whether the modifications are intentional or not.

| | | |
|---|---|---|
| Characters from another alphabet | | 0 |
| Spreads | A→ | 0 |
| Micro spaces | | 0 |
| Hidden characters | | 0 |
| Paraphrases (SmartMarks) | a | 58 |

## Active lists of similarities

This list of sources below contains sources from various databases. The color of the text indicates in which source it was found. These sources and Similarity Coefficient values do not reflect direct plagiarism. It is necessary to open each source, analyze the content and correctness of the source crediting.

The 10 longest fragments — Color of the text

| NO | TITLE OR SOURCE URL (DATABASE) | NUMBER OF IDENTICAL WORDS (FRAGMENTS) |
|---|---|---|
| 1 | Sentiment Analysis of the Dana Application Reviews with the Random Forest Method Hanggara Buce Trias, Dian Eka Ratnawati,Larasati Fanka Angelina; | 41 1.03 % |
| 2 | Sentiment Analysis of E-Grocery Application Reviews Using Lexicon-Based and Support Vector Machine Dian Ardiansyah, Eka Fitriani,Riska Aryanti, Atang Saepudin, Royadi Royadi; | 37 0.93 % |
| 3 | Sentiment Analysis of the Dana Application Reviews with the Random Forest Method Hanggara Buce Trias, Dian Eka Ratnawati,Larasati Fanka Angelina; | 35 0.88 % |

| | | |
|---|---|---|
| 4 | SENTIMENT ANALYSIS OF POST-COVID-19 INFLATION BASED ON TWITTER USING THE K-NEAREST NEIGHBOR AND SUPPORT VECTOR MACHINE CLASSIFICATION METHODS<br>Ratih Puspitasari, Yulian Findawati, Rosid Mochamad Alfan; | 30 0.76 % |
| 5 | Analisis Sentimen Kepuasan Pengguna Terhadap Layanan Streaming Mola Menggunakan Algoritma Random Forest<br>Nanda Setya, Fitri Diah Angraina, Desti Mualfah; | 29 0.73 % |
| 6 | SENTIMENT ANALYSIS OF POST-COVID-19 INFLATION BASED ON TWITTER USING THE K-NEAREST NEIGHBOR AND SUPPORT VECTOR MACHINE CLASSIFICATION METHODS<br>Ratih Puspitasari, Yulian Findawati, Rosid Mochamad Alfan; | 28 0.71 % |
| 7 | Analisis Sentimen Masyarakat Terhadap Penerima Beasiswa Kartu Indonesia Pintar Kuliah Dengan Metode Support Vector Machine<br>Yudha Dwi Putra Negara,Jannan M Fahriz Zain; | 27 0.68 % |
| 8 | Analisis Sentimen Kepuasan Pengguna Terhadap Layanan Streaming Mola Menggunakan Algoritma Random Forest<br>Nanda Setya, Fitri Diah Angraina, Desti Mualfah; | 26 0.65 % |
| 9 | Analisis Sentimen Kepuasan Pengguna Terhadap Layanan Streaming Mola Menggunakan Algoritma Random Forest<br>Nanda Setya, Fitri Diah Angraina, Desti Mualfah; | 25 0.63 % |
| 10 | http://36.95.239.66/1209/8/Bab2_D1041171012.pdf | 23 0.58 % |

## from RefBooks database (14.78 %)

| NO | TITLE | NUMBER OF IDENTICAL WORDS (FRAGMENTS) |
|---|---|---|
| **Source: Paperity** | | |
| **1** | Analisis Sentimen Kepuasan Pengguna Terhadap Layanan Streaming Mola Menggunakan Algoritma Random Forest<br>Nanda Setya, Fitri Diah Angraina, Desti Mualfah; | **175 (10) 4.41 %** |
| **2** | Sentiment Analysis of the Dana Application Reviews with the Random Forest Method<br>Hanggara Buce Trias, Dian Eka Ratnawati,Larasati Fanka Angelina; | **146 (11) 3.68 %** |
| 3 | ANALISIS SENTIMEN ULASAN PENGGUNA APLIKASI SIREKAP PADA PLAY STORE MENGGUNAKAN ALGORITMA RANDOM FOREST CLASSIFER<br>Carudin Carudin,Herjanto Muhamad Fajar Yudhistira; | 98 (8) 2.47 % |
| **4** | SENTIMENT ANALYSIS OF POST-COVID-19 INFLATION BASED ON TWITTER USING THE K-NEAREST NEIGHBOR AND SUPPORT VECTOR MACHINE CLASSIFICATION METHODS<br>Ratih Puspitasari, Yulian Findawati, Rosid Mochamad Alfan; | **85 (6) 2.14 %** |
| **5** | Sentiment Analysis of E-Grocery Application Reviews Using Lexicon-Based and Support Vector Machine<br>Dian Ardiansyah, Eka Fitriani,Riska Aryanti, Atang Saepudin, Royadi Royadi; | **45 (2) 1.13 %** |
| **6** | Analisis Sentimen Masyarakat Terhadap Penerima Beasiswa Kartu Indonesia Pintar Kuliah Dengan Metode Support Vector Machine<br>Yudha Dwi Putra Negara,Jannan M Fahriz Zain; | **27 (1) 0.68 %** |
| 7 | Analisis Sentimen Ulasan Pengguna Aplikasi Mobile Gapura UB pada Google Play Store Menggunakan Algoritma Support Vector Machine<br>Viriya Aurelius Alexaner, Setiawan Nanang Yudi, Maghfiroh Intan Sartika Eris; | 6 (1) 0.15 % |
| 8 | Application Of Text Mining In Grouping Thesis Topics Using TF-IDF Method Based On Thesis Abstract<br>Dewi Suranti, Trianggara Dimas Aulia,Delsy Andreswari; | 5 (1) 0.13 % |

## from the home database (0.00 %)

| NO | TITLE | NUMBER OF IDENTICAL WORDS (FRAGMENTS) |
|---|---|---|

## from the Database Exchange Program (0.00 %)

| NO | TITLE | NUMBER OF IDENTICAL WORDS (FRAGMENTS) |
|---|---|---|

## from the Internet (4.63 %)

| NO | SOURCE URL | NUMBER OF IDENTICAL WORDS (FRAGMENTS) |
|---|---|---|
| 1 | https://media.neliti.com/media/publications/496361-comparison-of-random-forest-and-support-6b4df15e.pdf | 47 (4) 1.18 % |
| 2 | https://media.neliti.com/media/publications/442344-none-1c546cb3.pdf | 38 (3) 0.96 % |
| 3 | http://36.95.239.66/1209/8/Bab2_D1041171012.pdf | 36 (2) 0.91 % |
| 4 | https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/download/14482/6461/103373 | 14 (2) 0.35 % |
| 5 | https://journal.eng.unila.ac.id/index.php/jitet/article/download/5155/2096 | 13 (2) 0.33 % |
| 6 | https://ejournal.itn.ac.id/index.php/jati/article/download/13675/7595 | 13 (2) 0.33 % |
| 7 | https://jom.fti.budiluhur.ac.id/index.php/SKANIKA/article/download/3193/1369/ | 9 (1) 0.23 % |
| 8 | https://publikasiilmiah.unwahas.ac.id/JINRPL/article/download/10302/pdf | 8 (1) 0.20 % |
| 9 | https://www.pkm.tunasbangsa.ac.id/index.php/kesatria/article/download/577/572 | 6 (1) 0.15 % |

## **List of accepted fragments** (no accepted fragments)

| NO | CONTENTS | NUMBER OF IDENTICAL WORDS (FRAGMENTS) |
|---|---|---|





Sentiment Analysis of Shopee App Users on Google Play Store
Using **the Random Forest Method**
**Analisis** Sentimen Pengguna **Aplikasi** Shopee Pada **Google Play Store Menggunakan** Metode Random Forest

Abstract. This study focuses on sentiment analysis to evaluate customer satisfaction with the Shopee app, using comments posted on the Google Play Store as the primary data source. A total of 5,000 comment data were collected over a relevant timeframe, from December 2024 to March 2025. The methodology applied was classification using the Random Forest Classifier algorithm. The analysis results show that the dominant sentiment expressed by users is positive, indicating a good level of satisfaction with the app. The Random Forest model successfully achieved an accuracy of 88%. This figure indicates that the algorithm is quite effective in classifying user comment sentiment. As a key contribution, this study provides up-to-date insights into customer perceptions thanks to the use of very recent data. These findings not only validate the effectiveness of Random Forest in sentiment analysis tasks but also provide valuable information for Shopee to understand user views and make strategic decisions to improve services.
Keywords - Sentiment Analysis; Random Forest; Customer Satisfaction; Shopee; Play Store.

Abstrak. Penelitian ini berfokus pada analisis sentimen untuk mengevaluasi kepuasan pelanggan terhadap aplikasi Shopee, dengan menggunakan komentar-komentar yang diunggah di Google Play Store sebagai sumber data utama. Sebanyak 5.000 data komentar dikumpulkan dalam rentang waktu yang relevan, yaitu dari Desember 2024 hingga Maret 2025. Metodologi yang diterapkan adalah klasifikasi dengan menggunakan algoritma Random Forest Classifier. Hasil analisis menunjukkan bahwa sentimen dominan yang diekspresikan oleh pengguna adalah positif, yang mengindikasikan tingkat kepuasan yang baik terhadap aplikasi tersebut. Model Random Forest yang digunakan berhasil mencapai nilai akurasi sebesar 88%. Angka ini menunjukkan bahwa algoritma tersebut cukup efektif dalam mengklasifikasikan sentimen komentar pengguna. Sebagai kontribusi utama, penelitian ini menyediakan wawasan terkini mengenai persepsi pelanggan berkat penggunaan data yang sangat baru. Temuan ini tidak hanya memvalidasi

efektivitas Random Forest dalam tugas analisis sentimen, tetapi juga memberikan informasi berharga bagi pihak Shopee untuk memahami pandangan pengguna dan membuat keputusan strategis guna meningkatkan layanan.

Kata Kunci - Analisis Sentimen; Random Forest; Kepuasan Pelanggan; Shopee; Play Store

## I. PENDAHULUAN

Aplikasi e-commerce seperti Shopee telah menjadi bagian integral dari kehidupan sehari-hari, memfasilitasi aktivitas jual beli secara online dengan beragam fitur menarik. Kepuasan pelanggan menjadi kunci dalam lanskap digital yang kompetitif ini, dipengaruhi oleh kemudahan navigasi, keamanan transaksi, kelengkapan fitur, dan kecepatan layanan (Rizky & Mahfudz, 2022. Ulasan pengguna di Google Play Store menjadi indikator krusial yang mencerminkan pengalaman mereka, meliputi penilaian terhadap fitur, kemudahan penggunaan, kecepatan transaksi, hingga kualitas layanan pelanggan. Analisis sentimen terhadap ulasan ini menjadi sangat penting untuk memahami opini publik dan memperoleh umpan balik berharga, yang kemudian dapat dimanfaatkan untuk perbaikan dan pengembangan aplikasi guna meningkatkan loyalitas pelanggan dan menarik pengguna baru. Data ulasan aplikasi Shopee diperoleh melalui **scraping menggunakan API Google-Play-Scraper. Google-Play-Scraper adalah API** yang memungkinkan dengan mudah **mengekstraksi data informasi aplikasi dan ulasan aplikasi dari Google Play** Store **tanpa** bergantung pada eksternal (Fahmi & Sumarno, 2022). Data yang diekstrak kemudian akan diproses melalui proses text preprocessing. Pendekatan ini melibatkan pengolahan teks semi-terstruktur dari ulasan menjadi data terorganisir, yang kemudian akan diklasifikasikan ke dalam kategori sentimen positif, negatif, atau netral (Arsi et al., 2021). Untuk proses klasifikasi, penelitian ini akan memanfaatkan algoritma Random Forest. Algoritma ini dipilih karena performanya yang unggul dalam klasifikasi data dan kemampuannya menangani data besar serta kompleks tanpa memerlukan banyak penyesuaian model (Amaliah et al., 2022). Proses analisis akan mencakup preprocessing data, ekstraksi fitur, pemodelan, dan evaluasi untuk memastikan akurasi klasifikasi sentimen.

Analisis sentimen, juga dikenal sebagai opinion mining, adalah metode otomatis untuk memahami, mengekstraksi, dan mengolah data teks guna menentukan penilaian positif atau negatif suatu pernyataan (Gifari et al., 2022). Teknik ini krusial untuk mengidentifikasi opini publik terhadap produk, layanan, atau merek, yang dapat menjadi umpan balik vital bagi pengembang. Random Forest merupakan algoritma machine learning supervised yang terbukti efektif dalam klasifikasi teks. Algoritma ini membangun banyak pohon keputusan yang dilatih dengan subset acak data dan fitur, menggabungkan prediksi dari setiap pohon untuk meningkatkan akurasi dan stabilitas model, serta mengurangi overfitting (Mahmuda, 2024).





Penelitian sebelumnya telah banyak meneliti sentimen dan mengklasifikasikan opini pengguna terhadap berbagai objek atau sistem, seringkali memanfaatkan platform media sosial seperti Twitter sebagai sumber data utama karena kemampuannya untuk dilacak dan dievaluasi secara instan (Vonega et al., 2022). Beberapa studi telah menunjukkan bahwa metode Random Forest memiliki performa yang unggul dalam klasifikasi data pada berbagai kasus. Amaliah et al. (2022) dan Mahmudah (2024), secara spesifik menyoroti efektivitas Random Forest dalam klasifikasi teks tradisional. Meskipun demikian, masih terdapat kesenjangan dalam membandingkan sistem penjabaran sentimen menggunakan berbagai pendekatan algoritma, mendorong penelitian ini untuk melakukan perbandingan lebih lanjut. **Penelitian ini bertujuan untuk menganalisis sentimen** pengguna yang diungkapkan dalam **ulasan aplikasi Shopee di Google Play Store** dari Desember 2024 hingga Maret 2025 sebanyak 5.000 data ulasan. Evaluasi yang digunakan yaitu akurasi, recall, presisi dan f1-score. Tujuan utamanya adalah menilai dinamika sentimen masyarakat terhadap aplikasi e-commerce ini. Kebaharuan penelitian ini terletak pada analisis sentimen terkini terhadap pengalaman pengguna Shopee setelah perubahan tren belanja digital, serta upaya membandingkan berbagai prosedur penyelesaian algoritma untuk menemukan pendekatan terbaik dalam memahami opini pengguna. Hasil analisis ini diharapkan memberikan wawasan mendalam dan masukan berguna bagi pengembang aplikasi untuk meningkatkan layanan Shopee.

## II. METODE

Penelitian ini akan mengikuti serangkaian langkah sistematis **untuk menganalisis sentimen ulasan pengguna** aplikasi Shopee **di Google Play Store menggunakan metode** Random Forest. Tahapan yang dilakukan yaitu identifikasi masalah, **pengumpulan** data, text preprosessing, pelabelan **data, pemodelan klasifikasi dengan Random Forest,** dan evaluasi data dan hasil validasi. Diagram alir berikut memberikan gambaran umum prosesnya:

Gambar 1 Alur Penelitian

### A. Analisis Sentimen

Analisis sentimen adalah sebuah teknik komputasi untuk menentukan dan mengklasifikasikan polaritas sentimen yang terkandung dalam suatu teks atau dokumen (Wardani et al., 2020). Tujuan utamanya adalah untuk mengkategorikan teks tersebut sebagai positif, negatif, atau netral. Dalam praktiknya, analisis ini sering diterapkan pada data dari jaringan media sosial seperti Twitter untuk mengevaluasi pandangan masyarakat secara luas

(Kurniawan & Susanto, 2019). Hal ini juga serupa dengan opini mining, di mana fokusnya adalah menambang data tekstual untuk mengekstrak dan menganalisis pendapat yang diungkapkan, baik itu tentang suatu produk, layanan, atau topik spesifik lainnya

B. Random Forest
Random Forest Classifier adalah algoritma klasifikasi ensemble yang berevolusi dari decision tree. Setiap pohon keputusan dalam model ini dilatih secara independen menggunakan subset data dan atribut yang dipilih secara acak (Alvanof & Dinata, 2024). Algoritma ini unggul dalam meningkatkan akurasi, bahkan ketika dihadapkan pada data yang tidak lengkap atau memiliki nilai ekstrem (outliers) (Siregar et al., 2023). Selain efisiensi dalam pengelolaan data, Random Forest juga mampu mengidentifikasi fitur-fitur yang paling relevan untuk meningkatkan kinerja model klasifikasi secara keseluruhan (Erkamim et al., 2023).

C. Identifikasi Masalah
Penelitian ini dilakukan berdasarkan permasalahan yang telah dijelaskan sebelumnya, yaitu:
a. Penerapan metode Random Forest pada analisis ulasan pengguna di aplikasi Shopee di Play Store, untuk mengklasifikasikan teks ulasan ke dalam label positif, negatif, dan netral.
b. Mengetahui tingkat akurasi metode Random Forest dalam menganalisis sentimen pengguna terhadap aplikasi Shopee di tahun 2024.
c. Memberi hasil analisa berupa saran pengembangan aplikasi mengenai ulasan yang telah dihasilkan untuk meningkatkan kepuasan pelanggan pengguna Shoppe.





D. **Pengumpulan** Data
Langkah **pertama** dalam metode penelitian **ini** adalah pengumpulan **data yang diambil dari hasil** crawling ulasan pengguna di Play Store terkait aplikasi Shopee (Puspitasari et al., 2023). Proses crawling dilakukan dengan menggunakan kata kunci seperti "Shopee Review", "Shopee User Experience", "Shopee Delivery Service", dan "Shopee Payment Issues". Data yang dikumpulkan mencakup ulasan dari periode pada jangka Desember 2024 sampai Maret 2025. Pengumpulan data pada penelitian ini diimplementasikan dengan menggunakan bahasa pemrograman Python dalam lingkungan Google Colab. Metode crawling yang digunakan tidak memanfaatkan API, namun memanfaatkan library Scrapy untuk mengumpulkan hingga 5.000 ulasan pengguna.
Dalam penelitian ini, data dikumpulkan dari ulasan pengguna aplikasi Shopee di Play Store. Proses pengumpulan data dilakukan menggunakan Python dengan bantuan library Scrapy sebagai alat untuk scraping data terkait opini pengguna terhadap aplikasi Shopee dalam jangka Desember 2024 sampai Maret 2025. Beberapa teknik diterapkan, seperti menghilangkan data duplikat untuk memastikan data yang diperoleh bersih dan relevan. Proses ini juga menyertakan langkah pemilihan atribut penting, seperti teks ulasan pengguna, untuk dianalisis lebih lanjut. Data yang telah terkumpul kemudian disimpan dalam format Excel untuk memudahkan pengolahan dan analisis pada tahap berikutnya.

E. Text Preprocessing
Text Preprocessing dilakukan untuk membentuk kumpulan data yang siap untuk dianalisis. Pada penelitian ini, dilakukan langkah preprocessing dengan empat langkah, yaitu (Fauzan et al., 2021):
a. Data Cleansing yaitu tahap data pada penelitian ini dibersihkan (cleansing) dari anomali semisal jejak tanda baca dan karakter-karakter yang tidak relevan. Proses cleansing ini dimulai dengan mengeliminasi simbol "@", tagar, karakter- karakter asing, **dan** sebagainya. **Kemudian,** seluruh **huruf besar** diubah menjadi **huruf kecil (case folding)."**
b. Tokenizing merupakan langkah fragmentasi dokumen atau kalimat menjadi unit-unit yang disebut token. Tahap tokenisasi ini menyegmentasi kalimat, **kata, simbol, dan entitas penting** lainnya.
c. Stopwords **Remove merupakan** leksikon **yang tidak** distingtif dalam **dokumen.** Contohnya, "pun", "akan", " di", "nya", "me", "oleh", "dan", dan lain sebagainya. Eliminasi stopword dilakukan untuk mengoptimalkan performa analisis sentimen.
d. Stemming merupakan teknik normalisasi teks dengan cara mereduksi kata menjadi bentuk dasarnya.

F. Pelabelan Data
Ada dua metode untuk pelabelan data yaitu secara manual dan sentiment menggunakan Bahasa pemrograman python.
a. Pelabelan Manual adalah proses mengklasifikasikan sentimen ulasan secara langsung (Wahyuningsih et al., 2022). Dari 100 ulasan Shopee yang diuji, 60 positif, 25 netral, dan 15 negatif ditemukan. Ini menunjukkan kepuasan pengguna terhadap pengiriman, kualitas produk, dan kemudahan transaksi, meski ada keluhan. Mayoritas positif mengindikasikan Shopee memenuhi ekspektasi belanja online pengguna.
b. Pelabelan Phyton digunakan alam penelitian ini dilakukan secara otomatis menggunakan kode Python,

mengidentifikasi ulasan sebagai positif, netral, atau negatif[16]. Data ulasan yang telah diproses akan diolah oleh pustaka VADER Sentiment, yang mendukung bahasa Indonesia. Pendekatan berbasis leksikon VADER (Valence Aware Dictionary for Sentiment Reasoning) diterapkan untuk memberikan label. VADER tidak hanya mengukur polaritas, tetapi juga intensitas sentimen. Nilai sentimen VADER berkisar dari -0.05 hingga 0.05: ≥0.05 untuk positif, =0 untuk netral, dan ≤−0.05 untuk negatif.

G. Klasifikasi Data

Random Forest diukur untuk optimalisasi analisis sentimen ulasan Shopee. Metode klasifikasi supervised learning ini membangun banyak pohon keputusan dari subset acak data dan fitur. Setiap pohon berkontribusi pada hasil klasifikasi akhir melalui mayoritas suara. Keunggulan Random Forest terletak pada kemampuannya menangani data kompleks, menghasilkan prediksi sentimen yang stabil dan akurat pada ulasan pengguna (Wahyuningsih et al., 2022).





Gambar 2 Random Forest

H. Evaluasi dan Validasi Data

Evaluasi kinerja model Random Forest dilakukan setelah pembagian data terstruktur, dengan mengukur **akurasi, presisi, recall, dan F1-Score yang** dihitung menggunakan confusion matrix (Fahmi & Sumarno, 2022). Nilai akurasi yang lebih tinggi **menunjukkan** bahwa algoritma tersebut efektif **dalam mengklasifikasikan sentimen ulasan pengguna aplikasi** Shopee. Kinerja algoritma ini akan diuji untuk menentukan mana yang lebih baik dalam menganalisis sentimen positif, netral, atau negatif dalam konteks ulasan aplikasi Shopee. Metode evaluasi ini bertujuan untuk memilih algoritma yang paling sesuai untuk pengklasifikasian sentimen pengguna berdasarkan data yang ada. Hasil evaluasi ini akan menjadi dasar untuk menyimpulkan algoritma mana yang paling efektif dalam menganalisis sentimen dalam penelitian ini (Warhiyono et al., 2024).

Accuracy = TP + TNTP + FP = TN + FN (3.1)
Precision = TPTP + FP (3.2)
Recall = TPTP + FN (3.3)
f1 − Score = 2TP2TP + FP + FN (3.4)

III. **HASIL** DAN PEMBAHASAN

Hasil **dan pembahasan** analisis sentiment **pada aplikasi** Shopee menggunakan **metode klasifikasi Random Forest sebagai berikut.**

A. Scraping data

Pengumpulan data pada penelitian ini menggunakan API google- **play-store dengan mengumpulkan** sebanyak 5.000 **data ulasan** terkait **aplikasi** Shopee dengan **menggunakan skrip** berikut ini.

Gambar 3 Tahap Coding

**Hasil output dari skrip tersebut dapat dilihat pada gambar** 4.

**Gambar** 4 Hasil **Coding**

B. Text preprocessing

**Selanjutnya** merupakan proses preprocessing, **pada tahap ini data yang telah** terkumpul akan dilakukan pembersihan data **melalui beberapa tahapan yaitu** data cleansing, case **folding,** tokenizing, stopword remove dan **stemming** (Aditya et al., 2024).





1. **Case Folding Tahapan case** folding bertujuan untuk **mengkonversikan keseluruhan teks dari** huruf kapital **menjadi** huruf kecil. Tahapan ini membutuhkan **library lower** untuk **mengubah huruf** kapital yang **ada didokumen menjadi huruf kecil** (Hidayat et al., 2023). **Berikut hasil case folding terdapat pada** gambar 5 berikut.

Gambar 5 Case Folding

2. Data Cleansing

Pada tahapan ini dilakukan **penghapusan elemen-elemen yang tidak relevan** seperti symbol dan **emotikon. Pada tahapan cleansing ini adalah untuk menghapus karakter yang tidak memberikan pengaruh terhadap proses** klasifikasi sentiment seperti **menghapus tanda baca koma (,), titik (.), hastag (#), mention (@), link dan angka** (Erkamim et al., 2023). Tahap ini bertujuan untuk **meningkatkan kualitas data dan membersihkan data dari** noise[ 10]. **Berikut** hasil **dari proses data** cleansing dapat dilihat pada gambar 6

Gambar 6 Data Cleansing

3. Tokenizing

Tahapan **tokenizing** adalah **metode untuk melakukan pemisahan kata dalam suatu kalimat dengan tujuan untuk proses analisis teks lebih lanjut** (Bin et al., 2020). Pada gambar 7 dibawah ini merupakan hasil dari tahapan tokenizing.

Gambar 7 Tokenizing

4. Stopword Remove

Berikut merupakan tahapan **stopword yaitu membersihkan kata penghubung contohnya saya, lagi, dan, aku, yang, dan imbuhan serta kata sambung** lainnya (Dikiyanti et al., 2021). Berikut hasil dari tahapan stopword remove dapat dilihat pada gambar 8 dibawah ini.

Gambar 8 Stopword Remove

5. **Stemming**

**Tahapan Stemming adalah suatu proses pengembalian suatu kata berimbuhan ke dalam bentuk dasarnya** (Ardhani et al., 2021). **Proses** ini menghilangkan **awalan, akhiran, sisipan dan confixes (kombinasi dari awalan dan akhiran).** Stemming yang digunakan pada proses ini **adalah stemming Sastrawi, yang merupakan stemmer pengembangan** dari Algoirtma Nazief **dan Adriani. Sastrawi sangat bergantung pada kamus kata dasar yang diambil dari kateglo.com dengan perubahan** (Khairunnisa et al., 2021).

Gambar 9 Stemming

C. Pelabelan Data





Ulasan masih belum mempunyai sentimen sehingga sulit untuk mencari tahu apakah pengguna memberikan ulasan positif atau negatif (Mukarromah et al., 2021). **Proses** pemberian sentimen tidak **mungkin dilakukan secara manual dengan melihat ulasan secara satu per satu karena membutuhkan waktu yang lama dan memerlukan seorang ahli di bidang bahasa yang dapat menafsirkan ulasan kemudian mengelompokkan ke sentimen positif dan negatif** (Dwiki et al., 2021). Maka dari itu, setelah melalui proses text preprosesing kemudian dilakukan tahapan berikutnya yaitu melakukan **proses pemberian label sentimen dengan menerapkan** metode yang berbasis lexicon atau biasa dikenal dengan lexicon-based method (Hidayati et al., 2021). Kamus yang digunakan adalah InSet Lexicon yang terdiri dari kamus positif dan negatif. Berikut data ditunjukkan pada gambar dibawah ini.

Gambar 10 Pelabelan Data

D. Visualisasi Data

Pada **pengujian ini menggunakan ulasan** aplikasi Shopee yang **telah** dikelompokkan menjadi 3 kelas sentiment **yaitu positif, negatif dan netral bertujuan untuk mencari kata yang sering** muncul Afdal & Elita, 2022). **Dari** 5.000 data **ulasan, ulasan yang** mengandung sentimen **positif** sebanyak 1.470 **ulasan, sentimen** negative sebanyak 2.539 ulasan **dan sentimen** netral sebanyak 991 ulasan. **Analisis ini dilakukan menggunakan wordcloud agar dapat menampilkan kata yang sering muncul pada** ulasan (Laurenz & Sediyoni, 2021). **Pada** sentimen positif kata **yang sering** muncul pada dilihat pada gambar 11.

Gambar 11 Sentimen Positif

**Pada** Gambar 11 didapatkan beberapa kata **yang sering muncul yaitu** “shopee”, “aplikasi”, “gratis”, “ongkos”, “barang”, “kirim”, “suka”, “bagus”, belajar”, dan “beli”. Kata tersebut merupakan **kata yang sering digunakan** dalam memberikan ulasan **pengguna** terkait **aplikasi** shopee di **google play store. Hal ini** mengartikan bahwa pengguna **merasa puas terhadap pelayanan** aplikasi shopee dikarenakan **memberikan** kemudahan, gratis ongkir dan **membantu pengguna** dalam belanja barang kebutuhan.

Selanjutnya, **kata yang sering muncul pada** kategori **negative dapat dilihat pada Gambar** 12





Gambar 12 Sentimen Negatif

Pada Gambar 12 terdapat beberapa kata yang sering muncul yaitu “shopee”, “barang”, “kirim”, “aplikasi”, “kecewa”, “tolong”, “lambat”, “pesan”, “kurir”, “buruk”, “rugi” dan “paket”. Hal ini menunjukan bahwa kata “aplikasi” pada ulasa negatif memberikan informasi tentang keluhan pengguna terhadap aplikasi Shopee yang berat karena aplikasi ini memerlukan memori yang besar, aplikasi lambat saat mengaksesnya meskipun internet lancar dirasa karena masih berhubungan dengan memori tadi, serta aplikasi bermasalah dan sering logout dirasa karena terdapat masalah dari sistem aplikasi Shopee itu sendiri. **Lalu, kata ulasan yang sering muncul pada** kategori **netral dapat dilihat pada Gambar** 13

Gambar 13 Sentimen Netral

Pada Gambar 13 terdapat beberapa kata yang sering muncul yaitu “shopee”, “barang”, “bayar, “pakai”, “aplikasi”, “belanja”, “pesan” ,“kurir”, “mudah”, “kirim”, “banget” dan “cepat” menunjukan **bahwa pengguna sering bertanya atau** memberikan masukan dan saran terkait beberapa kata tersebut. Sedangkan kata “mudah” merepresentasikan bahwa pengguna merasa mudah dalam mengakses aplikasi shopee.

Gambar 14 Grafik Sentimen Netral





E. Pembobotan TF-IDF

**Setelah melalui proses preprocessing, data kemudian ditransformasikan menjadi** representasi vector dengan **menggunakan** metode TF-IDF, **representasi** vector yang **dihasilkan dari transformasi tersebut akan diproses oleh model. Nilai TF-IDF dari sebuah kata merupakan kombinasi dari nilai tf dan nilai idf dalam perhitungan bobot metode** TF-IDF ini **menggabungkan dua** konsep yaitu **frekuensi kemunculan sebuah kata di dalam sebuah** dokumen **dan** inverse **frekuensi dokumen yang mengandung kata tersebut.** Berikut merupakan script **dari metode TF-IDF pada** Gambar 15

Gambar 15 TF-IDF

F. **Algoritma Random Forest Pada tahapan klasifikasi Random Forest ini dilakukan pembagian data, yaitu data** training (latih) **dan data** testing (test). **Perbandingan** antara data **testing dan training** adalah 80:20. Tahap selanjutnya melakukan klasifikasi sentiment menggunakan Random Forest. **Pada tahapan pengujian menggunakan confusion matrix untuk mengetahui informasi tentang nilai-nilai yang diprediksi dan hasil yang sebenarnya, biasa digunakan untuk perhitungan accuracy, recall, precision, dan f1-score[** 28]. Confusion matrix **menyatakan jumlah data uji yang benar diklasifikasikan dan jumlah data** uji salah **diklasifikasikan.** Berikut hasil pengujian **menggunakan confusion matrix dapat dilihat pada** Gambar 16

Gambar 16 Algoritma Random Forest

Dari hasil penelitian, disimpulkan bahwa model ini dapat melakukan analisis sentimen pada dataset model ini dapat melakukan analisis sentiment pada dataset ulasan pengguna aplikasi Shopee dengan akurasi sebesar 88%. Hal ini menandakan bahwa menggunakan algoritma random forest dapat menghasilkan hasil analisis sentiment yang baik.

G. Evaluasi Model

**Untuk mengevaluasi model yang** telah di buat, digunakan **metode confusion matrix dan cross validation score. Confusion matrix digunakan untuk menunjukkan jumlah hasil klasifikasi yang benar dan salah dari model.** Sedangkan, cross validation score digunakan untuk mengukur seberapa baik model yang dibuat dalam memprediksi data yang

belum dilihat sebelumnya (Hidayati et al., 2021).

H. Visualisasi

Pada penelitian ini, dilakukan visualisasi data menggunakan wordcloud untuk menganalisis frekuensi tentang kata-kata yang paling dominan dalam ulasan pengguna, sehingga memudahkan pemahaman mengenai preferensi dan pengalaman pengguna terkait dengan pengguna e-commerce (Kurniasih & Seseno, 2022). Visualisasi wordcloud ini membantu dalam mengidentifikasi kata-ata kunci yang paling signifikan dan memperoleh wawasan yang berguna dalam analisis sentiment pada dataset tersebut.

VII. SIMPULAN

Dari hasil penelitian ini, dapat disimpulkan bahwa metode klasifikasi Random Forest mampu menghasilkan nilai akurasi yang cukup tinggi, yaitu sebesar 88%. Meskipun begitu, akurasi ini masih bisa ditingkatkan lebih lanjut dengan menggunakan data yang **lebih baik atau lebih bersih.** Peningkatan akurasi maksimum juga dapat dicapai dengan menambahkan lebih banyak data ke dalam set pelatihan atau dengan mencoba metode klasifikasi lain. **Data yang digunakan dalam penelitian ini** terdiri dari 5.000 komentar yang dikumpulkan dari aplikasi Shopee di Play Store. **Adapun saran mengenai ide untuk memperluas penelitian yang serupa adalah penentuan data training dapat mempengaruhi hasil klasifikasi, maka dari itu untuk penelitian diharapkan menambahkan atribut atau data training yang lebih lengkap karena penentuan tingkat akurat dapat dibentuk oleh data training. Karena data training dapat mempengaruhi hasil klasifikasi. Selain itu, pada saat pelabelan data juga ketika menggunakan vader sentiment juga mempunyai pengaruh ketika kata dalam bahasa Indonesia** diberi palabelan **dalam menyatakan sebuah sentimen pada kalimat.**