



NATURAL LANGUAGE PROCESSING

Angella Ananta Batubara



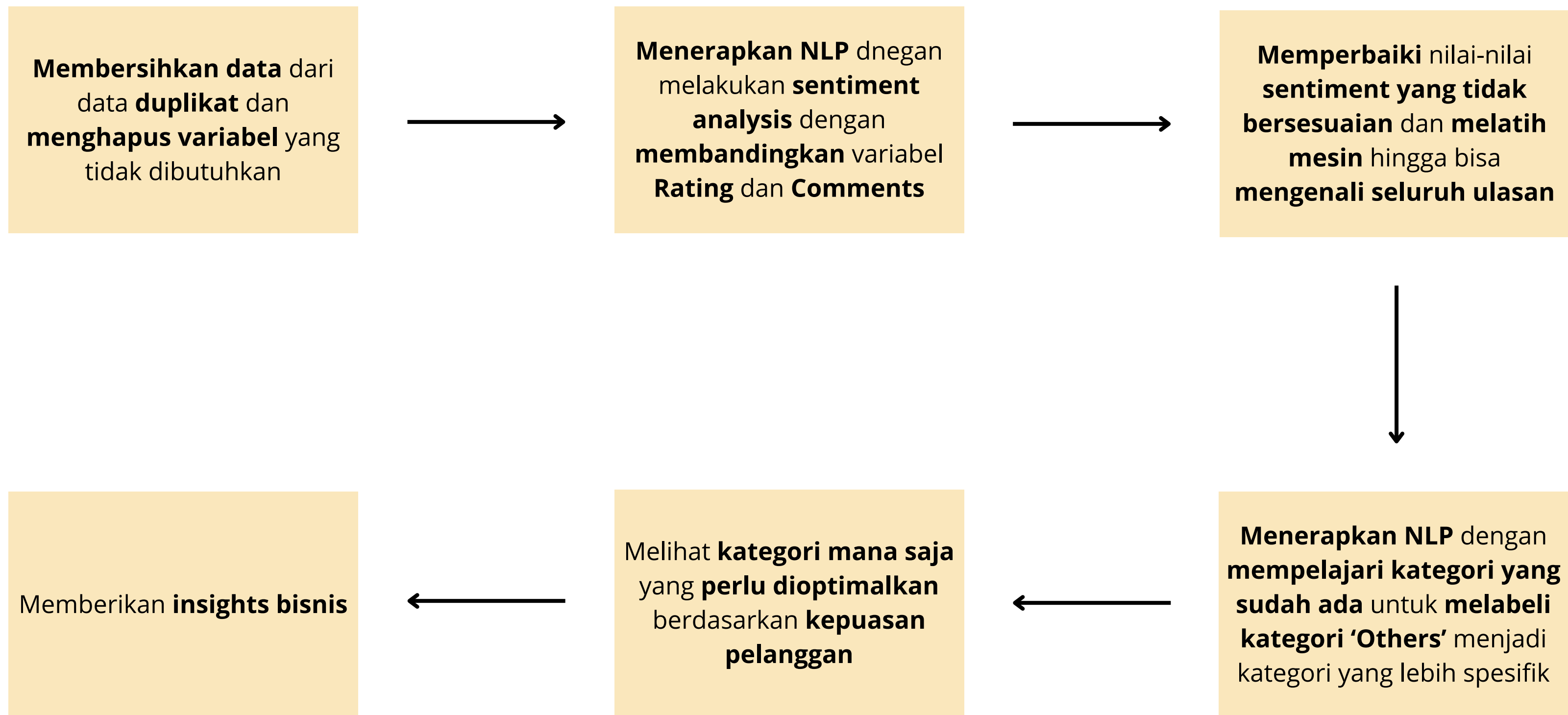
BACKGROUND

Melakukan Natural Language Processing (NLP) dari data penilaian produk Redmi (dataset Redmi6.csv)

GOALS

1. Melihat kepuasan pelanggan yang membeli produk berdasarkan rating dan komentar yang diberikan
2. Menganalisis masalah pada kategori tertentu dari hasil analisis sentimen yang ada

WORK FLOW



CLEANING DATA

- Dari 280 baris data, terdapat **35 baris data duplikat** yang **perlu dihilangkan**. Sehingga **data akhir** yang dapat dianalisis sebanyak **245 baris data**.
- Analisis **hanya memerlukan** variabel **Rating**, **Comments**, dan **Category**. Maka dari itu, variabel lain akan dihapus.
- Pada variabel **Rating**, **hanya diambil nilai depannya saja**. (Contoh: **4.0 out of 5 stars** diubah menjadi **4.0**)
- Pada variabel **Comments**, **data distandarisasi** dengan **menghapus URL, angka dan simbol**, serta **spasi berlebih** (jika ada).

SENTIMENT ANALYSIS

Melakukan labelling sentimen untuk variabel Comments.

- 1. Satisfied — Positive
- 2. Mixed > Neutral
- 3. Neutral > Negative
- 4. Angry > Negative
- 5. Unsatisfied > Negative

Melakukan labelling sentimen untuk variabel **Rating**.

- Positive untuk nilai Rating > 3
- Neutral untuk nilai Rating $= 3$
- Negative untuk nilai Rating < 3

SENTIMENT ANALYSIS

Mencocokkan hasil sentimen Rating dengan Comments menggunakan metode BART.

Terdapat **satu baris** yang **hanya memasukkan Rating** dan **tidak menambahkan Comment**.
Data tersebut **diasumsikan** memberi penilaian yang **positive**.

```
Perbandingan hasil sentiment:  
Sentiment_Match  
True          171  
False         74
```

Hasil pencocokan pertama menunjukkan **171 data sudah cocok** hasil sentimennya, sedangkan **74 data masih menghasilkan hasil analisis yang berbeda**.

SENTIMENT ANALYSIS

Hasil dari analisis **BART** digunakan untuk **melatih model Regresi Logistik**. Hal ini dilakukan agar data yang masih salah dapat terprediksi lebih baik.

```
Perbandingan hasil sentiment:  
Sentiment_Match  
True      220  
False     25
```

Akurasi 74%

Hasil pencocokan kedua menunjukkan **220 data sudah cocok** hasil sentimennya, sedangkan **25 data masih menghasilkan hasil analisis yang berbeda**.

SENTIMENT ANALYSIS

Hasil dari analisis yang sudah benar kembali dijadikan bahan untuk machine learning belajar. **Metode BART kembali diterapkan dengan bantuan manual labeled**, di mana nilai sentimen langsung diberikan ke ulasan tanpa bantuan sistem.

Sentiment_Match	# count
True	229
False	16

Hasil pencocokan kedua menunjukkan **229 data sudah cocok** hasil sentimennya, sedangkan **16 data masih menghasilkan hasil analisis yang berbeda**.

SENTIMENT ANALYSIS

Terakhir, **metode Regresi Logistik** kembali diterapkan. Di sini, **mesin dibantu untuk mengenali kata-kata negatif dan transisi opini** (dengan adanya kata 'tapi' pada ulasan). Hal ini dilakukan dengan harapan machine learning dapat lebih mengenali pola kata yang ada.

Sentiment_Match	
True	243
False	2

Akurasi 83%

Hasil pencocokan kedua menunjukkan **243 data sudah cocok** hasil sentimennya, sedangkan **2 data masih menghasilkan hasil analisis yang berbeda**.

2 data ini akan dilabeli secara manual.

ANALISIS KATEGORI

Category	# count
Others	161
Display	29
Camera	25
Battery	24
Delivery	6

Dari sebaran data yang ada, **kategori 'Others'** memiliki **jumlah terbanyak**. Hal ini akan menyulitkan proses evaluasi dari rating yang telah diberikan oleh pelanggan.

ANALISIS KATEGORI

Maka dari itu, **dilakukan NLP** untuk **mengelompokkan komentar** dari kategori '**Others**' ke dalam kategori yang lain, baik itu **Display**, **Battery**, **Camera**, ataupun **Delivery**.

Mengingat satu komentar bisa memberikan ulasan ke beberapa fitur, maka komentar yang ada perlu ditinjau lebih dalam menggunakan **BART**.

Saat komentar tidak ditinjau lebih dalam

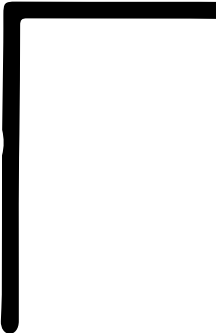
Category	
Display	162
Camera	30
Delivery	28
Battery	25

VS

Saat komentar ditinjau lebih dalam

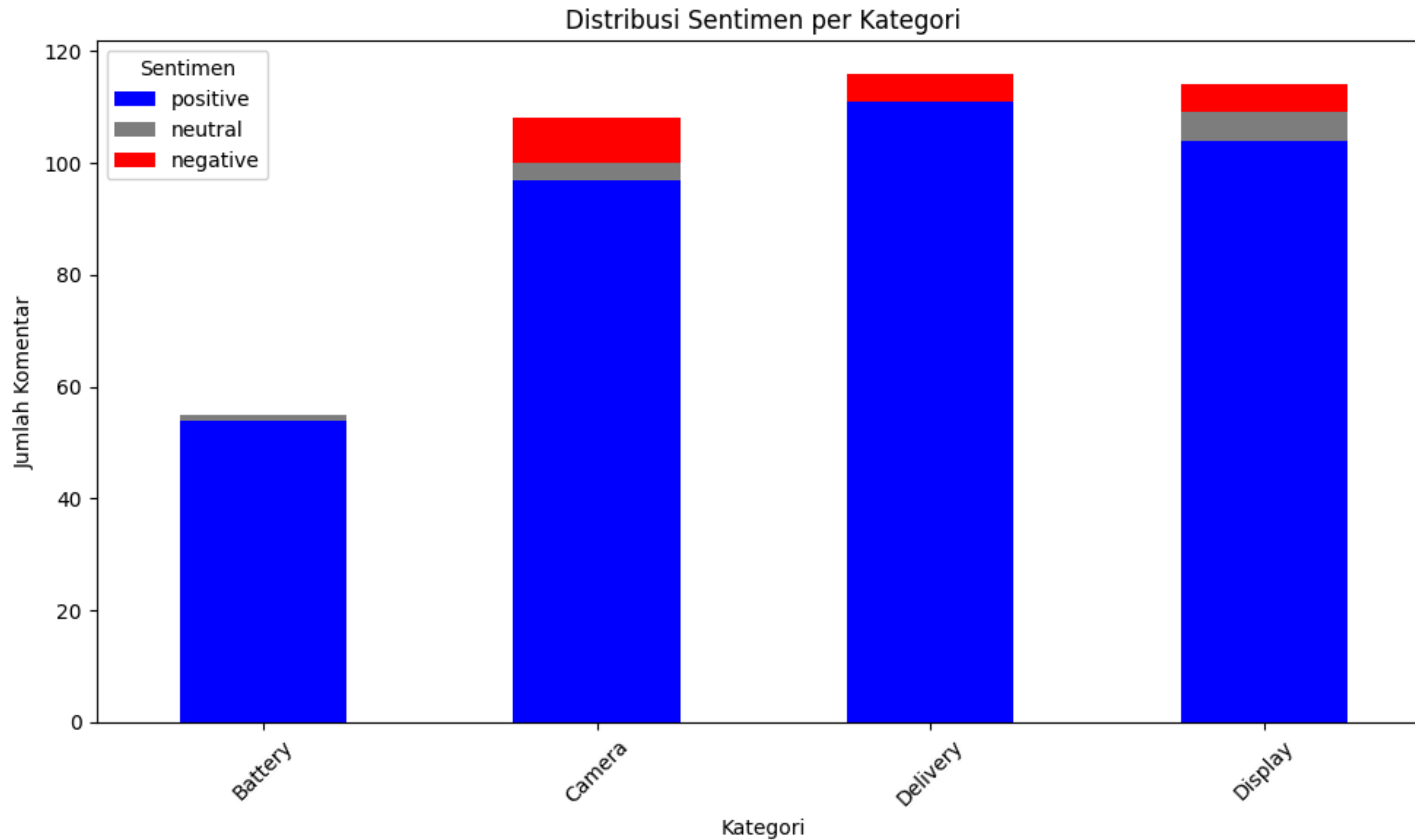
Category	
Delivery	116
Display	114
Camera	108
Battery	55

A hand-drawn black arrow pointing to the right. The arrow has a long, slightly curved shaft and a simple V-shaped head. A large, loopy flourish is drawn above the shaft, starting from the middle, looping back to the left, and then crossing over itself before continuing to the right.



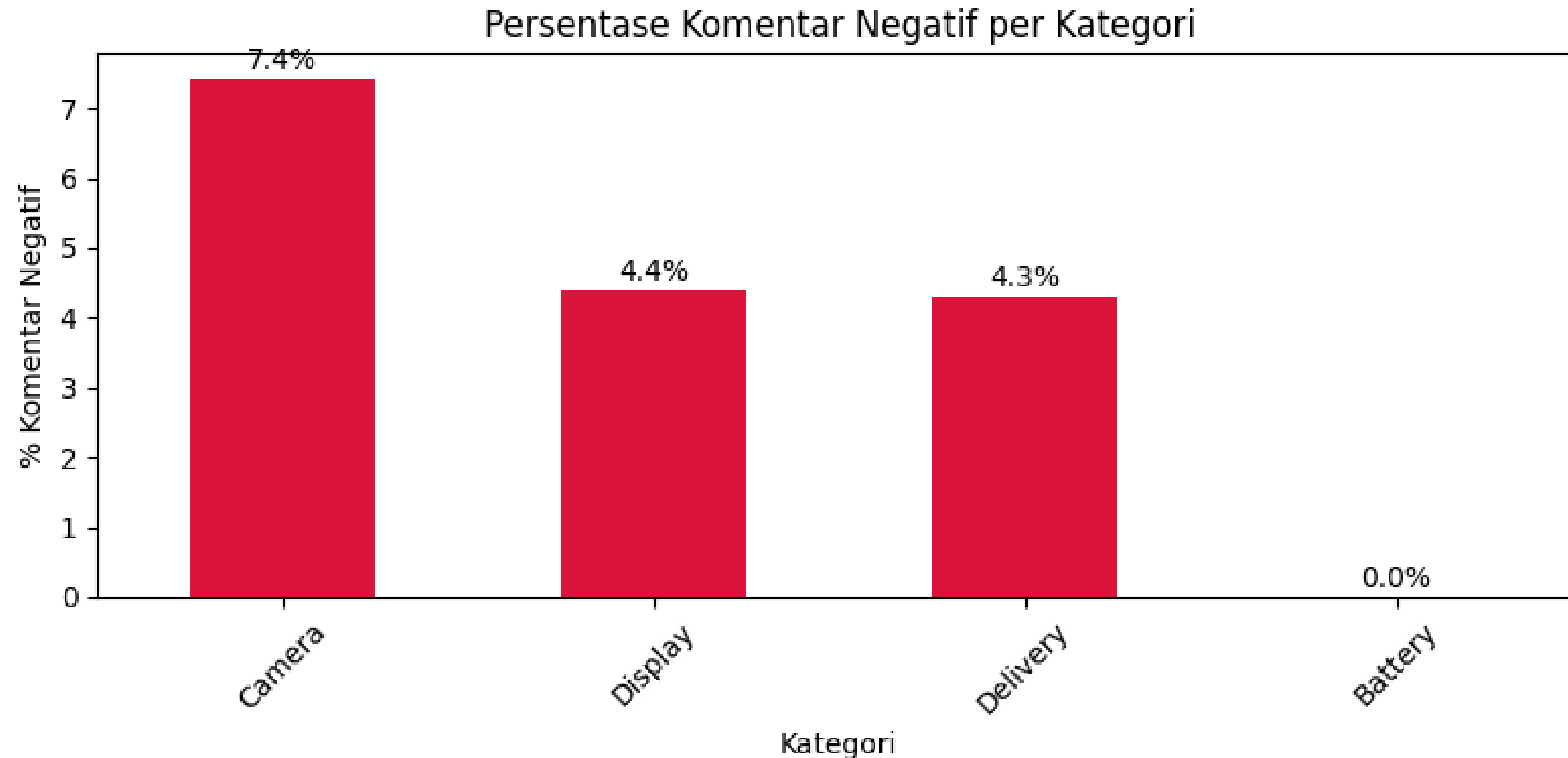
ANALISIS KATEGORI

Berikut merupakan **distribusi sentimen** pelanggan perkategori



ANALISIS KATEGORI

Berikut merupakan **persentase sentimen negatif** pelanggan perkategori



HASIL

- Dari analisis sentimen, **lebih banyak** pelanggan yang menyampaikan **ulasan positif** dibandingkan ulasan negatif ataupun netral
- Pelanggan umumnya menyampaikan **ulasan** mengenai **delivery**
- **Kamera** merupakan kategori yang persentase **ulasan negatifnya paling banyak**

INSIGHTS

- **Battery** merupakan kategori yang tidak memiliki sentimen negatif, menandakan bahwa baterai dapat menjadi **selling point** utama dari Redmi.
- **Camera** merupakan kategori yang **paling banyak memiliki sentimen negatif**, menandakan bahwa kualitas atau performa kamera perlu ditingkatkan untuk menarik lebih banyak pelanggan.
- **Display** juga **perlu diperhatikan** kualitasnya walaupun sentimen negatifnya tidak sebanyak camera.

SARAN

- Analisis lebih dalam dapat dilakukan dengan membedah kembali komentar-komentar yang ditulis oleh pelanggan.
- Pengelompokkan sentimen dapat lebih dioptimalkan untuk menghasilkan analisis yang lebih baik, mengingat akurasi prediksi sentimen tidak lebih dari 85%



THANK YOU