

# Deteksi Sarkasme Menggunakan *Bidirectional Encoder Representations from Transformers* (BERT) Pada Teks Bahasa Indonesia

Azka Fauzi Al Parisi<sup>1</sup>, Aradea<sup>2</sup>, Rahmi Nur Shofa<sup>3</sup>

Universitas Siliwangi<sup>1,2,3</sup>

197006057@student.unsil.ac.id<sup>1</sup>, aradea@unsil.ac.id<sup>2</sup>, rahmi.shofa@unsil.ac.id<sup>3</sup>

**Abstrak –** Sarkasme merupakan bahasa figuratif yang banyak digunakan di media sosial. Deteksi sarkasme memiliki kompleksitas tinggi karena memerlukan pemahaman terhadap konteks kata dalam sebuah kalimat. Terdapat celah yang signifikan dalam penelitian deteksi sarkasme berbahasa Indonesia, yaitu keterbatasan dataset sarkasme bahasa Indonesia dan model bahasa yang hanya fokus pada probabilitas kata dasarnya saja tanpa memperhatikan informasi kontekstual yang penting. Tujuan penelitian ini yaitu membangun dataset sarkasme bahasa Indonesia yang dikumpulkan dari berbagai sumber seperti Twitter, Letterboxd, *benchmark* sentimen analisis, serta jurnal linguistik dan sastra Indonesia. Dataset yang terkumpul dilakukan anotasi manual berdasarkan fitur tanda baca dan gaya bahasa. Sedangkan, algoritma yang digunakan sejauh ini adalah algoritma klasifikasi tradisional yang memiliki keterbatasan dalam menangkap nuansa dan kompleksitas sarkasme. Sehingga algoritma yang digunakan pada penelitian ini berfokus memanfaatkan representasi kontekstual dari model BERT, dengan tokenisasi IndoBERT dibantu pustaka PyTorch. Pengukuran evaluasi menggunakan *confusion matrix*. Proses eksperimen model dilakukan sebanyak empat kali dengan fokus *hyperparameter* yang diubah yaitu *dropout* dan *warmup\_steps*. Berdasarkan dari keempat eksperimen yang telah dilakukan, model yang dianggap lebih baik pada penelitian ini yaitu model dari hasil eksperimen ketiga dengan nilai *accuracy* sebesar 84.77%, *precision* sebesar 84.58%, *recall* sebesar 84.97% dan *F1-Score* sebesar 84.77%. Model bahasa yang telah dibuat diberi nama IndoSarcasm.

**Kata Kunci:** *deteksi sarkasme, BERT, IndoBERT, PyTorch, confusion matrix, dropout, num\_warmup\_steps*

## I. PENDAHULUAN

Manusia memiliki pola bahasa yang unik, baik secara lisan maupun tulisan. Keunikannya terdapat pada penggunaan bahasa figuratif atau bahasa kiasan. Penggunaan bahasa kiasan dilakukan sebagai bentuk ekspresi yang sengaja disampaikan dengan menyimpangkan makna literal, sehingga sebuah kata akan memiliki pemaknaan dan penafsiran berbeda [1]. Contoh bahasa kiasan yang sering ditemukan di media sosial adalah penggunaan bahasa sarkasme dan ironi [2]. Perbedaan sarkasme dan ironi terletak pada bunyi atau gaya bahasa. Sarkasme memiliki emosi dan intensitas tinggi, juga gaya bahasa yang eksplisit. Sedangkan ironi memiliki intensitas emosi lebih rendah dengan gaya bahasa yang lebih implisit [3]. Namun, keduanya memiliki kesamaan yaitu dapat menyamarkan pesan negatif, sebab tujuannya adalah untuk mengejek seseorang atau situasi tertentu dengan cara memutarbalikan makna kata yang sebenarnya [4]. Selain itu, keduanya menjadi subjek yang banyak diteliti di bidang *Natural Language Processing* (NLP), terutama dalam sub-bidang *Sentiment Analysis* dan *Sarcasm Detection* [5].

Kepopuleran deteksi sarkasme dipengaruhi oleh jumlah penggunaan media sosial dan perkembangan analisis sentimen terhadap berbagai jenis ulasan, *feedback*, komentar serta postingan lainnya yang beragam dengan tujuan untuk mengetahui sentimen positif maupun negatif sehingga dapat digunakan untuk mengambil keputusan. Namun, yang menjadi permasalahan tidak semua ekspresi positif dimaksudkan positif, begitupun tidak semua ekspresi negatif memiliki konteks yang sepenuhnya negatif [6]. Ekspresi seperti itulah yang disebut sarkasme, sekaligus menjadi tantangan pada perkembangan sentimen analisis atau umumnya pada bidang NLP [7]. Selain itu, deteksi sarkasme memiliki kompleksitas yang tinggi bahkan untuk manusia sekalipun, sebab setiap kata tidak cukup dicari berdasarkan strukturnya saja, tetapi perlu diketahui konteks kata tersebut dalam sebuah kalimat utuh.

Sedangkan konteks kata sangat beragam, tergantung waktu, daerah, sejarah, budaya dan bahasa yang digunakan [8][9].

Terdapat empat pendekatan yang umum dilakukan pada deteksi sarkasme yaitu pendekatan berbasis aturan (*Rules-based Approaches*), pendekatan berbasis pembelajaran mesin tradisional (*Traditional Machine Learning-based Approaches*), pendekatan berbasis deep learning (*Deep Learning-based Approaches*) dan pendekatan berbasis transformer (*Transformer-based Approaches*) [10]. Pada penelitian ini, pendekatan yang dilakukan adalah pendekatan berbasis *transformer*. Hal ini dilakukan karena pendekatan berbasis *transformer* menerapkan *attention* dalam model arsitekturnya, yang memungkinkan kata dipahami secara mendalam. Pada kasus deteksi sarkasme, umumnya menggunakan model *Bidirectional Encoder Representations from Transformers* (BERT) karena memiliki performa rata-rata yang tinggi.

Terdapat celah yang signifikan dalam penelitian deteksi sarkasme berbahasa Indonesia, yaitu penelitian sebelumnya yang hanya fokus pada probabilitas kata dasarnya saja dan tidak memperhatikan informasi kontekstual yang penting. Selain itu, algoritma yang digunakan sejauh ini adalah algoritma klasifikasi sederhana yang memiliki keterbatasan dalam menangkap nuansa dan kompleksitas sarkasme. Sehingga, fokus target perbaikan pada penelitian ini adalah mengembangkan metode deteksi sarkasme dengan memanfaatkan representasi kontekstual dari model BERT. Model BERT yang akan digunakan pada penelitian ini adalah IndoBERT beserta optimalisasi menggunakan AdamW yang diharapkan mampu memberikan pemahaman yang lebih baik tentang konteks dan niat dibalik kalimat sarkastik berbahasa Indonesia. Selain itu, dataset yang dikumpulkan pada penelitian ini diharapkan mampu menjawab keterbatasan yang disebabkan tidak adanya dataset khusus untuk deteksi sarkasme berbahasa Indonesia.

## II. PEKERJAAN TERKAIT

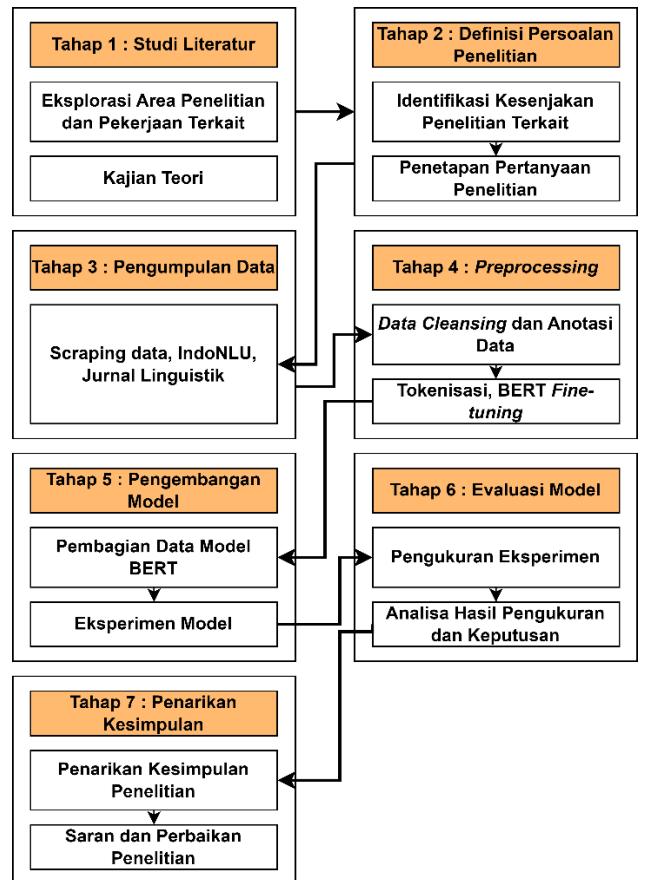
Penelitian yang berfokus untuk mendeteksi kalimat sarkasme dilakukan dengan berbagai macam pendekatan dan algoritma. Secara umum, BERT menghasilkan kinerja lebih baik dibandingkan algoritma klasifikasi teks lainnya [11]–[16]. Adapun penelitian serupa BERT, yaitu RoBERTa memiliki kinerja yang tinggi dalam menangani permasalahan deteksi sarkasme [17]. Bahkan, penelitian lainnya mengkomparasi model *recurrent CNN* RoBERTA

yang disebut RCNN-RoBERTa. Hasilnya lebih tinggi dibandingkan *neural network classifiers* lain [18].

Penelitian deteksi sarkasme berbahasa Indonesia masih terbatas dan jarang dilakukan. Dataset yang digunakan masih terbatas dataset untuk sentimen analisis, belum ada dataset khusus sarkasme bahasa Indonesia [19]–[22]. Proses anotasi data pada penelitian [19] dilakukan dengan memberi label positif dan negatif. Anotasi tersebut kurang tepat, sehingga penelitian lainnya [20]–[23] melakukan anotasi data menjadi label sarkasme dan tidak sarkasme. Selain itu, algoritma yang digunakan masih terbatas pada pendekatan berbasis pembelajaran mesin tradisional [19]–[22]. Meskipun demikian, terdapat penelitian yang menerapkan algoritma *neural network* yaitu BiLSTM untuk mendeteksi sarkasme pada teks bahasa Indonesia, hasilnya masih sebanding dengan penelitian yang menggunakan algoritma klasifikasi sederhana [23].

## III. METODOLOGI PENELITIAN

Penelitian menggunakan pendekatan eksperimen, algoritma berbasis *transformers* dengan model BERT *classifier*, tepatnya IndoBERT pada kalimat bahasa Indonesia. Detail tahapan penelitian tertera pada Gambar 1.



Gambar 1. Tahapan Penelitian

**Studi Literatur** meliputi pengumpulan konsep, teori, metode, algoritma serta data-data dari berbagai sumber seperti jurnal, buku dan lain-lain yang berhubungan dengan penelitian.

**Perancangan Persoalan Penelitian** dilakukan dengan mengidentifikasi kesenjangan, kekurangan dan saran dari penelitian-penelitian yang pernah dilakukan sebelumnya sehingga dapat dilakukan perbaikan. Setelah itu, dilakukan perancangan dengan tujuan yang jelas dengan menetapkan pertanyaan penelitian, batasan penelitian dan manfaat penelitian.

**Pengumpulan Data** bertujuan untuk mengumpulkan data dari berbagai sumber yaitu *scrapping* Twitter dan Letterboxd, *benchmark* IndoNLU serta jurnal-jurnal bahasa dan sastra Indonesia.

**Preprocessing** meliputi *data cleansing*, anotasi data menggunakan ekstraksi fitur tanda baca dan fitur gaya bahasa. Kemudian, dilakukan tokenisasi, konversi token menjadi ID, *padding* dan *truncation* serta pembuatan *attention mask* untuk model BERT.

**Pengembangan Model** meliputi pembagian data menjadi data *train*, *test* dan *validation*. Pada tahapan ini dibuat beberapa rancangan eksperimen model berdasarkan *hyperparameter* yang ditetapkan. Model ini dibantu dengan *Google Colaboratory* dan *Pytorch*.

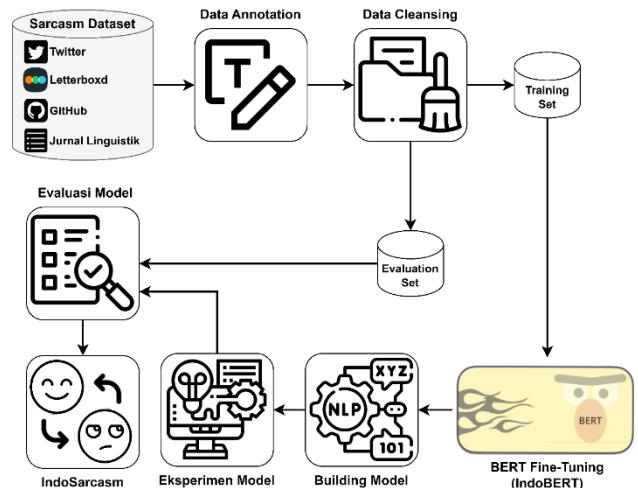
**Evaluasi Model** bertujuan untuk mengukur hasil akhir model menggunakan *matrix* dengan parameter performa yang diukur adalah *accuracy*, *precision*, *recall* dan *F1 score*. Evaluasi bertujuan untuk membandingkan hasil beberapa eksperimen dengan nilai *hyperparameter* yang berbeda.

**Penarikan Kesimpulan** untuk memberikan gambaran umum, kelebihan, kekurangan terhadap hasil evaluasi modul yang mencakup keseluruhan penelitian.

## IV. HASIL DAN PEMBAHASAN

### A. Arsitektur Sistem

Pada penelitian ini, arsitektur sistem atau kerangka kerja yang berfungsi untuk memberikan informasi umum secara terstruktur terkait penelitian yang dilakukan. Data annotation, data cleansing dan BERT Fine-Tuning merupakan bagian dari tahapan preprocessing yang dilanjutkan ke tahap pengembangan model, eksperimen model dan evaluasi model. Arsitektur sistem ini dibuat untuk menghasilkan model akhir yang dinamakan sebagai IndoSarcasm.



Gambar 2. Arsitektur Sistem IndoSarcasm

### B. Datasets

Datasets yang digunakan merupakan hasil gabungan dari data-data yang diambil dari Twitter, Letterboxd, *benchmark* IndoNLU [24] serta jurnal-jurnal bahasa dan sastra Indonesia [25] [26] [27] [28] [29] [30] [31] [32] [33] [34] [35] [36] [37] [38]. Datasets berjumlah 4400 data dengan proporsi seimbang, 2200 berlabel sarkas dan 2200 berlabel tidak sarkas. Labelisasi datasets dibuat seimbang agar pelatihan dan evaluasi model dapat dilakukan lebih baik dan lebih objektif.

### C. Data Annotation

Anotasi data dilakukan secara manual dengan mempertimbangkan ciri-ciri bahasa dan makna yang dapat menentukan suatu teks sarkastik. Anotasi data dibagi menjadi sarkasme (1) dan tidak sarkasme (0) Adapun dua fitur yang digunakan yaitu fitur *punctuation-related* (tanda baca) dan fitur gaya bahasa. Berikut adalah penjelasan untuk kedua fitur tersebut beserta contoh data yang mengilustrasikannya :

#### 1. Punctuation-related

Kriteria tanda baca yang dipakai pada penelitian untuk mengidentifikasi sarkasme, diantaranya sebagai berikut [39] :

- Jumlah tanda seru (!). Penggunaan tanda seru tiga atau lebih pada sebuah kalimat dapat mencerminkan nada tertentu sehingga dimasukan labelisasi sebagai sarkasme.
- Jumlah huruf vokal atau konsonan yang berlebihan. Kalimat yang sengaja dilebihkan dengan menambahkan huruf tersebut untuk mempertegas emosi.

Berikut Tabel 1 yang merupakan contoh dataset yang dilabeli sarkas beserta penjelasannya.

Tabel 1. Teks Sarkasme *Punctuation-related*

No	Teks Sarkasme	Penjelasan
1.	Manusia yg paling suci di negri ini,di arab sana tukang jaga WC umum pun pakai sorban santai aja mereka !!!	Selain konteks kalimatnya bernada sarkasme. Tanda seru pada teks tersebut mempertegas emosi penulis, teks tersebut memiliki maksud kekesalan dan kemarahan.
2.	Awal nonton bareng temen: "AWKOKWOKWOKWOKWOK OWKOKWOK" lama-kelamaan: "kapan selesainya ini film wooooiiiiii plisssssss selesai plisssssss"	"AWKOKWOKWOKWOKWOK WK" yang tertulis di awal teks memberikan informasi bahwa tontonan tersebut menyenangkan. Namun, kalimat seterusnya menggugurkan informasi itu. Terlebih "woooooiiiiii plisssssss selesai plisssssss", merupakan salah satu ciri-ciri sarkasme dengan contoh jumlah huruf yang dilebih-lebihkan.
3.	wkwkwkw...anak TK masih mendengar pohon mahoni di relokasi dan di krantina.: fix ya ini orang kok goblooog ya naudzu bilaah tenan	Kata "goblooog" pada kalimat tersebut, selain maknanya yang negatif dan menghina. Penambahan huruf "o" mempertegas emosi merendahkan sehingga memberikan nuansa sarkasme.

## **2. Gaya Bahasa**

Gaya bahasa sarkasme memiliki ciri khusus yaitu memprovokasi pendengar atau pembaca untuk mengikuti komentar yang dituliskan. Ciri kalimatnya, yaitu terdapat celaan, perkataan kasar, humor menyinggung dan kemarahan [33].

Berikut Tabel 2 yang merupakan contoh dataset yang dilabeli sarkas beserta penjelasannya.

Tabel 2. Teks Sarkasme Gaya Bahasa

No	Teks Sarkasme	Penjelasan
1.	Dasar muka tembok gak punya malu	Kalimat ini termasuk sarkasme karena terkandung gaya bahasa sindiran yang pernyataan yang sifatnya menyamakan dengan benda lain. Wajah manusia disepertikan dengan tembok karena tidak punya rasa malu.
2.	Masyaallah selain cantik dia juga pinter maen piano dan lebih pinter lagi menangin hati laki yang punya istri	Kalimat ini termasuk sarkasme karena menampilkan hinaan yang disembunyikan dengan puji-pujian. Kata “cantik” dan “pintar” merupakan puji-pujian baik, sedangkan “pintar menangin hati laki yang punya istri” merupakan perbuatan negatif yang sifatnya menyindir.
3.	Anjir barusan gue nonton apaan?! Yang gue ingat cuma Tokopedia 😊	“Anjir barusan gue nonton apaan?!“ merupakan ungkapan sarkasme. Karena selain kata yang kasar, terdapat ekspresi ketidakpercayaan terhadap apa yang sudah ditonton. Dilanjutkan dengan kalimat “Yang gue ingat cuma Tokopedia 😊”.

Memperkuat kesan sarkasme bahwa yang diingat bukan konten atau tontonan aslinya tetapi iklan Tokopedia.

#### D. *Data Cleansing*

Pada penelitian ini, pembersihan data berfokus pada beberapa hal, diantaranya sebagai berikut:

- a. Mengubah semua karakter teks menjadi huruf kecil.
  - b. Menghapus spasi yang mungkin ada di awal atau akhir teks.
  - c. Menghapus URL atau tautan yang ada dalam teks seperti "http://", "https://" dan "www.".
  - d. Menghapus angka dalam teks.
  - e. Menghapus semua karakter *non-alphanumeric* kecuali spasi dan tanda seru (!). Tanda seru tidak dihapus karena dapat memberikan informasi tentang sarkasme atau perubahan emosi pada teks. Selain itu, dalam kehidupan nyata penggunaan tanda seru pada teks bertujuan untuk mengekspresikan perasaan tertentu. Jika tanda seru dihapus, maka akan menyebabkan kehilangan representasi sebenarnya dari teks yang mencerminkan perasaan atau ekspresi penulisnya.

Berikut contoh hasil teks sebelum dan sesudah dilakukan pembersihan data tertera pada Tabel 3.

Tabel 3. Hasil Pembersihan Data

No	Teks Sebelum Dibersihkan	Teks Setelah Dibersihkan
1.	tingkat laughable-nya melebihi Bisikan Jenazah ternyata 🎉	tingkat laughablenya melebihi bisikan jenazah ternyata
2.	ASHIAPPPP THE REAL CINEMA MASTERPIECE!!!! Ni film harusnya masuk TOP 1 IMDB dan menang Oscar sihh. The BEST superhero movie ever ASHIAP MAN >>>> Marvel, DC	ashiapppp the real cinema masterpiece!!!! ni film harusnya masuk top imbd dan menang oscar sihh the best superhero movie ever ashiap man marvel dc
3.	<a href="https://youtu.be/Ki7LogqUMKI">https://youtu.be/Ki7LogqUMKI</a> gelar wan bacot buat gabener rasanya kurang menghargai keunggulan beliau	gelar wan bacot buat gabener rasanya kurang menghargai keunggulan beliau

### E. Pembagian Data

Dataset dapat dibagi menjadi dua bagian yaitu data pelatihan dan data validasi [40] [41]. Selain itu, penelitian lain membaginya menjadi tiga bagian yaitu data pelatihan, data pengujian dan data validasi [12] [14] [42] [43]. Penelitian ini membagi data menjadi tiga bagian yang bertujuan untuk mengevaluasi kinerja model dan kemampuan algoritma dalam mempelajari data.

## F. BERT Fine-Tuning

Model BERT yang digunakan yaitu BertForSequenceClassification dengan tokenisasi dari *benchmark* IndoBERT. Implementasi model BERT disertai dengan penerapan Adam *optimizer*.

### 1. Tokenisasi

Tokenisasi yang digunakan adalah IndoBERT yang merupakan model BERT yang sebelumnya sudah dilatih pada dataset berbahasa Indonesia. Pada penelitian ini, tokenisasi IndoBERT merupakan model *indobenchmark* pada HunggingFace [44].

Token dikonversi menjadi ID, lalu, menyematkan token khusus pada teks input untuk IndoBERT. Token-token tersebut meliputi token [CLS] yang menandakan awal kalimat dan token [SEP]. Setelah itu, dilakukan *padding* dan *truncation* agar semua teks dalam satu batch memiliki panjang yang sama dengan cara menambahkan token dan memotongnya.

### 2. Inisialisasi Model dan Optimizer

Nilai *learning rate* yang digunakan adalah 2e-5 sesuai dengan rekomendasi yang terdapat dalam penelitian asli mengenai BERT [45]. Epsilon yang dipilih yaitu 1e-8. Sedangkan *hyperparameter* model yang dirubah berfokus pada nilai *dropout* dan *warmup steps* karena hasil akhir dari pelatihan pertama menunjukan kecenderungan *overfitting*. Nilai *dropout* yang ditetapkan pada penelitian ini yaitu 0.5 untuk menghasilkan nilai regularisasi maksimum [46].

## G. Pengembangan Model

Pengembangan model dilakukan menggunakan bantuan *google colaboratory pro* dan beberapa *library* untuk membantu pengembangannya. *Library* yang digunakan untuk proses pemodelan tertera pada Tabel 4.

Tabel 4. Library yang digunakan

No	Library	Penggunaan
1.	Pytorch	Pembuatan DataLoader
2.	Transformers	Pemanggilan model BERT
3.	Random, time, datetime	Persiapan eksperimen

Sedangkan *hardware google colaboratory pro* yang digunakan pada tahap pengembangan model dan perancangan eksperimen tertera pada Tabel 5.

Tabel 5. Spesifikasi *google colaborator pro*

CPU	Intel(R) Xeon(R) CPU @ 2.30GHz
GPU	Tesla T4
RAM	12 GB
Storage	141 GB available

Pada pengembangan model dibuat *DataLoader* untuk mengubah data menjadi tensor menggunakan modul PyTorch yang difokuskan untuk untuk mengelola ukuran *batch* data. Penelitian ini memilih ukuran batch 16 dan pengujian data validasi menggunakan *RandomSampler* karena dataset pada kasus klasifikasi teks tidak memerlukan urutan waktu dan struktur dari setiap data yang ada.

## H. Perancangan Eksperimen Model

Eksperimen pada penelitian ini akan dilakukan dengan tujuan untuk menguji *hyperparameter* model BERT terbaik dan membandingkan setiap model dalam mendekripsi kalimat sarkasme bahasa Indonesia. Fokus eksperimen ini adalah untuk mengukur *accuracy* dan *loss* saat proses pelatihan serta evaluasi akhir berupa matriks dengan performa yang diukur yaitu *accuracy*, *precision*, *recall* dan *F1-score*. Berikut tujuan dan indikator untuk membandingkan dan membuktikan variasi model dengan hasil terbaik, tertera pada Tabel 6.

Tabel 6. Library yang digunakan

No	Tujuan	Indikator
1.	Mengukur kinerja model dalam melakukan generalisasi data	<i>Accuracy Loss</i> , <i>Validation Loss</i>
2.	Mengkonstruksi <i>hyperparameter</i> model dengan matriks performa tinggi.	<i>Accuracy</i> , <i>Precision</i> , <i>Recall</i> dan <i>F1-Score</i>

Tujuan tersebut dapat dicapai dengan melakukan beberapa eksperimen menggunakan *hyperparameter* yang berbeda untuk mencapai tujuan yang sudah ditentukan. Eksperimen dan detail perbedaan *hyperparameter* yang digunakan tertera pada Tabel 7.

Tabel 7. Skenario Eksperimen

Eksperimen	Model dan Hyperparameter
Pertama	<i>epoch</i> = 3 <i>output attentions</i> = False, <i>output hidden states</i> = False, <i>learning rate</i> = 2e-5, <i>epsilon</i> = 1e-8, <i>num warmup steps</i> = 0
Kedua	<i>epoch</i> = 3 <i>output attentions</i> = False, <i>output hidden states</i> = False, <i>attention probs dropout prob</i> = 0.5, <i>hidden dropout prob</i> = 0.5, <i>learning rate</i> = 2e-5, <i>epsilon</i> = 1e-8, <i>num warmup steps</i> = 0
Ketiga	<i>epoch</i> = 3 <i>output attentions</i> = False, <i>output hidden states</i> = False,

	<i>attention probs dropout prob = 0.4, hidden dropout prob = 0.4, learning rate = 2e-5, epsilon = 1e-8, num warmup steps = 0</i>
Keempat	<i>epoch = 3 output attentions = False, output hidden states = False, attention probs dropout prob = 0.4, hidden dropout prob = 0.4, learning rate = 2e-5, epsilon = 1e-8, num warmup steps = 150</i>

## I. Evaluasi Model

### 1. Perbandingan Model

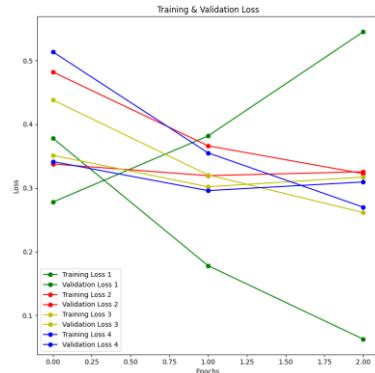
Eksperimen bertujuan untuk mencari dan membandingkan performa proses pelatihan dan evaluasi akhir, parameter yang dibandingkan yaitu *training accuracy* dengan *validation accuracy* dan *training loss* dengan *validation loss*. Sedangkan pada evaluasi akhir menggunakan *confusion matrix*, parameter yang dibandingkan yaitu *accuracy*, *precision*, *recall* dan *F1-Score*.

Hasil evaluasi pada eksperimen pertama menunjukkan nilai *accuracy* pelatihan dan *accuracy* akhir yang tinggi sebesar 86.14%. Pada proses pelatihannya model cenderung mengalami *overfitting*. Sehingga eksperimen kedua ini menambahkan nilai *dropout*, yaitu teknik yang umum dipakai untuk mencegah *overfitting*. Nilai *dropout* yang diterapkan yaitu 0.5. Pada eksperimen ini nilai *dropout* diterapkan pada lapisan tersembunyi (*hidden layers*) dan nilai *attention*.

Berdasarkan eksperimen kedua, indikasi *overfitting* dapat teratas. Namun, permasalahan lain yaitu akurasi akhir dan sensitivitas model terhadap kalimat sarkasme terlalu tinggi dibandingkan kalimat non-sarkasme sehingga memungkinkan kalimat sederhana diprediksi sebagai sarkasme. Hasil ini dapat ditemukan pada *accuracy* akhir sebesar 82.84%. Upaya untuk mengatasi hal tersebut yakni pada eksperimen ketiga nilai *dropout* diubah nilainya yang semula 0.5 menjadi 0.4. Meskipun pada eksperimen ketiga, model sudah cukup baik dengan nilai *accuracy* akhir sebesar 84.77%, eksperimen keempat pada model dilakukan untuk menemukan pengetahuan baru, dengan mengubah nilai *num\_warmup\_steps* yang semula bernilai 0 menjadi 150. Nilai 150 ini dihitung 10% dari total langkah, kemudian dibulatkan berdasarkan jumlah langkah pada pelatihan.

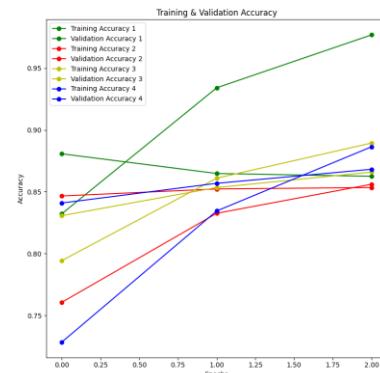
Detail perubahan nilai *training loss* dan *validation loss* pada semua eksperimen dapat dilihat

lebih jelas menggunakan diagram garis seperti tertera pada Gambar 3.



Gambar 3. *Training loss* dan *validation loss*

Sedangkan detail perubahan nilai *training accuracy* dan *validation accuracy* pada semua eksperimen dapat dilihat lebih jelas menggunakan diagram garis seperti tertera pada Gambar 4.



Gambar 4. *Training accuracy* dan *validation accuracy*

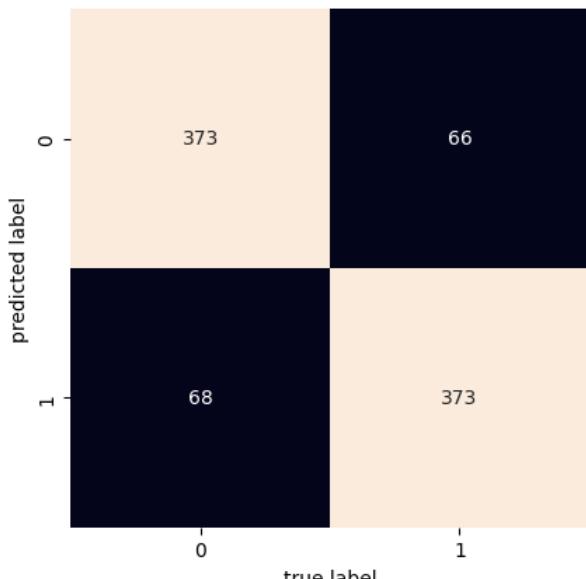
Berdasarkan dari keempat eksperimen yang telah dilakukan, model yang dianggap lebih baik pada penelitian ini yaitu model dari hasil eksperimen ketiga, meskipun akurasi akhirnya sama dengan model keempat yaitu sekitar 84.77%. Model ketiga ini memiliki nilai *F1-Score* yang seimbang untuk label 0 maupun label 1. Secara detail, perbedaan matriks hasil akhir setiap eksperimen diketahui seperti tertera pada Tabel 8.

Tabel 8. Hasil Perbandingan Nilai Matriks Setiap Eksperimen

Eksperimen	Accuracy	Precision	Recall	F1 - Score
1	86.14%	85.30%	87.24%	86.26%
2	82.84%	81.30%	85.19%	83.20%
3	84.77%	84.58%	84.97%	84.77%
4	84.77%	84.12%	85.65%	84.88%

## 2. Analisis Model IndoSarcasm

Berdasarkan hal tersebut, pembuktian model yang dipakai yaitu model IndoSarcasm hasil prediksi pada eksperimen ketiga. Secara detail *confusion matrix* eksperimen ketiga tertera pada Gambar 7.



Gambar 7. *Confusion Matrix* Eksperimen Ketiga

Evaluasi kinerja model BERT untuk kasus deteksi sarkasme yaitu menggunakan *confusion matrix*. Pada penelitian ini label berjumlah dua, yaitu 1 (sarkasme) dan 0 (non-sarkasme). Hasil *confusion matrix* eksperimen pertama yaitu *True Positive* (TP) sebanyak 373 data, *False Positive* (FP) sebanyak 66 data, *True Negative* (TN) sebanyak 373 data dan *False Negative* (FN) sebanyak 68 data.

Contoh kalimat atau teks yang hasilnya *True Positive* (TP) atau diprediksi benar sebagai kalimat sarkasme. Dari data tersebut dapat dipahami bahwa model memahami bentuk sarkasme sesuai dengan tanda seru yang belebihan, kalimat dengan nada mengejek dan komentar dengan gaya hiperbola. Detail teks tertera pada Tabel 13.

Tabel 13. Hasil Data *True Positive* (TP)

Teks	Label	Prediksi
saatnya menggatot!!!!	1	1
yap! saat ini aku tidak punya tenaga untuk berpurapura suka	1	1
alam berubah menjadi bencana atas nama investasi dikeruk diperkosa	1	1

*True Negative* (TN) atau teks non-sarkasme (0) yang diprediksi benar sebagai non-sarkasme tersebut dapat dipahami bahwa model memahami bentuk non-sarkasme dengan baik. Detail teks tertera pada Tabel 14.

Tabel 14. Hasil Data *True Negative* (TN)

Teks	Label	Prediksi
bukti jokowi orang nya jauh lebih rendah hati rumah makan nya enak masakan nya lengkap dan bervariasi harga nya lumayan dan sesuai dengan harga yang ditawarkan cocok untuk makan dengan keluarga yang ingin makan hidangan jawa barat	0	0
mohon ditindak lanjuti	0	0

Sedangkan data yang termasuk *False Positive* (FP) atau teks non-sarkasme yang diprediksi salah sebagai kalimat sarkasme tersebut dapat dipahami bahwa model sangat sentitif dalam memahami sarkasme. Hasilnya ini dapat memberikan dua perspektif untuk perbaikan, yaitu model salah dalam melakukan prediksi atau model lebih sensitif terhadap teks sarkasme sehingga kedepannya dataset dapat diperbaiki dan dilabelisasi ulang. Detail teks tertera pada Tabel 15.

Tabel 15. Hasil Data *False Positive* (FP)

Teks	Label	Prediksi
kalaup sampai ada garansi segala berarti capres yg diusung tidak meyakinkan dan meragukan kualitas serta integritasnya	0	1
paling suka sama orang yang tidak pernah mengeluh kalau lagi kerja	0	1
cerita si kancil itu identik dengan kecerdikan seorang kancil menyelesaikan masalah yang ada di hadapan nya	0	1

Terakhir, data yang termasuk *False Negative* (FN) atau teks sarkasme yang diprediksi salah, sehingga model memprediksinya menjadi sebagai kalimat non-sarkasme tersebut dapat dipahami bahwa model cenderung tidak memahami teks sarkasme yang diselipkan pada kalimat yang panjang, sehingga model memprediksinya sebagai non-sarkasme. Hasilnya ini dapat memberikan dua perspektif untuk perbaikan, yaitu dataset dapat diperbaiki dengan mempersingkat atau melabelisasi ulang jika ada kesalahan dalam proses labelisasi sebelumnya. Kedua, model dapat ditingkatkan agar mengetahui bentuk tertentu dari kalimat sarkasme yang disematkan pada teks panjang. Detail teks tertera pada Tabel 16.

Tabel 16. Hasil Data *False Negative* (FN)

Teks	Label	Prediksi
yang saya rispek beliau adalah suami yang sangat sayang isterinya meskipun beliau lupa tanggal pernikahannya makanya sekarang saya paham kenapa cak imin cocok banget jadi pejabat karena sangat sayang dengan rakyat tapi lupa dengan janjianjinya	1	0
ternyata ini makna posisi demokrat yang selalu netral pasti mereka maksud tertentu ini diamdiam netral habis itu ketika ada	1	0

kesempatan langsung deh pak presiden di jatuhkan	
bintang lima aku kasih buat diriku sendiri karena bangga bgt sama diri sendiri yang udah berhasil nonton film ini sampe abis tanpa skipskip tanpa playback speed dan ga bajakan padahal nontonnya gaperlu mikir tapi pusing bgt anjirit bagus banget filmnya aldi taher	1 0

## J. Ancaman Terhadap Validitas

### 1. Ancaman Internal

Eksperimen yang dilakukan terbatas pada perubahan beberapa *hyperparameter* saja. Penelitian ini hanya menambahkan *dropout* dan *num\_warmup\_steps*, merubahnya serta mengujinya dengan dua opsi nilai. Perubahan yang hanya fokus pada dua *hyperparameter* tersebut tidak dapat menggambarkan kemampuan model secara komprehensif apabila terdapat tujuan yang lebih khusus. Selain itu, matriks evaluasi yang digunakan terbatas menggunakan *confusion matrix* sehingga pada penelitian selanjutnya matriks evaluasi dapat dikembangkan lebih lanjut seperti menggunakan evaluasi *Matthews Correlation Coefficient* (MCC). Keterbatasan lainnya terdapat pada dataset. Karena Dataset yang digunakan merupakan dataset baru hasil *scraping* dan pencarian dari berbagai sumber. *Platform* sumber dataset masih terbatas pada Twitter dan Letterboxd sedangkan sarkasme memerlukan keberagaman dari berbagai *platform* media sosial yang ada.

### 2. Ancaman Konstruksi

Proses anotasi data secara manual memungkinkan terjadi kesalahan saat pelabelan, meskipun proses labelisasi dilakukan secara hati-hati berdasarkan referensi tanda baca dan gaya bahasa sarkasme berdasarkan berbagai bentuk penyampaiannya. Sedikit kesalahan dalam melakukan labelisasi akan mempengaruhi validitas dan kualitas dataset sehingga dapat menggugurkan sarkasme secara definitif dan teoritis serta mengancam validitas performa model yang sudah dibuat. Selain itu, labelisasi fokus terhadap bentuk teks dalam kalimat utuh, yang pada kasus nyata di lapangan, bentuk teks tidak selalu menggambarkan sarkasme, bahkan teks yang netral dapat berarti sarkasme, tergantung subjek yang menulis dan objek yang dituju.

### 3. Ancaman Konklusi

Pemilihan model ketiga sebagai model terbaik dengan akurasi sekitar 84.77% dilakukan secara hati-hati. Namun, uji model dengan memasukan kalimat sederhana untuk memprediksi kalimat tersebut merupakan contoh yang diasumsikan. Sehingga perlu dibuktikan secara mendalam. Seperti pada tahap pengujian dengan melakukan input kalimat

menggunakan model tersebut perlu mempertimbangkan keragaman bentuk kalimat sarkasme untuk menguji model secara objektif dan memungkinkan ditemukan berbagai kelemahan model terhadap bentuk kalimat sarkasme tertentu.

## 4. Ancaman Eskternal

Eksperimen berjalan pada *virtual machine* yang digunakan di *google colab*. *Google colab* yang digunakan merupakan *google colab pro* dengan GPU T4. Meskipun begitu, pengukuran terhadap eksperimen dan data akan bervariasi jika dipraktekan dan diuji coba pada mesin lainnya. Selain itu, *google colab* menggunakan sistem *sharing hardware* dimana satu perangkat keras akan digunakan untuk beberapa *virtual machine* sehingga bisa jadi hal ini akan mempengaruhi kinerja. Dalam hal ini penelitian selanjutnya disarankan untuk dapat menggunakan *virtual machine* terdedikasi dengan spesifikasi yang lebih tinggi sehingga performa dan pengukuran dapat dilakukan dengan lebih tepat.

## V. KESIMPULAN

Kebaharuan penelitian ini berfokus pada dataset yang secara khusus dirancang untuk kasus sarkasme berbahasa Indonesia. Dataset berjumlah total 4400 data yang bersumber dari Twitter, Letterboxd, *benchmark* IndoNLU bagian sentimen analisis yang dilabelisai ulang dan jurnal-jurnal linguistik serta sastra dengan topik pembahasannya yaitu analisa komentar sarkasme di media sosial.

Model IndoSarcasm yang dirancang menggunakan BERT, utamanya IndoBERT membuktikan dengan baik bahwa IndoSarcasm dapat mendeteksi kalimat sarkasme bahasa Indonesia. Hasil eksperimen membuktikan performa model IndoSarcasm memiliki akurasi sebesar 88.77% dengan nilai parameter lain yang relatif seimbang yaitu *precision* sebesar 84.58%, *recall* sebesar 84.97% dan *F1-Score* sebesar 84.77%. Hal ini diperkuat dengan hasil eksperimen selama pelatihan, yaitu tidak ditemukannya indikasi *overfitting*.

*Hyperparameter* pada penelitian ini terbatas pada menambahkan *dropout* dan *num\_warmupsteps*, merubahnya serta mengujinya dengan dua opsi nilai. Sehingga pada penelitian di masa depan pelatihan dapat diuji dengan mengganti berbagai macam *hyperparameter* dan membandingkannya secara komprehensif. Matriks evaluasi yang digunakan-pun terbatas menggunakan *confusion matrix* sehingga pada penelitian di masa depan matriks evaluasi dapat dikembangkan lebih lanjut menggunakan evaluasi *Matthews Correlation Coefficient* (MCC) atau matriks

evaluasi lainnya. Selain itu, model dan pendekatakan deteksi sarkasme dapat dikembangkan dengan menerapkan model lain, seperti RoBERTa, T5 (*Text-to-Text Transfer Transformer*), XLNet atau pendekatakan yang bukan hanya mempertimbangkan teks atau konten komentarnya saja tetapi pendekatakan berdasarkan akun.

## REFERENSI

- [1] R. W. Gibbs dan H. L. Colston, “Chapter 21 Figurative Language,” *Language (Baltim)*, , 2006.
- [2] M. del Pilar Salas-Zárate, G. Alor-Hernández, J. L. Sánchez-Cervantes, M. A. Paredes-Valverde, J. L. García-Alcaraz, dan R. Valencia-García, “Review of English literature on figurative language applied to social networks,” *Knowl. Inf. Syst.*, vol. 62, no. 6, 2020, doi: 10.1007/s10115-019-01425-3.
- [3] E. Sulis, D. Irazú Hernández Farías, P. Rosso, V. Patti, dan G. Ruffo, “Figurative messages and affect in Twitter: Differences between #irony, #sarcasm and #not,” *Knowledge-Based Syst.*, vol. 108, 2016, doi: 10.1016/j.knosys.2016.05.035.
- [4] S. Frenda, A. T. Cignarella, V. Basile, C. Bosco, V. Patti, dan P. Rosso, “The unbearable hurtfulness of sarcasm,” *Expert Syst. Appl.*, vol. 193, 2022, doi: 10.1016/j.eswa.2021.116398.
- [5] A. C. Băroiu dan Ştefan Trăuşan-Matu, “Automatic Sarcasm Detection: Systematic Literature Review,” *Inf.*, vol. 13, no. 8, hal. 1–17, 2022, doi: 10.3390/info13080399.
- [6] D. Maynard dan M. A. Greenwood, “Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis,” 2014.
- [7] M. Sykora, S. Elayan, dan T. W. Jackson, “A qualitative analysis of sarcasm, irony and related hashtags on Twitter,” *Big Data Soc.*, vol. 7, no. 2, 2020, doi: 10.1177/2053951720972735.
- [8] D. Jain, A. Kumar, dan G. Garg, “Sarcasm detection in mash-up language using soft-attention based bi-directional LSTM and feature-rich CNN,” *Appl. Soft Comput. J.*, vol. 91, 2020, doi: 10.1016/j.asoc.2020.106198.
- [9] L. Ren, B. Xu, H. Lin, X. Liu, dan L. Yang, “Sarcasm Detection with Sentiment Semantics Enhanced Multi-level Memory Network,” *Neurocomputing*, vol. 401, 2020, doi: 10.1016/j.neucom.2020.03.081.
- [10] F. B. Kader, N. H. Nujat, T. B. Sogir, M. Kabir, H. Mahmud, dan K. Hasan, “Computational Sarcasm Analysis on Social Media: A Systematic Review,” 2022, [Daring]. Tersedia pada: <http://arxiv.org/abs/2209.06170>.
- [11] B. Shmueli, L. W. Ku, dan S. Ray, “Reactive supervision: A new method for collecting sarcasm data,” 2020, doi: 10.18653/v1/2020.emnlp-main.201.
- [12] C. I. Eke, A. A. Norman, dan L. Shuib, “Context-Based Feature Technique for Sarcasm Identification in Benchmark Datasets Using Deep Learning and BERT Model,” *IEEE Access*, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3068323.
- [13] A. Abaskohi, A. Rasouli, T. Zeraati, dan B. Baharak, “UTNL at SemEval-2022 Task 6: A Comparative Analysis of Sarcasm Detection Using Generative-based and Mutation-based Data Augmentation,” *SemEval 2022 - 16th Int. Work. Semant. Eval. Proc. Work.*, hal. 962–969, 2022, doi: 10.18653/v1/2022.semeval-1.135.
- [14] E. Savini dan C. Caragea, “Intermediate-Task Transfer Learning with BERT for Sarcasm Detection,” *Mathematics*, vol. 10, no. 5, 2022, doi: 10.3390/math10050844.
- [15] A. Baruah, K. Das, F. Barbhuiya, dan K. Dey, “Context-Aware Sarcasm Detection Using BERT,” 2020, doi: 10.18653/v1/2020.figlang-1.12.
- [16] A. Benlahbib, H. Alami, dan A. Alami, “LISACTeam at SemEval-2022 Task 6: A Transformer based Approach for Intended Sarcasm Detection in English Tweets,” 2022, doi: 10.18653/v1/2022.semeval-1.139.
- [17] A. K. Jayaraman, T. E. Trueman, G. Ananthakrishnan, S. Mitra, Q. Liu, dan E. Cambria, “Sarcasm Detection in News Headlines using Supervised Learning,” 2022, doi: 10.1109/AIDE57180.2022.10060855.
- [18] R. A. Potamias, G. Siolas, dan A. G. Stafylopatis, “A transformer-based approach to irony and sarcasm detection,” *Neural Comput. Appl.*, vol. 32, no. 23, 2020, doi: 10.1007/s00521-020-05102-3.
- [19] L. Septiani dan Y. Sibaroni, “Sentiment Analysis Terhadap Tweet Bernada Sarkasme Berbahasa Indonesia,” *J. Linguist. Komputasional*, vol. 2, no. 2, hal. 62–67, 2019.
- [20] D. Alita dan A. R. Isnain, “Pendeteksian Sarkasme pada Proses Analisis Sentimen Menggunakan Random Forest Classifier,” *J. Komputasi*, vol. 8, no. 2, hal. 50–58, 2020, doi: 10.23960/komputasi.v8i2.2615.
- [21] Y. Yunitasari, A. Musdholifah, dan A. K. Sari, “Sarcasm Detection For Sentiment Analysis in Indonesian Tweets,” *IJCCS (Indonesian J. Comput. Cybern. Syst.)*, vol. 13, no. 1, hal. 53, 2019, doi: 10.22146/ijccs.41136.
- [22] D. Alita dan Y. Fernando, “Multiclass SVM Algorithm for Sarcasm Text in Twitter,” *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 8, no. 1, hal. 118–128, 2021, doi:

- 10.35957/jatisi.v8i1.646.
- [23] Y. V. Aritonang, D. P. Napitupulu, M. H. Sinaga, dan J. Amalia, "Pengaruh Hyperparameter pada Fasttext terhadap Performa Model Deteksi Sarkasme Berbasis Bi-LSTM," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 9, no. 3, hal. 2612–2625, 2022, doi: 10.35957/jatisi.v9i3.1331.
- [24] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," 2020, [Daring]. Tersedia pada: <http://arxiv.org/abs/2009.05387>.
- [25] Sarli, Nurhadi, dan E. S. Sari, "Analisis Penggunaan Gaya Bahasa Sarkasme Netizen Di Media Sosial Tiktok," vol. 3, no. Maret, hal. 84–92, 2023.
- [26] N. A. Budi, S. A. Aziz, dan S. S. Rimang, "Gaya Bahasa Sindiran Pada Media Sosial," *J. Sinestesia*, vol. 13, no. 1, hal. 163–174, 2023.
- [27] M. Afrodita, D. Ismawati, D. L. Sari, Lazfihma, dan F. Hiasa, "Penggunaan gaya bahasa sindiran kiky saputri untuk kritik sosial pada tayangan 'lapor pak,'" *J. Membaca Bhs. Sastra Indones.*, vol. 8, no. April, hal. 87–96, 2023.
- [28] Annisa, T. M. Tantika, dan Ngatma'in, "Sarkasme Netizen Di Media Sosial Tiktok," *Pros. Samasta*, hal. 94–103, 2022.
- [29] A. N. Febriyanto dan A. D. Savitri, "Sarkasme komentar netizen sepak bola pada akun instagram @Trolliyah," *Bapala*, vol. 9, no. 8, hal. 215–222, 2022.
- [30] D. Paramita, M. R. Aldiano, K. I. S. Siregar, H. Sazali, dan M. Andinata, "Analisis Penggunaan Gaya Bahasa Sarkasme Netizen Di Media Sosial Facebook," *J. Pendidik. Tambusai*, vol. 6, no. 2, hal. 14438–14445, 2022.
- [31] A. Safitri, S. Sumardjijati, Y. Candrasari, dan H. Arviani, "Analisis Komentar Sarkasme Netizen Terhadap Tayangan Youtube Deddy Corbuzier Bersama Agung Suprio," *Nusant. J. Ilmu Pengetah. Sos.*, vol. 9, no. 8, hal. 3127–3138, 2022, doi: 10.31604/jips.v9i8.2022.3127-3138.
- [32] I. Rahmayanti dan E. Nusivera, "Gaya Bahasa Sindiran Pandemi Covid-19 Pada Media Online," *Ling. Fr. Bahasa, Sastra, dan Pengajarannya*, vol. 6, no. 2, hal. 181, 2022, doi: 10.30651/lf.v6i2.7568.
- [33] U. Hasanah, A. R. Rahim, dan A. S. Syamsuri, "Analisis Penggunaan Gaya Bahasa Sarkasme Netizen di Media Sosial Instagram," *J. Onoma Pendidikan, Bahasa, dan Sastra*, vol. 7, no. 2, hal. 411–423, 2021, doi: 10.30605/onomा. v7i2.1255.
- [34] S. Lutfiyani, B. E. Purwanto, dan S. Anwar, "Sarkasme pada Media Sosial Twitter dan Implikasinya Terhadap Pembelajaran Bahasa Indonesia di SMA," *Tabasa J. Bahasa, Sastra Indones. dan Pengajarannya*, vol. 1, no. 2, hal. 269–284, 2021, doi: 10.22515/tabasa.v1i2.2628.
- [35] K. T. Syarifuddin, "Sarkasme pada Masyarakat Indonesia selama Pandemi Covid-19 dalam Media Sosial Twitter," *Senasbasa*, no. 4, hal. 66–79, 2020.
- [36] A. S. Cahyanti dan A. Sabardila, "Analisis Penggunaan Kalimat Sarkasme oleh Netizen di Media Sosial Instagram," *Ling. Fr. J. Bahasa, Sastra, dan Pengajarannya*, hal. 186–195, 2020.
- [37] S. N. Halimah dan H. Hilaliyah, "Gaya Bahasa Sindiran Najwa Shihab dalam Buku Catatan Najwa," *Deiksis*, vol. 11, no. 02, hal. 157–165, 2019, doi: 10.30998/deiksis.v11i02.3648.
- [38] E. Iderasari, F. Achsani, dan B. Lestari, "Bahasa Sarkasme Netizen Dalam Komentar Akun Instagram 'Lambe Turah,'" *Sematik*, vol. 8, no. 1, hal. 37–49, 2019, doi: 10.22460/semantik.vXiX.XXX.
- [39] M. Bouazizi dan T. Otsuki, "A Pattern-Based Approach for Sarcasm Detection on Twitter," *IEEE Access*, vol. 4, 2016, doi: 10.1109/ACCESS.2016.2594194.
- [40] A. Kumar, V. T. Narapareddy, V. A. Srikanth, A. Malapati, dan L. B. M. Neti, "Sarcasm Detection Using Multi-Head Attention Based Bidirectional LSTM," *IEEE Access*, vol. 8, hal. 6388–6397, 2020, doi: 10.1109/ACCESS.2019.2963630.
- [41] A. Kumar Jena, A. Sinha, dan R. Agarwal, "C-Net: Contextual Network for Sarcasm Detection," 2020, doi: 10.18653/v1/2020.figlang-1.8.
- [42] D. E. Birba, "A Comparative study of data splitting algorithms for machine learning model selection," *DEGREE Proj. Comput. Sci. Eng.*, 2020.
- [43] A. M. A. Barhoom, B. S. Abunasser, dan S. S. Abu-Naser, "Sarcasm Detection in Headline News using Machine and Deep Learning Algorithms," 2022.
- [44] IndoBenchmark, "indobenchmark/indobert-base-p1," 2020. <https://huggingface.co/indobenchmark/indobert-base-p1>.
- [45] J. Devlin, M. W. Chang, K. Lee, dan K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2019, vol. 1.
- [46] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, dan R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, 2014.