Klasifikasi Ujaran Kebencian Menggunakan Metode FeedForward Neural Network (IndoBERT)

Steven Dharmawan [1] Viny Christanti Mawardi [2] Novario Jaya Perdana [3]

[1) 2) 3)] Teknik Informatika Universitas Tarumanagara
Jl. Letjen S. Parman No. 1, Jakarta Barat 11440 Indonesia
[1)] steven.535180075@stu.untar.ac.id [2)] viny@fti.untar.ac.id [3)] novariojp@fti.untar.ac.id

ABSTRACT

Everyone in Indonesia has freedom of speech, both in real life and on social media. However, freedom of speech carried out without filtering can lead to hate speech. Hate speech is a form of discrimination directed against individuals or groups of individuals based on race, religion, gender, sexual orientation, or other identities. Hate speech can harm other parties which as a result can trigger conflict, violence, and can even cost a person's life. Therefore, it is important to be able to identify and manage this hate speech effectively. One way to manage hate speech on social media is to classify it. In this study, a web-based application was created that can classify a sentence to determine whether the sentence is hate speech or a normal sentence. The model created for classification uses the feedforward neural network method with IndoBERT. Based on the test results, the model created using the feedforward neural network method with IndoBERT provides the best accuracy of 89.52%.

Key words: Hate Speech, Feedforward Neural Network, IndoBERT

1. Pendahuluan

Di era digital ini, penggunaan media sosial telah menjadi hal yang umum. Banyak orang yang telah menggunakan media sosial. Bahkan tidak sedikit orang yang sulit untuk lepas dari menggunakan media sosial. Media sosial merupakan sarana untuk berinteraksi orang-orang satu sama lain dengan cara menciptakan, berbagi, serta bertukar informasi dan gagasan melalui kata-kata, gambar, dan video dalam sebuah jaringan dan komunitas virtual [1].

Setiap orang di Indonesia mempunyai kebebasan dalam berbicara, baik di kehidupan yang nyata maupun di media sosial. Namun kebebasan dalam berbicara yang dilakukan tanpa penyaringan dapat menyebabkan terjadinya ujaran kebencian. Menurut Surat Edaran Mabes Polri No:SE/6/X/2015, tanggal 8 Oktober 2015, ujaran kebencian di definisikan sebagai "tindak pidana yang berbentuk, penghinaan, pencemaran nama baik, penistaan, perbuatan yang tidak menyenangkan, memprovokasi, menghasut, penyebaran berita bohong, dimana semua tindakan di atas memiliki tujuan atau bisa berdampak pada tindak diskriminasi, kekerasan, penghilangan nyawa, dan atau konflik sosial".

Maraknya konten ujaran kebencian di media sosial terjadi karena canggihnya teknologi yang tidak diimbangi oleh budaya literasi dan kecerdasan emosional para pengguna akun [2]. Karna kurangnya pengetahuan tentang ujaran kebencian, banyak orang tidak menyadari bahwa kata-kata yang telah dikeluarkan merupakan ujaran kebencian yang dapat merugikan pihak lain. Untuk mengatasi permasalahan di atas, dibutuhkan sistem yang dapat melakukan klasifikasi apakah sebuah kalimat mengandung ujaran kebencian atau tidak.

Sistem di atas merupakan salah satu tugas dari domain Natural Language Processing yaitu Klasifikasi teks (*text classification*). Termasuk klasifikasi teks karena sistem mengategorikan kalimat menjadi ujaran kebencian atau bukan ujaran kebencian. Terdapat dua kelas atau kategori pada *output* dari sistem di atas yaitu "Ujaran Kebencian" dan "Bukan Ujaran Kebencian". Natural Language Processing adalah salah satu bidang dari kecerdasan buatan (*artificial intelligence*), dan Klasifikasi teks adalah tugas menetapkan kalimat atau dokumen kategori yang sesuai.

Sistem dibuat menggunakan pre-trained model yaitu IndoBERT. Alasan penggunaan IndoBERT adalah karena IndoBERT merupakan model yang dilatih khusus menggunakan Bahasa Indonesia sehingga IndoBERT sangat cocok digunakan untuk melakukan tugas *Natural Language Processing* dalam Bahasa Indonesia. IndoBERT dilatih menggunakan dataset berbahasa Indonesia yang terdiri dari sekitar 4 miliar kata, dengan sekitar 250 juta kalimat. IndoBERT mempunyai arsitektur yang sama dengan BERT, yang membedakan hanyalah pada dataset yang digunakan untuk pelatihan *unsupervised*.

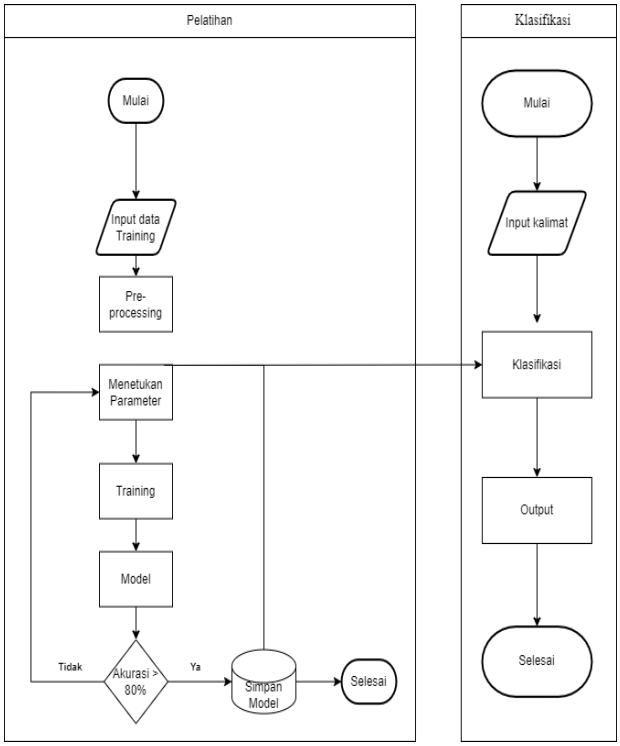
Penelitian sebelumnya [3] yang mendeteksi ujaran kebencian dan kalimat kasar di Twitter Indonesia memberikan akurasi terbaik 77,36% dengan menggunakan metode Random Forest Decision Tree, yang mana dataset nya akan digunakan dalam

penelitian ini. Adapun penelitian yang dilakukan oleh [4] yaitu deteksi ujaran kebencian di Twitter Indonesia memberikan akurasi terbaik sebesar 84,77% dengan menggunakan metode Bi-GRU dengan IndoBERT. Penelitian lain dilakukan oleh [5] yaitu deteksi ujaran kebencian di Instagram memberikan performa terbaik dengan F1-score 93,70% menggunakan metode TextCNN. Penelitian lainnya dilakukan oleh [6] yaitu deteksi ujaran kebencian Indonesia menggunakan *deep learning* memberikan performa terbaik dengan F1-*score* 87.98% menggunakan *word embedding* dengan arsitektur CBOW.

2. Metode Penelitian

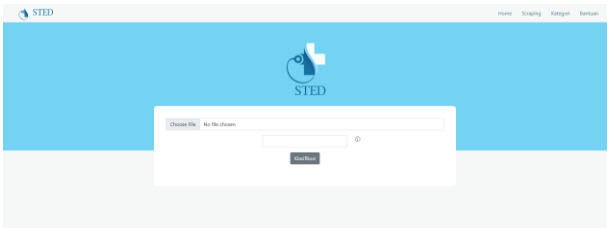
Terdapat beberapa tahapan yang dilakukan dalam pembuatan aplikasi pada penelitian ini. Pembuatan terdiri dari pengumpulan data, perancangan *flowchart*, pembuatan *user interface*, dan pengujian sistem. Seperti namanya, tahap pengumpulan data merupakan tahap mengumpulkan data yang akan digunakan untuk melatih model yang akan dibuat. Tahap perancangan *flowchart* merupakan tahapan untuk mengambarkan setiap proses yang akan dilakukan saat pembuatan aplikasi. *Flowchart* yang dibuat dapat dilihat pada Gambar 1. Tahap pembuatan *user interface* merupakan tahapan untuk membuat tampilan agar program dapat digunakan oleh pengguna. Tahap pengujian sistem dilakukan untuk menguji apakah model dapat digunakan untuk melakukan klasifikasi dan memberikan hasil yang sesuai.

Aplikasi yang dirancang menggunakan bahasa pemrograman Python. Aplikasi yang dirancang merupakan aplikasi untuk melakukan klasifikasi kata ujaran kebencian. Klasifikasi dapat dilakukan dengan menggunakan data sendiri (menggunggah file CSV sendiri). Selain dengan menggunakan data sendiri, aplikasi juga dapat mengambil komentar dari Youtube atau tweet dari Twitter dan melakukan klasifikasi.

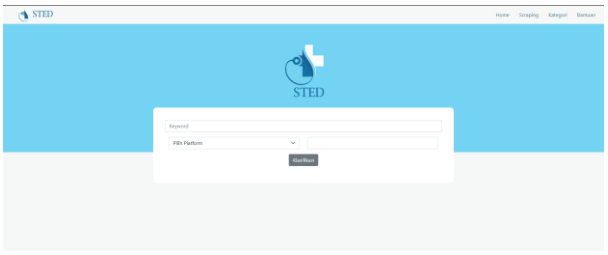


Gambar 1 *Flowchart*

Pengambilan komentar dari Youtube dilakukan menggunakan Youtube API v3 yang menggunakan bahasa pemrograman Python. Sedangkan untuk pengambilan tweet menggunakan library yang terdapat pada bahasa pemrograman Python yaitu Snscrape. Flowchart pada Gambar 1 terbagi menjadi dua yaitu, proses pelatihan dan proses klasifikasi. Pada proses pelatihan, dilakukan *pre-processing* pada dataset yang akan digunakan untuk melatih model. Proses selanjutnya adalah menentukan parameter yang akan digunakan pada model. Setelah parameter ditentukan, dilakukan pelatihan pada model dan didapatkan akurasi dari model. Jika akurasi lebih besar dari 80% maka model akan disimpan, namun jika akurasi model lebih kecil dari 80% maka akan dilakukan pelatihan ulang dengan mengubah parameter yang digunakan. Pada proses klasifikasi, sistem akan menerima kalimat dan melakukan klasifikasi dengan menggunakan model yang telah disimpan pada proses pelatihan. Lalu akan didapatkan hasil klasifikasi berupa "Ujaran Kebencian" atau "Bukan Ujaran Kebencian".



Gambar 2 Halaman Utama



Gambar 3 Halaman Scraping

Aplikasi yang dirancang adalah aplikasi berbasis web. Gambar 2 merupakan halaman utama yang digunakan untuk melakukan klasifikasi dengan menggunakan data sendiri dan Gambar 3 merupakan halaman untuk melakukan klasifikasi dengan mengambil komentar dari Youtube atau mengambil tweet dari Twitter.

2.1 *Pre-Processing*

Sebelum data digunakan untuk melatih model, perlu dilakukan *pre-processing* data terlebih dahulu. *Pre-processing* pada data bertujuan untuk memastikan data memiliki kualitas yang baik untuk digunakan dalam melatih model. *Pre-processing* yang dilakukan pada penelitian ini adalah *data cleaning* dan tokenisasi.

Data cleaning merupakan proses untuk membersihkan data. Yang dilakukan pada data cleaning adalah menghapus emoji, menghapus URL, dan menghapus *whitespace*. Sedangkan tokenisasi Tokenisasi merupakan proses pemisahan teks menjadi unit-unit yang lebih kecil yang disebut token contohnya adalah kata-kata [7].

2.2 Feedforward Neural Network

Feedforward neural network merupakan salah satu kategori dari arsitektur jaringan. Jika tidak terdapat "*feedback"* dari output neuron terhadap input di seluruh jaringan, maka jaringan tersebut disebut sebagai *feedforward neural network* [8]. *Feedforward neural network* terbagi dalam dua kategori tergantung pada jumlah lapisan, yaitu "*single layer*" dan "*multi-layer*" [8]. Pada *single layer* terdapat dua lapisan jika termasuk lapisan input, namun lapisan input tidak dihitung karena tidak ada komputasi yang dilakukan di lapisan input [8]. Sinyal input diteruskan ke lapisan output melalui bobot dan neuron di lapisan output menghitung sinyal output [8]. Perbedaan antara *single layer feedforward neural network* dan *multi-layer feedforward neural network* adalah terdapat setidaknya satu lapisan "*hidden neuron*" antara *input* dan *output layer* [8]. Fungsi *neuron* tersembunyi adalah untuk mengintervensi antara *input* eksternal dan *output* jaringan dalam beberapa cara yang berguna [8]. Keberadaan satu atau lebih lapisan tersembunyi memungkinkan jaringan untuk mengekstrak statistik tingkat tinggi [8].

2.3 Transformer

Transformer merupakan model arsitektur yang menghindari pengulangan dan mengandalkan mekanisme *attention* untuk menarik ketergantungan global antara input dan output [9]. Transformer adalah model transduksi yang sepenuhnya mengandalkan *attention* untuk menghitung representasi input dan outputnya tanpa menggunakan RNN atau konvolusi [9]. *Attention* dapat digambarkan sebagai pemetaan kueri dan satu set pasangan *key-value* ke output, di mana kueri, *key*, *value*, dan output semuanya berbentuk vector [9]. Output dihitung sebagai jumlah bobot nilai, di mana bobot yang ditetapkan untuk setiap *value* dihitung oleh fungsi kompatibilitas kueri dengan *key* yang sesuai [9]. Matriks *output* dapat dihitung dengan persamaan 1 [9]. Transformer terdiri atas dua bagian yaitu Encoder dan Decoder. Terdapat *Positional Encoding* di bawah tumpukan *encoder* dan *decoder*. *Positional Encoding* memiliki dimensi yang sama dengan *embedding*, sehingga keduanya dapat dijumlahkan. Karna Transformer tidak melakukan perulangan, Positional Encoding dibutuhkan untuk memberikan urutan pada model [9]. Transformer menggunakan rumus sinus dan kosinus untuk Positional Encoding. Rumus sinus dapat dilihat pada persamaan 2 [9] dan rumus kosinus dapat dilihat pada persamaan 3 [9].

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

Dengan keterangan sebagai berikut:
Q = Query
K = Key
V = Value
d_k = Dimensi k

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right) \qquad (2)$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right) \qquad (3)$$

Dengan keterangan sebagai berikut:
pos = posisi
i = dimensi ke-i
d = dimensi model

2.4 IndoBERT

BERT merupakan singkatan dari *Bidirectional Encoder Representations from* Transformers. BERT dirancang agar dapat melakukan pelatihan dua arah dari teks yang tidak berlabel dengan mengkondisikan bersama pada konteks kiri dan kanan di semua lapisan [10]. Arsitektur model BERT adalah *encoder*

Transformer dua arah multi-layer. Terdapat dua tahapan yang dilakukan pada BERT, yaitu tahapan *pre-training* dan *tahapan fine-tuning*. Selama pre-training, model dilatih dengan menggunakan data yang tidak berlabel dan dilatih dengan tugas yang berbeda, sedangkan untuk tahapan *fine-tuning*, model BERT pertama kali diinisialisasi dengan parameter yang telah dilatih sebelumnya, dan semua parameter disetel dengan menggunakan data berlabel [10].

IndoBERT mempunyai arsitektur yang sama dengan BERT, namun yang membedakan IndoBERT dengan BERT adalah dataset yang digunakan pada tahapan *pre-training*. Dataset yang digunakan untuk melatih IndoBERT disebut sebagai Indo4B yang terdiri dari sekitar 4 miliar kata dengan sekitar 250 juta kalimat, dataset Indo4B mencakup kalimat bahasa Indonesia formal dan sehari-hari yang disusun dari 15 dataset, yang dua di antaranya mencakup bahasa sehari-hari Indonesia, delapan mencakup bahasa Indonesia formal, dan sisanya memiliki gaya campuran antara bahasa sehari-hari dan formal [11].

2.5 Ujaran Kebencian

Tujuan dari definisi hukum sederhana: untuk mengidentifikasi pesan yang melanggar norma hukum yang ada dan memerlukan peraturan pemerintah, yaitu, pesan yang dibagikan secara publik, menghasut, mempromosikan, atau membenarkan kebencian, diskriminasi, atau permusuhan terhadap kelompok dan/atau individu tertentu, berdasarkan atribut tertentu, seperti ras atau asal suku, agama, disabilitas, jenis kelamin, usia, orientasi seksual/identitas gender [12]. Tidak ada definisi hukum yang diterima secara universal tentang ujaran kebencian dan negara yang berbeda, pengemban tugas utama di bawah hak asasi manusia internasional, berada di bawah yurisdiksi yang berbeda [13].

Istilah ujaran kebencian digunakan dalam berbagai disiplin ilmu seperti ekonomi, filsafat, sosiologi, psikologi, atau ilmu komputer, dan, meskipun tidak ada konsensus definisi, dapat melacak beberapa karakteristik konstituen yang umum [12]. ketika mendefinisikan ujaran kebencian, perlu membedakan ujaran kebencian dari istilah terkait yang tidak sesuai dengan definisi tersebut, seperti mengungkapkan ketidaksukaan, kurangnya rasa hormat, pandangan merendahkan orang lain, ketidaksetujuan, penggunaan kata-kata kasar atau ucapan yang menghina, dan ucapan yang "tidak menyerukan tindakan" [14]. Perbedaan halus ini dapat diselesaikan dengan memiliki definisi yang tepat tentang ujaran kebencian. Ujaran kebencian "mengungkapkan, mendorong, membangkitkan, atau menghasut kebencian terhadap sekelompok individu yang dibedakan oleh ciri atau serangkaian ciri tertentu seperti ras, etnis, jenis kelamin, agama, kebangsaan, dan orientasi seksual, dan sering (tetapi tidak perlu) diekspresikan dalam bahasa yang menyinggung, marah, kasar, dan menghina" [12]. Menurut Surat Edaran Mabes Polri No:SE/6/X/2015, tanggal 8 Oktober 2015, ujaran kebencian di definisikan sebagai "tindak pidana yang berbentuk, penghinaan, pencemaran nama baik, penistaan, perbuatan yang tidak menyenangkan, memprovokasi, menghasut, penyebaran berita bohong, dimana semua tindakan di atas memiliki tujuan atau bisa berdampak pada tindak diskriminasi, kekerasan, penghilangan nyawa, dan atau konflik sosial".

3. Hasil Percobaan

Pengujian model dilakukan untuk menguji tingkat keakuratan model dalam melakukan klasifikasi. Pengujian dilakukan dengan menggunakan dua data yang berbeda. Pengujian yang pertama dilakukan dengan menggunakan dataset yang sama seperti yang digunakan untuk melatih model. Sedangkan data untuk pengujian kedua menggunakan data yang diambil langsung dari platform Youtube dan Twitter.

3.1 Hasil Pengujian Model Menggunakan Dataset

Pengujian menggunakan dataset dilakukan untuk mendapatkan parameter dengan akurasi yang terbaik. Parameter yang akan dilakukan pengujian adalah Epoch, Learning Rate dan Batch Size. Hasil dari pengujian model dapat dilihat pada Table 1.

Tabel 1 Hasil Percobaan Dataset

Epoch	Learning Rate	Batch Size	Akurasi Test
2	5e-5	16	0,8694
		32	0,8808
	3e-5	16	0,88
		32	0,8679
	2e-5	16	0,8732
		32	0,8884
3	5e-5	16	0,8793
		32	0,8793
	3e-5	16	0,8755
		32	**0,8952**
	2e-5	16	0,8633
		32	0,88
4	5e-5	16	0,8664
		32	0,88
	3e-5	16	0,8755
		32	0,8922
	2e-5	16	0,8664
		32	0,8709

Berdasarkan dari hasil pengujian yang dilakukan kepada model seperti pada Tabel 1, hasil terbaik dapat

dilihat pada penggunaan parameter 3 untuk Epoch, 3e-5 untuk Learning Rate, dan 32 untuk Batch Size. Pengujian dengan menggunakan tiga parameter tersebut memberikan hasil akurasi sebesar 0,8952. Namun pada penggunaan parameter lain, model tidak memberikan perbedaan yang cukup signifikan. Akurasi terendah dari predikasi model adalah 0,8633. Akurasi terendah dari model tidak memiliki perbedaan yang cukup signifikan dari akurasi tertinggi dengan perbedaan 0,0319. Perubahan pada parameter tidak memberikan peningkatan yang cukup signifikan.

3.2 Hasil Pengujian Model menggunakan Data Asli

Pengujian ini dilakukan untuk melihat apakah model dapat digunakan untuk melakukan prediksi terhadap data yang asli. Pengujian ini dilakukan menggunakan model dengan parameter Epoch 3, Learning Rate 3e-5, dan Batch Size 32. Dilakukan empat pengujian, yaitu dua pengujian dilakukan dengan menggunakan komentar dari "Youtube" dan dua pengujian lainnya dilakukan dengan menggunakan tweet dari "Twitter". Hasil dari pengujian ini dapat dilihat pada Tabel 2.

Tabel 2 Hasil Percobaan Data Asli

Youtube			
No	**Judul**	**Channel**	**Akurasi**
1	AKHIRNYA BALIK KE JEPANG LAGI SETELAH LULUS KULIAH! - JEROME BACK TO JAPAN JP	Nihongo Mantappu	0.85
2	BERDIRI DI SAYAP PESAWAT SUHU EKSTRIM -14℃! JEROME WING WALKING RED BULL CHALLENGE	Nihongo Mantappu	1
Twitter			
No	**Keyword**		**Akurasi**
1	Pemerintah anjing		0.9
2	Pemerintah		0.9

Pada Tabel 2 dapat dilihat bahwa model dengan parameter Epoch 3, Learning Rate 3e-5, dan Batch Size 32 dapat melakukan klasifikasi dengan baik. Model telah dapat digunakan untuk melakukan klasifikasi menggunakan data yang asli. Komentar pada "Youtube" dan tweet pada "Twitter" memiliki sedikit perbedaan karakteristik, namun model dapat melakukan klasifikasi dengan baik pada kedua *platform* tersebut. Meski dapat digunakan untuk data asli, namun model masih belum dapat melakukan klasifikasi secara akurat 100%. Masih terdapat beberapa kesalahan prediksi yang dilakukan oleh model.

Tabel 3 Hasil Klasifikasi

No	Teks	Hasil Klasifikasi	Hasil Sesungguhnya
1	Bang jerr Kakuningshite menit 17:26 Ureshi = Senang Tapi hiragagananya itadakimasu (いただきます) wkwkw.. tetep mantapuu jawaa.. yeayy..	Normal	Normal
2	Kangen kak konten kakak kek dulu	Normal	Normal
3	Di biarkan lama-lama makin menindas aja ini negara. Sebat aja dibatasi ya anjing. Pemerintah tolol! https://t.co/IzvZybLZHl	Ujaran Kebencian	Ujaran Kebencian
4	Laika adalah seekor anjing astronot luar angkasa pertama di dunia. diluncurkan satelit bumi buatan di tahun 1957 oleh pemerintah Rusia.	Normal	Normal
5	Saya akan menunjukan kepada pemerintah, bagaimana caranya menjadi perempuan j*l*ng, dan bagaimana caranya menjadi rakyat yang b*ngs*t\n#data \n#Hacking \n#Hacked \n#CyberSec \n#databases https://t.co/jxnnCrRNB8	Ujaran Kebencian	Normal

Jika dilihat pada Tabel 3, masih terdapat kesalahan klasifikasi pada kalimat ke lima. Kesalahan klasifikasi terjadi karena terdapat kata kasar pada kalimat ke lima. Kalimat ke lima tidak termasuk kalimat ujaran kebencian namun hanya merupakan kalimat yang mengandung bahasa kasar. Model cenderung melakukan klasifikasi ujaran kebencian jika terdapat kata kasar pada kalimat yang ingin diklasifikasikan. Namun pada kalimat ke empat, model dapat melakukan klasifikasi dengan tepat meski terdapat kata "anjing" di dalam kalimatnya. Model dapat membedakan kata "anjing" yang mempunyai arti hewan dengan kata "anjing" yang digunakan untuk kata kasar.

4. Kesimpulan

Berdasarkan hasil pengujian terhadap metode feedforward neural network dengan IndoBERT didapatkan beberapa kesimpulan, yaitu:

1. Perubahan parameter Epoch, Learning Rate dan Batch Size tidak memberikan perubahan akurasi yang signifikan.
2. Metode feedforward neural network dengan IndoBERT berhasil melakukan klasifikasi dengan nilai akurasi terbaik sebesar 89,52%.
3. Model memiliki kecenderungan memberi hasil klasifikasi "ujaran kebencian" jika terdapat kata kasar.

REFERENSI

[1] Saputra, Andi. "Survei Penggunaan Media Sosial Di Kalangan Mahasiswa Kota Padang Menggunakan Teori Uses and Gratifications". Baca: Jurnal Dokumentasi Dan Informasi. Vol. 40, No. 2, 2019.

[2] Bina, Muhammad Arif Hidayatullah. "FENOMENA HATE SPEECH DI MEDIA SOSIAL DAN KONSTRUK SOSIAL MASYARAKAT." Jurnal Peurawi:Media Kajian Komunikasi Islam. Vol. 4, No. 2, (2020).

[3] Ibrohim, Muhammad Okky; and Indra Budi, "Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter". Proceedings of the Third Workshop on Abusive Language Online. 2019

[4] Marpaung, Angela; Rita Rismala; and Hani Nurrahmi. "Hate Speech Detection in Indonesian Twitter Texts Using Bidirectional Gated Recurrent Unit." 13th International Conference Knowledge and Smart Technology. 2021.

[5] Putra, I. Gede Manggala; and Dade Nurjanah. "Hate Speech Detection in Indonesian Language Instagram". International Conference on Advanced Computer Science and Information Systems. 2020.

[6] Sutejo, Taufic Leonardo; and Dessi Puji Lestari. "Indonesia Hate Speech Detection Using Deep Learning". Proceedings of the 2018 International Conference on Asian Language Processing. 2019.

[7] Song, Xinying; Alex Salcianu; Yang Song; Dave Dopson; and Denny Zhou. Fast WordPiece Tokenization. http://arxiv.org/abs/2012.15524, tanggal akses 9 Agustus 2022.

[8] SAZLI, Murat Hüsnü. A Brief Review of Feed-Forward Neural Networks. https://www.researchgate.net/publication/228394623, tanggal akses 10 Agustus 2022.

[9] Vaswani, Ashish; Noam Shazeer; Niki Parmar; Jakob Uszkoreit; Llion Jones; Aidan N. Gomez; Łukasz Kaiser; and Illia Polosukhin. "Attention Is All You Need." Advances in Neural Information Processing Systems. 2017.

[10] Devlin, Jacob; Ming Wei Chang; Kenton Lee; and Kristina Toutanova. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference. 2019.

[11] Wilie, Bryan; Karissa Vincentio; Genta Indra Winata; Samuel Cahyawijaya; Xiaohong Li; Zhi Yuan Lim; Sidik Soleman; Rahmad Mahendra; Pascale Fung; Syafri Bahar; Ayu Purwarianti. IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding. http://arxiv.org/abs/2009.05387, diakses tanggal 9 Agustus 2022.

[12] Papcunová, Jana; Marcel Martončik; Denisa Fedáková; Michal Kentoš; Miroslava Bozogáňová; Ivan Srba; Robert Moro; Matúš Pikuliak; Marián Šimko; and Matúš Adamkovič. Hate Speech Operationalization: A Preliminary Examination of Hate Speech Indicators and Their Structure. https://doi.org/10.1007/s40747-021-00561-0, tanggal akses 6 Agustus.

[13] Brown, Alexander. "What Is Hate Speech? Part 2: Family Resemblances." Law and Philosophy. Vol.36, No. 5, 2017.

[14] Parekh, Bhikhu. Is There a Case for Banning Hate Speech?. https://doi.org/10.1017/CBO9781139042871.006, tanggal akses 8 Agustus 2022.

Steven Dharmawan, seorang Mahasiswa program studi Teknik Informatika Universitas Tarumanagara, Jakarta.

Viny Christanti Mawardi, Memperoleh gelar S.Kom. dari Universitas Tarumanagara tahun 2004. Kemudian memperoleh gelar M.Kom. dari Universitas Indonesia tahun 2008. Saat ini aktif sebagai Dosen Tetap Fakultas Teknologi Informasi Tarumanagara, Jakarta.

Novario Jaya Perdana, Memperoleh gelar S.Kom. dari Institut Teknologi Sepuluh Nopember tahun 2011. Kemudian memperoleh gelar M.T. dari Universitas Indonesia tahun 2016. Saat ini aktif sebagai Dosen Tetap Perjanjian Fakultas Teknologi Informasi Tarumanagara, Jakarta.