Progresif: Jurnal Ilmiah Komputer
https://ojs.stmik-banjarbaru.ac.id/index.php/progresif/index
Jl. Ahmad Yani, K.M. 33,5 - Kampus STMIK Banjarbaru
Loktabat – Banjarbaru (Tlp. 0511 4782881), e-mail: puslit.stmikbjb@gmail.com
e-ISSN: 2685-0877

Pemanfaatan Fitur Tambahan Emosi Untuk Deteksi *Hate Speech* Media Sosial Bahasa Indonesia

DOI: http://dx.doi.org/10.35889/progresif.v22i1.3338

Michael Joy Clement[1*], Hafiz Irsyad[2]
Informatika, Universitas Multi Data Palembang, Palembang, Indonesia
**e-mail Corresponding Author:* michaeljoyclement_2226250041@mdp.ac.id

Abstract

This study examines the importance of incorporatring emotion features and enhancing the temporal robustness of hate-speech detection models to improve classification accuracy. The research aims to analyze the impact of emotion features on an IndoBERT based model and to evaluate the model's adaptability using an unsupervised self-learning approach. The dataset consists of two corpora, a public dataset from 2019 and twitter data from 2025, each divided into training, validation, and test sets with an 80%, 10%, 10% split. Model performance is evaluated using accuracy, precision, recall and F1-score calculated from confusion matrix. Experimental results show that adding emotion features increases accuracy by 1-2% across all scenarios. In cross-temporal testing, the supervised model performance declines duet o linguistic shifts whereas the self-learning method improves accuracy up to 77.67%. These findings indicate that emotion features and self-learning effectively enhance the model's ability to adapt to evolving language and social context.

Keyword: *Emotion; Hate speech detection; IndoBERT*

Abstrak

Penelitian ini membahas pentingnya penambahan fitur emosi dan peningkatan ketahanan model deteksi ujaran kebencian terhadap perubahan bahasa lintas waktu guna memperkuat akurasi klasifikasi. Tujuan penelitian adalah menganalisis pengaruh fitur emosi pada model berbasis IndoBERT dan mengevaluasi kemampuan adaptasi model menggunakan pendekatan *unsupervised self-learning*. Data menggunakan dua korpus yaitu dataset publik tahun 2019 dan data Twitter tahun 2025, yang masing-masing dibagi menjadi data latih dan data latih, validasi, dan uji dengan proporsi 80%, 10%, dan 10%. Model dievaluasi menggunakan *accuracy*, *precision, recall,* dan *F1-score* yang dihitung melalui *confusion matrix*. Hasil pengujian menunjukkan bahwa penambahan fitur emosi meningkatkan akurasi sebesar 1-2% di seluruh skenario. Pada pengujian lintas waktu, performa model supervised menurun akibat perubahan konteks linguistik, namun metode *self-learning* meningkatkan akurasi hingga 77.67%. temuan ini menunjukkan bahwa fitur emosi dan *self-learning* efektif meningkatkan adaptasi model terhadap dinamika bahasa serta konteks sosial.

Kata kunci: *Seteksi ujaran kebencian; Emosi; IndoBERT*

1. Pendahuluan

Perkembangan media sosial di Indonesia telah memberikan dampak besar terhadap cara masyarakat berkomunikasi dan berinteraksi [1]. Media sosial menjadi ruang terbuka bagi siapa saja untuk menyampaikan pendapat, berbagi informasi, dan juga mengekspresikan diri secara bebas [2]. Namun, kebebasan ini juga menghadirkan tantangan baru, salah satunya adalah meningkatnya kasus ujaran kebencian (*hate speech*) [3]. *Hate speech* tidak hanya menyinggung individu, tapi juga dapat menyerang kelompok berdasarkan suku agama, ras, gender, maupun orientasi politik [4]. Ujaran kebencian yang tersebar secara masif di ruang digital berpotensi memperburuk polarisasi sosial, menimbulkan konflik, dan juga dapat mengancam kerukunan Masyarakat [5].

Pentingnya deteksi otomatis terhadap *hate speech* telah mendorong banyak penelitian dalam ranah *Natural Language Processing* (*NLP*) seperti penelitian yang dilakukan oleh F. Andy Kusuma dan E. Wahyu Pamungkas [6], yang melakukan penelitian dengan menggunakan algoritma *SVM*, dan *Decision Tree*. Di Indonesia, penelitian ini semakin berkembang sejak tersedianya dataset publik seperti karya M. O. Ibrohim dan I. Budi [7] yang memperkenalkan korpus *multi-label* untuk deteksi *hate speech* dan *abusive language* dalam Bahasa Indonesia. Namun, dari keseluruhan penelitian tersebut sebagian besar hanya menggunakan satu input saja yaitu teks murni tanpa menggunakan fitur tambahan lainnya yang bisa saja relevan. Selain itu penelitian terdahulu yang menggunakan dataset yang sudah lama (2019) [7] masih belum bisa dipastikan hasilnya terhadap data sekarang karena tata Bahasa pada media sosial yang selalu berkembang baik dalam bahasa slang maupun resmi.

Salah satu alasan penting mengapa emosi relevan adalah karena ujaran kebencian sering kali memuat luapan emosi. Ungkapan kebencian biasanya muncul dalam bentuk kemarahan, ejekan, atau ekspresi kebencian yang kuat terhadap suatu individu atau kelompok [8]. Dengan demikian, emosi dapat dianggap sebagai sinyal tambahan yang memiliki potensi untuk memperkuat klasifikasi teks. Beberapa penelitian internasional mendukung asumsi ini. Misalnya penelitian yang dilakukan oleh [9], yang mengusulkan pendekatan *multi-task learning* (*MTL*) dengan *shared encoder* yang secara bersamaan menangani deteksi *hate speech* dan emosi. Penelitian lainnya yang dilakukan oleh Plaza-del-Arco et.al [10] melaporkan bahwa model *MTL* yang menggabungkan tugas deteksi *hate speech*, analisis emosi, dan identifikasi target dapat meningkatkan performa model. Penelitian-penelitian sebelumnya di Indonesia juga memang sudah banyak yang membangun model untuk deteksi *hate speech* seperti yang dilakukan oleh Bagestra et. al [11] yang menggunakan IndoBERT dan terbukti mampu mengungguli pendekatan klasik dengan Tingkat akurasi sampai 93% dibandingkan dengan metode seperti *SVM* yang hanya 76% dan *Naïve Bayes* yang hanya 51%. Namun, hampir semua penelitian tersebut hanya menggunakan teks sebagai input utama tanpa memanfaatkan fitur tambahan lain. Padahal dengan kekuatan IndoBERT dalam merepresentasikan teks, cukup menarik untuk menguji apakah model ini juga mampu memanfaatkan informasi emosional jika diberikan sebagai input tambahan. Selain penambahan fitur emosi, kualitas model dasar (*backbone*) juga berperan penting dalam menentukan keberhasilan dalam *text processing*. Untuk Bahasa Indonesia, salah satu model yang unggul adalah IndoBERT, sebuah model berbasis arsitektur *Bidirectional Encoder Representations from Transformers* (BERT) yang telah dilatih pada korpus besar Bahasa Indonesia. Studi sebelumnya yang dilakukan oleh Koto et.al [12] menunjukkan bahwa IndoBERT mencapai *state-of-the-art* pada berbagai tugas linguistik dasar seperti morfosintaksis, semantik, dan penalaran, termasuk analisis sentimen, dan *named entity recognition* (*NER*). Selain itu, penelitian yang dilakukan oleh Dhendra dan V. Gayuh Utomo [13] juga membuktikan keunggulan IndoBERT dalam tugas sentimen publik, dengan nilai *F1-score* mencapai 0.881, mengungguli model transformer multibahasa maupun model *deep learning* konvensional seperti *CNN* dan *BiLSTM* yang dimana keduanya memiliki *F1-score* sebesar 0.836 dan 0.833. Keunggulan IndoBERT terletak pada kemampuannya memahami konteks kalimat secara mendalam melalui mekanisme *self-attention* [14], sehingga lebih mampu menangkap makna dalam teks berbahasa Indonesia dibandingkan metode tradisional atau model berbasis kata seperti *Word2Vec*.

Berdasarkan pembahasan di atas, penelitian ini bertujuan untuk menguji pemanfaatan fitur emosi dalam meningkatkan performa model deteksi ujaran kebencian berbasis IndoBERT serta mengevaluasi kemampuan adaptasi model terhadap perubahan bahasa lintas waktu melalui pendekatan *unsupervised self-learning*. Selain itu, penelitian ini juga berfokus membandingkan performa model pada data lama dan data baru untuk mengidentifikasi sejauh mana degradasi akurasi terjadi akibat dinamika linguistik. Penelitian ini memberikan beberapa manfaat, yaitu memahami penerapan IndoBERT dalam tugas deteksi *hate speech*, menganalisis pengaruh fitur emosi terhadap performa model, serta mengetahui perbedaan kinerja model pada berbagai skenario eksperimen. Selain itu, penelitian ini berkontribusi pada pengembangan riset NLP berbahasa Indonesia, khususnya terkait integrasi analisis emosi ke dalam sistem deteksi ujaran kebencian.

2. Tinjauan Pustaka

Penelitian yang dilakukan oleh Glenn et.al [15] yang berjudul "*Emotion Classification of Indonesian Tweets using Bidirectional LSTM*" menggunakan metode *LSTM* dalam deteksi emosi

berbahasa Indonesia yang memiliki akurasi 70.83%. di sisi lain ada juga penelitian yang dilakukan oleh Habib et.al [16] yang menggunakan arsitektur *BERT+LSTM*, dan juga *BERT+CNN* mendapatkan performa jauh di bawah *baseline* nya, namun peneliti memberikan alasan dari penurunan performa tersebut, dikarenakan adanya perbedaan cara *split* dataset dan juga perbedaan dalam *preprocessing* data. Dari sini dapat kita lihat bahwa proses pembagian dataset, dan juga *preprocessing* dapat memberikan dampak yang tinggi terhadap performa dari modelnya.

Penelitian yang dilakukan oleh Ansyah et.al [17] mengombinasikan dua pendekatan yaitu dengan menggunakan *word* embedding dengan *lexicon-based features*, yang mendapatkan hasil akurasi hanya 60.1%. Di sisi lain ada penelitian yang dilakukan oleh Saputra *et. al* [18] yang menggunakan model IndoBERT dalam proses deteksi emosinya, dan mendapatkan nilai akurasi sebesar 73%, dari sini bisa dilihat bahwa model *IndoBERT* menghasilkan performa yang lebih baik dalam mendeteksi emosi, yang kemudian sistem pendeteksian emosi ini akan digunakan untuk *labeling* emosi pada dataset 2019 nantinya.

Penelitian yang dilakukan oleh Mnassri et.al. [9] menyatakan bahwa penggunaan fitur emosi sebagai variabel tambahan dapat meningkatkan performa dari model pendeteksian *hate speech*, di sisi lain, penelitian yang dilakukan oleh Plaza-del-Arco et.al [10] menggunakan metode *Multi-Task Learning*, yang melakukan *training* model secara paralel antara deteksi emosi, sentimen, dan target dari *hate speech* itu sendiri. Jika dilihat dari sisi emosinya saja terjadi penurunan performa, namun saat semua fiturnya digabungkan performanya meningkat, dari sini bisa disimpulkan bahwa fitur emosi bisa memiliki pengaruh dalam deteksi *hate speech* itu sendiri.

Penelitian dari J. Forry Kusuma and A. Chowanda [19] yang membahas mengenai pendeteksian *hate speech* menggunakan model *IndoBERTweet*, dan *BiLSTM*, saat menggunakan *IndoBERTweet* saja, akurasinya berada di 88.5%, sedangkan saat digabungkan menggunakan *BiLSTM*, performanya hanya meningkat menjadi 88,6%, dari sini bisa dikatakan bahwa arsitektur dari *IndoBERTweet* sudah dapat menghasilkan model yang cukup baik tanpa adanya tambahan arsitektur lainnya, penelitian yang dilakukan oleh Bagestra et. al [11] juga merujuk ke arah yang sama yang di mana peneliti membandingkan kinerja dari berbagai algoritma *machine learning* klasik terhadap *transformer* modern seperti *IndoBERT*, dan *IndoBERTweet* dalam tugas untuk mendeteksi *multi-label hate speech* berbahasa Indonesia, dan hasilnya menunjukkan bahwa model modern seperti *IndoBERT* dan *IndoBERTweet* memiliki hasil yang jauh lebih baik dengan *F1-Score* sebesar 84% dibandingkan dengan metode-metode lainnya yang rata-ratanya di 68%. Ada juga penelitian yang membangun model deteksi *hate speech* multi-bahasa, seperti penelitian yang dilakukan oleh Usman et.al [20] yang memanfaatkan model *GPT 3.5 turbo* untuk melakukan pendeteksian dan menghasilkan performa yang jauh lebih tinggi daripada *baseline* yang menggunakan metode *SVM*, yang peningkatan performanya sampai 9,19%. Penelitian lainya dari (Taradhita & Putra, 2021) menggunakan metode *CNN* untuk menghasilkan model deteksi *hate speech*, dan mendapatkan akurasi tertinggi di 82.5% dengan data tes berjumlah 100 dengan label yang sama rata (50-50 untuk *hate speech* dan *non hate-speech*), namun seiring meningkatnya data tes akurasinya semakin menurun, sampai dengan data tes berjumlah 400 (200-200) akurasinya menurun sampai 73%. Dari sini bisa dilihat bahwa model *transformer* modern seperti *IndoBERT*, *GPT 3.5 turbo*, dan *IndoBERTweet* unggul dibandingkan dengan model-model *machine learning* lainnya.

State of the art dari penelitian ini terletak pada pemanfaatan fitur emosi sebagai variabel tambahan dalam model deteksi *hate speech* berbasis IndoBERT. Berbeda dengan penelitian-penelitian sebelumnya yang sebagian besar hanya menggunakan teks sebagai satu-satunya sumber informasi [9], [10], [11], penelitian ini menguji apakah sinyal emosional dapat meningkatkan performa klasifikasi pada model *transformer* modern yang sudah kuat secara representasi. Perbedaan lain terletak pada aspek metode yang digunakan, yaitu menggunakan pendekatan *unsupervised self-learning* untuk mengevaluasi kemampuan adaptasi model terhadap perubahan bahasa lintas waktu, sesuatu yang belum menjadi fokus dalam penelitian terdahulu. Selain itu, penelitian ini mengombinasikan dua sudut analisis, yaitu pengaruh emosi dan pengaruh perbedaan distribusi bahasa pada dataset lama dan baru, sehingga menghasilkan kontribusi baru dalam meningkatkan *robustness* model IndoBERT terhadap dinamika linguistik.

3. Metodologi

Penelitian ini menggunakan tiga skenario untuk menganalisis perbedaan performa dari setiap metode dan kondisi. Dua sumber utama dataset yang digunakan berasal dari dataset publik tahun 2019, masing-masing emosi dan ujaran kebencian, untuk dataset emosi diambil dari

repositori github *IndoNLU* [21]. Dataset ini diberi label dengan menggunakan metode *Shaver Basic Emotion*, yang terdiri dari 5 emosi utama di antaranya *anger*, *sadness*, *happy*, *fear*, dan *love*. Input yang digunakan hanya berupa satu kolom tweet. Dataset ini sudah dipisah antara data *train*, *validation*, dan *test* oleh pengembang, yang menjadi tiga bagian, dengan proporsi 80%, 10%, 10%, yang masing-masing berjumlah 3521 data latih, 440 data validasi, dan 440 data uji. Untuk *dataset hate speech* diambil dari repositori *id-multi-label-hate-speech-and-abusive-language-detection* [22], proses anotasi dataset ini dilakukan melalui dua tahap, yang pertama setiap tweet akan diberi label *hate speech*, *abusive*, atau *neither* oleh tiga anotator dengan latar belakang pengguna Twitter berusia 20-30 tahun, penutur asli bahasa Indonesia. Kemudian pada tahap kedua, tweet yang diidentifikasi sebagai *hate speech* akan dilakukan anotasi yang lebih detail oleh tiga anotator untuk menentukan targetnya (individu atau kelompok), dan kategorinya (agama, ras, fisik, atau gender), serta tingkat keparahannya (lemah, sedang, atau kuat). Pedoman yang digunakan sebagai landasan dalam anotasi ini adalah *Buku Saku Penanganan Ujaran Kebencian KomNas HAM (2015)*, dan divalidasi melalui *Focus Group Discussion* Bersama Direktorat Tindak Pidana Siber Bareskrim Polri. Hasil akhir ditentukan dengan *majority voting*, dan yang tidak mencapai kesepakatan minimal akan dibuang dari dataset. Dalam penelitian ini, label pada dataset tersebut akan disederhanakan menjadi dua kelas, yaitu HS (*Hate Speech*) apabila salah satu kolom *hate speech* bernilai 1, dan Non-HS (*Non Hate-Speech*) apabila seluruh kolom *hate speech* bernilai 0. Dataset asli berjumlah 13.170 data, namun untuk keperluan penelitian ini hanya digunakan sebanyak 3000 data agar sebanding dengan dataset lainnya. Pembagian data dilakukan dengan proporsi 80% data latih (2400 data), 10% data validasi (300 data), dan 10% data uji (300 data).

Untuk *dataset* 2025 dikumpulkan secara mandiri dengan melakukan *scrapping* pada platform Twitter menggunakan *selenium*. *Scrapping* dilakukan dengan beberapa *query*, *query* ini dipisah menjadi 2 kategori yaitu kategori yang mayoritas menyuarakan *hate speech*, dan yang netral. Contoh *query* yang mayoritas menyuarakan *hate speech* ini salah satunya adalah "(dasar OR tolol OR idiot) lang:id", di sisi lain untuk yang netral contohnya seperti "(hari ini OR cuaca OR selamat pagi OR selamat malam) lang:id". *Query* yang digunakan masing-masing untuk *hate speech* dan netral ada 8 *query* yang kemudian dari proses scrapping didapat 14523, kemudian akan dilakukan tahap preprocessing dengan mengapus teks yang duplikat untuk mencegah data berulang mengganggu keseimbangan distribusi data. Lalu teks akan diubah menjadi huruf kecil, dan simbol simbol seperti URL, *mention* (@) dan *hashtag* (#) akan dihapus, namun teks dibelakang *hashtag* akan dipertahankan. Emoji juga akan dikonversi menjadi deskripsi teks menggunakan *demojize* sehingga sinyal emosi tetap terjaga sebelum akhirnya akan dihapus pada tahap konversi karakter. Pembersihan karakter dilakukan dengan menghapus tanda baca, menghilangkan karakter non-alfanumerik berlebih, menormalkan spasi ganda, serta melakukan normalisasi variasi pada kata yang tidak baku menggunakan kamus singkatan (misalnya “gk/ga/gak/g” menjadi “tidak”, “bgt” menjadi “banget”, “skrng” menjadi “sekarang” dab berbagai bentuk kata informal lainnya). Setelah pembersihan dasar, kemudian dilakukan seleksi teks berdasarkan panjang teks, teks yang memiliki kurang dari tiga kata setelah proses pembersihan akan dihapus dari dataset. Kemudian untuk mencegah adanya tweet yang merupakan span atau iklan masuk kedalam dataset maka akan dilakukan pembersihan lebih lanjut dengan menggunakan kata kunci komersial seperti (“jual”, “promo”, “open bo”, “shopee” dan lainnya). Dari total data bersih, 3000 tweet akan dipilih untuk digunakan dalam penelitian karena proses anotasi manual baru dilakukan pada sebagian data. Dari fitur emosi di anotasi menggunakan *Shaver Basic Emotion* sebagai dasar teorinya yang terdiri dari 5 emosi mayor yaitu, *joy*, *anger*, *fear*, *love*, dan *sadness*. Di sisi lain untuk anotasi *hate speech* akan menggunakan *Surat Edaran Kapolri Tentang Penanganan Ujaran Kebencian (Hate Speech) dalam Kerangka Hak Asasi Manusia*, yang berfokus dalam identifikasi apakah suatu teks mengandung ujaran kebencian atau tidak. Data yang sudah diberi label kemudian akan dibagi menjadi data latih, validasi, dan uji dengan proporsi yang sama seperti dataset lainnya yaitu 80% data latih, 10% data validasi, dan 10% data uji.

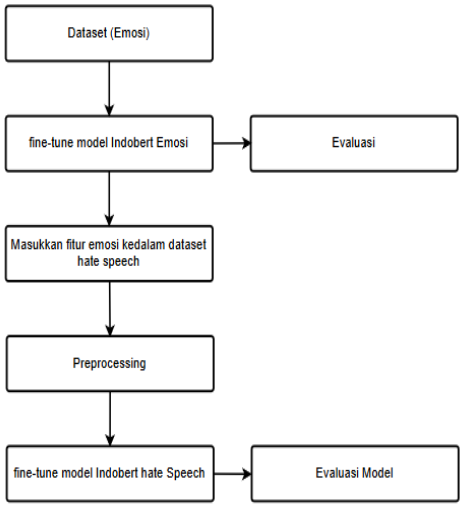
Untuk menjaga konsistensi dari jumlah data yang digunakan pada setiap skenario eksperimen, baik itu dataset 2019, maupun 2025 masing-masing akan dibatasi sebanyak 3000 data. Dengan demikian, setiap model akan dilatih dan diuji menggunakan jumlah data yang seimbang, sehingga perbandingan performa antar skenario dapat dilakukan secara adil, objektif, dan terukur.

Untuk memberikan gambaran yang lebih jelas mengenai karakteristik dari kedua dataset yang digunakan, yaitu dataset tahun 2019 dan dataset tahun 2025, detail dari dataset-dataset tersebut dapat dilihat pada Tabel 1.

Tabel 1. Tabel Perbandingan Karakteristik Dataset Tahun 2019 dan 2025

Aspek	Dataset 2019 (Publik)	Dataset 2025 (Mandiri)
Sumber Data	Repository publik IndoNLU (emosi), dan id-multi-label-hate-speech-and-abusive-language (hate speech)	Hasil *scrapping* Twitter menggunakan Selenium
Jenis Data	Tweet dalam bahasa Indonesia	Tweet dalam bahasa Indonesia
Jumlah Data Awal	13.170 (*hate speech*), 4.401 (emosi)	14.523 (hasil *scrapping* mentah)
Jumlah Data Setelah *Preprocessing*	3000 data digunakan untuk penelitian	3000 data digunakan untuk penelitian
Skema Label Emosi	*Shaver Basic Emotion: anger, sadness, happy, fear, love*	*Shaver Basic Emotion: anger, sadness, happy, fear, love*
Skema Label *Hate Speech*	Berdasarkan Buku Saku KomNas HAM (2015): *Hate Speech* vs *Non-Hate Speech*	Berdasarkan Buku Saku KomNas HAM (2015): *Hate Speech* vs *Non-Hate Speech*
Proses Anotasi	Dilakukan oleh 3 anotator independen, dengan *majority voting*	Dilakukan mandiri, yang kemudian divalidasi oleh ahli bahasa Indonesia
Proporsi Data (*Train/Validation/Test*)	80%/10%/10% (2.400/300/300)	80%/10%/10% (2.400/300/300)
Jumlah Kelas	2 kelas: HS, Non-HS (*hate speech*); 5 kelas emosi	2 kelas: HS, Non-HS (*hate speech*); 5 kelas emosi

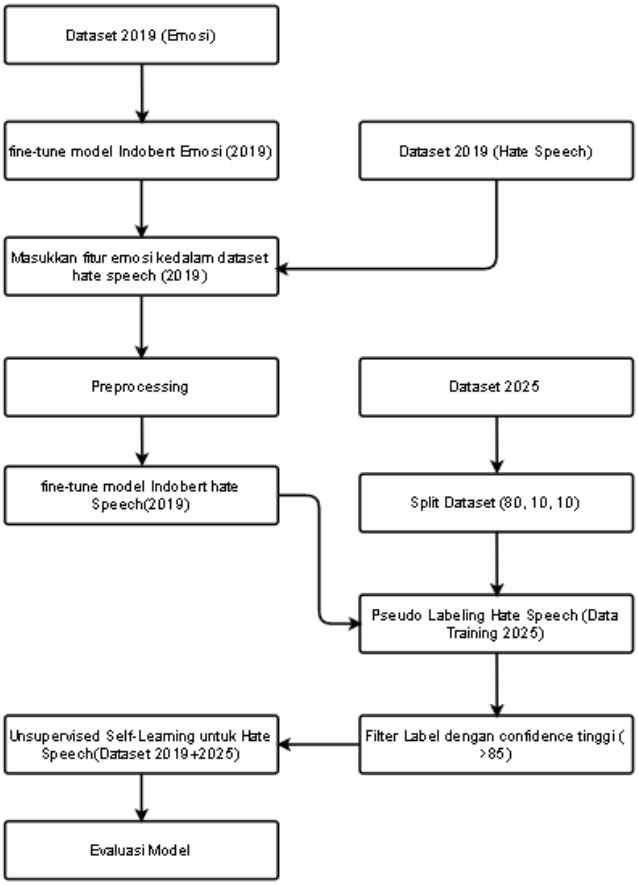
Setelah tahap pengumpulan dan persiapan dataset selesai dilakukan, Langkah berikutnya adalah menjelaskan metode yang digunakan dalam penelitian ini. Bagian ini membahas secara rinci arsitektur sistem, tahapan pelatihan model, serta pendekatan yang diterapkan untuk masing-masing skenario eksperimen, baik secara *supervised*, dan juga *unsupervised*.



Gambar 1. Arsitektur *training* model 2019

Urutan dari arsitektur pertama untuk melatih model 2019 secara *supervised* dapat dilihat pada Gambar 1. yang diawali dengan *load dataset* emosi. Berikutnya karena dataset emosi sudah

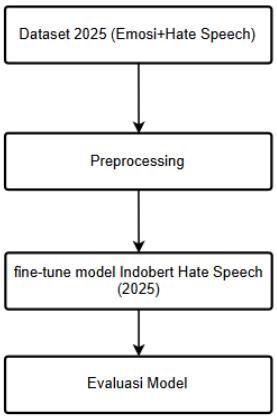
di *preprocess* dan juga di *split* menjadi data *train*, *validate*, dan *test* maka tidak perlu dilakukan tahap *preprocess*. Setelah itu akan dilakukan proses *fine-tune* model menggunakan *pretrained* model IndoBERT untuk mendeteksi emosi. Model *fine-tune* yang sudah selesai dibuat kemudian akan di evaluasi untuk mengetahui performa dari model ini, setelah itu model *fine-tune* ini akan digunakan untuk melakukan proses *labeling* emosi di *dataset hate speech*. Setelah fitur emosi dimasukkan ke dalam dataset melalui proses prediksi dari model *fine-tuned* emosi, maka proses *training* model untuk deteksi *hate speech* dengan tambahan fitur emosi dapat dilakukan.



Gambar 2. Arsitektur *self-training* model 2025

Urutan dari arsitektur untuk melatih model 2025 secara *unsupervised* menggunakan *pretrained* model dari 2019 dapat dilihat pada Gambar 2. Yang diawali dengan proses *training* model 2019 sampai menjadi model *hate speech*. Kemudian model *fine-tune hate speech* ini akan digunakan untuk *pseudo labeling* terhadap dataset 2025. *Pseudo labeling* ini adalah proses *labeling* data *train* sementara yang menggunakan model 2019 untuk *labeling* di dataset 2025. Sebelum *pseudo labeling* dilakukan, dataset akan di *split* terlebih dahulu menjadi 80% data *train*, 10% data *validasi*, dan 10% data *test*. Kemudian label *hate speech* di data *train* akan ditutup untuk melihat performa model dalam belajar secara *unsupervised*, untuk label di dataset *validation*, dan *test*, akan dibiarkan sebagai nilai dari *ground truth* dalam penilaian performa modelnya. lalu kemudian setelah *dataset* di *split*, akan dilakukan *pseudo labeling*, kemudian akan dilihat persentase *confidence* dari tiap-tiap label hasil dari *pseudo labeling*, label yang memiliki Tingkat *confidence* di bawah *threshold* (0.85) akan dihapus labelnya menjadi *unlabeled* lagi, lalu kemudian proses *training* dapat dimulai. Proses *training* dilakukan dengan cara iterasi, yang di mana dataset yang memiliki *confidence* tinggi akan digabungkan dengan dataset 2019 dan akan digunakan di setiap iterasi untuk membangun model. Model yang sudah dibentuk setelah satu iterasi selesai akan melakukan *pseudo labeling* kembali terhadap dataset 2025 yang label sebelumnya dihapus karena tidak mencapai *threshold*, kemudian akan dilihat lagi hasil prediksinya. Apabila tingkat *confidence* nya lebih tinggi dari *threshold* maka jumlah data *train*

akan ditambah, dan proses *training* akan dimulai lagi dengan data *train* yang sudah ditambah, dan seterusnya sampai iterasi yang ditentukan. Setelah proses *self-training* selesai, proses *testing* akan dilakukan dengan menggunakan data *test* 2025 yang berperan sebagai *ground truth* untuk melihat performa model.



Gambar 3. Arsitektur *training* model 2025

Urutan dari arsitektur untuk melatih model 2025 secara *supervised* dapat dilihat pada Gambar 3. yang akan diawali dengan *load* dataset 2025, didalam dataset ini sudah ada label emosi dan juga *hate speech*. Dataset tersebut kemudian akan dilakukan proses *cleaning* yang menghapus tweet yang duplikat, kosong, dan juga tweet yang terlalu pendek, dan juga akan dilakukan normalisasi bahasa. Kemudian akan di *split* menjadi 80% data *train*, 10% data *validasi* dan 10% data *test*. Kemudian dataset tersebut akan digunakan dalam proses *training fine-tune* menggunakan model *IndoBERT* untuk deteksi *hate speech*. Hasil dari *training* tersebut kemudian akan di evaluasi untuk melihat performa dari model tersebut. Setelah semua model selesai dibangun, akan dilakukan perbandingan antara model hasil dari arsitektur satu dengan yang lain untuk menentukan yang mana model yang memberikan performa terbaik dalam mendeteksi *hate speech*.

Input yang digunakan dalam penelitian ini berupa teks mentah dari tweet yang sudah melalui proses normalisasi, tokenisasi, dan padding sesuai dengan format IndoBERT *tokenizer*. Setiap teks diubah menjadi vektor *embedding* berdimensi 768, yang kemudian diproses oleh model untuk menghasilkan output berupa probabilitas kelas *Hate Speech (HS)*, atau *Non-Hate Speech (Non-HS)*.

Evaluasi performa akan dilakukan dengan menggunakan data uji (*test dataset*) yang tidak digunakan selama proses pelatihan. Pengukuran kinerja model dilakukan menggunakan empat metrik utama, yaitu Akurasi, *Precision*, *Recall*, dan *F1-Score*, yang dihasilkan melalui fungsi *classification_report* dari Pustaka *scikit-learn*. Validasi hasil dilakukan dengan membandingkan metrik performa antar skenario (*supervised* 2019, *supervised* 2025, dan *self learning* 2025) untuk menilai pengaruh penambahan fitur emosi serta efektivitas metode *self learning* dalam meningkatkan kemampuan generalisasi model.

4. Hasil dan Pembahasan

4.1 Hasil

Bab ini menyajikan hasil empiris dari seluruh rangkaian eksperimen yang dilakukan dalam penelitian ini, dimulai dari gambaran data yang digunakan, keluaran proses-proses inti seperti *preprocessing*, penambahan fitur emosi, hingga hasil evaluasi model pada berbagai skenario pengujian. Penyajian ini berfokus pada apa yang dihasilkan pada setiap proses, bukan pada prosedurnya sebagaimana yang telah dijelaskan pada bab Metodologi sebelumnya.

Data yang digunakan dalam penelitian ini terdiri dari tiga kelompok utama, yaitu dataset emosi 2019, dataset *hate speech* 2019, dan dataset *hate speech* 2025. Dataset emosi 2019 berjumlah 4401 data dengan lima kategori emosi yaitu *anger, fear, joy, sadness,* dan *love*. Dataset emosi ini tidak melalui tahap *preprocessing* lagi karena telah dibersihkan dan dibagi menjadi data *train*, *validation*, dan *test* oleh pembuat dataset. Beberapa sampel data dari dataset emosi dapat dilihat pada Tabel 2.

Tabel 2. Dataset emosi 2019

Teks	Label
Romantis itu MANIS lebih manis lagi kalo dilakuin orang tg kita SAYANG semua bakal terasa seneng dan berbunga bunga wkwk seperti ada ribuan kupu kupu terbang yg ada dii perut kalo lagi ada yang ngeromantisin gitu	love
gue yg dulu bingung knp org bisa trauma berkepanjangan skrg baru sadar setelah ngerasain sendiri, kita ga speak up kdg krn takut dibilang lebay lah boong lah bla bla ngatasin traumanya aja udh susah tmbh lg di judge org lain jd mending diem	fear
sama kayak bahagia mau gimanapun persepsi orang tentang bahagia kalau bahagiamu memang beda dari persepsi orang2 itu tetap bahagia kamu orang lain gak ngerasain tapi kamu yg ngerasain ttg bahagia kamu	joy
ini kakinya diapain sih harusnya ya ampun sakit banget nyeri orang rumah udah pada jengah kali ya aku aduh-aduh mulu gara2 serangan mendadak di kaki	sadness
bicara asal ngejiplak saja sok tahu sangat sih	anger
tapi iya sih dosa lo banyak sama gue tapi aku tetep cinta kok muahh	love

Dataset *hate speech* 2019 berjumlah 13.170 data. Sampel dari dataset *hate speech* 2019 dapat dilihat pada Tabel 3.

Tabel 3. Dataset *hate speech* 2019

Teks	Label
- disaat semua cowok berusaha melacak perhatian gue. loe lantas remehkan perhatian yg gue kasih khusus ke elo. basic elo cowok bego ! ! !'	HS
Jgn lupa juga kang.... Umat katolik dan minorutas lainnya diperhatika. Gubernur untuk semua golongan bukan satu golongan.'	Non-HS
Dari habis sahur sampe jam 10. Sibayik udah nete 4x. Skg rasanya lemas tak terkira \xf0\x9f\xa4\xa2'	Non-HS

Untuk sampel dari dataset *hate speech* 2025 dapat dilihat pada Tabel 4. Dataset 2025 sudah memiliki label emosi didalamnya sehingga tidak perlu menggunakan *pretrained* model emosi untuk melakukan labeling nya.

Tabel 4. Dataset *hate speech* 2025

Teks	Label	Emotion
Cok goblok, mouse aja kok pajaknya 8,9 juta, kalian menilainya berapa puluh juta? Bayar banyak malah tololl	HS	Anger
sangattt sangat bersyukur dan excited di setiap hal kecil yg terjadi di hidupku!! Aku sangat happy ketika belanja trus ga perlu antri banyak, ketikaa aku bisa makan enak... Trus pas di kasih sesuatu tuh kaya, omg I'm so lucky? Trus ketika ada org yg cerita--	Non-HS	Joy
Kewajiban orang tua memberikan makanan bergizi 3x sehari buat anak2nya DIRAMPAS OLEH NEGARA menjadi 1x sehari yg berakhir menjadi TAIK. Semestinya TUGAS NEGARA MENYEDIAKAN LAPANGAN PEKERJAAN. 1 lapangan pekerjaan bisa memberi makan 1 keluarga 3x sehari seumur hidup	Non-HS	Anger

Penelitian ini diawali dengan melakukan *preprocessing* terhadap dataset *hate speech* 2019 dan dataset *hate speech* 2025. Dataset emosi 2019 tidak perlu dilakukan *preprocessing* karena datasetnya sudah dibersihkan dan dibagi menjadi *train*, *validate*, dan *test* oleh pembuat dataset. *Preprocessing* dilakukan sesuai dengan yang dibahas pada bagian Metodologi dengan menghapus teks duplikat, teks yang terlalu pendek, membuat teks menjadi huruf kecil, menghapus simbol-simbol, dan juga menghapus teks yang berunsur spam atau iklan. Untuk contoh data hasil praproses dapat dilihat pada Tabel 5.

Tabel 5. Hasil *Preprocessing* dataset hate speech 2019 dan 2025

Teks	Hasil *preprocessing*
Cok goblok, mouse aja kok pajaknya 8,9 juta, kalian menilainya berapa puluh juta? Bayar banyak malah tololl	cok goblok mouse saja kok pajaknya 8 9 juta kalian menilainya berapa puluh juta bayar banyak malah tololl
sangattt sangat bersyukur dan excited di setiap hal kecil yg terjadi di hidupku!! Aku sangat happy ketika belanja trus ga perlu antri banyak, ketikaa aku bisa makan enak... Trus pas di kasih sesuatu tuh kaya, omg I'm so lucky? Trus ketika ada org yg cerita--	sangattt sangat bersyukur dan excited di setiap hal kecil yang terjadi di hidupku aku sangat happy ketika belanja terus tidak perlu antri banyak ketika aku bisa makan enak terus pas di kasih sesuatu itu kaya omg i m so lucky terus ketika ada orang yang cerita
China cina vs jubir morowali hgjviral 7 menit doodstream terbaru dood really Ometv tunjangan #evanurasyifa #izzafadhila #amaliamutya #amaliamutia	Teks dihapus dari dataset karena berunsur iklan
Iyakan sistt... Wkwkw. Akutuh udah di tahap ngerasa malah Nam Yejun itu ada di Korea sana. Mereka cerita ttg makanan korea, gym dan pakai bahasa cina itu kek---REAL HIDUP GITUU	iyakan sistt wkwkw akutuh udah di tahap ngerasa malah nam yejun itu ada di korea sana mereka cerita ttg makanan korea gym dan pakai bahasa cina itu kek real hidup gituu
Kewajiban orang tua memberikan makanan bergizi 3x sehari buat anak2nya DIRAMPAS OLEH NEGARA menjadi 1x sehari yg berakhir menjadi TAIK. Semestinya TUGAS NEGARA MENYEDIAKAN LAPANGAN PEKERJAAN. 1 lapangan pekerjaan bisa memberi makan 1 keluarga 3x sehari seumur hidup	kewajiban orang tua memberikan makanan bergizi 3x sehari buat anak2nya dirampas oleh negara menjadi 1x sehari yang berakhir menjadi taik semestinya tugas negara menyediakan lapangan pekerjaan 1 lapangan pekerjaan bisa memberi makan 1 keluarga 3x sehari seumur hidup

Setelah tahap *preprocess*, teks-teks tersebut kemudian siap untuk digunakan dalam proses *training*. Pada skenario *supervised* 2019, model *fine tuned* emosi akan digunakan untuk memprediksi emosi yang terkandung didalam teks-teks di dataset *hate speech* 2019. Model deteksi emosi tersebut memiliki performa akurasi 72.27%, *precision* 73.51%, *recall* 72.27%, dan *F1-score* 72.12%. Setelah prediksi tersebut dilakukan, dataset *hate speech* 2019 memiliki 2 *input* yaitu *tweet* dan juga *emotion*, dan 1 *output* yaitu *label* dari *hate speech* itu sendiri. Sampel dari dataset *hate speech* 2019 yang sudah ditambahkan fitur emosi dapat dilihat pada Tabel 6.

Tabel 6. Dataset *hate speech 2019* dengan fitur emosi

Teks	*Emotion*	Label
disaat semua cowok berusaha melacak perhatian gue loe lantas remehkan perhatian yang gue kasih khusus ke elo basic elo cowok bego	*anger*	HS
jangan lupa juga kang umat katolik dan minoritas lainnya diperhatika gubernur untuk semua golongan bukan satu golongan	*anger*	Non-HS
dari habis sahur sampe jam 10 sibayik sudah nete 4x skg rasanya lemas tak terkira	*sadness*	Non-HS

Setelah fitur emosi dimasukkan kemudian dataset 2019 ini akan diambil 3000 data untuk menjaga konsistensi jumlah data dalam proses *training*, karena dataset 2025 hanya memiliki 3000 data. Pengambilan 3000 data ini dilakukan dengan mempertahankan keseimbangan kedua label output baik itu HS (1500 data), dan Non-HS (1500 data) supaya proses *training* tidak jadi berat sebelah. Setelah fitur emosi dimasukkan dan dataset dipotong menjadi 3000 data, maka bentuk dari dataset 2019 dan juga 2025 sudah sama yang membuat performa model dapat dievaluasi dengan lebih akurat.

Fase pelatihan dilakukan dengan konfigurasi *hyperparameter* yang konsisten di seluruh skenario untuk menjaga *fairness* perbandingan. Proses *training* menggunakan *batch size* 16,

learning rate 3e-5, *weight decay* 0.01, dan lima *epoch* pelatihan. Optimasi dilakukan dengan AdamW, sementara skema pembelajaran dikendalikan oleh *linear warmup scheduler* dengan porsi *warmup* 10% dari total langkah pelatihan. Pemilihan konfigurasi ini bertujuan menjaga stabilitas *fine-tuning* model *transformer* sekaligus menghindari *overfitting*, mengingat ukuran dataset 2019 dan 2025 memiliki karakter yang berbeda.

Dalam proses *training supervised* 2019, dataset *hate speech* 2019 yang sudah ditambahkan fitur emosi akan di split dengan proporsi 80% data train (2400 data), 10% data *validation* (300 data), dan 10% data *test* (300 data). Kemudian evaluasi akan dilakukan untuk melihat performa model dengan menggunakan fitur emosi, dan tanpa menggunkaan fitur emosi. Hasil evaluasi menunjukkan bahwa model 2019 tanpa fitur emosi dan diuji pada data uji tahun yang sama menghasilkan akurasi sebesar 88%, *precision* 88.62%, *recall* 86.87%, dan *F1-Score* 87.46%. Ketika fitur emosi ditambahkan pada dataset 2019, performa model meningkat dengan akurasi 89%, *precision* 88.90%, *recall* 89.83%, dan *F1-Score* 88.92%. Hasil ini menunjukkan bahwa penambahan fitur emosi memberikan kontribusi positif terhadap peningkatan kinerja model dalam mendeteksi ujaran kebencian.

Proses *training supervised* 2025 juga mengikuti prosedur yang sama dengan *supervised* 2019, hanya saja dataset *hate speech* 2025 sudah memiliki fitur emosi didalamnya sehingga tidak perlu menggunakan model emosi untuk melakukan labeling. Evaluasinya mendapatkan hasil bahwa model yang dilatih dan diuji menggunakan dataset 2025 tanpa fitur emosi memperoleh akurasi 90.67%, *precision* 90.72%, *recall* 90.16%, dan *F1-Score* 90.40%. Ketika fitur emosi ditambahkan, performa model kembali meningkat menjadi akurasi 91.67%, *precision* 91.37%, *recall* 91.73%, dan *F1-Score* 91.52%. konsistensi peningkatan performa pada kedua periode data ini memperkuat bukti bahwa emosi berperan dalam membantu model memahami konteks ujaran kebencian secara lebih baik, terutama pada teks dengan ambiguitas tinggi.

Setelah dilakukan evaluasi pada data uji di tahun yang sama, langkah berikutnya adalah melakukan *cross-evaluation* untuk mengukur kemampuan generalisasi model terhadap perubahan data lintas waktu. Model 2019 yang diuji pada dataset 2025 tanpa fitur emosi menghasilkan akurasi 72.33%, *precision* 75.94%, *recall* 68.78%, dan *F1-Score* 68.64%. sementara itu dengan penambahan fitur emosi, performa model meningkat dengan akurasi 76.33%, *precision* 75.88%, *recall* 75.46%, dan *F1-Score* 75.63%. hasil ini menunjukkan bahwa performa model menurun secara signifikan ketika diuji pada data di dimensi waktu yang berbeda. Fenomena ini membuktikan bahwa perubahan bahasa, istilah, serta gaya komunikasi dari waktu ke waktu dapat mempengaruhi kinerja model deteksi ujaran kebencian, sehingga model yang dilatih pada satu periode tertentu belum tentu optimal untuk periode berikutnya.

Dari penurunan performa dalam *cross-evaluation* tersebut, maka digunakan metode *unsupervised self-learning* dengan memanfaatkan model deteksi *hate speech* tahun 2019 sebagai *teacher* model awal. Model ini digunakan untuk memberikan *pseudo label* terhadap dataset 2025, kemudian dilakukan proses pelatihan ulang menggunakan data berlabel otomatis tersebut. Pada tahap ini digunakan *threshold* 85% untuk memastikan bahwa sampel dengan tingkat keyakinan tinggi yang masuk sebagai data pelatihan baru. Proses *self-learning* dibatasi sebanyak 15 iterasi dan pada setiap iterasi, jumlah setiap *pseudo-label* yang lolos meningkat seiring stabilnya prediksi emosi, namun jumlah *pseudo-label* yang lolos stabil pada iterasi ke-7.

Hasilnya menunjukkan bahwa dengan fitur emosi, model dapat mencapai akurasi 77.67%, *precision* 77.20%, *recall* 77.03%, dan *F1-Score* 77.10%. performa ini meningkat dibandingkan dengan model 2019 yang dilatih secara *supervised* dan langsung diuji pada dataset 2025. Sebaliknya, tanpa fitur emosi, metode *unsupervised* tidak menunjukkan peningkatan yang berarti dan cenderung menurun dibandingkan dengan model *supervised* sebelumnya, berdasarkan hasil ini, model terbaik yang digunakan tetap berasal dari pelatihan supervised tahun 2019. Hal ini menegaskan bahwa fitur emosi memiliki peran penting dalam meningkatkan kemampuan adaptasi model terhadap perubahan data lintas waktu khususnya pada skenario *unsupervised self-training*.

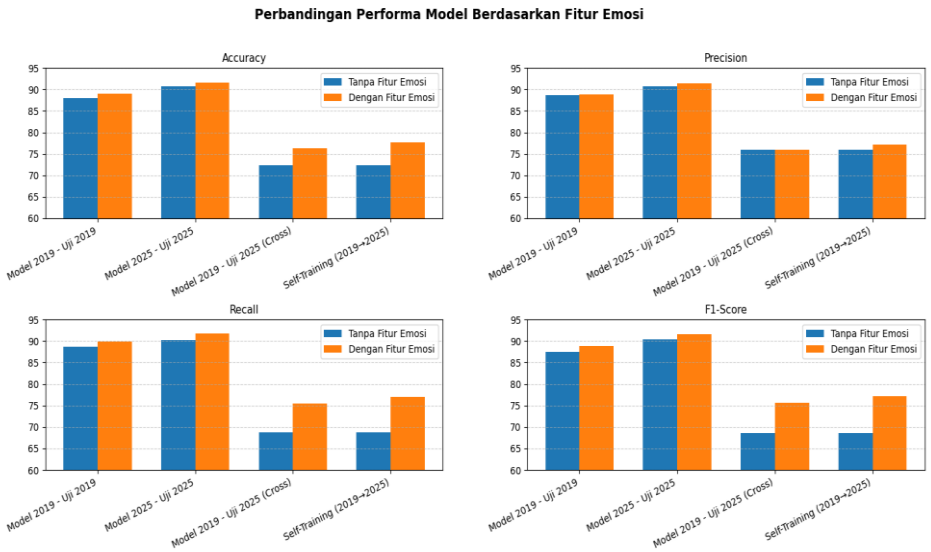
Tabel 7 berikut merangkum hasil pengujian dari seluruh skenario yang dilakukan. Tabel ini menampilkan perbandingan hasil dengan dan tanpa menggunakan fitur emosi berdasarkan metrik evaluasi utama, yaitu *accuracy*, *precision*, *recall*, dan *F1-Score*.
Dari ketiga skenario pengujian yang dilakukan, diperoleh hasil sebagai berikut:

Tabel 7. Tabel Hasil Pengujian Model Deteksi Ujaran Kebencian

Skenario Pengujian	Fitur Emosi	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Model 2019 - Uji 2019	Tidak	88.00	88.62	88.67	87.46
Model 2019 - Uji 2019	Ya	89.00	88.90	89.83	88.92
Model 2025 - Uji 2025	Tidak	90.67	90.72	90.16	90.40
Model 2025 - Uji 2025	Ya	91.67	91.37	91.73	91.52
Model 2019 – Uji 2025 (Cross)	Tidak	72.33	75.94	68.78	68.64
Model 2019 - Uji 2025 (Cross)	Ya	76.33	75.88	75.46	75.63
Self-Training (model 2019 - dataset 2025)	Tidak	72.33[1]	75.94[1]	68.78[1]	68.64[1]
Self-Training (model 2019 - dataset 2025)	Ya	77.67	77.20	77.03	77.10

[1]Tidak menunjukkan peningkatan dibandingkan model supervised 2019 terhadap data uji 2025

Tabel 7 menampilkan hasil evaluasi performa model deteksi ujaran kebencian pada berbagai skenario pengujian, baik dengan maupun tanpa penambahan fitur emosi. untuk mempermudah interpretasi performa antar skenario, hasil dari Tabel 2 akan divisualisasikan dalam grafik yang dapat dilihat pada Gambar 4.



Gambar 4. Grafik Hasil Pengujian Model Deteksi Ujaran Kebencian

4.2 Pembahasan

Hasil penelitian menunjukkan pola yang sangat jelas bahwa penambahan fitur emosi memberikan kontribusi substansial terhadap peningkatan performa model deteksi ujaran kebencian, baik pada pengujian intra waktu maupun lintas waktu sebagaimana yang dapat dilihat pada Gambar 4. Temuan ini memperlihatkan bahwa informasi emosional tidak hanya menjadi fitur pelengkap, tetapi justru berperan sebagai sinyal kognitif yang memperkuat pemahaman model terhadap konteks suatu teks. Pada data 2019 maupun 2025, model yang dilengkapi fitur emosi mampu mengidentifikasi ujaran kebencian dengan lebih presisi, terutama pada kasus-kasus yang sebelumnya ambigu atau memiliki struktur bahasa yang tidak secara eksplisit mengandung kata-kata kasar. Dengan kata lain, emosi bukan sekedar metadata, namun merupakan jembatan antara makna literal dan intensi psikologis suatu teks.

Jika memperhatikan perilaku model pada skenario intra waktu, peningkatan performa yang muncul setelah penambahan fitur emosi menandakan bahwa sinyal emosional mengisi celah yang tidak dapat dijangkau oleh fitur linguistik murni. Banyak data dalam kedua dataset memperlihatkan fenomena ujaran kebencian yang tidak dapat dikenali hanya dari permukaan

teks. Di sinilah peran fitur emosi menjadi sangat penting. Ketika model memproses teks yang bernada negatif tanpa kata kasar yang eksplisit, label emosi seperti *anger* atau *sadness* memberikan petunjuk tambahan mengenai intensi penulis teks. Temuan ini menguatkan bahwa ujaran kebencian sering disampaikan melalui nada emosional yang intens. Fakta bahwa akurasi dan *F1-score* meningkat secara konsisten setelah penambahan fitur emosi menjadi bukti empiris bahwa pola tersebut dapat dimodelkan secara sistematis.

Penurunan performa yang terjadi pada *cross-evaluation* antara model 2019 dan dataset 2025 mengungkapkan salah satu temuan kunci penelitian ini, yang dimana model deteksi ujaran kebencian sangat sensitif terhadap pergeseran bahasa lintas waktu. Perubahan istilah populer, gaya komunikasi, dan struktur dari teks digital dari tahun 2019 ke 2025 menyebabkan model yang dilatih pada data lama kesulitan mengikuti representasi bahasa yang lebih baru. Kondisi ini menjadi bukti konkret bahwa pembaruan dataset atau adaptasi model bukan sekedar kebutuhan opsional, melainkan keharusan. Dalam konteks ini, fitur emosi kembali menunjukkan perannya sebagai mekanisme kompensasi. Meskipun akurasi tetap menurun dibandingkan pengujian intra waktu, penambahan emosi berhasil menahan degradasi performa dan bahkan meningkatkan hasil hingga 4% dibandingkan model tanpa emosi. Artinya sinyal emosinya bekerja sebagai representasi yang lebih stabil dari waktu ke waktu dibandingkan fitur linguistik yang cenderung berubah mengikuti tren bahasa.

Temuan penting lainnya muncul dari eksperimen *self-learning*. Ketika model 2019 digunakan sebagai *teacher* untuk menghasilkan *pseudo-label* pada data 2025, performa model tanpa fitur emosi tidak mengalami peningkatan yang berarti. Bahkan dalam beberapa metrik, performanya cenderung stagnan atau menurun. Fenomena ini menunjukkan bahwa tanpa bantuan fitur yang lebih stabil, *pseudo-label* yang dihasilkan model lama menjadi kurang akurat untuk data baru. Namun ketika fitur emosi disertakan, proses *self-learning* menjadi lebih efektif. Model mampu memilih sampel *pseudo-label* dengan keyakinan tinggi sehingga lebih konsisten, yang berujung pada peningkatan performa akhir. Fakta bahwa peningkatan berhenti pada iterasi ke-7 memberikan indikasi bahwa model telah mencapai titik stabil, dan bahwa fitur emosi berperan dalam mempercepat proses konvergensi. Hal ini menegaskan bahwa emosi tidak hanya bermanfaat dalam skenario *supervised*, tetapi juga memperluas kegunaannya ke ranah *semi-supervised*.

Secara keseluruhan hasil penelitian ini menunjukkan bahwa penambahan fitur emosi mampu secara konsisten meningkatkan performa model, baik pada pengujian intra-waktu maupun lintas waktu. Selain itu, metode *self-learning* memberikan peningkatan moderat ketika fitur emosi digunakan, menunjukkan potensi pendekatan *semi-supervised* untuk adaptasi data baru. Temuan ini sejalan dengan penelitian terdahulu yang menekankan pentingnya konteks emosional dalam memahami ujaran kebencian [8], serta menunjukkan bahwa kombinasi fitur lingustik dan afektif dapat memperkuat kemampuan generalisasi model berbasis bahasa alami. Dengan demikian, pendekatan yang diusulkan pada penelitian ini telah berhasil menjawab permasalahan utama yang diidentifikasi di awal, yaitu meningkatkan ketahanan model terhadap variasi bahasa dan konteks temporal.

Hasil yang diperoleh melalui pengujian ini menunjukkan pola yang konsisten bahwa penambahan fitur emosi tidak hanya berfungsi sebagai atribut tambahan, tetapi berperan sebagai sinyal konteks yang mampu memperbaiki sensitivitas model terhadap teks berunsur ujaran kebencian. Peningkatan yang muncul pada skenario intra-waktu (2019-2019 dan 2025-2025) mengindikasikan bahwa sinyal emosi efektif dalam membantu model membedakan ujaran yang secara linguistik mirip, namun berbeda intensitas emosinya, sebuah fenomena yang banyak terjadi pada ujaran sarkastis atau implisit. Temuan sejalan serta memperkuat penelitian sebelumnya memanfaatkan IndoBERT untuk deteksi *hate speech* [19] namun belum mengintegrasikan fitur emosi, serta mendukung temuan studi lain yang menunjukkan bahwa informasi emosional dapat meningkatkan performa deteksi ujaran kebencian *[23].* Hasil penelitian ini juga konsisten dengan karya *[9]* yang sama-sama menunjukkan bahwa penggunaan fitur emosi memberikan kontribusi nyata dalam meningkatkan ketepatan model dalam mengidentifikasi *hate speech*.

5. Simpulan

Berdasarkan hasil penelitian dan pengujian yang dilakukan, dapat disimpulkan bahwa penambahan fitur emosi pada model deteksi ujaran kebencian berbasis IndoBERT memberikan pengaruh positif dan signifikan terhadap peningkatan performa model. Dibandingkan dengan

model tanpa fitur emosi, seluruh metrik evaluasi baik itu *accuracy*, *precision*, *recall*, dan *F1-score* menunjukkan peningkatan yang konsisten pada berbagai skenario pengujian.

Pada pengujian model 2019-Uji 2019 akurasi meningkat dari 88.00% menjadi 89.00%, sedangkan pada model 2025-Uji 2025 akurasi meningkat dari 90.67% menjadi 91.67%., di sisi lain pengujian lintas waktu (Model 2019-Uji 2025) menunjukkan penurunan performa akibat perbedaan konteks linguistik, istilah populer, serta gaya komunikasi antar periode. Meskipun demikian, penggunaan fitur emosi juga berhasil dalam meningkatkan performa model *unsupervised self-learning* yang akurasinya 72.33% meningkat menjadi 77.67%. Peningkatan serupa juga terlihat pada seluruh metrik lainnya (*precision*, *recall*, dan *F1-Score*) dengan rata-rata kenaikan sekitar 1-1.5%. hal ini menunjukkan bahwa informasi emosional dalam teks dapat membantu model memahami konteks ujaran kebencian dengan lebih baik, sehingga mampu membedakan ujaran kebencian dari teks netral secara lebih akurat. Secara keseluruhan, integrasi fitur emosi ini terbilang dapat meningkatkan performa model dalam mendeteksi ujaran kebencian secara lebih akurat, dengan demikian penelitian ini

Daftar Referensi

[1] A. W. Syakhrani, R.K. Amuntai and E. K. Widijatmoko, “Perkembangan Komunikasi Digital: Dampak Media Sosial Pada Interaksi Sosial di Era Modern,” *Jurnal Komunikasi*, vol. 2, no. 12, pp. 919–925, Dec. 2024.

[2] I. W. Zega, I. P.S.B. Purba, M. Iqbal, K. I. Ainurridho, Y. Madarusman, and R. S. Gueci, “Penggunaan Media Sosial yang Bijak Dalam Kebebasan Berekspresi dan Berpendapat,” *Abdi Laksana: Jurnal Pengabdian Kepada Masyarakat*, vol. 5, pp. 498–504, May 2024.

[3] S. Khodijah, N. Syifa, Y. Sembiring, N. R. Fauzan, and S. dan Teknologi, “Tinjauan Dampak Negatif Fenomena Kebencian di Media Sosial di Indonesia,” *Senashtek 2024*, vol. 2, no. 1, pp. 77–80, Jul. 2024.

[4] T. A. Azis *et al.*, “Hate Speech against Javanese in Social Media: A Case Study of Instagram Platform,” *Jurnal Ilmiah Multidisiplin*, vol. 4, no. 3, p. 46, May 2025, doi: 10.56127/jukim.v4i03.

[5] P. Madriaza *et al.*, “Exposure to hate in online and traditional media: A systematic review and meta-analysis of the impact of this exposure on individuals and communities,” *Campbell Systematic Reviews*, vol. 21, no. 1, p. e70018, Mar. 2025, doi: 10.1002/cl2.70018.

[6] F. Andy Kusuma and E. W. Pamungkas, “Pendeteksian Hate Speech Pada Sosial Media Indonesia dengan Algoritma Support Vector Machine (SVM) dan Decision Tree,” Tesis, Program Studi Teknik Informatika, Universitas Muhammadiyah, Surakarta, 2023.

[7] M. O. Ibrohim and I. Budi, “Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter,” in *Preceedings of the Third Workshop on Abusive Language Online*, Florence, Italy, Aug. 2019, pp. 46–57.

[8] A. Ghenai, Z. Noorian, H. Moradisani, P. Abadeh, C. Erentzen, and F. Zarrinkalam, “Exploring hate speech dynamics: The emotional, linguistic, and thematic impact on social media users,” *Inf Process Manag*, vol. 62, no. 3, p. 104079, May 2025, doi: 10.1016/j.ipm.2025.104079.

[9] K. Mnassri, P. Rajapaksha, R. Farahbakhsh, and N. Crespi, “Hate Speech and Offensive Language Detection using an Emotion-aware Shared Encoder,” in *ICC 2023 - IEEE International Conference on Communications*, Rome, Italy, Feb. 2023, pp. 2852–2857. doi: 10.1109/ICC45041.2023.10279690.

[10] F. M. Plaza-del-Arco, S. Halat, S. Padó, and R. Klinger, “Multi-Task Learning with Sentiment, Emotion, and Target Detection to Recognize Hate Speech and Offensive Language,” in *CEUR Workshop Proceedings*, Gandhinagar, India, Dec. 2021, pp. 297–318.

[11] R. Bagestra, A. Misbullah, Z. Zulfan, R. Rasudin, L. Farsiah, and S. A. Nazhifah, “Performance Assessment of Machine Learning and Transformer Models for Indonesian Multi-Label Hate Speech Detection,” *Infolitika Journal of Data Science*, vol. 2, no. 2, pp. 62–71, Nov. 2024, doi: 10.60084/ijds.v2i2.235.

[12] 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 *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain, Dec. 2020, pp. 757–770.

[13] Dhendra and V. G. Utomo, “Benchmarking IndoBERT and Transformer Models for Sentiment Classification on Indonesian E-Government Service Reviews,” *Jurnal Transformatika*, vol. 23, no. 1, pp. 86–95, Jul. 2025, doi: 10.26623/transformatika.v23i1.12095.

[14] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Preceedings of NAACL-HLT 2019*, Mineapolis, Minnesota, Jun. 2019, pp. 4171–4186.
[15] A. Glenn, P. LaCasse, and B. Cox, “Emotion classification of Indonesian Tweets using Bidirectional LSTM,” *Neural Comput Appl*, vol. 35, no. 13, pp. 9567–9578, May 2023, doi: 10.1007/s00521-022-08186-1.
[16] M. H. Algifari and D. Nugroho, “Emotion Classification of Indonesian Tweets using BERT Embedding,” *Journal of Applied Informatics and Computing (JAIC)*, vol. 7, no. 2, pp. 2548–6861, Dec. 2023.
[17] A. S. S. Ansyah, A. P. Kurniawan, A. N. Kholifah, and D. Purwitasari, “A Hybrid Method on Emotion Detection for Indonesian Tweets of COVID-19,” *Jurnal RESTI*, vol. 7, no. 2, pp. 254–262, Apr. 2023, doi: 10.29207/resti.v7i2.4816.
[18] A. C. Saputra *et al.*, “Prediksi Emosi Dalam Teks Bahasa Indonesia Menggunakan Model Indobert,” *Jurnal Teknologi Informasi : Jurnal Keilmuan Dan Aplikasi Bidang Teknik Informatika*, vol. 19, no. 1, pp. 1–15, Jan. 2025, doi: 10.47111/JTI.
[19] J. F. Kusuma and A. Chowanda, “International Journal On Informatics Visualization journal homepage : www.joiv.org/index.php/joiv International Journal On Informatics Visualization Indonesian Hate Speech Detection Using IndoBERTweet and BiLSTM on Twitter,” *International Journal on Informatics Visualization*, vol. 7, no. 3, pp. 773–780, Sep. 2023, [Online]. Available: www.joiv.org/index.php/joiv
[20] M. Usman, M. Ahmad, G. Sidorov, I. Gelbukh, and R. Q. Tellez, “A Large Language Model-Based Approach for Multilingual Hate Speech Detection on Social Media,” Multidisciplinary Digital Publishing Institute (MDPI), Jul. 2025. doi: 10.3390/computers14070279.
[21] IndoNLP Team, “IndoNLU — Emotion Twitter Dataset (emot_emotion-twitter).” Accessed: Aug. 04, 2025. [Online]. Available: https://github.com/IndoNLP/indonlu/tree/master/dataset/emot_emotion-twitter
[22] O. Ibirohim, “ID-MultiLabel Hate Speech and Abusive Language Detection Dataset.” Accessed: Aug. 04, 2025. [Online]. Available: https://github.com/okkyibrohim/id-multi-label-hate-speech-and-abusive-language-detection/commits/master/
[23] M. R. Awal, R. Cao, R. K.-W. Lee, and S. Mitrovic, “AngryBERT: Joint Learning Target and Emotion for Hate Speech Detection,” Mar. 2021, [Online]. Available: http://arxiv.org/abs/2103.11800