Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.1148719
Vol. 11, No. 4, Agustus 2024, hlm. 939-948 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 36/E/KPT/2019 e-ISSN: 2528-6579

BIJAKAWEB: PLATFORM BERBASIS WEB UNTUK DETEKSI *HATE SPEECH* PADA KOMENTAR BERITA BAHASA INDONESIA

Moh. Firdaus[1], Permata Nur Miftahur Rizki [*2]

[1,2]Program Studi Rekayasa Perangkat Lunak, Universitas Prasetiya Mulya, Kabupaten Tangerang
Email: [1]mohammedfirdaus1404@gmail.com, [2]permata.nmr@prasetiyamulya.ac.id
[*]Penulis Korespondensi

(Naskah masuk: 9 Februari 2024, diterima untuk diterbitkan: 19 Agustus 2024)

Abstrak

Jumlah pengguna internet di Indonesia telah mencapai lebih dari 221 juta jiwa, mayoritas penduduk Indonesia menggunakan internet dengan tujuan agar tetap *update* dengan berita terbaru. Detik, Kompas, dan CNNIndonesia merupakan portal berita daring favorit sebagian besar penduduk Indonesia. Fitur komentar pada portal berita yang ada saat ini memungkinkan pembaca berita dapat memberikan umpan-balik terhadap berita, namun sering kali tidak terkontrol, memicu munculnya ujaran kebencian. Meskipun tersedia fitur moderasi seperti "Laporkan", pendekatan manual ini sering kali lambat dan kurang efektif. Penelitian ini bertujuan untuk mengembangkan sistem deteksi otomatis terhadap ujaran kebencian pada komentar berita daring. Proses penelitian dimulai dengan *scraping* lebih dari 15 ribu data komentar dari portal berita menggunakan *library* Python, dilanjutkan dengan pelabelan manual ke dalam dua kategori: “Hate” dan “Non-Hate,” dengan jumlah data yang berhasil dilabeli sebanyak 11.478, yang dibagi ke dalam dua kelas seimbang. *Dataset* yang telah berlabel kemudian digunakan untuk *fine-tuning* model IndoBERT selama 14 *epoch*, dengan akurasi terbaik sebesar 95,91% yang dicapai pada *epoch* ke-14. Model dengan akurasi terbaik diimplementasikan pada platform web yang diberi nama BijakaWeb (Web Bijak Dalam Berkomentar) dengan menggunakan *framework* Django. Penelitian ini menghasilkan beberapa kontribusi penting, termasuk tersedianya *dataset* baru untuk penelitian relevan, model *fine-tuned* IndoBERT baru yang dapat diakses publik di HuggingFace, serta pengembangan platform *Website* Bijaka dengan menggunakan *framework fullstack* Django yang mampu melakukan *scraping* dan prediksi ujaran kebencian secara real-time. Harapannya, penelitian ini dapat membantu portal berita dalam moderasi komentar berita daring dalam melawan komentar yang mengandung ujaran dan menyediakan model yang dapat digunakan serta diadaptasi oleh platform berita daring lainnya untuk mencegah penyebaran ujaran kebencian di internet.

Kata kunci: *IndoBERT, Ujaran Kebencian, Django, Web Scraping, Portal Berita*

BIJAKAWEB: A WEB-BASED PLATFORM FOR DETECTING HATE SPEECH IN INDONESIAN NEWS COMMENTS

Abstract

The number of internet users in Indonesia has surpassed 221 million, with the majority of the population using the internet to stay updated with the latest news. Detik, Kompas, and CNNIndonesia are among the most popular online news portals for many Indonesians. The comment features on these news portals allow readers to provide feedback on news articles; however, this is often unregulated, leading to the spread of hate speech. Although moderation features like "Report" are available, these manual approaches are often slow and ineffective. This study aims to develop an automatic detection system for hate speech in online news comments. The research process began by scraping over 15,000 comment data from news portals using Python libraries, followed by manually labeling the comments into two categories: "Hate" and "Non-Hate." A total of 11,478 labeled data points were obtained, which were divided into two balanced classes. The labeled dataset was then used to fine-tune the IndoBERT model over 14 epochs, with the best accuracy of 95.91% achieved on the 14th epoch. The model with the best accuracy was implemented on a web platform named BijakaWeb (Web Bijak Dalam Berkomentar) using Django fullstack framework. This research has produced several significant contributions, including the availability of a new dataset for relevant research, a fine-tuned IndoBERT model accessible to the public on HuggingFace, and the development of the BijakaWeb platform using the full-stack Django framework, capable of real-time scraping and hate speech prediction. It is hoped that this research can assist news portals in moderating online news comments to combat hate speech and provide a model that can be used and adapted by other online news platforms to prevent the spread of hate speech on the internet.

***Keywords**: IndoBERT, Hate Speech, Django, Web Scraping, News Portal*

1. PENDAHULUAN

Saat ini, pengguna internet di Indonesia telah mencapai lebih dari 221 juta jiwa, atau sekitar 79,5% dari total populasi Indonesia saat ini (APJII, 2024). Tujuan utama pengguna internet di Indonesia cukup beragam, mulai dari mengakses media sosial, mendapatkan *update* terbaru mengenai berita, bekerja atau belajar secara daring, hingga mengakses layanan publik dan juga hiburan (APJII, 2024).

Mengakses berita terbaru secara daring telah menjadi salah satu tujuan utama pengguna internet di Indonesia. Berdasarkan hasil laporan dari Badan Pusat Statistika melalui Laporan Statistik Telekomunikasi Indonesia (BPS, 2022), sekitar 74,9% penduduk Indonesia menggunakan internet agar dapat mengakses berita terbaru secara daring.

Perkembangan internet telah memengaruhi cara penduduk Indonesia untuk mengakses berita terbaru. Dulu, berita umumnya hanya dapat diperoleh melalui media cetak, televisi, ataupun radio. Saat ini, dengan perkembangan internet, berita dapat diakses secara daring melalui berbagai platform seperti situs dan aplikasi berita, maupun media sosial. Menurut Survei Reuters Institute, setidaknya 84% penduduk Indonesia lebih memilih media daring untuk mengakses berita dibandingkan televisi (54%) dan media cetak (15%) (Newman et al., 2023). Data ini menunjukkan bahwa pergeseran menuju media daring sebagai sumber utama berita semakin menguat, mencerminkan perubahan signifikan dalam preferensi konsumen berita.

Perubahan preferensi penduduk Indonesia dalam mengakses berita dari media tradisional ke media daring telah mendorong munculnya banyak portal berita daring di Indonesia. Dari banyaknya portal berita daring yang ada saat ini, Detik, Kompas, dan CNNIndonesia merupakan tiga portal yang populer dan terpercaya yang sering diakses penduduk Indonesia (Newman et al., 2023). Selain menjadi sumber informasi yang andal, portal-portal berita ini juga terus berinovasi dengan melibatkan interaksi pembaca sehingga terdapat timbal-balik dari berita yang dipublikasikan. Interaksi ini dilakukan pada fitur "Komentar".

Fitur "Komentar" memungkinkan pembaca untuk tidak hanya menerima informasi secara pasif, tetapi juga berkontribusi dalam memberikan tanggapan, saran, dan opini terhadap berita yang dibacanya. Meskipun memberikan ruang untuk berekspresi, kolom komentar juga membuka celah bagi potensi penyalahgunaan, khususnya dalam bentuk penggunaan bahasa kasar, intimidasi, hingga ujaran kebencian atau *hate speech* (Kiasati Desrul and Romadhony, 2019).

Hate speech adalah bentuk ekspresi yang bertujuan merendahkan dan diskriminatif terhadap individu atau kelompok, seringkali berdasarkan pada SARA atau aspek lainnya (Herwanto et al., 2019). Asosiasi Jurnalis Indonesia berkerja sama dengan Monash Data & Democracy Research Hub melakukan pemantauan ujaran kebencian (*hate speech*) menjelang Pemilu 2024. Hasil dari pemantauan tersebut terdapat lebih dari 200 ribu ujaran kebencian yang tersebar diberbagai media sosial seperti Facebook, X (Sebelumnya Twitter), Facebook, serta artikel yang berasal dari Cek Fakta (Idris, PhD et al., 2024). Data ini memberikan gambaran yang cukup mengkhawatirkan terkait tingginya jumlah persebaran ujaran kebencian yang berasal dari internet.

Dengan meningkatnya penyebaran *hate speech* di internet, tantangan besar muncul dalam upaya menciptakan lingkungan internet yang lebih positif dan aman. Salah satu langkah yang telah diambil oleh media daring khususnya portal berita untuk mengurangi arus ujaran kebencian adalah dengan menerapkan moderasi konten melalui fitur "Laporkan" atau "*Report*" pada komentar berita daring. Fitur ini memungkinkan pembaca untuk melaporkan konten yang dianggap tidak pantas atau mengandung ujaran kebencian, yang kemudian akan ditinjau secara manual oleh tim moderasi. Namun, proses verifikasi manual ini sering kali memerlukan waktu yang cukup lama, sehingga mengurangi efektivitas dalam mengatasi dan menanggulangi penyebaran konten negatif secara cepat.

Dengan demikian, kebutuhan akan deteksi *hate speech* atau ujaran kebencian pada komentar di portal berita daring semakin penting. Pendekatan dengan menggunakan teknologi *machine learning* dapat menjadi solusi dengan menyediakan sistem deteksi otomatis yang mampu mengidentifikasi konten *hate speech* dengan lebih cepat dan efektif.

Beberapa penelitian terdahulu yang melakukan penerapan teknologi *machine learning* dalam mengidentifikasi konten *hate speech* seperti pada penelitian (Aulia and Budi, 2019) melakukan deteksi *hate speech* pada media sosial Facebook dan memberikan akses publik *dataset* yang digunakan. Lalu penelitian yang dilakukan oleh (Isnain, Sihabuddin and Suyanto, 2020) dalam mendeteksi *hate speech* menggunakan *dataset tweet* dari platform X (sebelumnya Twitter) dengan menggunakan pendekatan *deep learning*. Dan penelitian yang dilakukan oleh (Putra and Nurjanah, 2020) melakukan deteksi *hate speech* pada komentar media sosial Instagram.

Penelitian sebelumnya, seperti yang dilakukan oleh (Aulia and Budi, 2019; Isnain, Sihabuddin and Suyanto, 2020; Putra and Nurjanah, 2020) umumnya terfokus hanya pada deteksi *hate speech* di media sosial seperti Facebook, Instagram, dan X (sebelumnya Twitter). Hal ini berkaitan dengan penelitian studi literatur yang dilakukan oleh (Pamungkas, Putri and Fatmawati, 2023) yang menunjukkan mayoritas ketersediaan *dataset hate speech* dalam Bahasa Indonesia hanya bersumber dari ketiga media sosial tadi. Untuk mengatasi keterbatasan *dataset*, penelitian ini akan menggunakan *dataset* baru yang diperoleh melalui

scraping komentar di portal berita daring. *Dataset* ini juga akan disediakan untuk akses publik, guna mendukung penelitian lebih lanjut di bidang ini.

Penelitian ini akan menggunakan model *pre-trained language model* untuk melakukan deteksi *hate speech*. *Pre-trained language model* disarankan untuk digunakan dalam tugas-tugas pemrosesan bahasa alami, termasuk klasifikasi teks (Mathew and Bindu, 2020) karena telah dilatih pada korpus teks besar sehingga mampu memahami makna semantik kata atau kalimat. *Pre-trained language model* yang populer dan telah dilatih dalam Bahasa Indonesia yaitu IndoBERT. IndoBERT merupakan *language model* terbaru yang dikembangkan khusus dalam bahasa Indonesia yang dapat melakukan berbagai tugas *natural language processing* salah satunya yaitu klasifikasi teks (Wilie et al., 2020).

IndoBERT merupakan *pre-trained languange model* khusus Bahasa Indonesia (*mono language*) dengan menggunakan arsitektur BERT (Devlin et al., 2019) yang telah dilatih dengan setidaknya empat miliar korpus kata bahasa Indonesia (Indo4B *corpus*) yang bersumber dari berbagai platform Indonesia antara lain Wikipedia bahasa Indonesia, artikel berita dari Kompas, Tempo, Liputan6, dan Korpus Web Indonesia (Wilie et al., 2020). IndoBERT terdiri dari dua varian: IndoBERT-base (12 lapisan, 12 attention heads, 110 juta parameter) dan IndoBERT-large (24 lapisan, 16 attention heads, 340 juta parameter). IndoBERT-large dapat memberikan akurasi lebih tinggi karena menggunakan lebih banyak parameter, akurasi yang tinggi berbanding lurus dengan waktu yang dibutuhkan untuk melakukan proses *fine-tuning* varian IndoBERT-large. Beberapa penelitian yang pernah dilakukan dengan menggunakan IndoBERT seperti analisis sentimen (Geni, Yulianti and Sensuse, 2023), *fake news detection* (Isa, Nico and Permana, 2022), dan *detection clickbait headline news* (Fakhruzzaman and Gunawan, 2021).

Model IndoBERT yang telah melalui tahapan proses *fine-tuned* akan diimplementasikan pada *website-based platform* dengan menggunakan *framework fullstack* Django dengan nama produk BijakaWeb (*Website* Bijak Dalam Berkomentar). *Website* Bijaka akan dirancang dengan tampilan antarmuka yang mudah digunakan dan dapat mendeteksi *hate speech* pada komentar berita secara real-time dan mudah.

Dari pemaparan latar belakang sebelumnya, penelitian ini bertujuan untuk mengembangkan sistem deteksi ujaran kebencian (*hate speech*) secara otomatis berbasis web pada komentar berita daring. Sistem deteksi otomatis ini dapat membantu dan menggantikan moderasi konten yang dilakukan secara manual sehingga lebih cepat dan efektif. Selain itu, penelitian ini setidaknya memiliki tiga kontribusi seperti tersedianya *dataset* baru untuk penelitian yang releval, *fine-tuned model* IndoBERT baru khusus deteksi ujaran kebencian yang dapat diakses publik di HuggingFace, serta *web-based platform* Bijaka yang mampu melakukan *scraping* dan prediksi ujaran kebencian secara real-time dan mudah.

2. METODE PENELITIAN

Pada Gambar 1 memperlihatkan tiga tahapan utama yang akan dilakukan pada penelitian ini yang terdiri dari: Pengumpulan Data, Pengembangan Model (*Fine-tuning Model*), dan Perencanaan serta Pengembangan Website. Fase pertama berfokus pada pengumpulan dan pelabelan *dataset* yang diperlukan. Fase kedua mencakup *pre-processing* dan pembagian *dataset*, *fine-tuning* dan evaluasi model. Dan, fase terakhir melibatkan perancangan dan pengembangan website, di mana model yang telah *fine-tuning* diintegrasikan pada website untuk menghasilkan produk akhir yang siap digunakan.

2.1 *Data Collection* (Pengumpulan Data)

2.1.1 Pengumpulan *Dataset*

Penelitian ini menggunakan data komentar pada portal berita Kompas.com untuk dijadikan *dataset* pelatihan model yang diambil dengan menggunakan metode *web scraping*.

2.1.2 Pelabelan *Dataset*

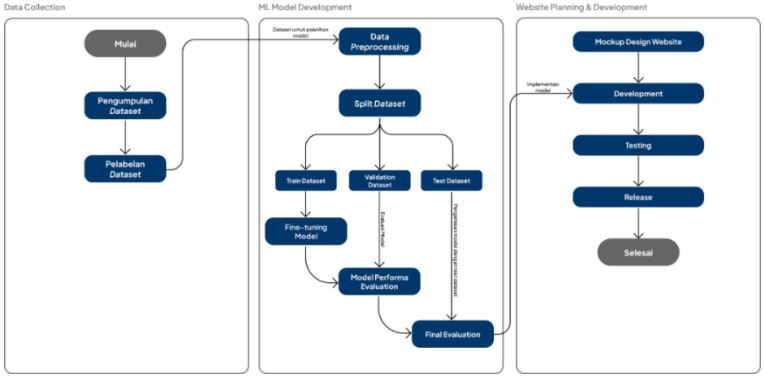
Komentar dari berita-berita yang telah dikumpulkan dilakukan pelabelan manual kedalam dua kelas yaitu “Hate” untuk komentar yang mengandung *hate speech* dan “Non-Hate” untuk komentar yang tidak mengandung *hate speech*. Panduan tentang proses pelabelan diberikan, termasuk definisi dan contoh dari komentar "Hate" dan "Non-Hate".

2.2 Pengembangan Model (*Fine-tuning Model*)

2.2.1 *Data Preprocessing*

Tahapan data *preprocessing* merupakan tahapan pembersihan data untuk meminimalisir atau menghilangkan *noise*. Proses data *preprocessing* sangat penting untuk mendapatkan akurasi model terbaik (Lubis et al., 2023). Pada penelitian ini terdapat beberapa tahapan *data preprocessing* seperti:

- Penghapusan emotikon,
- Penghapusan komentar yang duplikat,
- Penghapusan *whitespace*,
- *Case folding*,
- Penghapusan tanda baca yang berlebihan,
- Normalisasi kata, dan
- *Tokenization* (adalah proses membagi teks dalam bentuk yang lebih kecil, seperti kata atau sub-kata).



Gambar 1. Alur Penelitian

2.2.2 Pembagian *Dataset*

Dataset yang berhasil diberikan label dibagi menjadi tiga jenis *dataset* yaitu *data training*, *data validation*, dan *data testing* dengan rasio 80:10:10. Dengan jumlah komentar masing-masing *dataset* sebanyak 9.181:1.148:1.148.

2.2.3 Fine-tuning Model IndoBERT

Dalam penelitian ini, digunakan *pre-trained language model* IndoBERT varian IndoBERT-large untuk mencapai tingkat akurasi yang optimal, sehingga diperlukan infrastruktur *hardware* yang handal, seperti Google Colab. Spesifikasi *hardware* Google Colab yang digunakan pada penelitian ini dapat dilihat pada Tabel 1.

Tabel 1. Google Colab Runtime Environment

Hardware	***Spesification***
GPU	Tesla T4
RAM	16 GB
DISK	76.2 GB

2.2.4 Evaluasi Model

Evaluasi model pada penelitian ini menggunakan *confusion matrix* untuk menganalisis performa model dalam melakukan *text classification*. *Confusion matrix* memberikan gambaran rinci tentang seberapa baik model dapat mengklasifikasikan komentar ke dalam kategori yang benar. Terdapat empat elemen utama dalam *confusion matrix*, yaitu *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Dari empat elemen tersebut, berbagai metrik evaluasi dapat dihitung, seperti *Accuracy* (1), *Precision* (2), *Recall* (3), dan *F1-score* (4).

$$Accuracy: \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision: \frac{TP}{TP + FP} \quad (2)$$

$$Recall: \frac{TP}{TP + FN} \quad (3)$$

$$F1\ Score: \frac{2 \times (Precision \times Recall)}{Precision \times Recall} \quad (4)$$

2.3 Pengembangan BijakaWeb (Bijak Dalam Berkomentar Web)

Pengembangan *website* Bijaka dengan menggunakan *framework fullstack* Django. Pada tahapan *testing* atau pengujian akan menggunakan metode *black-box testing* dengan menggunakan metode UAT *Testing* untuk dilakukan pengujian fungsional fitur.

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan dan Persiapan *Dataset*

Jumlah *dataset* yang berhasil dikumpulkan pada penelitian ini sebanyak sebanyak 25.512 data dari 1.266 berita yang disimpan dengan format *file CSV* (*Comma Separated Values*). Informasi yang *scrape* pada portal berita meliputi tautan berita, judul berita, tanggal posting berita, kategori berita, nama pengomentar beserta isi komentarnya.

Pada Tabel 2 merupakan lima berita dengan jumlah komentar terbanyak selama pengumpulan komentar berita. Dari *dataset* yang ada, pembaca berita cenderung aktif berinteraksi dengan berita yang berkaitan dengan politik, hal ini berkaitan dengan pelaksanaan Pemilu 2024 yang semakin dekat.

Tabel 2. Lima Berita Kompas.com dengan Jumlah Komentar Terbanyak

Judul Berita	**Jumlah Komentar**
Cak Imin ke Kader PKB: Kalau Amin Tidak Menang, Indonesia Dalam Bahaya	380
Kaesang Tanggapi Megawati soal Penguasa Orde Baru: Menghina Presiden Ditangkap Enggak?	317
Jengkel Merasa Tak Dihormati, Megawati: Saya Pernah Jadi Presiden, Lho!	298
Anwar Usman Ajukan Keberatan Suhartoyo Jadi Ketua MK	225
Megawati: Kenapa Sekarang Penguasa Ingin Bertindak seperti Waktu Orde Baru?	208

Pelabelan manual dari *dataset* yang telah dikumpulkan, terdapat 12.123 data komentar berhasil dibagi menjadi dua kelas yaitu “Hate” dan “Non-

Hate" seperti pada Tabel 3. Jumlah data komentar "Hate" yang terbatas sehingga jumlah data "Non-Hate" yang digunakan disesuaikan kembali agar tidak terjadi *imbalanced dataset*.

Tabel 3. Pembagian Kelas Label Komentar Berita

Label	Jumlah Komentar
Non-Hate	5.738
Hate	5.739

Tabel 4 merupakan contoh hasil dari pelabelan manual yang dilakukan.

Tabel 4. Contoh Komentar Berita dengan Hasil Pelabelan

Komentar	Label
antek dan fanboy zionis halal darahnya ayo sodara² musnahkan mereka dari negeri ini,, takbir !!	Hate
penyembah puting payudara pada teriak2 kaum komunis pencuri tanah tetangganya uuurraakan wkwk....	Hate
pak ignasius jonas hrs turun gunung	Non-Hate
jangan subsidi dulu gunakan untuk membayar hutang aja biar bunga pinjaman baru di pakai subsidi petani.	Non-Hate

Setelah melakukan pelabelan manual, sebelum melakukan proses *fine-tuning model* IndoBERT. Dilakukan tahapan data *preprocessing.* Tabel 5 merupakan beberapa tahapan data *preprocessing* yang dilakukan.

Tabel 5. Tahapan Data *Preprocessing*

Data Preprocessing	Sebelum	Sesudah
Penghapusan Emotikon	Kok ga bs diderek? kan diangkut pake truk kan bisa, ga bs diderek atau takut???? □□□ %^%$%$^%$%	Kok ga bs diderek? kan diangkut pake truk kan bisa, ga bs diderek atau takut???? %^%$%$^%$%
Penghapusan Karakter yang Tidak Relevan	Kok ga bs diderek? kan diangkut pake truk kan bisa, ga bs diderek atau takut???? %^%$%$^%$%	Kok ga bs diderek? kan diangkut pake truk kan bisa, ga bs diderek atau takut????
Penghapusan *Punctuation* yang Berulang	kok ga bs diderek? kan diangkut pake truk kan bisa, ga bs diderek atau takut????	kok ga bs diderek? kan diangkut pake truk kan bisa, ga bs diderek atau takut?
Normalisasi Kata	kok ga bs diderek? kan diangkut pake truk kan bisa, ga bs diderek atau takut?	Kok tidak bisa diderek? kan diangkut pakai truk kan bisa, tidak bisa diderek atau takut?

Setelah melakukan data *preprocessing* dilakukan pembagian *dataset* sebelum melakukan proses *fine-tuning* model IndoBERT untuk melakukan *hate speech detection*, *dataset* yang telah diberikan label dibagi menjadi tiga bagian yaitu *training*, *validation, dan test dataset* dengan rasio 80:10:10 seperti pada Tabel 6.

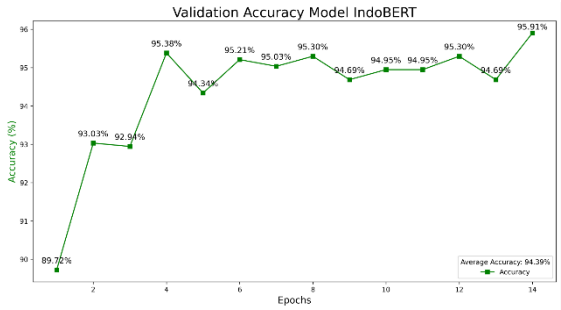
Tabel 6. Pembagian Dataset

Training Dataset	*Validation Dataset*	*Test Dataset*
9.181	1.148	1.148

3.2 *Fine-tuning* IndoBERT

Hasil dari proses *data preprocessing* akan dilakukan proses pelatihan model IndoBERT atau dengan kata lain proses *fine-tuning model*. Proses ini bertujuan untuk melatih model IndoBERT sesuai dengan *spesific domain* yang ingin diatasi yaitu klasifikasi teks untuk mendeteksi ujaran kebencian (*hate speech*)

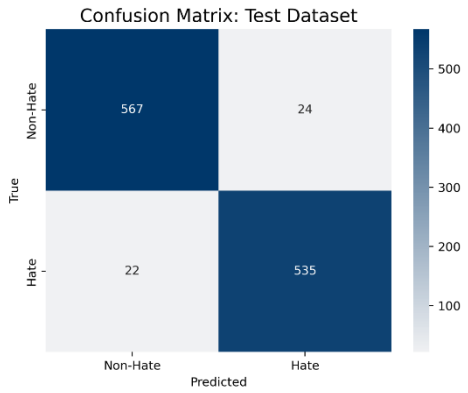
Dalam proses *fine-tuning* IndoBERT dilakukan sebanyak 14 *epoch* dengan menggunakan varian IndoBERT-large dari Hugging Face (Wolf et al., 2020). Gambar 2 menunjukkan bahwa akurasi terbaik selama proses *fine-tuning* IndoBERT pada *epoch* ke-14. Pada proses *fine-tuning*, akurasi model IndoBERT terus menunjukkan peningkatan hingga diakhir proses *fine tuning* pada *epoch* ke-14 dengan akurasi terbaiknya sebesar 95,91%.



Gambar 2. Akurasi Model IndoBERT Saat Proses *Fine-tuning*

Setelah mendapatkan akurasi terbaik selama proses *fine-tuning model*, dilakukan pengetesan model dengan menggunakan *test dataset* yang merupakan *dataset* yang belum dilihat model sebelumnya. Proses ini bertujuan untuk melihat seberapa baik *fine-tuned model* IndoBERT melakukan klasifikasi pada data baru. Gambar 3 merupakan *confusion matrix* model yang telah dilakukan *fine-tuning* dalam memprediksi *test dataset*.

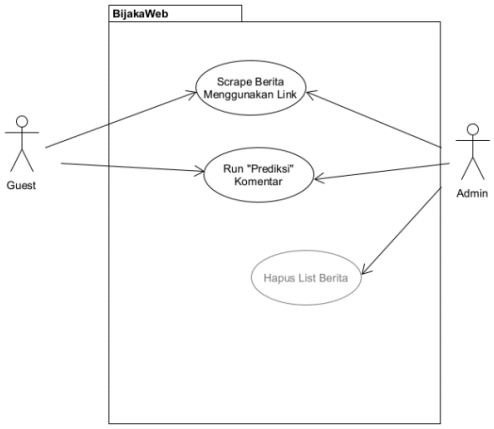
Confusion matrix pada Gambar 3 memperlihatkan performa model IndoBERT varian *large* dalam klasifikasi teks untuk dua kategori: "Non-Hate" dan "Hate". Akurasi model pada *test dataset* yang sangat tinggi, yaitu sekitar 95.99%. Model ini mampu mengklasifikasikan teks sebagai "Hate" atau "Non-Hate" dengan tingkat kesalahan yang sangat rendah, hanya 46 dari 1148 prediksi yang salah. Dengan demikian, model ini menunjukkan performa yang baik dalam mendeteksi konten kebencian.



Gambar 3. Confusion Matrix pada *Test Dataset*

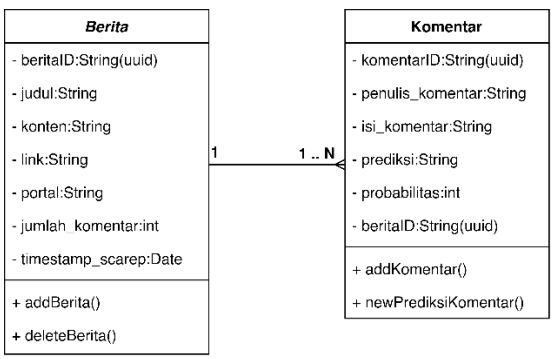
3.3 Perancangan BijakaWeb

Pada tahapan perancangan BijakaWeb dilakukan pembuatan *use case diagram* dan perancangan *class diagram*. *Use case diagram* BijakaWeb seperti pada Gambar 4.



Gambar 4. *Use Case Diagram* BijakaWeb

Use case diagram pada Gambar 4 menunjukkan terdapat dua aktor utama pada BijakaWeb, *Guest* dan Admin. BijakaWeb menyediakan dua fitur utama: "*Scrape* Berita" untuk mengekstraksi informasi dari berita dengan *web scraping* menggunakan PythonSelenium dan/atau BeautifulSoup dan "Prediksi Komentar" untuk menjalankan model *machine learning* pada berita yang telah di-*scrape* untuk mendeteksi *hate speech*.

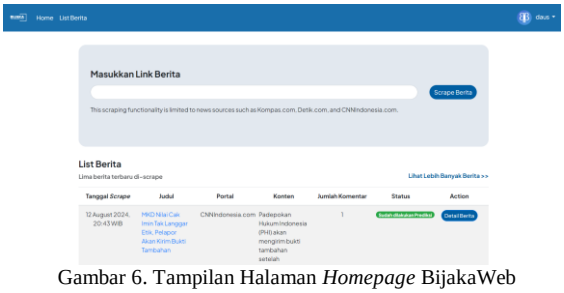


Gambar 5. *Class Diagram* BijakaWeb

Gambar 5 merupakan *Class Diagram* BijakaWeb yang menunjukkan hubungan antara tabel data Berita dan Komentar. Tabel Berita mencakup atribut seperti beritaID, judul, konten, link, portal, jumlah_komentar, dan timestamp_scrape, yang merepresentasikan informasi dasar artikel berita. Sementara itu, Tabel Komentar memuat atribut seperti komentarID, penulis_komentar, isi_komentar, prediksi, probabilitas, dan beritaID, yang tidak hanya berisi teks komentar tetapi juga hasil analisis prediksi ujaran kebencian. Hubungan antara kedua entitas ini adalah *one-to-many* di mana satu berita dapat memiliki banyak komentar.

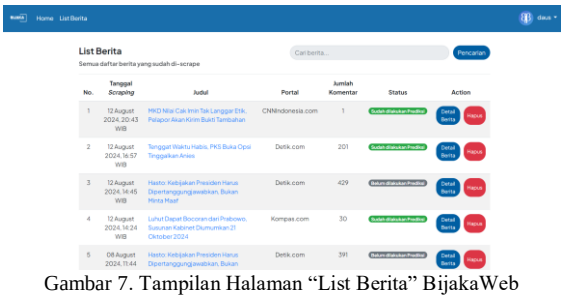
3.4 Pengembangan BijakaWeb

Berikut tampilan *user interface* BijakaWeb yang telah diimplementasikan pada *web-based platform* menggunakan *framework* Django dengan bahasa pemrograman Python.



Gambar 6. Tampilan Halaman *Homepage* BijakaWeb

Pada tampilan *homepage* BijakaWeb terdapat *field input* URL untuk melakukan *scrape* berita yang ingin dilakukan analisis *hate speech* pada komentarnya. Selain itu terdapat *top 5* berita yang baru saja di-*scrape*, jika ingin melihat keseluruhan berita bisa pada halaman "List Berita" atau "Lihat Lebih Banyak Berita".

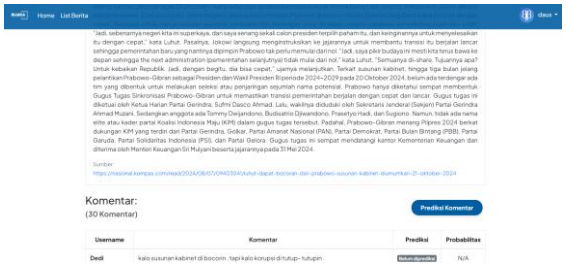


Gambar 7. Tampilan Halaman "List Berita" BijakaWeb

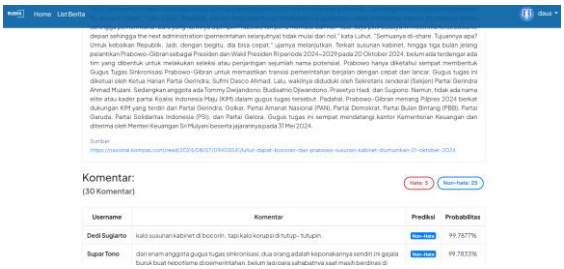
Pada Gambar 7 akan menampilkan seluruh berita yang telah di-*scrape*. Pada "List Berita" terdapat *list* berita yang berisi judul berita, portal, konten atau isiberita, jumlah komentar, status prediksi, dan *action button* (untuk melihat detail berita). Terdapat juga *field search* yang dapat mempermudah pengguna untuk mencari berita dengan kata kunci tertentu.

Pada Gambar 8 merupakan tampilan setelah proses *scrape* berita berhasil dilakukan, pada halaman ini terdapat informasi judul berita, isi berita, sumber

berita, dan tentunya komentar berita yang belum dilakukan prediksi *hate speech*.



Gambar 8. Tampilan Halaman "Detail Berita" BijakaWeb (Sebelum "Prediksi Komentar")



Gambar 9. Tampilan Halaman "Detail Berita" BijakaWeb (Setelah "Prediksi Komentar")

Setelah menjalankan analisis *hate speech* dengan menekan *button* "Prediksi Komentar". Model *machine learning* yang sudah diimplementasikan pada BijakaWeb akan melakukan pendeteksian *hate speech*. Hasil dari deteksi *hate speech* akan tampil seperti Gambar 9, jumlah yang teridentifikasi *hate* dan *non-hate* akan muncul. Serta probabilitas terhadap kelas yang dideteksi seberapa besar.

3.5 Batasan Penggunaan Data Portal Berita

Merujuk pada ketentuan layanan pengguna dari portal yang digunakan seperti CNN Indonesia, Kompas.com, dan Detik.com tidak ada yang menyatakan secara eksplisit tentang legalitas melakukan *scraping* data dari portal berita daring tersebut. Disisi lain, setelah ditinjau kembali *robots.txt* tercantum pada masing-masing portal memiliki beberapa batasan *bot crawler* untuk melakukan *indexing* pada portal berita tersebut seperti GoogleBot, ChatGPT-User, OpenAI, CCBot, GPTBot, dan beberapa robot *crawler* lainnya. Maka dari itu, perlu dipertimbangan kembali bagi penelitian serupa yang ingin menggunakan data dari sumber portal berita untuk kebutuhan non-pribadi atau komersil untuk mematuhi syarat dan ketentuan layanan portal berita daring.

4. KESIMPULAN DAN SARAN

Penelitian ini berhasil mengumpulkan *dataset* komentar berita dan telah diberikan label "Hate" dan "Non-Hate" sebanyak 15.356 data komentar. Komentar berita yang berhasil dikumpulnya mayoritas bernuansa politik hal ini berkaitan dengan dekatnya kontestasi Pemilu 2024 serta konflik Israel - Palestina. *Dataset* yang digunakan pada penelitian ini dapat diakses secara publik pada *link* https://www.github.com/mohfirdaus/datasethateindonesia.

Pemilihan *pre-trained language model* IndoBERT untuk mendeteksi *hate speech* pada penelitian ini merupakan langkah yang tepat. Model IndoBERT memiliki performa paling baik pada *epoch* ke-14 dengan akurasi model 95,91%. Model IndoBERT dilakukan ujicoba pada *test dataset* dan dapat memberikan akurasi yang konsisten pada 95,99%. *Fine-tuned model* IndoBERT ini dapat digunakan kembali untuk penelitian dan/atau pengembangan yang serupa dengan mengakses *link* https://huggingface.co/sidaus/hatespeech-commentnews-large-ind-2.

Fine-tuned model IndoBERT berhasil diimplementasikan pada *web-based platform* menggunakan *framework fullstack* Django. Sehingga dapat melakukan *scraping* berita daring beserta komentarnya. Selain itu, deteksi *hate speech* berhasil diimplementasikan pada BijakaWeb dengan menggunakan *fine-tuned model* IndoBERT dengan bantuan API HuggingFace. BijakaWeb dapat dijalankan terbatas pada *local environment*.

Saran penelitian selanjutnya agar memperluas variasi *dataset* yang digunakan dari berbagai topik berita agar model dapat mendeteksi *hate speech* untuk topik yang lebih umum. Dapat menggunakan sumber *dataset* lain diluar portal berita daring seperti media sosial, forum, hingga platform *streaming*. Dapat mempertimbangkan beberapa metode pelabelan *dataset* selain dilakukan secara manual yang cenderung memakan waktu yang lama. Serta, perlu diperhatikan juga pemilihan *pre-trained language model* dapat disesuaikan dengan ketersediaan infrastruktur perangkat keras yang akan digunakan selama proses *fine-tuning* karena dapat mempengaruhi lama proses *fine-tuning*.

UCAPAN TERIMA KASIH

Ucapan terima kasih yang mendalam atas dukungan penuh dari civitas akademika Program Studi Rekayasa Perangkat Lunak, Universitas Prasetiya Mulya sehingga penelitian ini terlaksana dengan baik.

DAFTAR PUSTAKA

Asosiasi Penyelenggara Jasa Internet Indonesia, 2024. *Survei Penetrasi Internet Indonesia 2024*. Asosiasi Penyelenggara Jasa Internet Indonesia.

Aulia, N. and Budi, I., 2019. Hate Speech Detection on Indonesian Long Text Documents Using Machine Learning Approach. In: *Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence*. [online] ICCAI '19: 2019 5th International

Conference on Computing and Artificial Intelligence. Bali Indonesia: ACM. pp.164–169. https://doi.org/10.1145/3330482.3330491.

Badan Pusat Statistik. 2022. *Statistik Telekomunikasi Indonesia*. Badan Pusat Statistik.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. pp.4171–4186.

Fakhruzzaman, M.N. and Gunawan, S.W., 2021. Web-based Application for Detecting Indonesian Clickbait Headlines using IndoBERT. [online] https://doi.org/10.48550/ARXIV.2102.10601.

Geni, L., Yulianti, E. and Sensuse, D.I., 2023. Sentiment Analysis of Tweets Before the 2024 Elections in Indonesia Using IndoBERT Language Models. 9(3).

Hafidz, 2021. Kejahatan Siber Meningkat di Masa Pandemi. *Universitas Indonesia*. Available at: <https://www.ui.ac.id/kejahatan-siber-meningkat-di-masa-pandemi/> [Accessed 1 February 2024].

Herwanto, G.B., Maulida Ningtyas, A., Nugraha, K.E. and Nyoman Prayana Trisna, I., 2019. Hate Speech and Abusive Language Classification using fastText. In: *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*. [online] 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI). Yogyakarta, Indonesia: IEEE. pp.69–72. https://doi.org/10.1109/ISRITI48646.2019.9034560.

Idris, PhD, I., Wijaya, PhD, D., Izzanardi Wijanarko, S.Si, M. and Tim AJI Indonesia, 2024. *Laporan Pemantauan Ujaran Kebencian Terhadap Kelompok Rentan pada Pemilu 2024*. [online] AJI Indonesia - Monash Data & Democracy Research Hub. p.33. Available at: <https://aji.or.id/system/files/2024-08/laporan-pemantauan-ujaran-kebencian-terhadap-kelompok-rentan-pada-pemilu-2024bahasa-indonesia.pdf>.

Isa, S.M., Nico, G. and Permana, M., 2022. *IndoBERT for Indonesian Fake News Detection*. https://doi.org/10.24507/icicel.16.03.289.

Isnain, A.R., Sihabuddin, A. and Suyanto, Y., 2020. Bidirectional Long Short Term Memory Method and Word2vec Extraction Approach for Hate Speech Detection. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, 14(2), p.169. https://doi.org/10.22146/ijccs.51743.

Kiasati Desrul, D.R. and Romadhony, A., 2019. Abusive Language Detection on Indonesian Online News Comments. In: *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*. [online] 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI). Yogyakarta, Indonesia: IEEE. pp.320–325. https://doi.org/10.1109/ISRITI48646.2019.9034620.

Lubis, A.R., Lase, Y.Y., Rahman, D.A. and Witarsyah, D., 2023. Improving Spell Checker Performance for Bahasa Indonesia Using Text Preprocessing Techniques with Deep Learning Models. *Ingénierie des systèmes d information*, 28(5), pp.1335–1342. https://doi.org/10.18280/isi.280522.

Mathew, L. and Bindu, V.R., 2020. A Review of Natural Language Processing Techniques for Sentiment Analysis using Pre-trained Models. In: *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*. [online] 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC). Erode, India: IEEE. pp.340–345. https://doi.org/10.1109/ICCMC48092.2020.ICCMC-00064.

Newman, N., Fletcher, R., Eddy, K., Robinson, C.T. and Nielsen, R.K., 2023. *Reuters Institute digital news report 2023*. [online] Reuters Institute for the Study of Journalism. https://doi.org/10.60625/RISJ-P6ES-HB13.

Pamungkas, E.W., Putri, D.G.P. and Fatmawati, A., 2023. Hate Speech Detection in Bahasa Indonesia: Challenges and Opportunities. *International Journal of Advanced Computer Science and Applications*, [online] 14(6). https://doi.org/10.14569/IJACSA.2023.0140625.

Putra, I.G.M. and Nurjanah, D., 2020. Hate Speech Detection In Indonesian Language Instagram. In: *2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. [online] 2020

International Conference on Advanced Computer Science and Information Systems (ICACSIS). Depok, Indonesia: IEEE. pp.413–420. https://doi.org/10.1109/ICACSIS51025.2020.9263084.

Wilie, B., Vincentio, K., Winata, G.I., Cahyawijaya, S., Li, X., Lim, Z.Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S. and Purwarianti, A., 2020. *IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding.* Available at: <http://arxiv.org/abs/2009.05387> [Accessed 1 February 2024].

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q. and Rush, A., 2020. Transformers: State-of-the-Art Natural Language Processing. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. [online] Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Online: Association for Computational Linguistics. pp.38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6.

Halaman ini sengaja dikosongkan