DETEKSI *HATE SPEECH* UNSUR SARA PADA KOMENTAR MEDIA SOSIAL MENGGUNAKAN PENDEKATAN *TWO-STAGE CLASSIFICATION* DENGAN ALGORITMA *INDOBERT* DAN *SUPPORT VECTOR MACHINE*

Ovy Marsya Zieera[1)], Monica Cinthya[2)]
[1)] Sistem Informasi, Fakultas Teknik, Universitas Negeri Surabaya, Surabaya, Indonesia
Email: ovy.22137@mhs.unesa.ac.id
[2)] Sistem Informasi, Fakultas Teknik, Universitas Negeri Surabaya, Surabaya, Indonesia
Email: monicacinthya@unesa.ac.id

Abstract
The rapid development of social media in Indonesia has increased public interaction on platforms such as YouTube, Instagram, and TikTok. However, this has also driven the proliferation of hate speech, particularly content containing elements of (SARA). This study proposes a Two-Stage Classification approach to address this challenge. In the first stage, the IndoBERT model (indobenchmark/indobert-base-p1) is fine-tuned to classify comments into Hate Speech and Non-Hate Speech. In the second stage, Support Vector Machine (SVM) with TF-IDF feature extraction and a custom SARA lexicon is used to further classify hate speech comments into SARA-based hate speech (HS_SARA) and general hate speech (HS_Umum). The dataset consists of 36,000 comments scraped from YouTube, Instagram, and TikTok on viral SARA-related topics. Data labeling was conducted using LLM Ensemble Voting involving three AI models followed by validation by three human annotators. The results show that IndoBERT in Stage 1 achieved an accuracy of 82.56% on the test set. In Stage 2, the SVM model achieved an accuracy of 95.07%, precision of 95.31%, recall of 95.07%, and F1-score of 95.07%, with cross-validation confirming stability at a mean accuracy of 96.74% (std = 0.19%). These findings demonstrate that the Two-Stage Classification approach effectively improves the specificity of hate speech detection by separating tasks in a sequential manner.
Keywords: *Hate Speech, SARA, IndoBERT, Support Vector Machine, LLM Ensemble Voting.*

Abstrak
Perkembangan media sosial di Indonesia telah meningkatkan interaksi masyarakat dalam menyampaikan opini melalui platform seperti YouTube, Instagram, dan TikTok. Di balik kemudahan tersebut, penyebaran ujaran kebencian (hate speech) yang mengandung unsur (SARA) menjadi tantangan yang berpotensi memicu konflik sosial. Penelitian ini mengusulkan pendekatan *Two-Stage Classification* untuk menjawab tantangan tersebut. Pada tahap pertama, model *IndoBERT (indobenchmark/indobert-base-p1)* digunakan untuk mengklasifikasikan komentar menjadi *Hate Speech dan Non-Hate Speech.* Pada tahap kedua, algoritma *Support Vector Machine (SVM)* dengan ekstraksi fitur TF-IDF dan lexicon SARA digunakan untuk membedakan ujaran kebencian berbasis SARA (HS_SARA) dan ujaran kebencian umum (HS_Umum). Dataset terdiri dari 36.000 komentar hasil scraping dari YouTube, Instagram, dan TikTok pada topik viral bermuatan SARA. Pelabelan data dilakukan menggunakan metode LLM *Ensemble Voting* yang melibatkan tiga model AI kemudian divalidasi oleh tiga annotator manusia. Hasil penelitian menunjukkan bahwa IndoBERT pada tahap pertama mencapai akurasi 82,56% pada data uji. Pada tahap kedua, SVM memperoleh akurasi 95,07%, precision 95,31%, recall 95,07%, dan F1-score 95,07%, dengan *cross-validation* mengkonfirmasi stabilitas model pada rata-rata akurasi 96,74% (std = 0,19%). Temuan ini membuktikan bahwa pendekatan *Two-Stage Classification* secara efektif meningkatkan spesifisitas deteksi ujaran kebencian dengan memisahkan tugas klasifikasi secara bertahap.
Kata Kunci: Ujaran Kebencian, SARA, *IndoBERT*, *Support Vector Machine*, Pemungutan Suara Ensemble LLM.

1. PENDAHULUAN

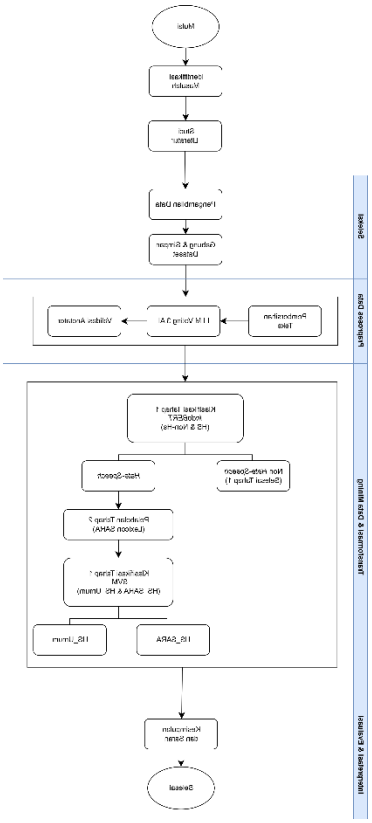
Perkembangan teknologi informasi yang pesar telah mendorong peningkatan interaksi social di ruang digital. Platform media social seperti YouTube, Instagram, dan TikTok kini menjadi ruang publik baru bagi masyarakat untuk mengekspresikan opin, kritik, dan kreativitas. Namun di balik kebebasa berekspresi tersebut muncul permasalahan serius berupa penyebaran ujaran kebencian *(hate speech),* khususnya yang mengandung unsur SARA. Fenomena ini tidak hanya memicu konflik di ruang digital, tetapi juga berpotensi mengancam kerukunan social di dunia nyata (Aurora Az Zahra et al., 2023). Sejak tahun 2018, Kementerian Komunikasi dan Informatika telah menangani sebanyak 3.640 konten ujaran kebencian berbasis SARA, sementara riset SAFEnet (2022) menemukan bahwa isu SARA mendominasi 62% dari seluruh kasus ujaran kebencian yang terdokumentasi di ruang digital Indonesia (Darussalam Gontar, 2023).

Karakteristik bahasa Indonesia informal di media sosial yang kaya akan slang, singkatan, dan variasi linguistik membuat deteksi otomatis menjadi tantangan tersendiri. Selain itu, ketersediaan dataset berbahasa Indonesia yang khusus menyoroti ujaran kebencian berbasis SARA masih sangat terbatas (Fetahi et al., 2025). Sebagian besar penelitian terdahulu juga hanya menggunakan pendekatan klasifikasi satu tahap (single-stage classification), sehingga belum mampu membedakan secara spesifik antara ujaran kebencian umum dan ujaran kebencian berbasis SARA (Caselli et al., 2021).

Penelitian terdahulu menunjukkan bahwa model berbasis transformer seperti IndoBERT terbukti efektif untuk klasifikasi teks bahasa Indonesia. Kusuma dan Chowanda (2023) mencapai akurasi 93,7% menggunakan IndoBERTweet dan BiLSTM untuk deteksi hate speech di Twitter. Dharmawan et al. (2022) memperoleh akurasi terbaik 89,52% menggunakan IndoBERT dengan feedforward neural network. Sementara itu, Khatib Sulaiman et al. (2024) membuktikan bahwa SVM dengan TF-IDF efektif untuk mendeteksi ujaran kebencian pada komentar TikTok. Namun, penelitian-penelitian tersebut belum secara khusus menangani dengan metode *Two-Stage Classification* dataset berbahasa Indonesia SARA juga masih sangat jarang ditemui.

Berdasarkan kesenjangan tersebut penelitian ini mengusulkan pendekatan *Two-Stage Classification* yang menggabungkan IndoBERT pada tahap pertama untuk memisahkan *hate speech* dari *non-hate speech*, dan SVM dengan lexicon SARA pada tahap kedua untuk mengidentifikasi ujaran kebencian berbasis SARA secara lebih spesifik. Dataset dibentuk oleh peneliti sendiri dan pelabelan dataset dilakukan menggunakan metode LLM Ensemble Voting yang melibatkan tiga model AI (LLaMA, Qwen, dan Mistral) yang kemudian divalidasi oleh tiga annotator manusia, sehingga menghasilkan dataset berlabel berkualitas tinggi yang representatif terhadap kondisi aktual media sosial berbahasa Indonesia (He et al., 2024).

2. METODOLOGI PENELITIAN



Gambar 1. Alur Penelitian

Penelitian ini menggunakan *framework* KDD *(Knowladge Discovery in Database)* yang terdiri dari tahap *selection, preprocessing, transformation, data mining,* serta interpretasi, evaluasi, dan implementasi pada website dengan *streamlit.*

2.1 Pengumpulan Data *(Selection)*

Data dikumpulkan melalui proses *scrapping* komentar dari 3 platform media sosial, yaitu YouTube, Instagram, dan TikTok, pada konten-konten viral yang berkaitan dengan topik sensitif SARA, meliputi Kasus Yai Mim, Kasus Gus Elham, Kasus Korupsi Dana Haji Yaqut Cholil, serta Kasus Ormas LDII, BANSER, dan MADAS. Total data mentah yang berhasil dikumpulkan berjumlah 58.429 komentar. Selanjutnya digunakan sebanyak 12.000 komentar per platform dengan distribusi seimbang 6.000 Hate Speech dan 6.000 Non-Hate Speech, sehingga total dataset yang digunakan berjumlah 36.000 komentar. Berikut atribut dataset yang telah diperoleh:

Tabel 1. Atribut Dataset Utama

No	Nama Atribut	Tipe Data	Penjelasan
1.	source	Object	Sumber data awal
2.	platform	Object	Sumber data per-platform
3.	comment	Object	Teks komen asli
4.	label_llama	Float64	Hasil label AI
5.	label_qwen	Float64	Hasil label AI
6.	label_mistral	Float64	Hasil label AI
7.	label_hs	Int64	Label hasil vote AI
8.	label_valid	Int64	Label final
9.	conflict	Int64	Tanda label ambigu

Kemudian selanjutnya untuk dataset untuk tahap 2 dapat dilihat pada tabel berikut:

Tabel 2. Atribut Dataset Tahap 2

No	Nama Atribut	Tipe Data	Penjelasan
1.	Platform	Object	Sumber data
2.	Comment_original	Object	Teks Asli
3.	Comment_clean	Object	Teks Bersih
4.	Confidence_tahap1	Float64	Nilai prediksi dari model 1
5.	Label_tahap2	Float64	Label tahap 2

2.2 Preprocessing

Tahap *preprocessing* dilakukan untuk membersihkan data dari elemen yang tidak relevan. Proses ini mencakup penghapusan data duplikat, data kosong (missing value), URL, mention, hashtag, karakter khusus, angka, dan emoji. Selain itu dilakukan pula normalisasi unicode, normalisasi huruf berulang, penghapusan teks pendek (kurang dari dua kata), case folding, dan pembersihan spasi berlebih.

2.3 Pelabelan Datase: LLM Ensemble Voting

Pelabelan data dilakukan menggunakan pendekatan *LLM Ensemble Voting* yang melibatkan tiga model Large Language Model (LLM), yaitu Meta LLaMA 3.3-70B-Instruct, Qwen3-32B, dan Mistral Nemo, yang diakses melalui platform OpenRouter API. Setiap model memberikan label secara independen (0 = Non-Hate Speech, 1 = Hate Speech), kemudian label akhir ditentukan melalui majority voting (minimal 2 dari 3 model sepakat). Data yang menghasilkan konflik ditandai dan diselesaikan melalui validasi manual oleh tiga annotator manusia. Distribusi akhir dataset menghasilkan 18.628 data Hate Speech (51,7%) dan 17.372 data Non-Hate Speech (48,3%).

2.4 Tahap 1: Klasifikasi dengan Algoritma IndoBERT

Model yang digunakan adalah indobenchmark/indobert-base-p1, yaitu model berbasis transformer yang dilatih khusus pada korpus bahasa Indonesia. Dataset dibagi menggunakan stratified split dengan proporsi 80% data latih (28.798 data), 10% data validasi (3.600 data), dan 10% data uji (3.600 data). Proses fine-tuning dilakukan dengan menambahkan lapisan klasifikasi dengan dua kelas output. Delapan layer pertama IndoBERT dibekukan (freeze) untuk mempertahankan representasi bahasa umum sekaligus mengurangi risiko overfitting. Konfigurasi hyperparameter yang digunakan meliputi:

Tabel 3. Konfigurasi Model

Parameter	Nilai
Model	indobenchmark/indobert-base-p1
Epoch	3
Batch Size	16
Learning Rate	1e-5
Warmup Ratio	0.2
Label Smoothing	0.1
Early Stopping	Patience = 1
Max Length	128
Threshold Tunning	0.5 ROC Curve

Pelatihan dilakukan menggunakan GPU Tesla T4 di Google Colab dengan teknik mixed precision (FP16).

2.5 Tahap 2: Klasifikasi menggunakan Lexicon SARA dan Algoritma SVM

Input pada tahap kedua adalah adalah 19.098 komentar yang diprediksi sebagai Hate Speech oleh model IndoBERT pada tahap pertama. Pelabelan dilakukan menggunakan pendekatan distant supervision berbasis lexicon SARA yang disusun secara mandiri berdasarkan empat dimensi:

Tabel 4. Dimensi Lexicon SARA

No	Dimensi	Frasa
1.	Agama	162 Frasa
2.	Suku	107 Frasa
3.	Ras	20 Frasa
4.	Antargolongan	42 Frasa

Hasil auto-labeling divalidasi secara manual pada 10% sampel stratifikasi. Distribusi label tahap kedua menghasilkan 9.727 data HS_SARA (51,01%) dan 9.342 data HS_Umum (48,99%). Representasi fitur teks dilakukan menggunakan TF-IDF dengan konfigurasi unigram dan

 ISSN 3063-864X (E)

bigram (max_features = 8.000). Dengan konfigurasi model yakni:

Tabel 5. Konfigurasi Model SVM

Parameter	Nilai
Model	LinearSVC
C	1.0
Max_iter	2000
Random_state	42

Dataset dibagi dengan proporsi 80% data latih (15.277 data) dan 20% data uji (3.820 data) atau 80:10 menggunakan stratified split. Evaluasi tambahan dilakukan menggunakan 5-Fold Stratified Cross Validation pada data latih.

2.6 Evaluasi

Evaluasi model dilakukan menggunakan metrik accuracy, precision, recall, F1-score, dan Confusion Matrix. Pada tahap pertama, evaluasi tambahan dilakukan melalui pemantauan kurva Training dan Validation Loss per epoch, serta threshold tuning menggunakan metode ROC Curve (Youden's J Statistic).

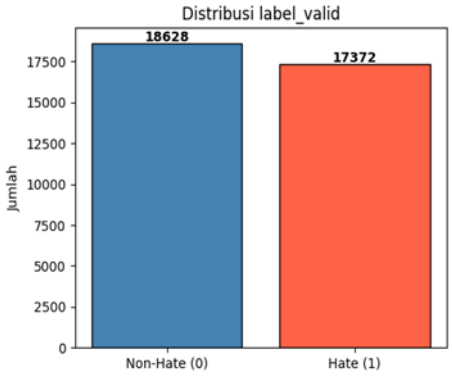
3. HASIL DAN PEMBAHASAN

3.1 Hasil Pelabelan Dataset

Proses pelabelan menggunakan LLM Ensemble Voting menghasilkan tingkat kesepakatan yang sangat tinggi antar ketiga model AI. Sebagai contoh pada platform TikTok, dari 9.418 data hanya terdapat 45 data (0,5%) yang masuk kategori konflik dan memerlukan validasi manual. Proporsi konflik yang sangat kecil ini menunjukkan bahwa mekanisme ensemble voting tiga model AI menghasilkan label yang konsisten dan andal. Setelah melalui validasi oleh tiga annotator manusia, distribusi akhir dataset berjumlah 18.628 data Hate Speech (51,7%) dan 17.372 data Non-Hate Speech (48,3%), sehingga distribusi kelas relatif seimbang dan tidak memerlukan teknik oversampling tambahan.

	A	B	C	D	E	F	G	H
1	source	platform	comment	label_llama	label_qwen	label_mistral	label_hs	label_val
2	datasetga	instagram	si paling pancasila	1	1	0	1	
3	datasetga	instagram	lepaskan mereka dari ind	1	1	1	1	
4	datasetga	instagram	usir dari pulau jawa	1	1	1	1	
5	datasetga	instagram	mabul manusia nya	1	1	0	1	
6	datasetga	instagram	maling kurupsi haji di bel	1	1	1	1	
7	datasetga	instagram	kasih tau madas ga berpe	1	0	1	1	
8	datasetga	instagram	ni manusia yg di halal kan	1	1	1	1	
9	datasetga	instagram	lhaa ngrombeng d kota or	1	0	1	1	
10	datasetga	instagram	ini juga yahudi ya	1	1	1	1	

Gambar 2. Hasil Pelabelan Dataset Tahap 1



Gambar 3. Distribusi Label Tahap 1

Tahap kedua penelitian menggunakan metode Support Vector Machine (SVM). Dataset tahap kedua terdiri dari 19.098 data Hate Speech hasil klasifikasi tahap pertama. Distribusi label terdiri dari 9.727 (51.01%) data HS_SARA dan 9.342 (48.99%) data HS_Umum.

	A	B	C	D	E	F	G
1	comment_clean	label_taha	dimensi_sara	kata_pemicu			
2	madas ga berpendidikan da	HS_SARA	['Suku/Etnis', 'Antargolon	['madas', 'madura', 'madas']			
3	madas ga berpendidikan da	HS_SARA	['Suku/Etnis', 'Antargolon	['madas', 'madas']			
4	heran ya dimana diusir loh	HS_SARA	['Suku/Etnis', 'Ras']	['usir', 'rasis']			
5	komedi org jawa	HS_SARA	['Suku/Etnis']	['jawa']			
6	madas ga berpendidikan da	HS_SARA	['Suku/Etnis', 'Antargolon	['madas', 'madura', 'bali', 'usir', 'madas']			
7	mgkin itu isu oknum dari m	HS_SARA	['Agama']	['Idii']			
8	madas ga berpendidikan da	HS_SARA	['Suku/Etnis', 'Antargolon	['madas', 'madas']			
9	coba ngomong kek gitu lagi	HS_SARA	['Suku/Etnis']	['sampit']			
10	pendatang kok ngamok	HS_SARA	['Suku/Etnis']	['pendatang']			
11	kebanyakan masiid Idii sud	HS_SARA	['Agama']	['Idii', 'eus']			

Gambar 4. Hasil Pelabelan Dataset Tahap 2

```
...
📊 Distribusi Label Tahap 2:
label_tahap2
HS_SARA    9727
HS_Umum    9342
Name: count, dtype: int64

label_tahap2
HS_SARA    51.01 %
HS_Umum    48.99 %
Name: proportion, dtype: object
```

Gambar 5. Distribusi Label Tahap 2

3.2 Tokenisasi

Sebelum masuk ke model pada tahap 1, dilakukan juga tokenization itu tahap untuk memecah teks menjadi unit-unit kecil yang disebut token sebelum diproses oleh model. (Koto Jey Han Lau Timothy Baldwin, 2021) Tahapan tokenisasi dilakukan menggunakan tokenizer bawaan IndoBERT. Proses tokenisasi dilakukan dengan panjang maksimum (max length) sebesar 128 token. Pembagian data *train, test,* dan *validation* dengan pembagian (80:10:10) agar nantinya siap dimasukkan kedalam tahap pelatihan model training.

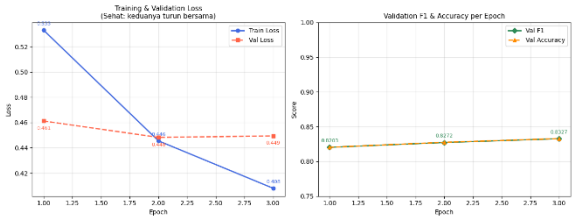
 ISSN 3063-864X (E)

```
Map: 100% 28798/28798 [00:03<00:00, 10892.58 examples/s]
Map: 100% 3600/3600 [00:00<00:00, 7289.15 examples/s]
Map: 100% 3600/3600 [00:00<00:00, 14154.57 examples/s]
✅ Tokenisasi selesai!
Train : 28798 sampel
Val   : 3600 sampel
Test  : 3600 sampel
```

Gambar 6. Hasil Tokenisasi

3.3 Hasil Tahap 1: Algoritma IndoBERT

Proses pelatihan model IndoBERT menunjukkan dinamika yang sehat, ditandai dengan peningkatan Validation F1-Score secara konsisten dari 0,8203 pada epoch pertama, menjadi 0,8272 pada epoch kedua, dan mencapai 0,8327 pada epoch ketiga. Validation Loss juga menunjukkan tren penurunan dari 0,4612 menjadi 0,4483, kemudian stabil di 0,4494, mengindikasikan bahwa model tidak mengalami overfitting.



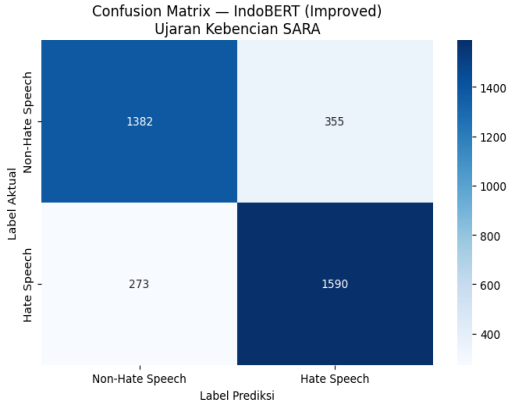
Gambar 7. *Kurva Loss & F1 per-Epoch*

Berdasarkan hasil evaluasi pada 3.600 data uji, model IndoBERT memperoleh accuracy sebesar 82,56% dengan F1-Score, precision, dan recall yang berada pada rentang 82,5%–82,6%. Hasil inferensi terhadap keseluruhan 36.000 data menghasilkan accuracy sebesar 87,69%, dengan 19.100 komentar diprediksi sebagai Hate Speech (53,1%) dan 16.898 komentar sebagai Non-Hate Speech (46,9%). Rata-rata confidence score prediksi sebesar 0,8479 menunjukkan bahwa model memiliki tingkat keyakinan yang tinggi dalam memberikan prediksi.

```
=== EVALUASI DATA TEST ===

=== CLASSIFICATION REPORT ===
                    precision    recall  f1-score   support

Non-Hate Speech (0)       0.84      0.80      0.81      1737
    Hate Speech (1)       0.82      0.85      0.84      1863

           accuracy                           0.83      3600
          macro avg       0.83      0.82      0.82      3600
       weighted avg       0.83      0.83      0.83      3600

----------------------------------------
Accuracy  : 0.8256 (82.56%)
F1-Score  : 0.8253
Precision : 0.8260
Recall    : 0.8256
----------------------------------------

=== PERBANDINGAN DENGAN VERSI SEBELUMNYA ===
Accuracy sebelumnya : 82.00%
Accuracy sekarang   : 82.56%
Perubahan           : ⬆️ 0.56%
```

Gambar 8. Hasil Evaluasi Model

Hasil Confusion Matrix menunjukkan bahwa model berhasil memprediksi 1.590 data sebagai True Positive dan 1.382 data sebagai True Negative, dengan 355 False Positive dan 273 False Negative.



Gambar 9. *Confusion Matrix*

Proses threshold tuning menggunakan metode ROC Curve (Youden's J) menghasilkan nilai AUC sebesar 0,9189 dan threshold optimal sebesar 0,5765, yang meningkatkan accuracy dari 82,28% (default 0,5) menjadi 82,67%. Peningkatan yang relatif kecil (~0,4%) mengindikasikan bahwa keterbatasan performa lebih bersumber dari sifat ambigu dan subjektif data komentar media sosial. Akurasi pada kisaran 82–83% merupakan hasil yang realistis dan sejalan dengan benchmark penelitian deteksi hate speech berbahasa Indonesia pada literatur sejenis.

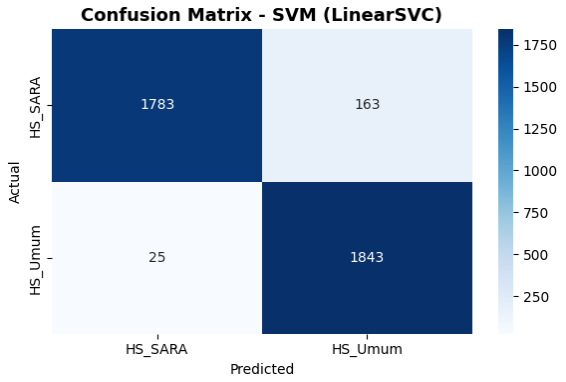
3.4 Hasil Tahap 2: Klasifikasi dengan SVM

Dataset hasil labeling dibagi menjadi data training dan data testing menggunakan metode stratified split. Pembagian dataset dilakukan dengan proporsi 80% data training dan 20% data testing. Berdasarkan hasil evaluasi pada 3.820 data uji, model SVM memperoleh accuracy sebesar 95,07%, precision sebesar 95,31%, recall sebesar 95,07%, dan F1-score sebesar 95,07%. Hasil per kelas menunjukkan bahwa kelas HS_SARA memperoleh precision sebesar 0,99 dan recall sebesar 0,92, sedangkan kelas HS_Umum memperoleh precision sebesar 0,92 dan recall sebesar 0,99. Kedua kelas memperoleh F1-score sebesar 0,95.

```
📊 Hasil Evaluasi SVM (LinearSVC):
Accuracy  : 0.9507 (95.07%)
Precision : 0.9531 (95.31%)
Recall    : 0.9507 (95.07%)
F1-Score  : 0.9507 (95.07%)

📋 Classification Report:
              precision    recall  f1-score   support

     HS_SARA       0.99      0.92      0.95      1946
     HS_Umum       0.92      0.99      0.95      1868

    accuracy                           0.95      3814
   macro avg       0.95      0.95      0.95      3814
weighted avg       0.95      0.95      0.95      3814
```

Gambar 10. Hasil Evaluasi SVM

Berdasarkan hasil confusion matrix diketahui bahwa sebagian besar data berhasil diprediksi dengan benar oleh model. Hal ini menunjukkan bahwa kombinasi fitur TF-IDF dan algoritma SVM mampu mempelajari pola ujaran kebencian berbasis SARA secara efektif. Berhasil memprediksi 1843 data sebagai True Negative, dengan 163 False Positive dan 25 False Negative.



Gambar 11. *Confusion Matrix*

Untuk mengukur stabilitas performa model dilakukan proses 5-Fold Stratified Cross Validation menggunakan metode StratifiedKFold pada data training. Untuk mengevaluasi stabilitas model, dilakukan 5-Fold Stratified Cross Validation menggunakan pipeline TF-IDF dan LinearSVC pada data pelatihan. Hasil evaluasi menunjukkan bahwa model memperoleh rata-rata accuracy sebesar 95,24% (±0,22%), precision sebesar 95,44% (±0,22%), recall sebesar 95,24% (±0,22%), dan F1-Score sebesar 95,24% (±0,22%). Berdasarkan evaluasi pada setiap fold, performa terbaik diperoleh pada Fold ke-5 dengan accuracy 95,48%, precision 95,70%, recall 95,48%, dan F1-Score 95,47%, sedangkan performa terendah diperoleh pada Fold ke-4 dengan accuracy 94,85%, precision 95,07%, recall 94,85%, dan F1-Score 94,85%. Selisih performa antarfold yang relatif kecil menunjukkan bahwa model memiliki performa yang stabil dan kemampuan generalisasi yang baik terhadap variasi pembagian data.

```
⏳ Menjalankan 5-Fold Cross Validation...

📊 Hasil 5-Fold Cross Validation SVM:
Accuracy  – Mean: 0.9524 | Std: 0.0022
            Fold : ['0.9535', '0.9515', '0.9538', '0.9485', '0.9548']

Precision – Mean: 0.9544 | Std: 0.0022
            Fold : ['0.9546', '0.9537', '0.9558', '0.9507', '0.9570']

Recall    – Mean: 0.9524 | Std: 0.0022
            Fold : ['0.9535', '0.9515', '0.9538', '0.9485', '0.9548']

F1-Score  – Mean: 0.9524 | Std: 0.0022
            Fold : ['0.9535', '0.9515', '0.9538', '0.9485', '0.9547']
```

Gambar 12. *Cross Validation*

3.5 Pembahasan Keseluruhan

Pendekatan Two-Stage Classification terbukti efektif dalam meningkatkan spesifisitas deteksi ujaran kebencian berbasis SARA. IndoBERT berhasil menjalankan fungsinya sebagai penyaring awal (filtering layer) yang memisahkan komentar hate speech dari komentar umum, sehingga SVM pada tahap kedua dapat lebih fokus mempelajari pola perbedaan antara HS_SARA dan HS_Umum. Performa SVM yang sangat tinggi (95,07%) dipengaruhi oleh penggunaan fitur TF-IDF unigram dan bigram yang mampu mengenali pola frasa seperti “dasar kafir”, “agama sesat”, atau “madas tolol”, serta penerapan lexicon SARA yang terstruktur berdasarkan empat dimensi.

Dibandingkan penelitian terdahulu, Kusuma dan Chowanda (2023) mencapai akurasi 93,7% pada klasifikasi hate speech satu tahap, sedangkan penelitian ini berhasil melakukan pemisahan lebih spesifik pada tingkat SARA dengan akurasi tahap kedua 95,07%. Hidayatulloh et al. (2025) melaporkan akurasi 97% untuk kombinasi IndoBERT dan SVM pada tugas analisis sentimen, sementara penelitian ini menerapkan kombinasi serupa pada domain yang lebih menantang, yaitu deteksi ujaran kebencian berbasis identitas kelompok. Hasil ini membuktikan bahwa dataset yang dibentuk mandiri dan arsitektur bertingkat *(two-stage)* memberikan performa yang lebih tangguh dan terstruktur dibandingkan metode klasifikasi tunggal konvensional.

KESIMPULAN

Penelitian ini berhasil membangun membangun sistem deteksi ujaran kebencian berbasis SARA menggunakan pendekatan Two-Stage Classification yang menggabungkan IndoBERT pada tahap pertama dan Support Vector Machine (SVM) pada tahap kedua. Pada Tahap 1 yakni menggunakan IndoBERT mencapai accuracy sebesar 82,56% pada data uji membuktikan kemampuannya sebagai penyaring hate speech yang efektif. Pada tahap kedua, SVM dengan fitur TF-IDF dan lexicon SARA berhasil mengklasifikasikan ujaran kebencian menjadi HS_SARA dan HS_Umum dengan berdasarkan metrik accuracy adalah 95,24% dengan standar deviasi 0,22%. Selain itu, model memperoleh rata-rata precision sebesar 95,44%, recall sebesar 95,24%, dan F1-Score sebesar 95,24%. Selain itu, metode pelabelan pada tahap 1 dengan AI yakni LLM Ensemble Voting yang melibatkan LLaMA, Qwen, dan Mistral dengan validasi tiga annotator manusia terbukti menghasilkan dataset berlabel berkualitas tinggi secara efisien. Serta pada tahap kedua dilakukan kembali pelabelan dengan LEXICON SARA yang dibangun mandiri oleh peneliti. Secara keseluruhan, pendekatan bertahap ini terbukti lebih efektif dibandingkan klasifikasi satu tahap karena memungkinkan optimasi model pada setiap tahap

secara terpisah dan menghasilkan klasifikasi yang lebih spesifik terhadap konten ujaran kebencian berbasis SARA di media sosial Indonesia.

Ucapan Terima Kasih

Terimakasih kepada seluruh Keluarga, Teman, Orang yang sudah mendampingi saya 2020 - jurnal ini dibuat sudah memberi support luar biasa yang tak terhingga dan selalu menemani saya. Terkhususkan juga untuk para idol Kpop saya Hearts2Hearts dan SEVENTEEN sebagai titik balik hidup saya, dan akan selalu menjadi pilihan yang saya pilih sekarang, esok, maupun sampai kapanpun.

DAFTAR PUSTAKA

Aurora Az Zahra, E., Sibaroni, Y., & Suryani Prasetyowati, S. (2023*). Classification of Multi-Label of Hate Speech on Twitter Indonesia using LSTM and BiLSTM Method. JINAV: Journal of Information and Visualization, 4(2),* 170–178. https://doi.org/10.35877/454ri.jinav1864

Caselli, T., Basile, V., Mitrović, J. M., & Granitzer, M. (2021). *HateBERT: Retraining BERT for Abusive Language Detection in English.*

Darussalam Gontar, U. (2023). *Ujaran Kebencian Berbasis Agama: Kebebasan Berbicara dan Konsekuensi Terhadap Kerukunan Umat Beragama.*

Dharmawan, S., Mawardi, V. C., & Perdana, N. J. (2022). *Klasifikasi Ujaran Kebencian Menggunakan Metode FeedForward Neural Network (IndoBERT). Jurnal Ilmu Komputer dan Sistem Informasi.*

Fetahi, E., Susuri, A., Hamiti, M., Kastrati, Z., Canhasi, E., & Misini, A. (2025*). Enhancing social media hate speech detection in low-resource languages using transformers and explainable AI. Social Network Analysis and Mining, 15(1).* https://doi.org/10.1007/s13278-025-01497-w

Findawati, Y., Raharjo, A. B., Navastara, A., Yonathan, V., Yatestha, A., & Purwitasari, D. (2025). *Multi-label Aspect Dangerous Speech Classification Using Keyword-Driven Ensemble Classifier on Imbalanced Data. International Journal on Informatics Visualization.*

He, X., Lin, Z., Gong, Y., Jin, A.-L., Zhang, H., Lin, C., Jiao, J., Yiu, S. M., Duan, N., & Chen, W. (2024). AnnoLLM: *Making Large Language Models to Be Better Crowdsourced Annotators.* http://arxiv.org/abs/2303.16854

Hidayatulloh, S., Muflikhah, L., & Perdana, R. S. (2025). *Implementasi Embedding IndoBERT dan Support Vector Machine (SVM) Untuk Analisis Sentimen Publik terhadap Layanan Biznet. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 9(9).*

Ibrohim, M. O., & Budi, I. (2019). *Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter. Proceedings of the Third Workshop on Abusive Language Online.*

Khatib Sulaiman, J., Ariska, A., Kamayani, M., & Muhammadiyah DrHamka, U. (2024). *Deteksi Hate Speech pada Kolom Komentar TikTok dengan Menggunakan SVM. Indonesian Journal of Computer Science.*

Koto, F., Lau, J. H., & Baldwin, T. (2021). IndoBERTweet: *A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization. Proceedings of EMNLP 2021.*

Kusuma, F., & Chowanda, A. (2023). *Indonesian Hate Speech Detection Using IndoBERT and BiLSTM on Twitter. Proceedings of the International Conference on Information Technology.*

Novandian, Y. D., Luthfiarta, A., Assyifa, D. S., Setiawan, J., Cahyaningrum, L., Althoff, N., Rahayu, M., Nugraha, A., & Rismiyati. (2024). *IndoBERT-based Indonesian Cyberbullying Detection with Multi-stage Labeling. Proceedings — 2024 International Seminar on Application for Technology of Information and Communication (ISemantic 2024), 515–521.* https://doi.org/10.1109/iSemantic63362.2024.10762553