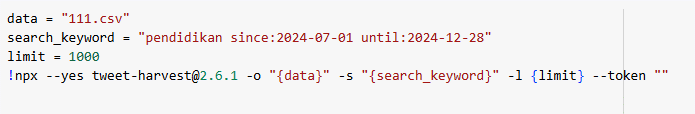
**LAPORAN TOPIC MODELLING LDA**

**Proses Pembuatan**

Untuk mengimplementasikan topic modelling LDA, pertama-tama, dilakukan scraping data 1000 tweet di X tentang pendidikan dalam rentang waktu Juni 2024 – Desember 2024 dengan menggunakan tweet harvest. Berikut adalah coding yang digunakan.







Namun karena per putaran scrape data hanya diperoleh 100 data per hari. Maka dilakukan iterasi beberapa kali hingga terkumpul 1000 data.

Selanjutnya, dilakukan preprocessing data dengan membersihkan data dari kata-kata kurang penting yang sering muncul seperti “ga, tdk, ya, gua”. Kata-kata juga dibuat ke bentuk paling dasarnya dengan menghapus imbuhan.

import pandas as pd

import re

from nltk.corpus import stopwords

from nltk.tokenize import word\_tokenize

from nltk.stem import WordNetLemmatizer

from sklearn.feature\_extraction.text import CountVectorizer

import gensim

from gensim import corpora

from gensim.models import LdaModel

from sklearn.feature\_extraction.text import TfidfVectorizer

from sklearn.decomposition import LatentDirichletAllocation

import nltk

nltk.download('punkt')

nltk.download('stopwords')

nltk.download('wordnet')

stop\_words = stopwords.words('indonesian')

custom\_stop\_words = ['yang', 'ga', 'https', 'co', 'yg','t', 'ya', 'nya', 'gue','gw', 'gua', 'aja', 'kalo', 'ga', 'pemerintahnaikkangajiguru', 'amp','sih']

stop\_words.extend(custom\_stop\_words)

wtk = nltk.tokenize.RegexpTokenizer(r'\w+')

wnl = WordNetLemmatizer()

from google.colab import files

uploaded = files.upload()

file\_name = list(uploaded.keys())[0]

app\_reviews\_df = pd.read\_csv(file\_name)

print(app\_reviews\_df.head())

reviews = app\_reviews\_df['description']

reviews.dropna(inplace=True)

reviews.reset\_index(drop=True, inplace=True)

def preprocess\_text(text):

    tokens = wtk.tokenize(text.lower())

    filtered\_tokens = [word for word in tokens if word not in stop\_words]

    lemmatized\_tokens = [wnl.lemmatize(word) for word in filtered\_tokens]

    return " ".join(lemmatized\_tokens)

processed\_reviews = reviews.apply(preprocess\_text)

print(processed\_reviews.head())

Kemudian, program menjalankan algoritma LDA untuk menghasilkan 5 buah topik yang paling sering tentang keseluruhan tweet.

from gensim import corpora

from gensim.models import LdaModel

processed\_reviews\_list = processed\_reviews.tolist()

dictionary = corpora.Dictionary([review.split() for review in processed\_reviews\_list])

corpus = [dictionary.doc2bow(review.split()) for review in processed\_reviews\_list]

num\_topics = 5

lda\_model = LdaModel(corpus=corpus, id2word=dictionary, num\_topics=num\_topics, passes=10, random\_state=42)

topics = lda\_model.show\_topics(num\_topics=num\_topics, num\_words=7, formatted=False)

topics = sorted(topics, key=lambda x: x[0])  # Mengurutkan topik berdasarkan indeks

for i, (\_, word\_list) in enumerate(topics):

    topic\_words = ', '.join([word for word, \_ in word\_list])

    print(f"Topic {i}: {topic\_words}")

**Hasil dan Pembahasan**

Berikut adalah topik yang dihasilkan dari ke-1000 tweet.

Topic 0: pendidikan, guru, indonesia, kenaikan, kesehatan, berkualitas, gaji

Topic 1: pendidikan, wayang, indonesia, budaya, batik, orang, pelatihan

Topic 2: pendidikan, anak, makanbergizibangunnegeri, guru, orang, sekolah, pemerintah

Topic 3: pendidikan, orang, kerja, dunia, kesehatan, sma, sekolah

Topic 4: pendidikan, orang, banget, susah, butuh, menteri, tu

Berikut adalah analisis untuk masing-masing topik berdasarkan kata kunci.

*Topic 0: pendidikan, guru, indonesia, kenaikan, kesehatan, berkualitas, gaji*

Fokus utama dari topik ini adalah isu kesejahteraan guru di Indonesia, termasuk gaji, kesehatan, dan kualitas pendidikan. Berarti, ada perhatian terhadap kesejahteraan guru dan bagaimana hal ini memengaruhi kualitas pendidikan di Indonesia. Selain itu, kata "kenaikan" kemungkinan mengacu pada harapan atau diskusi terkait kenaikan gaji guru.

*Topic 1: pendidikan, wayang, indonesia, budaya, batik, orang, pelatihan*

Fokus utama dari topik ini adalah pendidikan berbasis budaya dan pelestarian budaya lokal Indonesia. Selain itu, kata "wayang" dan "batik" menyoroti pentingnya memperkenalkan budaya lokal dalam pendidikan. Berikutnya, pelatihan mungkin berhubungan dengan upaya memperkenalkan budaya melalui pendidikan formal atau non-formal.

*Topic 2: pendidikan, anak, makanbergizibangunnegeri, guru, orang, sekolah, pemerintah*

Fokus utama dari topik ini adalah pendidikan anak, nutrisi, dan peran pemerintah dalam mendukungnya. Isu nutrisi ("makanbergizibangunnegeri") menunjukkan perhatian terhadap pentingnya gizi dalam mendukung keberhasilan pendidikan anak. Ada pula penekanan pada guru dan sekolah sebagai faktor penting dalam membangun negeri.

*Topic 3: pendidikan, orang, kerja, dunia, kesehatan, sma, sekolah*

Fokus utama dari topik ini adalah pendidikan yang relevan dengan dunia kerja dan kesehatan. Ada hubungan antara pendidikan (terutama di tingkat SMA) dan persiapan kerja. Kesehatan mungkin menjadi faktor yang memengaruhi kesiapan siswa dalam menghadapi dunia kerja.

*Topic 4: pendidikan, orang, banget, susah, butuh, menteri, tu*

Fokus utama dari topik ini adalah keluhan atau tantangan dalam sistem pendidikan. Kata seperti "susah", "butuh", dan "banget" menunjukkan adanya keluhan terkait akses atau kualitas pendidikan. "Menteri" mungkin mengacu pada ekspektasi terhadap kinerja Menteri Pendidikan dalam menyelesaikan masalah pendidikan.

**Laporan Visualisasi: Rata-rata Panjang Tweet per Bulan**

**Proses Lengkap**

Proyek ini mencakup serangkaian langkah untuk memvisualisasikan data rata-rata panjang tweet berdasarkan bulan, mulai dari pengumpulan data hingga deployment aplikasi visualisasi. Berikut adalah penjelasan lengkapnya:

**Langkah 1: Scraping Data**

**1. Load Data**

* Dataset awal dalam format CSV (all.csv) dimuat menggunakan **pandas**.
* Dilakukan pengecekan awal untuk memastikan kolom description tersedia.
* Baris dengan nilai kosong (NaN) pada kolom description dihapus untuk menghindari data kosong yang memengaruhi analisis.

**2. Hasil Awal**

* Dataset divisualisasikan dengan menampilkan 5 baris pertama sebagai validasi awal.

**Langkah 2: Pembersihan Data**

**1. Preprocessing Data**

* Dibuat fungsi clean\_text untuk pembersihan data yang meliputi:
  + Mengubah teks menjadi huruf kecil.
  + Menghapus URL, angka, dan karakter yang tidak relevan.
  + Melakukan tokenisasi menggunakan **WordTokenizer**.
  + Menghapus stop words untuk mengurangi kata-kata yang tidak memiliki informasi penting.
  + Melakukan lemmatization agar kata-kata dalam bentuk dasar.

**2. Aplikasi Pembersihan**

* Kolom description diproses menggunakan fungsi clean\_text.
* Hasilnya disimpan ke dalam kolom baru bernama cleaned\_description.

**3. Simpan Hasil**

* Dataset bersih disimpan ke file CSV baru dengan nama all\_cleaned.csv.

**Langkah 3: Insert Data ke Supabase**

**1. Persiapan Kolom**

* Pastikan dataset memiliki kolom berikut:

CopyEdit

conversation\_id\_str, created\_at, favorite\_count, description,

id\_str, image\_url, in\_reply\_to\_screen\_name, lang, location,

quote\_count, reply\_count, retweet\_count, tweet\_url, user\_id\_str

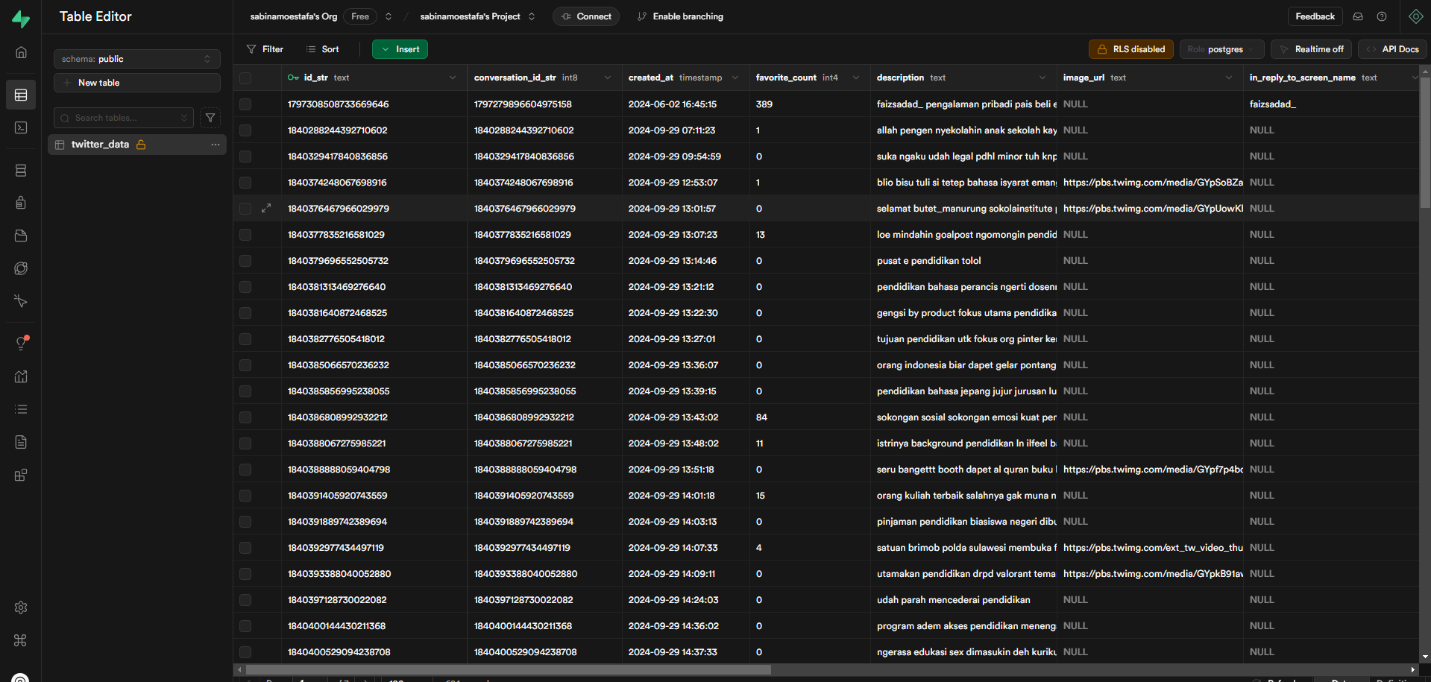
* Kolom created\_at diubah ke format string untuk memudahkan proses upload.

**2. Persiapan Data**

* Dataset dikonversi menjadi format dictionary menggunakan to\_dict(orient="records").
* Kolom description diisi dengan data dari cleaned\_description.

**3. Insert ke Supabase**

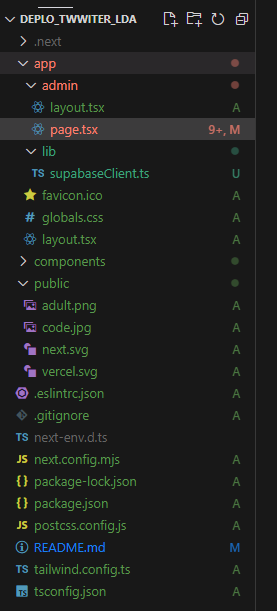
* Data dimasukkan ke tabel **twitter\_data** di database Supabase menggunakan API insert.
* Validasi dilakukan setelah setiap proses insert, dengan mencatat kesalahan bila terjadi.



**Langkah 4: Visualisasi di Next.js**

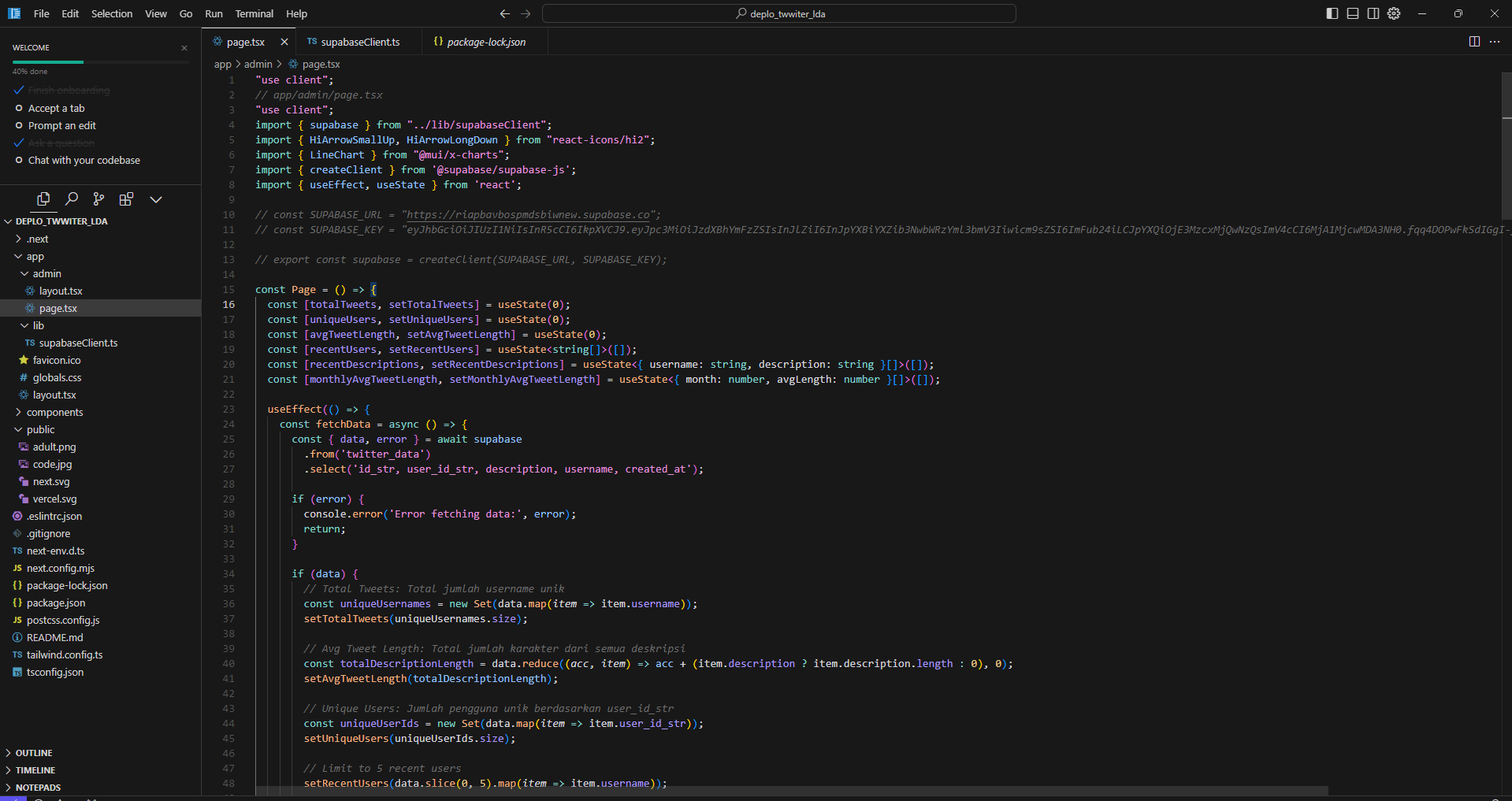
**1. Membuat Komponen Visualisasi**

* Dibuat aplikasi frontend menggunakan **Next.js**.
* Data diambil dari Supabase melalui API dan divisualisasikan menggunakan library **Chart.js** atau **Recharts**.



**2. Integrasi dengan Supabase**

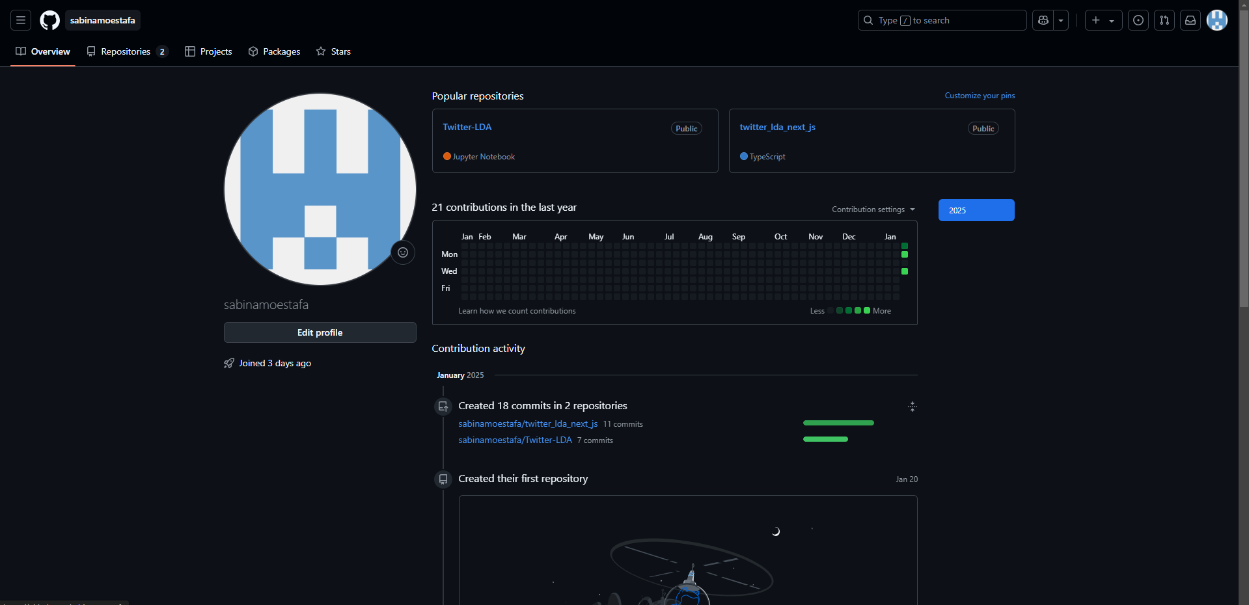
* Ditambahkan kredensial Supabase ke proyek Next.js untuk mengakses database.
* Data yang ditampilkan pada grafik dihitung sebagai rata-rata panjang tweet per bulan.



**Langkah 5: Deploy ke Vercel**

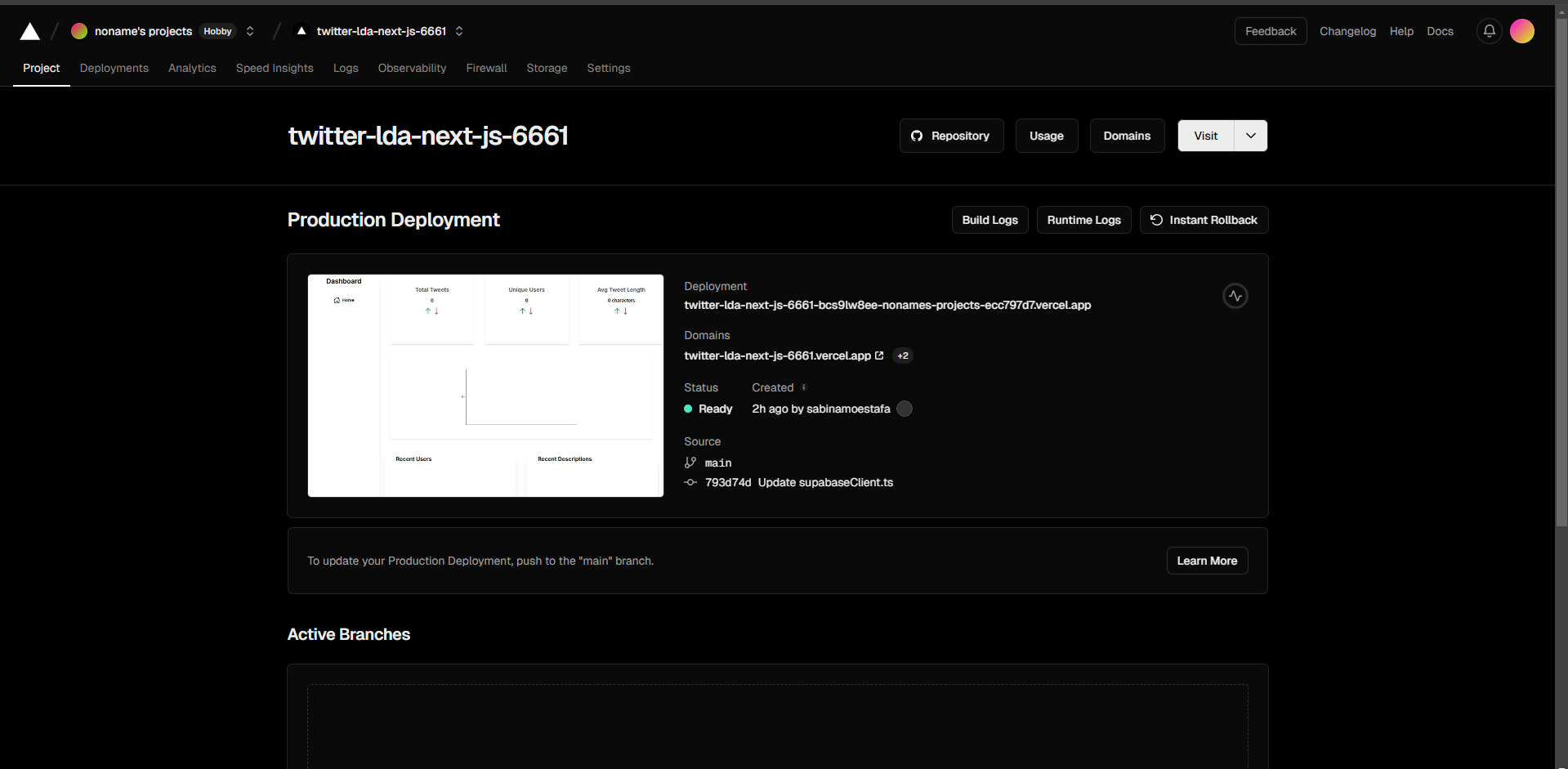
**1. Persiapan GitHub**

* Proyek Next.js diunggah ke GitHub, memastikan file konfigurasi .env memuat kredensial Supabase.



**2. Deploy ke Vercel**

* Login ke Vercel menggunakan akun GitHub.
* Proyek dideploy ke Vercel dengan URL publik untuk memudahkan akses.



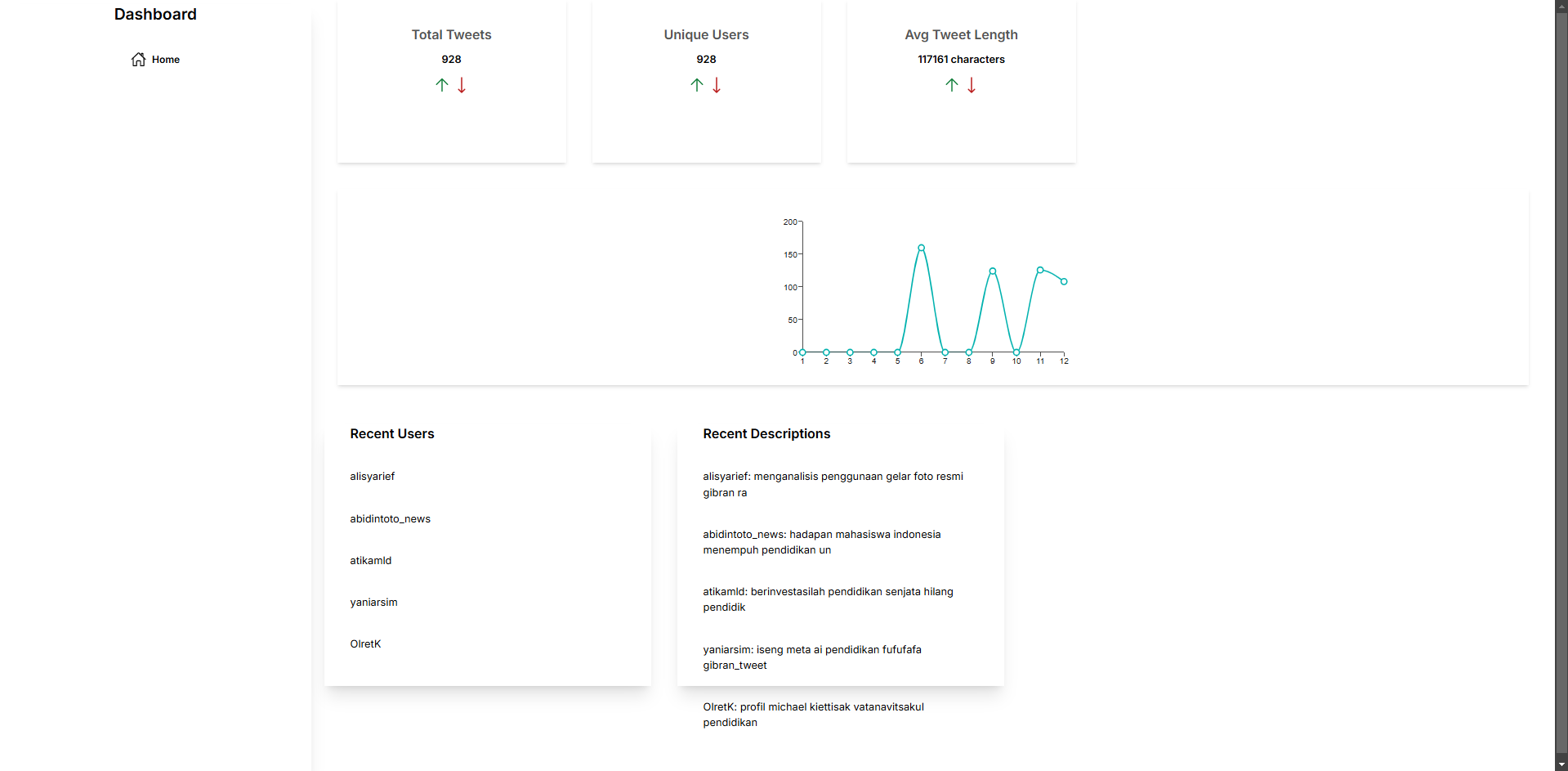
**Hasil Akhir**

**Visualisasi: Rata-rata Panjang Tweet per Bulan**

* Data rata-rata panjang tweet per bulan ditampilkan menggunakan grafik garis (**line chart**).
* Grafik menunjukkan pola perubahan panjang rata-rata tweet selama periode tertentu.

**Aplikasi yang Sudah Jadi**

* Aplikasi berbasis Next.js berhasil di-deploy di Vercel.
* Data dari Supabase divisualisasikan secara dinamis, memberikan wawasan yang mudah diakses melalui URL publik.



**Kesimpulan**

Semua topik berpusat pada pendidikan, namun dengan fokus berbeda: kesejahteraan guru, pelestarian budaya, nutrisi anak, relevansi pendidikan dengan dunia kerja, dan tantangan sistem pendidikan.

Link Project nya : <https://colab.research.google.com/drive/1OcRx2R5bYCbN6c9Zvd1-2ybYMtCQyapI?usp=sharing>