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

Deep Learning / LC01 / Semester 5

Disusun Oleh:

Vallerie Alexandra Putra	2702265793
Gabriel Jonathan	2702243142
Edward Rivaldi Kosasih	2702353770

Universitas Bina Nusantara

Kemanggisan

2025

Chapter 1: Introduction

1.1 Background

Perkembangan media sosial dan platform digital telah meningkatkan jumlah interaksi berbasis teks secara signifikan. Di sisi lain, kemudahan dalam menyampaikan opini juga mendorong meningkatnya penyebaran *hate speech* dan *abusive language*, termasuk dalam konteks bahasa Indonesia. Ujaran kebencian atau *hate speech* adalah penggunaan bahasa yang menyerang atau merendahkan, serta menghasut kekerasan atau kebencian terhadap suatu kelompok berdasarkan karakteristik tertentu (misalnya penampilan fisik, agama, keturunan, asal-usul nasional atau etnis, orientasi seksual, identitas gender, dan lain-lain), dan dapat muncul dalam berbagai gaya bahasa, termasuk bentuk yang halus maupun yang dibungkus humor [1]. Konten semacam ini dapat memicu konflik sosial, diskriminasi, serta berdampak negatif terhadap individu maupun masyarakat secara luas, sehingga diperlukan sistem klasifikasi otomatis yang mampu mendeteksi konten berbahaya secara efektif.

Pendekatan klasifikasi teks berbasis deep learning, khususnya model transformer seperti BERT, telah menunjukkan performa yang baik dalam berbagai tugas Natural Language Processing (NLP). IndoBERT [2] sebagai model pre-train untuk bahasa Indonesia mampu memahami karakteristik linguistik lokal dengan lebih baik. Namun, sebagian besar metode yang ada masih bergantung pada *supervised learning*, yang membutuhkan data berlabel dalam jumlah besar, sementara proses anotasi *hate speech* dan *abusive language* bersifat mahal dan subjektif.

Untuk mengatasi keterbatasan tersebut, kami memanfaatkan pendekatan *unsupervised representation learning* seperti Unsupervised SimCSE (Simple Contrastive Learning of Sentence Embeddings) [3] di atas model IndoBert untuk menghasilkan embedding kalimat yang lebih kaya dan berkualitas secara semantik tanpa memerlukan label. Dengan mengkombinasikan Unsupervised SimCSE dan IndoBERT, penelitian ini bertujuan untuk membangun sistem klasifikasi *hate speech* dan *abusive language* berbahasa Indonesia yang lebih akurat dan efektif.

1.2 Problem Statement

Meskipun berbagai metode telah dikembangkan untuk klasifikasi *hate speech* dan *abusive language*, sebagian besar pendekatan yang ada masih bergantung pada data berlabel dalam jumlah besar. Pada praktiknya, proses pelabelan data teks berbahasa Indonesia membutuhkan waktu, biaya, serta melibatkan subjektivitas anotator, sehingga sulit untuk diskalakan. Selain itu, model *supervised* seringkali kurang adaptif terhadap variasi bahasa informal dan konteks baru yang terus berkembang di media sosial. Oleh karena itu, diperlukan pendekatan yang dapat meningkatkan kualitas *sentence embeddings* IndoBERT, seperti melalui unsupervised SimCSE, dan mengevaluasi sejauh mana pendekatan tersebut dapat memperbaiki performa deteksi ujaran kebencian dan bahasa yang tidak pantas dalam Bahasa Indonesia.

1.3 Objectives

Tujuan dari pengembangan project ini adalah sebagai berikut:

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

- Menerapkan pendekatan **Unsupervised SimCSE** untuk menghasilkan representasi (embedding) kalimat yang lebih informatif tanpa ketergantungan penuh pada data berlabel.
- Mengintegrasikan embedding hasil SimCSE dengan model **IndoBERT** yang dilengkapi layer classifier untuk melakukan klasifikasi teks secara biner.
- Mengevaluasi performa sistem dalam membedakan teks bermuatan *hate speech* atau *abusive language* dengan teks netral berdasarkan metrik evaluasi standar klasifikasi.

1.4 Significance

Manfaat dan nilai dari project ini antara lain:

- Mengurangi ketergantungan pada proses pelabelan data secara manual yang membutuhkan waktu dan sumber daya besar.
- Mendukung pengembangan sistem moderasi konten otomatis berbahasa Indonesia yang dapat digunakan pada platform media sosial atau aplikasi berbasis teks.
- Menunjukkan penerapan pendekatan **unsupervised learning**, **supervised learning** dan **transformer-based model** dalam pengembangan aplikasi Natural Language Processing (NLP).
- Menjadi referensi implementatif bagi pengembang atau mahasiswa yang ingin membangun sistem klasifikasi teks berbahasa Indonesia menggunakan SimCSE dan IndoBERT.

Chapter 2: Related Works

2.1 BERT

BERT [4] adalah model encoder berbasis Transformer yang mempelajari representasi *bidirectional* (konteks kiri dan kanan) pada seluruh lapisan encoder melalui tahap pre-training di data teks besar tanpa label, lalu diadaptasi ke berbagai tugas *downstream* dengan fine-tuning end-to-end menggunakan tambahan layer output yang minimal. BERT menggunakan *Masked Language Modeling* (MLM) sehingga representasi dapat menggabungkan konteks dua arah dan lebih cocok untuk tugas pemahaman kalimat maupun token-level.

Arsitektur BERT sendiri adalah multi-layer bidirectional Transformer encoder yang mengacu pada implementasi Transformer [5]. Secara umum, BERT mendefinisikan jumlah layer sebagai L, dimensi hidden sebagai H, dan jumlah attention head sebagai A. Salah satu konfigurasi yang paling umum digunakan adalah BERT-Base (L=12, H=768, A=12) dengan sekitar 110M parameter, dan feed-forward size = 4H (3072).

Pada tahap input, BERT menggunakan tokenisasi WordPiece (vocabulary 30.000) serta dua token khusus: [CLS] di awal urutan dan [SEP] sebagai pemisah antar kalimat. Representasi [CLS] pada hidden state terakhir digunakan sebagai representasi agregat untuk tugas klasifikasi. Selain token embedding, input BERT juga menjumlahkan *segment embedding* dan *position embedding*.

2.2 SimCSE

SimCSE (*Simple Contrastive Learning of Sentence Embeddings*) adalah metode *contrastive learning* untuk menghasilkan sentence embedding yang lebih selaras dengan kemiripan semantik. Intuisi utamanya adalah representasi dua kalimat yang maknanya mirip perlu “ditarik mendekat”, sedangkan representasi kalimat yang tidak berhubungan perlu “dijauhkan” di ruang embedding. Pada

Unsupervised SimCSE, pasangan positif (*positive pair*) dibentuk dari kalimat yang sama yang dilewatkan dua kali ke encoder dengan dropout mask berbeda (h_i^{zi} dan h_i^{zi+}), sehingga dropout berperan sebagai augmentasi minimal pada representasi tersembunyi. Metode ini menggunakan *in-batch negatives*, yaitu contoh lain dalam batch diperlakukan sebagai negatif dan model dilatih untuk memilih pasangan positif di antara negatif-negatif tersebut. Berikut *objective* dari unsupervised SimCSE training:

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

dengan τ sebagai *temperature* hyperparameter dan $\text{sim}(\cdot, \cdot)$ adalah *cosine similarity*.

Secara analitis, SimCSE dijelaskan meningkatkan kualitas embedding melalui *trade-off alignment* (positif makin dekat) dan *uniformity* (distribusi embedding makin menyebar merata), sehingga ruang embedding menjadi lebih ekspresif.

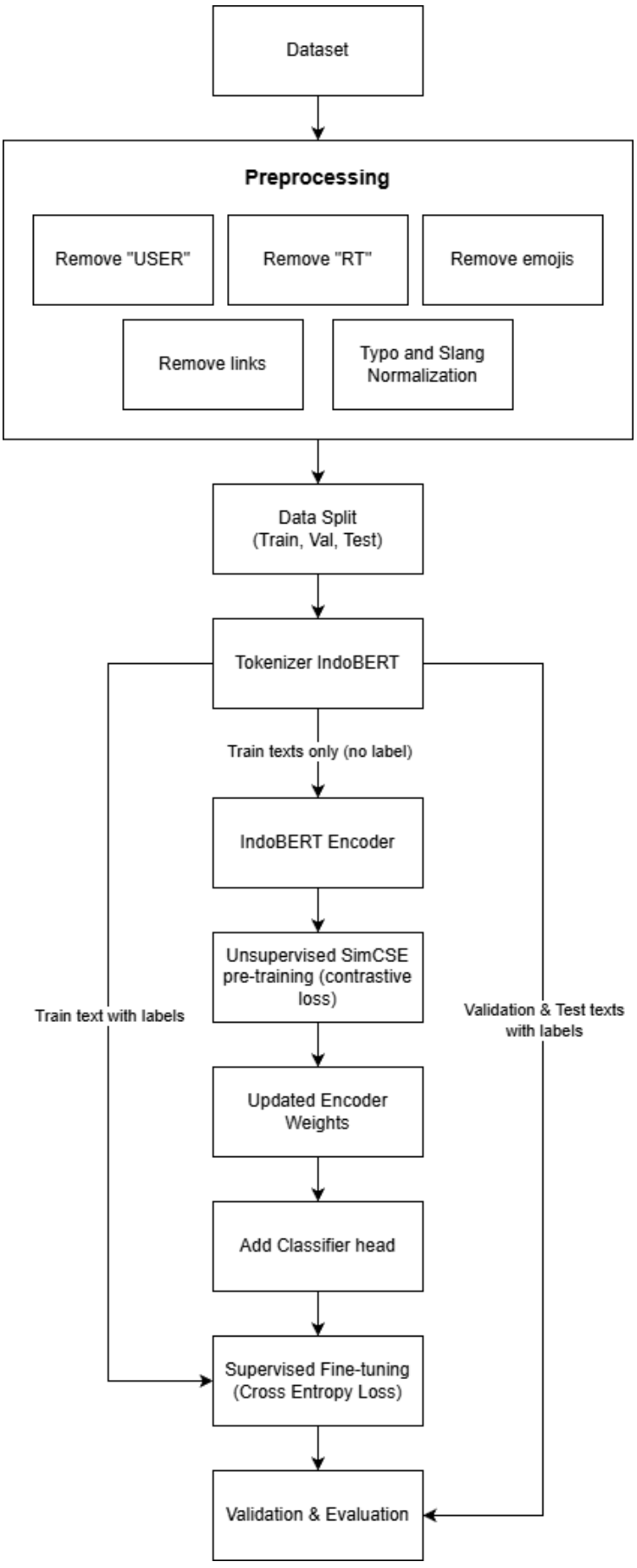
2.3 Relevant Works

[6] melakukan penelitian multi-label *hate speech* dan *abusive language detection* dengan skema label yang mencakup kombinasi seperti *no hate/no abusive, hate only, abusive only, dan hate + abusive* beserta variasi target dan kategori. Mereka membandingkan beberapa pendekatan machine learning seperti SVM, Naive Bayes, dan Random Forest Decision Tree (RFDT) dengan strategi problem transformation serta fitur n-gram dan leksikon. Hasilnya, kombinasi RFDT + Label Powerset (LP) dilaporkan sebagai konfigurasi terbaik, dan pada skenario deteksi hate + abusive tanpa rincian target atau kategori mencapai akurasi 77,36%.

[7] membangun model deep learning berbasis LSTM untuk klasifikasi hate speech menggunakan dataset yang sama dengan pembagian 80% data latih dan 20% data validasi. Arsitektur yang digunakan mencakup Embedding layer, LSTM, Dense/Dropout, Fully Connected (softmax) dan dilatih menggunakan Binary Cross Entropy. Konfigurasi terbaik dilaporkan pada penggunaan 256 neuron LSTM, dengan akurasi training 86,23% dan akurasi validasi 87,10% pada 10 epoch.

Chapter 3: Methodology

3.1 Pipeline


Dataset
Preprocessing
Remove "USER"
Remove "RT"
Remove emojis
Remove links
Typo and Slang Normalization
Data Split (Train, Val, Test)
Tokenizer IndoBERT
Train texts only (no label)
IndoBERT Encoder
Unsupervised SimCSE pre-training (contrastive loss)
Train text with labels
Validation & Test texts with labels
Updated Encoder Weights
Add Classifier head
Supervised Fine-tuning (Cross Entropy Loss)
Validation & Evaluation

3.2 Dataset

Dataset **Multi Label Hate Speech and Abusive Language Detection in Indonesian Twitter** adalah dataset publik yang bersumber dari Twitter (X) dan berisi teks berbahasa Indonesia untuk deteksi *hate speech* dan *abusive language*. Dataset tersebut dikurasi dan dibangun oleh [6] dengan bantuan anotasi hasil *crowdsourcing*. Dataset ini awalnya dirancang untuk skema *multi-label classification*, namun pada proyek ini disederhanakan menjadi **klasifikasi biner** (Hate Speech or Abusive, Neutral).

- **Sumber data**: Twitter (X)
- **Bahasa**: Indonesia
- **Jumlah kelas** (custom): 2 (Neutral, Hate_Abusive)

Distribusi Label:

- **Hate_Abusive**: 7.309 data (**55,50%**)
- **Neutral**: 5.860 data (**44,50%**)

Skema label:

- **Hate_Abusive**: teks yang mengandung *hate speech* dan/atau *abusive language*
- **Neutral**: teks tanpa konten *hate speech* atau *abusive language*

Pembagian Data

- **Training**: 10.535 data (**80%**)
- **Validation**: 1.317 data (**10%**)
- **Testing**: 1.317 data (**10%**)
- **Anonimisasi data**: Seluruh username dan URL telah diganti dengan token khusus sesuai kebijakan platform Twitter (X)

3.3 Preprocessing

Tahapan preprocessing pada proyek ini dilakukan untuk membersihkan teks sebelum diproses oleh model. Proses yang diterapkan bersifat minimal untuk menjaga konteks dan makna kalimat tetap utuh. Langkah preprocessing yang dilakukan meliputi:

- Menghapus username atau mention pengguna.
- Menghapus token RT atau *retweet*.
- Menghapus emoji yang terdapat pada teks.
- Menghapus tautan.
- Mengganti atau menormalisasikan kata-kata *typo* atau *slang* (dilakukan dengan menggunakan tabel daftar kata *typo/slang* beserta perbaikannya yang tersedia di dalam dataset)

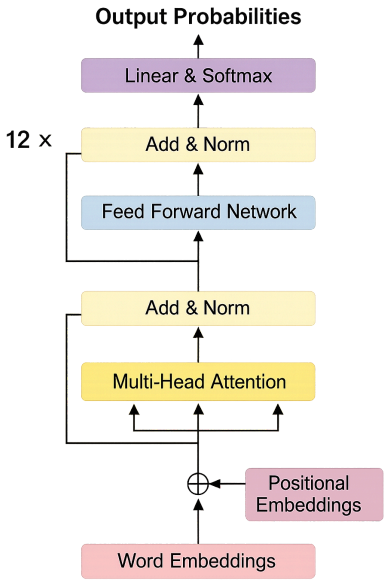
Tahapan ini bertujuan untuk mengurangi *noise* pada data teks tanpa menghilangkan informasi semantik yang penting bagi proses pembelajaran model.

Contoh hasil preprocessing:

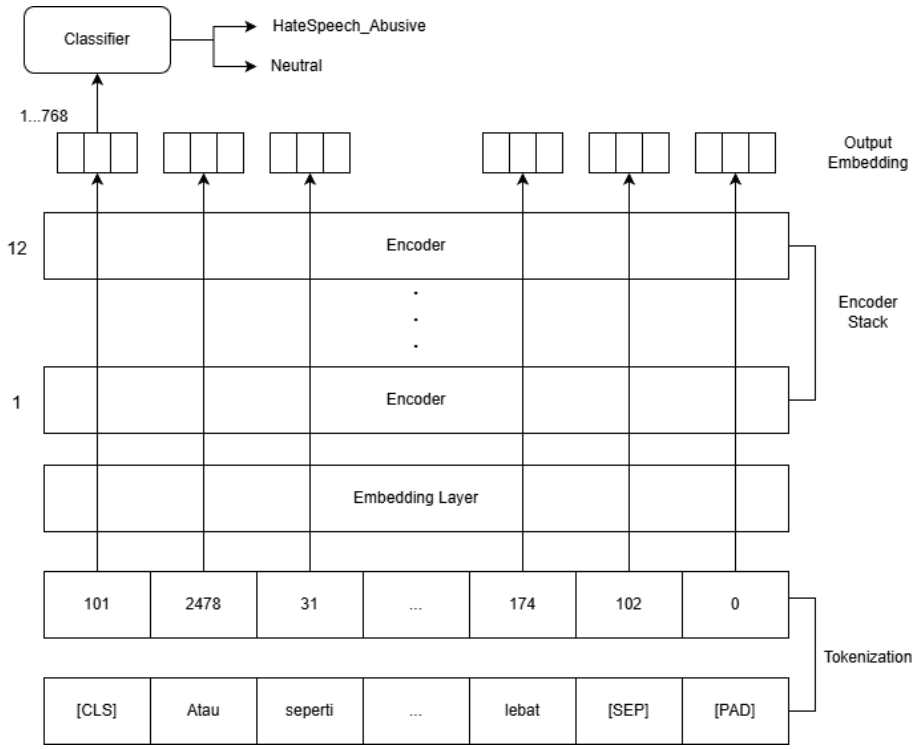
Sebelum preprocessing	Setelah preprocessing
USER Burik. Iya yg gitu burik panuan kutuan jamuran. Hah emang rang ajar.	Burik. Iya yang begitu burik panuan kutuan jamuran. ha memang rang ajar.
RT USER: Manusia kampret bernama Syahroni B Daud ini akhirnya MENGAKU SALAH SUDAH MENUDUH ADA TELUR PALSU dan meminta maaf. https:\xe2\x80\xa6	: Manusia kampret bernama Syahroni B Daud ini akhirnya MENGAKU SALAH SUDAH MENUDUH ADA TELUR PALSU dan meminta maaf.

3.4 Architecture

Model utama yang digunakan pada project ini adalah **IndoBERT**, yaitu model berbasis **Transformer** yang berperan sebagai *encoder*. Sebagai model encoder, BERT dirancang untuk **memahami dan merepresentasikan teks**, bukan untuk menghasilkan teks baru seperti model generatif (misalnya GPT). Output dari IndoBERT berupa **representasi vektor (embedding)** yang menangkap makna dan konteks dari kalimat input. Pada tugas klasifikasi, representasi token khusus [CLS] digunakan sebagai representasi keseluruhan kalimat.



BERT-Base Architecture. Gambar diambil dari https://pub.towardsai.net/bert-in-depth-exploration-of-architecture-workflow-code-and-mathematical-foundations-0c67ad24725b



Model Architecture. Gambar diadaptasi dari https://towardsdatascience.com/interpreting-the-prediction-of-bert-model-for-text-classification-5ab09f8ef074/

IndoBERT sendiri merupakan hasil *pre-training* dengan Masked Language Modeling (MLM) menggunakan jutaan data teks berbahasa Indonesia yang berasal dari Wikipedia Indonesia, artikel berita Kompas, Tempo, dan Liputan6, dan sebagainya, sehingga model ini telah memiliki pemahaman yang baik terhadap struktur, kosakata, dan konteks bahasa Indonesia. IndoBERT dilatih dengan konfigurasi BERT-Base yaitu 12 layer Transformer encoder, ukuran *hidden state* sebesar 768, dan 12 *attention heads* pada setiap layer, serta *position-wise feed-forward network* dengan dimensi hidden sebesar 3.072. Model ini memiliki kurang lebih 110 juta parameter.

Pada project ini, IndoBERT digunakan sebagai **backbone model** dan **sebelum proses supervised fine-tuning**, encoder IndoBERT terlebih dahulu **dilatih ulang secara unsupervised** menggunakan **metode SimCSE** selama 2 epochs untuk meningkatkan kualitas representasi kalimat. Selanjutnya, encoder hasil SimCSE di-*fine-tune* untuk tugas spesifik (*downstream task*), yaitu klasifikasi hate speech dan abusive language selama 3 epochs. Agar representasi teks yang dihasilkan IndoBERT dapat diubah menjadi prediksi kelas, ditambahkan **classifier layer** berupa *fully connected layer* di atas output encoder. Layer ini berfungsi untuk memetakan embedding hasil IndoBERT menjadi **probabilitas kelas**, yaitu **Neutral** atau **Hate_Abusive**. Dengan kombinasi IndoBERT dan classifier, sistem dapat mengubah pemahaman semantik teks menjadi keputusan klasifikasi secara langsung.

Secara lebih detail, teks input terlebih dahulu melalui tokenisasi dengan menambahkan token khusus [CLS] di awal kalimat, [SEP] sebagai penanda akhir, serta [PAD] untuk *padding*. Token-token tersebut dipetakan ke embedding layer dan kemudian diproses oleh 12 lapisan Transformer encoder yang masing-masing terdiri dari mekanisme multi-head self-attention dan position-wise feed-forward network, dilengkapi dengan residual connection dan layer normalization. Mekanisme multi-head self-attention memungkinkan setiap token untuk memperhatikan token lain dalam kalimat secara paralel sehingga konteks global dapat ditangkap. Pada keluaran encoder terakhir, setiap token

menghasilkan embedding kontekstual berdimensi 768, namun untuk keperluan klasifikasi, proyek ini secara langsung menggunakan embedding token [CLS] sebagai representasi keseluruhan kalimat tanpa menerapkan metode pooling tambahan seperti mean atau max pooling. Embedding [CLS] tersebut kemudian diteruskan ke classifier berupa fully connected layer untuk menghasilkan prediksi kelas, yaitu Neutral atau HateSpeech_Abusive.

3.5 Training Setup

Backbone Model: IndoBERT (Transformer Encoder)

Representation Learning: Unsupervised SimCSE

Device: GPU (CUDA)

Optimizer: AdamW

Loss Function:

- **SimCSE**: Contrastive Loss
- **Classification**: Cross Entropy Loss

Data Split:

- **Training**: **80%** (10.535 data)
- **Validation**: **10%** (1.317 data)
- **Testing**: **10%** (1.317 data)

Training Configuration

- Maximum sequence length: 256
- Truncation and padding enabled
- Seed: 42
- **SimCSE Pretraining**:
    - Epochs: 2
    - Learning rate: 5e-5
    - Batch size: 128
- **Supervised Fine-tuning (Classifier)**:
    - Epochs: 3
    - Learning rate: 3e-5
    - Weight decay: 0.01
    - Batch size: 64

3.6 Evaluation Metrics

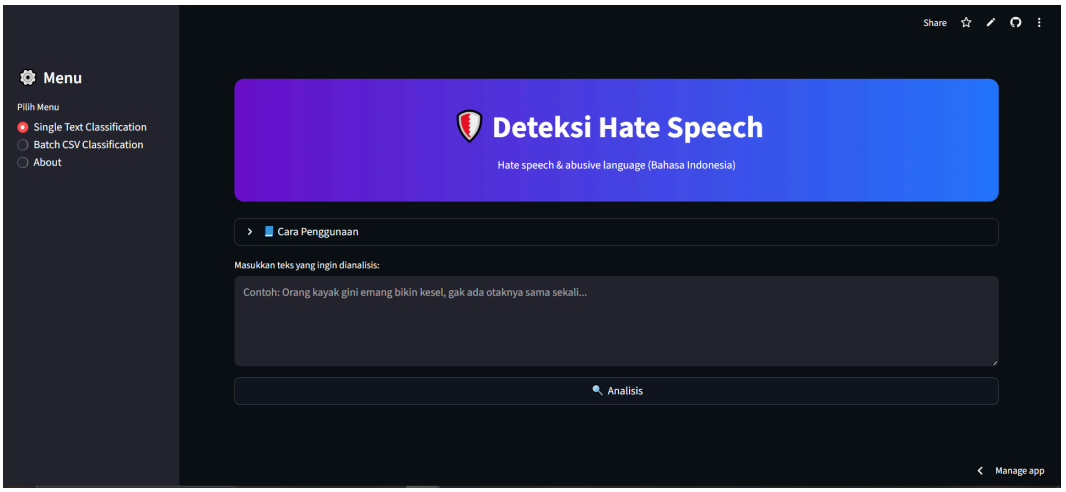
Precision, Recall, F1 Score, Accuracy

Chapter 4: Implementation

Setelah training selesai, model disimpan menggunakan **PyTorch** dalam format .pt. Karena ukuran file model cukup besar, penyimpanan di GitHub dilakukan menggunakan **Git Large File Storage (Git LFS)** agar repository tetap ringan dan mudah dikelola.

Untuk deployment, model diintegrasikan ke dalam aplikasi **Streamlit**. Meskipun model dilatih menggunakan CUDA, proses inferensi pada Streamlit dijalankan di **CPU**, yang tetap cukup untuk kebutuhan prediksi. Aplikasi ini menyediakan interface sederhana dan dapat diakses secara online melalui Streamlit Cloud.

Link aplikasi: https://indobert-hate-speech-classifier.streamlit.app/



4.1 Results

Classification Report

a. Classification Report untuk 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

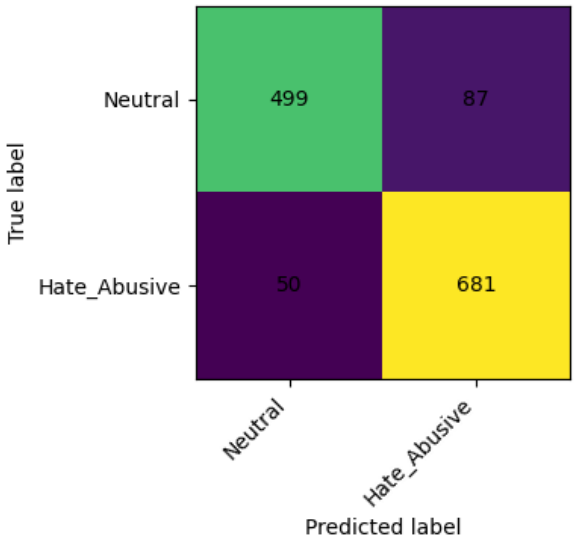
Accuracy: 0.8960

b. Classification Report untuk 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
```

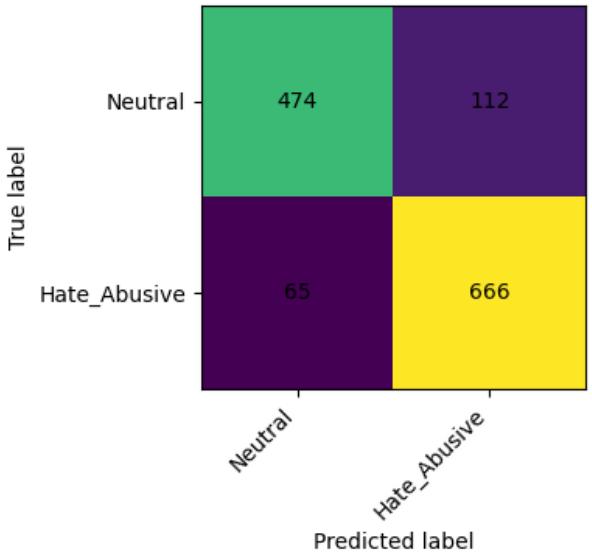
Confusion Matrix

a. **Confusion Matrix untuk Unsupervised SimCSE Pre-training dengan Fine-tuning IndoBERT**



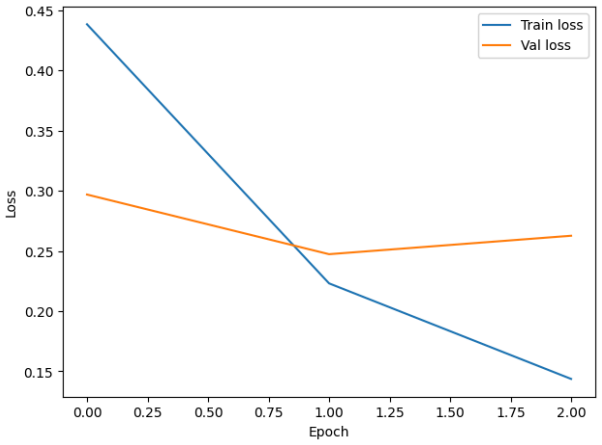
Keterangan:
True Negatives (Data Neutral yang berhasil diidentifikasi dengan benar): 499
False Positives (Data Neutral yang salah diidentifikasi sebagai Hate_Abusive): 87
False Negatives (Data Hate_Abusive yang salah diidentifikasi sebagai Neutral): 50
True Positives (Data Hate_Abusive yang berhasil diidentifikasi dengan benar): 681

b. **Confusion Matrix untuk Fine-tuning IndoBERT tanpa Unsupervised SimCSE Pre-training**

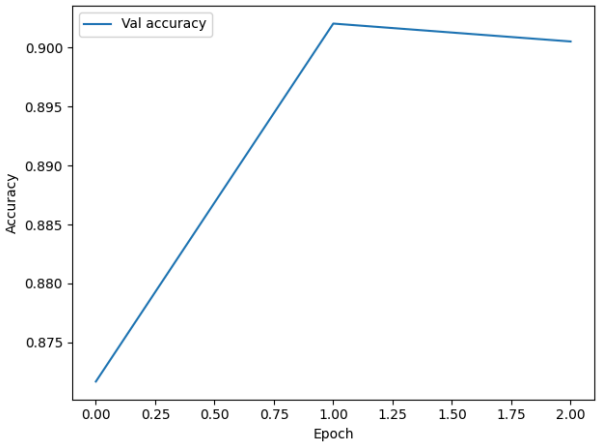


Keterangan:
True Negatives (Data Neutral yang berhasil diidentifikasi dengan benar): 474
False Positives (Data Neutral yang salah diidentifikasi sebagai Hate_Abusive): 112
False Negatives (Data Hate_Abusive yang salah diidentifikasi sebagai Neutral): 65
True Positives (Data Hate_Abusive yang berhasil diidentifikasi dengan benar): 666

Loss Curve untuk Supervised Fine-tuning



Validation Accuracy



Chapter 5: Discussion & Limitation

Berdasarkan hasil eksperimen, penambahan tahap Unsupervised SimCSE pre-training sebelum supervised fine-tuning IndoBERT meningkatkan performa dibanding baseline fine-tuning biasa, yang terlihat dari kenaikan akurasi dari 0.8656 menjadi 0.8960 (naik sekitar +3.04). Selain peningkatan akurasi, perbaikan juga dapat dilihat pada penurunan jenis kesalahan: False Positive menurun dari 112 ke 87 (lebih sedikit teks Neutral yang salah diprediksi sebagai Hate_Abusive) dan False Negative menurun dari 65 ke 50 (lebih sedikit teks Hate_Abusive yang lolos sebagai Neutral).

Secara teknikal, peningkatan tersebut masuk akal karena SimCSE melakukan *contrastive representation learning* pada encoder (tanpa label) dengan membuat dua "view" atau variasi dari input yang sama menggunakan dropout default dari model, lalu melatih model agar embedding kedua view tersebut menjadi lebih dekat, sementara embedding dari contoh lain dalam batch menjadi lebih jauh. Ketika ruang embedding menjadi lebih terstruktur, kalimat-kalimat yang semantik mirip cenderung membentuk klaster atau kelompok yang lebih rapat. Contohnya, dua kalimat bernada hinaan seperti "kamu goblok banget" dan "dasar bodoh", meskipun kata-katanya berbeda, diharapkan berada lebih dekat dibanding kalimat netral seperti "terima kasih atas informasinya". Dengan demikian, classifier di atas embedding kalimat akan lebih mudah menarik *decision boundary* yang memisahkan Neutral vs Hate_Abusive dengan lebih bersih. Dalam proyek ini, prediksi kelas dibuat dari embedding token [CLS] sebagai representasi keseluruhan kalimat (tanpa pooling mean/max), lalu diteruskan ke linear classifier, sehingga kualitas embedding [CLS] menjadi sangat krusial.

Namun, terdapat beberapa keterbatasan dalam proyek ini. Pertama, dataset asli bersifat multi-label namun pada proyek ini disederhanakan menjadi klasifikasi biner (Hate_Abusive vs Neutral). Penyederhanaan ini membantu implementasi tetapi berpotensi menghilangkan detail kategori seperti *abusive* tanpa *hate* atau *hate* dengan target tertentu, sehingga model tidak diuji pada label yang lebih spesifik. Lalu, arsitektur klasifikasi menggunakan embedding token [CLS] secara langsung tanpa pooling tambahan. Pendekatan ini kami anggap paling sederhana untuk diimplementasi, tetapi mungkin tidak paling optimal. Metode pooling lain dapat menjadi alternatif untuk studi atau eksperimen lebih lanjut.

Chapter 6: Conclusion & Future Work

Proyek ini membangun sistem klasifikasi teks berbahasa Indonesia untuk mendeteksi *hate speech* dan *abusive language* menggunakan IndoBERT sebagai *backbone* dan menambahkan tahap Unsupervised SimCSE pre-training sebelum supervised fine-tuning. Pipeline yang diterapkan mencakup preprocessing teks, pembagian data, representasi teks melalui IndoBERT, lalu klasifikasi biner menggunakan classifier head berbasis fully connected layer dengan input representasi kalimat dari token [CLS] tanpa metode pooling tambahan. Hasil eksperimen menunjukkan bahwa strategi SimCSE yang diikuti dengan fine-tuning memberikan peningkatan performa dibanding baseline fine-tuning biasa, dengan akurasi meningkat dari 0,8656 menjadi 0,8960 pada data uji. Peningkatan ini mengindikasikan bahwa *contrastive representation learning* melalui SimCSE mampu memperbaiki kualitas embedding kalimat sehingga pemisahan kelas menjadi lebih efektif pada tahap klasifikasi, yang pada akhirnya membuat model lebih akurat dalam membedakan teks Neutral dan HateSpeech_Abusive. Selain itu, model telah diintegrasikan ke aplikasi Streamlit sehingga dapat digunakan untuk inferensi secara praktis.

Untuk pengembangan selanjutnya, terdapat beberapa arah yang dapat dilakukan. Pertama, melakukan klasifikasi dengan skema multi-label classification sesuai bentuk dataset awal, sehingga

model dapat membedakan kategori yang lebih spesifik dan tidak hanya biner. Kedua, melakukan eksperimen representasi kalimat yang lebih kaya, misalnya membandingkan penggunaan embedding [CLS] dengan mean pooling/max pooling untuk melihat apakah informasi pada kalimat yang panjang dapat ditangkap lebih baik. Ketiga, melakukan tuning hyperparameter yang lebih menyeluruh untuk SimCSE dan fine-tuning agar konfigurasi lebih optimal. Keempat, menganalisis kesalahan prediksi untuk melihat eksistensi sarkasme, bahasa implisit, atau kata-kata slang baru yang belum tercakup normalisasi.

Team Contribution

Tugas	PIC	Penjelasan Tugas
Mencari dataset	Vallerie, Gabriel, Edward	Mencari dataset untuk tema
EDA	Vallerie, Gabriel	Melakukan EDA pada dataset
Preprocessing	Vallerie	Melakukan preprocessing
Model Creation & Training	Vallerie, Gabriel	Melakukan training model
Model Evaluation	Edward	Melakukan evaluation pada model
Deployment	Gabriel	Deployment model ke streamlit
Documentation and Report	Vallerie, Gabriel, Edward	Membuat report dan dokumentasi project
Presentation	Vallerie, Gabriel, Edward	Membuat slide presentasi

References

[1] P. Fortuna and S. Nunes, “A survey on automatic detection of hate speech in text,” ACM Computing Surveys (CSUR), vol. 51, no. 4, pp. 1-30, 2018. doi: https://doi.org/10.1145/3232676.

[2] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, “IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP,” In Proc. of the 28th International Conference on Computational Linguistics, 2020, pp. 757-770. doi: https://doi.org/10.18653/v1/2020.coling-main.66.

[3] T. Gao, X. Yao, and D. Chen. “SimCSE: Simple Contrastive Learning of Sentence Embeddings,” In Proc. of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 6894-6910. doi: https://doi.org/10.18653/v1/2021.emnlp-main.552

[4] J. Devlin, M. Chang, K. Lee, and K. Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” In Proc. of the 2019 Conference of

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171-4186. doi: https://doi.org/10.18653/v1/N19-1423

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017, pp. 5998-6008.

[6] M. O. Ibrohim and I. Budi, "Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter," in Proc. 3rd Workshop on Abusive Language Online (ALW), 2019, pp. 46-57, doi: 10.18653/v1/W19-3506.

[7] B. A. H. Kholifatullah and A. Prihanto, "Penerapan Metode Long Short Term Memory Untuk Klasifikasi Pada Hate Speech," Journal of Informatics and Computer Science (JINACS), vol. 4, no. 3, 2023, pp. 292-297. doi: 10.26740/jinacs.v4n03.p292-297.

Appendix