JIMP : Jurnal Informatika Merdeka Pasuruan
Vol.7 No.3 Desember 2022, P-ISSN : 2502-5716, E-ISSN : 2503-1945

Terakreditasi Peringkat Sinta 4 berdasarkan Petikan dari Keputusan Menteri Riset dan Teknologi/ Kepala Badan Riset dan Inovasi Nasional Nomor 200/M/KPT/2020, masa berlaku mulai Vol.3 No.1 tahun 2018 s.d Vol.7 No.2 tahun 2022

Implementasi Fine-Tuning BERT untuk Analisis Sentimen terhadap Review Aplikasi PUBG Mobile di Google Play Store

Alex Sander Prasetya Braja[1], Achmad Kodar [2]
[1,2] *Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana Jakarta, Indonesia*
[1]41518110021@student.mercubuana.ac.id
[2]achmad.kodar@mercubuana.ac.id

Received: 19-12-2022; Accepted: 27-08-2023; Published: 12-09-2023

Abstrak— ***Game online adalah salah satu hal yang paling relevan untuk beradaptasi dengan teknologi internet. PUBG Mobile adalah salah satu game online terpopuler di Indonesia, telah diunduh lebih dari 500 juta kali dengan 41,8 juta ulasan pengguna pada tahun 2022 di Google Play. Ulasan pengguna memainkan peran penting dalam keberhasilan pengembangan aplikasi.. Ulasan pengguna berupa teks dalam format data tidak terstruktur yang menimbulkan kerumitan saat bekerja dengan analisis sentimen. Ada sebuah pendekatan baru yang disebut BERT. BERT Ini model transfer-learning memperkenalkan model pre-training yang diperlukan untuk lebih baik dalam representasi konteks tekstual. Penelitian ini menguji kinerja BERT untuk analisis sentimen menggunakan dua model pre-training. Kami menggunakan model pre-training IndoBERT$_{BASE}$ dan BERT$_{BASE}$ Multilingual. Data yang digunakan adalah ulasan pengguna untuk aplikasi PUBG Mobile di Google Play Store. Kami juga melakukan pengaturan hyperparameter untuk menemukan model pencarian yang optimal menggunakan dua pendekatan pelabelan data: pelabelan berbasis skor dan pelabelan berbasis TextBlob untuk menentukan efisiensi model. Hasil percobaan menunjukkan bahwa model fine-tuned IndoBERT memiliki akurasi yang lebih baik dalam pelabelan data berbasis Textblob dengan akurasi tertinggi 94 % pada learning rate 0.00002, batch size 32, jumlah epoch 5, dan waktu pelatihan 12 menit.***

Kata kunci— ***Game Online, Analisis sentimen, BERT, PUBG Mobile***

Abstract— Online games are one of the best things to adapt to Internet technology. PUBG Mobile is among the most popular online games in Indonesia, having been downloaded over 500 million times with 41.8 million user reviews in 2022 on Google Play. User feedback plays an important role in the successful development of the app. User feedback is the text as unstructured data, which creates complexity when working with sentiment analysis. There is a new approach called BERT, this transfer-learning model introduces the necessary pre-training models to get a better contextual representation. This study examines BERT's performance in sentiment analysis using two pre-training models. We use the IndoBERTBASE and BERTBASE Multilingual pre-training models. The data used are user reviews for the PUBG Mobile app on Google Play Store. We also make hyperparameter adjustments to find the optimal search model using two data labeling approaches: score-based labeling and TextBlob-based labeling to determine model effectiveness. The experimental results show that IndoBERT's fine-tuned model has better accuracy in labeling Textblob-based data with the highest accuracy of 94% at a learning rate of 0.00002, batch size of 32, number of epochs of 5, and training time of 12 minutes.
.

Keywords— Online Games, Sentiment Analysis, BERT, PUBG Mobile

I. PENDAHULUAN

Evolusi teknologi yang serba cepat telah mengubah dunia selama beberapa dekade terakhir. Salah satu perubahan besar adalah perkembangan teknologi internet. Game online merupakan salah satu hal yang paling relevan untuk beradaptasi dengan teknologi internet. Game online kini telah menjadi pilihan hiburan yang populer bagi orang-orang dari semua lapisan masyarakat.

Seiring dengan semakin populernya game online, begitu pula teknologi yang membuat smartphone semakin populer. Menurut laporan *Limelight Networks* pada Januari 2021, Orang Indonesia bermain game sedikit lebih banyak (8,54 jam/minggu) daripada rata-rata global (8,45 jam/minggu). Dalam angka ini, Indonesia menempati urutan keempat, di belakang China, Vietnam, dan India [1]. Pada tahun 2022, jumlah pengguna smartphone di Indonesia juga diperkirakan akan tumbuh dari 47,3% pada tahun 2015 menjadi sekitar 55% pada tahun 2019. Pada Januari 2021, Per Januari 2021, terdapat 202,6 juta pengguna internet di Indonesia. Dari tahun 2020 hingga 2021, jumlah pengguna internet di Indonesia akan meningkat sebanyak 27 juta orang [2]. Dengan maraknya pengguna smartphone dan internet, Game online dapat terus mendominasi pasar game Indonesia.

Perkembangan game online di Indonesia sangat pesat. Hal ini dapat dibuktikan dengan banyaknya pemain yang aktif pada masing-masing game online tersebut. Menurut riset dan data yang dikumpulkan oleh tim GGWP.ID, game online seperti *Mobile Legends*, *Free Fire*, dan *PUBG Mobile* berkontribusi terhadap peningkatan jumlah pemain platform mobile dari tahun 2016 hingga 2020 [3]. Salah satu game online yang paling populer di Indonesia adalah *Playerunknown's Battleground Mobile* atau lebih dikenal dengan PUBG Mobile. PUBG Mobile adalah game online bergenre *Battle Royale* yang dikelola oleh *Tencent Games* & *Krafton, Inc.* terletak di Singapura.

Google play store adalah salah satu platform aplikasi android terbesar dan terpopuler. *Google play store* menyediakan mekanisme bagi pengguna aplikasi android untuk menilai aplikasi dan memberikan umpan balik dari ulasan pengguna. Ulasan pengguna sering kali

DOI http://dx.doi.org/10.51213/jimp.v7i3.779

menyertakan informasi penting seperti laporan bug dan permintaan fitur yang dapat digunakan pengembang untuk meningkatkan dan memelihara aplikasi mereka [4], [5].

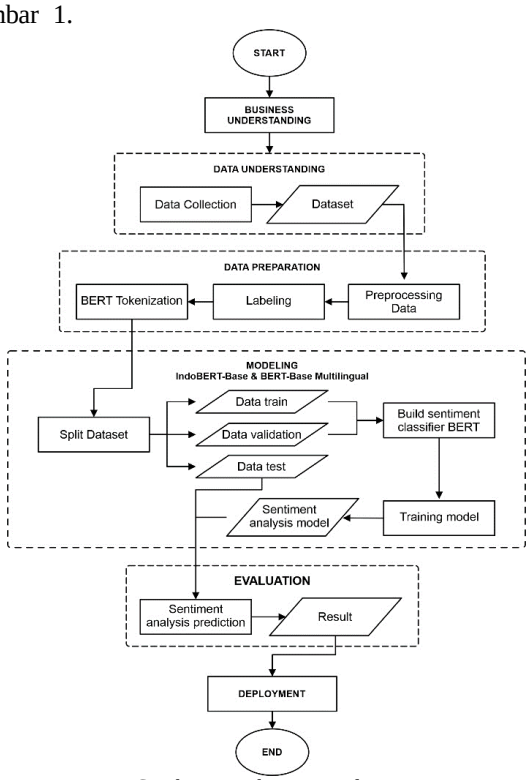
Analisis sentimen adalah tugas mendasar dalam *natural language processing*. Analisis sentimen disebut *opinion mining* dan *opinion analysis* yang bertujuan untuk menemukan polaritas teks dan mengklasifikasikannya menjadi positif, netral atau negatif. Area analisis sentimen meliputi analisis komentar berita, analisis komentar produk, analisis ulasan film, dan bidang lainnya [6], [7]. Analisis sentimen dapat digunakan untuk menyimpulkan konten tekstual dari setiap pendapat dalam sebuah ulasan, apakah pengguna suka atau tidak suka ketika menggunakan aplikasi. Dengan mempelajari preferensi pengguna melalui ulasan, pengembang dapat meningkatkan faktor kegunaan untuk semakin memperkuat bisnis.

Banyak penelitian telah dilakukan sebelumnya tentang analisis sentimen seperti penelitian [8], dengan menggunakan metode *naive bayes* dalam analisis sentimen pada vaksin Covid-19 di Indonesia. Penelitian [9], menggunakan algortima *Naive Bayes Classifier* dan *Support Vector Machine*. Penelitian [10], memperkenalkan model representasi bahasa baru yang paling akurat disebut BERT (*Bidirectional Encoder Representations From Transformers)*, BERT adalah model representasi bahasa baru yang mencapai *state-of-art* pada sebagian besar tugas *Natural Language Processing* yang diimplementasikannya. BERT dikembangkan berdasarkan berbagai metode seperti *Deep Learning* dan *Semi-Supervised Learning.* Selain itu, BERT mengungguli teknik *Natural Language Processing* tradisional [11]. Penelitian [12], akurasi yang diperoleh untuk analisis sentimen dengan BERT adalah 73%. Data yang digunakan adalah Data yang terdiri dari 2000 review, 1000 review positif dan 1000 review negatif yang tersedia dalam bahasa Inggris di *Website Cornelledu*. Namun, dalam penelitian ini sentimen diklasifikasikan menjadi sentimen positif dan sentimen negatif. Hasil penelitian menunjukkan bahwa ini adalah hasil yang sangat baik dibandingkan dengan model lainnya. Untuk mendapatkan hasil model klasifikasi yang lebih efektif, model BERT membutuhkan data dalam jumlah besar. *Pre-training* model BERT dari awal pada dataset dalam jumlah kecil akan menyebabkan *overfitting*. Oleh karena itu, *pre-training* BERT membutuhkan dataset dalam jumlah besar [7]. Penelitian [13], melakukan eksperimen mendalam untuk menyelidiki keefektifan *fine-tuning* BERT untuk klasifikasi teks menunjukkan hasil terbaru dalam analisis sentimen ulasan. Penelitian [14], menggunakan *fine-tuning* BERT untuk analisis sentimen pada dataset ulasan Vietnam menunjukkan bahwa BERT sedikit mengungguli model lain yang menggunakan *GloVe* dan *FastText*.

Berdasarkan hal tersebut maka penelitian yang dilakukan kami menggunakan metode *deep learning* untuk salah satu tugas *natural language processing* yaitu *sentiment analysis* dengan dua model *fine-tuning* yang berbeda yaitu $\text{BERT}_{\text{BASE}}$ Multilingual dan $\text{IndoBERT}_{\text{BASE}}$ [10], [15]. Selanjutnya kami lakukan pengaturan *Hyperparameters* untuk mendapatkan hasil yang terbaik. Kami mengumpulkan 15.000 ulasan dari pengguna dalam bahasa Indonesia tentang Game PUBG Mobile yang diambil dari *google play store* berdasarkan ulasan terbaru. Dataset yang kami kumpulkan tidak diberi label, jadi kami menggunakan dua pendekatan pelabelan untuk mengklasifikasikan dataset berdasarkan polaritasnya, yaitu pelabelan data berbasis *Score* dan pelabelan data berbasis *TextBlob* kemudian diklasifikasikan menjadi tiga kategori positif, neutral dan negatif. Pada penelitian ini kami memilih model BERT karena berdasarkan penelitian sebelumnya dianggap memiliki tingkat akurasi yang tinggi dibandingkan dengan model lainnya.

II. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan CRISP-DM (*Cross-Industry Standard Process for Data Mining*) [16]. Ada beberapa tahapan yaitu *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation* dan *Deployment*. Tahapan penelitian ditunjukkan pada Gambar 1.



Gambar 1. Tahapan Penelitian

A. Bussiness Understanding

Pada langkah ini, kami menetapkan tujuan kami: Analisis sentimen ulasan pengguna pada Aplikasi PUBG Mobile di *Google Play Store*. *Google play store. Google Play Store* mewakili berbagai aplikasi Android yang sebagian besar gratis. Jadi seluruh bagian komentar adalah cerminan dari pendapat orang-orang dengan selera dan pola pikir yang berbeda. Sebuah *scraper web* mengekstraksi informasi dari halaman web dengan cara otomatis dibuat untuk tujuan pengumpulan data yang memberikan informasi tinjauan kritis: nama pengguna, komentar, dan peringkat setiap aplikasi. Dalam penelitian ini langkah yang dilakukan adalah mencari data ulasan pengguna yang terdapat pada *google play store*.

DOI http://dx.doi.org/10.51213/jimp.v7i3.779

B. Data Understanding

Pada tahap ini, kenali dan pahami data yang dimilikinya. Data yang digunakan adalah data ulasan yang terdapat pada *platform google play store* terkait dengan aplikasi PUBG Mobile dalam bahasa Indonesia. Data diperoleh dengan menggunakan *package* dari *google-play-scraper* yang menyediakan API untuk Python, data akan disimpan dalam format file CSV.

C. Data Preparation

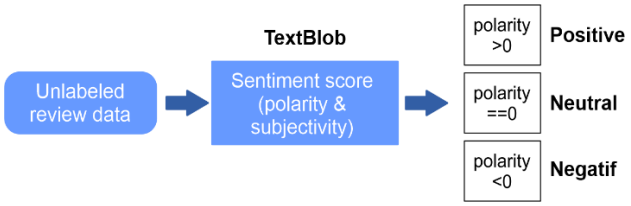
1) Pre-Processing:

Ulasan teks diambil dalam bentuk mentah dari *Google Play* dan sebagian besar ulasan berisi informasi yang tidak relevan. Teknik *preprocessing* digunakan untuk membersihkan kumpulan data yang akan diekstraksi menjadi informasi yang tepat dan membantu meningkatkan performa model BERT. kami menggunakan langkah *Data Cleaning*, langkah-langkah ini termasuk menghapus atribut yang tidak diperlukan dalam analisis sentimen seperti *reviewId*, *userName*, *userImage*, *thumbUpCount*, *reviewCreatedVersion*, *at*, *replyContent*, dan *repliedAt*, tokenisasi kata, menghapus *Stopwords*, semua kata diubah menjadi huruf kecil (*case folding*), menghapus angka, tanda baca, emoji, spasi ganda dan normalisasi kata.

2) Labeling:

Proses pelabelan data dapat memudahkan model untuk mengklasifikasikan emosi. Input model adalah konteks teks ke angka, sehingga data teks harus diberi label terlebih dahulu. Dalam proses pelabelan data, kami menggunakan pelabelan data berbasis *score* yang dilakukan dengan asumsi nilai (1-2) adalah negatif, nilai (3) adalah netral, dan nilai (4-5) adalah positif. dan kami juga menggunakan pelabelan berbasis TextBlob dengan menentukan nilai *polarity* mulai dari (−1, 0, +1). nilai yang kurang dari 0 adalah negatif, nilai yang sama dengan 0 adalah netral, dan nilai yang lebih besar dari 0 adalah positif [17].

Gambar 2. Pelabelan Data Berbasis Score



Gambar 3. Pelabelan Data Berbasis TextBlob

3) BERT Tokenization:

Setelah data diberi label, berikut ini harus dipertimbangkan untuk melatih model BERT [10], Pertama tambahkan token khusus di awal dan akhir komentar. token. [SEP]: token khusus yang ditambahkan di akhir setiap kalimat. [CLS]: Tugas klasifikasi memerlukan prefiks token CLS khusus di awal setiap kalimat, dan token ini memiliki arti khusus. Model BERT terdiri dari 12 *Layers Transformers*. Setiap *transformers* mengambil daftar *embbeding tokens* dan menghasilkan jumlah *Embedding* yang sama dalam output tetapi dengan nilai fitur yang diubah. Pada *output transformers* terakhir, pengklasifikasi hanya menggunakan *embedding* pertama sesuai dengan token [CLS].

Kedua,Token *padding* [PAD]: memotong komentar menjadi satu panjang konstan. Kalimat dalam dataset memiliki panjang yang berbeda, sehingga BERT memiliki dua batasan: Semua kalimat harus memiliki panjang yang sama sehingga dipadatkan atau dipotong menjadi satu panjang yang tetap dan panjang kalimat maksimum adalah 512 token.

Ketiga, membedakan token *padding* [PAD] secara eksplisit, "*attention mask*" hanyalah sebuah array dari 1 dan 0 yang menunjukkan token mana yang token *padding* [PAD] dan mana yang tidak. token [MASK] ini memberi tahu mekanisme "*attention mask*" di BERT untuk tidak *embedding tokens*, Proses ini mengonversi setiap kalimat menjadi angka, kemudian menyandikannya dan memasukkannya ke model BERT. *Tokenization* menggunakan *package* dari *hugging face* adalah *BertTokenizer*.

D. Modeling

Tahap *modeling*, data yang telah disiapkan setelah tahap *data preparation* akan di input ke dalam model. Ada beberapa tahap yaitu:

1) Split Dataset:

Proses *split dataset* ini memisahkan data menjadi tiga bagian yaitu *data train*, *data validation* dan *data test*. Pemisahan data ini penting karena menentukan data yang dilatih dan diuji sehingga model dikelompokkan dengan rasio 80%:10%:10% berdasarkan *data train*, *data validation* dan *data test*.

TABEL I
DISTRIBUSI LABEL BERBASIS SCORE DI KETIGA SET DATA

	Positive	Neutral	Negative	Total
Train	3701	3879	3899	11.479
Valid	463	485	488	1435
Test	462	485	487	1435
Total	4.626	4.849	4.874	14.349

TABEL II
DISTRIBUSI LABEL BERBASIS TEXTBLOB DI KETIGA SET DATA

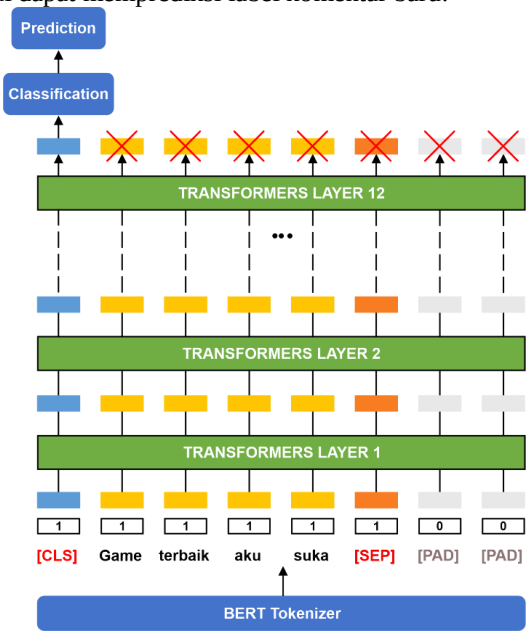
	Positive	Neutral	Negative	Total
Train	4651	3333	3488	11472
Valid	581	417	436	1434
Test	582	417	436	1435
Total	5.814	4.167	4.360	14.341

2) Build Sentiment Classifier:

Setelah split dataset, model dibangun BERT yang dilatih sebelumnya dimuat dan dipasang, membuat *single layer* baru untuk tugas analisis sentimen. Pada lapisan *dropout* ini, fungsi *Softmax* diterapkan untuk menurunkan prediksi *probability* dari model terlatih ke output model BERT, dan untuk membantu mencegah *overfitting* [18]. Terakhir mengonversi label ke encoder sekaligus, jadi model yang

DOI http://dx.doi.org/10.51213/jimp.v7i3.779

dilatih pada komentar dengan label negatif, positif dan netral dapat memprediksi label komentar baru.



Gambar 4. Arsitektur Model BERT

3) Training Model:

Selanjutnya training model BERT, untuk melatih model yang diusulkan perlu mendefinisikan parameter yang berbeda untuk penelitian ini, hyperparameters yang digunakan adalah yang direkomendasikan dalam penelitian [15].

TABEL III
HYPERPARAMETERS YANG DIGUNAKAN DALAM PENELITIAN INI

Parameters	Value
Pre-trained	$BERT_{BASE}$ Multilingual, $IndoBERT_{BASE}$
Optimizer	Adam
Learning rate	0.00002
Batch size	16 & 32
Max epoch	5
Dropout	0.3
Max length	80

E. Evaluation

Tahap Evaluation ini menjelaskan prediksi hasil analisis sentimen untuk kalimat-kalimat dalam dataset. Nilai akurasi tertinggi yang diperoleh selama proses Training Model sebelumnya akan digunakan sebagai nilai akurasi model. Confusion Matrix digunakan untuk mendapatkan prediksi dari model. Dalam penelitian ini, confusion matrix untuk menentukan hasil dari analisis sentimen ini. Penelitian ini menggunakan empat kriteria evaluasi: f1-score, precision, recall dan accuracy yang banyak digunakan dalam tugas klasifikasi teks dan analisis sentimen [19].

1) Accuracy:

Accuracy didefinisikan sebagai rasio prediksi yang benar terhadap total prediksi. *Accuracy* dihitung sebagai (1):

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

True Positive (TP) ini menunjukkan jumlah ulasan yang diprediksi model milik kelas tertentu dan benar-benar milik kelas tersebut. *True Negative* (TN) ini menunjukkan jumlah ulasan yang diprediksi model bukan milik kelas tertentu dan sebenarnya bukan milik kelas. *False Positive* (FP) ini menunjukkan jumlah ulasan model yang diklasifikasikan dalam kategori tertentu sedangkan ulasan tersebut tidak termasuk dalam kategori tersebut. *False Negative* (FN) ini menunjukkan jumlah ulasan yang tidak diklasifikasikan model dalam kategori tertentu sedangkan ulasan tersebut sebenarnya termasuk dalam kategori tersebut.

2) Precision:

Precision adalah proporsi positif terhadap jumlah total positif yang teridentifikasi dalam sampel data test. *Precision* mengukur berapa banyak jumlah positif yang diidentifikasi dengan benar dalam data *test*. Jika semakin tinggi nilai *precision*, maka semakin baik sampel positif dapat diprediksi. *Precision* dihitung sebagai (2):

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

3) Recall:

Recall mengukur tingkat positif sebenarnya dengan menentukan jumlah positif yang teridentifikasi dengan benar dalam data *test*. *Recall* dihitung sebagai (3):

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

4) F1-Score:

F1-score adalah *performance metrics* untuk algoritma klasifikasi yang membantu menunjukkan ketidakseimbangan kelas. Karena *Precision* dan *Recall* digunakan, *F1-score* ini dianggap lebih penting daripada *Precision* dan *Recall*. Jika semakin tinggi nilai *precision* dan *recall* tinggi, maka semakin tinggi juga nilai *f1-score*. Jika keduanya rendah, maka *f1-score* juga akan rendah. *f1-score* memberikan nilai sedang jika salah satunya rendah dan yang lainnya tinggi.

$$F1 = \frac{2^{*}precision^{*}recall}{precision + recall} \quad (4)$$

F. Deployment

Pada tahap *deployment* ini setelah mengetahui hasil analisis sentimen, dilakukan *final project* (laporan tugas akhir). Laporan tugas akhir ini menggambarkan alur analisis sentimen dari pengambilan data hingga proses pembuatan model dan output hasil dari analisis sentimen.

III. HASIL DAN PEMBAHASAN

Pada Bab ini terdapat beberapa topik yang berkaitan dengan hasil penelitian, antara lain:

A. Dataset

Data yang digunakan terdiri dari *review* mengenai Aplikasi PUBG Mobile dan diekstraksi dari *google play store*. Untuk itu penelitian ini menggunakan *package google_play_scraper*. Filter untuk *scraper* diatur, seperti batas jumlah *review* yang diatur menjadi 15.000 dan bahasa diatur ke dalam bahasa indonesia. Dataset diekstraksi pada 4 April 2022 dengan ulasan terbaru. Tabel I menunjukkan beberapa sampel dari kumpulan data yang dikumpulkan.

TABEL IV
DESKRIPSI DATASET

No	Column	Count	Dtype
1	reviewId	15000	object

DOI http://dx.doi.org/10.51213/jimp.v7i3.779

2	userName	15000	object
3	userImage	15000	object
4	content	15000	object
5	score	15000	Int64
6	thumbsUpCount	15000	Int64
7	reviewCreatedVersion	8110	object
8	at	15000	object
9	replyContent	14859	object
10	repliedAt	14859	object

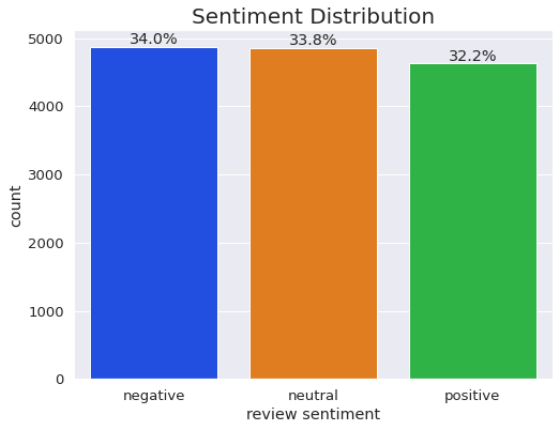
B. Pre-Processing

Data yang dikumpulkan tidak terstruktur. Oleh karena itu, kami menggunakan pendekatan *preprocessing* yang dapat diintegrasikan dengan berbagai tugas *natural language processing* menggunakan *package* dari *natural language toolkit*. kami melakukan *preprocessing* teks pada dataset untuk membersihkan data termasuk menghapus angka, tanda baca, emotikon, menghapus kata yang tidak perlu (*stopwords*) dan menghapus *stemming* tidak dilakukan karena dapat mengubah arti dari keseluruhan kalimat.

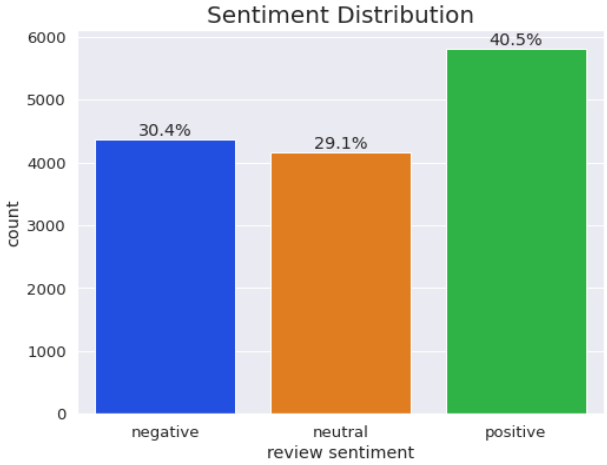
TABEL V
HASIL PRE-PROCESSING DATASET

Pre-Processing	Result
Original Text	Game bagus,tapi masih lemah. Hacker masih dengan mudahnya dapat bermain. Padahal Game Mobile. Koreksi! Dan Sensitivitas tombolnya juga mau berubah sendiri diwaktu tertentu buat capek nyusun ulang sesuai keinginan! 😭
Cleaning Data	game bagus tapi masih lemah hacker masih dengan mudahnya dapat bermain padahal game mobile koreksi dan sensitivitas tombolnya juga mau berubah sendiri diwaktu tertentu buat capek nyusun ulang sesuai keinginan
Normalization Word	game bagus tapi masih lemah hack masih dengan mudahnya dapat bermain padahal game mobile koreksi dan sensitivitas tombolnya juga mau berubah sendiri diwaktu tertentu buat lelah menyusun ulang sesuai keinginan
Stopwords Removal	game bagus lemah hack mudahnya bermain game mobile koreksi sensitivitas tombolnya berubah diwaktu lelah menyusun ulang sesuai

Setelah data bersihkan, data diberi label menggunakan dua metode berbeda. Pertama, hasil pelabelan data dengan menggunakan metode berbasis *Score* menunjukkan rata-rata distribusi kelas sentimen seperti pada Gambar 5. Sedangkan kedua hasil pelabelan data dengan menggunakan metode berbasis *TextBlob* menunjukkan distribusi kelas sentimen yang tidak seimbang, Jumlah sentimen positif dalam pelabelan data berbasis *TextBlob* lebih banyak daripada yang lain, seperti pada Gambar 6.



Gambar 5. Distribusi Sentimen Keseluruhan dengan Menggunakan Pelabelan Berbasis Score



Gambar 6. Distribusi Sentimen Keseluruhan dengan Menggunakan Pelabelan Berbasis TextBlob

C. Build Models

Semua eksperimen dijalankan di *Google Colabs* menggunakan GPU dan CUDA. Kami menggunakan *PyTorch* versi 1.12.1+cu113 untuk tugas analisis sentimen dan menginstal *package* dari *transformers* versi 4.23.1 dari *hugging face*. Kami menggunakan model *pre-trained indobert-base-p2* dari *IndoNLU* dan *bert-base-multilingual-uncased* [10], [15]. Kami *fine-tuning* model BERT dilatih dengan *Optimizer Adam* untuk setiap tugas dengan *learning rate* = 0.00002, *batch size* = 16 dan 32, jumlah *epoch* =5, *dropout* = 0,3 dan *max length* = 80. *CrossEntropyLoss* untuk menghitung fungsi kerugian (loss) multi [10]. Terapkan fungsi *softmax* ke output untuk mendapatkan probabilitas prediksi dari model yang dilatih.

D. Testing and Evaluation

Kami menetapkan jumlah *epoch* menjadi 5 dan menyimpan model terbaik di *data validation* untuk pengujian. Percobaan *fine-tuning* pertama untuk dua model *pre-trained* dilakukan dengan menggunakan *learning rate* = 0.00002 pada *batch size* = 16 dan 32 seperti yang ditunjukkan pada Tabel VI. $\text{BERT}_{\text{BASE}}$ Multilingual dengan *Batch Size* 32 menunjukkan akurasi pengujian terbaik di 0.71149. Tetapi, hasil yang diperoleh menunjukkan bahwa *fine-tuning* model BERT pada dataset yang menggunakan pelabelan berbasis *score*.

DOI http://dx.doi.org/10.51213/jimp.v7i3.779

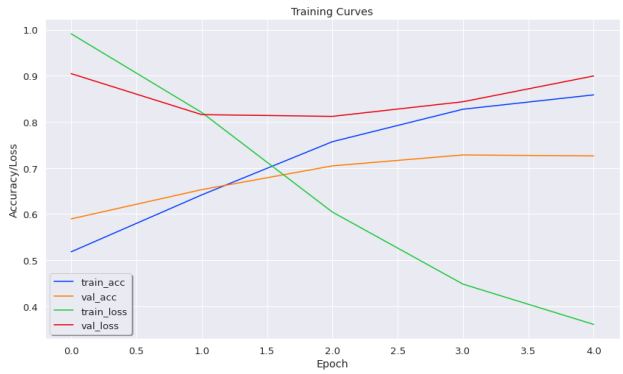
Dalam percobaan *fine-tuning* kedua, *hyperparameter* yang sama seperti sebelumnya digunakan untuk pelabelan data berbasis *TextBlob* dan hasilnya ditunjukkan pada Tabel VII. IndoBERT$_{BASE}$ memperoleh hasil akurasi pengujian terbaik pada *batch size* 32 sebesar 0.93519. Hal ini dikarenakan BERT$_{BASE}$ Multilingual hanya dilatih pada dataset multibahasa yang besar, padahal dataset yang digunakan banyak mengandung kalimat informal dan slang. Sementara IndoBERT$_{BASE}$ dilatih dengan data yang lebih besar dalam bahasa Indonesia dan mencakup bahasa formal dan slang (kata gaul).

TABEL VI
HASIL FINE-TUNING BERT DARI DUA MODEL UNTUK DATA PELABELAN BERBASIS SCORE

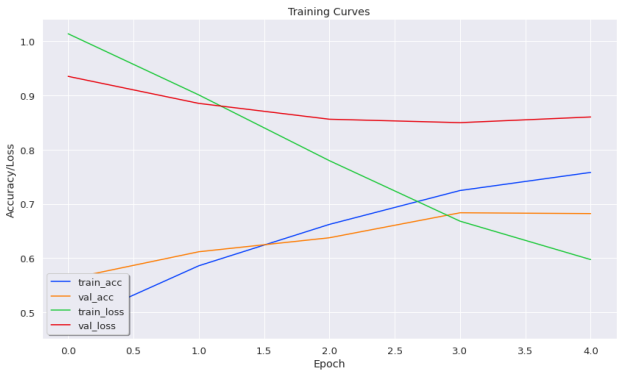
Pre-trained	Batch Size	Acc		Loss		Time Training	Test Acc
		Train	Valid	Train	Valid		
BERT$_{BASE}$ Multilingual	16	0.75799	0.68222	0.59747	0.86005	15min 12s	0.66550
	32	0.73621	0.67804	0.64659	0.83250	13min 36s	0.66341
IndoBERT$_{BASE}$	16	0.83909	0.72334	0.42338	0.89413	15min 24s	0.69895
	32	**0.84162**	**0.71498**	**0.41195**	**0.86319**	**12min 25s**	**0.71149**

TABEL VII
HASIL FINE-TUNING BERT DARI DUA MODEL UNTUK DATA PELABELAN BERBASIS TEXTBLOB

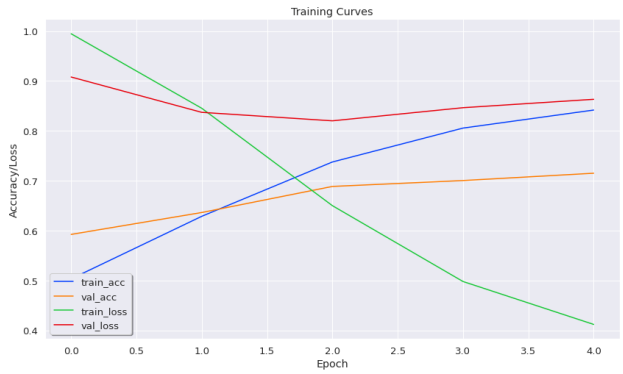
Pre-trained	Batch Size	Acc		Loss		Time Training	Test Acc
		Train	Valid	Train	Valid		
BERT$_{BASE}$ Multilingual	16	0.97332	0.92956	0.10105	0.38371	16min 37s	0.92404
	32	0.96469	0. 91910	0.11276	0.32677	13min 16s	0.92055
IndoBERT$_{BASE}$	16	0.98919	0.93375	0.04084	0.36125	13min 52s	0.92613
	32	**0.98814**	**0.93514**	**0.04332**	**0.31221**	**12min 00s**	**0.93519**



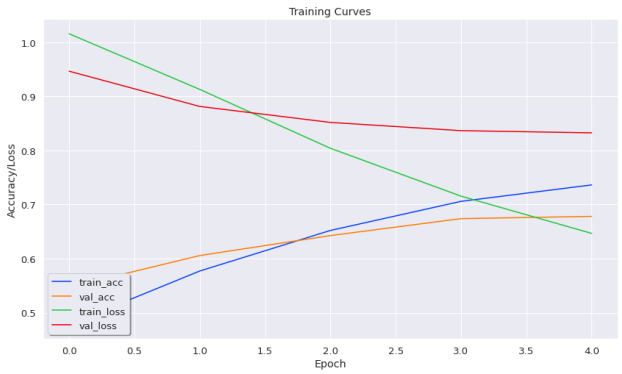
Gambar 7. Training Curves IndoBERT$_{BASE}$ dengan Batch Size 16 Mengunakan Pelabelan Berbasis Score



Gambar 9. Training Curves BERT$_{BASE}$ Multilingual dengan Batch Size 16 Mengunakan Pelabelan Berbasis Score

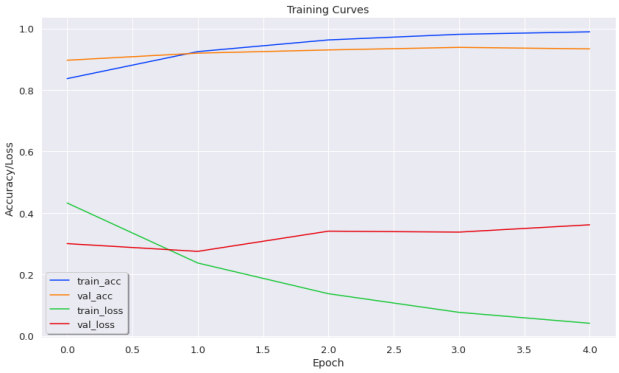


Gambar 8. Training Curves IndoBERT$_{BASE}$ dengan Batch Size 32 Mengunakan Pelabelan Berbasis Score

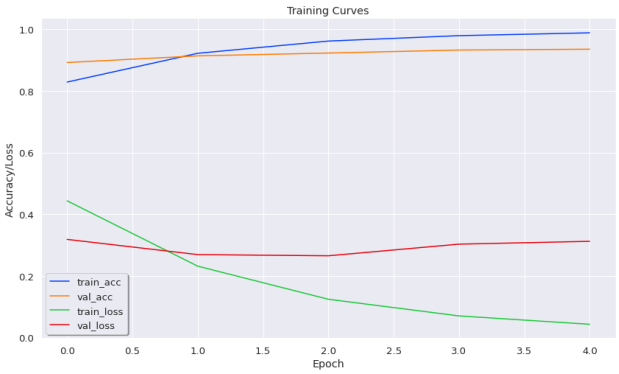


Gambar 10. Training Curves BERTBASE Multilingual dengan Batch Size 32 Mengunakan Pelabelan Berbasis Score

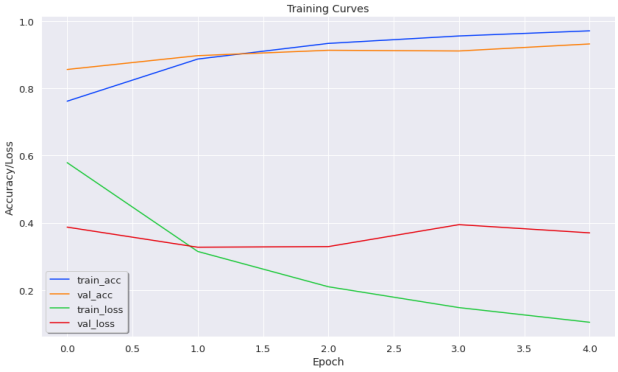
DOI http://dx.doi.org/10.51213/jimp.v7i3.779



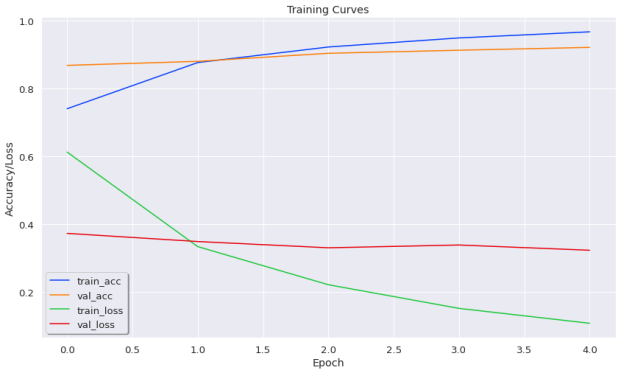
Gambar 11. Training Curves IndoBERT$_{BASE}$ dengan Batch Size 16 Mengunakan Pelabelan Berbasis TextBlob



Gambar 12. Training Curves IndoBERT$_{BASE}$ dengan Batch Size 32 Mengunakan Pelabelan Berbasis TextBlob



Gambar 13. Training Curves BERT$_{BASE}$ Multilingual dengan Batch Size 16 Mengunakan Pelabelan Berbasis TextBlob



Gambar 14. Training Curves BERT$_{BASE}$ Multilingual dengan Batch Size 32 Mengunakan Pelabelan Berbasis TextBlob

Fungsi *accuracy* dan *loss* yang terkait dengan *training* model dari dataset yang diusulkan ditunjukkan pada gambar 7-14. Setiap epoch, nilai *train_loss*, mulai menurun secara bertahap hingga *epoch* 5, namun *val_loss* tidak stabil dan setelah epoch 2 model cenderung *overfitting* dan Kami melihat bahwa *train_acc* dan *val_acc* meningkat hingga *epoch* 5. Hal ini menunjukkan bahwa model memperoleh lebih banyak *knowledge* dan memiliki kinerja yang unggul.

Akurasi pengujian terbaik diperoleh pada *learning rate* 0.00002 pada *batch size* 32 dengan nilai 0.93519, seperti terlihat pada Tabel VII. Hasil *fine-tuning* terbaik dari percobaan kedua ditunjukkan pada Gambar 12 sebagai hasil dari semua percobaan,dengan *batch size* yang lebih besar dapat meningkatkan akurasi dan mempengaruhi waktu pemrosesan secara keseluruhan. IndoBERT$_{BASE}$ sebagai model *pre-trained* khusus yang dilatih dalam bahasa Indonesia dan menunjukkan hasil terbaik dibandingkan dengan model BERT$_{BASE}$ Multilingual.

Dalam percobaan terakhir, kami hanya fokus pada pelabelan data berbasis *TextBlob* karena dataset dengan pelabelan berbasis *Score* memberikan hasil yang buruk. Berikut *Classification Report* dan *Sentiment Prediction* setelah *fine-tuning* kedua model dengan pelabelan data berbasis TextBlob ditunjukkan pada Tabel VIII dan Tabel IX.

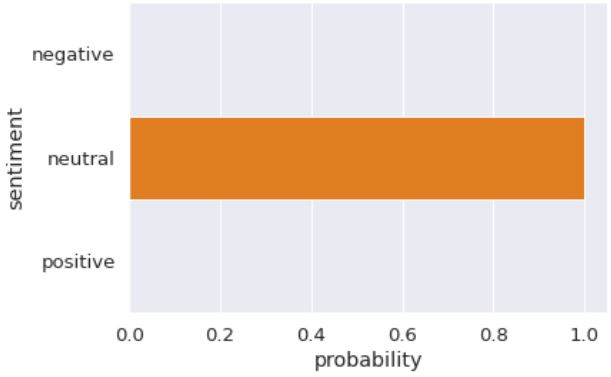
TABEL VIII
CLASSIFICATION REPORT DARI DUA MODEL UNTUK DATA PELABELAN BERBASIS TEXTBLOB

Pre-trained	Batch Size	Precision			Recall			F1-Score			Accuracy
		Pos	Net	Neg	Pos	Net	Neg	Pos	Net	Neg	
BERT$_{BASE}$ Multilingual	16	0.92	0.94	0.92	0.91	0.91	0.95	0.91	0.92	0.93	0.92
	32	0.89	0.94	0.93	0.91	0.91	0.93	0.90	0.93	0.93	0.92
IndoBERT$_{BASE}$	16	0.90	0.93	0.95	0.92	0.91	0.94	0.91	0.92	0.95	0.93
	32	**0.90**	**0.95**	**0.96**	**0.95**	**0.91**	**0.94**	**0.92**	**0.93**	**0.95**	**0.94**

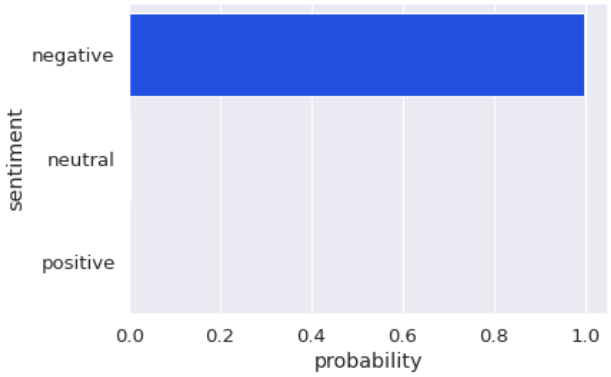
DOI http://dx.doi.org/10.51213/jimp.v7i3.779

TABEL IX
SENTIMENT PREDICTION DARI DUA MODEL UNTUK DATA PELABELAN BERBASIS TEXTBLOB

Review Text	Sentiment Prediction	
	IndoBERT$_{BASE}$	BERT$_{BASE}$ Multilingual
setelah pembaruan terakhir gim ini menjadi sangat buruk Meskipun saya memiliki internet yang layak ini menunjukkan masalah jaringan Selain itu dalam mode solo tidak dapat menemukan saya menunggu selama 3 menit tetapi tidak bisa Saya berharap para pengembang akan melakukan sesuatu tentang masalah ini	Negative	Positive
playernya pubg mobile pakai cheat suntik tindakan game	Negative	Neutral
download habis buka suruh download giga byte sialan	Negative	Neutral



Gambar 15. Nilai Probability BERTBASE Multilingual Review Text pada Tabel IX



Gambar 16. Nilai Probability IndoBERTBASE Review Text pada Tabel IX

Pada Tabel IX. Tiga ulasan yang diberikan sentimen *Positive* & *Neutral* oleh model BERT$_{BASE}$ Multilingual telah diklasifikasikan secara tidak benar sedangkan model IndoBERT$_{BASE}$ yang mengacu pada ulasan *Negative* telah diklasifikasikan dengan benar. sementara BERT memberikan nilai *probabilistic* untuk semua kalimat dari kedua metode tersebut yang ditunjukan pada Gambar 15 dan 16. Bagi manusia mudah untuk memahami komentar negatif, sedangkan BERT$_{BASE}$ Multilingual dapat menginterpretasikan posisi menonjol dari komentar positif atau netral sebagai penanda sentimen yang gagal dalam klasifikasi teks. Secara keseluruhan, mungkin untuk menggeneralisasi bahwa kesalahan yang disebabkan oleh BERT tampaknya terkait dengan keunggulan penanda sentimen berorientasi teks yang terkadang menyesatkan.

IV. KESIMPULAN

Penelitian ini tentang analisis sentimen pada ulasan pengguna aplikasi PUBG Mobile. BERT sebagai model *deep learning* untuk salah satu tugas *natural language processing* yang telah menunjukkan hasil yang optimal seperti analisis sentimen. Melalui *fine-tuning*, model BERT yang dilatih sebelumnya dapat *transfer-learning* yang memberikan lebih banyak *knowledge* untuk model BERT. Metode pelabelan data dapat mempengaruhi hasil akhir. Pelabelan data berbasis *TextBlob* terbukti menunjukkan akurasi yang jauh lebih baik daripada pelabelan berbasis *Score*. Model IndoBERT$_{BASE}$ dapat prediksi sentimen dengan baik. Hasil terbaik diperoleh dengan model IndoBERT$_{BASE}$ dengan akurasi tertinggi sebesar 94%, menggunakan *hyperparameters* yaitu *learning rate* 0.00002, *batch size* 32, jumlah *epoch* 5 dan waktu pelatihan 12 menit.

REFERENSI

[1] Biasramadhan Pandu, "Indonesia Gaming Market: A Window Opportunity for Chinese Developers," *www.cekindo.com*, 2021. https://www.cekindo.com/blog/indonesia-gaming-market (accessed Dec. 16, 2022).

[2] All Correct Games, "The Indonesian Gaming Market | Allcorrect Games," *allcorrectgames.com*, 2022. https://allcorrectgames.com/insights/indonesia/ (accessed Dec. 16, 2022).

[3] M. A. Andrhasa, "Prediksi Esports 2022: Menuju Puncak Gemilang Esports Indonesia?," *nawalakarsa.id*, 2022. https://nawalakarsa.id/game-teknologi/prediksi-esports-2022/ (accessed Dec. 16, 2022).

[4] R. da Silva, J. de Oliveira Liberato Magalhães, I. R. Rodrigues Silva, R. Fagundes, E. Lima, and A. Maciel, "Rating Prediction of Google Play Store apps with application of data mining techniques," *IEEE Lat. Am. Trans.*, vol. 19, no. 01, pp. 26–32, 2021, doi: 10.1109/TLA.2021.9423823.

[5] E. Noei and K. Lyons, "A study of gender in user reviews on the Google Play Store," *Empir. Softw. Eng.*, vol. 27, no. 2, p. 34, 2021, doi: 10.1007/s10664-021-10080-8.

[6] Z. Gao, A. Feng, X. Song, and X. Wu, "Target-dependent sentiment classification with BERT," *IEEE Access*, vol. 7, pp. 154290–154299, 2019, doi: 10.1109/ACCESS.2019.2946594.

[7] L. Yang, Y. Li, J. Wang, and R. S. Sherratt, "Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning," *IEEE Access*, vol. 8, pp. 23522–23530, 2020, doi: 10.1109/ACCESS.2020.2969854.

[8] Pristiyono, M. Ritonga, M. A. Al Ihsan, A. Anjar, and F. H. Rambe, "Sentiment analysis of COVID-19 vaccine in Indonesia using Naïve Bayes Algorithm," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1088, no. 1, p. 12045, Feb. 2021, doi: 10.1088/1757-899X/1088/1/012045.

[9] A. M. Rahat, A. Kahir, and A. K. M. Masum, "Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis

DOI http://dx.doi.org/10.51213/jimp.v7i3.779

Using Review Dataset," in *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*, 2019, pp. 266–270. doi: 10.1109/SMART46866.2019.9117512.
[10] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, no. Mlm, pp. 4171–4186, 2019.
[11] S. González-Carvajal and E. C. Garrido-Merchán, "Comparing BERT against traditional machine learning text classification," May 2020, [Online]. Available: http://arxiv.org/abs/2005.13012
[12] C. A. Putri, "Analisis Sentimen Review Film Berbahasa Inggris Dengan Pendekatan Bidirectional Encoder Representations from Transformers," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 6, no. 2, pp. 181–193, 2020, doi: 10.35957/jatisi.v6i2.206.
[13] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to Fine-Tune BERT for Text Classification?," in *Chinese Computational Linguistics*, 2019, pp. 194–206.
[14] Q. T. Nguyen, T. L. Nguyen, N. H. Luong, and Q. H. Ngo, "Fine-Tuning BERT for Sentiment Analysis of Vietnamese Reviews," in *2020 7th NAFOSTED Conference on Information and Computer Science (NICS)*, 2020, pp. 302–307. doi: 10.1109/NICS51282.2020.9335899.
[15] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding." arXiv, 2020. doi: 10.48550/ARXIV.2009.05387.
[16] C. G. Skarpathiotaki and K. E. Psannis, "Cross-Industry Process Standardization for Text Analytics," *Big Data Res.*, vol. 27, p. 100274, 2022, doi: https://doi.org/10.1016/j.bdr.2021.100274.
[17] M. Singh, A. K. Jakhar, and S. Pandey, "Sentiment analysis on the impact of coronavirus in social life using the BERT model," *Soc. Netw. Anal. Min.*, vol. 11, no. 1, p. 33, 2021, doi: 10.1007/s13278-021-00737-z.
[18] S. El Anigri, M. M. Himmi, and A. Mahmoudi, "How BERT's Dropout Fine-Tuning Affects Text Classification?," in *Business Intelligence*, 2021, pp. 130–139.
[19] R. Qasim, W. H. Bangyal, M. A. Alqarni, and A. Ali Almazroi, "A Fine-Tuned BERT-Based Transfer Learning Approach for Text Classification," *J. Healthc. Eng.*, vol. 2022, p. 3498123, 2022, doi: 10.1155/2022/3498123.

DOI http://dx.doi.org/10.51213/jimp.v7i3.779