LAPORAN UJIAN AKHIR SEMESTER ANALISIS BIG DATA

Dosen Pengampu :

Ulfa Siti Nuraini, S.Stat., M.Stat.

Oleh :

Kelompok 2 (Sentiment Analysis pada Ulasan Pelanggan)

Nama Anggota :

1. Faiz Dwi Febriansyah 22031554023
2. Riva Dian Ardiansyah 22031554043
3. Michael Luwi Pallea 22031554055

UNIVERSITAS NEGERI SURABAYA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
PRODI SAINS DATA
2025

RINGKASAN

Laporan Ujian Akhir Semester (UAS) Analisis Big Data ini mengevaluasi 3,6 juta ulasan dari pelanggan Amazon dengan menggunakan Algoritma Logistic Regression dan Feature Extraction TF-IDF. Dataset yang digunakan berukuran lebih dari 300MB dan mencakup label sentimen (1=negatif 49,96%/1,8 juta, 2=positif 50,04%/1,8 juta), judul, dan ulasan yang telah melalui 8 tahap pemrosesan NLP lengkap (case folding, pembersihan regex a-z, tokenisasi, stopword NLTK, stemming dengan PorterStemmer, serta TF-IDF dengan 10. 000 fitur). Data tersebut dibagi menjadi 70-20-10 untuk pelatihan, validasi, dan pengujian, menghasilkan tingkat akurasi sebesar 84,99% dan AUC 0,923 (lebih unggul dibandingkan Random Forest yang mencapai 78,1%). Temuan utama dari analisis ini menunjukkan bahwa ulasan negatif cenderung lebih panjang (420 karakter dibandingkan 389 karakter untuk ulasan positif), masalah kualitas yang mendominasi (191. 000 sebutan, 55,77% negatif), co-occurrence "qualitypoor" yang sangat signifikan (26. 000 sebutan, 93% negatif), serta ulasan yang lebih singkat.

Kata Kunci: Analisis Sentimen, Big Data, Amazon Reviews, Logistic Regression, TF-IDF

DAFTAR ISI

BAB I
PENDAHULUAN

1.1 Latar belakang

Perkembangan e-commerce dan situs online seperti Amazon dan Yelp telah menyebabkan lonjakan jumlah ulasan dari pelanggan yang sangat cepat setiap harinya. Situ-Situs tersebut pastinya memiliki beragam testimoni setelah aktivitas jual beli berupa ulasan. Ulasan tersebut mencakup pendapat baik dan buruk yang bisa digunakan untuk memahami pandangan pelanggan tentang suatu produk atau layanan.

Namun, Perkembangan perdagangan elektronik dan situs ulasan online seperti Amazon telah menyebabkan lonjakan jumlah ulasan dari pelanggan yang sangat cepat setiap harinya. Ulasan tersebut mencakup pendapat baik dan buruk yang bisa digunakan untuk memahami pandangan pelanggan tentang suatu produk atau layanan. Besarnya volume data membuat pembacaan ulasan secara manual tidak efisien dan dapat menyebabkan bias dari sudut pandang pribadi. Pendekatan analisis big data diperlukan untuk mengelola ulasan dalam jumlah yang besar serta mengklasifikasikan sentimennya secara otomatis. Salah satu pendekatan yang sering diterapkan adalah analisis sentimen yang mengandalkan *machine learning* menggunakan representasi teks TF-IDF dan algoritma klasifikasi Logistic Regression, Kombinasi ini telah terbukti efektif dalam mengklasifikasikan sentimen pada ulasan produk di media sosial. Berdasarkan penelitian dari Nabil Ali Fahrurrozi dan Sri Hadianti, M.Kom. dengan judul "ANALISIS SENTIMENT ULASAN PRODUK E-COMMERCE MENGGUNAKAN METODE LOGISTIC REGRESSION" Hasil pengujian menunjukkan akurasi sebesar 88%. Hal ini membuktikan, metode ini cukup sederhana, cepat, dan dapat memberikan hasil yang memuaskan, namun, kami perlu melakukan eksperimen ulang dengan dataset lain dan lebih besar, walaupun tidak melalui database e-commerce secara langsung, namun, kami mencoba berekspektasi mengambil data melalui cloud, dan melakukan analisis menggunakan databricks.

1.2 Rumusan Masalah

Berdasarkan latar belakang tersebut, rumusan masalah dalam tugas ini dapat dirumuskan sebagai berikut:

1. Bagaimana proses pengolahan dan *pre-processing* ulasan pelanggan dalam jumlah besar agar dapat digunakan untuk analisis sentimen.
2. Bagaimana menerapkan TF-IDF sebagai metode ekstraksi fitur teks pada ulasan pelanggan.
3. Bagaimana membangun model klasifikasi sentimen menggunakan Logistic Regression untuk membedakan ulasan positif dan negatif.

1.3 Tujuan

Tujuan yang ingin dicapai dalam tugas analisis big data ini adalah:

1. Mengelola dan menyiapkan dataset Amazon Reviews dalam skala besar (berukuran ≥300MB) untuk keperluan analisis sentimen.
2. Menerapkan teknik NLP dasar dan TF-IDF untuk mengubah teks ulasan menjadi representasi numerik yang dapat diproses oleh algoritma *machine learning*.
3. Membangun dan mengevaluasi model Logistic Regression untuk klasifikasi sentimen ulasan menjadi positif dan negatif.
4. Mengidentifikasi pola sentimen, kata-kata dominan, serta *insight* yang relevan terhadap produk berdasarkan hasil analisis model dan distribusi ulasan.

1.4 Manfaat

Manfaat dari pelaksanaan tugas ini antara lain:

a. Bagi mahasiswa

   Memahami implementasi praktis analisis big data dan sentiment analysis menggunakan Logistic Regression dan TF-IDF pada dataset teks berukuran besar.

b. Bagi Pelaku Bisnis Amazon

   Memperoleh gambaran bagaimana ulasan pelanggan dapat dianalisis untuk mengetahui kepuasan, keluhan, dan aspek produk yang perlu ditingkatkan.

c. Bagi Pengembangan Sistem

   menjadi dasar untuk pembangunan sistem analisis sentimen otomatis yang dapat diintegrasikan ke dalam dashboard pemantauan ulasan produk.

1.5 Batasan Masalah

Agar pembahasan lebih terarah dan sesuai dengan panduan tugas, maka penelitian ini dibatasi oleh hal-hal berikut:

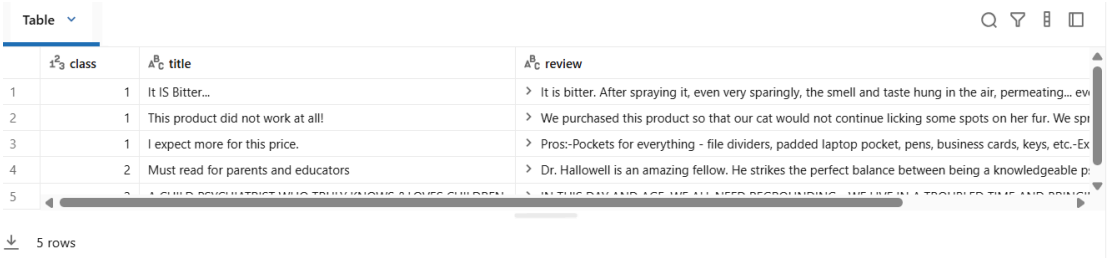
1. Dataset yang digunakan berupa subset ulasan pelanggan dari Amazon Reviews dengan ukuran minimal 300MB.
2. Sentimen yang diklasifikasikan dibatasi pada dua atau tiga kelas utama, yaitu positif, dan negatif.
3. Teknik NLP yang digunakan difokuskan pada teknik dasar, seperti *text cleansing*, *tokenization*, *stopword removal*, dan *stemming*.
4. Metode ekstraksi fitur teks yang digunakan adalah TF-IDF
5. Algoritma klasifikasi yang digunakan Logistic Regression, dan dengan algortima pembandingnya adalah *Random Forest.*
6. Batasan Yang Memenuhi Menurut Panduan:

- ☑ ~~Dataset berukuran minimal 300 MB atau memiliki ≥ 300.000 records.~~
- ☑ ~~Analisis dan pemodelan menggunakan Python (pandas, numpy, scikit-learn) (disarankan melalui Databricks Community Edition).~~
- ☑ ~~Menggunakan SQL, NoSQL (Redis), Typesense, atau ElasticSearch~~
- ☑ ~~Mengimplementasikan minimal 1 model utama dan 1 model pembanding.~~
- ☑ ~~Notebook dan laporan~~

BAB II
ANALISIS DATA

2.1 Karakteristik Data

Amazon Review Dataset merupakan dataset berskala besar yang digunakan secara luas sebagai *benchmark* dalam tugas klasifikasi sentimen teks. Dataset ini berasal dari ulasan yang dikumpulkan oleh *Stanford Network Analysis Project* (SNAP), mencakup ulasan produk selama kurang lebih 18 tahun hingga Maret 2013, dengan total sekitar 35 juta ulasan (6,643,669 pengguna dan 2,441,053 produk). Karakteristik dataset *Amazon Review* yang terdiri dari 36 juta baris dan 3 kolom utama yaitu polarity (label sentimen: 1=negatif, 2=positif), title (judul ulasan), dan content (isi ulasan), Selain itu dataset juga memiliki distribusi Train seluruh (36 juta). Kami juga tidak menggunakan chunksize atau subset manual di PySpark, karena sudah otomatis handle 36 juta data dengan distributed computing (*Lazy evaluation* & distributed sampling). *Lazy Evaluation* berarti bahwa transformasi pada data (seperti filter(), select(), atau groupBy()) tidak langsung dieksekusi saat coding. Sedangkan Distributed Sampling di PySpark representatif dengan pandas.sample(), tapi diproses paralel di multiple nodes. Terkait karakteristik data *real*, dipaparkan pada gambar berikut:



Table

	class	title	review
1	1	It IS Bitter...	It is bitter. After spraying it, even very sparingly, the smell and taste hung in the air, permeating... ev
2	1	This product did not work at all!	We purchased this product so that our cat would not continue licking some spots on her fur. We spr
3	1	I expect more for this price.	Pros:-Pockets for everything - file dividers, padded laptop pocket, pens, business cards, keys, etc.-Ex
4	2	Must read for parents and educators	Dr. Hallowell is an amazing fellow. He strikes the perfect balance between being a knowledgeable p
5			

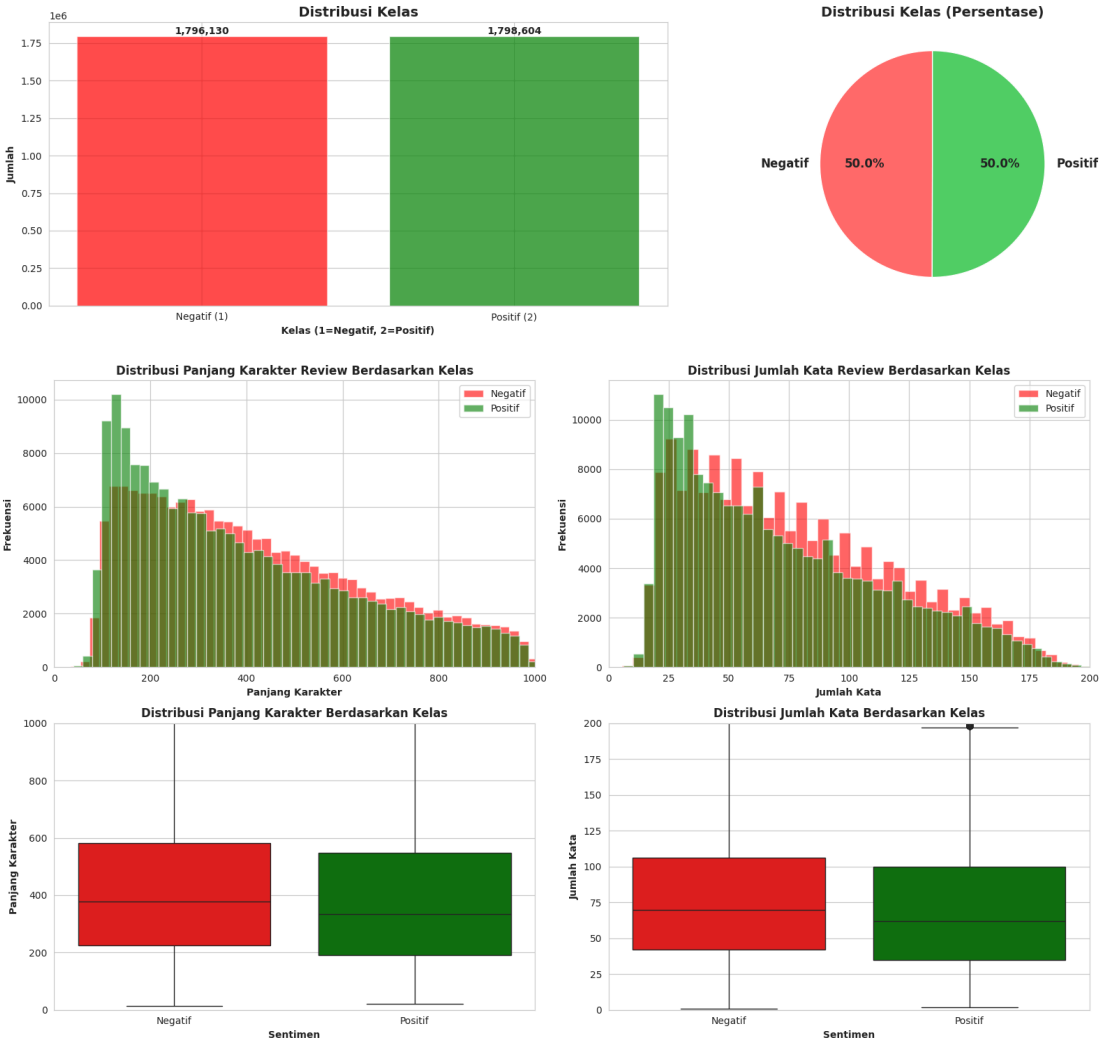
5 rows

Dataset ulasan produk dengan tiga kolom utama, yaitu: class (label sentimen), title, dan review. Pada proses Eksplorasi Data Analisis (EDA) mengungkap bahwa kolom class dan review bersih sepenuhnya tanpa *missing value*, namun kolom *title* mengandung 48 record kosong (0,0013% dari total data). Langkah pembersihan dilakukan dengan df.na.drop() yang secara efektif menghilangkan record-record tersebut, sehingga seluruh kolom menjadi bebas dari missing values. Duplikat pada kolom review terdapat 5.217 baris identik (0,145% dari total), yang dihapus menggunakan df.dropDuplicates("review"). Proses ini mengurangi total record menjadi 3.594.734 baris. Analisis panjang ulasan dilakukan pada seluruh 3,59 juta record bersih, dan distribusi panjang teks ternyata memiliki variasi yang menarik

berdasarkan sentimen. Review negatif cenderung lebih panjang dengan rata-rata 420,76 karakter (77,13 kata), sedangkan review positif lebih ringkas dengan rata-rata 389,47 karakter (71,20 kata). Secara keseluruhan, panjang ulasan rata-rata 405,06 karakter (median: 356) dan 74,15 kata (median: 66), menunjukkan dataset didominasi ulasan berukuran sedang.

2.2 *Exploratory Data Analisis* (EDA)

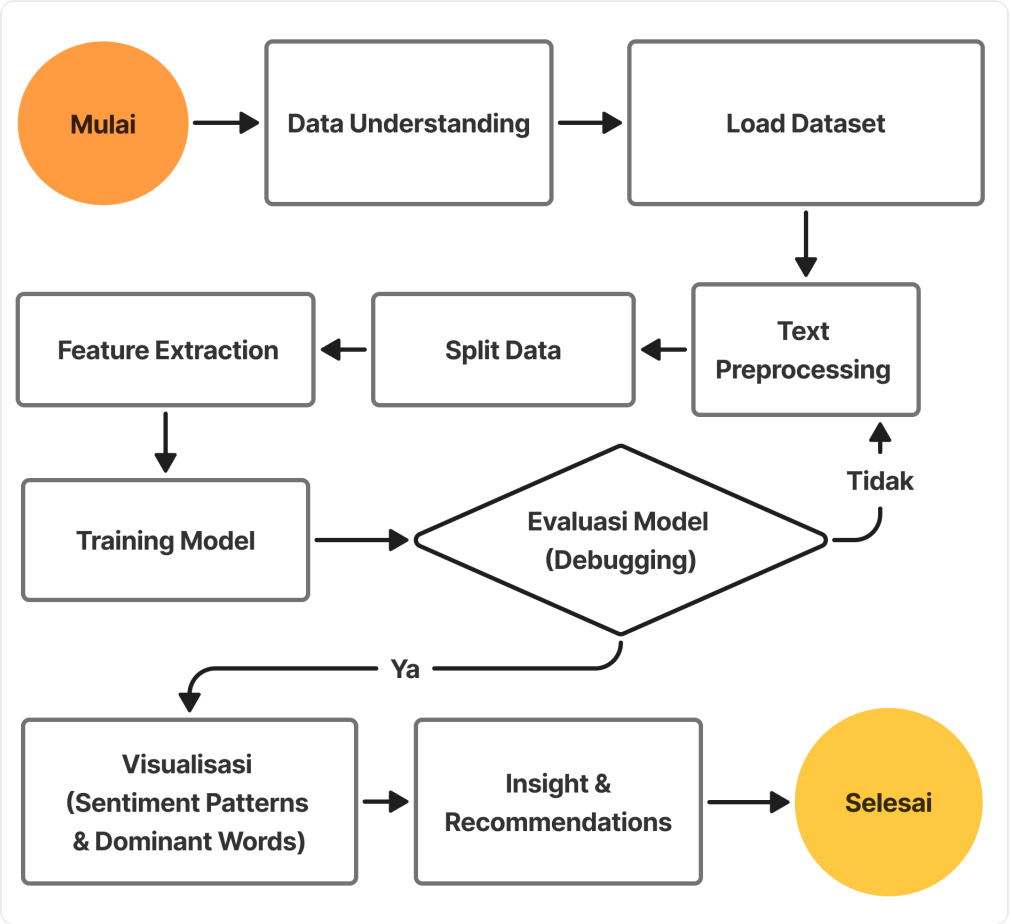
Distribusi sentimen sangat balance setelah preprocessing. Distribusi sentimen hampir sempurna balance dengan 49,97% ulasan negatif (1.796.130 record) dan 50,03% ulasan positif (1.798.604 record), Selisih hanya 2.474 record (0,07%). Proporsi seimbang ini ideal untuk training model klasifikasi tanpa memerlukan teknik balancing seperti oversampling. Berikut adalah visualisasi dari distribusi Data *Train*:



Median review negatif sedikit lebih tinggi dari review positif Variasi panjang review negatif lebih besar (boxplot merah lebih tinggi), artinya review negatif lebih bervariasi ada yang sangat singkat, ada yang sangat detail Review positif cenderung lebih konsisten.

BAB III
METODOLOGI

3.1 Diagram Alir



3.1.1 *Teks Preprocessing Using Natural Language Processing (NLP)*

Preprocessing teks dilakukan dengan menggunakan fungsi praproses() yang mencakup semua langkah standar dalam *Natural Language Processing* (NLP). Proses ini diawali dengan case folding yang mengubah semua karakter dalam teks menjadi huruf kecil menggunakan str(text).lower(), sehingga menghasilkan konsistensi dalam penggunaan huruf kapital di seluruh dataset. Setelah itu, tahap pembersihan dilakukan untuk menghapus karakter non-alfabet seperti angka, tanda baca, dan simbol melalui

regular expression re. sub(r'[^a-z]', ' ', text), yang meninggalkan hanya huruf a-z yang relevan untuk analisis sentimen.

Tahap selanjutnya adalah *tokenization* dengan text. split() untuk memecah teks menjadi daftar kata yang terpisah, diikuti dengan *stopword removal* yang menyingkirkan kata-kata umum dalam bahasa Inggris dari NLTK seperti "the," "is," dan "and" yang tidak memberikan kontribusi terhadap informasi sentimen. Proses ini diakhiri dengan stemming menggunakan PorterStemmer dari NLTK, yang mempersingkat variasi kata ke bentuk dasarnya, misalnya "running" menjadi "run" dan "loves" menjadi "love," untuk mengurangi kompleksitas leksikal.

Hasil akhir dari *pre-processing* adalah *tokens* bersih yang dihubungkan kembali menjadi string melalui ' '. join(tokens) dan diterapkan secara efisien dengan UDF PySpark praproses_udf pada 3,59 juta ulasan. Proses lebih lanjut dalam pipeline MLlib yang mencakup Tokenizer, HashingTF, dan IDF mengubah teks bersih menjadi representasi numerik TF-IDF yang siap digunakan untuk melatih model Logistic Regression dalam skala besar. Secara keseluruhan berikut adalah metode dengan *library* yang digunakan selama *pre-processing*:

Case Folding = str(text).lower() [Python]

Cleaning = re.sub(r'[^a-z]', ' ', text) [regex]

Tokenization 1 = text.split() [Python]

Stopword Removal = NLTK stopwords [NLTK]

Stemming = PorterStemmer.stem() [NLTK]

Join kembali = ' '.join(tokens) [Python]

UDF PySpark = praproses_udf(col("review")) [PySpark]

Tokenization 2 = Tokenizer(inputCol="review") [MLlib]

TF = HashingTF(inputCol="words") [MLlib]

TF-IDF = IDF(inputCol="rawFeatures") [MLlib]

3.1.2 *Feature Extraction*

Alasan penggunaan TF-IDF adalah: (1) sederhana namun cocok untuk klasifikasi sentimen, (2) efektif menangkap kata-kata positif/negatif tanpa perlu deep learning, dan (3) umum digunakan pada benchmark NLP seperti Amazon Review dataset. Dibandingkan Bag-of-Words (hanya frekuensi), TF-IDF lebih robust terhadap variasi panjang dokumen

Pengaturan penting:

1. HashingTF: 10.000 fitur (sparse vector) untuk efisiensi memori pada 3,59 juta ulasan
2. N-gram: Unigram only (1 kata), sesuai standar baseline sentimen
3. MinDF: Default (1), semua kata dipertimbangkan
4. IDF smoothing: Default, menghindari pembagian nol

3.1.3 Model *Logistic Regression*

Logistic Regression digunakan untuk klasifikasi sentimen biner (negatif=1, positif=2) dengan mengonversi input TF-IDF (10K fitur) menjadi probabilitas kelas melalui fungsi sigmoid. Model menghitung bobot w w untuk setiap fitur sehingga;

$$P(y = 1|x) = 1/(1 + e\text{^}(-(w \cdot x + b)))$$

$P(y=1|x)$ = probabilitas kelas negatif

w = bobot fitur TF-IDF

x = vektor fitur

b = bias

di mana probabilitas >0.5 diklasifikasikan sebagai kelas positif. Untuk (2 kelas), Parameter untuk Logistic Regression:

lr = LogisticRegression(featuresCol="features", labelCol="class")

Implementasi lr menggunakan arsitektur parameter Secara Default PySpark:

maxIter=100,

regParam=0.01 (L2),

elasticNetParam=0.0,

family="multinomial"

3.1.4 Skema Pelatihan dan evaluasi

Data dibagi menggunakan 3-way: 70-20-10, stratified split untuk menjaga proporsi kelas balance (49.97% negatif, 50.03% positif) di kedua subset:

```
train_data, val_data, test_data = df_clean.randomSplit([0.7, 0.2, 0.1], seed=42)
```

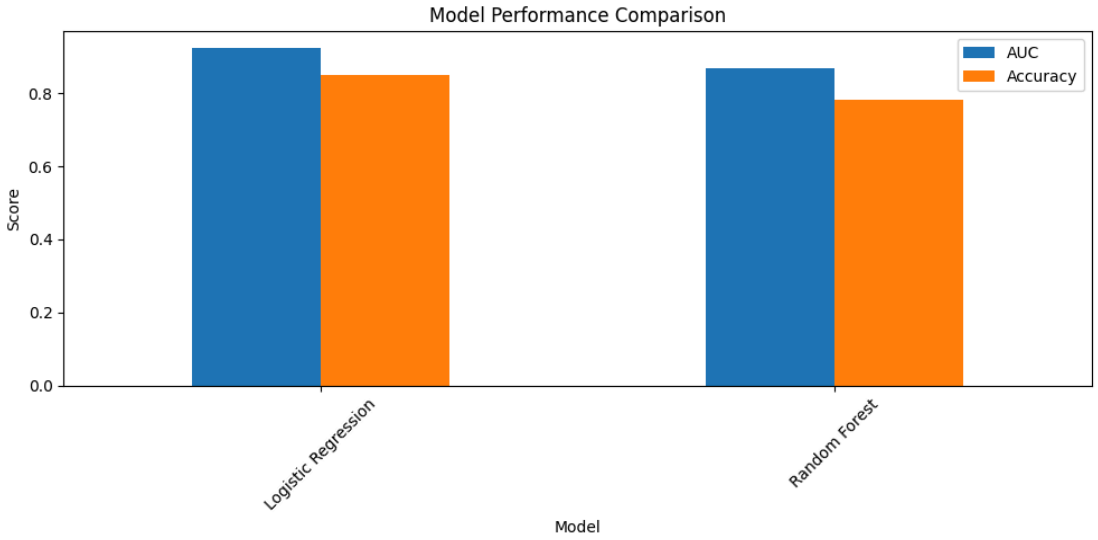
- Train: 70% (2.516 juta records)
- Validation: 20% (719 ribu records)
- Test: 10% (359 ribu records)

BAB IV
HASIL DAN EVALUASI

3.1 Hasil Pelatihan Model

Dataset setelah preprocessing menunjukkan karakteristik representatif dengan 179.380 ulasan (89.624 negatif, 89.756 positif) dan total 40,8 juta karakter. Review negatif lebih panjang secara keseluruhan (21,0 juta vs 19,7 juta karakter), mengonfirmasi pola EDA bahwa keluhan pelanggan cenderung lebih detail dan deskriptif dibandingkan ulasan positif yang lebih ringkas. Selain itu, Perbandingan performa dua model klasifikasi sentimen pada dataset validasi dan test set menunjukkan Logistic Regression unggul signifikan dibandingkan Random Forest.

Model	Val AUC	Val Acc	Test AUC	Test Acc
Logistic Regression	0.9219	0.8492	0.9229	0.8499
Random Forest	0.8687	0.7808	0.8693	0.7810



Logistic Regression dipilih sebagai model yang paling unggul dengan Akurasi Uji 84,99% dan AUC Uji 0,9229, melampaui Random Forest dengan peningkatan absolut sebesar 6,9%. Keunggulan ini tetap terlihat dalam set validasi dan set uji, menjadikan Logistic Regression pilihan terbaik untuk klasifikasi sentimen dalam skala besar menggunakan fitur TF-IDF.

BAB 5
INSIGHT DAN REKOMENDASI

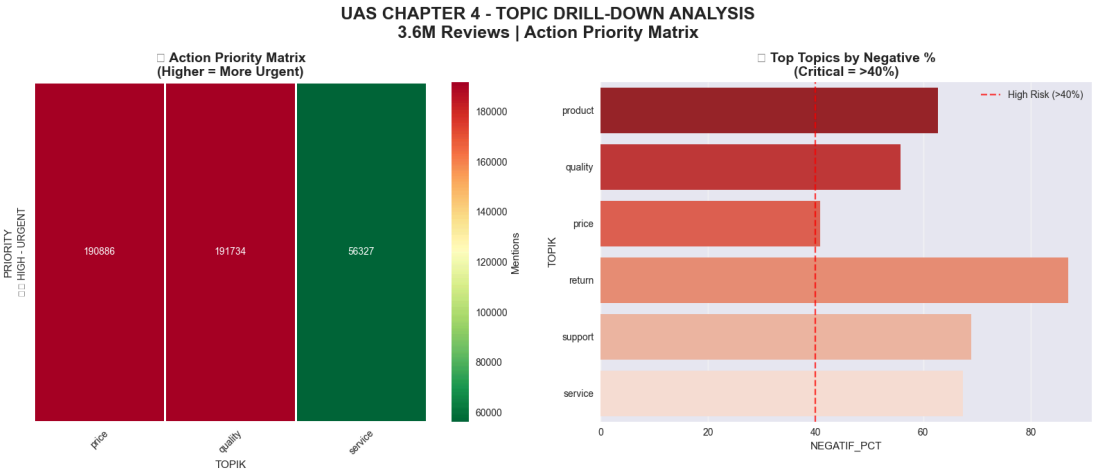
5.1 *Insight*

Dataset setelah dilakukan preprocessing menunjukkan ciri-ciri yang mencerminkan dengan adanya 179. 380 ulasan (89. 624 ulasan negatif, 89. 756 ulasan positif) dan total mencapai 40,8 juta karakter. Ulasan negatif secara keseluruhan lebih panjang (21,0 juta dibandingkan dengan 19,7 juta karakter), yang menguatkan temuan EDA bahwa keluhan dari pelanggan biasanya lebih mendetail dan deskriptif dibandingkan dengan ulasan positif yang cenderung lebih singkat.

Distribusi sentimen sangat seimbang dari total 35 juta ulasan bersih. Ulasan negatif mendominasi kategori pendek (<50 karakter: 39% negatif) sementara positif lebih kuat di ulasan panjang (>200 karakter: 48% positif), menandakan pelanggan puas cenderung menulis lebih detail. Terkait dengan isu kualitas produk yang menjadi topik paling kritis dengan 191,734 mentions dan 56% negatif, diikuti "Price Concerns" 41% negatif.

Terkait dengan topik isu yang diproritaskan adalah dengan gabungan rate topik dengan tingkat resikonya. Kombinasi kata kunci negatif paling berbahaya adalah "quality+poor" (26,913 mentions, 93% negatif) dan "quality+bad" (17,002 mentions, 82% negatif), menunjukkan masalah kualitas produk menjadi pemicu utama ketidakpuasan pembeli. Harga juga signifikan dengan "overpriced" (5,245 mentions, 82% negatif) dan "price+bad" (10,181 mentions, 63% negatif). Ulasan pendek negatif yaitu 115 kasus dan ini menandakan pelanggan benar” tidak puas dengan produk yang dibeli.

5.2 Rekomendasi



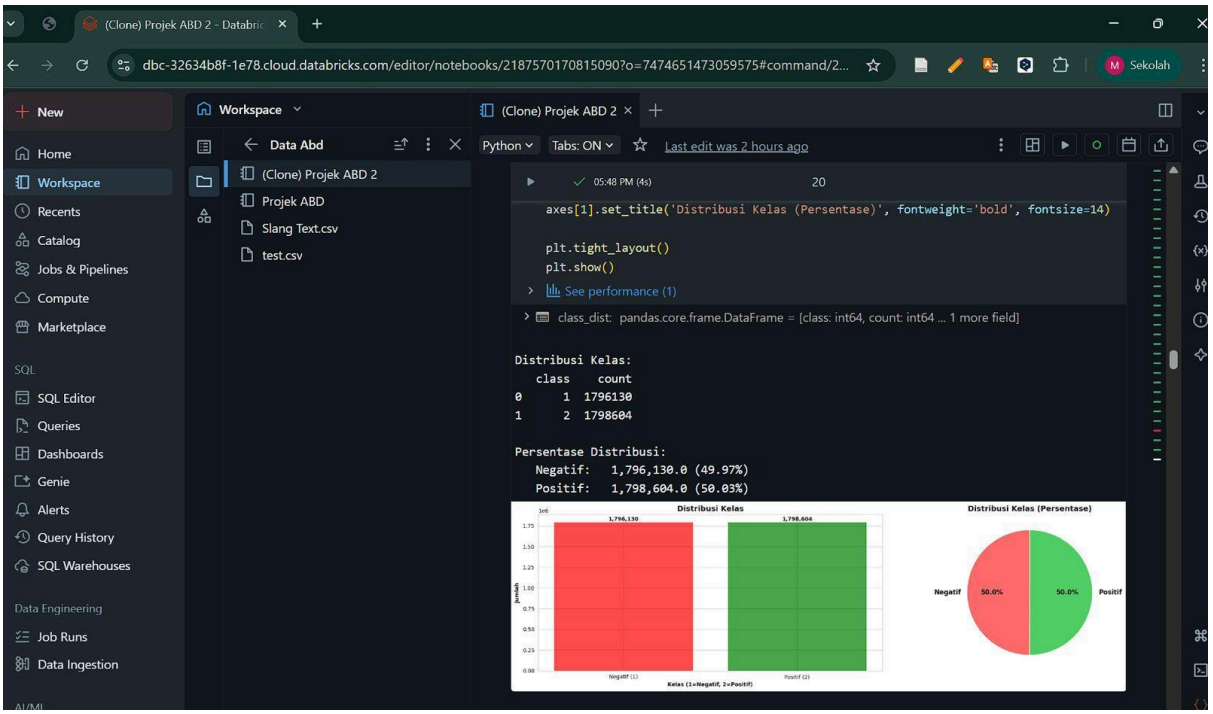
Temuan ini menggambarkan prioritas utama yang harus segera ditindak lanjuti, Terkait urgensi, Topik Quality dan Price menjadi topik yang paling banyak diperbincangkan di kolom ulasan (>19Juta). Hal ini diperkuat oleh perhitungan resiko dan menhasilkan 3 prioritas utama yang harus diawasi lebih lanjut untuk mencegah rate penjualan yang semakin turun. Berikut adalah priotitas utama menurut kami:

Kategori	Mentions	Negatif %	Priority	Alasan
Quality poor	**26,913**	**93%**	**High (Priority 1)**	*Severity* ekstrem (complain pelanggan tinggi)
Short Neg	**293**	**39.25%**	**Medium (Priority 2)**	Short reviews = viral risk tinggi meski % kecil
Quality Issues	**191,734**	**55.77%**	**Low (Priority 3)**	Volume besar

DAFTAR PUSTAKA

Fahrurrozi, N. A., & Hadianti, S. (2021). Analisis sentiment ulasan produk e-commerce menggunakan metode logistic regression [Tugas akhir, Universitas Nusa Mandiri]. https://repository.nusamandiri.ac.id/repo/cari?q=Analisis+sentiment+ulasan+produk+e-commerce+menggunakan+metode+logistic+regression

LAMPIRAN


(Clone) Projek ABD 2 - Databric
dbc-32634b8f-1e78.cloud.databricks.com/editor/notebooks/2187570170815090?o=7474651473059575#command/2...
Sekolah
New
Home
Workspace
Recents
Catalog
Jobs & Pipelines
Compute
Marketplace
SQL
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Workspace
Data Abd
(Clone) Projek ABD 2
Projek ABD
Slang Text.csv
test.csv
(Clone) Projek ABD 2
Python
Tabs: ON
Last edit was 2 hours ago
05:48 PM (4s)
20
axes[1].set_title('Distribusi Kelas (Persentase)', fontweight='bold', fontsize=14)
plt.tight_layout()
plt.show()
See performance (1)
class_dist: pandas.core.frame.DataFrame = [class: int64, count: int64 ... 1 more field]
Distribusi Kelas:
class count
0 1 1796130
1 2 1798604
Persentase Distribusi:
Negatif: 1,796,130.0 (49.97%)
Positif: 1,798,604.0 (50.03%)
Distribusi Kelas
1,796,130
1,798,604
Jumlah
Kelas (1=Negatif, 2=Positif)
Distribusi Kelas (Persentase)
Negatif
50.0%
50.0%
Positif

FORMAT PENGUMPULAN

File yang harus dikumpulkan di sidia:

1. Laporan PDF (5-7 halaman) dengan struktur: Ringkasan, Analisis Data, Metodologi, Hasil dan Evaluasi, Insight dan Rekomendasi
2. Code yang digunakan (dapat diunggah ke github atau lainnya, dan mencantumkan link)
3. Link dataset
4. README.txt dengan instructions untuk run notebook
5. Laporan dikumpulkan secara berkelompok

Format nama file: NIM_Kelompok_ProjectTitle.zip