

# Hasil dan Pembahasan Analisis Sentimen Review Snack

## 1. Tahapan Analisis

Dalam penelitian ini, analisis sentimen dilakukan dalam dua skema yang berbeda, yaitu klasifikasi 2 label (positif–negatif) dan klasifikasi 3 label (positif–netral–negatif). Pemisahan ini bertujuan untuk membandingkan performa model dan melihat konsistensi hasil klasifikasi pada tingkat granularitas yang berbeda.

### a. Penentuan Label

Proses pelabelan awal pada dataset tidak dilakukan secara manual, melainkan dengan memanfaatkan **pretrained transformer model** dari HuggingFace untuk label dari text review, yaitu cardiffnlp/twitter-roberta-base-sentiment. Model ini sudah dilatih khusus pada data Twitter dengan output berupa probabilitas untuk tiga kelas: Positive, Neutral, Negative.

- Dari output probabilitas, dipilih label dengan skor tertinggi sebagai kategori sentimen untuk setiap review.
- Pada eksperimen **2 label**, kategori Neutral digabungkan sesuai skema sehingga hanya tersisa Positive dan Negative.
- Sedangkan pada eksperimen **3 label**, seluruh kategori asli dari model tetap dipertahankan (Positive, Neutral, Negative).

Kemudian untuk label dari rating yang diberikan menggunakan aturan sebagai berikut.

- Data review memiliki skor numerik dari 1 hingga 5.
- Pada eksperimen 2 label, skor 1-2 dikategorikan sebagai Negative dan skor 3-5 dikategorikan sebagai Positive.
- Pada eksperimen 3 label, Skor 1–2 dikategorikan sebagai Negative. kemudian Skor 3 dikategorikan sebagai Neutral dan Skor 4–5 dikategorikan sebagai Positive.

### b. Preprocessing Data

Tahapan preprocessing dilakukan untuk membersihkan teks review sehingga dapat digunakan sebagai input model. Proses yang dilakukan meliputi:

- Normalisasi teks: menghapus tanda baca, karakter khusus, dan emoji.
- Case folding: mengubah seluruh huruf menjadi huruf kecil.
- Tokenisasi: memecah kalimat menjadi kata.
- Stopword removal: menghapus kata-kata umum yang tidak memiliki nilai sentimen.
- Stemming: mengubah kata ke bentuk dasar.

Tahapan preprocessing ini dilakukan pada kedua analisis (2 label dan 3 label) untuk menjaga konsistensi data.

### c. Word Embedding / Representasi Teks

Data yang sudah dibersihkan direpresentasikan dalam bentuk numerik menggunakan TF-IDF (Term Frequency–Inverse Document Frequency), agar kata-kata dengan bobot informasi

lebih tinggi mendapat prioritas. Representasi ini dipakai baik untuk eksperimen **2 label** maupun **3 label**.

#### d. Model Klasifikasi

Beberapa algoritma machine learning yang digunakan antara lain:

- Naive Bayes : baseline model yang sederhana dan cepat.
- Support Vector Machine (SVM) : efektif untuk data teks berdimensi tinggi.
- Voting Classifier (Naive Bayes + SVM) : mengombinasikan dua model untuk meningkatkan akurasi.

Semua model diuji pada dua eksperimen (2 label dan 3 label).

#### e. Evaluasi Model

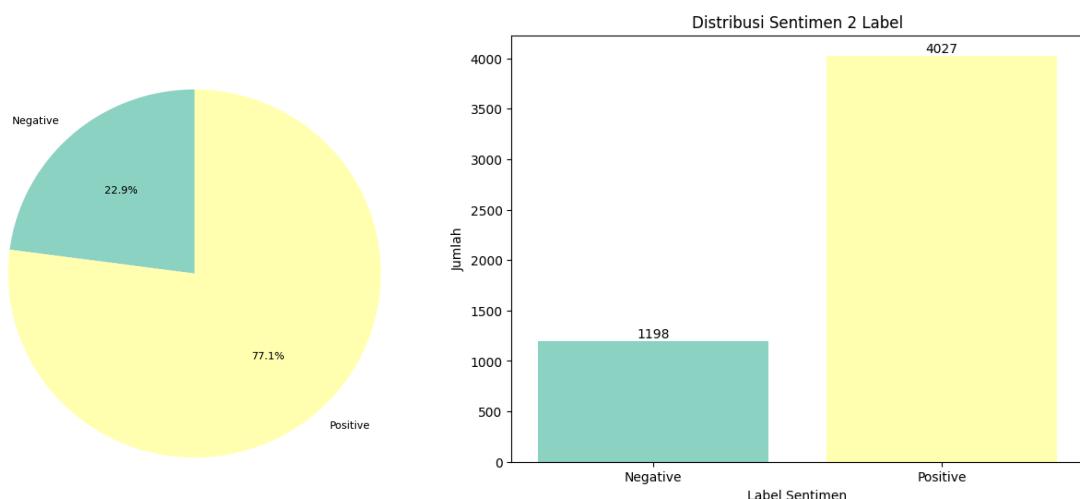
Evaluasi dilakukan dengan metrik berikut:

- Accuracy → mengukur ketepatan prediksi.
- Precision, Recall, F1-Score → menilai keseimbangan performa tiap kelas.
- Confusion Matrix → memperlihatkan distribusi prediksi benar dan salah.

Evaluasi dilakukan pada data training dan testing, untuk membandingkan performa model pada 2 label dan 3 label.

## 2. Hasil Analisis Sentimen (2 Label: Positive & Negative)

Setelah dilakukan labelling, berikut distribusi data label sentimen dengan 2 label yakni positive dan negative.



Berdasarkan hasil visualisasi distribusi sentimen dengan 2 label, terlihat bahwa mayoritas ulasan cenderung bernuansa positif. Hal ini ditunjukkan dengan jumlah sentimen positif sebanyak 4.027 ulasan (77,1%), sedangkan sentimen negatif berjumlah 1.198 ulasan (22,9%). Temuan ini mengindikasikan bahwa persepsi pengguna terhadap produk secara umum lebih dominan positif, meskipun masih terdapat sebagian kecil ulasan yang bersifat negatif.

Hasil pengujian model terbaik menggunakan model Ensemble Voting Classifier ditunjukkan melalui confusion matrix serta perhitungan metrik evaluasi utama, yaitu Accuracy, Precision, Recall, dan F1-Score sebagai berikut.

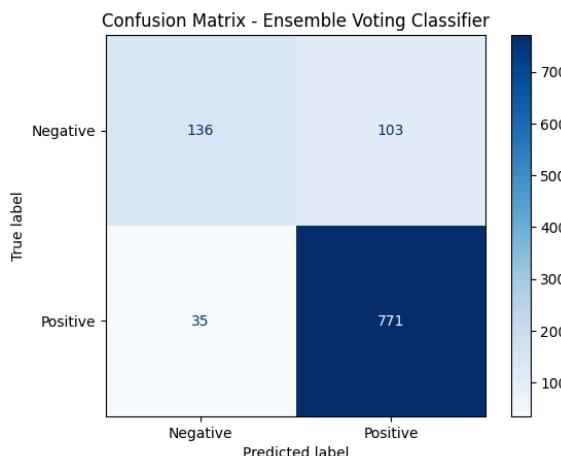
```

=====
ENSEMBLE VOTING CLASSIFIER RESULTS
=====

📊 ACCURACY SUMMARY:
Training Accuracy: 0.8871 (88.71%)
Testing Accuracy: 0.8679 (86.79%)

📈 DETAILED TESTING PERFORMANCE:
=====
precision    recall   f1-score   support
=====
Negative      0.80      0.57      0.66      239
Positive       0.88      0.96      0.92      806
accuracy          0.84      0.76      0.87      1045
macro avg        0.84      0.76      0.79      1045
weighted avg     0.86      0.87      0.86      1045

```



Analisis sentimen dengan menggunakan Ensemble Voting Classifier (kombinasi Naive Bayes dan SVM dengan voting soft) menghasilkan performa yang cukup baik pada klasifikasi dua kategori sentimen, yaitu Positive dan Negative. Akurasi pada data training sebesar 88.71%, sedangkan akurasi pada data testing sebesar 86.79%. Hal ini menunjukkan bahwa model mampu melakukan generalisasi dengan cukup baik, karena selisih antara akurasi training dan testing tidak terlalu besar sehingga mengindikasikan tidak terjadi overfitting yang signifikan.

#### - Kinerja pada Label Negative

Precision: 0.80 → Dari seluruh prediksi negatif, 80% benar-benar negatif.

Recall: 0.57 → Dari seluruh data negatif, hanya 57% yang berhasil diprediksi benar.

F1-score: 0.66 → Kinerja model pada kelas negatif masih cukup baik, tetapi terlihat bahwa recall rendah menunjukkan adanya banyak data negatif yang salah diprediksi menjadi positif.

#### - Kinerja pada Label Positive

Precision: 0.88 → Dari seluruh prediksi positif, 88% benar-benar positif.

Recall: 0.96 → Dari seluruh data positif, 96% berhasil diprediksi benar.

F1-score: 0.92 → Kinerja pada kelas positif sangat baik, baik dari sisi ketepatan maupun kelengkapan.

Macro Average F1-score: 0.79 menunjukkan performa rata-rata antar kelas masih seimbang walaupun kelas negatif lebih sulit diprediksi. Weighted Average F1-score: 0.86 menunjukkan bahwa secara keseluruhan model lebih condong pada performa kelas positif karena distribusi data positif lebih dominan.

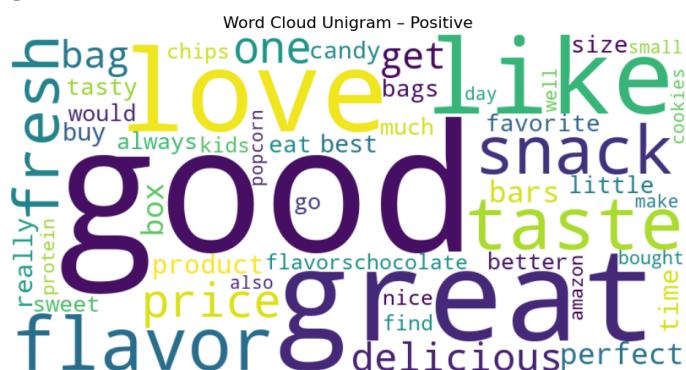
## - Confusion Matrix

Dari total 239 data Negative, sebanyak 136 berhasil diprediksi dengan benar sedangkan 103 salah diprediksi sebagai Positive. Dari total 806 data Positive, sebanyak 771 berhasil diprediksi dengan benar, sementara 35 salah diprediksi sebagai Negative. Pola ini mengonfirmasi bahwa model lebih unggul dalam mengenali data positif dibandingkan data negatif.

Secara Keseluruhan, model Ensemble Voting Classifier terbukti cukup efektif dalam mengklasifikasikan sentimen 2 label dengan akurasi testing mencapai 86.79%. Namun, terdapat tantangan dalam meningkatkan performa pada kelas negatif, terutama dari sisi recall, karena masih ada proporsi yang cukup besar dari data negatif yang salah diprediksi sebagai positif. Hal ini kemungkinan dipengaruhi oleh distribusi data yang tidak seimbang (jumlah positif jauh lebih besar dibanding negatif)

## Visualisasi Word Cloud untuk sentiment 2 label

### - Positive - Unigram



Kata dominan: good, great, love, fresh, flavor, snack, price.

Hal ini menunjukkan bahwa ulasan positif banyak menekankan pada kualitas rasa, harga yang sesuai, serta kepuasan konsumen terhadap produk snack.

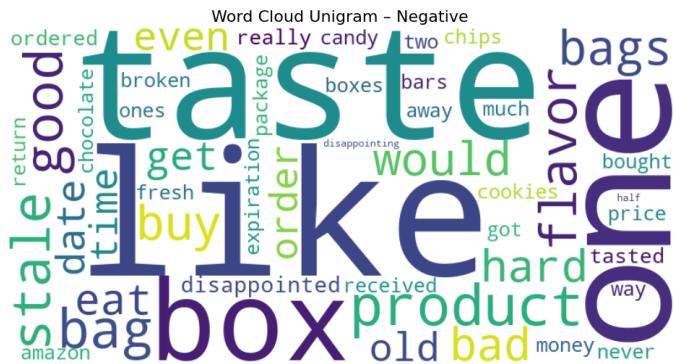
### - Positive - Bigram



Frasa dominan: *great price, really good, highly recommend, great flavor, good value.*

Dari bigram, terlihat lebih jelas bahwa aspek harga murah, rasa enak, dan rekomendasi pelanggan adalah faktor utama yang mendorong ulasan positif.

### - Negative - Unigram



Kata dominan: taste, stale, box, bags, old, hard, broken, refund.

Konsumen yang memberi ulasan negatif banyak menyoroti masalah kualitas produk (basi, keras, rusak) dan pengemasan.

### - Negative - Bigram



Frasa dominan: waste money, expiration date, never buy, stale taste, bad batch.

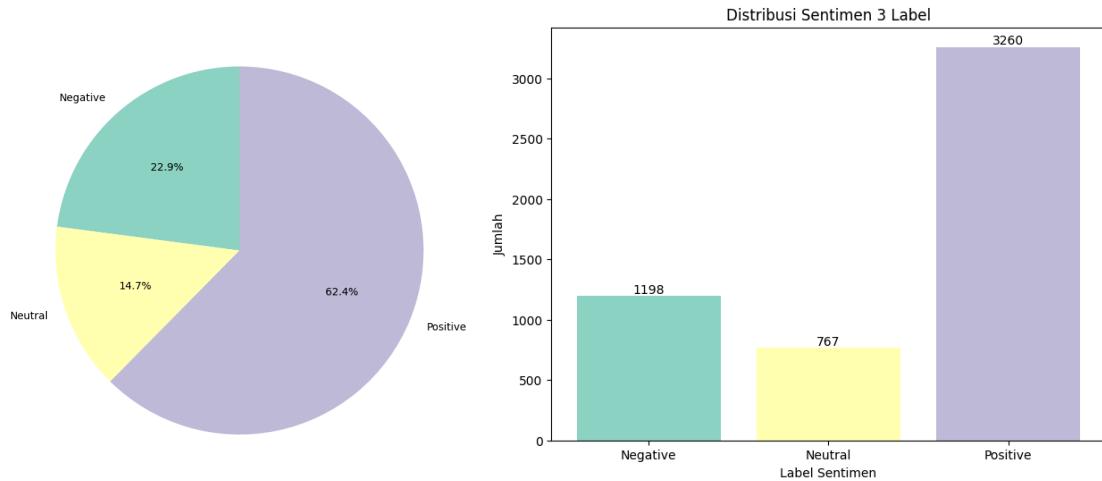
Bigram memperjelas keluhan konsumen, terutama terkait produk melewati masa kadaluwarsa, rasa basi, serta kerugian finansial (uang terbuang).

### Kesimpulan:

interpretasi bigram lebih informatif dalam memahami alasan konsumen memberikan sentimen positif maupun negatif. Karena mampu menampilkan frasa bermakna yang memberi konteks lebih jelas, seperti *great price* (nilai positif) atau *waste money* (nilai negatif).

### 3. Hasil Analisis Sentimen (3 Label: Positive, Neutral, Negative)

Setelah dilakukan labelling, berikut distribusi data label sentimen dengan 2 label yakni positive, neutral, dan negative.



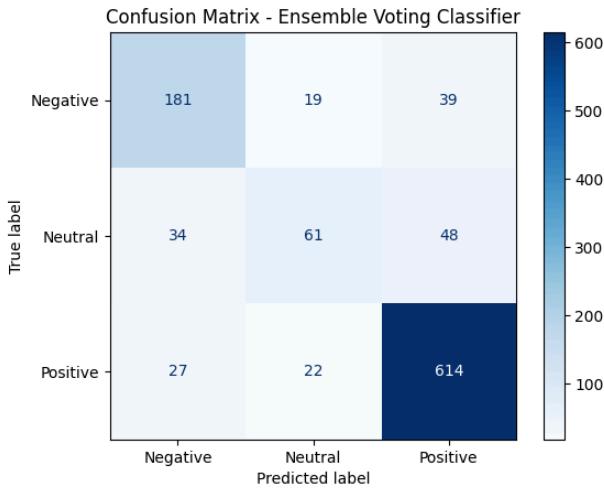
Berdasarkan hasil distribusi sentimen dengan 3 label, terlihat bahwa sentimen positif tetap mendominasi dengan jumlah 3.260 ulasan (62,4%). Sementara itu, sentimen negatif berada pada angka 1.198 ulasan (22,9%), dan sentimen netral mencatat 767 ulasan (14,7%). Temuan ini menunjukkan bahwa mayoritas pengguna masih memberikan tanggapan yang positif terhadap produk, namun keberadaan ulasan negatif dan netral tidak bisa diabaikan. Sentimen negatif memberikan sinyal adanya aspek produk yang kurang memuaskan, sedangkan sentimen netral dapat mencerminkan pengalaman pengguna yang biasa saja atau ambigu.

Hasil pengujian model terbaik menggunakan model Ensemble Voting Classifier ditunjukkan melalui confusion matrix serta perhitungan metrik evaluasi utama, yaitu Accuracy, Precision, Recall, dan F1-Score sebagai berikut.

```
=====
ENSEMBLE VOTING CLASSIFIER RESULTS
=====

📊 ACCURACY SUMMARY:
Training Accuracy: 0.8672 (86.72%)
Testing Accuracy: 0.8191 (81.91%)

📈 DETAILED TESTING PERFORMANCE:
=====
precision    recall    f1-score   support
Negative     0.75      0.76      0.75      239
Neutral       0.60      0.43      0.50      143
Positive      0.88      0.93      0.90      663
accuracy      0.74      0.70      0.72      1045
macro avg     0.74      0.70      0.72      1045
weighted avg  0.81      0.82      0.81      1045
```



Analisis sentimen dengan Ensemble Voting Classifier pada klasifikasi tiga kategori sentimen, yaitu Negative, Neutral, dan Positive, menghasilkan performa yang cukup baik meskipun tantangan lebih besar dibanding klasifikasi 2 label. Akurasi pada data training mencapai 86.72%, sedangkan pada data testing sebesar 81.91%. Hal ini menunjukkan adanya sedikit penurunan akurasi dibanding klasifikasi 2 label, yang wajar karena klasifikasi dengan 3 label memiliki tingkat kompleksitas lebih tinggi.

#### - Kinerja pada Label Negative

Precision: 0.75 → dari seluruh prediksi negatif, 75% benar-benar negatif.

Recall: 0.76 → dari seluruh data negatif, 76% berhasil diprediksi benar.

F1-score: 0.75 → performa cukup seimbang antara ketepatan dan kelengkapan.

#### - Kinerja pada Label Neutral

Precision: 0.60 → dari seluruh prediksi netral, 60% benar-benar netral.

Recall: 0.43 → dari seluruh data netral, hanya 43% berhasil diprediksi benar.

F1-score: 0.50 → ini merupakan performa terendah di antara tiga kelas, menunjukkan bahwa model masih kesulitan mengenali sentimen netral. Hal ini bisa terjadi karena kalimat netral sering memiliki kemiripan dengan sentimen positif maupun negatif sehingga model mengalami kebingungan dalam klasifikasi.

#### - Kinerja pada Label Positive

Precision: 0.88 → dari seluruh prediksi positif, 88% benar-benar positif.

Recall: 0.93 → dari seluruh data positif, 93% berhasil diprediksi benar.

F1-score: 0.90 → menunjukkan performa yang sangat baik pada kelas positif, baik dari segi ketepatan maupun kelengkapan.

Macro Average F1-score: 0.72, menandakan performa rata-rata antar kelas. Nilai ini lebih rendah dibanding klasifikasi 2 label karena adanya kesulitan pada kelas netral. Weighted Average F1-score: 0.81, cukup tinggi karena didominasi oleh performa baik pada kelas positif yang jumlah datanya lebih besar.

#### - Confusion Matrix

Dari total 239 data negatif, sebanyak 181 diprediksi benar, sedangkan sisanya salah diklasifikasikan ke netral (19) dan positif (39). Kemudian dari total 143 data netral, hanya 61 diprediksi benar, sedangkan 34 salah diprediksi negatif dan 48 salah diprediksi positif. Lalu dari 663 data positif,

sebanyak 614 diprediksi benar, hanya sedikit salah ke negatif (27) dan netral (22). Pola ini memperlihatkan bahwa kelas netral menjadi yang paling sulit dibedakan, dengan kesalahan prediksi yang cukup besar ke arah positif maupun negatif.

Secara keseluruhan, model Ensemble Voting Classifier menunjukkan performa yang baik dalam klasifikasi 3 label, dengan akurasi testing 81.91%. Model mampu mengenali sentimen positif dengan sangat baik, sedangkan netral menjadi kelas yang paling menantang dengan nilai recall yang rendah. Hal ini menunjukkan bahwa untuk meningkatkan kinerja pada skenario 3 label, diperlukan pendekatan tambahan seperti penyeimbangan data antar kelas, penggunaan model berbasis transformer (misalnya BERT/RoBERTa), atau feature engineering yang lebih mendalam untuk membedakan teks netral dari positif dan negatif.

## Visualisasi Word Cloud untuk sentiment 3 label

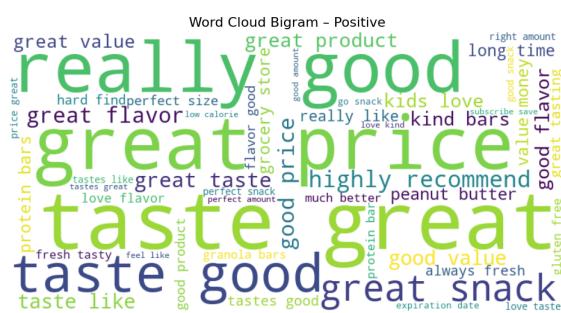
### - Positive - Unigram



Kata dominan: good, great, love, taste, fresh, price, delicious.

Konsumen menekankan kualitas rasa dan harga yang sesuai.

### - Positive - Bigram



Frasa dominan: really good, great price, highly recommend, great flavor.

Memberikan konteks lebih jelas bahwa konsumen puas karena kombinasi rasa enak, harga terjangkau, dan rekomendasi tinggi.

### - Neutral - Unigram



Kata dominan: *good, great, love, like, flavor, snack.*

Ulasan netral masih memunculkan kata positif, namun tanpa penekanan kuat pada rekomendasi atau keluhan.

### - Neutral - Bigram



Frasa dominan: *really good, great snack, good price.*

Bigram menampilkan konteks bahwa ulasan netral biasanya berupa komentar singkat atau deskriptif, tidak terlalu menekankan pada kepuasan ataupun kekecewaan.

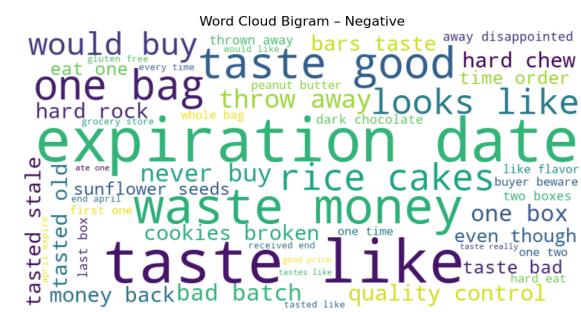
### - Negative - Unigram



Kata dominan: *taste, stale, box, waste, bad, refund, old.*

Keluhan konsumen berfokus pada kualitas buruk (rasa basi, produk lama) serta masalah kemasan.

### - Negative - Bigram



Frasa dominan: waste money, expiration date, never buy, stale taste, bad batch.

Bigram memberi gambaran lebih tegas tentang kerugian finansial, produk kedaluwarsa, dan rasa yang mengecewakan.

## Kesimpulan:

Dengan demikian, analisis bigram lebih kaya makna dan membantu memahami alasan konsumen memberikan label sentimen tertentu karena memberikan gambaran yang lebih jelas dalam bentuk frasa, misalnya *great price* (positif), *really good* (netral), atau *waste money* (negatif).

## **4. Hasil Data yang MisInform**

Hasil analisis juga menunjukkan adanya ketidaksesuaian antara teks ulasan dengan skor numerik yang diberikan oleh konsumen. Beberapa respon memiliki teks yang mengindikasikan sentimen positif, seperti penggunaan kata *good*, *delicious*, atau *great*, namun justru disertai dengan skor yang rendah. Fenomena ini dapat disebut sebagai miss informasi, karena penilaian numerik tidak sejalan dengan isi ulasan. Identifikasi terhadap ulasan-ulasan dengan pola seperti ini sangat penting, karena dapat menimbulkan bias dalam analisis sentimen berbasis skor saja. Oleh karena itu, daftar lengkap respon yang terindikasi mengalami miss informasi telah disajikan pada lampiran/drive hasil analisis sebagai bahan evaluasi lebih lanjut.

### **Catatan:**

Hasil pelabelan sentimen dalam penelitian ini masih memiliki potensi kesalahan meskipun sudah menggunakan model bahasa mutakhir, sehingga interpretasi sebaiknya dipahami sebagai gambaran umum, bukan kebenaran absolut dari setiap ulasan.

Lampiran File :

▣ Analisis Sentiment Snack Riview

<https://drive.google.com/drive/folders/1I9Jn4KFrJTRxsBXoMar-fzG6gANXGWmq?usp=sharing>

