Deteksi Penggunaan Kalimat *Abusive* Pada Teks Bahasa Indonesia Menggunakan Metode IndoBERT

Hadiyan Kundrat Putra[1], Moch Arif Bijaksana[2], Ade Romadhony[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[4]Divisi Digital Service PT Telekomunikasi Indonesia
[1]binggows@students.telkomuniversity.ac.id, [2]arifbijaksan@telkomuniversity.ac.id,
[3]aderomadhony@telkomuniversity.ac.id

Abstrak

Saat ini banyak contoh yang biasa digunakan dalam pengembangan masalah deep learning, seperti Transformers. Penelitian ini menggunakan salah satu arsitektur dari Transformers yaitu IndoBERT, IndoBERT itu sendiri terdiri dari BERT yaitu Bidirectional Encoder Representations from Transformers yang biasa digunakan untuk masalah deep learning. Metode IndoBERT diimplementasikan untuk mendeteksi penggunaan kalimat abusive pada teks bahasa Indonesia. Dataset yang digunakan dalam penelitian ini memiliki ketidakseimbangan jumlah data pada setiap kelas sehingga akan dilakukan penambahan data untuk mengetahui pengaruh penambahan jumlah data terhadap kinerja hasil arsitektur tersebut. Tahapan pengerjaan dalam penelitian ini dimulai dari dataset, pra-pemrosesan data, pembuatan model dengan metode IndoBERT untuk mendeteksi kalimat abusive, pelatihan dan pengujian. Pengujian dilakukan terhadap arsitektur KNN, SVM, Naive Bayes, BERT Multilingual Base dan BERT Base lalu dibandingkan dengan IndoBERT. Hasil pengujian menunjukkan bahwa selain BERT Multilingual Base, BERT Base dan IndoBERT hanya dapat memprediksi terhadap kelas mayoritas sehingga dilakukan penambahan penggunaan dataset. Hasil pengujian menunjukkan IndoBERT dapat lebih baik dalam mengklasifikasikan kalimat abusive pada teks bahasa Indonesia. Di model BERT Base, berhasil menghasilkan nilai F1 Score untuk semua kelas sebesar 0.6842 IndoBERT sudah dapat menghasilkan nilai F1 Score untuk semua kelas lebih baik.

Kata kunci : Kalimat Abusive, IndoBERT, F1 Score, SVM, Naive Bayes, BERT

Abstract

Nowadays many examples are commonly used in the development of deep learning problems, such as Transformers. This research uses one of transformers architecture namely IndoBERT, IndoBERT itself consists of BERT namely Bidirectional Encoder Representations from Transformers which is commonly used for deep learning problems. IndoBERT method is implemented to detect the use of abusive sentences in Indonesian text. The dataset used in this study has an imbalance in the amount of data in each class so that there will be additional data to find out the effect of increasing the amount of data on the performance of the architecture results. The stages of work in this research started from dataset, data pre-processing, modeling with IndoBERT method to detect abusive sentences, training and testing. Testing was carried out on the architecture of KNN, SVM, Naive Bayes, BERT Multilingual Base and BERT Base and then compared to IndoBERT. The test results showed that in addition to BERT Multilingual Base, BERT Base and IndoBERT can only predict against the majority class so that the addition of dataset usage is carried out. The test results showed IndoBERT could be better at classifying abusive sentences in Indonesian texts. In the BERT Base model, successfully generating an F1 Score for all classes of 0.6842 IndoBERT can already produce an F1 Score for all classes better.

Keywords : Abusive Sentences, IndoBERT, F1 Score, SVM, Naive Bayes, BERT

1. Pendahuluan

Latar Belakang

Keberadaan internet dan munculnya berbagai jenis jejaring sosial adalah salah satu hasil dari perkembangan sebuah teknologi[1]. Asosisasi Penyelenggara Jasa Internet Indonesia (APJII) pada Laporan Survei Internet APJII 2019 – 2020 (Q2) menuliskan jumlah pengguna internet di Indonesia mencapai 196,7 juta pengguna atau setara dengan 73,7% dari total populasi RI yang mencapai 266,9 juta penduduk[2]. Sehingga dengan penggunaan internet di Indonesia itu semakin meningkat dapat mengakibatkan juga penyalahgunaan dalam internet itu sendiri, khususnya media sosial. Banyak yang masih menggunakan media sosial dengan tidak bertanggung jawab dengan dalih ‘media sosial saya ini, ya terserah saya mau ngomong apa saja’. Kalimat-kalimat yang melecehkan seringkali

digunakan untuk menyerang partai politik tertentu atau bahkan hanya untuk bercanda. Jika kalimat *abusive* dapat dengan mudah ditemukan di internet, dapat mempengaruhi kesehatan mental pihak yang dituju jika pihak dituju masih tergolong remaja[3].

Kalimat *abusive* merupakan ungkapan yang mengandung kata-kata kasar dan frasa yang kotor baik dalam bahasa lisan maupun tulisan, dikarenakan kalimat *abusive* mungkin akan melibatkan sarkasme, pelecehan dll[4]. Alasan banyaknya penggunaan kalimat yang melecehkan di internet atau jejaring sosial adalah kurangnya alat yang efektif untuk menyaring penggunaan kalimat yang melecehkan, kurangnya empati di antara pengguna Internet, dan kurangnya pengawasan orang tua[5]. Sebagai contoh, *Hate Speech* atau ujaran kebencian yang berisi kalimat-kalimat *abusive* biasanya memicu konflik sosial karena akan membawa emosi kepada pihak atau pembaca yang dituju[6]. Contoh di *twitter* banyak sekali cuitan atau *tweet* yang mengandung *hate speech* baik secara individu (*cyberbullying)* atau juga kelompok tertentu (LGBT, Agama, Jenis Kelamin)[7], dan untuk kalimat yang mengarah pada tindakan *cyberbullying* yang mana tingkat depresi yang akan dialami korban dapat lebih tinggi daripada depresi yang didapat dari kekerasan secara fisik. Oleh karena itu penting mendeteksi adanya kalimat *abusive* atau *hate speech* seperti itu untuk menganalisis sentimen publik dari sekelompok pengguna terhadap kelompok yang lain[8].

Dalam mengatasi masalah tersebut, perlu dilakukan tindakan preventif dan penegakan hukum secara tegas menurut hukum yang berlaku. Salah satu tindakan yang bisa dilakukan ialah mendeteksi kalimat *abusive* pada media sosial, akan tetapi akan memakan waktu lama untuk mendeteksi secara manual. Oleh karena itu hal ini mendorong para peneliti untuk menciptakan cara – cara otomatis dengan cara membangun sistem yang dapat mendeteksi kalimat *abusive* pada media sosial. Penelitian sebelumnya ada penelitian yang mengenai kalimat *abusive* pada media sosial di Indonesia dengan dataset yang memiliki multi-label dengan menggunakan Support Vector Machine (SVM), Naive Bayes (NB), and Random Forest Decision Tree (RFDT) classifier and Binary Relevance (BR), Label Power-set (LP), and Classifier Chains (CC) sebagai data transformasinya[6]. Dalam beberapa tahun terakhir model bahasa menggunakan *pre-trained* telah menunjukkan terobosan besar dalam NLP dan pada Devlin dkk membuat BERT yaitu sebuah arsitektur untuk melatih model bahasa yang lebih cepat yang mehilangkan *recurrences* dengan menambahkan *multi-head attention layer*[9]. Sehingga pada penelitian kali ini digunakan metode IndoBERT untuk mengklasifikasikan kalimat inputan menjadi label *Not Abusive, Abusive But Not Offensive,* dan *Abusive And Offensive* serta akan dinilai akurasi sistem menggunakan *F1 Score*. IndoBERT itu sendiri merupakan *transformers-based model* yang bergaya BERT itu sendiri[10]. Akan tetapi terdapat *imbalance* atau ketidakseimbangan jumlah antar label pada dataset yang digunakan sehingga terdapat kelas mayoritas dan minoritas. Untuk itu peneliti juga ingin mengetahui pengaruh terhadap adanya penambahan jumlah dataset pada arsitektur yang dibangun.

Topik dan Batasannya

Topik penelitiannya adalah mendeteksi penggunaan kalimat-kalimat *abusive* melalui input berupa teks bahasa Indonesia kemudian, menggunakan metode IndoBERT untuk memproses bahasa Indonesia untuk mengklasifikasikan jenis kalimat. Ada 3 jenis label dalam kategori keluaran, seperti yang ditunjukkan di bawah ini:

Gambar 1 Gambaran Sistem Secara Umum

Batasan masalah dalam penelitian ini ialah jumlah dataset yang digunakan terhadap penelitian ini merupakan dataset yang bersifat imbalance atau ketidakseimbangan dari tiap labelnya.

Tujuan

Tujuan dari penilitian ini yaitu mengevaluasi penerapan metode IndoBERT untuk melakukan deteksi penggunaan kalimat *abusive* pada teks bahas Indonesia, perhitungan nilai akurasi nilai *F1 Score* untuk masing masing kelas dan pengaruh adanya penambahan data pada dataset kelas minoritas terhadap nilai akurasi *F1 Score*.

Organisasi Tulisan

Pada bagian 2 menjelaskan dasar teori yang digunakan sebagai pedoman penelitian. Pada bagian 3 menjelaskan alur penelitian yang dilakukan. Pada bagian 4 menjelaskan hasil yang didapat melalui algoritma dan

sistem yang sudah dibangun. Pada bagian 5 menjelaskan kesimpulan yang didapat dan saran yang bisa dikembangkan untuk penelitian selanjutnya

2. Studi Terkait

Penelitian ini dibangun berdasarkan beberapa referensi dari penelitian yang sudah dilakukan sebelumnya. Beberapa tahun terakhir penelitian mengenai *offensive language*, *hate speech* atau *abusive language* cukup banyak bahkan dijadikan salah satu task pada SemEval 2019. Pada penelitian [11], task utama dibagi menjadi 3 Sub-task yaitu Sub-task A: *Offensive language identification* (104 partisipan) Sub-task B: *Automatic categorization of offense types* (71 partisipan), Sub-task C : *Offense target identification* (66 partisipan). Sub-task A bertujuan untuk mendiskripsikan antara *offensive* dan *non offensive post* dengan 2 label yaitu Not Offensive (NOT) dan Offensive (OFF). Sub-task B bertujuan untuk dapat memprediksi tipe dari pelanggarannya seperti apa, dengan 2 label yaitu Target Insult (TIN) dan Untargeted (UNT), dan Sub-task C bertujuan untuk mencari fokus target pelanggarannya seperti apa, yang dibagi menjadi 3 label yaitu Individual (IND), Group (GRP) dan Other (OTH). Pada penelian tersebut model yang digunakan berkisar dari pembelajaran mesin tradisional seperti SVM dan *logistic regression* lalu *deep learning* seperti CNN, RNN, BiLSTM hingga model pembelajaran yang canggih seperti ELMo dan BERT

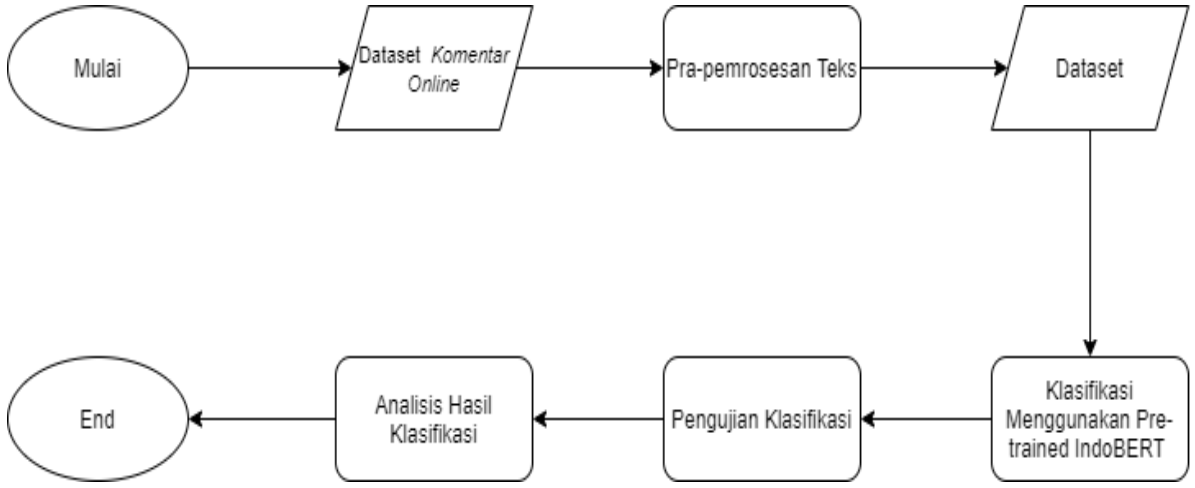
Pada penelitian [12], penelitian ini merupakan performansi terbaik pada Sub-task A SemEval 2019 task 6 dengan menggunakan BERT Base model. BERT Base terdiri dari 12 trasnformers blocks, 12 self-attention heads, dan 768 hidden dimension dengan total parameternya ialah 110 M dan di trained dengan BookCorpus (800M kata) serta English Wikipedia (2500M). Paramater yang digunakan selanjutnya ialah dropout 0.1 pada setiap layer dan juga digunakannya *learning rate* 2e-5. Dalam Sub-task A sistem tersebut berhasil mendapatkan nilai F1-Score 82.9% berbeda 1.4 point dari sistem selanjutnya yang berada di peringkat 2.

Pada penelitian [13], penelitian ini merupakan performanis terbaik pada Sub-task C SemEval 2019 task 6 serta nomor 4 untuk Sub-task A Semeval 2019 task 6, mereka menggunakan beberapa model untuk pengerjaannya seperti standar SVM dari *libary python*, CNN, BERT dan *Logistic Regression* serta mengatasi masalah *class imbalance* dengan *oversampling* serta *class weight*

Pada penelitian [14], penelitian ini bertujuan untuk mendeteksi ujaran kebencian atau *hate speech* pada kasus Pemilihan Presiden (Pilpres) 2019. Dataset yang digunakan berasal dari kolom komentar media sosial Facebook, dengan jumlah kalimat sebanyak 950. Dalam penelitian ini data dibagi menjadi 2 label, yaitu *hate speech* (HS) dan *non-hate speech* (Non_HS). Dalam penelitian ini, pengujian dilakukan dua pengujian dengan menggunakan jumlah *datatesting* yang berbeda. Penelitian ini menggunakan LSTM dan Word2Vec memiliki nilai *accuracy* 58.42%, nilai *recall* 0.7021, dan nilai *precision* 0.5641.

3. Sistem yang Dibangun

Sistem yang akan dibangun merupakan sistem yang dapat mendeteksi pengunaan kalimat *abusive* pada teks Bahasa Indonesia.



Gambar 2 Alur Pembuatan Sistem

3. 1 Pembangunan Dataset

Pada penelitian ini menggunakan dataset yang sudah dibangun dari penelitian[15], data yang digunakan berasal dari komentar berita *online*. Komentar dipilih berdasarkan berita yang sedang *trend* pada bulan Maret 2019 hingga September 2019. Total data yang didapatkan sebanyak 3184 komentar. Data terdiri dari tiga label dengan jumlah sebagai berikut. Dataset tersebut termasuk dalam *imbalance data*, maka dari itu ditambahkan *class weight* yang supaya label yang lebih sedikit akan ditambahkan bobot yang lebih tinggi dibandingkan label yang lebih banyak dalam hal ini *class weight* dari label 2 dan 3 lebih tinggi dibandingkan label 1.

Tabel 1 Jumlah Data Setiap Label Pada Dataset Pertama

Label	Keterangan	Jumlah
1	Not Abusive	2789
2	Abusive Not Offensive	110
3	Abusive And Offensive	285

3. 2 Pra-pemrosesan Teks

Pra-pemrosesan teks adalah tahapan untuk mempersiapkan teks menjadi data yang lebih terstruktur untuk bisa diolah ke tahapan berikutnya. Langkah-langkah yang dilakukan adalah *case folding, removal punctuation, tokenizing, stopword removal, replacing acronym, stemming*

A. Case Folding

Tahapan ini digunakan untuk mengubah semua karakter menjadi huruf kecil

Tabel 2 Pra-pemprossesan Teks Case Folding

Input	Output
Sebagai ketua KPAI mustinya anda ngurusin masalah anak2 dibawah umur yg ikut demo, tawuran atau yg dilecehkan. Ini kok malah cari2 masalah dng PB Djarum yg jelas2 sdh memberikan banyak sumbangan bagi prestasi bulutangkis nasional. Dasar otak udang loe!!!	sebagai ketua kpai mustinya anda ngurusin masalah anak2 dibawah umur yg ikut demo, tawuran atau yg dilecehkan. ini kok malah cari2 masalah dng pb djarum yg jelas2 sdh memberikan banyak sumbangan bagi prestasi bulutangkis nasional. dasar otak udang loe!!!

B. Removal Punctuation

Tahapan ini digunakan untuk menghilangkan angka, url, mention dan tanda baca yang ada pada kalimat

Tabel 3 Pra-pemrosesan Teks Removal Punctuation

Input	Output
sebagai ketua kpai mustinya anda ngurusin masalah anak2 dibawah umur yg ikut demo, tawuran atau yg dilecehkan. ini kok malah cari2 masalah dng pb djarum yg jelas2 sdh memberikan banyak sumbangan bagi prestasi bulutangkis nasional. dasar otak udang loe!!!	sebagai ketua kpai mustinya anda ngurusin masalah anak dibawah umur yg ikut demo tawuran atau yg dilecehkan ini kok malah cari masalah dng pb djarum yg jelas sdh memberikan banyak sumbangan bagi prestasi bulutangkis nasional dasar otak udang loe

C. Replacing Acronym

Tahapan ini digunakan untuk mengganti kata kata yang disingkat dengan kata aslinya berdasarkan sebuah kamus data singkatan[15].

Tabel 4 Pra-pemrosesan Teks Replacing Acronym

Input	Output
sebagai ketua kpai mustinya anda ngurusin masalah anak2 dibawah umur yg ikut demo, tawuran atau yg dilecehkan. ini kok malah cari2 masalah dng pb djarum yg jelas2 sdh memberikan banyak sumbangan bagi prestasi bulutangkis nasional. dasar otak udang loe!!!	sebagai ketua kpai mustinya anda ngurusin masalah anak dibawah umur yg ikut demo tawuran atau yg dilecehkan ini kok malah cari masalah dng pb djarum yg jelas sdh memberikan banyak sumbangan bagi prestasi bulutangkis nasional dasar otak udang loe

D. Stop Removal

Tahapan digunakan untuk menghapus kata-kata yang tidak diperlukan seperti kata bantu yang diantaranya adalah ‘akan’, ‘yang’, ‘juga’, ‘untuk’, ‘dan’, ‘dari’, ‘maka’, ’di’, ‘kan’ menggunakan data stopwords yang diambil dari [16] .

Tabel 5 Pra-pemrosesan Teks Stop Removal

Input	Output
sebagai ketua kpai mustinya anda ngurusin masalah anak dibawah umur yang ikut demo tawuran atau yang dilecehkan ini kok malah cari masalah dengan pb djarum yang jelas sudah memberikan banyak sumbangan bagi prestasi bulutangkis nasional dasar otak udang loe	sebagai ketua kpai mustinya anda ngurusin masalah anak dibawah umur ikut demo tawuran dilecehkan malah cari masalah dengan pb djarum jelas sudah memberikan banyak sumbangan bagi prestasi bulutangkis nasional dasar otak udang loe

E. Stemming

Tahapan ini digunakan untuk menghilangkan imbuhan yang terdapat pada kata, sehingga mengubahnya menjadi kata asli

Tabel 6 Pra-pemrosesan Teks Stemming

Input	Output
sebagai ketua kpai mustinya anda ngurusin masalah anak dibawah umur ikut demo tawuran dilecehkan malah cari masalah dengan pb djarum jelas sudah memberikan banyak sumbangan bagi prestasi bulutangkis nasional dasar otak udang loe	sebagai ketua kpai musti anda urus masalah anak bawah umur ikut demo tawuran dilecehkan malah cari masalah dengan pb djarum jelas sudah beri banyak sumbangan bagi prestasi bulutangkis nasional dasar otak udang loe

F. Tokenizing

Tahapan ini digunakan untuk memecah kalimat menjadi list kata

Tabel 7 Pra-pemrosesan Teks Tokenizing

Input	Output
sebagai ketua kpai mustinya anda ngurusin masalah anak dibawah umur ikut demo tawuran dilecehkan malah cari masalah dengan pb djarum jelas sudah memberikan banyak sumbangan bagi prestasi bulutangkis nasional dasar otak udang loe	'sebagai', 'ketua', 'kp', '##ai', 'musti', '##nya', 'anda', 'ngurus', '##in', 'masalah', 'anak', 'dibawah', 'umur', 'ikut', 'demo', 'tawuran', 'dil', '##ecehkan', 'malah', 'cari', 'masalah', 'dengan', 'pb', 'djarum', 'jelas', 'sudah', 'memberikan', 'banyak', 'sumbangan', 'bagi', 'prestasi', 'bulutangkis', 'nasional', 'dasar', 'otak', 'udang', 'loe'

3. 3 Dataset

Dataset yang sudah dilakukan pra-pemrosesan teks kemudian disimpan dalam *file* 'Dataset.csv'. Berikut contoh kalimat pada dataset berdasarkan label yang sudah melalui tahapan pra-pemrosesan teks. Yang nanti akan dibagi menjadi 60% data train, 20% untuk masing masing data test dan validasi.

Tabel 8 Contoh Kalimat Pada Dataset Pertama

Kalimat	Label
indosat udah rugi mulu ngapain beli lagi klo murah sih gpp 1 t mending buat besarin telkom telkomsel	1 (Not Abusive)
begini tipe org yg sdg cr citra tdk berfikir pjg bs nya ngebodohin dan bohong rakyat sj	2 (Abusive Not Offensive)
goblok ngapain beli indosat g ada untung jaring aja super lot di papua aja g ada jaring mentok2 cuma 2g harus pikir pakai otak bagaimana beli seluSruh saham telkomsel dari singtel	3 (Abusive And Offensive)

3. 4 IndoBERT

Pada penelitian [17], Dijelaskan bahwa IndoBERT sendiri merupakan modifikasi dari BERT Base yang sudah ada dengan mengikuti kofigurasi dari BERT-Base (uncased) yang memiliki 12 hidden layers masing masing memiliki 768d, 12 attention heads, and feed-forward hidden layers of 3,072d. Jika di total IndoBERT di train dengan lebih dari 220M kata. Terdiri dari 3 main resources Indonesia Wikipedia (74M kata), artikel Kompas Tempo dan Liputan6 (total 55M) dan Indonesia Web Corpus (90 M kata). IndoBERT merupakan salah satu monlingual BERT model untuk bahasa Indonesia , IndoBERT memiliki 3 model IndoBERT-lite$_{Base}$, IndoBERT$_{Base,}$ IndoBERT$_{Large}$. BERT memanfaatkan *transformer*, mekanisme perhatian yang mempelajari hubungan kontekstual antara kata (atau sub kata) dalam teks. Dalam bentuk, *transformer* menyertakan dua mekanisme terpisah yaitu encoder yang membaca input teks dan decoder yang menghasilkan prediksi untuk tugas tersebut. Karena tujuan BERT adalah untuk menghasilkan model bahasa, hanya diperlukan mekanisme encoder. Berlawanan dengan model sekuensial (dari kiri ke kanan atau kanan ke kiri), yang membaca teks masukan secara berurutan, pembuat encode *transformer* membaca seluruh urutan kata sekaligus. Oleh karena itu, meskipun lebih akurat untuk mengatakan bahwa itu non-arah, itu dianggap dua arah. Fitur-fitur ini memungkinkan model untuk mempelajari konteks kata berdasarkan semua lingkungan sekitar (kata kiri dan kanan) kata tersebut[18].

3. 5 Klasifikasi

Pada tahap ini akan dilakukan proses klasifikasi teks bahasa Indonesia dengan menggunakan IndoBERT$_{Base}$. Terdapat 2 langkah pada penggunaan metode IndoBERT yaitu *pre training* dan *fine tuning.*

1. Pre Training

Pada langkah pertama yaitu *pre training* terdapat 2 tahap yang akan dilalu yaitu *Masked Language Modeling (MLM)* dan *Next Sentence Prediction (NSP).*

A. Masked Language Modeling (MLM)

Disini kita akan mengganti 15% kata disetiap urutan token dengan [MASK] secara random. Kemudian, model akan mencoba memprediksi nilai asli dari kata yang di MASK tersebut. Berdasrkan konteks yang disediakan oleh kata lain, kata yang tidak di MASK dan kata kata dalam urutan tersebut. Contoh sebagai berikut,

Kalimat : "the man went to store he bought a gallon milk"

Encode : 'the', 'man', 'wen', '##t', 'to', 'store', 'he', 'bo', '##ught', 'a', 'gall', '##on', 'milk'

B. *Next Sentence Prediction (NSP)*

Dalam proses training BERT, model memilih kalimat A dan kalimat B untuk setiap contoh pre training sebelumnya, 50% dari pretraining tersebut memang benar Kalimat B merupakan kalimat selanjutnya dari kalimat A (dilabeli sebagai IsNext) dan 50 % lainnya merupakan kalimat acak dari korpus (dilabeli NotNext). Dan untuk membantu model membedakan dua kalimat pada training, input di proses dengan cara berikut sebelum masuk ke dalam model

1. Token [CLS] disisipkan diawal kalimat dan token [SEP] disisipkan diakhir kalimat

2. *Sentence Embedding* yang terindikasi merupakan kalimat A dan kalimat B ditambahkan ke setiap token.
3. Sebuah Embeddings posisi ditambahkan ke setiap token untuk menunjukkan posisinya dalam urutan. Berikut contoh kalimat yang menggambarkan proses dari *Next Sentence Prediction* (NSP),

```
Input = [CLS] the man went to [MASK] store [SEP]
        he bought a gallon [MASK] milk [SEP]
Label = IsNext

Input = [CLS] the man [MASK] to the store [SEP]
        penguin [MASK] are flight ##less birds [SEP]
Label = NotNext
```

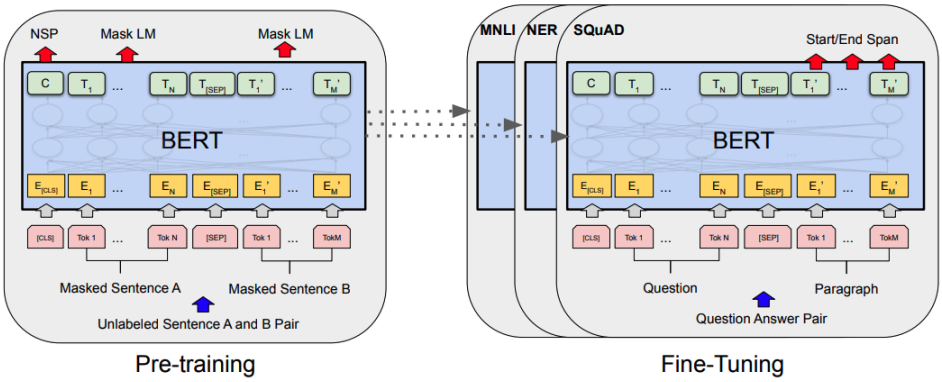
Gambar 3 Contoh Kalimat Pada Proses NSP[19]

2. Fine Tunning

Dalam tahap ini karena model yang diinginkan merupakan bentuk dari *sentiment analysis* dimana memiliki kesamaan dengan *Next Sentence Classification* dengan menambahkan layer diatas output Transformers untuk token [CLS]. Cukup mencolokkan input dan output spesifikasi tugas ke BERT dan menyempurnakan dengan semua parameter. Parameter yang digunakan sebagai berikut

Tabel 9 Nilai Hyperparameters Model IndoBERT dan BERT Lainnya

Parameter	Nilai
Dropout	0.1
Batch Size	32
Learning Rate (AdamW)	2e-5
Epoch	10



Gambar 4 Arsitektur Pre-Training dan Fine-Tuning

3. 6 Evaluasi Sistem

Evaluasi sistem dilakukan dengan menggunakan *confusion matrix multiclass* dan menghitung nilai akurasi, *precision, recall dan F1-score* dari setiap kelas. *Confusion Matrix Multiclass* adalah sebuah tabel yang sering digunakan untuk mendeskripsikan peformansi model klasifikasi pada kumpulan set data testing dengan nilai aktual

yang sudah diketahui[20]. Tabel *confusion matrix multiclass* yang diterapkan pada penelitian ini adalah sebagai berikut.

Tabel 10 Confussion Matrix Multiclass

		Predicted		
		Not Abusive	**Abusive Not Offensive**	**Abusive And Offensive**
Actual	**Not Abusive**	TP NotAbusive	X	X
	Abusive Not Offensive	X	TP Abusive Not Offensive	X
	Abusive And Offensive	X	X	TP Abusive And Offensive

Pada tabel 10 terdapat tiga label prediksi sesuai dengan dataset. TP adalah singkatan dari True Positive, yaitu kasus di mana nilai prediksi dan nilai sebenarnya adalah True atau benar. Pada matrix multi class kebingungan hanya TP yang dicantumkan, karena untuk penentuan FN (false negative) berasal dari semua baris tiap label, dan untuk penentuan FP (false positive) didapat dari total dari setiap label. Jumlah kolom. Label dan TN (True Negative) mengacu pada saat nilai prediksi tidak ada dan nilai sebenarnya salah[20]. Performa dari deteksi penggunaan kalimat abusive diukur dengan menggunakan beberapa parameter yaitu precision, recall dan F1 Score. Formula dari parameter-parameter tersebut diberikan pada persamaan (1-3).

Precision adalah perbandingan antara data yang diklasifikasikan secara benar dibandingkan dengan seluruh data yang diklasifikasikan secara benar.

$$Precision = \frac{TP}{(TP+FP)} \quad (1)$$

Recall adalah perbandingan antara data yang diklasifikasikan secara benar dengan jumlah data yang berada di kelas tersebut. Rumus untuk masing-masing perhitungan adalah sebagai berikut.

$$Recall = \frac{TP}{(TP+FN)} \quad (2)$$

F1 Score adalah perbandingan rata-rata nilai *precision* dan nilai *recall* yang dibobotkan.

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (3)$$

4. Evaluasi dan Analisis

Eksperimen dilakukan dengan membandingkan nilai F1 Score dari semua kelas menggunakan metode *non-neural network* yaitu KNN, Naive Bayes dan SVM dengan metode lain yaitu BERT Base, BERT Multilingual Base dan IndoBERT. Hasil perbandingan nilai F1 Score yang didapatkan dari kedua arsiterktur seperti pada Tabel 10 dibawah ini.

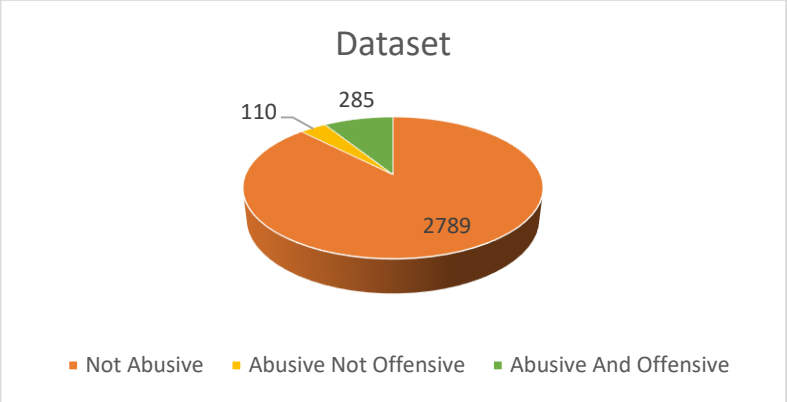
Tabel 11 Hasil F1 Score Masing Masing Model Pada Dataset Pertama

Arsitektur	F1 Score Kelas 1	F1 Score Kelas 2	F1 Score Kelas 3
KNN	0.9339	0	0
Naive Bayes	0.3319	0.0863	0.2368
SVM	0.9339	0	0
BERT Multilingual Base	0.9381	0.0800	0.4902
BERT Base	0.9268	0.0500	0.4737
IndoBERT	**0.9246**	**0.2182**	**0.5085**

Dari hasil eksperimen pertama menunjukkan bahwa arsitektur KNN, SVM hanya dapat mengklasifikasikan data terhadap kelas mayoritas atau kelas Not Abusive dan kedua arsitektur belum mampu mengklasifikasikan data terhadap kelas minoritas. Dari hasil F1 Score pada tabel 10 juga menunjukkan bahwa arsitektur BERT Base dan

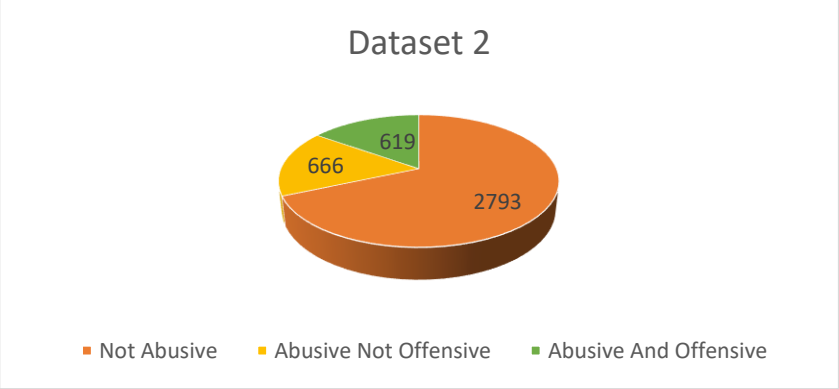
IndoBERT yang merupakan salah satu dari arsitektur lebih baik dalam mengklasifikasikan kalimat daripada arsitektur KNN, Naive Bayes dan SVM yang merupakan arsitektur non-neural network serta juga beberapa model BERT lainnya.

Agar tercapai tujuan penelitian dan mengetahui performansi arsitektur tersebut, dilakukan uji skenario untuk mengetahui dampak penambahan jumlah dataset pada arsitektur BERT dan arsitektur IndoBERT yang dibangun. Dalam skenario pengujian ini, dua dataset dengan nomor berbeda digunakan untuk pengujian. Pada pengujian pertama, dataset memiliki total 3184 kalimat, termasuk 2789 kalimat yang ditandai sebagai 1 atau Not Abusive, 110 kalimat yang ditandai sebagai 2 atau Abusive Not Offensive, dan 285 kalimat yang ditandai sebagai 3 atau Abusive And Offensive seperti Gambar 5.



Gambar 5 Jumlah Data Setiap Label Pada Dataset Pertama

Pada pengujian kedua, terdapat penambahan data pada dataset dari penelitianan[6]. Dataset tersebut dipilih karena dataset tersebut berisi kalimat bahasa indonesia dengan pelabelan pada kalimat tersebut memiliki kesamaan makna dengan dataset yang sudah digunakan pada awal penelitian ini. Penambahan dataset sebanyak 894 kalimat dimana terdiri dari 2.793 kalimat berlabel 1 atau Not Abusive, 666 kalimat berlabel 2 atau Abusive Not Offensive dan 619 Kalimat berlabel 3 atau Abusive And Offensive yang mana jumlah total data menjadi 4.078 kalimat. Jumlah data dalam dataset pengujian pertama dapat dilihat pada gambar 9 dan dataset pengujian kedua dapat dilihat pada Gambar 6



Gambar 6 Jumlah Data Setiap Label Pada Dataset Kedua

Dari pengujian dengan menggunakan jumlah dataset yang berbeda didapatkan hasil F1-Score pada setiap kelasnya seperti berikut :

Tabel 12 Hasil Pengujian Setelah Penambahan Dataset

Arsitektur	F1 – Score Kelas 1		F1 – Score Kelas 2		F1 – Score Kelas 3		Marco Avg F1 Score	
	Dataset 1	Dataset 2	Dataset 1	Dataset 2	Dataset 1	Dataset 2	Dataset 1	Dataset 2
IndoBERT	**0.9246**	**0.9399**	**0.2727**	**0.7331**	**0.5085**	**0.6165**	**0.5504**	**0.7632**
BERT Base	0.9268	0.9204	0.0500	0.5867	0.4737	0.5455	0.4835	0.6842

BERT Multilingual Base	0.9381	0.8988	0.0800	0.0567	0.4902	0.5049	0.5028	0.4868

Berdasarkan Tabel 11 pada pengujian menggunakan dataset pertama terlihat dari ketiga model BERT tersebut terlihat jika memang dengan keadaan dataset yang imbalance pun ketiga model BERT tersebut dapat mengklasifikasikan semua kelas walaupun dengan akurasi yang sangat kecil. Setelah dilakukannya percobaan dengan penambahan jumlah data pada dataset kedua, terlihat bahwa IndoBERT sudah mampu mengklasifikasi semua kelas dan menghasilkan nilai *F1 Score* untuk semua kelas dari dataset dengan nilai *F1 Score* diatas 50% per kelas.

Berdasarkan pengujian arsitektur terhadap data test, dilakukan analisis pada beberapa kalimat yang belum sesuai diklasifikasikan atau belum sesuai oleh arsitektur dan didapatkan beberapa alasan sebagai berikut:

- Terdapat beberapa kata kasar yang berubah makna setelah proses pra-pemrosesan teks. Contoh kata ‘bajingan’ setelah proses steeming maka katanya berubah jadi bajing
- Terdapat beberapa kata seperti ‘tai’,‘anjing’,’ yang memiliki bisa 2 makna yang sudah dikategorikan menjadi kata kasar sehingga jika dipadukan dengan kata yang lain tetap akan mendapat label ‘abusive’
- Terdapat beberapa kalimat yang ditujukan untuk menghina namun tidak menggunakan kata-kata kasar atau hinaan yang mana arsitektur belum mampu untuk mengenali konteks dengan kalimat hinaan yang dituliskan secara implisit seperti ‘wkwkwkkkk nek aku sing nyawang rupane zonktor persis upil coro kecoak’ dan ‘bukan lupa lagi judul tapi blaga budek iya pasti wkwkwkwk’

5. Kesimpulan

Arsitektur model IndoBERT sudah mampu mendeteksi kalimat *abusive* dengan cukup optimal terutama pada kasus dataset kedua walaupun pada kasus dataset pertama sebenernya jika kita bandingkan dengan model BERT yang lainnya IndoBERT jelas lebih unggul dari segi akurasi *F1 Score.*

Pada IndoBERT sudah dapat mengklasifikasikan ketiga jenis kelas dan menghasilkan semua nilai F1 Score untuk setiap kelas. Hal ini dikarenakan pada IndoBERT memanfaatkan transformer, mekanisme perhatian yang mempelajari hubungan kontekstual antara kata (atau sub-kata) dalam teks. Sehingga proses pembelajaran model lebih kompleks dalam mengenal kontekstual antar kata yang mana dalam hal ini akan meningkatkan keakuratan hasil klasifikasi pada setiap label.

Penambahan Dataset juga mempengaruhi pada hasil *F1 Score* dari arsitektur IndoBERT. Hasil *F1 Score* model IndoBERT mengalami peningkatan setelah dilakukan adanya penambahan data pada kelas minoritas dengan nilai rata- rata *F1- Score* dari tiap kelas dengan 76,32%

Reference

[1] Y. Fitriani, “Analisis Pemanfaatan Berbagai Media Sosial sebagai Sarana Penyebaran Informasi bagi Masyarakat,” *Paradig. - J. Komput. dan Inform.*, vol. 19, no. 2, pp. 148–152, 2017, [Online]. Available: http://ejournal.bsi.ac.id/ejurnal/index.php/paradigma/article/view/2120.

[2] APJII, “Laporan Survei Internet APJII 2019 – 2020,” *Asos. Penyelenggara Jasa Internet Indones.*, vol. 2020, pp. 1–146, 2020, [Online]. Available: https://apjii.or.id/survei.

[3] Z. Xu and S. Zhu, “Filtering offensive language in online communities using grammatical relations,” 2010.

[4] J. Pavlopoulos, P. Malakasiotis, and I. Androutsopoulos, “Deep learning for user comment moderation,” *arXiv*, 2017, doi: 10.18653/v1/w17-3004.

[5] S. Tuarob and J. L. Mitrpanont, “Automatic discovery of abusive thai language usages in social networks,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017, vol. 10647 LNCS, pp. 267–278, doi: 10.1007/978-3-319-70232-2_23.

[6] M. O. Ibrohim and I. Budi, “Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter,” pp. 46–57, 2019, doi: 10.18653/v1/w19-3506.

[7] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, “Deep learning for hate speech detection in tweets,” *26th Int. World Wide Web Conf. 2017, WWW 2017 Companion*, no. 2, pp. 759–760, 2019, doi: 10.1145/3041021.3054223.

[8] G. K. Pitsilis, H. Ramampiaro, and H. Langseth, “Detecting offensive language in tweets using deep learning,” *arXiv*, pp. 1–17, 2018, doi: 10.1007/s10489-018-1242-y.

[9] 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.

[10] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," 2020, [Online]. Available: http://arxiv.org/abs/2009.05387.

[11] J. Han, S. Wu, and X. Liu, "Identifying and Categorizing Offensive Language in Social Media," *SemEval-2019*, pp. 652–656, 2019.

[12] J. Zhu, Z. Tian, and S. Kübler, "UM-IU@LING at SemEval-2019 Task 6: Identifying offensive tweets using BERT and SVMs," *arXiv*, 2019, doi: 10.18653/v1/s19-2138.

[13] A. Nikolov and V. Radivchev, "Nikolov-Radivchev at SemEval-2019 Task 6: Offensive Tweet Classification with BERT and Ensembles," pp. 691–695, 2019, doi: 10.18653/v1/s19-2123.

[14] A. S. Talita and A. Wiguna, "Implementasi Algoritma Long Short-Term Memory (LSTM) Untuk Mendeteksi Ujaran Kebencian (Hate Speech) Pada Kasus Pilpres 2019," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 19, no. 1, pp. 37–44, 2019, doi: 10.30812/matrik.v19i1.495.

[15] D. R. K. Desrul and A. Romadhony, "Abusive Language Detection on Indonesian Online News Comments," in *2019 2nd International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2019*, 2019, pp. 320–325, doi: 10.1109/ISRITI48646.2019.9034620.

[16] F. Z. Tala, "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia," *M.Sc. Thesis, Append. D*, vol. pp, pp. 39–46, 2003.

[17] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," 2020, [Online]. Available: http://arxiv.org/abs/2011.00677.

[18] R. Horev, "BERT Explained: State of the art language model for NLP," *https://towardsdatascience.com/*, 2020. https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270 (accessed Jan. 25, 2020).

[19] Yashu seth, "BERT Explained – A list of Frequently Asked Questions," *https://yashuseth.blog/*, 2019. https://yashuseth.blog/2019/06/12/bert-explained-faqs-understand-bert-working/ (accessed Feb. 15, 2021).

[20] V. W. Siburian and I. E. Mulyana, "Prediksi Harga Ponsel Menggunakan Metode Random Forest," *Pros. Annu. Res. Semin.*, vol. 4, no. 1, pp. 144–147, 2018.