DOKUMENTASI CRISP-DM
ANALISIS BERITA DIY OKTOBER 2025

Disusun oleh:

Nama	: Assyfa Nur Fathona
NIM	: 23/513289/SV/22210
Kelas	: PL5BB
Dosen Pengampu	: Ahmad Iwan Fadli, S.Kom., M.Eng.

PROGRAM STUDI D-IV TEKNOLOGI REKAYASA PERANGKAT LUNAK
DEPARTEMEN TEKNIK ELEKTRO DAN INFORMATIKA
SEKOLAH VOKASI
UNIVERSITAS GADJAH MADA
YOGYAKARTA
2025

A. Business Understanding

A.1. Tujuan

Tahap Business Understanding bertujuan untuk memahami konteks bisnis dan tujuan analisis data yang akan dilakukan. Dalam konteks ini, data yang dianalisis berasal dari kumpulan berita daring yang dipublikasikan oleh berbagai media di wilayah Yogyakarta pada bulan Oktober 2025. Melalui proses pemahaman konteks ini, ditentukan informasi apa saja yang dapat diperoleh dari dataset untuk mendukung analisis pola pemberitaan media, khususnya dalam memahami karakteristik berita, sentimen terhadap institusi (seperti kepolisian), serta dinamika publikasi dan pembaruan berita di media daring.

A.2. Latar Belakang

Media daring merupakan salah satu sumber informasi utama masyarakat dalam mengikuti perkembangan isu sosial, politik, dan keamanan. Dalam era digital, arus berita yang cepat dan masif perlu dianalisis agar dapat diketahui:

- Bagaimana pola sentimen publik dan media terhadap isu tertentu, khususnya pemberitaan tentang kepolisian.
- Topik dan kategori berita apa yang paling sering muncul dalam periode tertentu.
- Seberapa sering berita diperbarui atau mengalami koreksi setelah dipublikasikan.
- Pola sebaran lokasi dan waktu publikasi berita untuk melihat fokus wilayah peliputan.
- Performa sistem publikasi (misalnya waktu berita masuk ke Elasticsearch dan diketahui oleh polisi).

Analisis ini penting untuk memberikan gambaran empiris mengenai citra kepolisian di media, aktivitas pemberitaan di wilayah Yogyakarta, dan efisiensi sistem penyebaran informasi digital.

A.3. Permasalahan Bisnis (Business Objectives)

Berdasarkan latar belakang di atas, tujuan utama dari analisis ini adalah:

1. Mengidentifikasi kategori, topik, dan jenis berita yang paling dominan dipublikasikan di wilayah Yogyakarta.
2. Menganalisis sentimen umum dan sentimen terhadap polisi pada berita daring.
3. Mengevaluasi emosi yang muncul dalam pemberitaan dan hubungannya dengan jenis/topik berita.
4. Mengukur perbedaan waktu antara pembuatan berita (created) dan publikasi (published) untuk mengetahui kecepatan proses editorial.
5. Menganalisis pola waktu berita diperbarui (is_updated) dan waktu diketahui oleh kepolisian (known_by_police).
6. Melihat sebaran geografis berita berdasarkan atribut kota, provinsi, dan koordinat (geo).
7. Mengevaluasi aktivitas publikasi berita ke sistem Elasticsearch (published_to_es) sebagai indikator efisiensi distribusi data.

A.4. Hasil yang Diharapkan

Dari analisis ini, diharapkan diperoleh insight sebagai berikut:

1. Visualisasi sebaran berita berdasarkan lokasi dan topik pemberitaan.
2. Identifikasi dominasi sentimen dan emosi, baik secara umum maupun terhadap polisi.

3. Informasi kecepatan publikasi dan pembaruan berita, termasuk jeda antara waktu dibuat, dipublikasikan, dan diperbarui.
4. Gambaran efektivitas sistem distribusi berita melalui Elasticsearch serta waktu berita diketahui oleh kepolisian.
5. Evaluasi aktivitas media daring di wilayah Yogyakarta selama periode analisis.

Hasil analisis ini akan disajikan dalam bentuk dashboard interaktif menggunakan Tableau, sehingga dapat digunakan oleh peneliti, analis media, atau lembaga pemerintahan untuk memahami pola pemberitaan, citra publik terhadap institusi tertentu, serta dinamika informasi di media daring.

A.5. Tabel Mapping Use Case

No_Usecase	Identifikasi Usecase	Deskripsi	Data yang diolah	Jenis Grafik
1	Distribusi Berita per Topik	Mengetahui sebaran jumlah berita berdasarkan topik.	topic	Pie Chart, Bar Chart
2	Sebaran Bahasa Berita	Melihat proporsi berita berdasarkan bahasa yang digunakan.	language	Pie Chart
3	Sentimen Berita	Menganalisis persebaran sentimen (positif, negatif, netral) dari berita.	sentiment	Pie Chart, Bar Chart
4	Emosi Dominan dalam Berita	Mengidentifikasi emosi yang paling sering muncul (marah, sedih, senang, takut, dll).	emotion	Bar Chart
5	Analisis Topik Berdasarkan SDGs	Melihat keterkaitan antara berita dan tujuan SDGs tertentu.	sdgs, topic	Treemap
6	Sebaran Sumber Media	Mengetahui 10 media teratas yang paling aktif mempublikasikan berita.	site	Bar Chart
7	Persebaran Sentimen Berdasarkan SDGs	Menilai apakah berita dengan tema SDGs cenderung bernada positif atau negatif.	sdgs, sentiment	Treemap
8	Keterkaitan	Melihat kota yang	topic, kota	Treemap

	Antara Topik dan Kota	memiliki berita terbanyak di setiap topik.		
9	Analisis Waktu Publikasi	Mengetahui waktu yang paling sering digunakan untuk publikasi berita.	published	Pie Chart
10	Jumlah Berita yang Dipublikasikan per Hari	Mengamati jumlah berita yang dipublikasikan tiap hari	published	Line Chart
11	Sinkronisasi data	Mengamati perbandingan berita yang sudah yang sudah dan yang belum dipublikasikan ke Elasticsearch	published_to_es	Pie Chart
12	Waktu Berita Dipublikasikan	Mengamati frekuensi lama waktu yang diperlukan berita untuk dipublikasikan setelah berita dibuat	published, created	Pie Chart
13	Berita yang Diperbarui	Mengamati frekuensi berita yang diperbarui (update) dan yang tidak diperbarui	is_updated	Pie Chart
14	Frekuensi update per media	Menampilkan 10 media dengan jumlah update berita tertinggi	is_updated, site	Bar Chart
15	Perbandingan Berita yang Di-update per Topic	Menganalisis perbandingan berita yang diperbarui dan yang tidak diperbarui per topik	topic, is_updated	Stacked Bar Chart
16	Persebaran berita di Indonesia	Menampilkan persebaran berita di seluruh kota	kota, provinsi	Map
17	Jumlah Berita di Tiap Kota	Menampilkan 10 kota dengan jumlah berita terbanyak	kota	Bar Chart

18	Citra polisi di media	Mengetahui sebaran sentimen polisi	sentiment_polisi	Bar Chart
19	Sentimen Berita per Topic	Mengetahui persebaran sentimen berita pada masing-masing topik	topic, sentiment	Stacked Bar Chart
20	Emotion Berita per Topic	Mengetahui persebaran emosi pembaca terhadap berita pada masing-masing topik	emotion, topic	Stacked Bar Chart

B. Data Understanding

B.1. Tujuan

Tahap Data Understanding bertujuan untuk melakukan eksplorasi awal terhadap dataset, memahami struktur dan isi data, serta mengidentifikasi potensi permasalahan yang mungkin timbul dalam proses analisis. Proses ini mencakup evaluasi terhadap tipe data, distribusi nilai, kelengkapan data, serta relevansinya terhadap kebutuhan analisis yang telah ditentukan pada tahap Business Understanding.

B.2. Struktur Dataset

Dataset yang dianalisis adalah data berita dari wilayah Jogja pada bulan Oktober 2025 yang berisi total 2109 baris (record) dan 37 atribut (kolom). Setiap baris mewakili satu berita yang dipublikasikan. Struktur dataset ini mencakup beberapa aspek utama, yaitu informasi isi berita, analisis teks dan emosi, lokasi pemberitaan, serta atribut waktu dan teknis sistem. Berikut ini rincian kategori atribut yang tersedia dalam dataset:

1. Informasi Umum Berita

   Atribut-atribut ini memuat informasi dasar tentang setiap berita yang dipublikasikan:
   - title – Judul berita.
   - summary – Ringkasan berita.
   - fulltext – isi lengkap teks berita.
   - desc – deskripsi tambahan atau hasil ekstraksi isi berita.
   - link – tautan (URL) sumber berita.
   - site – nama situs atau media penerbit berita.
   - favicon – URL ikon situs media.
   - media_url – URL media atau gambar.
   - highlight – cuplikan atau bagian penting berita yang disorot.
   - subs_name – nama sub-sumber atau akun publikasi.
2. Analisis Teks dan Emosi

   Atribut-atribut ini diperoleh dari proses analisis natural language processing (NLP):
   - sentiment – sentimen umum berita (positif, netral, negatif).
   - emotion – emosi dominan dalam berita (marah, sedih, senang, dll).
   - sentiment_polisi – sentimen yang secara khusus mengarah pada polisi.
   - emotion_polisi – emosi yang diarahkan terhadap polisi.

- topic – topic utama berita (misal: wisata, sport, lifestyle, dll).
- sdgs – klasifikasi berdasarkan tujuan pembangunan berkelanjutan (SDGs).
- phrases – frasa penting hasil ekstraksi teks.
- keywords – kata kunci utama dari isi berita.
- ners – Named Entities Recognition – entitas yang disebut dalam teks (misal: nama orang, tempat, lembaga).
- alchemy – skor atau hasil analisis semantik tambahan (hasil dari sistem NLP).
- summarization – ringkasan hasil pemrosesan otomatis.
- hashtitle – representasi hash dari judul berita (untuk identifikasi unik).

3. Informasi Lokasi
   Terdapat beberapa kolom yang menjelaskan lokasi terkait berita, baik secara teks maupun koordinat:
   - kota – nama kota tempat kejadian atau pemberitaan.
   - provinsi – nama provinsi (banyak nilai kosong, terutama untuk berita luar wilayah).
   - geo – data koordinat geografis dalam bentuk list.
   - location_aho – lokasi alternatif hasil ekstraksi sistem.
4. Atribut Waktu
   Kolom-kolom ini berhubungan dengan waktu publikasi, pembaruan, dan proses teknis sistem:
   - created – waktu berita pertama kali direkam dalam sistem.
   - publised – waktu berita dipublikasikan oleh media.
   - published_all – waktu berita diterbitkan ke sistem agregator.
   - publised_to_es – waktu berita dimasukkan ke Elasticsearch.
   - is_updated – waktu pembaruan terakhir (atau status update).
   - known_by_police – waktu berita diketahui oleh pihak kepolisian (atau nama situs pelapor).
5. Informasi Penulis dan Metadata Tambahan
   Beberapa atribut menjelaskan informasi penulis dan informasi berita tambahan:
   - authors – nama penulis berita.
   - username – nama akun atau pengguna yang terkait dengan berita.
   - hashtag – tagar yang digunakan dalam berita.
   - topic – tema besar pemberitaan.
   - jenis – jenis berita (misalnya: fwp, kriminal, sosial, dll).
   - language – bahasa yang digunakan dalam berita.

B.3. Kualitas dan Kelengkapan Data

Aspek	Temuan
Missing Values	Beberapa kolom memiliki data kosong (missing), terutama: highlight, geo, is_updated, published_to_es, published_all, location_aho, hashtag, subs_name.
Redundansi	Tidak ada data yang duplikat.

Format Waktu	Kolom waktu (created, published, dll) dalam format teks yang bisa dikonversi ke datetime.

Analisis awal menunjukkan bahwa terdapat beberapa atribut dengan nilai kosong (missing values) yang signifikan, antara lain:

- Atribut favicon, highlight, geo, location_aho, hashtag, dan subs_name memiliki lebih dari 1.000 data kosong, sehingga atribut ini kurang relevan untuk digunakan dalam analisis.
- Atribut is_updated, published_to_es, published_all, known_by_police (beberapa merupakan data waktu/text dan boolean) juga memiliki lebih dari 900 data kosong, tetapi bisa diubah menjadi type boolean.
- Atribut teks seperti language memiliki sebagian kecil nilai kosong (19 baris), yang dapat diisi dengan string deskriptif umum seperti "lainnya".
- Beberapa atribut waktu masih menggunakan format string dan memerlukan konversi ke format datetime untuk keperluan analisis waktu.

B.4. Evaluasi Kecocokan Data dengan Use Case

No_Usecase	Identifikasi Use Case	Ketersediaan Data	Evaluasi dan Catatan
1	Distribusi Berita per Topik	Lengkap	Kolom topic tersedia.
2	Sebaran Bahasa Berita	Sebagian Kosong	Sebagian kolom language memiliki missing values, dapat diisi dengan ‘lainnya’ sebagai default.
3	Sentimen Berita	Lengkap	Kolom sentiment lengkap.
4	Emosi Dominan dalam Berita	Lengkap	Kolom emotion lengkap.
5	Analisis Topik Berdasarkan SDGs	Lengkap	Kolom sdgs dan topic lengkap.
6	Sebaran Sumber Media	Lengkap	Kolom site lengkap.
7	Persebaran Sentimen Berdasarkan SDGs	Lengkap	Kolom sdgs dan sentiment lengkap.
8	Keterkaitan Antara Topik dan Kota	Sebagian Kosong	Kolom topic lengkap. Kolom kota sebagian kosong, dapat diisi dengan ‘lainnya.’

9	Analisis Waktu Publikasi	Lengkap	Kolom published lengkap, tetapi perlu diubah menjadi datetime
10	Jumlah Berita yang Dipublikasikan per Hari	Lengkap	Kolom published lengkap, tetapi perlu diubah menjadi datetime
11	Sinkronisasi data	Sebagian Kosong	Kolom published_to_es memiliki missing value dan nilai yang tidak konsisten (waktu dan boolean), nilai waktu diubah menjadi True, sedangkan nilai kosong diubah menjadi False.
12	Waktu Berita Dipublikasikan	Lengkap	Kolom published dan created lengkap, tetapi perlu diubah menjadi datetime
13	Berita yang Diperbarui	Sebagian Kosong	Kolom is_updated memiliki missing value dan nilai yang tidak konsisten (waktu dan boolean), nilai waktu diubah menjadi True, sedangkan nilai kosong diubah menjadi False.
14	Frekuensi update per media	Sebagian Kosong	Kolom site lengkap. Kolom is_updated memiliki missing value dan nilai yang tidak konsisten (waktu dan boolean), nilai waktu diubah menjadi True, sedangkan nilai kosong diubah menjadi False.
15	Perbandingan Berita yang Di-update per Topic	Sebagian Kosong	Kolom topic lengkap. Kolom is_updated memiliki missing value dan nilai yang tidak konsisten (waktu dan boolean), nilai waktu diubah menjadi True, sedangkan nilai kosong diubah menjadi False.
16	Persebaran berita di Indonesia	Sebagian Kosong	Kolom provinsi berisi nilai latitude dan longitude, perlu diubah menjadi 2 kolom latitude dan longitude.
17	Jumlah Berita di Tiap Kota	Sebagian Kosong	Kolom kota sebagian kosong, dapat diisi dengan ‘lainnya.’
18	Citra polisi di media	Lengkap	Kolom sentiment_polisi lengkap.
19	Sentimen Berita per Topic	Lengkap	Kolom topic dan sentiment lengkap.

20	Emotion Berita per Topic	Lengkap	Kolom emotion dan topic lengkap.

Secara keseluruhan, dataset ini mendukung kebutuhan analisis yang telah dirumuskan pada tahap Business Understanding. Berdasarkan hasil eksplorasi, dapat disimpulkan bahwa:

- Data memiliki cakupan yang luas dan mencakup informasi penting seperti waktu publikasi, waktu dibuat, waktu pembaruan, topik, bahasa, sentimen, emosi, SDGs, dan lokasi.
- Sebagian besar atribut yang diperlukan untuk menjawab usecase telah tersedia dan dapat digunakan setelah dilakukan pembersihan.
- Dataset ini juga memungkinkan analisis dari berbagai perspektif, baik dari sisi konten berita, persepsi publik terhadap institusi (khususnya kepolisian), maupun aktivitas media daring dalam menyebarkan informasi.
- Dari 20 usecase yang dirancang, seluruhnya dapat direalisasikan menggunakan data yang tersedia.
- Sebagian besar kendala berasal dari missing value, terutama pada kolom text seperti is_updated, published_to_es, published_all, dan known_by_police.
- Semua kendala dapat diatasi melalui tahap Data Preparation, dengan strategi pengisian nilai kosong dan konversi format data.

Dataset memiliki struktur yang lengkap dan cukup representatif untuk dilakukan analisis mendalam. Namun, perlu dilakukan tahap pembersihan data (data preparation) lebih lanjut untuk menangani missing values dan menyesuaikan format data dengan kebutuhan visualisasi. Tahap selanjutnya akan difokuskan pada proses tersebut sebelum data dimuat ke dalam dashboard visual menggunakan Tableau.

C. Data Preparation

C.1. Tujuan

Tahap Data Preparation bertujuan untuk membersihkan dan menyiapkan dataset agar dapat digunakan secara optimal dalam proses visualisasi data di Tableau. Kegiatan dalam tahap ini mencakup penghapusan atribut yang tidak relevan, penanganan nilai kosong, konversi tipe data, serta seleksi atribut berdasarkan kebutuhan masing-masing use case yang telah dirancang pada tahap Business Understanding.

C.2. Dokumentasi Pengolahan Data Berdasarkan Use Case

No_Usecase	Identifikasi Usecase	Atribut yang Digunakan	Proses Data Preparation
1	Distribusi Berita per Topik	topic	Tidak ada missing value.
2	Sebaran Bahasa Berita	language	Nilai kosong diisi dengan ‘lainnya.’ Menyamakan nilai yang tidak konsisten.

3	Sentimen Berita	sentiment	Tidak ada missing value.
4	Emosi Dominan dalam Berita	emotion	Tidak ada missing value.
5	Analisis Topik Berdasarkan SDGs	sdgs, topic	Tidak ada missing value.
6	Sebaran Sumber Media	site	Tidak ada missing value.
7	Persebaran Sentimen Berdasarkan SDGs	sdgs, sentiment	Tidak ada missing value.
8	Keterkaitan Antara Topik dan Kota	topic, kota	Nilai kosong pada kota diisi dengan 'lainnya.' Menangani data yang tidak konsisten.
9	Analisis Waktu Publikasi	published	Dibuat kolom baru (pagi, siang, sore, malam)
10	Jumlah Berita yang Dipublikasikan per Hari	published	Dikoversi ke datetime.
11	Sinkronisasi data	published_to_es	Nilai waktu diubah menjadi True, sedangkan nilai kosong diubah menjadi False.
12	Waktu Berita Dipublikasikan	published, created	Dikoversi ke datetime.
13	Berita yang Diperbarui	is_updated	Nilai waktu diubah menjadi True, sedangkan nilai kosong diubah menjadi False.
14	Frekuensi update per media	is_updated, site	Nilai waktu diubah menjadi True, sedangkan nilai kosong diubah menjadi False.
15	Perbandingan Berita yang Di-update per Topic	topic, is_updated	Nilai waktu diubah menjadi True, sedangkan nilai kosong diubah menjadi False.
16	Persebaran berita di Indonesia	kota, provinsi	Kolom provinsi dipisah menjadi kolom latitude dan longitude. Menangani

			data kota yang tidak konsisten.
17	Jumlah Berita di Tiap Kota	kota	Nilai kosong diisi dengan ‘lainnya.’ Menangani data yang tidak konsisten.
18	Citra polisi di media	sentiment_polisi	Tidak ada missing value.
19	Sentimen Berita per Topic	topic, sentiment	Tidak ada missing value.
20	Emotion Berita per Topic	emotion, topic	Tidak ada missing value.

C.3. Langkah Tambahan Lintas Use Case

- Penghapusan atribut tidak relevan seperti kolom authors, favicon, fulltext, hashtitle, jenis, keywords, link, media_url, summary, desc, ners, phrases, alchemy, highlight, geo, summarization, username, location_aho, hashtag, dan subs_name dihapus karena tidak berkontribusi terhadap analisis atau visualisasi.
- Penanganan kolom language dan kota yang memiliki nilai kosong diisi dengan string ‘lainnya.’
- Konversi kolom yang tidak konsisten seperti is_updated dan published_to_es dikonversi ke format boolean.

C.4. Kode Python

```
######### read dataset #########
import pandas as pd
import numpy as np
import ast
import re
df =
pd.read_csv('https://raw.githubusercontent.com/sifanurfa/dataset/refs/
heads/main/diy-news.csv', delimiter=',')
df.head()

######### konversi bahasa yang tidak konsisten #########
df['language'] = df['language'].replace('indonesian', 'id')
df['language'].unique()

######### menangani missing values #########
df['language'] = df['language'].fillna('tidak diketahui')
df['kota'] = df['kota'].fillna('lainnya')

######### menangani nilai kolom kota yang tidak konsisten #########
```

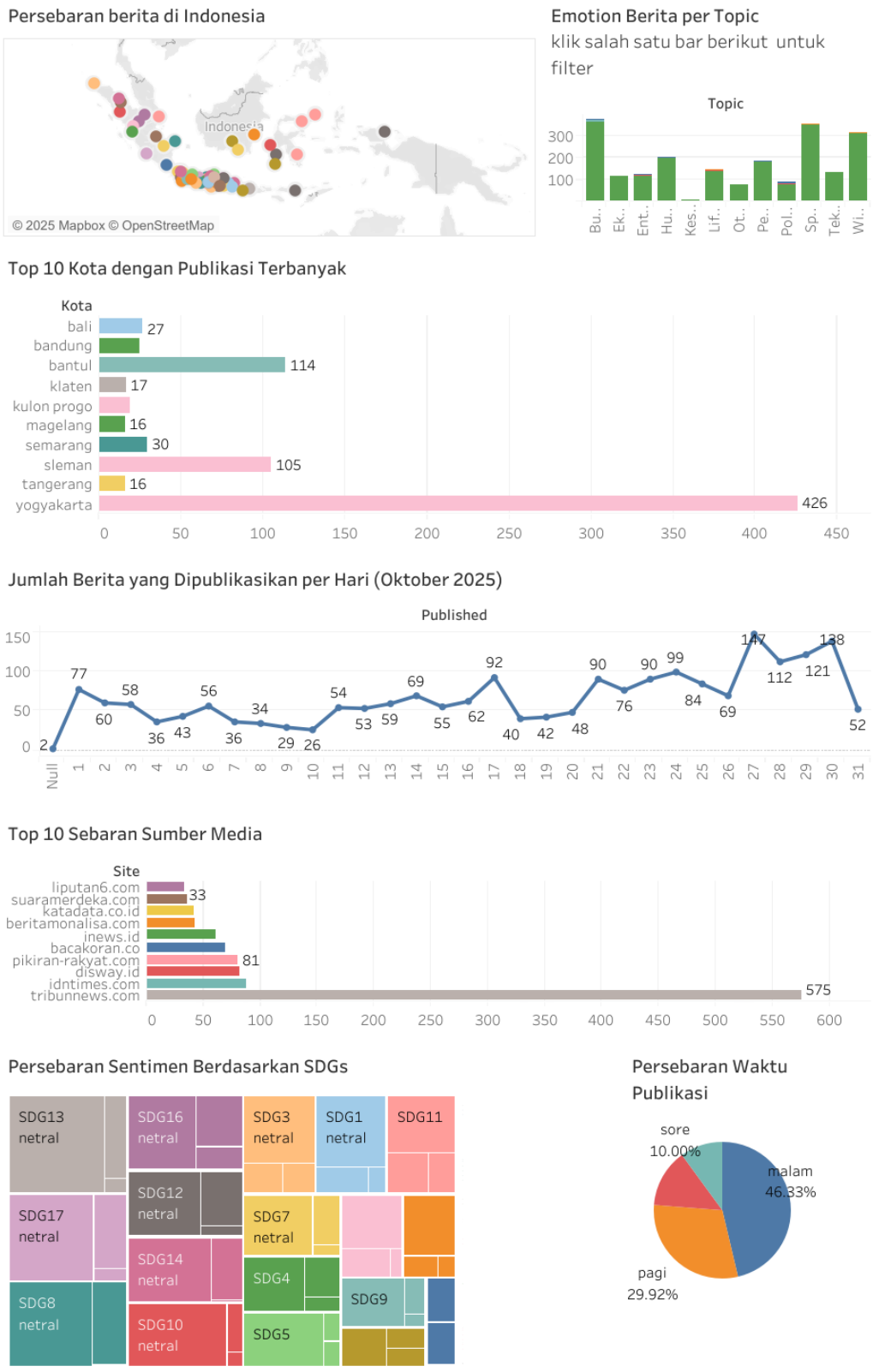
```
def handle_dict_string(x):
    if isinstance(x, str) and x.startswith('[{') and x.endswith('}]'):
        return 'lainnya'
    return x

def clean_list_string(x):
    # cek kalau berbentuk list string
    if isinstance(x, str) and (x.startswith('[') and x.endswith(']')):
        # hapus tanda [ ] dan ' , lalu strip spasi
        x_clean = re.sub(r"[\[\]']", '', x).strip()
        return x_clean
    return x

df['kota'] = df['kota'].apply(handle_dict_string)
df['kota'] = df['kota'].apply(clean_list_string)
df['kota'] = df['kota'].str.replace(r"^(kota|kabupaten|provinsi)\s+",
"", regex=True)
df['kota'] = df['kota'].replace("{'authors': None}", 'lainnya')

######### ubah kolom provinsi jadi latitude dan longitude #########
def extract_lat_lon(x):
    if isinstance(x, str) and x.startswith('[{'):
        try:
            coords = ast.literal_eval(x)
            return coords[0]['lat'], coords[0]['lon']
        except:
            return np.nan, np.nan
    else:
        return np.nan, np.nan

df[['lat', 'lon']] = df['provinsi'].apply(lambda x:
pd.Series(extract_lat_lon(x)))

######### konversi kolom published ke kategori jam #########
import pandas as pd

# ubah ke datetime
df['published'] = pd.to_datetime(df['published'], errors='coerce',
utc=True)

# buat kolom kategori jam
df['published_local'] = df['published'].dt.tz_convert('Asia/Jakarta')
```

```
def kategori_jam(hour):
    if 5 <= hour < 11:
        return 'pagi'
    elif 11 <= hour < 15:
        return 'siang'
    elif 15 <= hour < 18:
        return 'sore'
    else:
        return 'malam'

df['jam_published'] =
df['published_local'].dt.hour.apply(kategori_jam)

######### konversi published_to_es ke datetime dan boolean #########
def split_es(x):
    val = str(x)
    date = pd.to_datetime(x, errors='coerce', utc=True)
    if pd.notna(date):
        return pd.Series({'published_to_es_flag': True,
'published_to_es_date': date})
    elif val.lower() == 'true':
        return pd.Series({'published_to_es_flag': True,
'published_to_es_date': np.nan})
    else:
        return pd.Series({'published_to_es_flag': False,
'published_to_es_date': np.nan})

df[['published_to_es_flag', 'published_to_es_date']] =
df['published_to_es'].apply(split_es)

######### konversi is_updated ke datetime dan boolean #########
def split_is_updated(x):
    val = str(x)
    date = pd.to_datetime(x, errors='coerce', utc=True)
    if pd.notna(date):
        return pd.Series({'is_updated_flag': True, 'is_updated_date':
date})
    elif val.lower() == 'true':
        return pd.Series({'is_updated_flag': True, 'is_updated_date':
np.nan})
    else:
```

```
        return pd.Series({'is_updated_flag': False, 'is_updated_date':
np.nan})

df[['is_updated_flag', 'is_updated_date']] =
df['is_updated'].apply(split_is_updated)

# hapus timezone dari semua kolom datetime
for col in df.select_dtypes(['datetimetz']).columns:
    df[col] = df[col].dt.tz_localize(None)

######### hapus kolom yang tidak relevan #########
drop_columns = [
    'favicon', 'hashtitle', 'link', 'media_url', 'alchemy',
'highlight',
    'hashtag', 'subs_name', 'phrases', 'keywords', 'published_local',
    'fulltext', 'summary', 'desc', 'username', 'authors', 'ners',
    'summarization', 'location_aho', 'published_to_es', 'is_updated',
    'known_by_police', 'published_all', 'jenis', 'geo'
]
df.drop(columns=drop_columns, inplace=True)

######### unduh file #########
from google.colab import files
df.to_excel('news_cleaned.xlsx', index=False)
files.download('news_cleaned.xlsx')
```

Setelah melalui proses pembersihan dan transformasi, dataset kini bebas dari atribut tidak relevan dan missing value yang dapat mengganggu visualisasi. Data telah diformat dan disesuaikan dengan kebutuhan 20 use case yang telah dipetakan sebelumnya. File akhir dengan nama news_cleaned.xlsx siap digunakan dalam tahap berikutnya, yaitu Modeling dan pembuatan dashboard di Tableau.

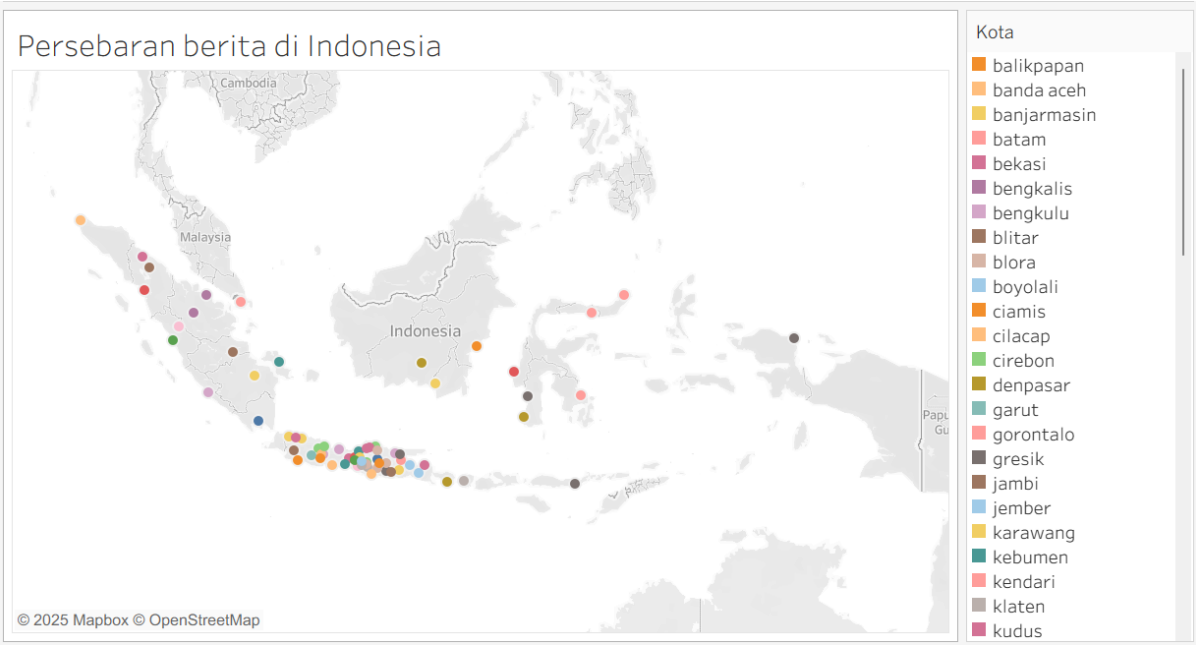
D. Modelling



link tableau:
https://public.tableau.com/views/PSD_17657771014070/Dashboard2?:language=en-US&publish=yes&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link

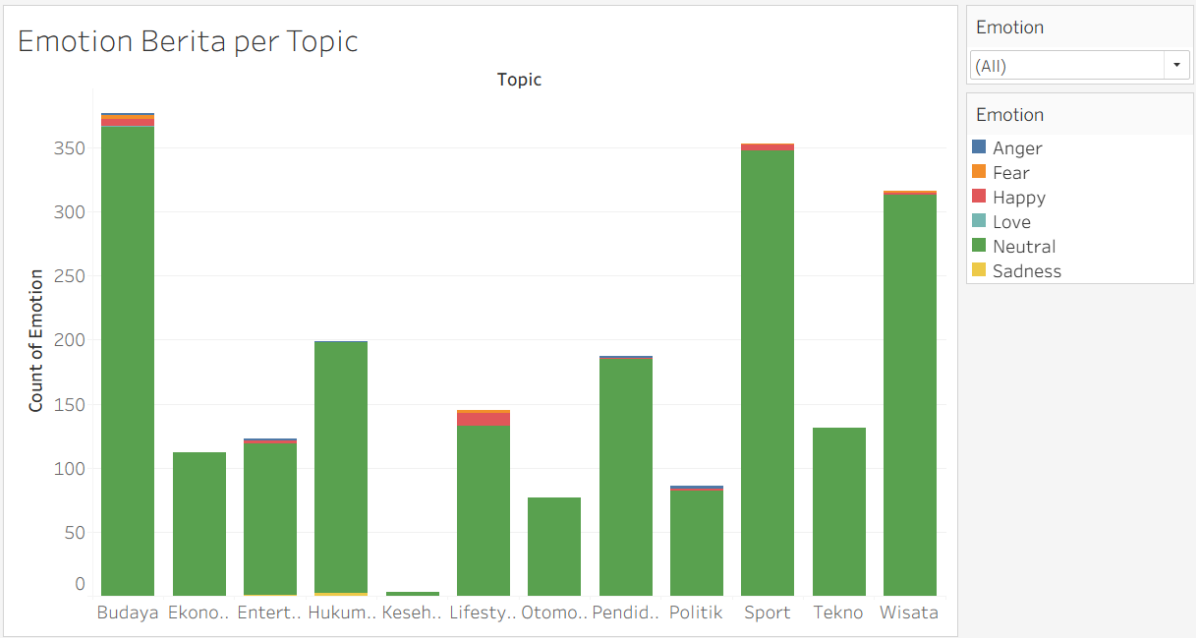
Dashboard yang dikembangkan dalam tahap ini menyajikan visualisasi komprehensif terhadap data berita daring yang dipublikasikan di wilayah Yogyakarta pada bulan Oktober 2025. Visualisasi ini dibangun berdasarkan hasil pemetaan use case pada tahap Business Understanding dan Data Understanding, dengan tujuan untuk mengidentifikasi pola, tren, serta insight penting dari aktivitas pemberitaan media daring. Dashboard terdiri atas beberapa komponen visual utama, antara lain:

1. Persebaran berita di Indonesia



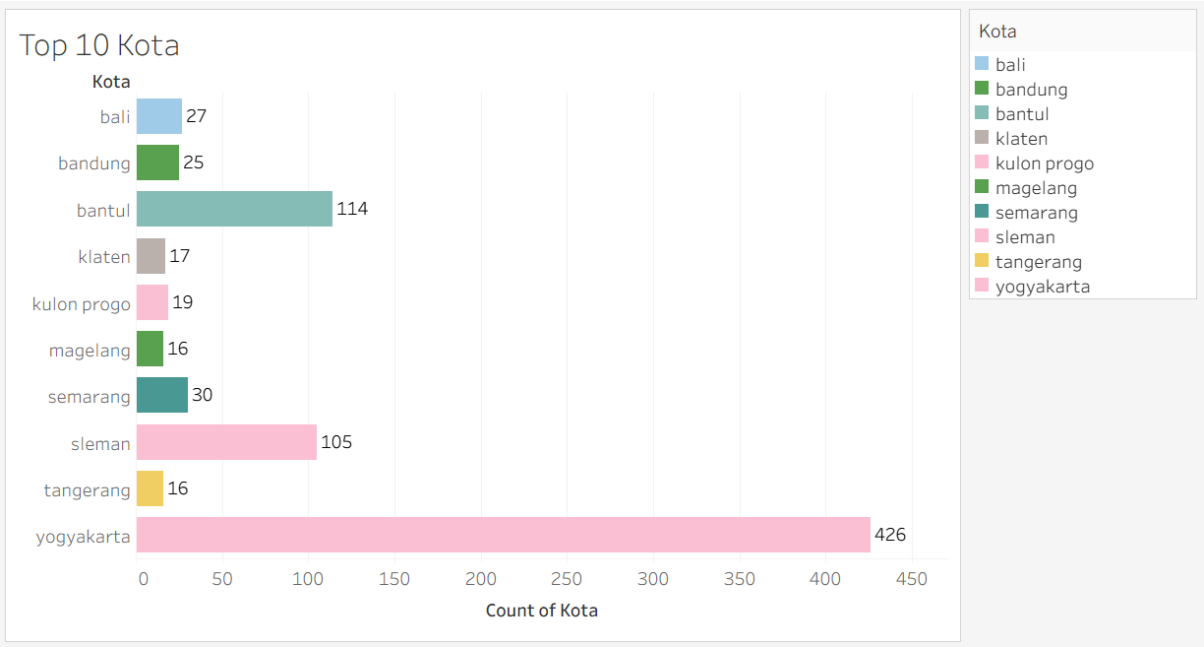
Peta interaktif ini menampilkan sebaran lokasi pemberitaan berdasarkan kota dan provinsi. Pulau Jawa menjadi wilayah dengan konsentrasi berita tertinggi dibandingkan dengan pulau lainnya.

2. Emotion Berita per Topic



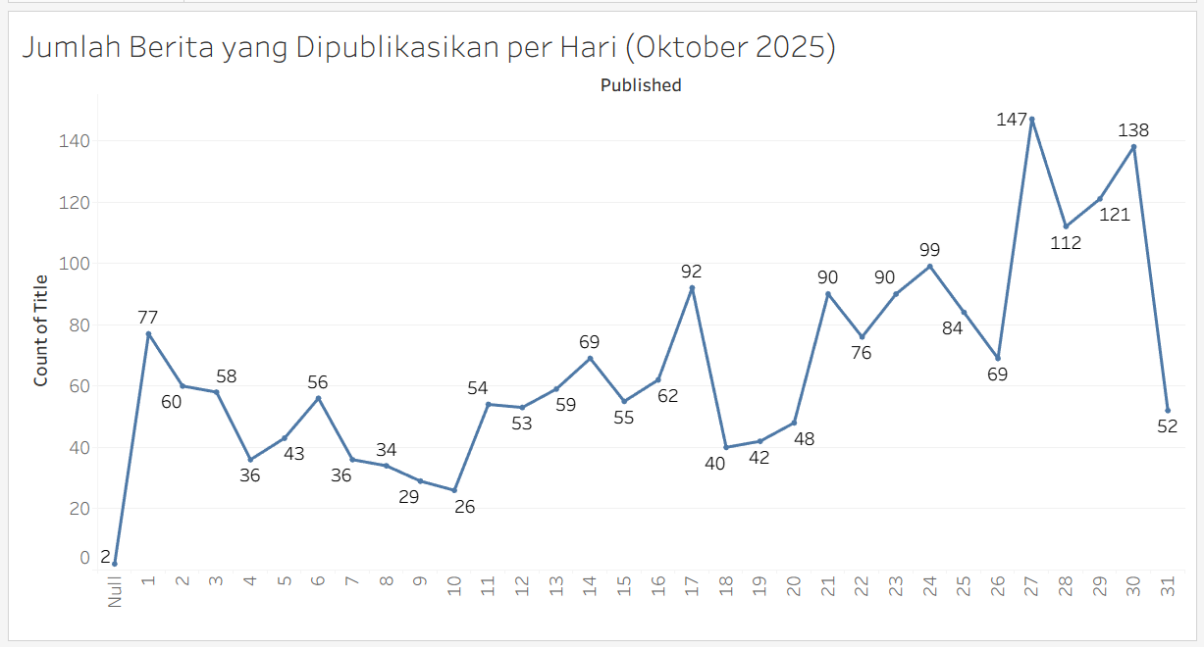
Stacked bar chart ini menampilkan distribusi emosi (marah, sedih, senang, takut, netral) berdasarkan topik berita. Semua topik didominasi oleh emosi netral.

3. Jumlah Berita di Tiap Kota



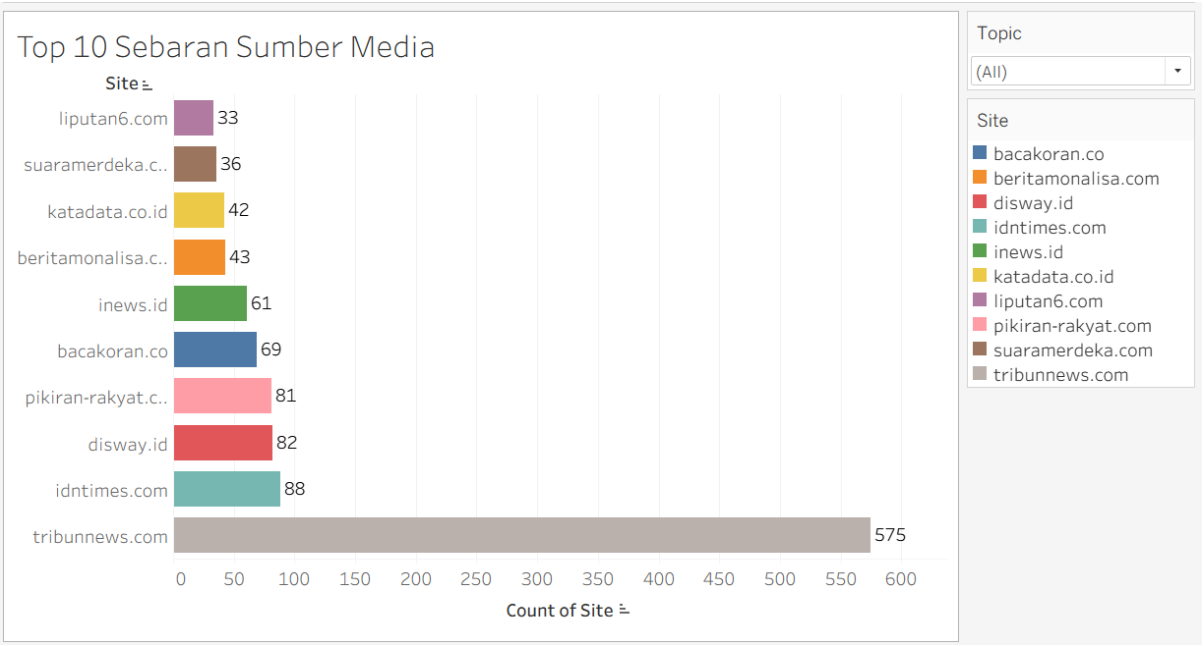
Bar chart ini memperlihatkan 10 kota dengan jumlah berita terbanyak. Yogyakarta dan sekitarnya mendominasi, diikuti oleh kota-kota besar seperti Semarang dan Bali.

4. Tren Berita yang Dipublikasikan per Hari



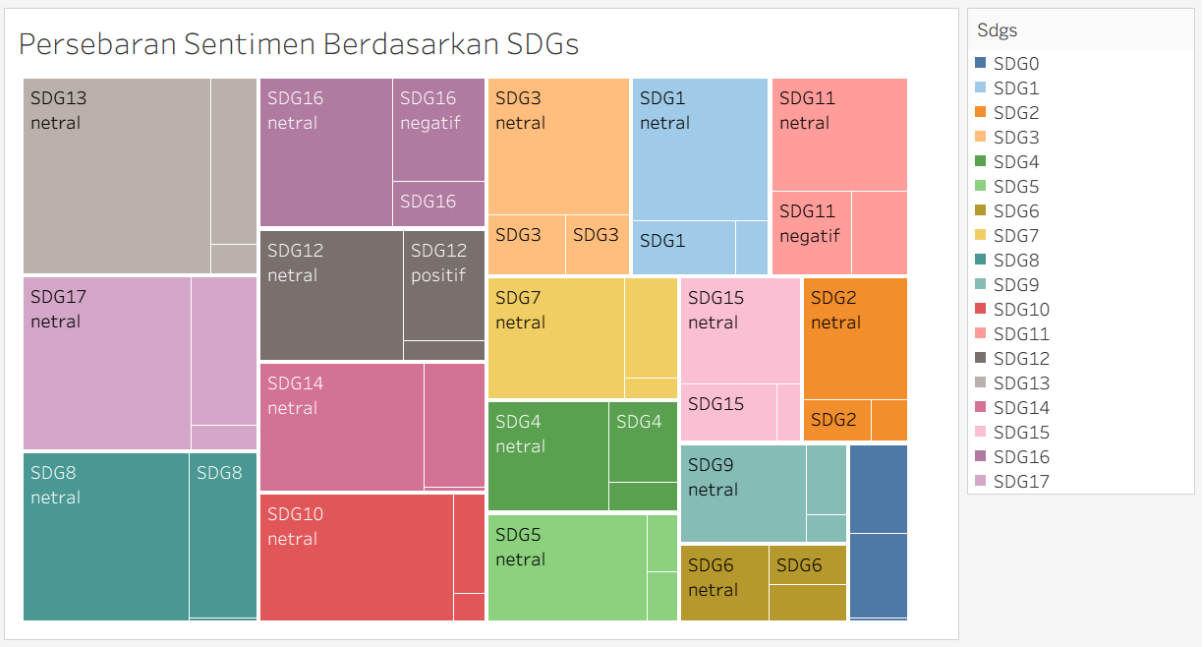
Grafik garis ini memperlihatkan jumlah berita yang diterbitkan setiap hari selama periode Oktober 2025. Terlihat adanya lonjakan pada beberapa tanggal tertentu yang berkaitan dengan isu besar di wilayah Yogyakarta.

5. Sebaran Sumber Media



Visualisasi batang ini menampilkan 10 media daring teratas yang paling aktif mempublikasikan berita. Media tribunnews.com paling mendominasi jumlah publikasi.

6. Persebaran Sentimen Berdasarkan SDGs



Treemap ini menggambarkan persebaran sentimen berita berdasarkan topik SDGs. Hasil analisis menunjukkan bahwa semua topik SDGs lebih sering bernada netral.

7. Analisis Waktu Publikasi

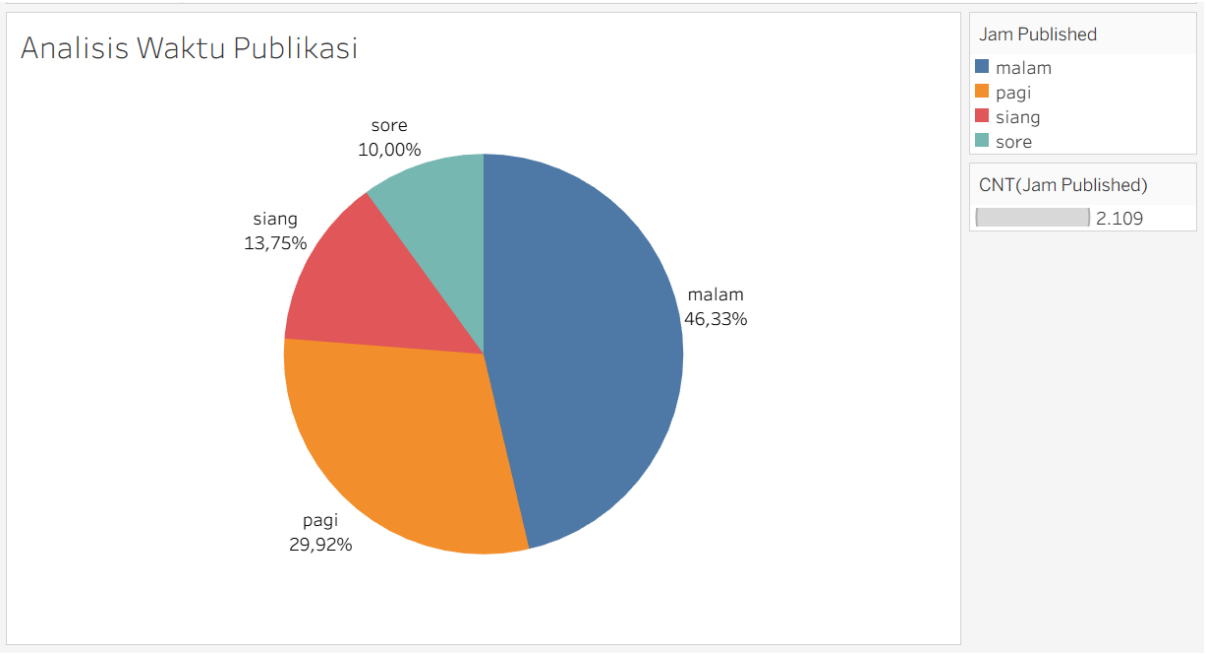
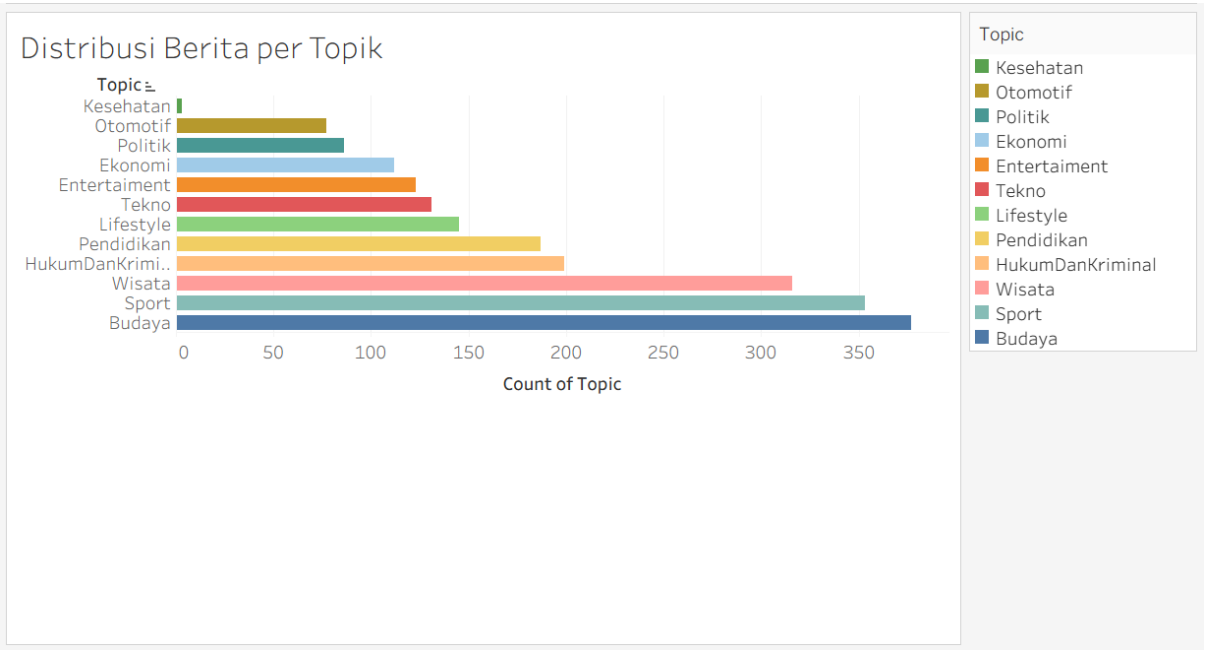


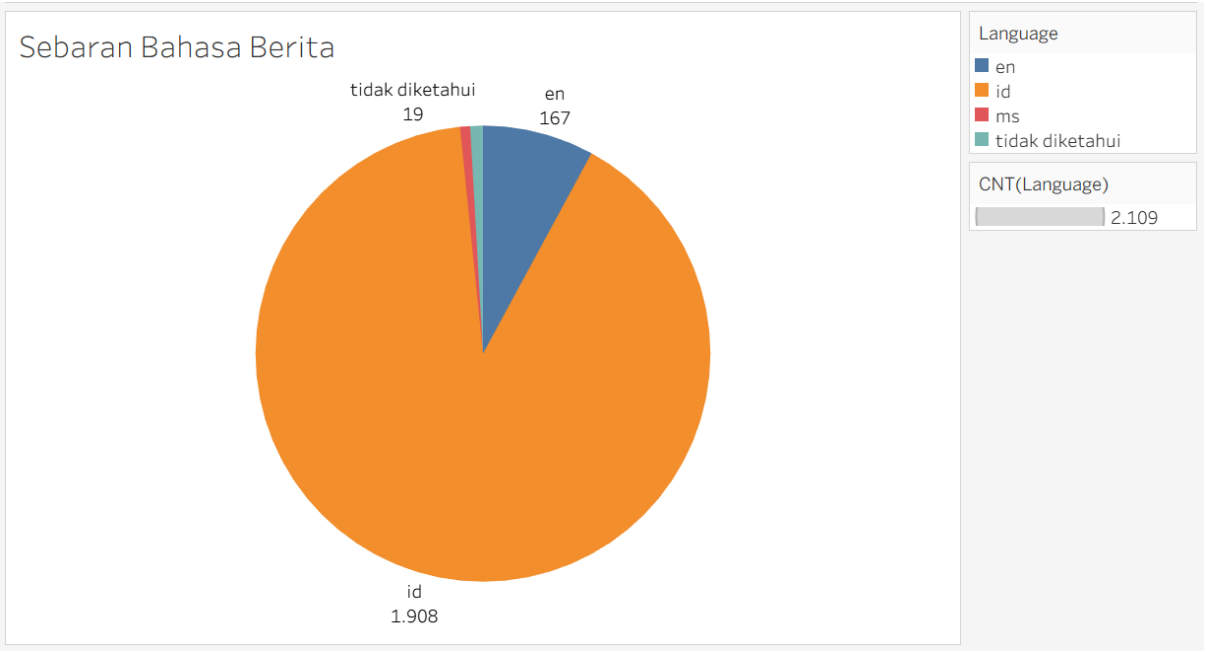
Diagram pie ini menunjukkan distribusi waktu publikasi berita dalam sehari (pagi, siang, sore, malam). Mayoritas berita dipublikasikan pada rentang waktu malam hari, menggambarkan pola aktivitas redaksi media daring.

8. Distribusi Berita per Topik



Visualisasi ini menampilkan sebaran jumlah berita berdasarkan atribut topic. Hasilnya menunjukkan bahwa beberapa topik seperti budaya, sport, dan wisata memiliki jumlah pemberitaan yang paling tinggi.

9. Sebaran Bahasa Berita



Visualisasi ini menggambarkan proporsi berita berdasarkan bahasa yang digunakan, seperti Bahasa Indonesia dan Bahasa Inggris. Sebagian besar berita ditulis dalam Bahasa Indonesia, mencerminkan konteks lokal wilayah Indonesia.

10. Sentimen Berita

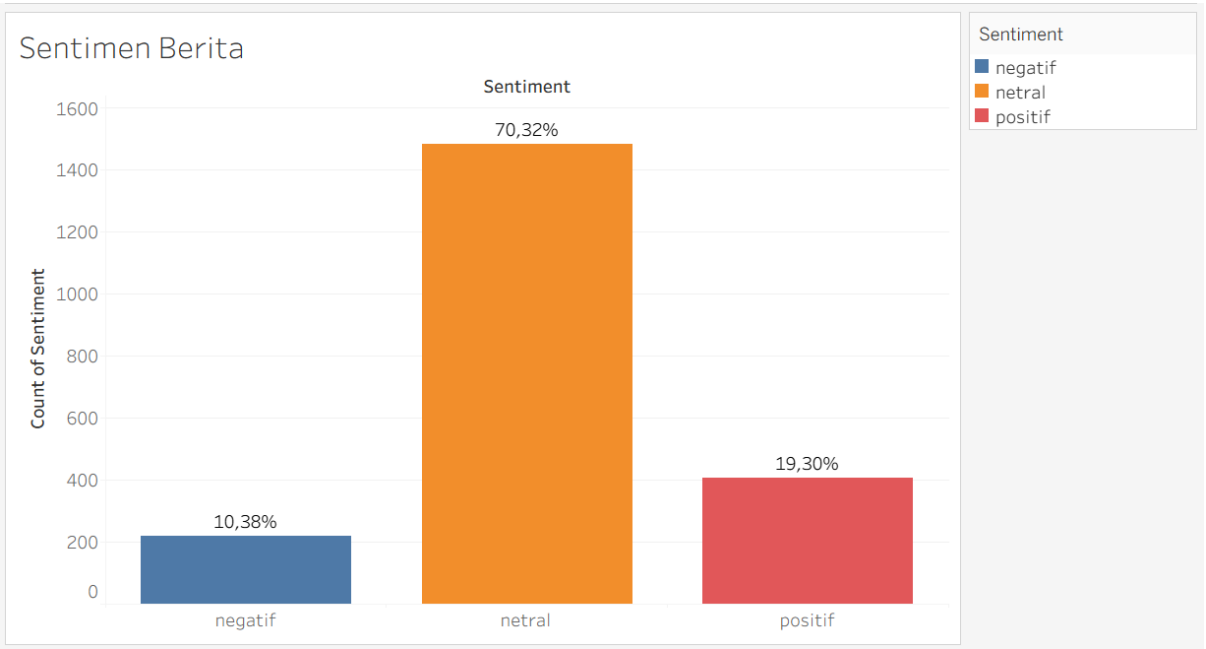


Diagram ini menunjukkan distribusi sentimen berita yang dikelompokkan menjadi positif, netral, dan negatif. Hasilnya memperlihatkan bahwa berita dengan sentimen netral mendominasi, diikuti oleh sentimen positif, sementara sentimen negatif relatif lebih sedikit.

11. Emosi Dominan dalam Berita

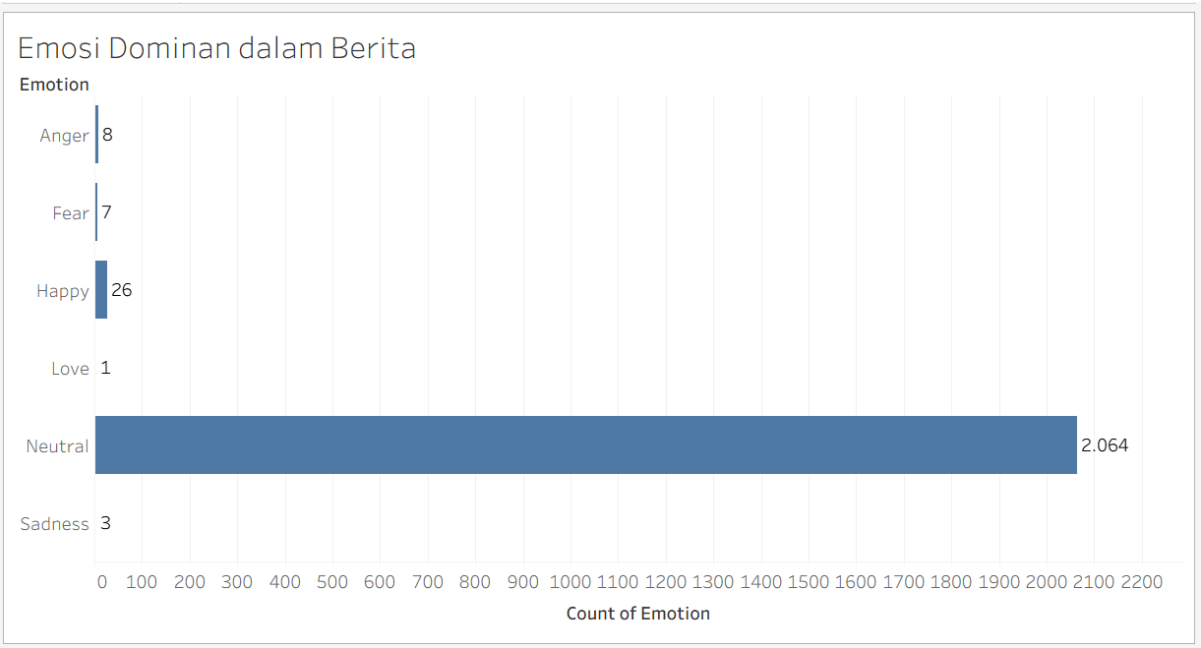
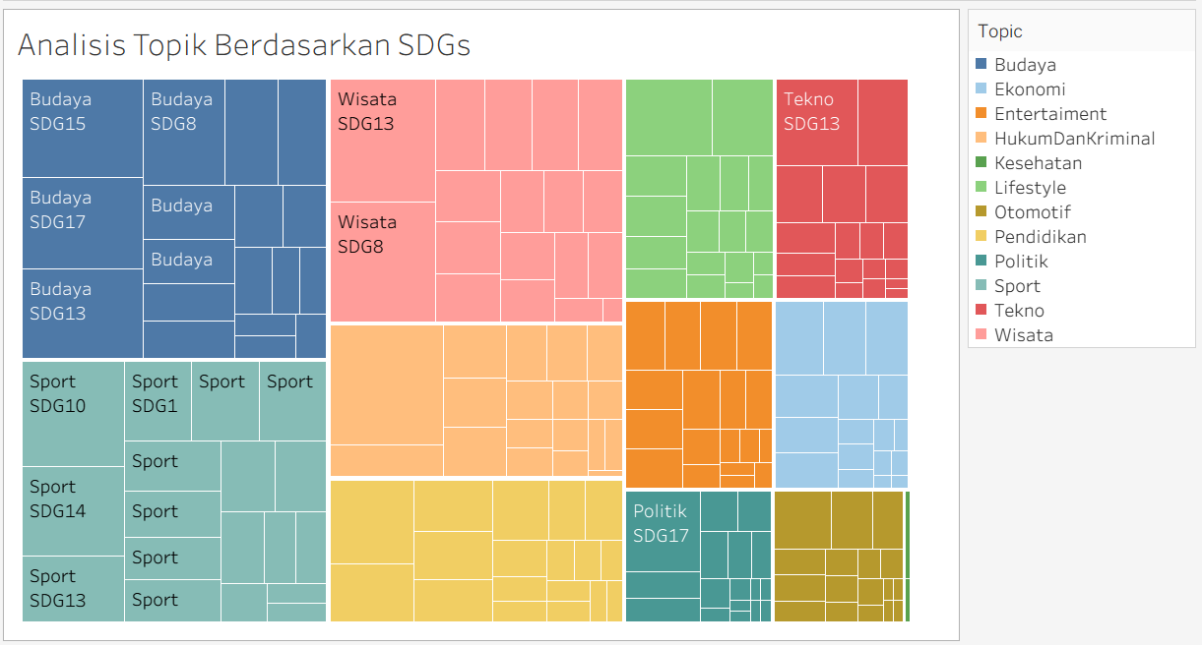


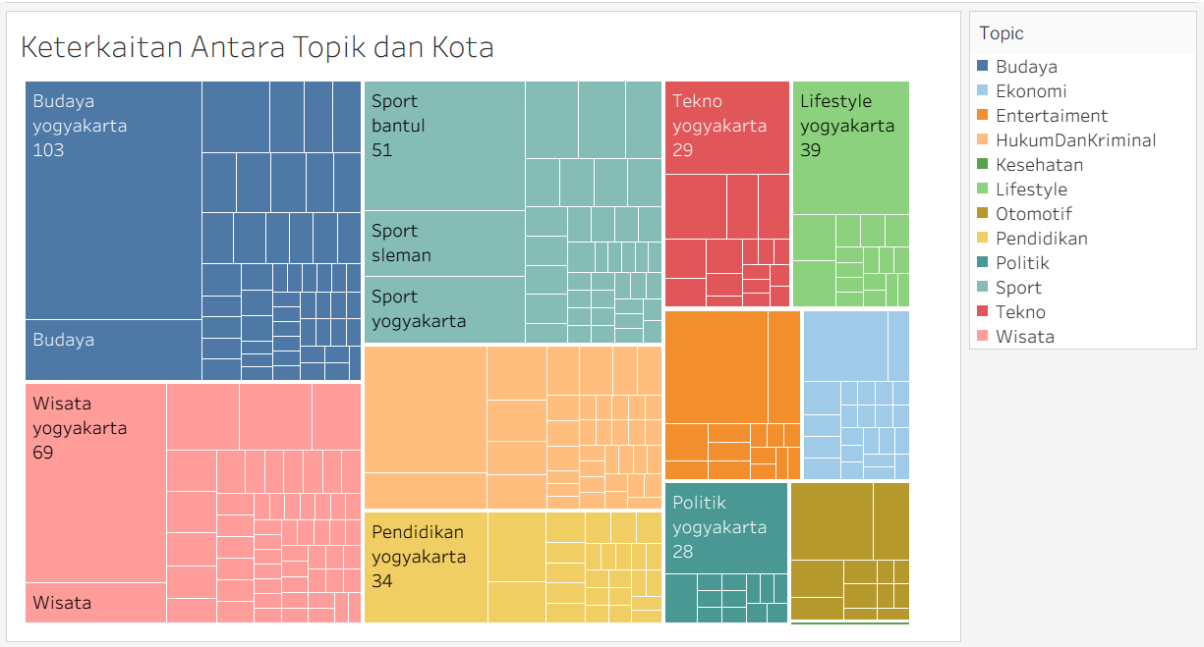
Diagram batang ini menampilkan emosi utama yang muncul dalam berita, seperti marah, takut, sedih, senang, dan netral. Dari hasil visualisasi, dapat diamati bahwa emosi netral paling sering muncul.

12. Analisis Topik Berdasarkan SDGs



Treemap ini memperlihatkan keterkaitan antara berita dengan kategori SDGs (Sustainable Development Goals). Dari visualisasi ini, dapat diamati topik SDGs apa yang paling sering dibahas dalam berita.

13. Keterkaitan Antara Topik dan Kota



Treemap ini menunjukkan kota yang paling sering muncul dalam pemberitaan untuk setiap topik berita. Misalnya, topik budaya dan wisata banyak muncul di Yogyakarta, sedangkan topik sport lebih tersebar di Bantul.

14. Sinkronisasi data

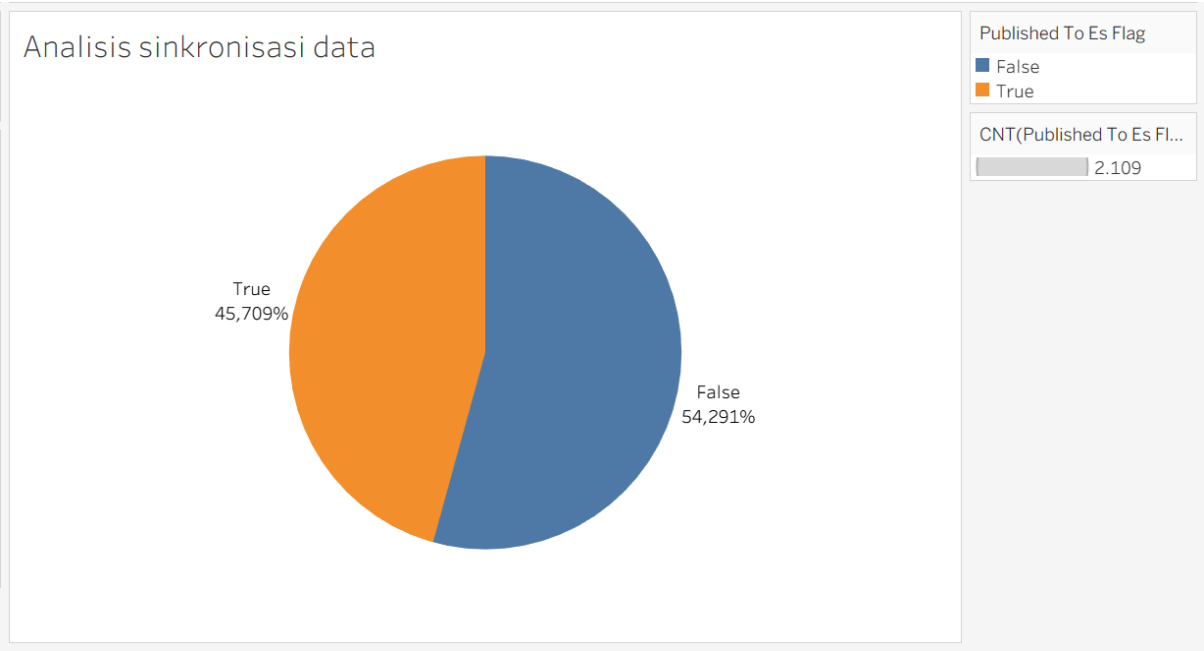
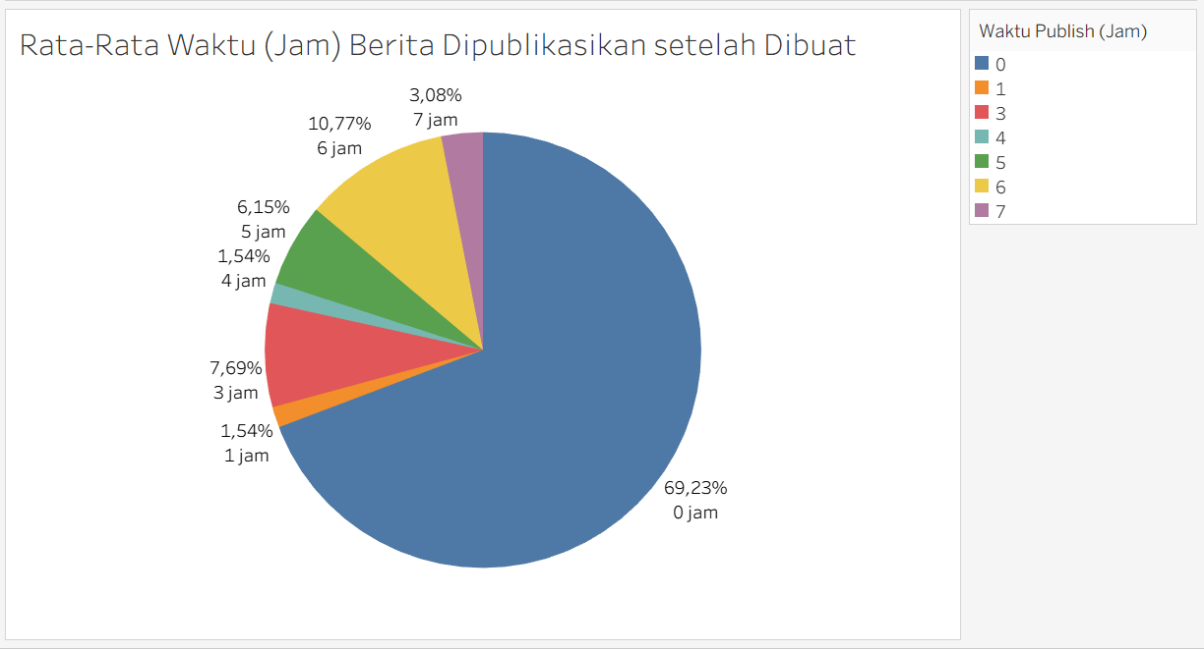


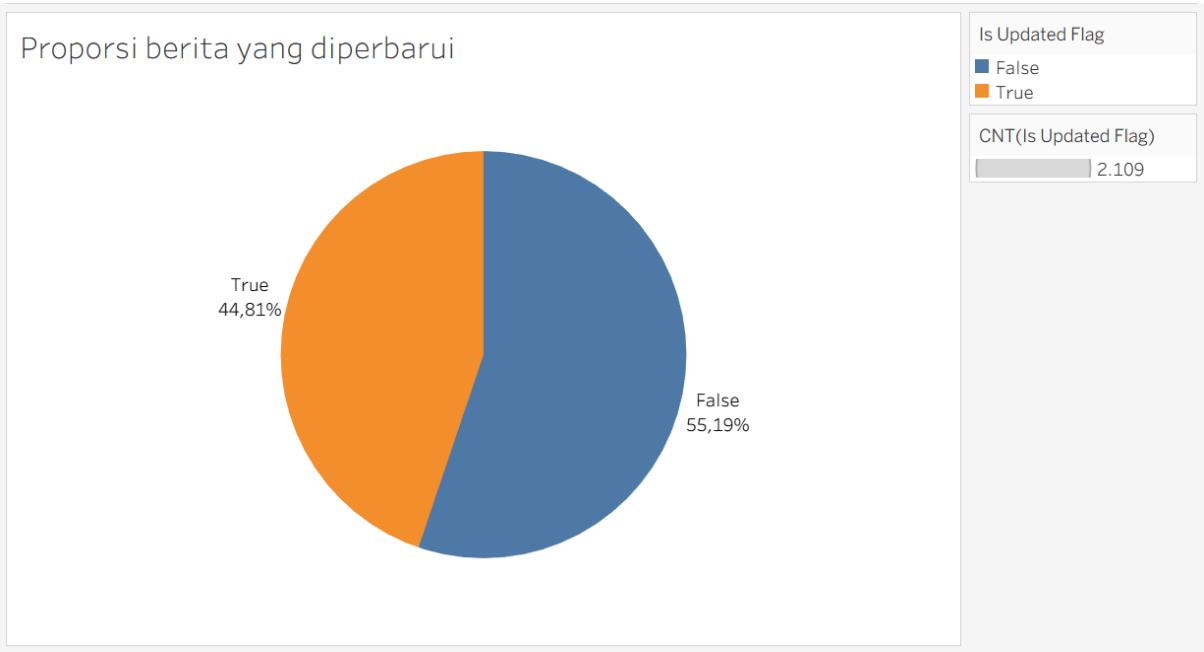
Diagram pie ini menampilkan perbandingan antara berita yang sudah dan belum dipublikasikan ke sistem Elasticsearch (published_to_es). Sebagian besar berita belum tersinkronisasi, menandakan proses publikasi digital belum berjalan baik.

15. Waktu Berita Dipublikasikan



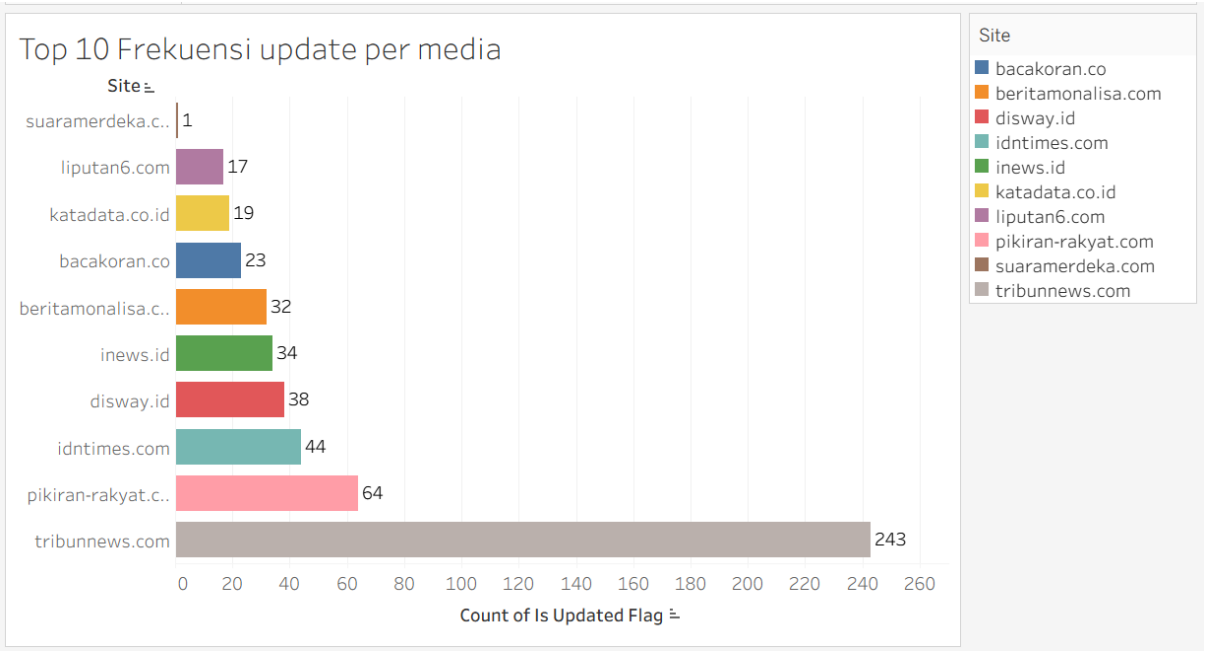
Visualisasi ini menunjukkan selisih waktu antara pembuatan (created) dan publikasi (published) berita. Sebagian besar berita diterbitkan dalam waktu yang relatif cepat setelah dibuat (tidak sampai 1 jam), menunjukkan efisiensi redaksi.

16. Berita yang Diperbarui



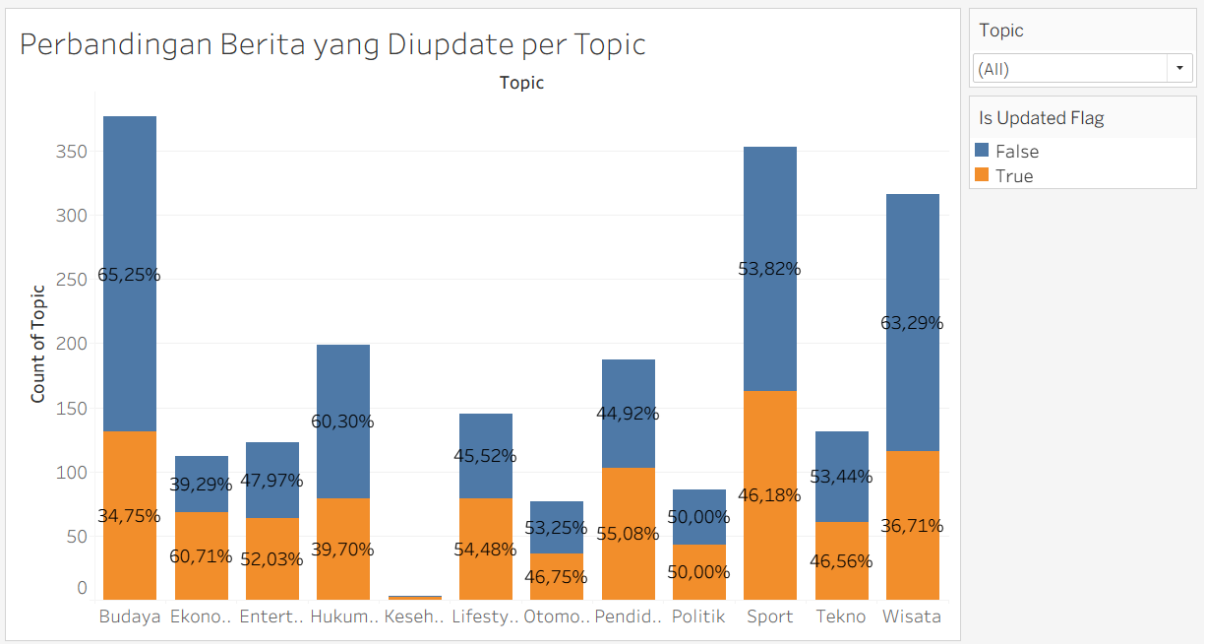
Pie chart ini menampilkan proporsi berita yang diperbarui (is_updated) dibandingkan dengan yang tidak diperbarui. Hasilnya menunjukkan sebagian kecil berita mengalami pembaruan setelah publikasi awal.

17. Frekuensi update per media



Bar chart ini menampilkan 10 media dengan jumlah pembaruan berita terbanyak. Media tribunnews.com tampak lebih aktif memperbarui berita mereka.

18. Perbandingan Berita yang Di-update per Topic



Stacked bar chart ini menampilkan perbandingan antara berita yang diperbarui dan yang tidak diperbarui untuk setiap topik. Beberapa topik seperti budaya, hukumdankriminal, dan wisata memiliki proporsi update yang lebih tinggi.

19. Citra polisi di media

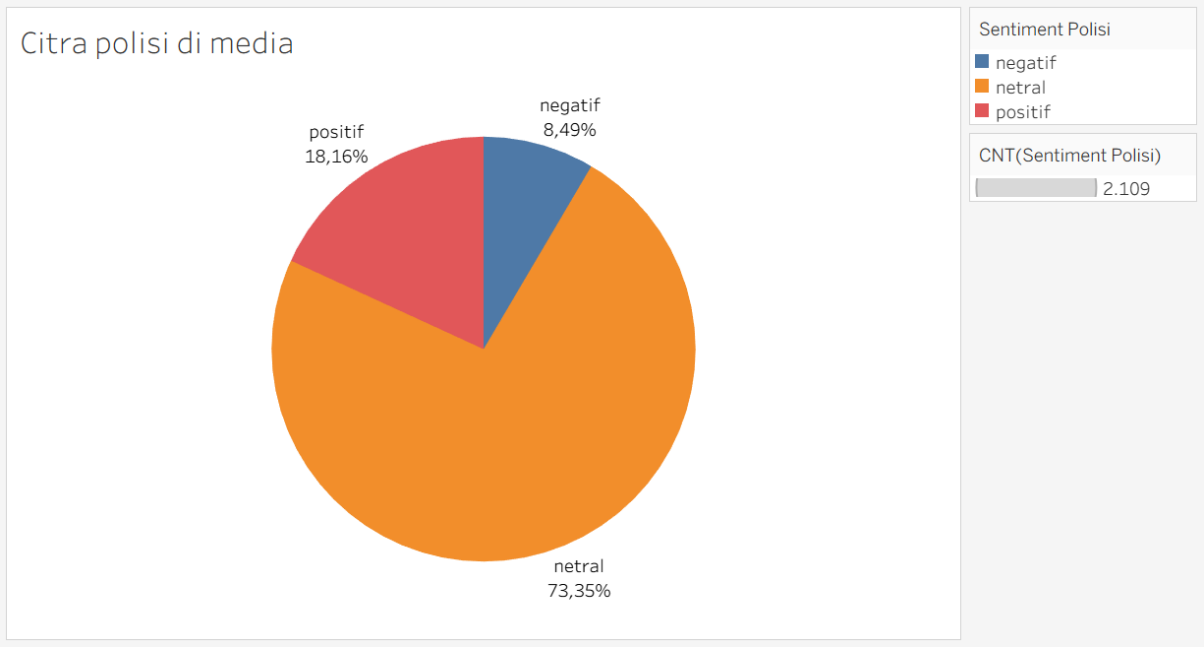
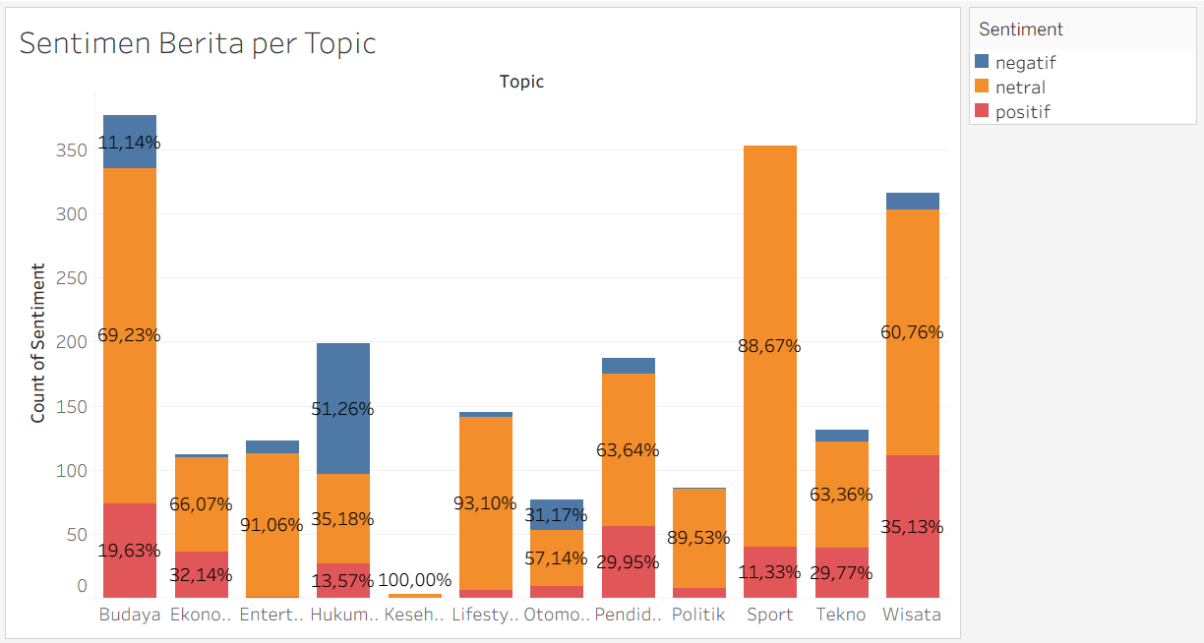


Diagram batang ini menampilkan distribusi sentimen polisi (sentiment_polisi), yaitu proporsi berita bernada positif, netral, dan negatif terhadap kepolisian. Hasilnya menunjukkan dominasi berita bernada netral, diikuti dengan berita bernada negatif.

20. Sentimen Berita per Topic



Stacked bar chart ini memperlihatkan distribusi sentimen (positif, netral, negatif) pada masing-masing topik berita. Topik hukumdankriminal cenderung lebih banyak mengandung sentimen negatif, sedangkan topik lifestyle dan sport didominasi sentimen netral. Topik yang mengandung proporsi sentimen positif terbanyak yaitu wisata.

Dashboard ini dirancang untuk memberikan gambaran menyeluruh mengenai pola dan dinamika pemberitaan media daring di wilayah Yogyakarta. Dengan memanfaatkan berbagai visualisasi interaktif, pengguna dapat mengeksplorasi informasi penting secara efisien dan memperoleh insight yang relevan terkait karakteristik berita, persepsi media terhadap

kepolisian, serta aktivitas publikasi dan pembaruan berita. Dashboard ini mendukung proses pengambilan keputusan berbasis data, baik untuk pemantauan citra institusi di media, analisis tren isu publik, maupun evaluasi aktivitas publikasi media daring selama periode analisis.