

Nama : Naufal Aqil

NPM : 2208107010043

## PROPOSAL UTS

### Analisis Sentiment Bersumber dari Video Youtube dengan Hashtag #indonesiagelap Menggunakan Library TextBlob dan Pre-train Model BERT

#### 1. Pendahuluan / Rationale

Perkembangan media sosial telah membuka peluang bagi masyarakat untuk mengekspresikan opini mereka terhadap berbagai isu sosial, politik, dan ekonomi. Salah satu platform yang sering digunakan untuk berdiskusi adalah YouTube, di mana komentar dalam video dapat memberikan wawasan terhadap persepsi publik.

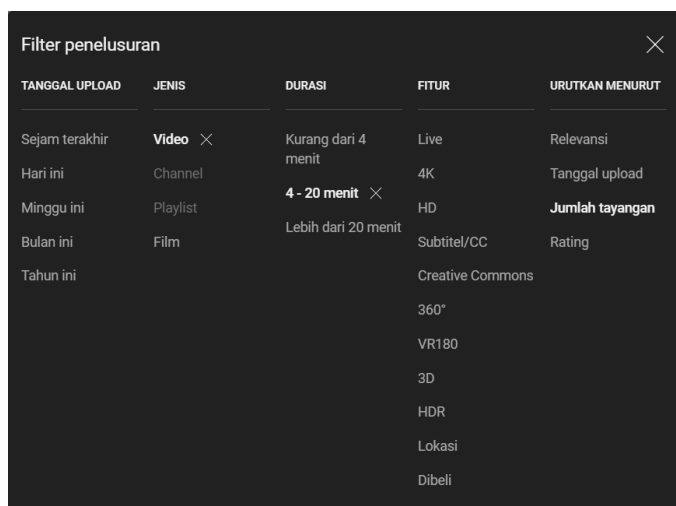
Hashtag **#indonesiagelap** menjadi salah satu topik yang menarik perhatian karena berisi diskusi mengenai berbagai permasalahan yang dihadapi Indonesia. Analisis sentimen terhadap komentar yang muncul pada video dengan hashtag ini dapat membantu memahami kecenderungan opini masyarakat, baik dalam kategori positif, negatif, maupun netral.

Dalam penelitian ini, dilakukan perbandingan dua pendekatan analisis sentimen, yaitu **TextBlob**, yang berbasis leksikon, dan model **BERT** (nlptown/bert-base-multilingual-uncased-sentiment), yang berbasis pembelajaran mesin. Dengan membandingkan kedua metode ini, diharapkan diperoleh pemahaman mengenai keakuratan dan efektivitas masing-masing model dalam menganalisis sentimen dalam bahasa Indonesia setelah diterjemahkan ke bahasa Inggris.

#### 2. Metodologi

##### a. Data

- Data dikumpulkan dengan metode **web scraping** dari komentar pada video YouTube yang memiliki **hashtag #indonesiagelap**.
- Pengambilan data dilakukan pada **10 video teratas** yang muncul dalam pencarian menggunakan hashtag tersebut.
- Scraping dilakukan menggunakan **API YouTube** untuk mengakses dan mengekstrak komentar.
- Filter yang digunakan untuk memilih video teratas sebagai berikut :



## b. Algoritma dan Teknik yang Digunakan

### 1. Preprocessing Data

- **Data Cleaning:** Menghapus karakter khusus, emoji, angka, dan tanda baca yang tidak diperlukan.
- **Normalisasi:** Mengonversi teks menjadi bentuk standar, seperti mengubah kata-kata tidak baku menjadi bentuk baku.
- **Stopword Removal:** Menghapus kata-kata umum yang tidak memberikan makna signifikan.
- **Tokenisasi:** Memecah teks menjadi kata-kata atau frasa kecil.
- **Stemming:** Mengubah kata menjadi bentuk dasarnya.
- **Penerjemahan ke Bahasa Inggris:** Menggunakan **API LLM Deepseek** untuk menerjemahkan teks hasil preprocessing agar sesuai dengan model analisis sentimen berbasis bahasa Inggris.

### 2. Analisis Sentimen

- **TextBlob:** Menggunakan pendekatan berbasis leksikon untuk menentukan polaritas komentar (positif, negatif, atau netral).
- **BERT (nlptown/bert-base-multilingual-uncased-sentiment):** Menggunakan model deep learning yang telah dilatih untuk analisis sentimen multibahasa.
- **Pelabelan data** dilakukan menggunakan kedua metode tersebut dan hasilnya dibandingkan.

### 3. Visualisasi Hasil

- Hasil analisis sentimen dari kedua metode akan divisualisasikan dalam bentuk grafik menggunakan **seaborn** dan **matplotlib** untuk mempermudah perbandingan tren sentimen.

## 3. Hasil yang Ingin Dicapai

- Memahami perbedaan hasil analisis sentimen antara metode berbasis leksikon (TextBlob) dan metode berbasis deep learning (BERT).
- Mengukur efektivitas dan akurasi kedua metode dalam menganalisis sentimen komentar berbahasa Indonesia setelah diterjemahkan ke bahasa Inggris.
- Menyediakan visualisasi data yang dapat digunakan untuk memahami opini publik terhadap topik dengan hashtag **#indonesiagelap**.

REVISI :

Menggunakan API LLM deepseek untuk melakukan translate data yang telah dilakukan preprocessing di banding dengan menggunakan library translate