Jutisi: Jurnal Ilmiah Teknik Informatika dan Sistem Informasi
Jl. Ahmad Yani, K.M. 33,5 - Kampus STMIK Banjarbaru
Loktabat – Banjarbaru (Tlp. 0511 4782881), e-mail: puslit.stmikbjb@gmail.com
e-ISSN: 2685-0893
p-ISSN: 2089-3787

Penerapan *Machine Learning* Pada Sistem Informasi Klasifikasi Informasi Penggalian Potensi Pajak

Esha Indra Sukmana[1], Lala Nilawati[2*]
Sistem Informasi, Universitas Bina Sarana Informatika, Jakarta, Indonesia
*e-mail *Corresponding Author:* lala.lni@bsi.ac.id

Abstract

The Directorate General of Taxes (DGT) faces challenges in managing and utilizing external data on the internet, such as news from online news portals to explore the potential of Taxpayers' taxes. This study produces an information system that automatically classifies news titles, based on their relevance to tax potential using machine learning algorithms. The algorithm chosen for this study is the CRISP-DM technique, which includes understanding business processes, collecting and exploring news data, text processing, and developing classification models using the BERT and distilBERT models. The results of the model evaluation matrix test show that the distilBERT model obtained an accuracy of 0.8763, precision 0.8776, Recall 0.8763, and F1-Score 0.8768. While the results for the confusion matrix obtained the highest accuracy, recall, precision, and F1-Score values with a value of 0.78. It is concluded that this model is the best, with balanced performance and higher metrics especially for the "Potential" class. The information system built is web-based by implementing the waterfall method, and using Python software.

Keywords: *Information System; Machine learning Algorithms; News Classification; Tax Potential Analysis; Directorate General of Taxes.*

Abstrak

Direktorat Jenderal Pajak (DJP) menghadapi tantangan dalam pengelolaan, dan pemanfaatan data eksternal yang ada di internet, seperti berita dari portal berita daring untuk penggalian potensi pajak Wajib Pajak. Penelitian ini menghasilkan sistem informasi yang secara otomatis mengklasifikasikan judul berita, berdasarkan relevansinya dengan potensi pajak menggunakan algoritma machine learning. Algoritma yang dipilih untuk penelitian ini adalah teknik CRISP-DM, yang meliputi pemahaman proses bisnis, pengumpulan dan eksplorasi data berita, pemrosesan teks, dan pengembangan model klasifikasi menggunakan model BERT dan distilBERT. Hasil pengujian matriks evaluasi model menunjukan model distilBERT didapat hasil accuracy 0,8763, precision 0,8776, Recall 0,8763, dan F1-Score 0,8768. Sedangkan hasil untuk confusion matrix didapat nilai accuracy, recall, precision, dan F1-Score tertinggi dengan nilai 0.78. Disimpulkan bahwa model ini adalah yang terbaik, dengan performa yang seimbang dan metrik yang lebih tinggi terutama untuk kelas "Potensi". Sistem informasi yang dibangun berbasis web dengan menerapkan metode waterfall, dan menggunakan *software Python.*

Kata kunci: *Sistem Informasi; Algoritma Machine learning; Klasifikasi Berita; Analisis Potensi Pajak; Direktorat Jenderal Pajak.*

1. Pendahuluan

Pada surat edaran yang dikeluarkan oleh Direktur Jenderal Pajak dengan Nomor SE-11/PJ/2020, menyebutkan bahwa Direktorat Jenderal Pajak (DJP) selalu melakukan tindakan dalam peningkatan pengawasan pemenuhan kewajiban perpajakan, yaitu melalui pengumpulan data yang akurat dari Wajib Pajak, dengan kondisi sudah ataupun belum mempunyai NPWP. Teknik pengumpulan data yang dilakukan diambil dari sumber internal DJP, maupun data dari instansi lain, serta melalui Kegiatan Pengumpulan Data Lapangan (KPDL). KPDL mencakup pengamatan, pengambilan gambar, dan wawancara di lokasi kediaman atau usaha WP untuk memperbarui dan memperoleh data baru, meningkatkan kualitas data, serta mengidentifikasi potensi pajak untuk memperluas basis data dan membangun profil Wajib Pajak. Beberapa hasil penelitian terkait kebutuhan data perpajakan ada yang menyebutkan, bahwa terdapat

kebutuhan akan data tambahan selain yang dimiliki oleh DJP. Metode pencarian data penunjang yang paling umum dilakukan adalah melalui internet, diikuti oleh pengamatan langsung di lapangan, serta adanya kolaborasi dengan pihak ketiga [1]. Pada penelitian ini dengan memanfaatkan teks berita di internet, akan dilakukan pencarian dan pengumpulan informasi dalam rangka penggalian potensi pajak sebagai salah satu Kegiatan Pengumpulan Data Lapangan (KPDL).

Saat ini, pada Direktorat Jenderal Pajak untuk data tidak terstruktur berupa teks berita di internet, belum dilakukan pencarian dan pengumpulan informasi dalam rangka penggalian potensi pajak melalui Kegiatan Pengumpulan Data Lapangan (KPDL). Pengumpulan data dari internet perlu dilakukan karena informasi yang terdapat di sana, dapat memberikan wawasan tambahan mengenai aktivitas dan transaksi yang mungkin terlewatkan dalam sumber data internal, serta memperkaya analisis potensi pajak yang dilakukan oleh DJP. Algoritma *machine learning* dapat digunakan untuk mengembangkan sistem informasi, yang mampu mempelajari pola dari data berita yang relevan dengan potensi pajak, sehingga dapat mengklasifikasikan berita-berita yang ditemukan secara otomatis. Dengan demikian, penggunaan algoritma *machine learning* dalam rancang bangun sistem informasi ini diharapkan bisa meningkatkan efektivitas, dan akurasi penggalian potensi pajak melalui pencarian berita di internet.

Ketersediaan berupa data penunjang bertujuan untuk meningkatkan analisis perpajakan dan pemahaman dinamikanya. Pengumpulan data secara manual sangat tidak efektif dan membutuhkan waktu. Menurut Suryadi dkk, "Penggunaan *web scraping*, dapat membantu dalam pengelolaan data, sehingga menjadi meningkat efektifitasnya walaupun hanya dengan mengandalkan fitur *library* yang disediakan oleh *software Python*" [2]. Pada penelitian ini, penulis melakukan observasi pada Direktorat Jenderal Pajak dengan melakukan pengumpulan data berupa teks berita dari internet. Hal ini perlu dilakukan karena informasi yang terdapat di sana, dapat memberikan wawasan tambahan mengenai aktivitas dan transaksi yang mungkin terlewatkan dalam sumber data internal, serta memperkaya analisis potensi pajak yang dilakukan oleh DJP. Pada hasil penelitian yang terkait kebutuhan data perpajakan, ada yang menyebutkan bahwa terdapat kebutuhan akan data penunjang selain yang dimiliki oleh DJP.

Pada penelitian ini dilakukan dengan tujuan menghasilkan sebuah sistem informasi dengan *machine learning*, yang dapat mengklasifikasikan secara otomatis hasil pencarian berita di internet yang berhubungan dengan penggalian potensi pajak. Selain itu agar instansi Direktorat Jenderal Pajak diharapkan akan bisa memanfaatkan hasil penelitian ini, dan menjadi alternatif dalam menemukan solusi untuk permasalahan yang ada. Sedangkan manfaat adanya penelitian ini salah satunya untuk meningkatkan efisiensi proses penggalian potensi pajak. Disamping itu untuk mengidentifikasi pola-pola yang sulit dikenali secara manual, dan membantu dalam pengembangan sistem informasi untuk analisis dan penggalian potensi pajak.

2. Tinjauan Pustaka

Saat ini penerapan *machine learning* menjadi salah satu solusi yang semakin diminati, untuk menangani berbagai tugas yang masuk kategori kompleks. Salah satunya adalah dalam klasifikasi sentimen sudah dengan menggunakan *machine learning* serta *deep learning* [3]. Penerapan *Machine learning* serta *Deep Learning* meningkatkan efisiensi analisis, dengan mendeteksi pola yang kompleks secara otomatis termasuk dalam analisis sentimen. Kelebihan metode klasifikasi pada *Machine learning* salah satunya dapat mengelompokkan teks menjadi beberapa kategori [4]. Dalam konteks pemilihan algoritma untuk kasus klasifikasi sentimen, penting untuk mempertimbangkan kinerja dan kemampuan adaptasi algoritma terhadap variasi dan tingkat kerumitan data yang ada. Algoritma atau model BERT menunjukkan kinerja yang superior ketika diterapkan pada proses klasifikasi untuk *dataset* berupa artikel berita CNN [5]. *Machine learning* dapat diartikan merupakan hasil, dari sebuah penggunaan algoritma untuk pengolahan data, mempelajarinya, dan kemudian memprediksinya [6]. *Machine learning* bisa dikatakan sebagai pengimplementasian kecerdasan buatan dalam menyelesaikan berbagai masalah [7], dan merupakan metode modern yang salah satunya bisa diterapkan adalah teknik klasifikasi [8]. Penggunaan *Machine learning* untuk proses klasifikasi, bisa menjadi alternatif pilihan untuk klasifkasi data [9]. Tampilan sistem algoritma klasifikasi *Machine learning* pada umumnya dapat dievaluasi [10]. Jadi, bisa dikatakan metode klasifikasi dapat dioptimalkan menggunakan kinerja m*achine learning* [11]*,* dan penggunaan metode klasfikasi juga telah memberikan kontribusi penting [12]. Banyak sekali pemodelan teknik klasifikasi *machine learning*, salah satunya menggunakan teknik resample dan terbukti mampu meningkatkan

ketepatan pada proses klasifikasi [13]. Klasifikasi email spam dengan beberapa algoritma menggunakan *machine learning,* menghasilkan metode *Naïve Bayes* menjadi metode terbaik dalam penerapan klasifikasi [14]. Pada kasus pengklasifikasian menggunakan *Machine learning* pada proses kategori yang menghasilkan penilaian akhir dalam penilaian tunjangan dari sisi kinerja pegawai, didapat bahwa algoritma C4.5 menjadi yang terbaik [15]. Pada penelitian memprediksi kemungkinan diabetes dengan tiga buah algoritma metode klasifikasi, didapatkan kesimpulan metode SVM dapat menghasilkan hasil yang paling tinggi [16]. Penerapan arsitektur *transfer learning Xception* dengan menerapkan metode klasifikasi, terbukti menghasilkan penampilan teknik *machine learning* yang dihasilkan [17]. Teknik kualifikasi yang diterapkan untuk menganalisa tingkat kualitas pada air minum dengan teknik algoritma *machine learning*, mendapatkan hasil yang sangat mudah dipahami, karena menghasilkan hasil yang sangat sederhana [18] dan juga dapat memberikan manfaat penting dalam bidang lingkungan dan kesehatan [19]. Teknik klasifikasi menggunakan *Machine learning* juga dianggap penting dalam mengatasi permasalahan menilai potensi mahasiswa, [20] dan bisa juga sebagai upaya cikal bakal untuk sistem alternatif yang dapat diambil calon mahasiswa baru ketika menentukan program studi yang diinginkan [21].

Pada penelitian ini akan dirancang sistem informasi dengan menggunakan algoritma *machine learning*, untuk klasifikasi informasi penggalian potensi pajak. Penelitian ini dilakukan untuk menghasilkan rancangan sistem yang bersifat otomatis dalam mengklasifikasikan judul berita, berdasarkan relevansinya dengan potensi pajak menggunakan algoritma *machine learning*. Penggunaan metode dalam penelitian ini adalah menggunakan teknik *Cross-Industry Standard Process for Data Mining* (CRISP-DM), yang meliputi pemahaman proses bisnis, pengumpulan dan eksplorasi data berita, pemrosesan teks, dan pengembangan model klasifikasi menggunakan *neural network* BERT. Evaluasi model yang akan diterapkan adalah dengan penggunaan metrik seperti *accuracy*, *precision*, *recall*, dan *F1-score*. Sistem yang dikembangkan akan dirancang akan berbasis web dengan penerapan metode *waterfall,* dan menerapkan *software Python* dan implementasi basis data dari perancangan ERD dan UML. Penerapan sistem informasi memiliki potensi untuk memberikan dukungan dalam menangani berbagai permasalahan, terkait kebutuhan informasi yang dihadapi oleh para penggunanya [22]. Algoritma *machine learning* dapat digunakan untuk mengembangkan sistem informasi, yang mampu mempelajari pola dari data berita yang relevan dengan potensi pajak, sehingga dapat mengklasifikasikan berita-berita yang ditemukan secara otomatis. Dengan demikian, penggunaan algoritma *machine learning* dalam rancang bangun sistem informasi ini, diharapkan bisa meningkatkan efektivitas dan akurasi penggalian potensi pajak melalui pencarian berita di internet.

3. Metodologi

Pada pemodelan algoritma *machine learning*, analisis data pada penelitian yang dilakukan ini, menerapkan teknik *Cross-Industry Standard Process for Data Mining* (CRISP-DM) yang akan melalui tahapan berikut:

1) Tahap Dalam Pemahaman Proses Bisnis (*Business Understanding*)
   Salah satu sumber data yang dapat digunakan sebagai data awal adalah berita pada portal berita daring. Data ini adalah produk jurnalisme yang telah melalui proses pengumpulan dan verifikasi fakta. Google sebagai mesin pencari melakukan pengindeksan terhadap berita-berita yang ada di internet dan ditampilkan pada Google News. Pencarian menggunakan kata kunci pada Google News dapat menghasilkan daftar berita yang relevan terhadap kata kunci tersebut. Namun tidak semua hasil pencarian relevan dengan pengawasan dan penggalian potensi pajak. Data hasil pencarian Google News perlu dilakukan penyaringan agar didapatkan data yang relevan dengan potensi pajak saja.
2) Tahap Pemahaman Data (*Data Understanding*)
   Google News menampilkan daftar berita hasil pencarian menggunakan kata kunci sebanyak 100 baris setiap pencarian. Adanya penerapan metode *web scraping* dengan dibantu *software Python* pada halaman yang merupakan hasil pencarian menggunakan kombinasi kata kunci “usaha”, “pengusaha”, ”bisnis”, ”omset” serta kata kunci wilayah “bekasi”, ”cikarang”, ”subang”, “indramayu”, ”cirebon”, dapat diambil data berupa judul berita dan URL dari berita hasil pencarian tersebut. Data tersebut kemudian disimpan dalam bentuk *Comma Separated Value* (CSV) dengan total baris data sebanyak 13.879 baris. Data CSV kemudian dilakukan pelabelan manual menggunakan klasifikasi kata pada judul berita.

3) Tahap Penyiapan Data (*Data Preparation*)
   Proses pada tahap ini adalah melakukan penyiapan data untuk proses pemodelan. Langkah pertama yaitu pemrosesan teks seperti penyamaan bentuk huruf menjadi huruf kecil, penghapusan tanda baca, normalisasi kata, serta penghilangan kata-kata umum dan tidak relevan (*stopwords*). Lalu dilakukan penanganan data yang duplikasi sehingga dihasilkan 13.856 baris data dan telah siap untuk digunakan sebagai data pelatihan model *machine learning*. Selanjutnya dilakukan pembagian data yang diperuntukan sebagai data pelatihan, dan data pengujian dengan komposisi data 80:20, dengan tetap mempertahankan distribusi kelas pada data latih dan data uji (*stratified split*).
4) Proses Pemodelan (*Modeling*)
   Pada tahap proses pemodelan akan dilakukan pengembangan model klasifikasi lewat arsitektur model BERT dan distilBERT dengan *software Python* dan kerangka kerja (*framework*) TensorFlow. Langkah-langkahnya meliputi pemuatan model *BERT Base Multilingual Uncased dari Google* maupun distilBERT *Base Indonesian* dari Cahya Wirawan sebagai model dasar (*base model*) yang kemudian dilakukan *transfer learning* menggunakan data latih yang sudah disiapkan sebelumnya. Proses pelatihan menggunakan proses dengan komposisi perbandingan data latih dan data validasi yaitu 80%:20%. Konfigurasi pelatihan sebanyak 200 siklus latih (*epochs)* dengan parameter penghentian dini (*early stopping*) untuk efisiensi waktu pelatihan dan mencegah model *overfitting*. Sebagai pengawasan atas hasil pelatihan, dilakukan visualisasi atas hasil matriks *accuracy* dan *loss* tiap siklus latih
5) Evaluasi (*Evaluation*)
   Model yang telah dilatih kemudian dilakukan evaluasi pada data uji dengan beberapa matriks evaluasi seperti *accuracy*, *precision*, *recall*, *F1-Score* dan dilakukan pelatihan kembali dengan penyesuaian parameter sampai ditemukan model dengan matriks evaluasi sesuai dengan tujuan awal analisis data.
6) Penyebaran (*Deployment*)
   Model dengan matriks evaluasi terbaik kemudian disimpan melalui fungsi pada kerangka kerja *TensorFlow* agar dapat digunakan kembali untuk memprediksi kelas data pada data baru hasil pencarian pengguna di sistem yang dirancang.

4. Hasil dan Pembahasan

4.1 Implementasi Algoritma

Hasil pencarian teks berita di internet akan diterapkan teknik *web scraping* dengan *software Python,* serta klasifikasi data menggunakan *machine learning* untuk mengefisienkan proses penyaringan data. Untuk pelatihan model klasifikasi, dibutuhkan pelabelan terhadap data latih. Kelas atau kategorinya adalah data yang relevan dengan potensi pajak (data potensi), data yang bersifat informasi misalnya informasi makroekonomi, serta data yang tidak relevan dengan potensi pajak. Berikut adalah *sample* data latih dengan label Potensi (Label 1):

Tabel 1. *Sample* Data Latih Dengan Label Potensi (Label 1)

No	URL	Label
1	https://news.google.com/articles/CAIiEAVZNlJxUFlUG....	1
2	https://news.google.com/articles/CAIiEAwRSj3lvW0hS....	1
3	https://news.google.com/articles/CAIiEAwUhbg0oc3Yf....	1
4	https://news.google.com/articles/CAIiEAwum8HEe3fJS....	1
5	https://news.google.com/articles/CAIiEAwVdZ4mNBPAO....	1
6	https://news.google.com/articles/CAIiEAwVS-eFKXvH7....	1
7	https://news.google.com/articles/CAIiEAXJ6qUzDHuV_....	1
8	https://news.google.com/articles/CAIiEAxORxZWD3RWY....	1
9	https://news.google.com/articles/CAIiEAxu5cj3lUicA....	1
10	https://news.google.com/articles/CAIiEAz4WgMnc6YtI....	1

Sumber: (Hasil Penelitian, 2024)

Sedangkan untuk *sample* data latih dengan label Informasi (Label 2) yaitu sebagai berikut:

Tabel 2. *Sample* Data Latih Dengan Label Informasi (Label 2)

No	URL	Label
1	https://news.google.com/articles/CAIiEC5z4P4_kkfwo....	2
2	https://news.google.com/articles/CAIiEC7_mt1aObvhu....	2
3	https://news.google.com/articles/CAIiECatJvw0LyopS....	2
4	https://news.google.com/articles/CAIiECB7FCqqgbhA1....	2
5	https://news.google.com/articles/CAIiECdNLsVMFQgCN....	2
6	https://news.google.com/articles/CAIiECHScsdDri2cr....	2
7	https://news.google.com/articles/CAIiECHz3dxxT0Xt9....	2
8	https://news.google.com/articles/CAIiECi_q1DI9tRGT....	2
9	https://news.google.com/articles/CAIiECIZNpyw3A0hv....	2
10	https://news.google.com/articles/CAIiECJXR6ezD1fBi....	2

Sumber: (Hasil Penelitian, 2024)

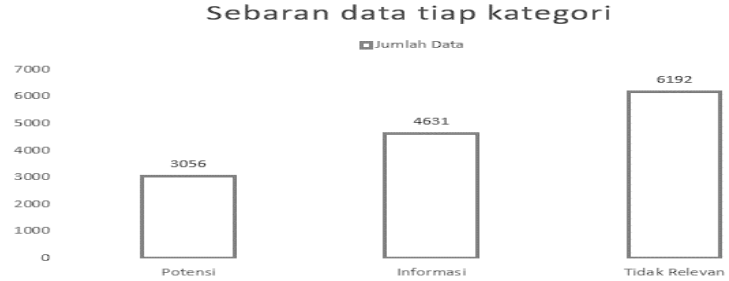
Sample data latih dengan label Tidak Relevan (Label 0) yaitu sebagai berikut:

Tabel 3. *Sample* Data Latih Dengan Label Tidak Relevan (Label 0)

No	URL	Label
1	https://news.google.com/articles/CAIiEAGXU5VuSUPiT....	0
2	https://news.google.com/articles/CAIiEAgXZBTfDz8Er....	0
3	https://news.google.com/articles/CAIiEAh2MUpCPt0GW....	0
4	https://news.google.com/articles/CAIiEAH2TbzS_E976....	0
5	https://news.google.com/articles/CAIiEAh8RZ8rR-6qu....	0
6	https://news.google.com/articles/CAIiEAhfK8Klfkj4u....	0
7	https://news.google.com/articles/CAIiEAhGV_TWGX6li....	0
8	https://news.google.com/articles/CAIiEAHHvrlOCKaFz....	0
9	https://news.google.com/articles/CAIiEAHmBWhGOzi6T....	0
10	https://news.google.com/articles/CAIiEAHmi3n9EyfDu....	0

Sumber: (Hasil Penelitian, 2024)

Dari ketiga kelas tersebut, kelas data yang relevan dengan potensi pajak adalah kelas yang paling penting dalam model klasifikasi ini untuk mengkategorikan data yang relevan dengan kegiatan penggalian potensi pajak. Kelas informasi mungkin masih memiliki nilai bagi analisis potensi pajak, meskipun tidak langsung menunjukkan potensi pajak. Hasil dari penggunaan teknik *web scraping* dengan bantuan *software* Python yang sudah disimpan dalam bentuk *Comma Separated Value* (CSV), dengan total baris data sebanyak 13.879 baris, kemudian dilakukan pelabelan manual menggunakan klasifikasi kata pada judul berita. Distribusi kelas setelah proses pelabelan akan tampil seperti berikut:



Sumber: (Hasil Penelitian, 2024)

Gambar 1. Distribusi Kelas Data Hasil *Web Scraping*

Untuk lebih memahami distribusi kata dalam data dilakukan visualisasi dengan menggunakan awan kata (*word clouds*) dengan hasil tampilan:

Sumber: (Hasil Penelitian, 2024)

Gambar 2. Visualisasi Awan Kata Pada Data Awal

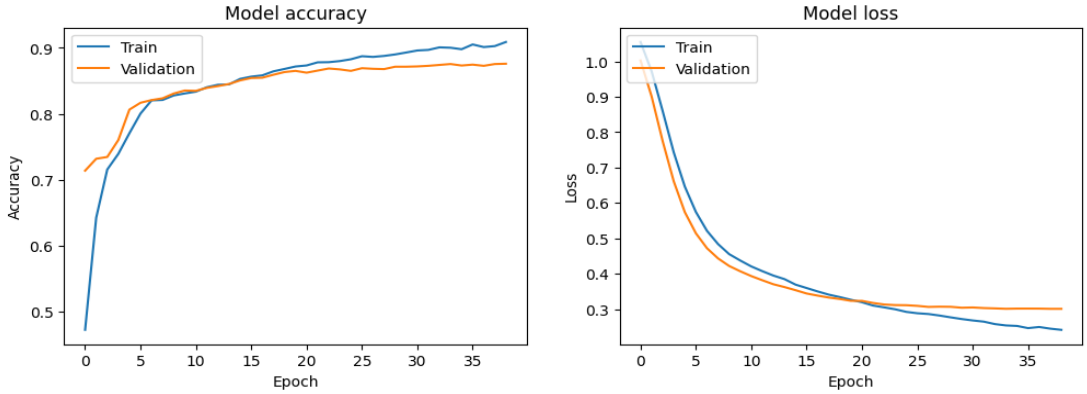
Selanjutnya pembagian data untuk pelatihan dan pengujian dengan proprosi 80:20, dengan tetap mempertahankan distribusi kelas yaitu kelas Tidak relevan, Potensi dan Informasi pada data latih dan data uji (*stratified split*). Berikut rekapitulasi data untuk tiap kelas dan prosesnya:

Tabel 4. Distribusi Data Per Kategori

Kelas	Data Awal	Data Setelah Diproses		
		Total	Data Latih (80%)	Data Uji (20%)
Tidak relevan (label 0)	6.192	6.188	4.950	1.238
Potensi (label 1)	3.056	3.048	2.438	610
Informasi (label 2)	4.631	4.620	3.696	924
Total	13.879	13.856	11.084	2.772

Sumber: (Hasil Penelitian, 2024)

Proses pelatihan menggunakan komposisi perbandingan data latih dan data validasi yaitu 80%:20%, dengan konfigurasi pelatihan sebanyak 200 siklus latih (*epochs)* dengan parameter penghentian dini (*early stopping*), untuk efisiensi waktu pelatihan dan mencegah model *overfitting*. Sebagai pengawasan atas hasil pelatihan, dilakukan visualisasi atas hasil matriks *accuracy* dan *loss* tiap siklus latih dengan hasil berikut:



Sumber: (Hasil Penelitian, 2024)

Gambar 3. Visualisasi Hasil Pelatihan Model Klasifikasi

4.2 Evaluasi Kinerja *Machine Learning*

Hasil pelatihan model kemudian akan dilakukan evaluasi pada data uji dengan beberapa matriks evaluasi seperti *accuracy*, *precision*, *recall*, *F1-Score*, dan dilakukan pelatihan kembali dengan penyesuaian parameter sampai ditemukan model dengan matriks evaluasi

sesuai dengan tujuan awal analisis data. Setelah 3 iterasi pelatihan didapatkan model dengan matriks evaluasi seperti berikut:

Tabel 2. Matriks Evaluasi Model

Model Yang Dilatih	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Model 1 BERT	0,8620	0,8618	0,8620	0,8619
Model 2 distilBERT	0,8750	0,8746	0,8750	0,8747
Model 3 distilBERT	**0,8763**	**0,8776**	**0,8763**	**0,8768**

Sumber: (Hasil Penelitian, 2024)

Adapun *confusion matrix* untuk ketiga model adalah berikut:

Tabel 3 *Confusion Matrix Model 1*

		Prediksi			Precision	Recall	F1-Score
		Tidak relevan (0)	Potensi (1)	Informasi (2)			
Aktual	Tidak relevan (0)	1189	17	33	0,96	0,96	0,96
	Potensi (1)	20	456	135	0,75	0,75	0,75
	Informasi (2)	36	142	748	0,83	0,80	0,81

Sumber: (Hasil Penelitian, 2024)

Tabel 4. *Confusion Matrix Model 2*

		Prediksi			Precision	Recall	F1-Score
		Tidak relevan (0)	Potensi (1)	Informasi (2)			
Aktual	Tidak relevan (0)	1181	29	29	0,96	0,96	0,96
	Potensi (1)	25	458	128	0,75	0,75	0,75
	Informasi (2)	32	104	790	0,83	0,85	0,84

Sumber: (Hasil Penelitian, 2024)

Tabel 5. *Confusion Matrix Model 3*

		Prediksi			Precision	Recall	F1-Score
		Tidak relevan (0)	Potensi (1)	Informasi (2)			
Aktual	Tidak relevan (0)	1171	26	41	0,96	0,95	0,96
	Potensi (1)	18	475	117	0,78	0,78	0,78
	Informasi (2)	25	116	783	0,83	0,85	0,84

Sumber: (Hasil Penelitian, 2024)

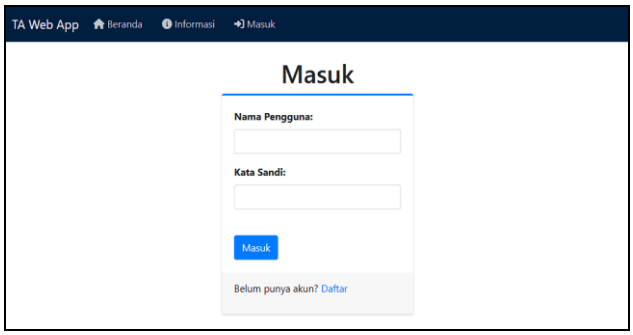
Dari matriks evaluasi di atas, model ketiga mempunyai hasil *accuracy*, *recall*, *precision*, dan *F1-Score* tertinggi di antara ketiga model. Model 3 juga memiliki *F1-score* tertinggi untuk kelas "Potensi" (1). Berdasarkan laporan klasifikasi ini, Model 3 adalah yang terbaik, dengan performa yang seimbang dan metrik yang lebih tinggi terutama untuk kelas "Potensi".

4.3 Antarmuka Pengguna

Hasil implementasi rancangan antar muka pada sistem informasi algoritma *machine learning* untuk klasifikasi informasi penggalian potensi pajak adalah seperti berikut:

1) Halaman *Login* Pengguna

Halaman yang berisi *form login* pengguna dan pesan kesalahan terkait proses *login* pengguna.

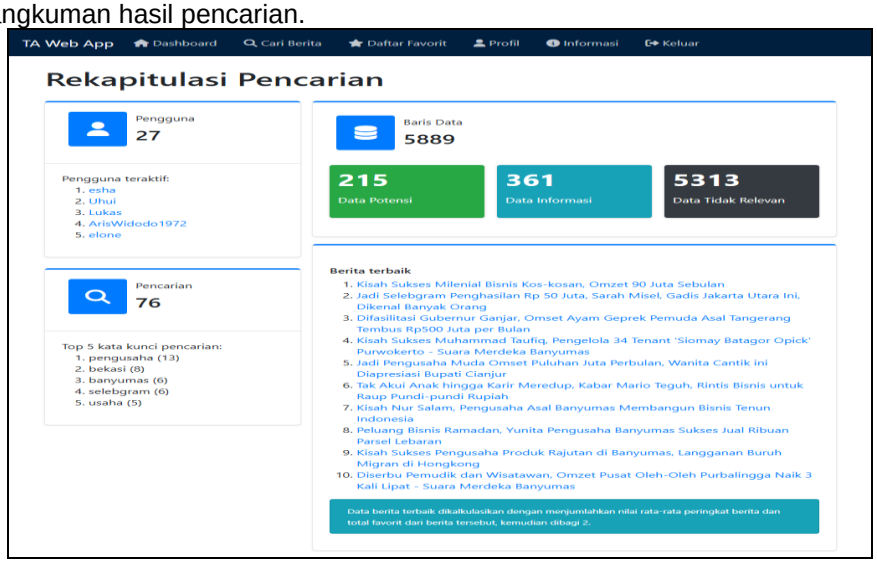


Sumber: (Hasil Penelitian, 2024)

Gambar 5. Implementasi Rancangan Halaman *Login* Pengguna

2) Halaman *Dashboard* Pengguna

Halaman yang menampilkan rekapitulasi pencarian yang memberikan informasi tentang tren dan rangkuman hasil pencarian.

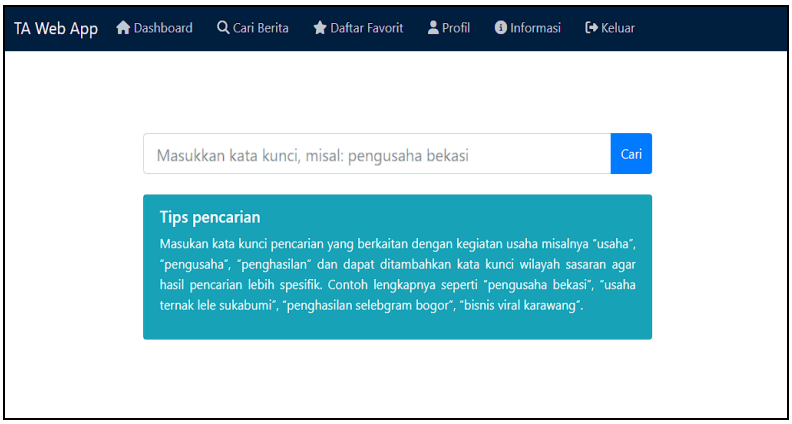


Sumber: (Hasil Penelitian, 2024)

Gambar 6. Implementasi Rancangan Halaman *Dashboard* Pengguna

3) Halaman Pencarian Berita

Halaman yang berisi *form* pencarian berita dan pesan kesalahan terkait proses pencarian berita.

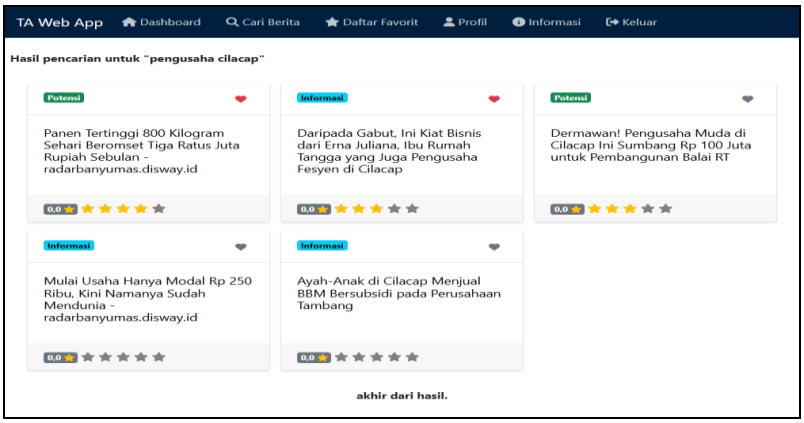


Sumber: (Hasil Penelitian, 2024)

Gambar 7. Implementasi Rancangan Halaman Pencarian Berita

4) Halaman Hasil Pencarian

Halaman yang menampilkan hasil pencarian di mana pengguna juga dapat menambahkan berita hasil pencarian ke daftar favorit serta memberikan peringkat pada berita.

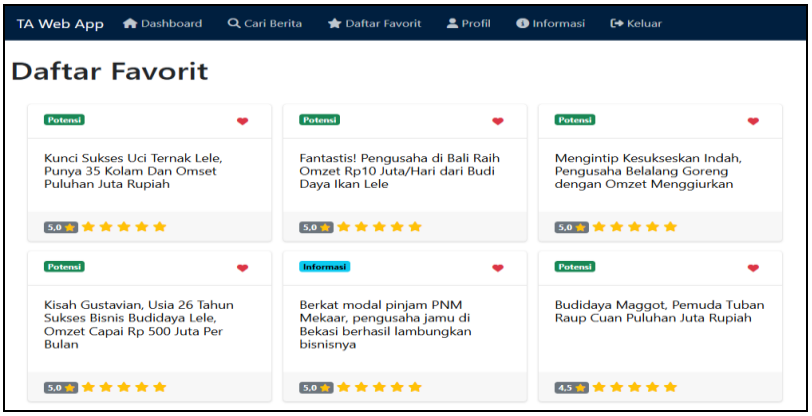


Sumber: (Hasil Penelitian, 2024)

Gambar 8. Implementasi Rancangan Halaman Hasil Pencarian

5) Halaman Daftar Favorit

Halaman yang menampilkan daftar berita favorit pengguna. Pengguna dapat melihat dan menghapus berita dari daftar favorit.



Sumber: (Hasil Penelitian, 2024)

Gambar 9. Implementasi Rancangan Halaman Daftar Favorit

4.3 Pembahasan

Pada penelitian terdahulu sudah banyak yang menunjukan hasil positif, dalam penggunaan algoritma CRISP-DM dan metode BERT dan Distil-BERT. Hasil penelitian pada proses klasifikasi untuk sentimen serta analisis, pada tingkah laku dalam pembelian yang diperuntukan pada layanan akomodasi hotel dengan penerapaan metode CRISP-DM, menunjukan hasil signifikan dalam upaya pengoptimalan Customer Relationship Management (CRM) [23]. Pada prose pengklasifikasian hasil data curah hujan menggunakan metode decision tree algoritma CART (*Classification And Regression Tree*) dengan CRISP-DM menunjukan hasil kerja proses yang cukup baik, terlihat dari akurasi yang didapatkan yaitu sebesar 89,4%, serta menerapkan proses evaluasi serta validasi menerapkan parameter uji *Confusion Matrix* [24]. Pada proses observasi untuk data isu pada tenaga kerja di lapangan dengan menerapkan CRISP-DM serta *K-Nearest Neighbor* (KNN), dapat disimpulkan bahwa tingkat keakuratan menghasilkan nilai yang tinggi sebesar 93.88% [25]. Penggunaan CRISP-DM juga diterapkan pada pengklasifikasian data hasil komentar yang diberikan pengunjung, pada destinasi di danau toba yang dikolaborasikan dengan algoritma *Naïve Bayes Classifier* (NBC) dan *Decision Tree* (DT), menghasilkan sebuah atraksi yang menjadi aspek penting pemberian persepsi wisatawan, hal ini menunjukan perlu adanya pengelolaan optimal [26].

Pada hasil penelitian tentang perbandingan tingkat akurasi BERT dengan DistilBERT pada *Dataset* yang ada *Twitter*, DistilBERT menghasilkan nilai akurasi lebih tinggi bila dibandingkan dengan BERT. Dimana DistilBERT merupakan salah satu teknik dari BERT yang memberikan kecepatan dan memaksimalkan klasifikasi [27]. Pada penelitian lain masih tentang perbandingan tingkat akurasi metode DistilBERT dengan BERT, yaitu pada penyajian dataset tentang penganalisaan sentimen pada lembaga kursus, menghasilkan tingkat akurasi yang mirip. Meskipun seperti itu, waktu eksekusi Arsitektur DistilBERT lebih cepat dibandingkan dengan BERT, dan menghasilkan kinerja yang hampir mirip [28]. Penerapan teknik BERT, RoBERTa, distilBERT, dan distilBERT-frezee dalam proses pengklasifikasian emosi pada data Bahasa Indonesia, hasilnya didapatkan model distilBERT-*frezee* menunjukan peningkatan untuk hasil *f1-score* yang secara signifikan [29]. Pada penelitian ini, penggunaan algoritma CRISP-DM dan model DistilBERT juga menunjukan hasil dengan performa yang seimbang, dan metrik yang lebih tinggi terutama untuk kelas "Potensi". Serta mempunyai nilai tertinggi dan terbaik untuk nilai *accuracy, precision, recall,* dan *F1-Score.* Hasil matriks evaluasi model DistilBERT menunjukan tingkat *accuracy* 0,8763, *precision* 0,8776, *Recall* 0,8763, dan *F1-Score* 0,8768. Sedangkan hasil untuk *confusion matrix* didapat hasil *accuracy, recall*, *precision,* dan *F1-Score* pada tingkat tertinggi dengan nilai 0.78.

5. Simpulan

Tujuan dari penelitian yang dilakukan ini adalah menekankan pentingnya pemanfaatan data dari internet untuk penggalian potensi pajak. Melalui penerapan algoritma *machine learning* yaitu menggunakan algoritma CRISP-DM dan model DistilBERT, menunjukan hasil dengan performa yang seimbang, dan metrik yang lebih tinggi terutama untuk kelas "Potensi". Hasil yang terlihat adalah menunjukan nilai tertinggi dan terbaik untuk tingkat *accuracy, precision, recall*, dan *F1-Score*. Hasil matriks evaluasi model DistilBERT menunjukan tingkat *accuracy* 0,8763, *precision* 0,8776, *Recall* 0,8763, dan *F1-Score* 0,8768. Sedangkan hasil untuk *confusion matrix* didapat nilai *accuracy, precision, recall*, dan *F1-Score* tertinggi dengan nilai 0.78. Penerapan algoritma *machine learning* dalam klasifikasi berita di internet diharapkan dapat menghasilkan informasi yang lebih terstruktur dan relevan, sehingga mendukung kebijakan pengawasan dan pengambilan keputusan yang lebih baik dalam hal pengawasan Wajib Pajak dan penggalian potensi pajak. Sistem yang dirancang ini memungkinkan pencarian dan pengumpulan data yang lebih cepat dan akurat, melalui penggunaan teknik *web scraping* dan algoritma *machine learning*. Dengan demikian, DJP dapat melakukan analisis potensi pajak secara lebih sistematis dan efektif. Adanya sistem informasi ini bertujuan untuk dapat meningkatkan efisiensi dalam pengawasan Wajib Pajak, dan penggalian potensi pajak dengan memanfaatkan data berita dari internet yang relevan. Hal ini diharapkan dapat memperluas cakupan pengawasan terhadap transaksi dan aktivitas, yang mungkin terlewatkan di sumber data internal.

Daftar Referensi

[1] M. Djufri, “Penerapan Teknik Web Scraping Untuk Penggalian Potensi Pajak (Studi Kasus pada Online Market Place Tokopedia, Shopee dan Bukalapak),” vol. 13, no. 2, pp. 65–75, 2020.

[2] A. Suryadi, W. A. Syb’an, N. Alfa’inna, and E. H. Hermaliani, “Implementasi Web Scraping dan Sentiment Analysis Terhadap Berita Menggunakan Machine Learning,” *Swabumi*, vol. 11, no. 1, pp. 28–34, 2023, doi: 10.31294/swabumi.v11i1.15145.

[3] S. M. P. Tyas, R. Sarno, and B. S. Rintyarna, “Analisis Perbandingan Metode Klasifikasi Sentimen Berita Saham: Pendekatan Machine Learning, Deep Learning, Transfer Learning, dan Graf,” *J. Penelit. IPTEKS*, vol. 9, no. 1, pp. 58–64, 2024, doi: 10.32528/penelitianipteks.v9i1.1479.

[4] Alfando and R. Hayami, “Klasifikasi Teks Berita Berbahasa Indonesia Menggunakan Machine Learning Dan Deep Learning: Studi Literatur,” *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 1, pp. 681–686, 2023.

[5] N. Husin, “Komparasi Algoritma Random Forest, Naïve Bayes, dan Bert Untuk Multi-Class Classification Pada Artikel Cable News Network (CNN),” *J. Esensi Infokom J. Esensi Sist. Inf. dan Sist. Komput.*, vol. 7, no. 1, pp. 75–84, 2023, doi: 10.55886/infokom.v7i1.608.

[6] E. Haerani, F. Syafria, F. Lestari, Novriyanto, and I. Marzuki, “Classification Academic Data Using Machine Learning for Decision Making Process,” *J. Appl. Eng. Technol. Sci.*,

vol. 4, no. 2, pp. 955–968, 2023, doi: 10.37385/jaets.v4i2.1983.
[7] S. Sunardi, A. Fadlil, and D. Prayogi, "Face Recognition Using Machine Learning Algorithm Based on Raspberry Pi 4b," *Int. J. Artif. Intell. Res.*, vol. ISSN, no. 1, pp. 2579–7298, 2022, doi: 10.29099/ijair.v7i1.321.
[8] Generosa Lukhayu Pritalia, "Analisis Komparatif Algoritme Machine Learning dan Penanganan Imbalanced Data pada Klasifikasi Kualitas Air Layak Minum," *KONSTELASI Konvergensi Teknol. dan Sist. Inf.*, vol. 2, no. 1, pp. 43–55, 2022, doi: 10.24002/konstelasi.v2i1.5630.
[9] F. Baharuddin and A. Tjahyanto, "Peningkatan Performa Klasifikasi Machine Learning Melalui Perbandingan Metode Machine Learning dan Peningkatan Dataset," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 11, no. 1, pp. 25–31, 2022, doi: 10.32736/sisfokom.v11i1.1337.
[10] S. Salma, F. Dewanta, and M. Abdillah, "Klasifikasi Beban Listrik Dengan Machine Learning Menggunakan Metode K-Nearest Neighbor," *Resist. (Elektronika Kendali Telekomun. Tenaga List. Komputer)*, vol. 5, no. 2, p. 163, 2022, doi: 10.24853/resistor.5.2.163-172.
[11] M. Yanto, Febri Hadi, and S. Arlis, "Optimization of Machine Learning Classification Analysis of Malnutrition Cases in Children," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 7, no. 6, pp. 1378–1386, 2023, doi: 10.29207/resti.v7i6.5278.
[12] Ramadhani, Ramadhanu, and Taufik Hidayat, "Metode Machine Learning untuk Klasifikasi Data Gizi Balita dengan Algoritma Naïve Bayes, KNN dan Decision Tree," *J. SIMETRIS*, vol. 15, no. 1, pp. 57–68, 2024.
[13] P. R. Sihombing and I. F. Yuliati, "Penerapan Metode Machine Learning dalam Klasifikasi Risiko Kejadian Berat Badan Lahir Rendah di Indonesia," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 20, no. 2, pp. 417–426, 2021, doi: 10.30812/matrik.v20i2.1174.
[14] H. Iswanto, E. Seniwati, Y. Astuti, and D. Maulina, "Comparison of Algorithms on Machine Learning For Spam Email Classification," *IJISTECH (International J. Inf. Syst. Technol.*, vol. 5, no. 4, p. 446, 2021, doi: 10.30645/ijistech.v5i4.164.
[15] I. D. S. Tarigan, Roni Habibi, and Rd. Nuraini Siti Fatonah, "Evaluasi Algoritma Klasifikasi Machine Learning Kategori Nilai Akhir Tunjangan Kinerja Pegawai," *J. Sist. Cerdas*, vol. 6, no. 3, pp. 251–261, 2023, doi: 10.37396/jsc.v6i3.246.
[16] I. M. Karo Karo and H. Hendriyana, "Klasifikasi Penderita Diabetes menggunakan Algoritma Machine Learning dan Z-Score," *J. Teknol. Terpadu*, vol. 8, no. 2, pp. 94–99, 2022, doi: 10.54914/jtt.v8i2.564.
[17] E. N. Cahyo, E. Susanti, and R. Y. Ariyana, "Model Machine Learning Untuk Klasifikasi Kesegaran Daging Menggunakan Arsitektur Transfer Learning Xception," *J. Sist. dan Teknol. Inf.*, vol. 11, no. 2, p. 371, 2023, doi: 10.26418/justin.v11i2.57517.
[18] L. Savitri and R. Nursalim, "Klasifikasi Kualitas Air Minum menggunakan Penerapan Algoritma Machine Learning dengan Pendekatan Supervised Learning," *Diophantine J. Math. Its Appl.*, vol. 2, no. 01, pp. 30–36, 2023, doi: 10.33369/diophantine.v2i01.28260.
[19] Stacyana Jesika, Suci Ramadhani, and Yohanna Permata Putri, "Implementasi Model Machine Learning dalam Mengklasifikasi Kualitas Air," *J. Ilm. Dan Karya Mhs.*, vol. 1, no. 6, pp. 382–396, 2023, doi: 10.54066/jikma.v1i6.1162.
[20] R. Haque, A. Quek, C. Y. Ting, H. N. Goh, and M. R. Hasan, "Classification Techniques Using Machine Learning for Graduate Student Employability Predictions," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 14, no. 1, pp. 45–56, 2024, doi: 10.18517/ijaseit.14.1.19549.
[21] A. R. Pratama, R. R. Aryanto, and A. T. M. Pratama, "Model Klasifikasi Calon Mahasiswa Baru Untuk Sistem Rekomendasi Program Studi Sarjana Berbasis Machine Learning," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 9, no. 4, pp. 725–734, 2022, doi: 10.25126/jtiik.2022934311.
[22] L. Nilawati and M. Martin, "Penerapan Metode RAD Pada Perancangan Sistem Informasi Permohonan Data Aduan Smartmaps Berbasis Web," *JURIKOM (Jurnal Ris. Komputer)*, vol. 10, no. 2, p. 648, 2023, doi: 10.30865/jurikom.v10i2.6041.
[23] Y. A. Singgalen, "Penerapan CRISP-DM dalam Klasifikasi Sentimen dan Analisis Perilaku Pembelian Layanan Akomodasi Hotel Berbasis Algoritma Decision Tree (DT)," *J. Sist. Komput. dan Inform.*, vol. 5, no. 2, p. 237, 2023, doi: 10.30865/json.v5i2.7081.
[24] M. A. Hasanah, S. Soim, and A. S. Handayani, "Implementasi CRISP-DM Model

Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir," *J. Appl. Informatics Comput.*, vol. 5, no. 2, pp. 103–108, 2021, doi: 10.30871/jaic.v5i2.3200.

[25] N. Ajijah and A. Kurniawan, "Klasifikasi Teks Mining Terhadap Analisa Isu Kegiatan Tenaga Lapangan Menggunakan Algoritma K-Nearest Neighbor (KNN)," *J-SAKTI (Jurnal Sains Komput. Inform.*, vol. 7, no. 1, pp. 254–262, 2023.

[26] Y. A. Singgalen, "Penerapan Metode CRISP-DM dalam Klasifikasi Data Ulasan Pengunjung Destinasi Danau Toba Menggunakan Algoritma Naïve Bayes Classifier (NBC) dan Decision Tree (DT)," *J. Media Inform. Budidarma*, vol. 7, no. 3, p. 1551, 2023, doi: 10.30865/mib.v7i3.6461.

[27] F. Fajri, B. Tutuko, and S. Sukemi, "Membandingkan Nilai Akurasi BERT dan DistilBERT pada Dataset Twitter," *JUSIFO (Jurnal Sist. Informasi)*, vol. 8, no. 2, pp. 71–80, 2022, doi: 10.19109/jusifo.v8i2.13885.

[28] A. C. Saputra, A. S. Saragih, and D. Ronaldo, "Perbandingan Nilai Akurasi DistilBERT Dan BERT Pada Dataset Analisis Sentimen Lembaga Kursus," *J. Teknol. Inf.*, vol. 18, no. 2, pp. 160–171, 2024.

[29] F. Basbeth and D. H. Fudholi, "Klasifikasi Emosi Pada Data Text Bahasa Indonesia Menggunakan Algoritma BERT, RoBERTa, dan Distil-BERT," *J. Media Inform. Budidarma*, vol. 8, no. 2, p. 1160, 2024, doi: 10.30865/mib.v8i2.7472.