# Klasifikasi *Hate Speech* dan *Abusive Language* Berbahasa Indonesia dengan Unsupervised SimCSE dan IndoBERT

# Problem Statement

01. Sebagian besar metode klasifikasi hate speech masih bergantung pada data berlabel dalam jumlah besar.

02. Proses pelabelan teks berbahasa Indonesia memakan waktu, biaya, dan bersifat subjektif.

03. Model supervised sering kurang adaptif terhadap bahasa informal dan konteks baru yang terus berkembang di media sosial.

01

# Objective

01. Mengembangkan aplikasi klasifikasi teks berbahasa Indonesia untuk mendeteksi konten hate speech dan abusive language.

02. Menerapkan pendekatan Unsupervised SimCSE untuk menghasilkan representasi (embedding) kalimat yang lebih informatif tanpa ketergantungan penuh pada data berlabel.

03. Mengintegrasikan embedding hasil SimCSE dengan model IndoBERT yang dilengkapi layer classifier untuk melakukan klasifikasi teks secara biner.

04. Mengevaluasi performa sistem dalam membedakan teks bermuatan hate speech atau abusive language dengan teks netral berdasarkan metrik evaluasi standar klasifikasi.

03

Deep Learning - Presentation 2025

# Dataset

Name: Multi Label Hate Speech and Abusive Language Detection in Indonesian Twitter Dataset
Source: Twitter (x)
Language: Indonesian
Original Num of Labels: 12
2 Types of Labels:

- Hate_Abusive (55.50%) ➜ 7.309 Data
- Neutral (45.40%) ➜ 5.860 Data

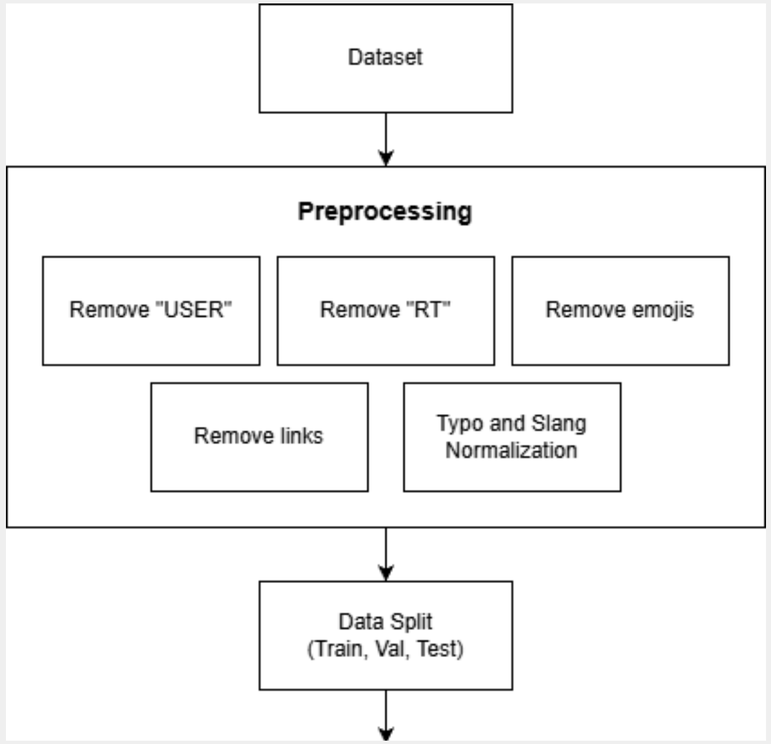
Total: 13.169 Data

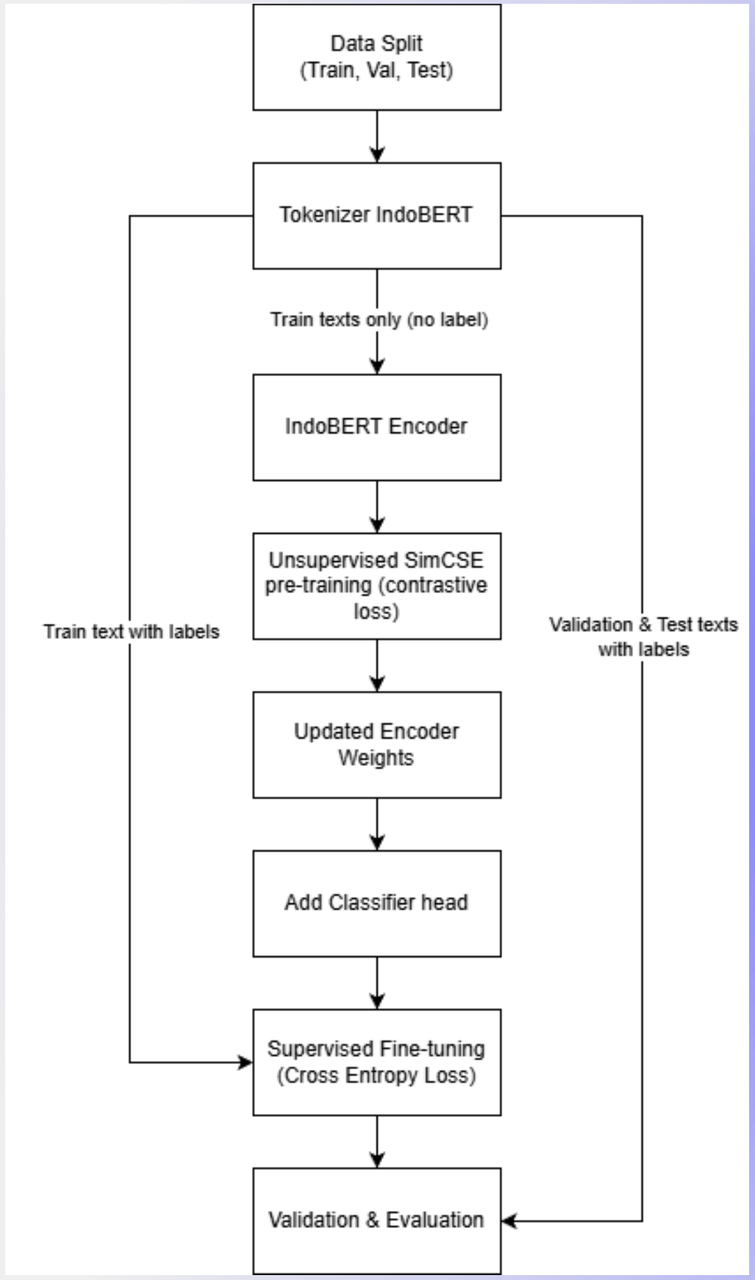
Data Splitting:
Train : Test : Validation ➜ 80 : 10 : 10

Anonimisasi Data - Seluruh username dan URL telah diganti dengan token khusus sesuai kebijakan platform twitter (X)

Final Project - Deep Learning

# Pipeline



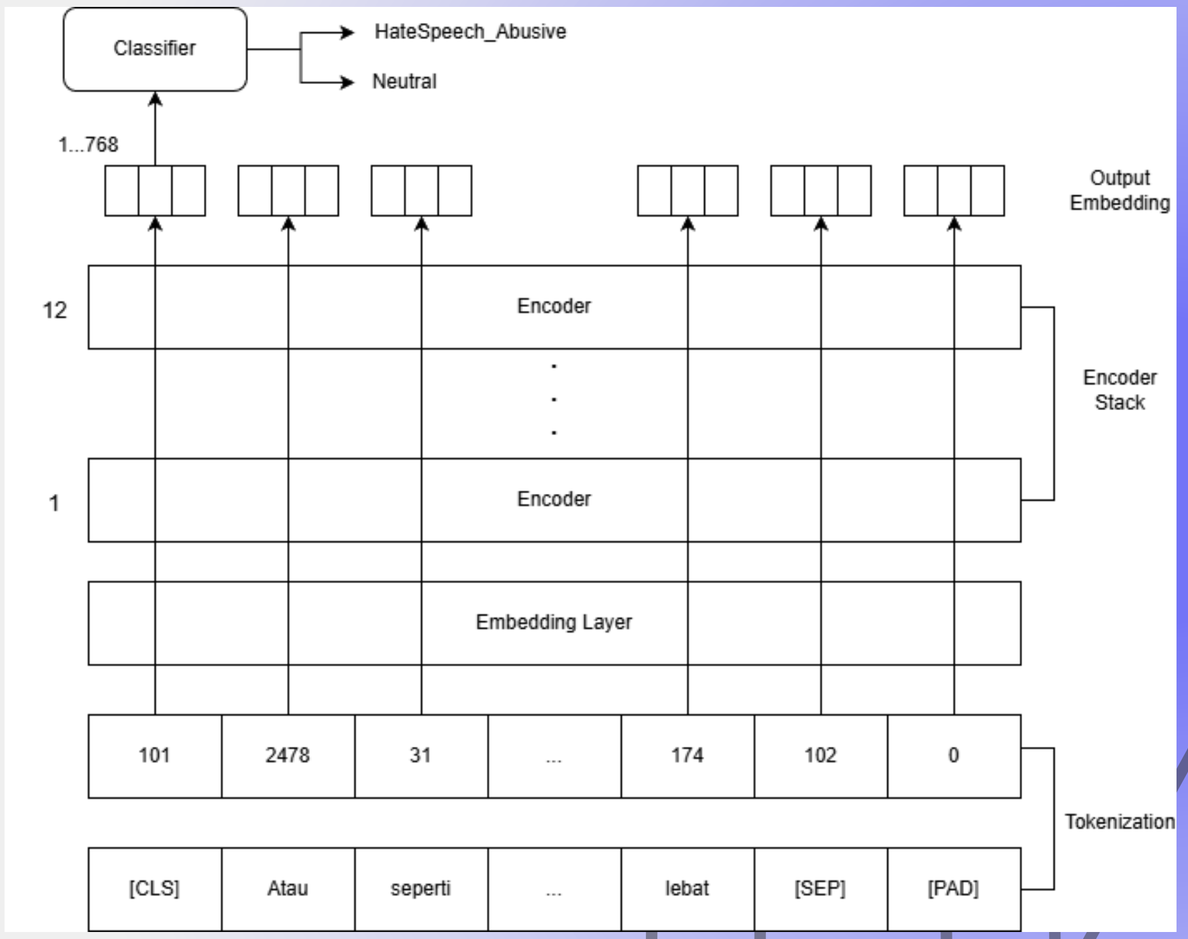
# Model

Model utama: IndoBERT (Transformer encoder)

- Arsitektur BERT-Base (12 encoder layer, hidden size 768, 12 attention heads)

Pada proyek ini:

- IndoBERT digunakan sebagai *backbone* model
- Unsupervised training dengan SimCSE (2 epoch)
- Fine-tuning untuk klasifikasi hate speech & abusive language (3 epoch)
- Fully Connected layer untuk klasifikasi → Output: Neutral | Hate/Abusive



# SimCSE (Simple Contrastive Learning of Sentence Embeddings)

- SimCSE adalah metode *contrastive learning* untuk meningkatkan kualitas *sentence embedding*
- Tujuan utama:
  - Kalimat bermakna mirip → embedding makin dekat
  - Kalimat tidak berhubungan → embedding makin jauh
- Pada Unsupervised SimCSE:
  - Satu kalimat (x) dilewatkan dua kali ke encoder
  - Menggunakan dropout aktif sebagai augmentasi ringan
- Kedua representasi tersebut membentuk positive pair ($h_i$ , $h_i^+$)
- Kalimat lain dalam batch dianggap sebagai negative samples
- Model dilatih untuk:
  - Memaksimalkan kemiripan positive pair
  - Meminimalkan kemiripan dengan negative samples

*Objective* dari Unsupervised SimCSE:

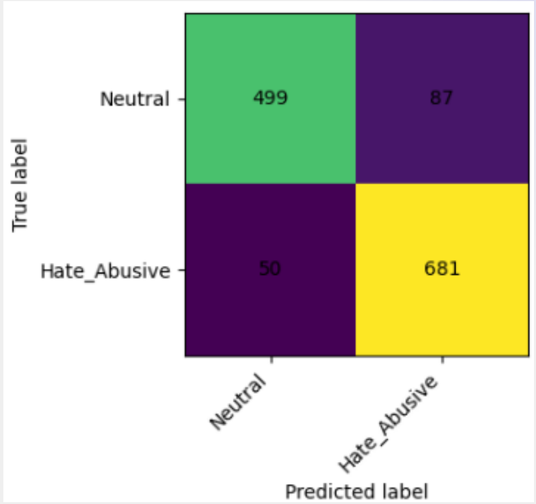
$$\ell_i = -\log \frac{e^{\text{sim}(\mathbf{h}_i^{z_i}, \mathbf{h}_i^{z_i'})/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(\mathbf{h}_i^{z_i}, \mathbf{h}_j^{z_j'})/\tau}},$$

# Result & Evaluation

a) Unsupervised SimCSE Pre-training dengan Fine-tuning IndoBERT

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Neutral | 0.9089 | 0.8515 | 0.8793 | 586 |
| Hate_Abusive | 0.8867 | 0.9316 | 0.9086 | 731 |
| accuracy | | | 0.8960 | 1317 |
| macro avg | 0.8978 | 0.8916 | 0.8940 | 1317 |
| weighted avg | 0.8966 | 0.8960 | 0.8956 | 1317 |

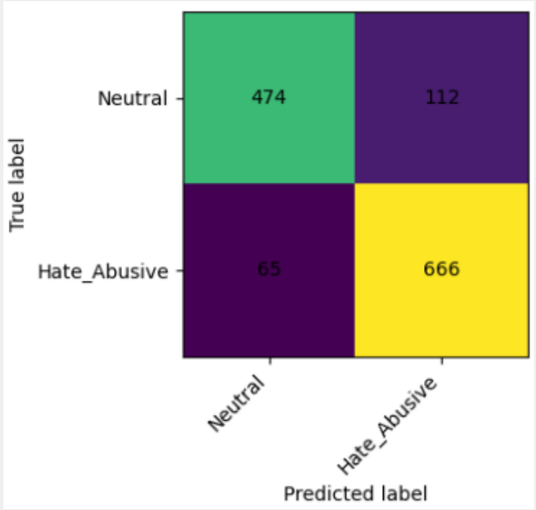
Accuracy: 0.8960




b) Fine-tuning IndoBERT tanpa Unsupervised SimCSE Pre-training

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Neutral | 0.8794 | 0.8089 | 0.8427 | 586 |
| Hate_Abusive | 0.8560 | 0.9111 | 0.8827 | 731 |
| accuracy | | | 0.8656 | 1317 |
| macro avg | 0.8677 | 0.8600 | 0.8627 | 1317 |
| weighted avg | 0.8664 | 0.8656 | 0.8649 | 1317 |

Accuracy: 0.8656




Penambahan SimCSE membantu meningkatkan kualitas embedding kalimat. Model lebih baik dalam membedakan teks Neutral dan Hate_Abusive, dengan kesalahan prediksi yang lebih sedikit.

kalimat-kalimat yang semantik mirip cenderung membentuk klaster atau kelompok yang lebih rapat. Maka, classifier akan lebih mudah menarik *decision boundary* yang memisahkan label dengan lebih bersih

# Application Demo

https://drive.google.com/drive/folders/1-Zb3VtAR7fgMLHYrNZCBrQUZF8bD_n1k



06

07

# Final reflections and future steps

- Unsupervised SimCSE pre-training meningkatkan performa IndoBERT secara konsisten dibanding baseline, dengan akurasi naik dari 0.8656 menjadi 0.8960 serta penurunan False Positive dan False Negative.
- Pendekatan contrastive learning pada SimCSE membantu membentuk ruang embedding yang lebih terstruktur, sehingga pemisahan kelas Neutral dan Hate/Abusive menjadi lebih jelas.
- Kualitas embedding [CLS] berperan penting karena digunakan langsung sebagai representasi kalimat dalam proses klasifikasi.
- Keterbatasan penelitian ini meliputi penyederhanaan dataset menjadi klasifikasi biner dan penggunaan strategi representasi kalimat yang masih sederhana.
- Pengembangan selanjutnya dapat mencakup multi-label classification, eksplorasi pooling alternatif, hyperparameter tuning, serta analisis error untuk menangani sarkasme, makna implisit, dan slang baru.

# GitHub Repositories Link

Repositories Link

# Thank You