



PENERAPAN METODE TRANSFER LEARNING PADA INDOBERT UNTUK ANALISIS SENTIMEN TEKS BAHASA JAWA NGOKO LUGU

Aziz Prabowo¹, Fadil Indra Sanjaya²

¹azizprabowo128@gmail.com, ²fadil.indra@staff.uty.ac.id

¹Program Studi Sains Data, Universitas Teknologi Yogyakarta

²Program Studi Informatika, Universitas Teknologi Yogyakarta

Abstrak

Peningkatan signifikan dalam penggunaan media sosial dan platform *online* di Indonesia telah menyebabkan lonjakan produksi dan penyebaran teks dalam berbagai bahasa daerah, termasuk Bahasa Jawa. Hal ini menciptakan kebutuhan mendesak akan sistem analisis sentimen yang dapat memahami nuansa Bahasa Jawa dalam konteks opini dan ekspresi pengguna. Penelitian ini bertujuan untuk menerapkan metode *transfer learning* pada model IndoBERT, dengan fokus pada evaluasi kinerjanya dalam tugas analisis sentimen teks Bahasa Jawa Ngoko Lugu. Hasil penelitian menunjukkan bahwa model IndoBERT mencapai akurasi, presisi, *recall*, dan F1-Score yang tinggi dalam mengidentifikasi sentimen positif dan negatif dalam teks Bahasa Jawa Ngoko Lugu, dengan tingkat akurasi 99% pada data pelatihan, 90% pada data validasi, dan 90% pada data uji. Hasil ini menunjukkan efektivitas model IndoBERT dalam memprediksi sentimen dengan dua kelas (“positif” dan “negatif”) dalam teks Bahasa Jawa Ngoko Lugu menggunakan metode *transfer learning* dan *fine-tuning*.

Kata kunci: Analisis Sentimen, Bahasa Jawa, IndoBERT, Transfer Learning

Abstract

The significant increase in the use of social media and online platforms in Indonesia has led to a surge in the production and dissemination of texts in various regional languages, including Javanese. This creates an urgent need for sentiment analysis systems that can understand the nuances of Javanese in the context of user opinions and expressions. This research aims to apply the transfer learning method to the IndoBERT model, with a focus on evaluating its performance in the task of sentiment analysis of Ngoko Lugu Javanese language texts. The research results show that the IndoBERT model achieves high accuracy, precision, recall, and F1-Score in identifying positive and negative sentiments in Ngoko Lugu Javanese language texts, with an accuracy rate of 99% on training data, 90% on validation data, and 90% on test data. These results show the effectiveness of the IndoBERT model in predicting sentiment with two classes (“positive” and “negative”) in Ngoko Lugu Javanese language texts using transfer learning and fine-tuning methods.

Keywords: Sentiment Analysis, Javanese Language, IndoBERT, Transfer Learning

1. Pendahuluan

Peningkatan penggunaan media sosial dan platform daring di Indonesia telah membawa dampak signifikan terhadap produksi dan penyebaran teks dalam berbagai bahasa, termasuk Bahasa Jawa. Menurut hasil riset *We Are Social* tahun 2024, penggunaan media sosial aktif di Indonesia mencapai 167 juta jiwa dari jumlah populasi sebesar 278,7 juta jiwa. Twitter menghasilkan 177 juta *tweet* per bulan dengan jumlah 25,5 juta pengguna di Indonesia. Twitter juga merupakan media sosial urutan ke-6 yang sering diakses di Indonesia [1]. Masyarakat Indonesia semakin aktif berpartisipasi dalam berbagai diskusi, memberikan komentar, dan berbagi pengalaman menggunakan bahasa daerahnya masing-masing. Seiring dengan itu, muncul kebutuhan untuk mengembangkan sistem analisis sentimen teks yang mampu memahami nuansa Bahasa Jawa dalam konteks opini dan ekspresi pengguna.

Bahasa Jawa merupakan salah satu bahasa daerah utama yang dituturkan oleh masyarakat yang tinggal di Pulau Jawa. Berdasarkan data persebaran penggunaan Bahasa Jawa di laman petabahasa.kemdikbud.co.id, diketahui bahwa Bahasa Jawa dituturkan oleh etnik Jawa yang di antaranya

tinggal di Provinsi Jawa Tengah, Daerah Istimewa Yogyakarta (DIY), Jawa Timur, Jawa Barat, dan Banten. Bahasa ini umumnya dikenal memiliki 3 ragam bahasa yaitu Bahasa Jawa Ngoko (BJN), Bahasa Jawa Madyo (BJM), dan Bahasa Jawa Kromo (BJK). Aturan penggunaan ragam Bahasa Jawa tersebut biasanya didasarkan pada siapa lawan tutur kita [2].

Berdasarkan hasil analisis kontrastif verba Bahasa Jawa dan Bahasa Indonesia, peneliti menemukan bahwa verba Bahasa Jawa memiliki kemiripan dengan verba Bahasa Indonesia [3]. Persamaan ini memunculkan tantangan baru dalam mengadaptasi model analisis sentimen umumnya yang telah dilatih dalam Bahasa Indonesia. Bahasa Jawa dan Bahasa Indonesia memiliki kesamaan dasar dalam unsur dan struktur kebahasaan. Menurut Suhardi, Bahasa Indonesia, Bahasa Banjar, dan Bahasa Jawa memiliki deskripsi kemiripan dan kesamaan unsur dan struktur kebahasaan [4]. Asumsi itu dilandasi oleh hasil studi Linguistik Historis Komparatif yang mengungkapkan bahwa Bahasa Indonesia dan Bahasa Jawa merupakan satu kelompok, yaitu kelompok Bahasa Austronesia Barat. Selain itu, menurut Fudiat (1981) melalui Suhardi, ada persamaan yang cukup besar antara Bahasa Banjar, Bahasa Indonesia, dan Bahasa Jawa. Segi-segi kesejarahan mengindikasikan adanya pengaruh Bahasa Melayu (sekarang Bahasa Indonesia), Jawa, dan Dayak terhadap Bahasa Banjar [4].

Beberapa studi sebelumnya telah dilakukan oleh beberapa peneliti dalam bidang pemrosesan teks Bahasa Jawa. Sebuah penelitian pernah melakukan klasifikasi dialek Bahasa Jawa menggunakan algoritma *Naive Bayes* yang menghasilkan tingkat akurasi sebesar 96,97%, presisi sebesar 97,53%, dan *recall* sebesar 96,83% [5]. Penelitian lain juga telah mengeksplorasi perbandingan antara 2 metode *Machine Learning* (ML) yaitu *Naive Bayes* dan *Support Vector Machine* (SVM) dalam konteks klasifikasi teks untuk 22 bahasa daerah di Indonesia, termasuk Bahasa Jawa, dan menghasilkan evaluasi akurasi model tertinggi mencapai 89,41% [6].

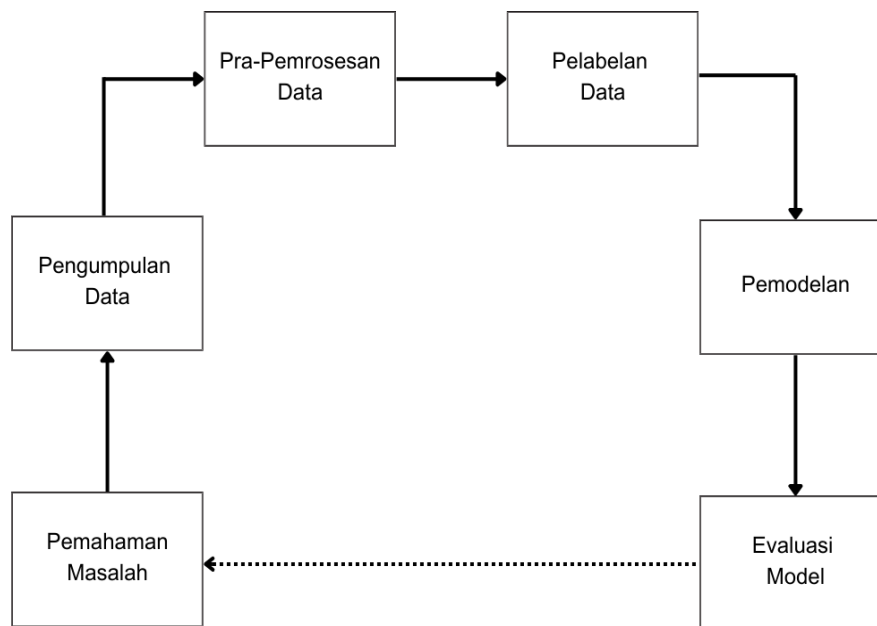
Dalam konteks analisis sentimen teks Bahasa Indonesia, penelitian terkait penerapan metode *transfer learning* juga pernah dilakukan sebelumnya. Dalam penelitian tersebut ditemukan bahwa model yang telah dilakukan proses *pre-train* seperti *Vader*, *Textblob*, dan *Flair* memiliki performa yang lebih unggul dibandingkan dengan model klasifikasi tradisional dalam kasus analisis sentimen [7]. Selain itu, sebelumnya juga pernah dilakukan beberapa penelitian terkait penerapan metode *transfer learning* pada model IndoBERT. Sebuah penelitian membahas tentang implementasi model IndoBERT dalam analisis sentimen untuk dataset *review* aplikasi *mobile* di Indonesia. Temuan penelitian menunjukkan bahwa model IndoBERT mencapai rata-rata akurasi yang lebih tinggi dalam memproses data berbasis leksikon dibandingkan dengan model BERT [8]. Pada penelitian lain, penerapan metode *transfer learning* pada model IndoBERT untuk analisis sentimen pada data *review* aplikasi kesehatan juga pernah dilakukan. Hasil penelitian menunjukkan bahwa penerapan metode *transfer learning* dengan model IndoBERT pada data kesehatan menghasilkan tingkat akurasi yang tinggi, mencapai 96% akurasi, 95% presisi, dan 95% F1-score [9].

Metode *transfer learning* merupakan teknik pembelajaran mesin yang memanfaatkan pengetahuan yang telah dipelajari dari model sebelumnya untuk menyelesaikan tugas baru [10]. Pada penelitian ini, ide untuk menerapkan metode *transfer learning* bertujuan untuk mengevaluasi kemampuan model IndoBERT dalam kasus analisis sentimen teks Bahasa Jawa Ngoko Lugu. Pengetahuan yang akan ditransfer adalah kemampuan IndoBERT dalam memahami konteks dan representasi kata. Model ini telah terbukti efektif dalam tugas analisis sentimen teks Bahasa Indonesia [8]. Maka dari itu, dengan menerapkan metode *transfer learning*, diharapkan model yang dihasilkan dapat lebih efektif dan akurat dalam menangkap makna serta sentimen dalam teks Bahasa Jawa Ngoko Lugu.

Penelitian ini juga didorong oleh kurangnya penelitian yang secara khusus fokus pada penerapan metode *transfer learning* untuk kasus analisis sentimen teks Bahasa Jawa. Oleh karena itu, penelitian ini memiliki relevansi yang signifikan untuk pengembangan teknologi NLP (*Natural Language Processing*) yang dapat memberikan pemahaman yang lebih baik terhadap konten berbahasa Jawa di lingkungan digital.

2. Metode

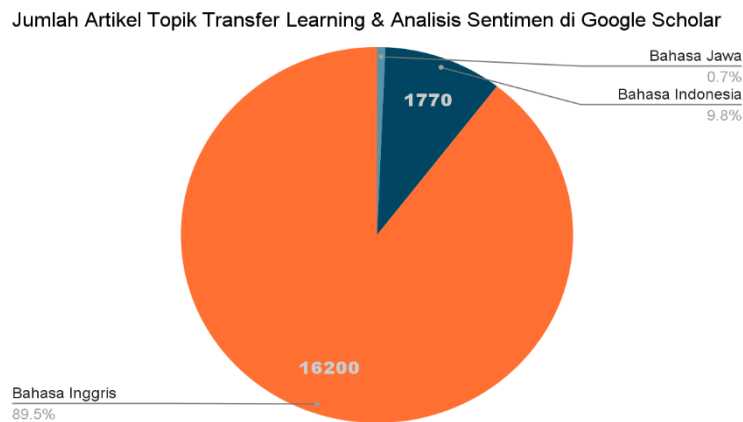
Penelitian ini akan mengikuti metode *lifecycle* sains data yang terdiri dari beberapa tahapan. Diagram pada Gambar 1 menggambarkan dengan terperinci setiap langkah dari tahap pengumpulan data, pembersihan data, pelabelan data, pembagian data, hingga tahap pembuatan dan evaluasi model. Diagram ini akan membantu dalam memvisualisasikan alur kerja penelitian dengan lebih jelas.



Gambar 1. Diagram Metode Penelitian

2.1. Pemahaman Masalah

Penelitian ini dilatarbelakangi oleh fenomena masifnya penggunaan media sosial, terutama Twitter, yang menjadi wadah bagi aktivitas harian dengan jumlah pengguna yang tinggi. Selain itu, penelitian ini juga didorong oleh kurangnya penelitian yang secara khusus fokus pada penerapan metode *transfer learning* untuk kasus analisis sentimen teks Bahasa Jawa berdasarkan dari jumlah artikel untuk topik *transfer learning* dan analisis sentimen pada Bahasa Jawa, Bahasa Indonesia, dan Bahasa Inggris di website Google Scholar menggunakan kueri pencarian: "*sentiment analysis*" ("*Javanese*" OR "*Indonesia*" OR "*English*") "*transfer learning*" seperti yang tertera pada Gambar 2. Oleh karena itu, penelitian ini memiliki relevansi yang signifikan untuk pengembangan teknologi NLP (*Natural Language Processing*) yang dapat memberikan pemahaman yang lebih baik terhadap konten berbahasa Jawa di lingkungan digital.

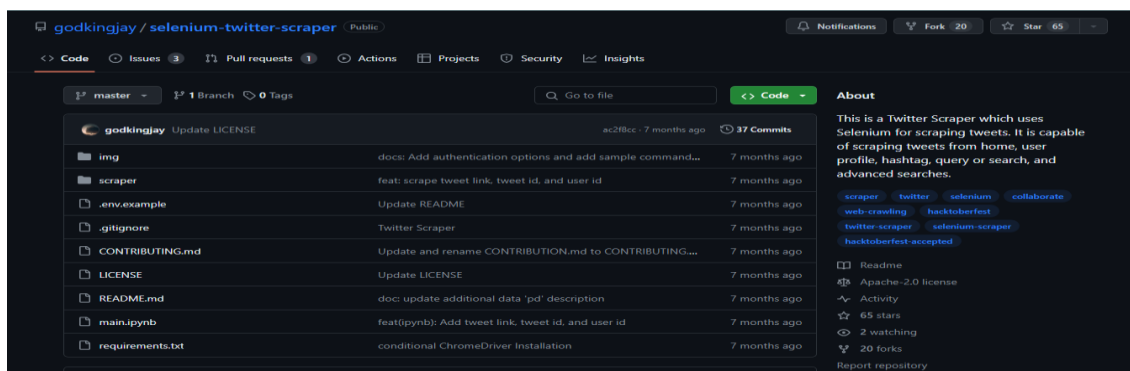


Gambar 2. Jumlah Artikel Topik *Transfer Learning* dan Analisis Sentimen di Website Google Scholar

Keterbatasan tersebut mendorong keinginan peneliti untuk mengeksplorasi analisis sentimen pada teks Bahasa Jawa Ngoko Lugu dengan memanfaatkan metode *transfer learning* pada model IndoBERT. Dengan demikian, penelitian ini bertujuan untuk mengisi kesenjangan pengetahuan tersebut dan memberikan kontribusi pada pengembangan analisis sentimen dalam konteks Bahasa Jawa Ngoko Lugu.

2.2. Pengumpulan Data

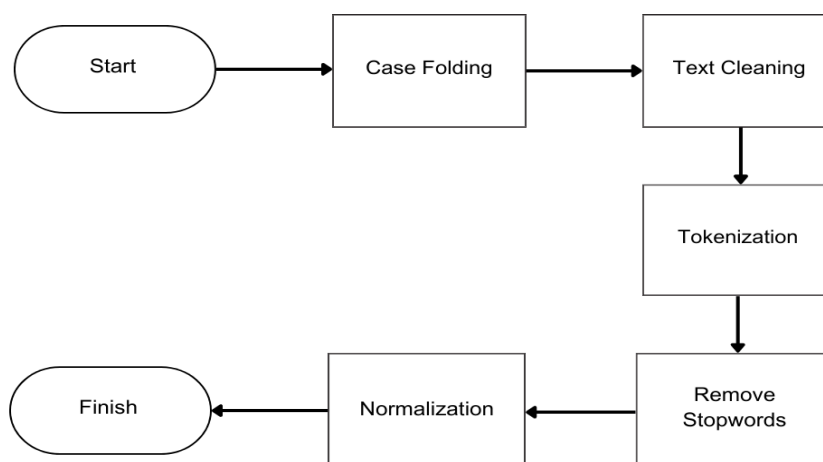
Data yang digunakan dalam penelitian ini diperoleh dari platform media sosial Twitter sebanyak 300 *tweets* menggunakan alat *open source* dari Github bernama *selenium-twitter-scraper* seperti yang tertera di Gambar 3. Alat tersebut dapat diperoleh melalui repositori Github <https://github.com/godkingjay/selenium-twitter-scraper>.



Gambar 3. Repositori Alat *Crawling* Twitter

2.3. Data Preprocessing

Pada tahap ini, data yang berhasil diperoleh akan melalui serangkaian proses persiapan agar dapat dimanfaatkan secara optimal pada tahap pemodelan. Proses persiapan data ini terdiri dari beberapa langkah penting seperti yang diilustrasikan pada Gambar 4. Pertama, dilakukan proses *case folding* untuk mengubah semua huruf menjadi huruf kecil guna konsistensi dalam analisis. Selanjutnya, langkah *text cleaning* dilakukan untuk membersihkan data dari karakter-karakter yang tidak relevan dan menghilangkan duplikasi pada kalimat yang sama. Selain itu, *tokenization* juga diterapkan untuk mengubah setiap kalimat menjadi kumpulan kata atau token yang lebih mudah dianalisis. Kemudian, dilakukan *remove stopwords* untuk menghilangkan kata-kata yang tidak memberikan kontribusi signifikan dalam analisis, seperti konjungsi, kata depan, kata tanya, kata sandang, atau kata ganti. Terakhir, proses *normalization* dilakukan untuk menstandarisasi kata-kata tidak formal ke dalam bentuk kata formal untuk mempermudah proses analisis selanjutnya.



Gambar 4. Diagram Proses *Preprocessing*

2.4. Pelabelan Data

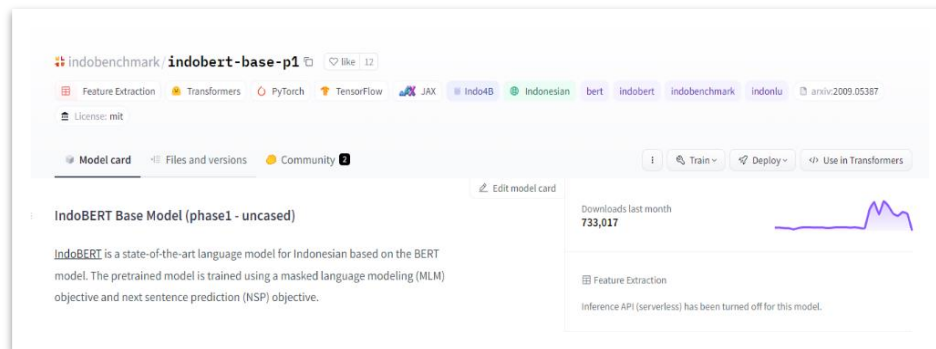
Setelah melewati proses *preprocessing*, data kemudian diekspor dan di-upload ke *Google Spread Sheet* untuk dianotasi sesuai dengan sentimen “*positive*” atau “*negative*” yang terdapat dalam setiap teks supaya dapat digunakan pada tahap pemodelan kasus analisis sentimen dua kelas (*binary class*). Proses anotasi ini dilakukan setelah proses pembersihan teks (*text cleaning*) untuk memastikan bahwa data yang diberi label mencerminkan sentimen yang diungkapkan dalam teks yang dibersihkan dan tidak dipengaruhi oleh kebisingan atau informasi yang tidak relevan. Proses anotasi ini dilakukan secara manual oleh peneliti menggunakan aplikasi *Google Spreadsheet* dengan mengisi kolom “*sentiment*” dengan label sentimen yang sesuai dengan sentimen yang terdapat dalam setiap teks di kolom “*text*” yang dapat dilihat seperti di Gambar 5.

	text	sentiment
1	ya allah aku ora kuat hawane iki kok yo puanase ngentang ngentang	negative
2	wong dudu sopo sopo kok njaluk dikabari terus yo ndak sopan depan itu	negative
3	aku tetep ora bakal ngapusi karo ora bakal korupsi pak	positive
4	halah rek aku iso ngomong jowo kok wkwk gawe gaya gaya wae	positive
5	raimu ki ketok e kesel tak idak idak gelem moco cap cap an nang lawang bis	negative
6	liane ndelok udan kelingan mantan aku ndelok udan kelingan pemean	positive
7	sehabis buang air besar aku lgbt lego banget	positive
8	nek mung njalok perhatian yo tak ke i neng yen njalok bondo sorry rung mapan wkwkwkk	positive
9	hayoben digoblok e ro wong ngenyekan ngomelan ngamukan soale arep dikandani koyo opo wae mesti ngeyelane wong koyo ngono kuwi alias percumah langsung blok njur tinggal nggletak rasah digagas blas	negative
10		

Gambar 5. Sampel Anotasi Data

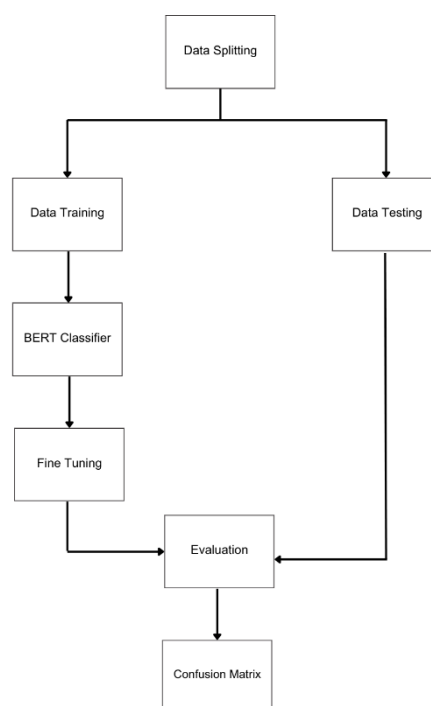
2.5. Pemodelan

Dalam proses ini, metode pembuatan model akan melibatkan implementasi metode *transfer learning* pada model IndoBERT untuk melakukan analisis sentimen pada data hasil *crawling* dari platform media sosial Twitter. Model yang akan digunakan dalam penelitian ini adalah *indobert-base-p1*. Model ini akan diimpor melalui repositori *indobenchmark* yang tersedia di platform *HuggingFace* seperti yang dapat dilihat pada Gambar 6.



Gambar 6. Repositori Model IndoBERT

Proses *fine-tuning* model akan melibatkan beberapa langkah yang telah ditetapkan sebelumnya, seperti pembagian dataset dan *parameter-tuning*. Pada tahap pertama, dataset akan dibagi menjadi dua bagian: data latih dan data uji untuk memaksimalkan keefektifan model yang dibangun. Selanjutnya, metode *transfer learning* akan diterapkan pada model IndoBERT untuk mengadaptasikan model tersebut dengan konteks Bahasa Jawa Ngoko Lugu. Proses *fine-tuning* model kemudian akan dilakukan, di mana model akan disesuaikan dengan *dataset* Bahasa Jawa Ngoko Lugu untuk meningkatkan kinerjanya dalam melakukan analisis sentimen.



Gambar 7. Diagram Tahap Pemodelan

Tahap pemodelan pada Gambar 7 akan dilakukan menggunakan bahasa pemrograman Python dengan memanfaatkan platform Google Colab sebagai *notebook* untuk menjalankan program. Model *pre-trained* yang akan digunakan adalah *indobert-base-p1*, dan akan diimpor langsung dari platform *Huggingface* menggunakan API *Transformers*. Hal ini dapat dilakukan dengan cara menginstal dan mengimpor *library* tersebut menggunakan Python.

2.6. Evaluasi Model

Model yang telah dibangun kemudian akan dievaluasi menggunakan dataset uji yang sudah disiapkan. Proses evaluasi dalam penelitian ini akan memanfaatkan metode evaluasi *confusion matrix* dengan parameter evaluasi yang mencakup akurasi, presisi, *recall*, dan F1-score. Parameter ini digunakan untuk mengukur kinerja model dalam mengklasifikasikan sentimen teks Bahasa Jawa Ngoko Lugu.

3. Hasil dan Pembahasan

3.1. Pengumpulan Data

Proses *crawling* data pada penelitian ini dilakukan dengan menggunakan sistem operasi Ubuntu melalui *Windows Subsystem for Linux 2* (WSL2) seperti yang ada di Gambar 8. Tahapan ini memanfaatkan terminal WSL2 untuk menjalankan proses *crawling* dengan mengikuti petunjuk yang telah disediakan di repositori *selenium-twitter-scraper*.

```
(scraper-venv) %python3 --help: python3 scraper.py --user= --password=
--query="(23bosojowo)20lang3Aid" --tweets 1000
Loading .env file
Loaded .env file

Initializing Twitter Scraper...
Setup WebDriver...
Initializing ChromeDriver...
WebDriver Setup Complete

Logging in to Twitter...
Re-attempting to input username...

Login Successful

Scraping Latest Tweets from (23bosojowo)20lang3Aid search...
Progress: [=====] 100.00% 1000 of 1000
Scraping Complete
Tweets: 1000 out of 1000

Saving Tweets to CSV...
CSV Saved: ./tweets/2024-03-12_18-15-25_tweets_1-1000.csv
```

Gambar 9. Alat *Crawling* Twitter

Hasil proses pengumpulan data dalam penelitian ini mencakup teks dalam Bahasa Jawa Ngoko Lugu yang diperoleh dari platform media sosial Twitter dengan jumlah 3000 *tweets*. Proses pengumpulan data dilakukan dengan melakukan *crawling* pada website Twitter menggunakan alat *selenium-twitter-scraper*, dimana setiap *query* pencarian menghasilkan 1000 *tweets*. Beberapa *keyword query* pencarian yang digunakan antara lain *(#bosojowo) lang:id*, *(#quotesjowo) lang:id*, dan *(#jowostory) lang:id* seperti yang dapat dilihat pada Gambar 9.

	Name	Handle	Timestamp	Verified	Content	Comments	Retweets	Likes	Analytics	Tags
0	gemoynyakamuh	@gemoynyakamuh	2023-08-11T06:53:20.000Z	False	Artine opo?InKadang mripat iso salah ndelok, k...	0	0	0	27	["#katabijak", "#bosojowo"]
1	TM - Indra	@TulMasbro	2023-07-13T04:00:48.000Z	False	BAHAGIA ITU SEDERHANA .. InDelok mbah google k...	0	0	0	64	["#bbcprester", "#bosojowo", "#mbahgugel", "..."]
2	sing waras ojo ngalah. #swong	@mncrgknski	2023-03-25T11:55:45.000Z	True	Istilah 'nakal' ini terlalu halus; diganti apa...	0	0	0	110	["#bosojowo"]
3	brur BRUR	@bravetrackindo	2023-01-23T15:42:27.000Z	False	Bhs mana yg paling ringkas pendek cepat?#bosojowo	0	1	5	75	["#bosojowo"]
4	neko	@neuvilette	2023-01-11T23:49:59.000Z	False	Katto 332 x/6 #BosoJowohttps://katto.vercel.app	0	0	0	11	["#BosoJowo"]

Gambar 8. Hasil *Crawling* Twitter

3.2. Preprocessing Data

Setelah semua data teks terkumpul, tahap selanjutnya adalah proses *preprocessing* yang melibatkan serangkaian langkah untuk mempersiapkan data tersebut agar dapat digunakan pada proses pemodelan. Tahapan ini bertujuan untuk membersihkan data dengan menghilangkan karakter yang tidak relevan sehingga menghasilkan data teks yang terstruktur. Proses ini akan dilakukan menggunakan bahasa pemrograman Python dan Google Colab. Beberapa proses yang akan dilakukan pada tahapan ini meliputi:

1) Transformasi semua teks ke dalam huruf kecil (*case folding*)

Tahapan ini merupakan tahap awal dalam *preprocessing* untuk memastikan konsistensi dalam teks yang akan dianalisis. Setelah proses pelabelan data, dimana setiap teks telah diberi label, langkah selanjutnya adalah mengubah semua huruf dalam teks menjadi huruf kecil seperti pada Tabel 1.

Tabel 1. Hasil Transformasi Teks ke Huruf Kecil

Sebelum	Sesudah
Artine opo? Kadang mripat iso slh ndelok, kpng iso salah krungu, lambe iso salah ngomong, tapi ati ora bakal iso diapusi.#katabijak #bosojowo	artine opo? kadang mripat iso slh ndelok,kpng iso salah krungu, lambe iso salah ngomong, tapi ati ora bakal iso diapusi.#katabijak #bosojowo

Proses ini dilakukan agar tidak ada perbedaan dalam interpretasi kata yang sebenarnya sama namun ditulis dengan huruf besar atau kecil. Misalnya, kata "Kucing" dan "kucing" sebenarnya merujuk pada entitas yang sama, tetapi jika tidak dilakukan transformasi menjadi huruf kecil, kedua kata tersebut akan dianggap berbeda dan dapat memengaruhi hasil analisis. Oleh karena itu, mengubah semua huruf menjadi huruf kecil diperlukan untuk memastikan konsistensi dalam data teks dan mempermudah proses analisis teks lebih lanjut.

2) Pembersihan Teks (*Text Cleaning*)

Pada tahap ini, proses pembersihan teks melibatkan penyaringan kata, karakter, URL, dan simbol yang tidak relevan. Ini termasuk menghilangkan karakter HTML, *hashtag*, *username*, URL, *email*, tanda baca, *retweet* (RT), dan angka dari teks yang sudah diubah ke huruf kecil seperti yang dapat dilihat pada Tabel 2.

Tabel 2. Hasil Pembersihan Teks dari Karakter Spesial

Sebelum	Sesudah
artine opo? kadang mripat iso slh ndelok, kpng iso salah krungu, lambe iso salah ngomong, tapi ati ora bakal iso diapusi.#katabijak #bosojowo	artine opo kadang mripat iso slh ndelok kpng iso salah krungu lambe iso salah ngomong tapi ati ora bakal iso diapusi katabijak bosojowo

3) Tokenisasi

Pada tahap ini, semua teks atau dokumen akan dibagi menjadi unit-unit yang lebih kecil atau kata-kata individual, yang disebut sebagai token yang dapat dilihat seperti pada Tabel 3. Proses ini berfungsi untuk mempersiapkan data teks sehingga setiap kata atau elemen penting dalam teks dapat diproses secara terpisah. Ini membantu dalam menyederhanakan teks sehingga dapat diolah dengan lebih efisien dalam analisis teks yang lebih lanjut.

Tabel 3. Hasil Tokenisasi Teks

Sebelum	Sesudah
artine opo kadang mripat iso slh ndelok kpng iso salah krungu lambe iso salah ngomong tapi ati ora bakal iso diapusi katabijak bosojowo	[artine,opo,kadang,mripat,iso,slh,ndelok,kpng,iso,salah,krungu,lambe,iso,salah,ngomong,tapi,ati,ora,bakal,iso,diapusi,katabijak,bosojowo]

4) Penghapusan kata-kata penghubung (*Stopwords*)

Setelah teks dibersihkan dari karakter dan simbol yang tidak relevan, langkah berikutnya adalah menghapus *stopwords* atau kata-kata yang tidak memiliki makna tambahan dalam konteks analisis yang dapat dilihat pada Tabel 4. Ini termasuk kata-kata seperti kata-kata dari *hashtag*, "dan", "atau", "adalah", "apa", "ke", dan sejenisnya. Proses ini membantu meningkatkan kualitas analisis teks dengan fokus pada kata-kata kunci atau kata-kata yang lebih informatif. Selain itu, tahap ini juga dapat mempercepat proses analisis karena mengurangi jumlah kata yang harus diproses.

Tabel 4. Hasil *Stopwords Removal*

Sebelum	Sesudah
[artine,opo,kadang,mripat,iso,slh,ndelok,kpng,iso, salah,krungu,lambe,iso, salah,ngomong,tapi,ati,ora ,bakal,iso,diapusi,katabijak,bosojowo]	[artine,mripat,kpng,lambe,ati,diapusi]

5) Normalisasi

Setelah proses tokenisasi dilakukan, tahap berikutnya adalah normalisasi dan menghapus kata duplikat. Proses ini bertujuan untuk menyamakan bentuk kata-kata menjadi bentuk standar untuk meningkatkan konsistensi data teks. Misalnya, kata "idn" dan "indo" dapat disamakan menjadi "indonesia" dan juga menghapus semua kata duplikat yang ada pada teks seperti yang dapat dilihat pada Tabel 5. Proses ini membantu dalam mengatasi variasi bentuk kata, seperti bentuk jamak, kehadiran tanda baca, variasi kapitalisasi, dan lainnya sehingga data teks dapat diolah dengan lebih konsisten dan akurat.

Tabel 5. Hasil Normalisasi Teks

Sebelum	Sesudah
[artine,mripat,kpng,lambe,ati,diapusi]	[artine,mripat,kuping,lambe,ati,diapusi]

3.3. Pelabelan Data

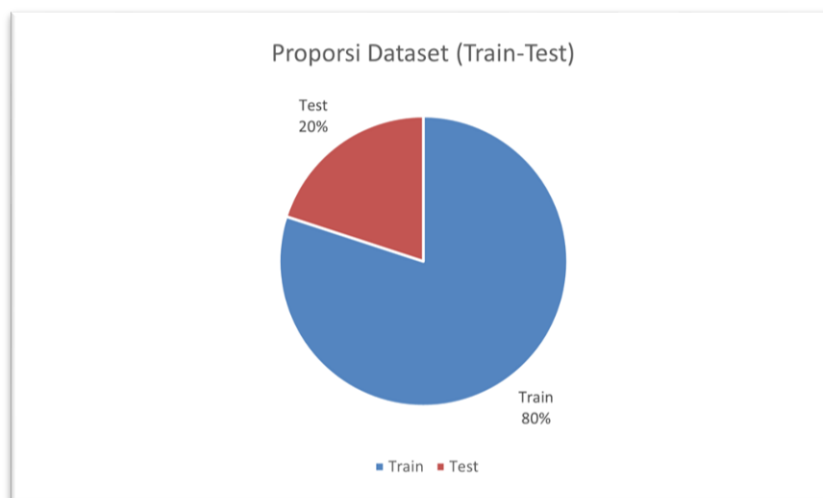
Setelah semua data dibersihkan, langkah selanjutnya melibatkan proses pelabelan sentimen pada teks yang telah dilakukan proses *preprocessing*. Proses ini dilakukan dengan cara menganotasi sentimen “negative” atau “positive” sesuai sentimen yang terdapat pada setiap teks. Proses anotasi ini dilakukan secara manual oleh peneliti menggunakan bantuan aplikasi Google Spread Sheet. Gambar 10 adalah sampel hasil akhir dari proses anotasi teks di *Google Spread Sheet*.

1	text	sentiment
2	ya allah aku ora kuat hawane iki kok yo puanase ngentang ngentang	negative
3	wong dudu sopo sopo kok njaluk dikabari terus yo ndak sopan depan itu	negative
4	aku tetep ora bakal ngapusi karo ora bakal korupsi pak	positive
5	halah rek aku iso ngomong jowo kok wkwk gawe gaya gaya wae	positive
6	raimu ki ketok e kesel tak idak idak gelem moco cap cap an nang lawang bis	negative
7	liane ndelok udan kelingan mantan aku ndelok udan kelingan pemea	positive
8	sehabis buang air besar aku lgbt lego banget	positive
9	nek mung njalok perhatian yo tak ke i neng yen njalok bondo sorry rung mapan wkwkwkk	positive
10	hayoben digoblok e ro wong ngenyekan ngomelan ngamukan soale arep dikandani koyo opo wae mesti ngeyelane wong koyo ngono kuwi alias percumah langsung blok njur tinggal nggletak rasah digagas blas	negative

Gambar 10. Hasil Anotasi Data

3.4. Splitting Data

Setelah proses pelabelan, dataset akan dibagi menjadi dua bagian, yaitu data latih dan data uji. Dalam penelitian ini, jumlah teks yang akan digunakan dalam proses pemodelan adalah 300 teks *tweets*, yang merupakan hasil penyaringan dari total 3000 *tweets* yang telah dikumpulkan. Proses penyaringan tersebut dilakukan untuk menghilangkan duplikasi, menyaring teks yang tidak relevan, dan meningkatkan efisiensi dalam proses pemodelan. Hal ini dilakukan untuk memastikan bahwa *dataset* yang digunakan dalam penelitian ini memiliki kualitas yang tinggi, relevan, dan efisien dalam proses pemodelan. Pembagian dataset ini akan mengikuti proporsi 80-20, di mana 80% dari data akan digunakan sebagai data latih dan 20% sisanya sebagai data uji seperti yang diilustrasikan pada Gambar 11. Langkah ini sangat penting untuk memastikan bahwa model yang dibangun dapat diuji dengan baik dan tidak mengalami *overfit* untuk penerapan analisis sentimen pada Bahasa Jawa Ngoko Lugu.



Gambar 11. Proporsi Pembagian Dataset untuk Pemodelan

3.5. Pemodelan

Setelah semua teks melalui proses *preprocessing*, pelabelan dan pembagian dataset, kemudian langkah selanjutnya adalah proses *modeling*. Pada proses ini akan dilakukan dua tahap, yaitu proses penerapan model IndoBERT dan juga proses *fine-tuning* model.

1) Penerapan Model IndoBERT

Pada tahap ini, model *pre-trained* yang akan digunakan adalah *indobert-base-p1*. Model ini dapat diperoleh dari repositori Indobenchmark melalui platform HuggingFace. Daftar model *pre-trained* yang terdapat di repositori Indobenchmark beserta spesifikasinya dapat dilihat pada Gambar 12.

All Pre-trained Models			
Model	#params	Arch.	Training data
indobenchmark/indobert-base-p1	124.5M	Base	Indo4B (23.43 GB of text)
indobenchmark/indobert-base-p2	124.5M	Base	Indo4B (23.43 GB of text)
indobenchmark/indobert-large-p1	335.2M	Large	Indo4B (23.43 GB of text)
indobenchmark/indobert-large-p2	335.2M	Large	Indo4B (23.43 GB of text)
indobenchmark/indobert-lite-base-p1	11.7M	Base	Indo4B (23.43 GB of text)
indobenchmark/indobert-lite-base-p2	11.7M	Base	Indo4B (23.43 GB of text)
indobenchmark/indobert-lite-large-p1	17.7M	Large	Indo4B (23.43 GB of text)
indobenchmark/indobert-lite-large-p2	17.7M	Large	Indo4B (23.43 GB of text)

Gambar 12. Spesifikasi Model *Pre-trained* di Repositori Indobenchmark

Pada proses implementasi model, semua konfigurasi dan utilitas model IndoBERT, seperti jumlah label sentimen dan metrik evaluasi yang digunakan akan disesuaikan untuk kasus analisis sentimen pada teks Bahasa Jawa Ngoko Lugu.

Secara *default*, model ini memiliki jumlah parameter sebanyak 124.5 juta parameter dengan jumlah data *training* sebesar 23.43 GB teks. Selain itu, model ini juga bisa digunakan untuk memprediksi 3 jenis sentimen yaitu positif, negatif, dan netral. Namun, karena pada penelitian ini hanya menggunakan 2 jenis sentimen saja, maka peneliti akan mengubah kode utilitas supaya model dapat digunakan hanya untuk memprediksi 2 jenis sentimen yaitu positif dan negatif.

Pada proses penerapan model, tahap pertama yang dilakukan adalah memuat semua modul dan utilitas yang diperlukan. Modul utama yang digunakan untuk proses penerapan model pada penelitian ini adalah *transformers* dan *pytorch*. *Transformers* adalah *library open-source* khusus untuk pemrosesan bahasa alami (NLP) yang dikembangkan oleh *Hugging Face*. *Library* ini digunakan untuk implementasi berbagai arsitektur *Transformer* atau *pre-trained* model seperti BERT, GPT (*Generative Pre-trained Transformer*), RoBERTa, dan sebagainya. Sedangkan *Pytorch* adalah *framework open-source* untuk komputasi ilmiah yang berfokus pada penerapan *deep learning*. *Framework* ini digunakan untuk memuat, menerapkan, dan mengevaluasi model IndoBERT dengan mudah untuk membangun aplikasi NLP berbasis IndoBERT.

2) *Fine-tuning* Model

Pada proses *fine-tuning*, model dilatih dengan cara melakukan penyesuaian parameter (*parameter-tuning*) seperti jumlah *learning rate*, *epoch*, *batch size*, *max sequence length* dan metode *multiprocessing*. Pada tahap ini, proses pelatihan model dilakukan menggunakan metode *Cross Validation* dengan *5-fold* dan proporsi dataset 80-20 (*train-valid*) untuk mengevaluasi kemampuan model dalam menggeneralisasi teks Bahasa Jawa Ngoko Lugu.

Tabel 6. Hasil Parameter Tuning Model

Parameter					Train (Avg)	Validation (Avg)
epoch	learning_rate	batch_size	max_seq_length	multiprocessing		
10	3×10^{-6}	8	512	spawn	50%	52%
10	4×10^{-6}	8	512	spawn	51%	51.5%
10	5×10^{-6}	8	512	spawn	99%	83.25%
10	6×10^{-6}	8	512	spawn	99%	90%
10	8×10^{-6}	8	512	spawn	55%	65%

Hasil *parameter-tuning* pada Tabel 6, diperoleh model dengan rata-rata matriks tertinggi dengan konfigurasi parameter yaitu: (*epoch=10*, *learning_rate=6* $\times 10^{-6}$, *batch_size=8*, *max_sequence_length=512*, *multiprocessing='spawn'*) seperti pada Tabel 6.

Tabel 7. Parameter Model Terbaik

Parameter	Nilai
learning rate	0,000006
epoch	10
batch size	8
max sequence length	512
multiprocessing method	spawn

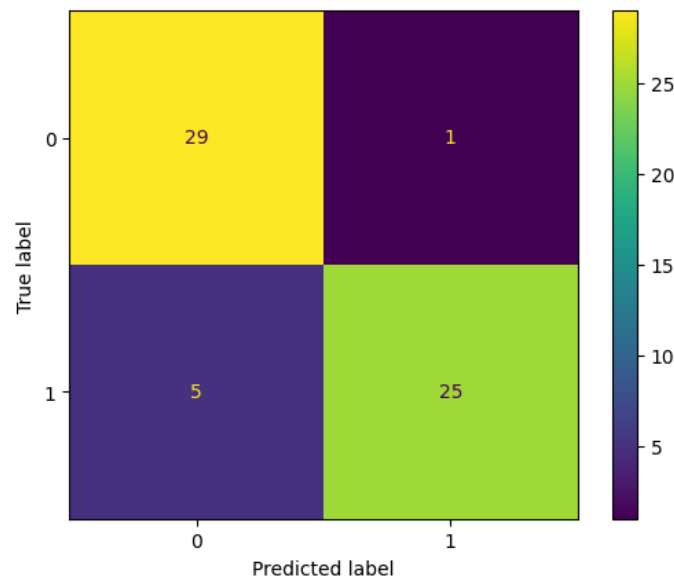
Pada proses *training* dengan konfigurasi *hyperparameters* di Tabel 7, model IndoBERT mendapatkan akurasi, presisi, *recall*, dan F1-Score tertinggi sebesar 99% pada *data training*, kemudian pada *data validation*, model mendapatkan akurasi, presisi, *recall*, dan F1-Score sebesar 90%. Tabel 8 adalah hasil rata-rata akurasi, presisi, *recall*, dan F1-Score pada *data training* dan *data validation* pada *5-folds*.

Tabel 8. Hasil Rata-Rata Cross Validation 5-Folds

Metrik	Training	Validation
Akurasi	99.4%	90.6%
Presisi	99.4%	90.8%
Recall	99.4%	90.4%
F1-Score	99.4%	90.2%

3.5.1. Evaluasi Model

Untuk melakukan proses evaluasi model, peneliti menggunakan API dari *Hugging Face* bernama *confusion_matrix* agar dapat menampilkan hasil evaluasi model dalam bentuk tabel *confusion matrix*. Proses ini dilakukan dengan cara menginstal dan mengimpor *library* bernama *evaluate*. Gambar 13 adalah hasil visualisasi tabel *confusion matrix* pada proses evaluasi model.



Gambar 13. Tabel *Confusion Matrix* Evaluasi Model

Grafik pada Gambar 13 menunjukkan bahwa model dapat memprediksi dengan benar pada kelas “positif” sebanyak 25 *instance* (TP), dan pada kelas “negatif” sebanyak 29 *instance* (TN). Selain itu, model salah dalam memprediksi 1 *instance* sebagai kelas “positif” (FP), dan 5 *instance* sebagai kelas “negatif” (FN).

Secara keseluruhan, menggunakan metode *confusion matrix*, model berhasil mendapatkan akurasi, presisi, *recall*, dan F1-score sebesar 90% pada *data testing* yang dapat dilihat pada Tabel 9.

Tabel 9. Hasil Evaluasi Model Pada Data Uji

Metrik	Testing
Akurasi	90%
Presisi	89.9%
Recall	90%
F1-Score	90.7%

4. Kesimpulan dan Saran

Berdasarkan hasil penelitian yang telah dilakukan, model IndoBERT memiliki kinerja yang cukup baik dalam memproses teks berbahasa Jawa Ngoko Lugu untuk kasus analisis sentimen dengan dua kelas (*binary class*). Hal ini ditunjukkan dengan peningkatan performa yang signifikan setelah melakukan proses *fine-tuning* dengan mengubah beberapa parameter seperti *learning rate*, jumlah *epoch*, *hidden layer*, *batch size*, dan metode *multiprocessing*. Model mendapatkan akurasi, presisi, *recall*, dan F1-score pada *data training* sebesar 99%. Pada *data validation* dan *data testing* model mendapatkan akurasi, presisi, *recall*, dan F1-score sebesar 90%. Penelitian ini menunjukkan bahwa model *pre-trained* IndoBERT dapat digunakan untuk memprediksi sentimen dengan dua kelas yaitu “positif” dan “negatif” pada teks Bahasa Jawa Ngoko Lugu melalui penerapan metode *transfer learning* dan proses *fine-tuning* model.

Dalam penelitian ini, penerapan metode *transfer learning* pada model IndoBERT untuk kasus analisis sentimen teks Bahasa Jawa Ngoko Lugu telah dilakukan menggunakan *dataset tweet* dari platform Twitter. Untuk penelitian selanjutnya, disarankan untuk menerapkan dan mengevaluasi model IndoBERT untuk kasus analisis sentimen teks Bahasa Jawa Ngoko Lugu dengan kelas lebih dari dua (*multiclass*), atau menerapkan model IndoBERT untuk memproses teks bahasa daerah lainnya.

Daftar Pustaka

- [1] A. Riyanto, “Hootsuite (We are Social): Data Digital Indonesia 2024.” Accessed: Jun. 11, 2024. [Online]. Available: <https://andi.link/hootsuite-we-are-social-data-digital-indonesia-2024/>
- [2] H. Rani, E. Retnaningtyas, and U. M. Merauke, “Bahasa Jawa Sebagai Identitas Generasi Muda Masyarakat Pendatang Dari Jawa DI Merauke,” vol. 6, no. 2, pp. 198–212, 2022, doi: <https://doi.org/10.20961/hsb.v6i2.60616>.
- [3] A. Bambang, H. Balai, B. Provinsi, S. Utara, and J. Kolam, “Analisis Kontrasif Afiksasi Verba Bahasa Jawa Dengan Bahasa Indonesia (Affixation Contrastive Analysis Between Javanese Language and Indonesian Language),” 2015. doi: <https://doi.org/10.26499/mm.v13i1.1221>.
- [4] S. Suhardi and J. Santoso, “Analisis Kontrasif Bahasa Indonesia, Jawa, Dan Banjar Sebagai Dasar Penyusunan Model Pembelajaran Bahasa Indonesia Permulaan,” *LITERA*, vol. 10, no. 2, 2013, doi: [10.21831/ltr.v10i2.1160](https://doi.org/10.21831/ltr.v10i2.1160).
- [5] G. Angeline, A. P. Wibawa, and U. Pujianto, “Klasifikasi Dialek Bahasa Jawa Menggunakan Metode Naives Bayes,” *Jurnal Mnemonic*, vol. 5, no. 2, pp. 103–110, 2022, doi: [10.36040/mnemonic.v5i2.4748](https://doi.org/10.36040/mnemonic.v5i2.4748).
- [6] B. Rakajati and E. Y. Hidayat, “Perbandingan Metode Naïve Bayes dan Support Vector Machine Pada Klasifikasi 22 Bahasa Daerah,” 2024, doi: [10.30865/mib.v8i1.7236](https://doi.org/10.30865/mib.v8i1.7236).
- [7] A. A. Arifiyanti, D. S. Y. Kartika, and C. J. Prawiro, “Using Pre-Trained Models for Sentiment Analysis in Indonesian Tweets,” in *Proceedings - International Conference on Informatics and Computational Sciences*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 78–83. doi: [10.1109/ICICoS56336.2022.9930599](https://doi.org/10.1109/ICICoS56336.2022.9930599).
- [8] K. S. Nugroho, A. Y. Sukmadewa, H. Wuswilahaken Dw, F. A. Bachtiar, and N. Yudistira, “BERT Fine-Tuning for Sentiment Analysis on Indonesian Mobile Apps Reviews,” *ACM International Conference Proceeding Series*, pp. 258–264, 2021, doi: [10.1145/3479645.3479679](https://doi.org/10.1145/3479645.3479679).
- [9] H. Imaduddin, F. Y. A’la, and Y. S. Nugroho, “Sentiment Analysis in Indonesian Healthcare Applications using IndoBERT Approach,” *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 8, pp. 113–117, 2023, doi: [10.14569/IJACSA.2023.0140813](https://doi.org/10.14569/IJACSA.2023.0140813).
- [10] Zaid Alyaeai, Maged Saeed AlShaibani, and Irfan Ahmad, “Transfer Learning in Natural Language Processing: A Survey,” *Mathematical Statistician and Engineering Applications*, 2020, doi: [10.17762/msea.v70i1.2312](https://doi.org/10.17762/msea.v70i1.2312).