BAB 2
LANDASAN TEORI

2.1 Studi Literatur

Sejumlah studi sebelumnya telah membahas peningkatan deteksi ujaran kebencian dengan mengoptimalkan berbagai model transformer modern. García et al. [9] menyelidiki peningkatan deteksi ujaran kebencian melalui optimalisasi model transformer seperti TwHIN, DistilBERT, mDeBERTa, BERT, dan MarIA untuk bahasa Spanyol dan Inggris. Studi mereka yang berjudul *"Leveraging Zero and Few-Shot Learning for Enhanced Model Generality in Hate Speech Detection in Spanish and English"* menunjukkan kemampuan model-model tersebut dalam menangani skenario ujaran kebencian yang kompleks secara efektif. Rincian metrik performa seperti presisi dan recall untuk masing-masing model terhadap dataset bahasa Spanyol dan Inggris disajikan secara lengkap dalam Tabel 2.1. Studi ini menyoroti adaptabilitas dan ketangguhan model transformer dalam mengelola tugas deteksi ujaran kebencian di lingkungan linguistik yang berbeda.

Kathiravan et al. [10] membandingkan efektivitas berbagai model klasifikasi NLP untuk mendeteksi konten ofensif dalam percakapan campuran bahasa Tamil-Inggris. Dalam studi terbarunya, mereka mengusulkan metode Sentence Transfer Fine-tuning (SetFit) yang dikombinasikan dengan regresi logistik. Metode ini terbukti mengungguli model-model tradisional seperti Multilingual BERT (mBERT), LSTM, BERT, IndicBERT, dan LaBSE. SetFit berhasil mencapai presisi sebesar 0,90, recall 0,87, dan F1-score 0,88, dengan akurasi keseluruhan 89,72%. Hasil ini menunjukkan bahwa SetFit memiliki kemampuan yang kuat dalam mendeteksi konten ofensif pada teks campuran bahasa dan membuka potensi besar bagi pengembangan aplikasi NLP dalam konteks alih kode.

Helmi Imaduddin et al. [11] menerapkan teknik deep learning LSTM (Long Short-Term Memory) yang dipadukan dengan representasi kata GloVe (Global Vectors for Word Representation) untuk mendeteksi ujaran kebencian di Twitter. Metode ini diterapkan pada dataset berisi 13.169 tweet berbahasa Indonesia yang telah diklasifikasikan ke dalam kategori ujaran kebencian dan bukan kebencian. Kombinasi antara LSTM dan GloVe terbukti sangat efektif, dengan model terbaik mencatatkan akurasi sebesar 94,24%, presisi 89%, recall 99%, dan F1-score 94%.

Hashmi et al. [12] membahas tantangan deteksi ujaran kebencian dalam

bahasa sumber daya rendah seperti Norwegia. Studi mereka yang berjudul *"Multi-class Hate Speech Detection in the Norwegian Language Using FAST-RNN and Multilingual Fine-tuned Transformers"* menekankan pentingnya peningkatan pemantauan dan kebijakan media sosial dalam melawan ujaran diskriminatif. Metode mereka menggabungkan embedding FastText dengan arsitektur Bidirectional LSTM dan GRU, serta transformer multibahasa yang disesuaikan dan dioptimalkan melalui penyetelan hiperparameter. Dengan membandingkan model mereka terhadap model-model mutakhir dan menambahkan pendekatan interpretabilitas LIME (Local Interpretable Model-Agnostic Explanations), mereka berhasil meningkatkan kinerja sekaligus transparansi model deteksi ujaran kebencian.

Siddiqui et al. [13] mengkaji deteksi ujaran kebencian multibahasa yang lebih terperinci dalam studi berjudul *"Fine-Grained Multilingual Hate Speech Detection Using Explainable AI and Transformers"*. Dengan menggunakan model transformer seperti XLM-RoBERTa serta pendekatan Explainable AI seperti LIME, mereka meningkatkan interpretabilitas model dan proses pengambilan keputusan. Dataset multibahasa yang digunakan diklasifikasikan ke dalam kategori ujaran kebencian yang lebih rinci, seperti disabilitas, gender, kewarganegaraan, ras, dan agama. Studi ini mencatatkan F1-score sebesar 91% untuk klasifikasi biner dan 86% untuk klasifikasi rinci, dengan performa tertinggi pada kategori agama.

Kumar et al. [14] mengevaluasi performa model GPT dan LLaMA-2 yang telah disesuaikan untuk berbagai tugas NLP, termasuk deteksi ujaran kebencian. Mereka menggunakan GPT-2, GPT-3, GPT-3.5, dan LLaMA-2 (7B, 13B, dan 70B) yang dioptimalkan dengan metode QLoRA dan mengaplikasikan dataset seperti HSOL, SST-2, dan FNC. Hasil penelitian menunjukkan bahwa model LLaMA-2 13B memberikan performa terbaik, dengan keseimbangan yang unggul antara efisiensi dan akurasi.

Tabel 2.1. Analisis komparatif model NLP untuk deteksi ujaran kebencian

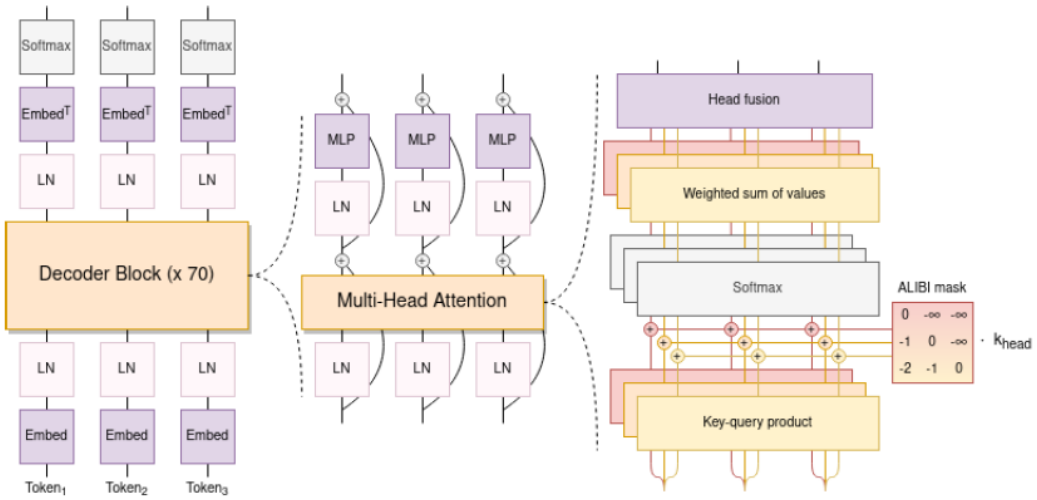
Studi	Model	Bahasa	Dataset	Akurasi	F1-score	Presisi	Recall	State Of The Art (SOTA)
García et al. [9]	TwHIN	Spanish	EXIST-2021-es	tidak disebutkan	82,26%	tidak disebutkan	tidak disebutkan	Menunjukkan adaptabilitas dan efektivitas model transformer dalam mendorong kemajuan deteksi ujaran kebencian multibahasa.
		English	HatEval	tidak disebutkan	63,98%	tidak disebutkan	tidak disebutkan	
			HASOC	tidak disebutkan	86,76%	tidak disebutkan	tidak disebutkan	
	DistilBETO	Spanish	EXIST-2021-es	tidak disebutkan	69,87%	tidak disebutkan	tidak disebutkan	
			HatEval	tidak disebutkan	65,37%	tidak disebutkan	tidak disebutkan	
	mDeBERTa	Spanish	HaterNET	tidak disebutkan	68,86%	tidak disebutkan	tidak disebutkan	
			MisoCorpus	tidak disebutkan	90,50%	tidak disebutkan	tidak disebutkan	
	MarIA	Spanish	Football	tidak disebutkan	85,18%	tidak disebutkan	tidak disebutkan	
	BERT	English	EXIST-2021-en	tidak disebutkan	79,77%	tidak disebutkan	tidak disebutkan	
			EXIST-2022-en	tidak disebutkan	79,68%	tidak disebutkan	tidak disebutkan	
			EDOS	tidak disebutkan	73,80%	tidak disebutkan	tidak disebutkan	
Kathiravan et al. [10]	SetFit	Tamil-English	Manual	89,00%	88,00%	90,00%	tidak disebutkan	Menunjukkan efektivitas unggul dari SetFit dibandingkan model tradisional, menyoroti kemampuannya yang tangguh dalam menangani tantangan khas deteksi konten ofensif pada bahasa campuran dengan akurasi tinggi (89,72%).
	mBERT			86,00%	84,00%	88,00%	tidak disebutkan	
	LSTM			76,00%	67,00%	70,00%	tidak disebutkan	
	BERT			78,00%	70,00%	72,00%	tidak disebutkan	
	indicBERT			80,00%	72,00%	74,00%	tidak disebutkan	
	LaBSE			84,00%	79,00%	80,00%	tidak disebutkan	
Imaduddin et al. [11]	GloVe + 6 Layer	Indonesia	Twitter	94,24%	94,00%	89,00%	tidak disebutkan	Menunjukkan efektivitas tinggi dalam mendeteksi nuansa linguistik pada ujaran kebencian berbahasa Indonesia di Twitter, dengan capaian recall dan performa keseluruhan yang sangat baik.
	LSTM + 6 Layer			93,61%	93,00%	88,00%	tidak disebutkan	
	GloVe + 5 Layer			94,24%	93,00%	88,00%	tidak disebutkan	
	LSTM + 5 Layer			93,80%	93,00%	89,00%	tidak disebutkan	
Hashmi et al. [12]	mBERT	Norwegia	Resett, Twitter, Facebook	82,00%	79,00%	78,00%	82,00%	Menggabungkan jaringan saraf lanjutan dan alat interpretabilitas untuk meningkatkan performa sekaligus pemahaman sistem deteksi ujaran kebencian dalam lingkungan data terbatas (low-resource).
	ELECTRAbase			83,00%	75,00%	69,00%	83,00%	
	ELECTRAlarge			83,00%	75,00%	69,00%	83,00%	
	scandi-BERT			81,00%	80,00%	79,00%	81,00%	
	nb-BERTbase			81,00%	81,00%	81,00%	81,00%	
	nb-BERTlarge			81,00%	81,00%	81,00%	81,00%	
	Nor-BERTsmall			82,00%	79,00%	77,00%	83,00%	
	Nor-BERTbase			82,00%	80,00%	78,00%	83,00%	
	Nor-BERTlarge			83,00%	81,00%	80,00%	82,00%	
	FLAN-T5-small			80,00%	77,00%	76,00%	80,00%	
	FLAN-T5-base			83,00%	80,00%	82,00%	82,00%	
	Nor-T5small			77,00%	78,00%	77,00%	78,00%	

Studi	Model	Bahasa	Dataset	Akurasi	F1-score	Presisi	Recall	State Of The Art (SOTA)
Siddiqui et al. [13]	mBERT	English	Public dataset	97,00%	97,00%	tidak disebutkan	tidak disebutkan	Mengembangkan deteksi ujaran kebencian multibahasa yang terperinci dengan AI yang dapat dijelaskan (explainable AI), meningkatkan interpretabilitas dan akurasi dalam mengklasifikasikan ujaran kebencian pada kategori seperti Disabilitas, Gender, Kewarganegaraan, Ras, dan Agama.
		Urdu	Translated	69,00%	68,00%	tidak disebutkan	tidak disebutkan	
		Sindhi	Translated from English	87,00%	86,00%	tidak disebutkan	tidak disebutkan	
		Combined	Combined All dataset	84,00%	84,00%	tidak disebutkan	tidak disebutkan	
	XLM-RoBERTa	English	Public dataset	97,00%	97,00%	tidak disebutkan	tidak disebutkan	
		Urdu	Translated	69,00%	68,00%	tidak disebutkan	tidak disebutkan	
		Sindhi	Translated from English	92,00%	92,00%	tidak disebutkan	tidak disebutkan	
		Combined	Combined All dataset	85,00%	86,00%	tidak disebutkan	tidak disebutkan	
	Distil-RoBERTa	English	Public dataset	97,00%	97,00%	tidak disebutkan	tidak disebutkan	
		Urdu	Translated	54,00%	51,00%	tidak disebutkan	tidak disebutkan	
		Sindhi	Translated from English	74,00%	73,00%	tidak disebutkan	tidak disebutkan	
		Combined	Combined All dataset	75,00%	74,00%	tidak disebutkan	tidak disebutkan	
Kumar et al. [14]	GPT-2 FT 500	English	HSOL	78,32%	77,54%	tidak disebutkan	tidak disebutkan	Menunjukkan pengaruh skala model dan ukuran data fine-tuning terhadap performa berbagai tugas NLP. LLaMA-2 70B secara konsisten mengungguli model yang lebih kecil, terutama pada ukuran data besar, menunjukkan keunggulan skalabilitas dan efisiensi.
			SST-2	74,41%	72,68%	tidak disebutkan	tidak disebutkan	
			FNC	80,43%	79,20%	tidak disebutkan	tidak disebutkan	
	GPT-3 FT 500	English	HSOL	96,29%	91,21%	tidak disebutkan	tidak disebutkan	
			SST-2	87,48%	86,09%	tidak disebutkan	tidak disebutkan	
			FNC	84,35%	83,65%	tidak disebutkan	tidak disebutkan	
	GPT-3.5 FT 500	English	HSOL	96,94%	96,25%	tidak disebutkan	tidak disebutkan	
			SST-2	90,01%	89,08%	tidak disebutkan	tidak disebutkan	
			FNC	97,11%	96,88%	tidak disebutkan	tidak disebutkan	
	LLaMA-2 7B FT 500	English	HSOL	91,50%	90,82%	tidak disebutkan	tidak disebutkan	
			SST-2	88,32%	87,85%	tidak disebutkan	tidak disebutkan	
			FNC	86,95%	84,99%	tidak disebutkan	tidak disebutkan	
	LLaMA-2 13B FT 500	English	HSOL	92,49%	91,87%	tidak disebutkan	tidak disebutkan	
			SST-2	89,06%	88,36%	tidak disebutkan	tidak disebutkan	
			FNC	87,23%	85,86%	tidak disebutkan	tidak disebutkan	
	LLaMA-2 70B FT 500	English	HSOL	92,59%	92,35%	tidak disebutkan	tidak disebutkan	
			SST-2	89,79%	89,48%	tidak disebutkan	tidak disebutkan	
			FNC	97,07%	96,32%	tidak disebutkan	tidak disebutkan	
	GPT-2 FT 1000	English	HSOL	82,06%	80,52%	tidak disebutkan	tidak disebutkan	
			SST-2	88,73%	87,22%	tidak disebutkan	tidak disebutkan	
			FNC	84,61%	83,11%	tidak disebutkan	tidak disebutkan	
	GPT-3 FT 1000	English	HSOL	92,98%	91,37%	tidak disebutkan	tidak disebutkan	
			SST-2	90,87%	89,40%	tidak disebutkan	tidak disebutkan	

Studi	Model	Bahasa	Dataset	Akurasi	F1-score	Presisi	Recall	State Of The Art (SOTA)
			FNC	92,52%	91,34%	tidak disebutkan	tidak disebutkan	
	GPT-3.5 FT 1000	English	HSOL	97,84%	97,33%	tidak disebutkan	tidak disebutkan	
			SST-2	92,84%	92,46%	tidak disebutkan	tidak disebutkan	
			FNC	97,51%	96,94%	tidak disebutkan	tidak disebutkan	
	LLaMA-2 7B FT 1000	English	HSOL	93,01%	92,11%	tidak disebutkan	tidak disebutkan	
			SST-2	91,11%	90,48%	tidak disebutkan	tidak disebutkan	
			FNC	92,68%	90,78%	tidak disebutkan	tidak disebutkan	
	LLaMA-2 13B FT 1000	English	HSOL	93,82%	93,00%	tidak disebutkan	tidak disebutkan	
			SST-2	91,36%	90,84%	tidak disebutkan	tidak disebutkan	
			FNC	92,95%	91,58%	tidak disebutkan	tidak disebutkan	
	LLaMA-2 70B FT 1000	English	HSOL	95,26%	95,02%	tidak disebutkan	tidak disebutkan	
			SST-2	92,21%	91,98%	tidak disebutkan	tidak disebutkan	
			FNC	97,50%	96,82%	tidak disebutkan	tidak disebutkan	

2.2 Model BLOOM

BLOOM (BigScience Large Open-science Open-access Multilingual Language Model) merupakan model bahasa multibahasa berbasis transformer yang mutakhir dan dirancang untuk menangani berbagai tugas pemrosesan bahasa alami dalam berbagai bahasa. Model ini memiliki 176 miliar parameter dan dilatih menggunakan 46 bahasa alami serta 13 bahasa pemrograman. BLOOM dikembangkan dan dirilis oleh kolaborasi ratusan peneliti internasional. Proses pelatihan BLOOM didukung oleh hibah publik dari pemerintah Prancis melalui GENCI (Grand équipement national de calcul intensif) dan IDRIS (Institut du développement et des ressources en informatique scientifique), dengan menggunakan superkomputer Jean Zay yang dimiliki oleh IDRIS [15].



Gambar 2.1. Arsitektur model BLOOM

Model ini dibangun berdasarkan arsitektur transformer, seperti yang ditunjukkan pada Gambar 2.1, yang sangat unggul dalam memproses data berurutan seperti teks melalui lapisan self-attention dan jaringan saraf feed-forward. Mekanisme self-attention memungkinkan model untuk memberikan bobot penting pada setiap kata dalam sebuah urutan terhadap kata-kata lainnya, sehingga dapat memahami konteks, keterkaitan, dan hubungan antar kata secara menyeluruh. BLOOM dilatih menggunakan data multibahasa berskala besar yang berasal dari berbagai sumber seperti buku, artikel, dan situs web, yang memungkinkan model ini untuk mengenali pola dan struktur kompleks dalam berbagai bahasa. Melalui pelatihan tersebut, BLOOM dapat memproses token dengan cara mengubah teks input menjadi satuan-satuan yang lebih kecil seperti kata atau sub-kata, lalu

menghasilkan respons yang koheren dan sesuai konteks, sebagaimana digambarkan pada Gambar 2.1 [16].

Aplikasi model ini dalam deteksi ujaran kebencian untuk bahasa Italia telah diteliti dalam [17], [18]. Pelatihan dilakukan menggunakan dua dataset yang dikenal luas, yaitu tugas EVALITA (deteksi misogini) dan HASPEEDE-v2-2020 (deteksi ujaran kebencian). Model BLOOM yang digunakan dalam penelitian tersebut adalah bloom-1b7, dengan strategi adaptasi bahasa MAD-X seperti dijelaskan dalam [19]. Hasil dari studi tersebut menunjukkan efektivitas BLOOM dalam konteks bahasa Italia, dengan nilai rata-rata macro F1-score terbaik sebesar 0,785.

Dalam penelitian ini, model yang digunakan adalah BLOOM-560m, yaitu varian ringan dari arsitektur BLOOM yang dikembangkan oleh BigScience, dengan jumlah parameter sebesar 560 juta — jauh lebih kecil dibandingkan versi penuhnya yang berisi 176 miliar parameter. Pemilihan model ini didasarkan pada pertimbangan praktis, khususnya keterbatasan sumber daya komputasi dan kebutuhan akan efisiensi dalam proses pelatihan dan eksperimen.

Meskipun berukuran lebih kecil, BLOOM-560m tetap menggunakan arsitektur transformer yang kuat seperti model aslinya, sehingga tetap mampu memproses data multibahasa maupun teks campuran bahasa seperti yang menjadi fokus dalam penelitian ini. Ukuran model yang lebih ringkas juga memungkinkan penggunaan platform komputasi yang lebih mudah diakses, seperti Google Colab Pro, yang sudah memadai untuk proses pelatihan dan penyetelan model ini tanpa memerlukan superkomputer.

Selain itu, evaluasi awal menunjukkan bahwa BLOOM-560m menawarkan keseimbangan yang baik antara performa dan kebutuhan komputasi, terutama dalam konteks mendeteksi ujaran kebencian yang bersifat halus dalam teks campuran bahasa Indonesia dan Inggris. Pendekatan ini menjadikan penelitian tetap efisien dan realistis, baik dari sisi linguistik maupun dari sisi teknis pelaksanaannya.