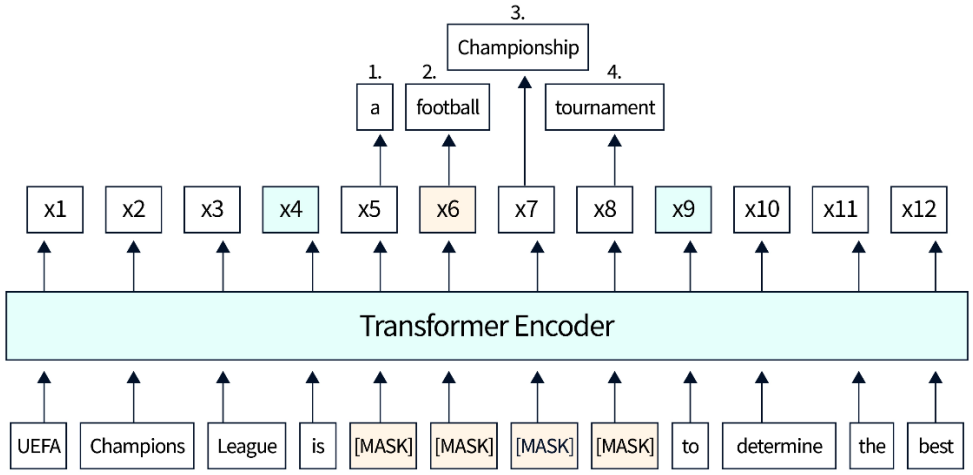
LSTM Model for classifying Harmful comments : Case study on Youtube datasets

A. Datasets

https://www.kaggle.com/datasets/atifaliak/youtube-comments-dataset

B. Transformer model (BERT) for word embedding



BERT adalah model berbasis *transformer* yang digunakan untuk memahami teks dengan membaca konteks kalimat secara 2 arah (Bidirectional) untuk mendapatkan pemahaman, ini bisa dimungkinkan dengan melatih model dengan cara membuat model menebak kata dalam sentence (Masked Language Model) dan Next Sentence Prediction (NSP)

1. Masked Language Model

Pada metode ini, beberapa token (kata) dalam kalimat akan disembunyikan (diberi label [MASK]), kemudian model dilatih untuk memprediksi kata yang tersembunyi tersebut berdasarkan konteks dua arah (sebelum dan sesudah kata yang tersembunyi).

Sentence	This tutorial is very **helpful** for learning Python programming.
Mask	This tutorial is very **[MASK]** for learning Python programming.
Model train	**Target Model:** helpful

2. Next Sentence Prediction

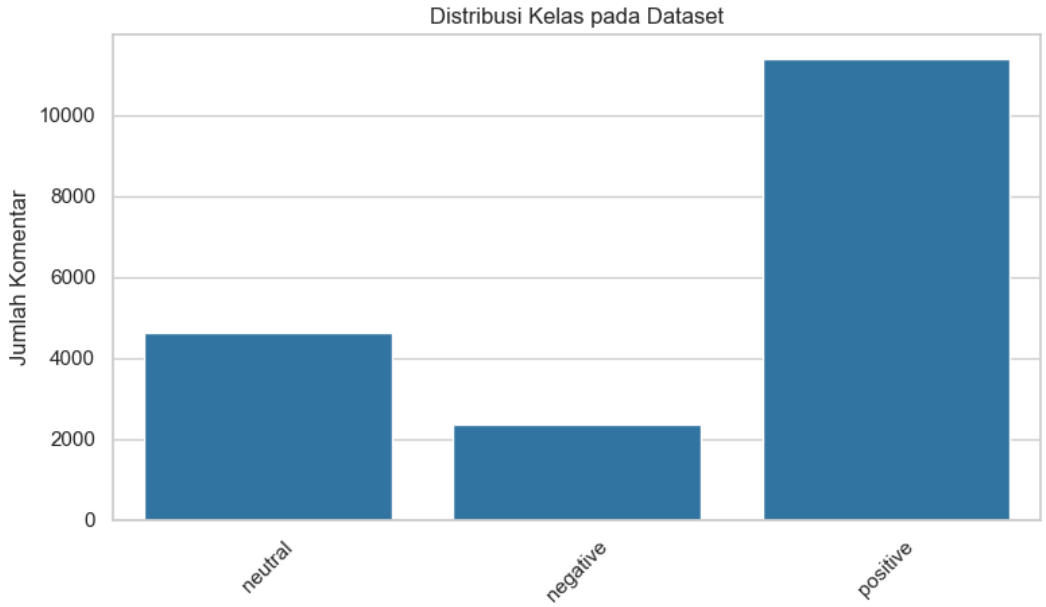
Dalam proses pelatihan NSP, model diberikan dua kalima, Kalimat A (Kalimat pertama) dan Kalimat B (Kalimat kedua). Model kemudian memprediksi apakah Kalimat B adalah kelanjutan logis dari Kalimat A (isNext) dan apakah Kalimat B bukan kelanjutan logis dari Kalimat A (notNext)

Sentence	This video explained how to analyze data in Python very clearly.
isNext	Now I can analyze data using python
notNext	Can anyone recommend a good pizza place?

3. **Kenapa menggunakan BERT model pada project**
    a. Dataset comment memiliki makna kontekstual yang kuat
    b. Struktur kalimat yang beragam dalam dataset comment sehingga butuh Bidirectional Model untuk benar benar memahami context dalam comment
    c. Handling unseen word (out of vocabulary) sehingga vector pada word yang out of vocabulary tidak di assign dengan random vector tetapi dapat di assign unknown
    d. Pre-trained via Transformers with large-scale text corpora (e.g., BooksCorpus, Wikipedia)

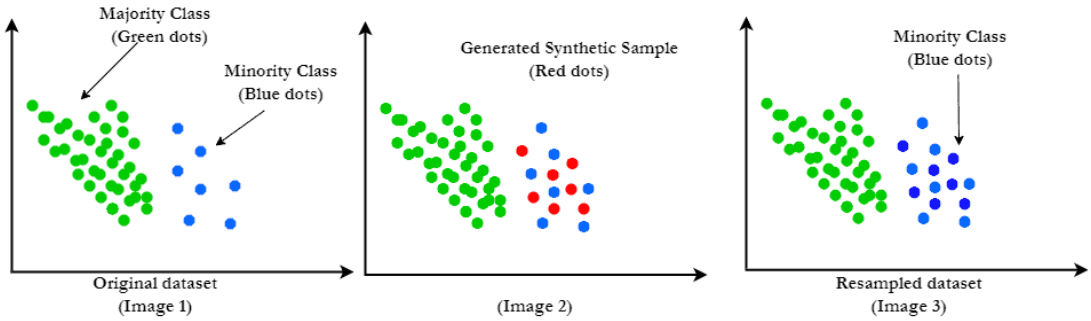
C. Re-Sampling data using SMOTE

1. Kenapa perlu resampling



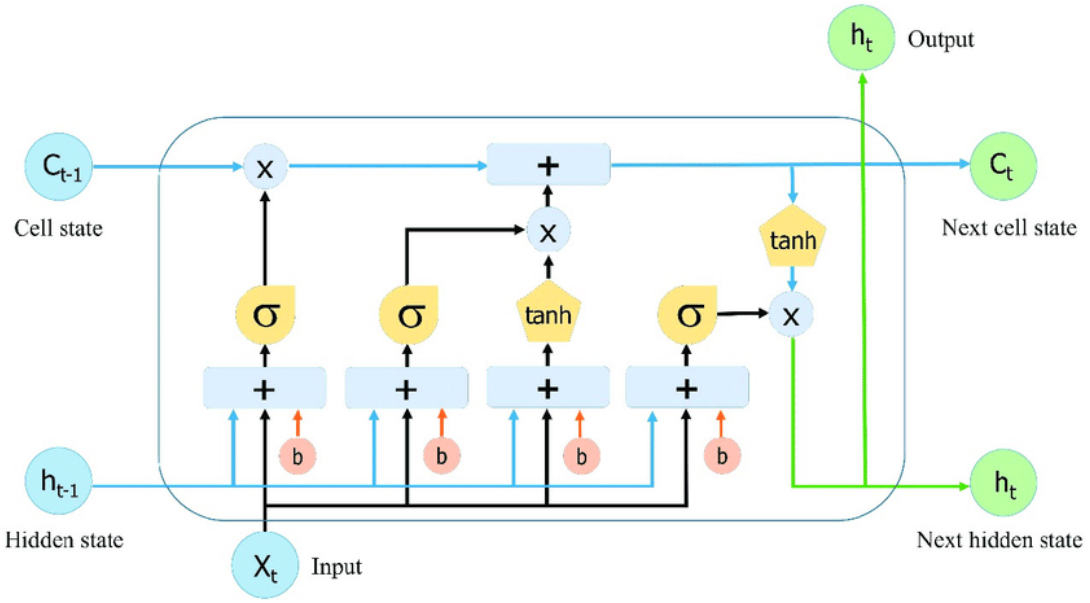
Jumlah comment dengan positive sentiment lebih banyak daripada class sentiment lain, ini menggambarkan bahwa dataset yang dipakai memiliki imbalanced data problem yang memungkinkan untuk model cenderung bias kepada class positive (majority class)

2. Kenapa menggunakan SMOTE



SMOTE adalah metode sampling yang digunakan untuk mengatasi masalah imbalanced data dengan membuat sampel sintetik pada kelas minoritas. Jadi SMOTE tidak menduplikasi data yang sudah ada (seperti oversampling konvensional), SMOTE menciptakan data baru yang unik dan realists. Ini dimungkinkan karena SMOTE menggunakan Algoritma K-nearest neighbour untuk identifikasi kelas dan dari selisih antar class di generate data sintetik (Data Baru=Sampel Asli+(Selisih)×Nilai Random (0 - 1))

D. LSTM Model



Karena dalam project ini menggunakan comment sebagai source of data, maka saya menggunakan LSTM model untuk algoritma deep learning yang digunakan untuk classification, karena comment merupakan data yang bersifat sequential dan ada context yang terkandung di dalam comment, LSTM bisa menangkap context tersebut baik context secara keseluruhan comment atau context untuk next word

1. **Input data**
   Misal dalam data ada comment yang seperti, : "Hari ini cuaca cerah, besok mungkin hujan." LSTM akan membaca teks ini per token (satu kata demi satu kata)

2. **Input gate**
   Kemudian LSTM akan menambahkan informasi baru yang penting dari data saat ini. Misal dari kata sebelumnya cuaca cerah adalah context untuk hari ini, maka informasi cuaca cerah di update dalam cell state

3. **Forget gate**
   Pada setiap iterasi, LSTM akan mengevaluasi informasi sebelumnya. Jika ada informasi yang sudah tidak relevan, ia akan "melupakan" data tersebut. Misal informasi tentang "Hari ini cerah" mungkin tidak penting jika kita hanya ingin memprediksi cuaca besok.

4. **Cell state**
   LSTM akan memperbarui memorinya berdasarkan hasil dari Forget Gate dan Input Gate dalam cell state. Inilah yang membuat LSTM mampu mengingat informasi dari waktu yang lama.

E. Project Pipeline