

# **ANALISIS SENTIMEN KONSUMEN TERHADAP PRODUK LOKAL DAN GLOBAL DI TOKOPEDIA: STUDI KASUS ADVAN DAN XIAOMI**

Muhammad Luthfi Aziz Sunarya – 140810230081

Danish Rahadian Mirza Effendi – 140810230058

Athallah Azhar Aulia Hadi – 140810230083

Naufal Fakhri Ilyas – 140810220068

**PRESENTATION**

# LATAR BELAKANG PROJEK

Pasar teknologi Indonesia menunjukkan dominasi produk impor, khususnya merek Xiaomi, yang mengindikasikan tingginya preferensi konsumen terhadap produk asing dan berpotensi menekan daya saing industri nasional, sementara merek lokal seperti Advan masih menghadapi tantangan dalam persepsi kualitas dan loyalitas pengguna. Ulasan pengguna di platform digital menjadi sumber data penting untuk memahami persepsi publik secara objektif, sehingga penelitian ini memanfaatkan pendekatan text mining dan machine learning untuk menganalisis sentimen serta mengidentifikasi aspek krusial yang membedakan tingkat kepuasan pengguna antara merek asing dan lokal. Dengan menerapkan Semi-Supervised Annotation dan klasifikasi sentimen berbasis Supervised Learning, penelitian ini diharapkan mampu memberikan gambaran komprehensif mengenai peta persaingan persepsi konsumen terhadap produk teknologi di Indonesia.

# TUJUAN PROJEK

01



## Compare

Menganalisis perbandingan sentimen publik terhadap brand teknologi asing (Xiaomi) dan lokal (Advan) untuk memahami persepsi konsumen di Indonesia.

02



## Aspect Identification

Mengidentifikasi aspek dominan seperti kualitas, inovasi, dan harga yang memengaruhi preferensi konsumen terhadap produk asing dan lokal.

03



## Aspect Mapping

Memetakan persepsi masyarakat terhadap produk teknologi asing dan lokal sebagai kontribusi ilmiah dalam kajian perilaku konsumen digital.

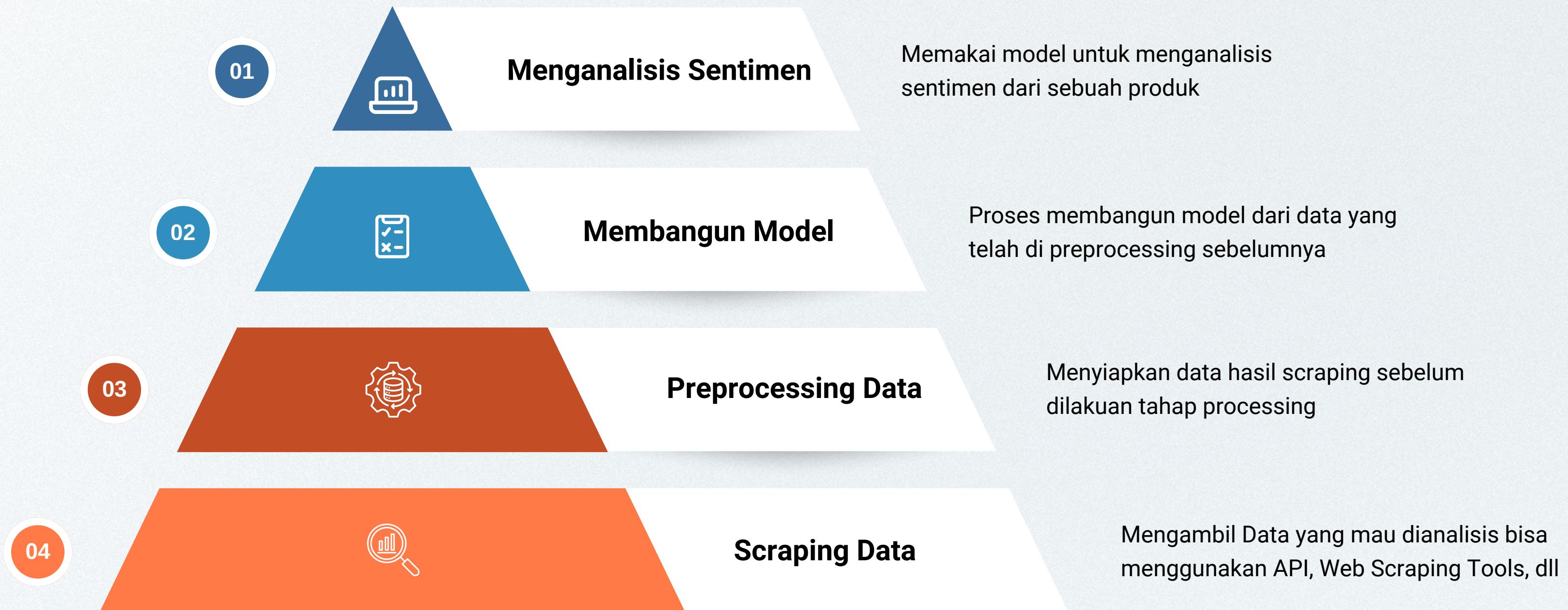
04



## Strategic Insight

Memberikan insight strategis bagi pengembangan industri teknologi lokal guna meningkatkan kualitas dan daya saing terhadap produk impor.

# Tahapan Analisis Sentimen



# METODE SCRAPPING DATA

Kami memakan metode parsing HTML dan selenium untuk membuka web drivernya



**Selenium**

01

Berfungsi untuk memuat dan mengontrol halaman web yang izingin di scraping



**BeautifulSoup**

02

Berfungsi untuk memarsing html dari halaman web yang ingin di scraping



**Pandas**

03

untuk menyimpan hasil scraping ke dalam dataframe untuk diolah nantinya

# Preprocessing Data

## Case Folding

mengubah seluruh teks menjadi huruf kecil agar konsisten dan menghindari perbedaan makna akibat perbedaan kapitalisasi.



Step 01

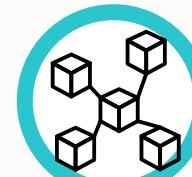
## Cleaning

menghapus URL, tag HTML, tanda baca, angka, dan spasi berlebih, sehingga hanya menyisakan huruf, emoji wajah, dan spasi.

Aa

## Tokenization

memecah teks menjadi daftar kata (token) berdasarkan spasi agar setiap kata bisa diproses atau dianalisis secara terpisah.



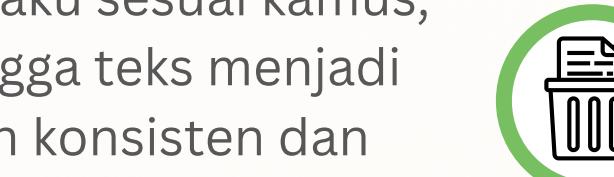
Step 02

## Normalisasi

menyediakan kata tidak baku dalam daftar token dengan padanan katanya yang baku sesuai kamus, sehingga teks menjadi lebih konsisten dan sesuai ejaan standar.



Step 03



Step 04

## Stopwords Removal

menghapus kata-kata umum yang tidak memiliki makna penting seperti "dan", "ya", atau "di", agar model fokus pada kata yang lebih bermakna dalam analisis.



Step 05

## Stemming

setiap kata direduksi ke bentuk dasarnya (misalnya, "meningkatkan" menjadi "tingkat") untuk mengkonsolidasi fitur leksikal



Step 06

## Merging

berfungsi untuk mengubah daftar token pada misalnya ['tidakpuas', 'sekali'] menjadi satu string utuh ("tidakpuas sekali") sehingga bisa digunakan untuk model klasifikasi.

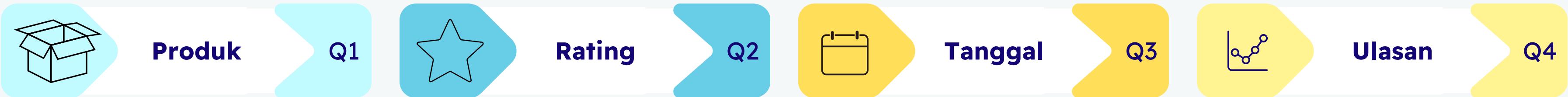


## Negation Handling

menggabungkan kata "tidak" dengan kata sesudahnya (misalnya menjadi "tidakpuas") agar makna negasi tetap utuh dan tidak terpisah saat pemrosesan teks.

# Struktur Dataset

Dataset tersebut berisi data hasil scraping ulasan produk Tokopedia yang terdiri dari nama produk, rating bintang, tanggal ulasan, dan isi teks ulasan pelanggan yang nantinya bisa digunakan untuk analisis sentimen.



Berisi teks panjang yang mendeskripsikan nama atau varian produk dengan tipe data string (object).

Biasanya tidak langsung digunakan untuk klasifikasi sentimen, tapi penting untuk filtering per produk.

Nilai numerik (1–5) yang merepresentasikan tingkat kepuasan pengguna. Bisa berfungsi sebagai label supervised learning (misal dikonversi ke kelas sentimen: negatif, netral, positif) atau sebagai variabel target untuk regression model. dengan tipe datanya adalah int

Explore digital platforms and social media to expand your audience. Strategic partnerships with established brands can increase market exposure, access new customer bases, share resources, and foster innovative marketing strategies for brand growth.

Teks bebas (unstructured data) hasil input pengguna. Ini fitur utama untuk analisis sentimen bisa terdapat emoji, teks, special character dan harus di preprocessing sebelum diolah

# Data Understanding

Data Understanding adalah tahap awal dalam proses Data Science / Machine Learning pipeline yang bertujuan untuk memahami karakteristik, struktur, dan kualitas data yang akan digunakan.

## General

Jumlah record data sebanyak 3331 records  
dan tidak ada data yang kosong

## Distribusi Data

Komentar dengan Rating >3 sebanyak 3205 records, rating 3 sebanyak 45, rating < 3 sebanyak 81 menandakan distribusi positif, negatif dan netral bisa jadi tidak seimbang

## Ulasan

Ulasan masih banyak kata yang tidak baku, emoji, penggunaan special character dan inkonsistensi penggunaan huruf kapital

## Tanggal

Bentuknya masih string atau objek dan bukan tipe data date



# Modeling

Proses ini merupakan bagian dari supervised learning, di mana model dilatih menggunakan data teks ulasan produk (text\_clean) yang sudah diberi label sentimen (positif, netral, negatif). Label tersebut dihasilkan secara otomatis melalui API OpenAI (ChatGPT-4o-latest) berdasarkan isi ulasan, sehingga setiap data memiliki target kelas yang jelas untuk dipelajari model. Dataset kemudian dibagi menjadi data latih dan data uji menggunakan metode train-test split dengan proporsi 80:20 serta stratified sampling agar distribusi label tetap seimbang. Selanjutnya, teks ulasan diubah menjadi representasi numerik menggunakan TF-IDF Vectorizer, yang mengekstraksi fitur penting dari kata dan frasa (unigram dan bigram) sehingga dapat diproses oleh algoritma klasifikasi machine learning.

## Logistic Regression



Model Logistic Regression menghasilkan akurasi sebesar 0.892, precision 0.905, recall 0.892, dan F1-score 0.897. Hasil ini mengindikasikan bahwa Logistic Regression memiliki performa yang cukup kuat dan relatif seimbang antara ketepatan dan kelengkapan prediksi. Model ini mampu mempelajari pola umum dalam data ulasan dengan baik, meskipun kemampuannya dalam menangkap hubungan kata yang kompleks masih lebih terbatas dibandingkan model berbasis margin seperti SVM.

## Naive Bayes



Model Naive Bayes memperoleh akurasi sebesar 0.858, precision 0.735, recall 0.858, dan F1-score 0.792. Performa ini menunjukkan bahwa meskipun Naive Bayes tergolong efisien dan sederhana, nilai precision yang lebih rendah mengindikasikan adanya kecenderungan kesalahan dalam memprediksi label sentimen tertentu. Hal ini sejalan dengan asumsi independensi fitur yang digunakan oleh model, yang kurang optimal untuk data teks dengan konteks yang saling berkaitan.

## Support Vector Machine (SVM)



Model Support Vector Machine (SVM) menunjukkan performa terbaik dengan akurasi 0.909, precision 0.901, recall 0.909, dan F1-score 0.903. Tingginya nilai pada seluruh metrik mengindikasikan bahwa SVM paling efektif dalam mengenali pola sentimen pada data teks. Model ini bekerja optimal pada representasi fitur TF-IDF berdimensi tinggi dan mampu memisahkan kelas sentimen secara lebih jelas dibandingkan model lain, sehingga menjadi model dengan performa paling unggul pada pengujian ini.

## Random Forest



Model Random Forest mencatat akurasi sebesar 0.871, precision 0.849, recall 0.871, dan F1-score 0.852. Hasil ini menunjukkan bahwa Random Forest memiliki performa yang cukup stabil, meskipun masih berada di bawah Logistic Regression dan SVM. Model ini mampu menangkap hubungan non-linear antar fitur, namun pada data teks dengan dimensi fitur yang besar, efektivitasnya cenderung menurun dibandingkan metode berbasis linear dan margin.

# VALIDATION

## Scraping Data Xiaomi

Memiliki struktur yang sama yaitu rating, produk, ulasan, tanggal



## Labeling

Mengirim data xiaomi ke LLM ChatGPT 4 dengan model yang sama untuk labeling sebagai ground truth

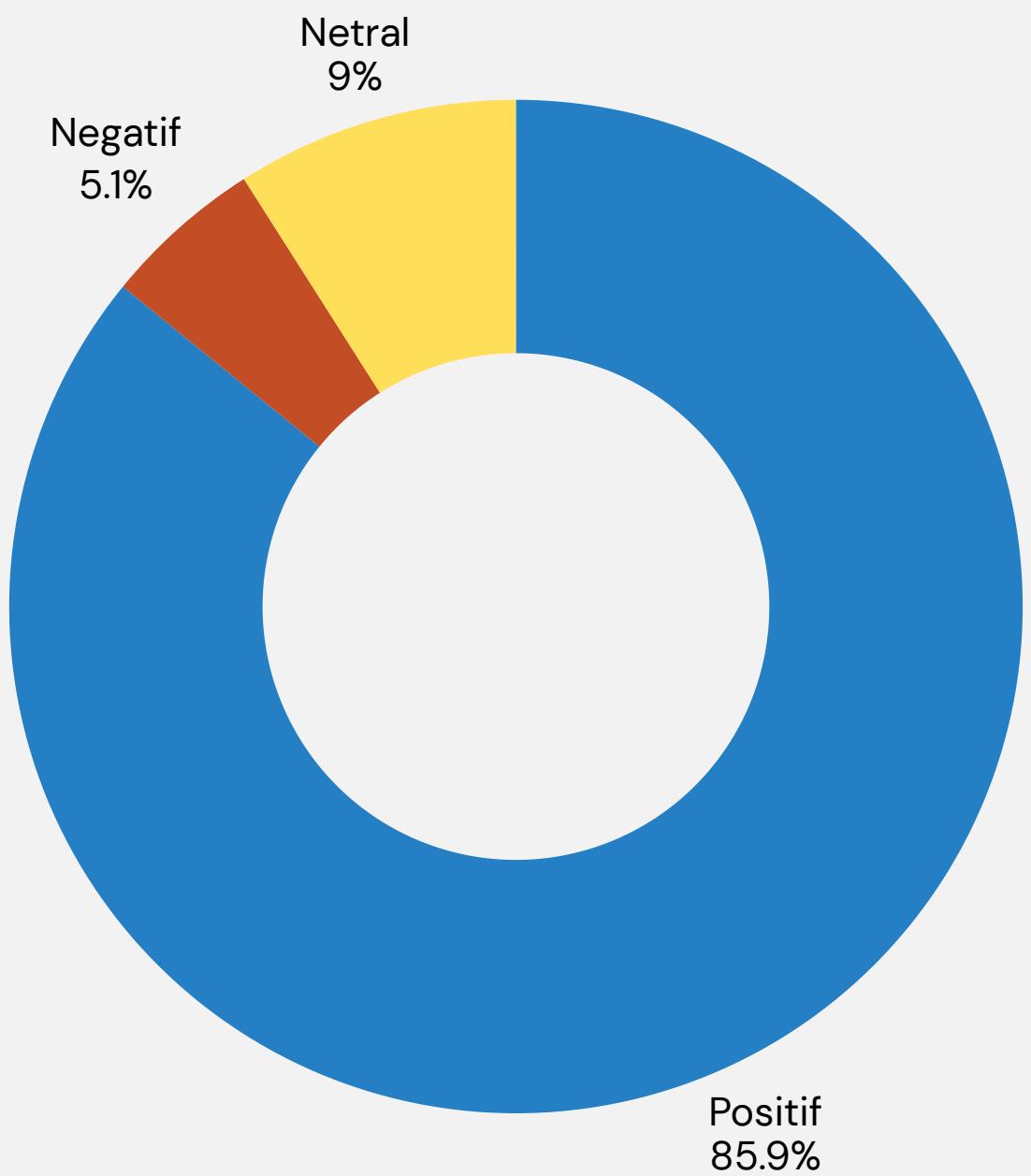


## Compare

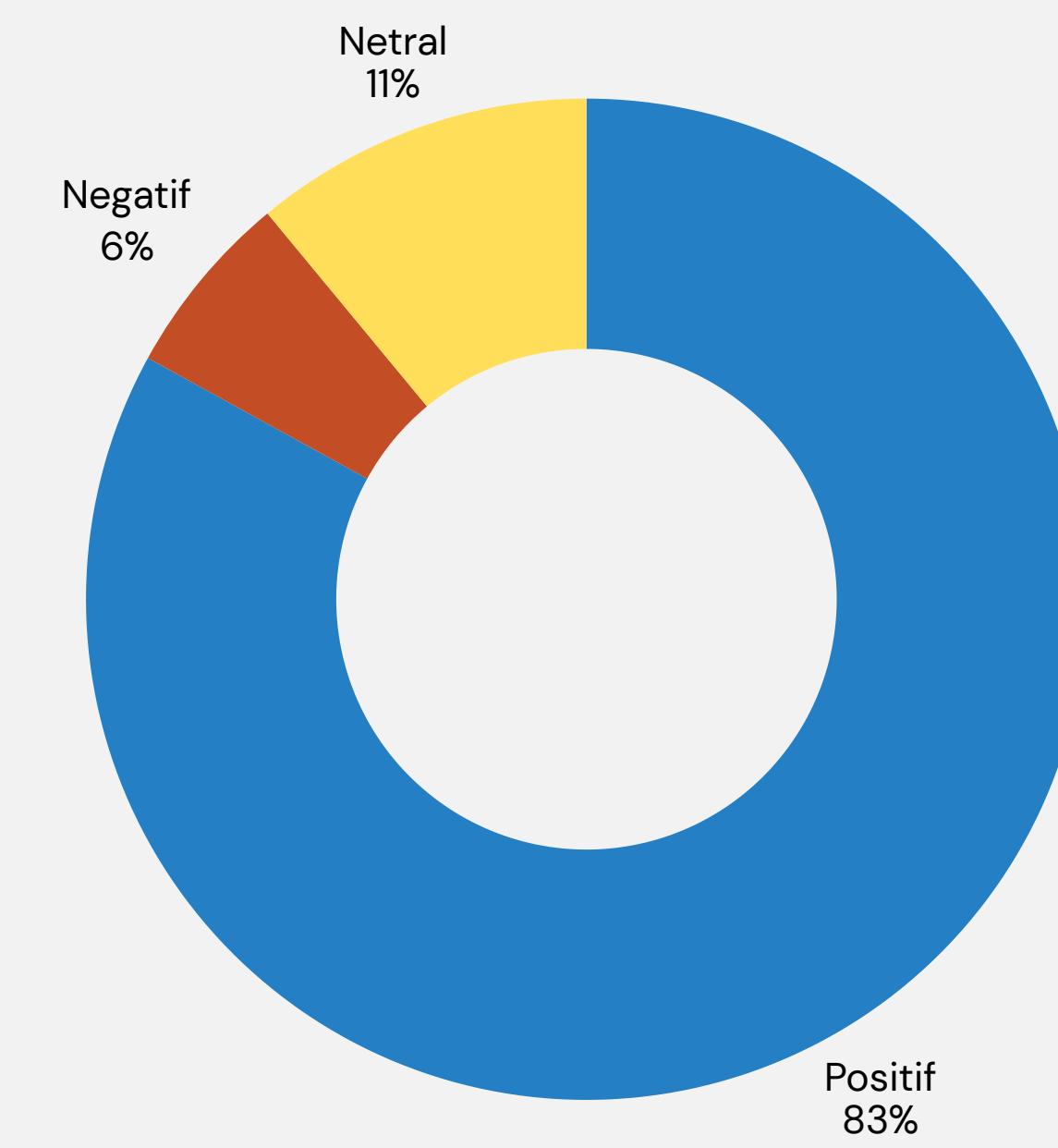
Membandingkan hasil prediksi model dengan hasil prediksi LLM ChatGPT 4 dan didapatkan akurasi 85% menandakan model cukup robust



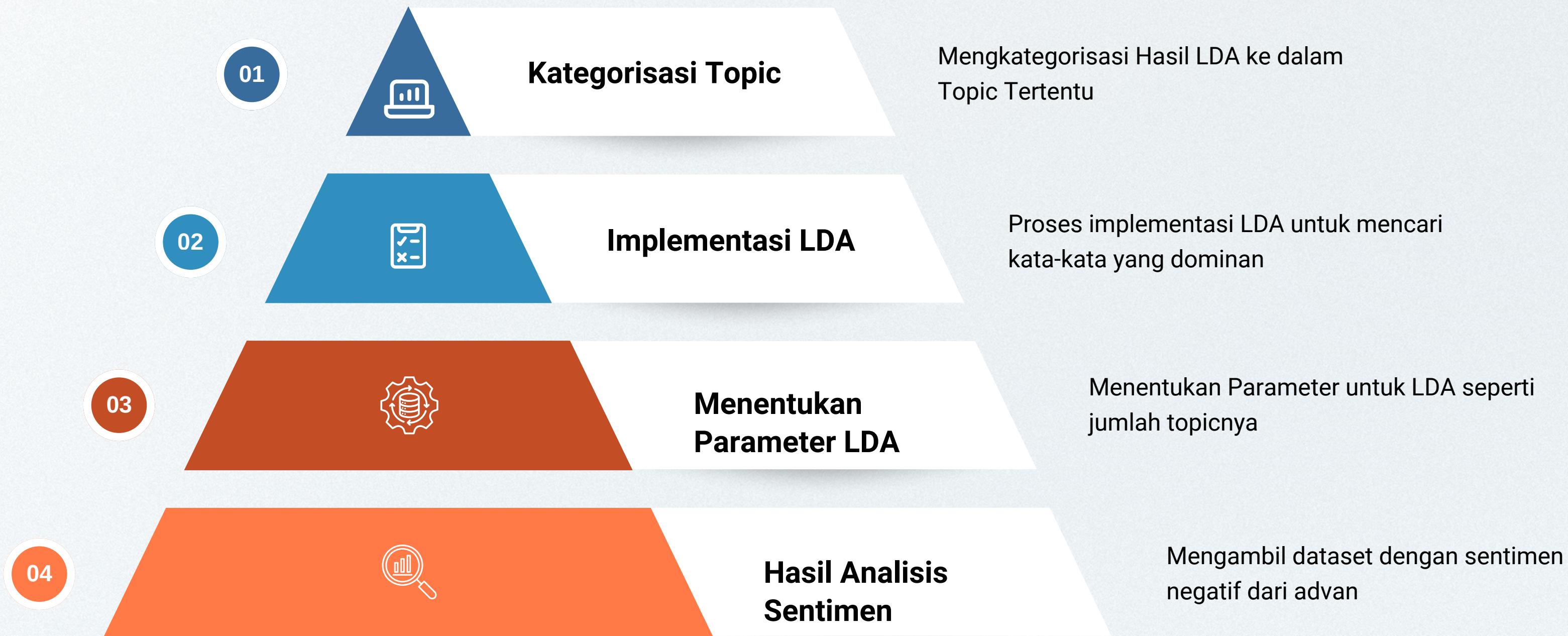
# Sentimen Advan



# Sentimen Xiaomi



# Tahapan Analisis Topic



# KATEGORISASI TOPIC

Berdasarkan hasil hyperparameter tuning, CountVectorizer menunjukkan koherensi semantik yang lebih baik dibandingkan TfIdfVectorizer, dengan Coherence Score tertinggi sebesar 0.425038 pada konfigurasi 5 topik ( $\alpha = 0.1$ ,  $\eta = 0.9$ ). Meskipun TfIdfVectorizer menghasilkan Silhouette Score positif, penelitian ini memprioritaskan koherensi topik demi interpretabilitas, sehingga konfigurasi terbaik CountVectorizer digunakan untuk mengekstraksi lima topik laten dari ulasan negatif berdasarkan sepuluh kata kunci dominan pada tiap topik.

01



## Produk Mati Total (Dead on Arrival)

bisa, sampai,  
tidakada, mati,  
belum, pakai, saya,  
kalau, beli, ada

02



## Kualitas Multimedia & Fisik

tapi, jam, layar,  
suara, pakai, baru,  
ada, padahal,  
banget, kecil

03



## Kerusakan Barang saat Diterima

pakai, tidakbisa,  
tapi, bagus, sih,  
datang, rusak, beli,  
belum, cuma

04



## Layanan Retur & Purna Jual

tidakbisa, tapi,  
cuma, hp, datang,  
retur, sama, laptop,  
layar, saja

05



## Kualitas Kamera & Performa

bagus, kali,  
kualitas, jernih,  
kerja, banget,  
performa, lensa,  
tapi, kecewa

# Kesimpulan

Penelitian ini berhasil membuktikan bahwa produk teknologi lokal (Advan) memiliki daya saing persepsi yang kompetitif di pasar Indonesia, ditunjukkan dengan dominasi sentimen positif sebesar 85,7% yang justru sedikit mengungguli produk asing (Xiaomi) di angka 82,9%. Temuan ini membantah stigma inferioritas produk lokal, di mana preferensi konsumen terhadap Advan sangat didorong oleh kepuasan atas kesesuaian spesifikasi dengan harga (value for money) dan kinerja fungsional. Sebaliknya, persepsi positif terhadap Xiaomi lebih banyak diasosiasikan dengan reputasi merek jangka panjang seperti keawetan dan orisinalitas produk, yang menunjukkan bahwa produk lokal telah diterima dengan baik secara fungsional namun masih perlu membangun kepercayaan terhadap durabilitas merek.

Meskipun demikian, analisis mendalam menggunakan Latent Dirichlet Allocation (LDA) pada sentimen negatif menyingkap bahwa hambatan utama bagi produk lokal untuk mendominasi pasar bukanlah kurangnya inovasi fitur, melainkan inkonsistensi pada kontrol kualitas dan layanan purna jual. Lima aspek krusial yang menjadi sumber keluhan utama pengguna Advan meliputi kegagalan fungsi saat barang diterima (Dead on Arrival), kerusakan fisik, defek pada komponen multimedia, serta kerumitan proses retur garansi. Oleh karena itu, strategi kunci untuk mendukung kemandirian ekonomi digital nasional harus berfokus pada pembenahan manajemen rantai pasok dan pengetatan Quality Control guna menekan tingkat produk cacat, serta reformasi layanan purna jual untuk meningkatkan loyalitas konsumen setara dengan standar global.mendominasi dan menjadi preferensi utama masyarakat indonesia dalam mencari layanan sesuai kebutuhan. Harapannya dengan hasil riset kami aplikasi layanan lokal bisa lebih diminati dan indonesia bisa menjadi negara dengan kemandirian digital yang akan berpengaruh ke indonesia sebagai negara yang memiliki kemandirian ekonomi.

**THANK**  
*You!*

