

Analisis NLP:

Dataset diambil langsung dari twitter dengan menggunakan auth_token twitter, yaitu di tunjukkan pada code `twitter_auth_token` = `'0c10a9b8508cc10f8b4116cf4d4fd38f36f7b852'`

Kemudian, ditentukan review film apa yang akan diambil. Dalam case ini dipilih review terkait film siksa kubur yang baru saja ditayangkan di bioskop. Dataset yang didapatkan yaitu berjumlah 292. Keseluruhan data yang telah didapatkan dari twitter dikelompokkan terlebih dahulu berdasarkan sentimennya (positif, netral, negatif). Kemudian dilakukan training dan testing dengan pembagian 0.7 keseluruhan data digunakan digunakan untuk training dan sisanya digunakan sebagai testing.

Pengolahan dataset dilakukan berdasarkan:

- **Cleaning text**
Tahap ini melibatkan penghapusan karakter atau elemen teks yang tidak relevan atau mengganggu, seperti tanda baca, karakter khusus, dan URL. Tujuannya adalah untuk membersihkan teks sehingga hanya sisanya adalah informasi yang relevan untuk analisis.
- **Remove stopwords**
Stopwords adalah kata-kata umum yang tidak memiliki nilai informatif yang tinggi dalam analisis teks, seperti "the", "is", "and", dan sebagainya. Dalam tahap ini, stopwords dihapus dari teks karena mereka tidak memberikan kontribusi signifikan terhadap pemahaman makna teks.
- **Stemming and lemmatization**
Stemming dan lemmatization adalah proses normalisasi kata-kata dalam teks. Stemming menghapus akhiran kata untuk menghasilkan bentuk dasarnya (stem), sedangkan lemmatization mengubah kata-kata ke bentuk dasarnya yang disebut lemma. Contohnya, kata-kata "running", "runs", dan "ran" akan diubah menjadi "run" dalam proses stemming atau lemmatization. Tujuannya adalah untuk mengurangi variasi kata yang sama sehingga model dapat dengan lebih mudah memahami teks.
- **Tokenization**
Tokenization adalah proses memecah teks menjadi unit-unit yang lebih kecil yang disebut token. Token bisa berupa kata, frasa, atau bahkan karakter. Biasanya, dalam konteks NLP, tokenisasi dilakukan dengan memecah teks menjadi kata-kata individual. Setelah tokenisasi, setiap kata dalam teks akan direpresentasikan sebagai token yang terpisah, yang kemudian dapat diolah lebih lanjut dalam tahap analisis atau pemodelan selanjutnya.

Kemudian setelah data diproses dan model telah didapatkan, selanjutnya dilakukan LSTM. Tujuannya yaitu untuk memodelkan urutan kata dalam bahasa alami. Dengan memahami konteks dan urutan kata. LSTM dapat digunakan untuk memprediksi kata berikutnya dalam sebuah kalimat atau teks. Ini sering digunakan dalam aplikasi seperti prediksi teks, penyelesaian otomatis, atau generasi teks. LSTM juga dapat digunakan untuk menganalisis sentimen dari teks, yaitu menentukan apakah teks tersebut memiliki sentimen positif, netral, dan negatif.

Kemudian dilakukan testing menggunakan komentar/review lainnya.