

**ANALISIS SENTIMEN SARA PADA *TWEET* BERBAHASA INDONESIA
MENGUNAKAN INDOBERT DAN SUPPORT VECTOR MACHINE (SVM)**

SKRIPSI



**UIN SUNAN AMPEL
S U R A B A Y A**

Disusun Oleh:

**AGUNG SUSILO
H96219037**

**PROGRAM STUDI SISTEM INFORMASI
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI SUNAN AMPEL
SURABAYA
2023**

PERNYATAAN KEASLIAN

Saya yang bertanda tangan di bawah ini,

Nama : AGUNG SUSILO
NIM : H96219037
Program Studi : SISTEM INFORMASI
Angkatan : 2019

Menyatakan bahwa saya tidak melakukan plagiat dalam penulisan skripsi saya yang berjudul: “ANALISIS SENTIMEN SARA PADA TWEET BERBAHASA INDONESIA MENGGUNAKAN INDOBERT DAN SUPPORT VECTOR MACHINE (SVM)”. Apabila suatu saat nanti terbukti saya melakukan tindakan plagiat, maka saya bersedia menerima sanksi yang telah ditetapkan.

Demikian pernyataan keaslian ini saya buat dengan sebenar-benarnya.

Surabaya, 12 Juli 2023

Yang menyatakan,


METERAI
PEMPEL
B2AKX544920849
(AGUNG SUSILO)
NIM H96219037

LEMBAR PERSETUJUAN PEMBIMBING

Skripsi oleh

NAMA : AGUNG SUSILO

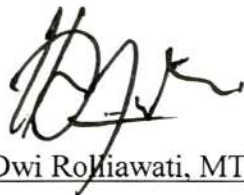
NIM : H96219037

JUDUL : ANALISIS SENTIMEN SARA PADA TWEET
BERBAHASA INDONESIA MENGGUNAKAN
INDOBERT DAN SUPPORT VECTOR MACHINE (SVM)

Ini telah diperiksa dan disetujui untuk diujikan.

Surabaya, 12 Juli 2023

Dosen Pembimbing 1



Dwi Rohawati, MT

NIP 197909272014032001

Dosen Pembimbing 2



Khalid, M.Kom

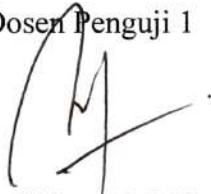
NIP 197906092014031002

PENGESAHAN TIM PENGUJI SKRIPSI

Skripsi Agung Susilo ini telah dipertahankan di depan tim penguji skripsi
di Surabaya, 14 Juli 2023

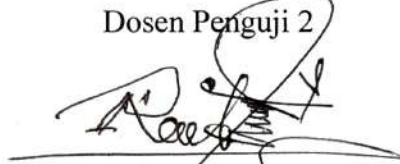
Mengesahkan, Dewan Penguji

Dosen Penguji 1



Ahmad Yusuf, M. Kom.
NIP 199001202014031003

Dosen Penguji 2



Mujib Ridwan, S.Kom., M.T.
NIP 198604272014031004

Dosen Penguji 3



Dwi Rolliawati, MT.
NIP 197909272014032001

Dosen Penguji 4



Khalid, M.Kom.
NIP 197906092014031002

Mengetahui,

Dekan Fakultas Sains dan Teknologi
UIN Sunan Ampel Surabaya



Agung Hamdani M.Pd
NIP 196507312000031002

LEMBAR PERNYATAAN PERSETUJUAN PUBLIKASI
KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademika UIN Sunan Ampel Surabaya, yang bertanda tangan di bawah ini, saya:

Nama : AGUNG SUSILO
NIM : H96219037
Fakultas/Jurusan : Sains dan Teknologi/Sistem Informasi
E-mail address : h96219037@student.uinsby.ac.id

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Perpustakaan UIN Sunan Ampel Surabaya, Hak Bebas Royalti Non-Eksklusif atas karya ilmiah :

Sekripsi Tesis Desertasi Lain-lain (.....)
yang berjudul :

ANALISIS SENTIMEN SARA PADA *TWEET* BERBAHASA INDONESIA

MENGGUNAKAN INDOBERT DAN SUPPORT VECTOR MACHINE (SVM)

beserta perangkat yang diperlukan (bila ada). Dengan Hak Bebas Royalti Non-Eksklusif ini Perpustakaan UIN Sunan Ampel Surabaya berhak menyimpan, mengalih-media/format-kan, mengelolanya dalam bentuk pangkalan data (database), mendistribusikannya, dan menampilkan/mempublikasikannya di Internet atau media lain secara **fulltext** untuk kepentingan akademis tanpa perlu meminta ijin dari saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan atau penerbit yang bersangkutan.

Saya bersedia untuk menanggung secara pribadi, tanpa melibatkan pihak Perpustakaan UIN Sunan Ampel Surabaya, segala bentuk tuntutan hukum yang timbul atas pelanggaran Hak Cipta dalam karya ilmiah saya ini.

Demikian pernyataan ini yang saya buat dengan sebenarnya.

Surabaya, 29 Juli 2023

Penulis



(Agung Susilo)
nama terang dan tanda tangan

ABSTRAK

ANALISIS SENTIMEN SARA PADA *TWEET* BERBAHASA INDONESIA MENGGUNAKAN INDOBERT DAN SUPPORT VECTOR MACHINE (SVM)

By:
Agung Susilo

Penelitian ini bertujuan untuk mengetahui sentimen pada *tweets* yang mengandung unsur SARA (Suku, Agama, Ras, dan Golongan) dengan melakukan analisis sentimen pada *tweets* berbahasa Indonesia di Twitter. Dengan demikian, pihak platform media sosial dan pemerintah dapat membatasi penggunaan kata-kata yang mengacu pada unsur SARA yang sering digunakan ketika membagikan *tweets*. Dataset yang digunakan berisi 103.180 *tweets*. Analisis penelitian ini menggunakan metode deep learning, khususnya model IndoBERT yang berbasis pada BERT dengan arsitektur Transformer. Selain itu, model SVM (Support Vector Machine) juga digunakan sebagai pembandingan performa. Hasil penelitian menunjukkan bahwa model IndoBERT memiliki akurasi lebih tinggi dibandingkan model SVM, dengan nilai 78% untuk IndoBERT dan 65% untuk SVM. Selain itu, analisis sentimen menggunakan kedua model tersebut mengungkapkan bahwa *tweets* dengan sentimen positif cenderung menggunakan kata "islam," "jawa," dan "muslim," yang merujuk pada identitas agama dan suku. Sedangkan kata-kata yang cenderung digunakan pada *tweets* dengan sentimen negatif adalah "agama," "jawa," dan "orang," yang juga mengacu pada identitas agama dan suku.

Kata kunci: analisis sentimen, *tweets* SARA, IndoBERT, *deep learning*, SVM

ABSTRACT

SARA SENTIMENT ANALYSIS ON INDONESIAN TWEET USING INDOBERT AND SUPPORT VECTOR MACHINE (SVM)

By:
Agung Susilo

This research aims to determine sentiments in tweets containing elements of SARA (Ethnicity, Religion, Race, and Group) by conducting sentiment analysis on Indonesian-language tweets on Twitter. By doing so, social media platforms and the government can take measures to restrict the use of words that refer to SARA elements, which are commonly used when sharing tweets. The dataset comprises 103,180 tweets, and deep learning, specifically the IndoBERT model based on BERT with a Transformer architecture, is employed for analysis. Additionally, an SVM (Support Vector Machine) model serves as a performance benchmark. The study concludes that IndoBERT outperforms the SVM model, with an accuracy of 78% compared to 65% for SVM. Furthermore, the sentiment analysis reveals that tweets with positive sentiments frequently use words like "islam," "jawa," and "muslim," referring to religious and ethnic identities, while negative sentiments are associated with words like "agama," "jawa," and "orang," which also pertain to religious and ethnic identities.

Keywords: *sentiment analysis, SARA tweets, IndoBERT, deep learning, SVM*

UIN SUNAN AMPEL
S U R A B A Y A

DAFTAR ISI

LEMBAR PERSETUJUAN PEMBIMBING	i
PENGESAHAN TIM PENGUJI SKRIPSI	ii
PERNYATAAN KEASLIAN	iii
KATA PENGANTAR	iv
ABSTRAK	vi
ABSTRACT	vii
DAFTAR ISI	viii
DAFTAR GAMBAR	x
DAFTAR TABEL	xi
BAB I PENDAHULUAN	1
1.1. Latar Belakang	1
1.2. Rumusan Masalah	4
1.3. Batasan Masalah	4
1.4. Manfaat Penelitian	4
1.5. Tujuan Penelitian	5
BAB II KAJIAN PUSTAKA	6
2.1. Tinjauan Penelitian Terdahulu	6
2.2. Dasar Teori	7
2.2.1 Twitter	7
2.2.2 SARA	8
2.2.3 Analisis Sentimen	9
2.2.4 Deep Learning	10
2.2.5 BERT	13
2.2.6 IndoBERT	17
2.2.7 SVM	18
2.2.8 Confusion Matrix	21
2.3. Integrasi Keilmuan	22
BAB III METODE PENELITIAN	24
3.1. Jenis Data	24
3.2. Tahapan Penelitian	24
3.2.1. Perumusan Masalah	24

3.2.2.	Studi Literatur.....	25
3.2.3.	Pengumpulan Data.....	25
3.2.4.	<i>Preprocessing</i>	26
3.2.5.	Pelabelan Data.....	27
3.2.6.	<i>Modelling</i>	29
3.2.7.	<i>Evaluating</i>	31
3.2.8.	Analisis Hasil	31
BAB IV HASIL DAN PEMBAHASAN		32
4.1	Pengumpulan Data.....	32
4.2	<i>Preprocessing</i>	33
4.2.1.	<i>Cleaning</i>	33
4.2.2.	<i>Case Folding</i>	33
4.2.3.	<i>Tokenizing</i>	34
4.2.4.	<i>Normalization</i>	34
4.3	Pelabelan Data	34
4.3.1.	<i>Filtering</i>	35
4.3.2.	<i>Stemming</i>	35
4.3.3.	<i>Labelling</i>	35
4.3.4.	<i>Validation</i>	36
4.4	<i>Modelling</i>	37
4.4.1.	IndoBERT	38
4.4.2.	SVM.....	44
4.5	<i>Evaluating</i>	46
4.6	Analisis Hasil.....	49
BAB V PENUTUP.....		56
5.1	Kesimpulan	56
5.2	Saran.....	56
DAFTAR PUSTAKA		58
LAMPIRAN.....		61

DAFTAR GAMBAR

Gambar 2.1. Proses Deep Learning.	11
Gambar 2.2. Arsitektur Transformer.	14
Gambar 2.3. Representasi Input BERT.	15
Gambar 2.4. Ilustrasi Alur Proses Pre-training dan Fine-tuning.	16
Gambar 2.5. Ilustrasi SVM Linier.....	18
Gambar 2.6. Ilustrasi SVM Non-linier.	20
Gambar 3.1. Alur Penelitian.....	24
Gambar 3.2. Alur Pelabelan Data.	28
Gambar 3.3. Alur <i>Modelling</i>	30
Gambar 4.1. Visualisasi Jumlah Tweet Bulan Februari.....	32
Gambar 4.2. Grafik Akurasi <i>Train Valid IndoBERT Batch Size 16 (5 Epoch)</i>	39
Gambar 4.3. Grafik Akurasi <i>Train Valid IndoBERT Batch Size 16 (10 Epoch)</i> ..	39
Gambar 4.4. Grafik Akurasi Setiap <i>Fold IndoBERT Batch Size 16 (5 Epoch)</i>	40
Gambar 4.5. Grafik Akurasi Setiap <i>Fold IndoBERT Batch Size 16 (10 Epoch)</i> ..	40
Gambar 4.6. Grafik Akurasi <i>Train Valid IndoBERT Batch Size 32 (5 Epoch)</i>	42
Gambar 4.7. Grafik Akurasi <i>Train Valid IndoBERT Batch Size 32 (10 Epoch)</i> ..	42
Gambar 4.8. Grafik Akurasi Setiap <i>Fold IndoBERT Batch Size 32 (5 Epoch)</i>	43
Gambar 4.9. Grafik Akurasi Setiap <i>Fold IndoBERT Batch Size 32 (10 Epoch)</i> ..	43
Gambar 4.10. Grafik Akurasi Setiap <i>Fold SVM</i>	45
Gambar 4.11. Diagram <i>Confusion Matrix IndoBERT Batch Size 16 (5 Epoch)</i> . .	46
Gambar 4.12. Diagram <i>Confusion Matrix IndoBERT Batch Size 16 (10 Epoch)</i> .47	
Gambar 4.13. Diagram <i>Confusion Matrix IndoBERT Batch Size 32 (5 Epoch)</i> . .	47
Gambar 4.14. Diagram <i>Confusion Matrix IndoBERT Batch Size 32 (10 Epoch)</i> .48	
Gambar 4.15. Diagram <i>Confusion Matrix SVM</i>	49
Gambar 4.16. Diagram Jumlah <i>Tweets</i> Berdasarkan Sentimen.	50
Gambar 4.17. <i>Wordcloud</i> Sentimen Positif.	51
Gambar 4. 18, <i>Wordcloud</i> Sentimen Negatif.	53

DAFTAR TABEL

Tabel 2.1. Daftar Penelitian Terdahulu.....	6
Tabel 2.2. Contoh Tabel Confusion Matrix.....	21
Tabel 3.1. Dataset Tweet.....	25
Tabel 3.2. Data Hyperparameter IndoBERT.....	30
Tabel 3.3. Data Parameter SVM.....	31
Tabel 4.1. <i>Tweet</i> Setelah <i>Cleaning</i>	33
Tabel 4.2. <i>Tweet</i> Setelah <i>Case Folding</i>	33
Tabel 4.3. <i>Tweet</i> Setelah <i>Tokenizing</i>	34
Tabel 4.4. <i>Tweet</i> Setelah <i>Normalization</i>	34
Tabel 4.5. <i>Tweet</i> Setelah <i>Filtering</i>	35
Tabel 4.6. <i>Tweet</i> Setelah <i>Stemming</i>	35
Tabel 4.7. <i>Tweet</i> Setelah <i>Labelling</i>	35
Tabel 4.8. Label <i>Tweet</i> oleh InSet Lexicon dan Pakar.....	36
Tabel 4.9. Standar Nilai Kesepakatan. (Landis & Koch, 1977).....	37
Tabel 4.10. <i>Tweet</i> Hasil <i>Tokenizing</i> dan <i>Encoding</i>	38
Tabel 4.11. Parameter yang Digunakan untuk <i>Fine-Tuning</i> IndoBERT.....	38
Tabel 4.12. Data Akurasi <i>Train Valid</i> Model IndoBERT dengan <i>Batch Size</i> 16.....	41
Tabel 4.13. Data Akurasi <i>Train Valid</i> Model IndoBERT dengan <i>Batch Size</i> 32.....	44
Tabel 4.14. Parameter untuk <i>Modelling</i> SVM.....	45
Tabel 4.15. Data Akurasi <i>Training</i> dan <i>Validasi</i> Model SVM.....	45
Tabel 4.16. Tabel Perbandingan Performa Model.....	49
Tabel 4.17. Detail Jumlah <i>Tweets</i>	51
Tabel 4.18. Jumlah Kata Sentimen Positif Berdasarkan Frekuensi Kemunculan.....	51
Tabel 4.19. Jumlah Kata Sentimen Negatif Berdasarkan Frekuensi Kemunculan.....	53

BAB I

PENDAHULUAN

1.1. Latar Belakang

Dewasa ini, informasi teknologi mengalami perkembangan yang signifikan. Dampak dari perkembangan tersebut menghasilkan produk yang bernama internet. Pengguna internet di Indonesia mengalami peningkatan pesat, dalam laporan yang disampaikan oleh *We Are Social* disebutkan bahwa jumlah pengguna internet di Indonesia mencapai 204,7 juta dari 277,7 juta masyarakat Indonesia pada tahun 2022. Jumlah pengguna internet tersebut memiliki selisih 1,03% dengan jumlah pengguna internet pada tahun 2021 dan 54,25% dengan jumlah pengguna internet pada tahun 2018. (We Are Social, 2021).

Media sosial menjadi salah satu hasil berkembangnya internet yang sangat pesat. Media sosial adalah media digital yang memberikan layanan kepada pengguna sehingga dapat saling berbagi informasi dan komunikasi dengan menggunakan jaringan internet. Informasi yang dibagikan oleh pengguna beraneka ragam juga bisa bermakna positif maupun negatif. Terdapat beragam media sosial seperti Twitter, Instagram, dan Facebook. Twitter merupakan salah satu media sosial dengan jumlah pengguna yang tinggi yaitu 18,45 juta pengguna pada tahun 2022. Artinya pengguna Twitter naik 31,3% dibandingkan tahun 2021. Selain itu Indonesia menjadi peringkat 6 sebagai pengguna Twitter terbanyak. (Kemp, 2023).

Dalam berbagi informasi di Twitter atau dapat disebut sebagai tweet terdapat etika yang perlu diperhatikan salah satunya yaitu larangan penyematan unsur SARA (suku, agama, ras, antar golongan) dalam berbagi informasi. Tweet dengan unsur SARA dilarang karena memiliki dampak negatif bagi kesehatan media sosial seperti timbulnya rasa benci hingga memicu perpecahan antar kelompok. Terlebih tweet yang disebarakan merupakan informasi palsu atau hoax dapat mengakibatkan dampak yang lebih parah. (Christine, 2018). Untuk membatasi tweet yang mengandung unsur SARA pihak Twitter dapat memberi peringatan pada postingan yang mengandung unsur SARA. Selain itu pemerintah dapat membuat peraturan tentang penyebaran informasi yang mengandung SARA.

Namun tidak semua tweet dengan unsur SARA bermakna negatif, ada juga tweet dengan unsur SARA yang memiliki makna positif. (Asrif, 2020). Untuk itu diperlukan analisis sentimen terhadap tweet untuk mengetahui apakah tweet tersebut SARA positif, SARA negatif, atau SARA netral.

Analisis sentimen merupakan jenis topik penelitian dalam *Text Mining* yang berfokus pada klasifikasi sentimen negatif dan positif terhadap objek tertentu seperti layanan, produk, isu, dan peristiwa. (Maulana et al., 2018). Penerapan analisis sentimen pada Twitter dapat mempermudah dalam menilai serta mengukur jenis sentimen pada tweet. Dalam sisi sosial, analisis sentimen pada Twitter digunakan untuk mengukur dan mengklasifikasi tweet yang mengandung ujaran kebencian di Twitter sehingga pemerintah dengan mudah melakukan penyaringan tweet yang berpotensi menimbulkan perpecahan. (Badjatiya et al., 2017). Selain itu pada sisi politik, sentimen analisis pada Twitter digunakan untuk mengukur sentimen masyarakat terhadap calon Pemilihan Presiden sehingga calon dapat mengetahui pandangan masyarakat terhadap calon tersebut. (Juanita, 2020). Tidak hanya itu, analisis sentimen pada Twitter juga digunakan pada sisi bisnis yaitu dengan mengukur review penonton terhadap film. (Lu & Wu, 2019).

Metode *machine learning* biasa digunakan dalam melakukan analisis sentiment. Metode *machine learning* terbagi menjadi dua yaitu metode *classic machine learning* dan metode *advanced machine learning* atau juga dikenal sebagai *deep learning*. Namun metode *deep learning* dinilai memiliki hasil yang lebih baik pada segi performa dibandingkan *classic machine learning*. Hal tersebut dikarenakan proses yang dimiliki *deep learning* tidak menggunakan algoritma statistika dasar melainkan algoritma artificial neural network (ANN). (Goodfellow et al., 2016).

Salah satu model *deep learning* yang sedang populer adalah Bidirectional Encoder Representations from Transformers (BERT). Keunggulan model BERT terletak pada kemampuan adaptasi dalam tugas *machine learning* khususnya Text Mining seperti *question answering* dan klasifikasi dengan memberi tambahan satu layer baru. Hal tersebut dikarenakan model BERT merupakan model pre-trained yang telah melakukan *training* secara dua arah terhadap data teks yang tidak berlabel atau disebut dengan *deep bidirectional*. (Devlin et al., 2019). Saat ini

model BERT telah dikembangkan untuk bisa memproses beragam bahasa salah satunya bahasa Indonesia. Salah satu hasil pengembangan model BERT dengan dukungan bahasa Indonesia yaitu model IndoBERT. (Wilie et al., 2020).

Masalah yang terjadi saat ini adalah bagaimana menentukan serta mengukur suatu kelas sentimen terhadap tweet SARA. Analisis sentimen terhadap tweet SARA perlu dilakukan untuk mengetahui sentimen pada tweet yang memiliki unsur SARA. Selain itu kebanyakan penelitian-penelitian terdahulu masih menggunakan metode classic machine learning seperti Support Vector Machine (SVM) dalam melakukan analisis sentimen SARA terhadap tweet dengan hasil yang kurang. Pada tahun 2022, Eka Karyawati dan temannya melakukan penelitian dengan topik yang sama berjudul "*Comparison of SVM and LIWC for Sentiment Analysis of SARA*", penelitian tersebut menghasilkan performa yang kurang memuaskan. (Karyawati et al., 2022). Penerapan deep learning dalam analisis sentimen SARA dirasa mampu menghasilkan akurasi yang lebih optimal seperti yang dijelaskan Norah Naver pada penelitiannya di tahun 2021 judul "*The past, present or future? : A comparative NLP study of Naive Bayes, LSTM and BERT for classifying Swedish sentences based on their tense*". Penelitian itu dilakukan untuk membandingkan akurasi dari ketiga model machine learning yaitu Naïve Bayes, BERTbase, dan LSTM untuk melakukan klasifikasi teks berbahasa Swedia dengan berdasarkan kosakatanya. Hasil penelitian tersebut menyatakan bahwa model BERTbase memiliki akurasi tertinggi yaitu sebesar 96,3%, diikuti dengan model LSTM dengan nilai akurasi sebesar 88,6% dan model Naive Bayes dengan nilai akurasi sebesar 78,3%. (Navér, 2021). Dalam hasil penelitian yang dilakukan oleh Ghinaa Zain Nabiilah, Simeon Yuda Prasetyo, dan yang lain pada tahun 2022 dengan judul "*BERT Base Model for Toxic Comment Analysis on Indonesian Social Media*" menyebutkan bahwa IndoBERT memiliki hasil yang lebih optimal dibandingkan dengan MBERT dan Indo RoBERTa dalam melakukan klasifikasi komentar toxic di media sosial. (Nabiilah et al., 2023). Selain itu, Herlina Jayadianti dan teman-temannya melakukan penelitian pada tahun yang sama dengan judul "*Analisis Sentimen Review Indonesia Menggunakan Fine-Tuning IndoBERT dan R-CNN*" juga menghasilkan akurasi 95,16% dengan menggabungkan IndoBERT dengan R-

CNN untuk analisis sentimen. (Jayadianti et al., 2022). Maka dari itu penelitian ini menyarankan untuk menerapkan salah satu model dari metode deep learning yaitu IndoBERT kemudian dikomparasi dengan metode terdahulu yaitu SVM dalam melakukan analisis sentimen SARA terhadap tweet.

1.2. Rumusan Masalah

Masalah yang diangkat dalam penelitian ini disebutkan dalam rumusan masalah sebagai berikut:

1. Bagaimana penerapan sentimen analisis SARA pada tweet berbahasa Indonesia menggunakan IndoBERT dan SVM?
2. Bagaimana hasil komparasi model IndoBERT dan SVM dalam analisis sentimen SARA pada tweet berbahasa Indonesia?

1.3. Batasan Masalah

Batasan masalah digunakan untuk memperjelas arah pembahasan dalam penelitian ini:

1. Penelitian ini menggunakan dataset tweet berbahasa Indonesia dengan kata kunci “buruh”, “lgbt”, “feminis”, “pki”, “komunis”, “khalifah”, “bule”, “pribumi”, “golongan”, “kaum”, “etnis”, “jawa”, “madura”, “cina”, “batak”, “ambon”, “arab”, “sipit”, “suku”, “agama”, “islam”, “kristen”, “hindu”, “budha”, “muslim” yang di-*crawling* pada tanggal 1 Februari 2023 hingga 28 Februari 2023.
2. Fokus dari penelitian ini yaitu perbandingan akurasi model IndoBERT dengan model SVM dan dilanjutkan dengan sentiment analisis SARA pada *tweet* berbahasa Indonesia menggunakan IndoBERT.

1.4. Manfaat Penelitian

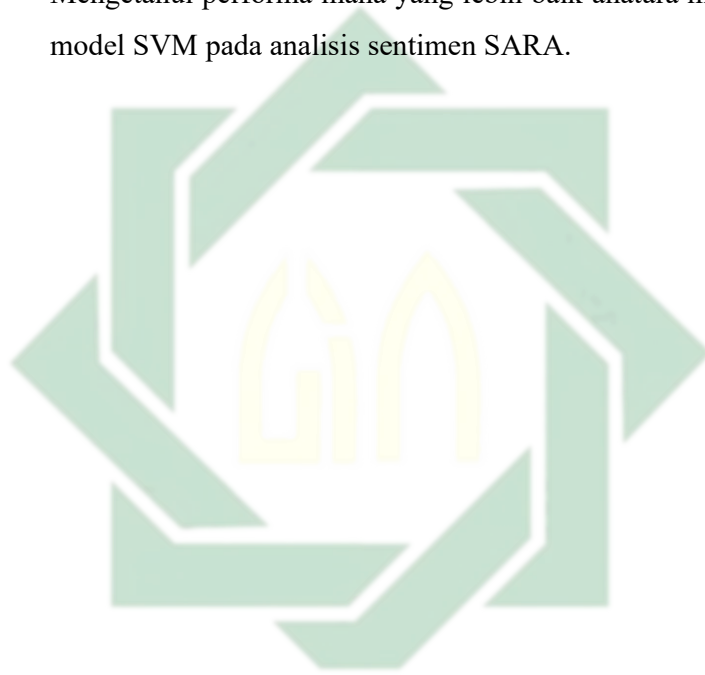
Penelitian ini diharapkan memberikan hasil yang bermanfaat bagi berbagai macam pihak yang terkait, diantaranya:

1. Manfaat Akademis
Penelitian ini dapat memberikan wawasan mengenai IndoBERT dan SVM dalam melakukan prediksi label pada *tweet* dengan unsur.

2. Manfaat Praktis
Sebagai rujukan penelitian masa depan yang dapat membantu pemerintah dalam analisa sentimen pada *tweet* masyarakat yang mengandung SARA.

1.5. Tujuan Penelitian

1. Mengetahui proses analisis sentimen SARA pada tweet menggunakan IndoBERT dan SVM.
2. Mengetahui performa mana yang lebih baik antara model IndoBERT dan model SVM pada analisis sentimen SARA.



UIN SUNAN AMPEL
S U R A B A Y A

BAB II

KAJIAN PUSTAKA

2.1. Tinjauan Penelitian Terdahulu

Dalam upaya memberikan pemahaman dan interpretasi yang lebih detail mengenai hubungan antara penelitian yang akan dilakukan dengan penelitian terdahulu yang memiliki relevansi. Informasi yang lebih lengkap ditunjukkan pada Tabel 2.1.

Tabel 2.1. Daftar Penelitian Terdahulu.

No	Judul	Metode	Hasil
1	Comparison of SVM and LIWC for Sentiment Analysis of SARA. (Karyawati et al., 2022)	SVM dan LIWC	Skor hasil evaluasi terbaik dimiliki oleh model SVM dengan tugas klasifikasi SARA negatif yaitu skor <i>precision</i> sebesar 0.9, skor <i>recall</i> sebesar 0.6, dan <i>F1-Score</i> sebesar 0.72. Skor hasil evaluasi terbaik dimiliki oleh model LIWC dengan tugas klasifikasi SARA positif yaitu skor <i>precision</i> 0.6, skor <i>recall</i> sebesar 0.8, dan <i>F1-Score</i> sebesar 0.69. Skor hasil evaluasi terbaik dimiliki oleh model SVM dengan tugas klasifikasi SARA netral yaitu skor <i>precision</i> sebesar 0.52, skor <i>recall</i> sebesar 0.87, dan <i>F1-Score</i> sebesar 0.65.
2	Implementation of BERT, IndoBERT, and CNN-LSTM in Classifying Public Opinion about COVID - 19 Vaccine in Indonesia. (Saadah et al., 2022)	BERT, IndoBERT, dan CNN-LSTM	IndoBERT berhasil mengklasifikasikan sentimen menjadi sentimen positif sekitar 80%, kemudian IndoBERTweet pada 68 %, diikuti dengan CNN-LSTM mencapai 53% dengan total penggunaan dataset tahun 2020 dari Twitter.
3	BERT base model for toxic comment analysis on Indonesian social media. (Nabiilah et al., 2023)	MBERT, IndoBERT, dan Indo RoBERTa Small	Dari hasil evaluasi ketiga model BERT, hasil optimal dimiliki oleh model IndoBERT dengan F1 Score sebesar 0,8897.
4	Sentiment analysis of Indonesian reviews using fine-tuning IndoBERT	IndoBERT dan R-CNN	IndoBERT dengan Recurrent Convolutional Neural Network (RCNN) memperoleh akurasi

No	Judul	Metode	Hasil
	and R-CNN. (Jayadianti et al., 2022)		95,16%, presisi 94,05%, recall 92,74% dan skor f1 93,27%.
5	Personality Detection on Twitter Social Media Using IndoBERT Method. (Rohmah & Maharani, 2022)	IndoBERT	Hasil dari penelitian ini diketahui bahwa penggunaan metode IndoBERT tanpa proses stemming untuk mendeteksi 5 kepribadian umum memiliki tingkat akurasi sebesar 0,46.
6	A comparative NLP study of Naive Bayes, LSTM and BERT for classifying Swedish sentences based on their tense. (Navér, 2021).	Naive Bayes, LSTM, dan BERT	Dengan dataset tanpa melalui proses lemmatization menghasilkan model dengan kinerja terbaik yaitu BERT dengan nilai accuracy sebesar 96,3% Diikuti dengan LSTM dengan nilai accuracy sebesar 88,6%. Dan model terburuk dimiliki oleh Naive Bayes dengan nilai accuracy sebesar 78,3%.

Berdasarkan data tinjauan penelitian terdahulu pada tabel 2.1, ditunjukkan bahwa metode *deep learning* menghasilkan skor evaluasi yang lebih baik dalam kasus analisis sentimen dibandingkan dengan *traditional machine learning*. Terdapat beberapa model deep learning yang cukup sering digunakan dalam penelitian terdahulu yaitu LSTM, BERT, dan IndoBERT, namun dari ketiga model tersebut, BERT dan IndoBERT memiliki performa dan akurasi yang lebih baik dibandingkan LSTM. Pemilihan model IndoBERT didasarkan pada performa dan akurasi yang dihasilkan dari penelitian-penelitian terdahulu dengan kasus analisis sentimen berbahasa Indonesia.

2.2. Dasar Teori

2.2.1 Twitter

Twitter merupakan salah satu media sosial yang sangat populer dan telah berhasil menarik banyak pengguna aktif sejak pertama kali dirilis pada tahun 2006, mengalami pertumbuhan yang signifikan hingga saat ini. Dengan memungkinkan pengguna untuk membagikan pemikiran, berita, dan informasi melalui pesan teks terbatas hingga 280 karakter yang dikenal sebagai "*tweet*",

Twitter memberikan sarana efektif bagi para pengguna untuk berinteraksi dan berpartisipasi dalam berbagai diskusi publik (Twitter, 2023).

Salah satu manfaat utama yang diberikan oleh Twitter adalah memberikan platform yang memungkinkan penggunanya untuk berbagi informasi secara cepat dan luas. Hal ini tidak hanya meningkatkan visibilitas dan eksposur dari individu atau organisasi, tetapi juga memungkinkan terbentuknya jaringan dan koneksi dengan orang lain yang memiliki minat atau tujuan yang serupa. Dalam lingkungan bisnis, Twitter juga berperan penting sebagai alat promosi yang efektif, memungkinkan pelaku usaha untuk mencapai audiens yang lebih luas dan berinteraksi langsung dengan konsumen potensial.

Selain itu, Twitter juga menjadi sarana yang sangat berguna untuk melacak dan memantau perkembangan terkini dalam berbagai bidang, seperti politik, bisnis, dan teknologi. Dengan banyaknya tokoh publik, perusahaan besar, dan media massa yang aktif di platform ini, informasi terbaru dapat dengan mudah diakses dan diikuti oleh para pengguna, menjadikan Twitter sebagai sumber berita yang relevan dan up-to-date. (Comm et al., 2015).

Namun, seperti halnya media sosial lainnya, pengguna Twitter perlu berhati-hati dalam menghadapi beberapa tantangan dan risiko yang mungkin timbul. Salah satu masalah yang sering muncul adalah penyebaran informasi yang salah atau tidak akurat. Keterbatasan karakter pada setiap tweet bisa menyebabkan informasi menjadi terpotong-potong dan kurang lengkap, sehingga terdapat potensi untuk terjadinya penyebaran berita palsu atau informasi yang tidak diverifikasi secara benar. Oleh karena itu, pengguna Twitter diharapkan bijaksana dalam memvalidasi kebenaran informasi sebelum membagikannya kepada audiens

2.2.2 SARA

SARA merupakan singkatan dari Suku, Agama, Ras, dan Antar Golongan, yang secara khusus mengacu pada isu-isu yang berkaitan dengan identitas dan seringkali menjadi sumber permasalahan sosial di Indonesia. Permasalahan SARA di Indonesia sangat kompleks dan mencakup berbagai aspek kehidupan masyarakat, termasuk dalam konteks media sosial seperti Twitter. Fenomena SARA pada platform ini, khususnya dalam bentuk cyberbullying, menimbulkan

dampak negatif yang serius bagi penerima pesan, dan menjadi hal yang harus diperhatikan secara serius untuk menciptakan lingkungan daring yang lebih aman dan inklusif.

Salah satu bentuk SARA yang kerap muncul di Twitter adalah cyberbullying. Cyberbullying adalah perilaku agresif dan merendahkan yang dilakukan secara elektronik melalui media sosial dan platform daring lainnya. Terlepas dari karakteristik dan bentuknya, cyberbullying berbasis identitas seperti SARA dapat menimbulkan dampak psikologis yang serius pada penerimanya. Korban cyberbullying cenderung merasa terisolasi, tidak aman, dan mengalami tingkat stres yang tinggi. Hal ini dapat mempengaruhi kesehatan mental mereka dan bahkan mengurangi kepercayaan diri serta kualitas hidup secara keseluruhan.

Penting untuk menyadari bahwa cyberbullying berbasis SARA juga bisa berdampak lebih luas pada masyarakat. Ketika pesan-pesan berisi SARA menyebar di platform seperti Twitter, dapat memicu konflik sosial, memperkuat stereotip negatif, dan memperburuk divisi sosial di masyarakat. Menciptakan lingkungan di media sosial yang bebas dari SARA dan cyberbullying menjadi tanggung jawab bersama, baik oleh perusahaan penyedia platform seperti Twitter, pemerintah, maupun pengguna aktifnya.

Dalam mengatasi permasalahan SARA di Twitter, kolaborasi antara pihak-pihak terkait sangatlah penting. Perusahaan Twitter, pemerintah, masyarakat sipil, dan pengguna aktif harus bekerjasama untuk mengembangkan strategi dan inisiatif yang efektif dalam menciptakan lingkungan daring yang aman, inklusif, dan bebas dari SARA. Upaya bersama ini akan memberikan kontribusi nyata dalam menciptakan media sosial yang lebih positif, berdampak baik, dan mendukung pembangunan masyarakat yang harmonis di Indonesia.. (Christine, 2018).

2.2.3 Analisis Sentimen

Analisis Sentimen adalah suatu proses pemodelan dan pengenalan pola dalam teks data yang memiliki tujuan untuk mengidentifikasi dan memahami emosi, perasaan, dan pendapat individu dalam teks. Tujuannya adalah untuk menentukan apakah suatu pernyataan atau teks memiliki sentimen positif, negatif,

atau netral. Ini sering digunakan dalam berbagai aplikasi seperti pemantauan merek, analisis pasar, dan pemasaran viral. (Liu, 2012).

Analisis Sentimen sangat penting dalam era digital saat ini, di mana banyak informasi dan opini dapat dengan mudah dibagikan melalui media sosial. Ini membantu perusahaan dan individu memahami perasaan dan pandangan masyarakat terhadap produk, layanan, dan topik tertentu, yang pada akhirnya dapat membantu mereka mengambil keputusan yang lebih informatif dan tepat. (Marshan et al., 2021).

Metode Sentimen Analisis meliputi beberapa teknik, seperti analisis frekuensi kata, berbasis leksikon, dan berbasis pembelajaran mesin. Analisis frekuensi kata menghitung jumlah kata positif dan negatif dalam sebuah teks dan membandingkannya untuk menentukan sentimen. Berbasis leksikon menggunakan daftar kata yang terkait dengan sentimen tertentu dan memilih kata-kata dalam teks untuk menentukan sentimen. *Machine learning-based* menggunakan algoritma seperti Naive Bayes, SVM, LSTM, dan BERT untuk mempelajari pola sentimen dalam data dan memprediksi teks sentimen baru. (Srivastava et al., 2022).

2.2.4 Deep Learning

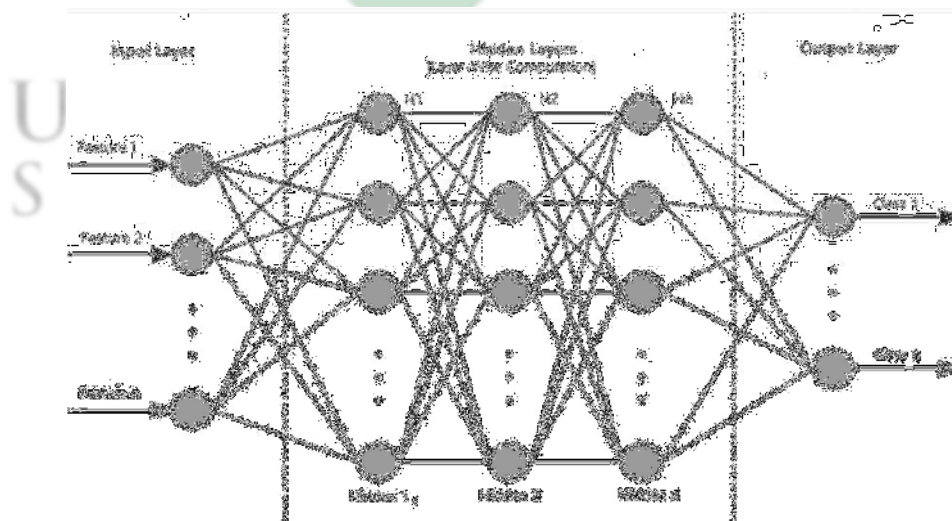
Deep learning merupakan salah satu pendekatan dalam bidang machine learning yang telah menunjukkan keunggulan dalam mengatasi masalah kompleks dan abstrak. Metode ini didasarkan pada upaya untuk meniru cara kerja jaringan saraf pada otak manusia, sehingga dikenal sebagai jaringan saraf tiruan atau neural network. Dalam deep learning, jaringan neural dikonstruksi dengan struktur yang sangat dalam, memiliki banyak lapisan node (neuron) yang terhubung secara kompleks, dan mampu memproses input data menjadi output dengan tingkat representasi informasi yang semakin kompleks dan abstrak. (Deng, 2014).

Teknisnya, deep learning menggunakan arsitektur jaringan neural yang terdiri dari beberapa lapisan. Pertama, ada input layer, yang merupakan lapisan pertama dari jaringan neural dan berfungsi untuk menerima data input. Data input dikirimkan dari node-node di lapisan ini ke lapisan selanjutnya, yaitu hidden layer. Hidden layer adalah bagian inti dari deep learning, karena di dalamnya

terdapat algoritma kompleks yang bertugas untuk memproses data dan membentuk representasi fitur yang semakin kompleks dan abstrak. Setiap lapisan hidden ini secara berurutan membangun tingkat representasi fitur yang lebih tinggi dari data input, sehingga semakin dalam jaringan, semakin abstrak dan kompleks representasi fitur yang dihasilkan.

Keberhasilan deep learning dalam memproses data dan membentuk representasi fitur yang kompleks dan abstrak terletak pada banyaknya hidden layer yang saling terhubung dan jumlah neuron dalam masing-masing layer tersebut. Semakin banyak hidden layer dalam jaringan dan semakin besar jumlah neuron dalam masing-masing lapisan, semakin tinggi performa dan kapabilitas model deep learning dalam menyelesaikan tugas-tugas yang kompleks (Goodfellow et al., 2016). Dalam beberapa kasus, jaringan deep learning bahkan memiliki ratusan atau ribuan hidden layer, yang memungkinkan mereka untuk mengekstraksi dan memahami pola yang sangat kompleks dalam data.

Satu lagi lapisan yang penting dalam jaringan deep learning adalah output layer. Output layer bertanggung jawab untuk menghasilkan output akhir dari jaringan setelah proses representasi fitur yang kompleks di lapisan-lapisan sebelumnya. Output ini berupa kesimpulan atau hasil dari pemrosesan data dan representasi fitur, dan dapat divisualisasikan untuk memahami hasil dari proses yang telah dilakukan oleh jaringan neural. (Sarker, 2021).



Gambar 2.1. Proses Deep Learning.

(Hagan et al., 2014)

Dalam dunia machine learning, terutama dalam konteks deep learning, proses pembentukan model merupakan bagian penting dari proses latihan yang memerlukan dataset yang besar. Untuk membangun model deep learning, jaringan neural dikonstruksi dengan banyak layer, dan selanjutnya jaringan ini akan dijalankan pada dataset yang besar untuk belajar dan menyesuaikan diri dengan pola yang ada dalam data tersebut. Proses ini melibatkan perbandingan hasil yang dihasilkan oleh jaringan dengan target yang diinginkan, dan kemudian bobot pada setiap node dalam jaringan disesuaikan secara berulang-ulang untuk mengurangi perbedaan antara hasil prediksi dan target yang diinginkan. (Goodfellow et al., 2016).

Proses latihan jaringan deep learning memerlukan banyak iterasi dan komputasi yang kompleks, terutama pada dataset yang besar. Sebagai hasilnya, waktu yang dibutuhkan untuk melatih model bisa menjadi sangat lama jika hanya menggunakan central processing unit (CPU) biasa. Oleh karena itu, untuk mempercepat proses ini, biasanya digunakan graphic processing unit (GPU). GPU merupakan unit pemrosesan grafis yang memiliki kemampuan paralel yang tinggi, yang memungkinkannya untuk menjalankan banyak operasi matematika secara bersamaan. Hal ini sangat menguntungkan dalam konteks deep learning, karena memungkinkan pelatihan model dilakukan dengan lebih cepat dan efisien. (Danukusumo, 2017).

Keuntungan menggunakan GPU dalam proses pembentukan model deep learning tidak hanya terbatas pada peningkatan kecepatan, tetapi juga pada kapabilitasnya dalam menangani tugas-tugas paralel yang rumit. Struktur jaringan deep learning dengan banyak layer dan neuron memungkinkan pelaksanaan banyak operasi matematika secara paralel dalam satu waktu, yang bisa dikerjakan dengan lebih efisien oleh GPU daripada oleh CPU. Sebagai contoh, saat memproses gambar, proses pelatihan jaringan deep learning melibatkan banyak operasi matriks dan konvolusi, dan ini dapat dijalankan dengan sangat efisien oleh GPU karena kemampuannya dalam paralelisme.

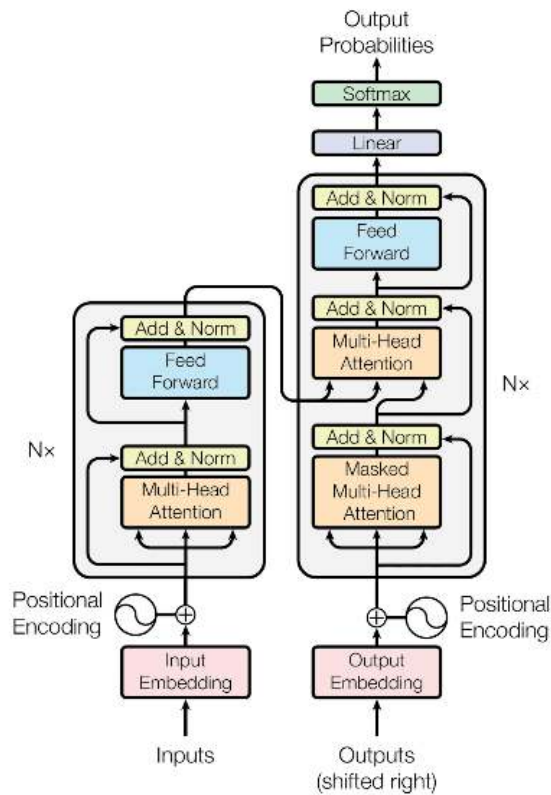
Selain GPU, terdapat pula field-programmable gate array (FPGA) dan application-specific integrated circuit (ASIC) yang juga dapat digunakan dalam deep learning untuk meningkatkan performa dan efisiensi. FPGA dan ASIC

adalah jenis perangkat keras yang dirancang khusus untuk tugas-tugas tertentu, sehingga mereka dapat memberikan kinerja yang lebih tinggi daripada CPU dan GPU dalam konteks yang sesuai. Namun, penggunaan FPGA dan ASIC dalam deep learning memerlukan penyesuaian khusus untuk tugas yang hendak dijalankan, dan ini dapat menjadi lebih rumit dan memerlukan waktu lebih lama untuk implementasinya dibandingkan dengan GPU. (Chen et al., 2018)

2.2.5 BERT

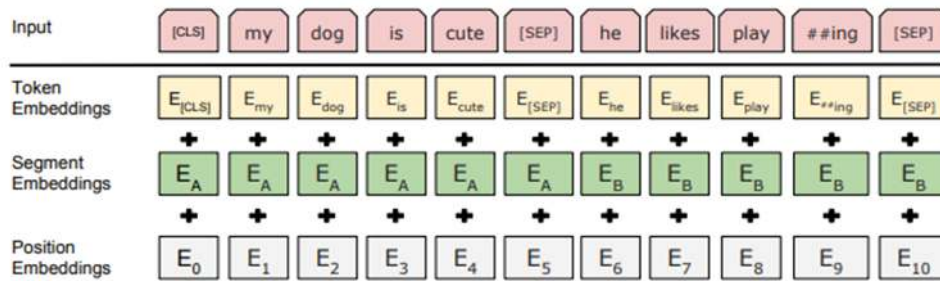
Bidirectional Encoders Representations from Transformers (BERT) merupakan salah satu model *machine learning* Natural Language Processing (NLP) yang dikembangkan oleh Google pada 11 Oktober 2018. Dalam BERT, teks di encode dengan menggunakan mekanisme *bidirectional self-attention* yang memproses data teks dari dua arah yaitu kanan ke kiri dan sebaliknya. Sehingga model BERT mampu melakukan *fine-tuning* dengan tambahan satu layer saja. Hal tersebut membuat model BERT memiliki pemahaman konteks yang lebih baik dan dapat memahami makna kata dalam konteks yang lebih luas. (Devlin et al., 2019).

BERT dikembangkan berdasarkan penggabungan antara arsitektur Transformer *encoder* dengan pemrosesan dua arah atau *bidirectional*. Setiap *layer* pada *encoder* memiliki dua *sub-layer* yaitu *multi-head self-attention mechanism* dan *fully connected feedforward network*. Setiap *input* pada *encoder* akan melalui *self-attention layer*. *Layer* ini dapat mempermudah encoder dalam memahami konteks lebih luas. *Output* dari *self-attention layer* diteruskan ke *feedforward neural network* sebagai *input*.



Gambar 2.2. Arsitektur Transformer.
(Devlin et al., 2019)

Representasi vector kata dalam BERT memakai WordPiece *embeddings* yang berisi 30.000 token kosa kata. Awal pengembangan WordPiece ditujukan untuk dapat memperbaiki masalah penyebab kehilangan kata pada *vocabulary* atau Out of Vocabulary (OOV) ketika segmentasi kata pada Google *speech recognition system* dengan bahasa Jepang dan Korea. (Wu et al., 2016). Maka WordPiece perlu dilatih untuk bisa memproses setiap unit kata dalam satu dataset besar. Unit kata yang memiliki frekuensi penggunaan paling sering digunakan akan disimpan secara penuh sedangkan unit kata dengan frekuensi penggunaan yang sedikit akan dipecah ke bentuk sub-kata hingga ke bentuk karakter.



Gambar 2.3. Representasi Input BERT.

(Devlin et al., 2019)

Token pada posisi pertama di setiap kalimat merupakan token *special class classification* ([CLS]). Pada akhir kalimat ditambahkan token *separate* ([SEP]) sebagai pemisah untuk membedakan antar kalimat. Selain itu, terdapat cara lain dalam membedakan kalimat yaitu dengan menambahkan *learning embedding* ke setiap token sehingga kalimat A dan B dapat dikenali dan dibedakan. *Input* pada *embedding* merupakan jumlah dari *token embedding*, *segment embedding*, dan *position embedding*.

Dalam model BERT terdapat dua tahap yaitu *pre-training* dan *fine-tuning*. Tahap *pre-training* bertujuan untuk menentukan representasi kata dan konteks dalam teks melalui proses Mask Language Modelling (MLM) dan proses Next Sentence Prediction (NSP). MLM membuat representasi dalam penggabungan konteks dari kiri dan kanan sehingga *transformator* dapat terlatih secara dua arah. NSP menggabungkan representasi tiap pasangan teks *pre-training*.

a. Masked Language Modelling (MLM)

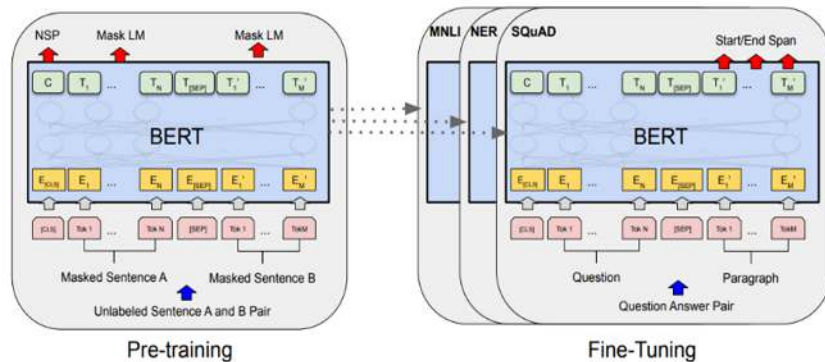
Dalam Masked Language Modeling (MLM), sebagian dari kata dalam kalimat menjadi kalimat yang ditutupi untuk kemudian dilakukan prediksi oleh model pada kalimat yang ditutupi. Proses MLM membantu model dalam memahami konteks kata secara luas pada kalimat.

Pada MLM, setiap kalimat dipilih 15% kata acak untuk kemudian diganti dengan token [MASK]. Selanjutnya model memprediksi kosa kata terhadap kata asli sesuai dengan konteks pada kalimat. *Layer classification* ditambahkan diatas *output encoder* untuk menjalankan prediksi. Kemudian vektor *output* dikalikan dengan *embedding* matriks dan diubah ke dimensi *vocabulary* untuk dihitung probabilitas dengan *softmax*.

b. Next Sentence Prediction (NSP)

Dalam Next Sentence Prediction (NSP), model diberikan *input* berupa dua kalimat dan diminta melakukan prediksi untuk mengetahui apakah kalimat kedua adalah kalimat berikutnya dari kalimat pertama atau tidak. NSP dapat mempermudah model memahami hubungan antar kalimat dalam teks dan memperkuat representasi kalimat. Representasi kalimat ini kemudian digunakan sebagai dasar untuk melakukan pemrosesan teks selanjutnya, seperti klasifikasi teks atau penandaan posisi.

Pada NSP, model dilatih untuk memprediksi label pada pasangan kalimat baru yang tidak pernah dilihat sebelumnya. Ketika model melakukan *training* data, pasangan kalimat diterima oleh model sebagai *input* untuk dilatih sehingga model mampu memprediksi hubungan antara kalimat kedua dengan kalimat pertama. Dalam proses training, 50% dari *input* merupakan pasangan kalimat asli. Sedangkan 50% kalimat lain diambil secara acak dari corpus sebagai kalimat kedua.



Gambar 2.4. Ilustrasi Alur Proses Pre-training dan Fine-tuning.

(Devlin et al., 2019)

Pelatihan model bahasa diperlukan tahap *fine-tuning* yaitu tahap pelatihan dengan melakukan beberapa perubahan pada model BERT. *Fine-tuning* dapat dilakukan dengan mudah karena arsitektur Transformer dalam model BERT memiliki mekanisme *self-attention*, sehingga model BERT dapat digunakan untuk melakukan berbagai tugas pada kalimat tunggal maupun kalimat berpasangan dengan menukar *input* dan *output* yang sesuai. Gambar 2.5 menunjukkan ilustrasi proses *pre-training* yang dilanjutkan dengan proses *fine-tuning*.

Selama tahap *fine-tuning*, *hyperparameter* yang digunakan sama seperti tahap *pre-training* kecuali untuk *batch size*, *learning rate*, dan jumlah *epoch* yang disesuaikan. *Dropout probability* selalu diatur pada angka 0.1. Meskipun nilai optimal untuk setiap tugas di bidang Natural Language Processing (NLP) bervariasi, namun model yang digunakan memberikan hasil yang cukup baik dengan *batch size* 16 dan 32, *learning rate* dengan optimasi Adam $5e-5$, $3e-5$, dan $2e-5$, serta *epoch* yang berjumlah 2, 3, atau 4 kali. (Devlin et al., 2019).

2.2.6 IndoBERT

IndoBERT adalah varian dari model BERT yang dikembangkan khusus untuk bahasa Indonesia. Model ini memiliki arsitektur yang sama dengan BERT, namun memiliki corpus teks dalam bahasa Indonesia yang besar sehingga memiliki representasi konsep dan konteks yang lebih baik dibandingkan BERT standar.

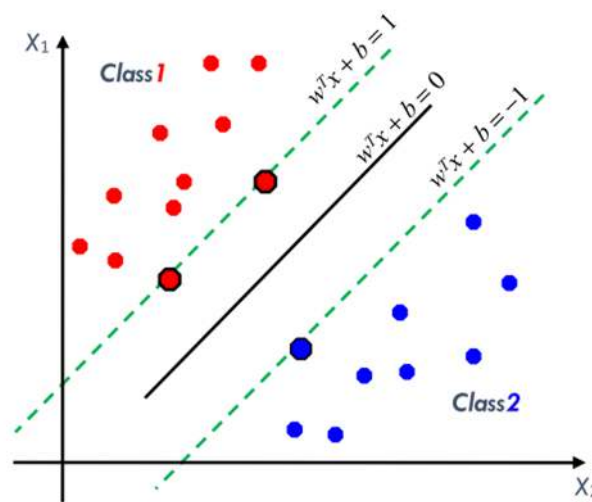
Teknik *fine-tuning* digunakan dalam IndoBERT untuk menyesuaikan diri dengan domain dan jenis teks dalam bahasa Indonesia. Penyempurnaan membuat model ini dapat memahami konteks dan membedakan makna kata-kata dalam bahasa Indonesia secara lebih baik dibandingkan standar BERT. Hal ini membuat IndoBERT memiliki kemampuan yang lebih baik dalam melakukan tugas *Natural Language Processing* (NLP) seperti penandaan posisi, klasifikasi teks, dan pemahaman bahasa alami. (Koto et al., 2020).

Dalam prosesnya, IndoBERT dilakukan *pre-training* dengan dataset Indo4B yang memiliki ukuran 23 GB data teks berbahasa Indonesia yang terdiri dari 4 miliar kata. Data teks tersebut berisi kata formal dan non-formal yang diambil dari berbagai sumber yaitu berita, media sosial, blog, dan website. Berbeda dengan BERT yang menggunakan WordPiece dalam pembentukan *vocabulary*. SentencePiece digunakan dalam IndoBERT dengan Byte Pair Encoding (BPE) *tokenizer* sebagai metode penyusunan *vocabulary*. SentencePiece merupakan algoritma *subword tokenizer* atau tokenisasi kata yang dibangun untuk memahami konteks pada kalimat yang kompleks. IndoBERT terbagi menjadi empat jenis model yang masing-masing memiliki kelemahan dan kelebihan yaitu IndoBERT_{BASE}, IndoBERT_{LARGE}., IndoBERT-lite_{BASE}, IndoBERT-lite_{LARGE}.

Semua model IndoBERT dilatih dengan *chipset* TPU v3-8 dalam dua tahap. Pada tahap awal, model dilatih menggunakan *maximum sequence length* 128. Tahap selanjutnya *maximum sequence length* diubah dari 128 ke 512 dan dilakukan pelatihan kembali. Dalam segi akurasi, IndoBERT_{LARGE} mendapat skor akurasi lebih tinggi dibanding dengan jenis model IndoBERT lain, namun pemrosesan menjadi lebih berat dan lama dikarenakan besarnya memori yang dibutuhkan. (Wilie et al., 2020).

2.2.7 SVM

SVM (Support Vector Machine) adalah salah satu algoritma *machine learning* yang digunakan untuk melakukan klasifikasi atau regresi pada data. Pertama kali SVM diperkenalkan oleh Vapnik pada tahun 1992. SVM bekerja dengan mencari sebuah *hyperplane* yaitu batas yang memisahkan dua kelas data berbeda secara optimal. Hyperplane yang optimal dipilih berdasarkan jarak terdekat dari setiap kelas data ke hyperplane tersebut. Oleh karena itu, SVM juga dikenal sebagai algoritma pembelajaran dengan dukungan vektor (Support Vector Learning). (Gholami & Fakhari, 2017)



Gambar 2.5. Ilustrasi SVM Linier.

(Gholami & Fakhari, 2017)

Terdapat dua tipe metode SVM yaitu SVM Linier dan SVM Non-Linier. SVM linier merupakan metode SVM yang menangani data linier, yaitu data yang

terbagi menjadi dua kelas dan terpisah oleh garis *hyperplane* secara linier dengan menggunakan *soft margin*. Sedangkan SVM Non-Linier yaitu metode SVM yang menangani data tidak linier sehingga perlu diterapkan fungsi kernel terhadap ruang yang berdimensi tinggi (Rachman, 2012).

Metode SVM Linier membagi dataset menjadi 2 kelas. Kelas pertama yang dipisah oleh *hyperplane* bernilai 1, sedangkan kelas lainnya bernilai -1. Menentukan garis *hyperplane* tiap kelas ditunjukkan pada rumus persamaan 2.1.

$$w^T x + b = \begin{cases} \geq 1 \\ \leq -1 \end{cases} \quad (2.1)$$

Dimana w merupakan bobot vektor yaitu garis vektor yang tegak lurus antara titik pusat koordinat dengan garis *hyperplane*. Urutan data ke- i direpresentasikan dalam variabel x . Sedangkan b merupakan nilai bias. Dalam keadaan ini, data dari setiap kelas bisa hanya berada di sisi kiri atau kanan *hyperplane*. Untuk menghitung bobot vektor dan nilai bias ditunjukkan pada rumus persamaan 2.2 dan 2.3.

$$b = -\frac{1}{2}(w \cdot x^+ + w \cdot x^-) \quad (2.2)$$

$$w = \sum_{i=1}^n a_i y_i x_i \quad (2.3)$$

Pada persamaan diatas, variabel x positif dan x negatif merupakan representasi kelas data positif dan negatif. Selain itu variable a_i merupakan nilai bobot tiap data ke- i , variabel y_i merepresentasikan kelas tiap data ke- i , dan variabel x_i adalah representasi data ke- i . Untuk menentukan *hyperplane* yang optimal dari kedua kelas yaitu dengan memaksimalkan jarak antara *hyperplane* pendukung dari kedua kelas yang ditunjukkan pada rumus persamaan 2.6. (Gholami & Fakhari, 2017).

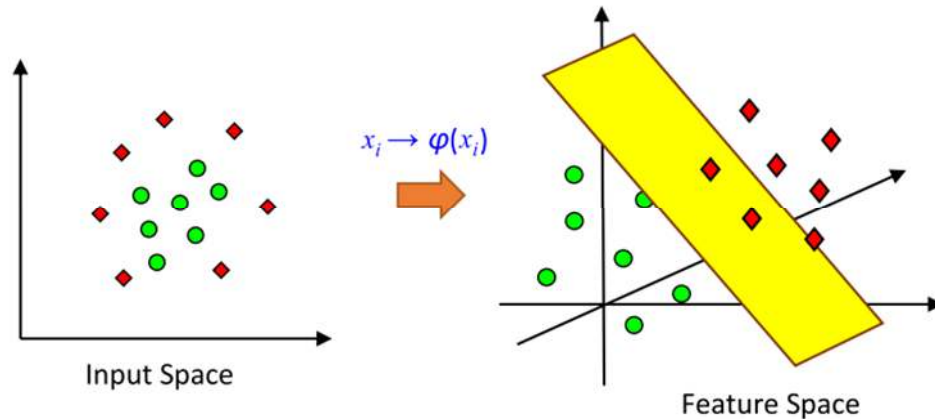
$$m = \frac{b+1}{\|w\|} - \frac{b-1}{\|w\|} \quad (2.4)$$

$$m = \frac{2}{\|w\|} \quad (2.5)$$

$$\min \frac{1}{2} \|w\|^2 \quad (2.6)$$

Dengan syarat :

$$y_i(x_i \cdot w + b) - 1 \geq 0, i = 1, \dots, n \quad (2.7)$$



Gambar 2.6. Ilustrasi SVM Non-linier.

(Gholami & Fakhari, 2017)

Sedangkan Metode Support Vector Machine Non-Linier digunakan pada permasalahan data yang tidak dapat dipisahkan secara *linier*. Sehingga diperlukan pendekatan kernel dalam metode SVM (Octaviani dkk., 2014). Konsep kerja kernel adalah dengan mengubah data ke dalam dimensi ruang fitur sehingga menjadi data yang *linier*. (Hamel, 2009). Beberapa fungsi kernel yang umumnya digunakan dalam SVM adalah sebagai berikut (Han dkk., 2012):

a. Kernel Linier

$$K(x_i, x_j) = x_i x_j \quad (2.8)$$

b. Kernel Polynomial

$$K(x_i, x_j) = (x_i x_j + 1)^d \quad (2.9)$$

c. Kernel Gaussian Radial Basis Function (RBF)

$$K(x_i, x_j) = e^{-\left(\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)}, \sigma > 0 \quad (2.10)$$

d. Sigmoid

$$K(x_i, x_j) = \tanh(\gamma x_i x_j - \delta) \quad (2.11)$$

2.2.8 Confusion Matrix

Confusion matrix adalah sebuah tabel 2 dimensi yang membandingkan model prediksi dengan hasil aktual. Matriks ini membantu mengevaluasi performa model pembelajaran mesin dalam melakukan klasifikasi data. *Confusion matrix* menyediakan informasi tentang jumlah prediksi yang benar dan salah, termasuk jumlah False Positive (FP), False Negative (FN), True Positive (TP), dan True Negative (TN). Contoh tabel *confusion matrix* ditunjukkan pada tabel 2.2.

Tabel 2.2. Contoh Tabel Confusion Matrix. (Grandini et al., 2020)

Classes		Predicted Classification		
		A	B	C
Actual Classification	A	TN	FP	TN
	B	FN	TP	FN
	C	TN	FP	TN

TP menunjukkan jumlah data yang diklasifikasikan sebagai positif oleh model dan juga teridentifikasi sebagai positif oleh data aktual. FN menunjukkan jumlah data yang diklasifikasikan sebagai negatif oleh model tetapi teridentifikasi secara positif oleh data aktual. FP menunjukkan jumlah data yang diklasifikasikan sebagai positif oleh model tetapi teridentifikasi sebagai negatif oleh data aktual. TN menunjukkan jumlah data yang diklasifikasikan sebagai negatif oleh model dan juga teridentifikasi sebagai negatif oleh data aktual. Dengan menggunakan data dari Confusion Matrix, berbagai metrik dapat dikalkulasikan untuk mengevaluasi performa model yaitu *accuracy*, *precision*, *recall*, dan *f1-score* dengan rumus sebagai berikut.

$$accuracy_k = \frac{TP_k + TN_k}{TP_k + FP_k + TN_k + FN_k} \quad (2.12)$$

$$precision_k = \frac{TP_k}{TP_k + FP_k} \quad (2.13)$$

$$recall_k = \frac{TP_k}{TP_k + FN_k} \quad (2.14)$$

$$F1 - Score = 2 \times \frac{precision \times recall}{precision + recall} \quad (2.15)$$

Dalam kasus klasifikasi *multi-class*, perhitungan *f1-score* harus mencakup seluruh *class* dengan menghitung *macro f1-score* dan *micro f1-score*. (Grandini et al., 2020).

$$\text{Macro F1 - Score} = 2 \times \frac{\text{Macro Average Precision} \times \text{Macro Average Recall}}{\text{Macro Average Precision} + \text{recall}} \quad (2.16)$$

$$\text{Macro Average Precision} = \frac{\sum_k^K \text{Precision}_k}{K} \quad (2.17)$$

$$\text{Macro Average Recall} = \frac{\sum_k^K \text{Recall}_k}{K} \quad (2.18)$$

$$\text{Micro Average F1 - Score} = \frac{\sum_k^K TP_k}{\text{Grand Total}} \quad (2.19)$$

2.3. Integrasi Keilmuan

Integrasi keilmuan Islam dengan penelitian ini dapat dilakukan dengan cara menggabungkan pendekatan multidisiplin antara ilmu agama Islam dan teknologi informasi. Dalam konteks ini, keilmuan Islam dapat digunakan sebagai landasan keilmuan islam dalam melakukan penelitian dengan topik analisis sentimen.

Peneliti menggunakan beberapa ayat Al-Qur'an sebagai landasan keilmuan islam yaitu surat Al-Hujurat ayat 6 dan surat Al-An'am ayat 108. Dalam surat Al-Hujurat ayat 6 yang berbunyi :

يَا أَيُّهَا الَّذِينَ ءَامَنُوا إِن جَاءَكُمْ فَاسِقٌ بِنَبَأٍ فَتَبَيَّنُوا أَن تُصِيبُوا قَوْمًا بِجَهَالَةٍ فَتُصْحَبُوا عَلَىٰ مَا فَعَلْتُمْ نَادِمِينَ ۖ
S U R A B A Y A

Artinya : "Hai orang-orang yang beriman, jika datang kepadamu orang fasik membawa suatu berita, maka periksalah dengan teliti agar kamu tidak menimpakan suatu kaum dengan cara tidak sadar, lalu kamu menyesal atas perbuatanmu itu." (QS. Al-Hujurat: 6).

Dalam ayat tersebut, terdapat pesan yang sangat relevan dengan konteks analisis sentimen SARA pada tweet. Ayat ini mengingatkan kita tentang pentingnya memeriksa kebenaran sebuah informasi sebelum menyebarkannya kepada orang lain. Hal ini menunjukkan bahwa kewaspadaan dalam mengonsumsi

dan menyebarkan informasi adalah prinsip yang telah diakui bahkan dalam ajaran agama, untuk mencegah penyebaran fitnah, prasangka, atau informasi yang tidak benar yang dapat merugikan individu atau kelompok tertentu.

Dalam konteks analisis sentimen SARA pada tweet, prinsip ini menjadi semakin relevan. Twitter sebagai platform media sosial telah menyediakan wadah bagi pengguna untuk menyampaikan pandangan dan opini mereka dengan cepat dan mudah. Namun, kerentanannya terhadap penyebaran informasi yang salah atau fitnah juga semakin meningkat. Oleh karena itu, penting bagi para pengguna untuk bersikap bijaksana dan bertanggung jawab dalam menyebarkan informasi yang berkaitan dengan identitas suku, agama, ras, dan golongan sosial.

Dalam surat Al-An'am ayat 108 yang berbunyi :

وَلَا تَسُبُّوا الَّذِينَ يَدْعُونَ مِنْ دُونِ اللَّهِ فَيَسُبُّوا اللَّهَ عَدْوًا بِغَيْرِ عِلْمٍ كَذَلِكَ زَيْنًا لِكُلِّ أُمَّةٍ عَمَلُهُمْ ثُمَّ إِلَىٰ رَبِّهِمْ
مَرْجِعُهُمْ فَيُنَبِّئُهُمْ بِمَا كَانُوا يَعْمَلُونَ ١٠٨

Artinya : "Dan janganlah kamu memaki sembah-sembahan yang mereka sembah selain Allah, karena mereka nanti dengan melampaui batas akan memaki Allah tanpa ilmu pengetahuan." (QS. Al-An'am: 108).

Ayat di atas mengandung pesan yang sangat relevan dalam konteks analisis sentimen SARA. Pesan tersebut menekankan untuk tidak memaki atau merendahkan sembah atau agama yang dipercayai oleh orang lain. Dalam era digital dan penggunaan media sosial yang semakin luas, pesan ini menjadi lebih penting daripada sebelumnya. Dalam konteks analisis sentimen SARA, pesan ini dapat diartikan sebagai panggilan untuk tidak menyebarkan informasi atau konten yang merendahkan atau melecehkan agama, suku, ras, atau golongan sosial tertentu.

BAB III

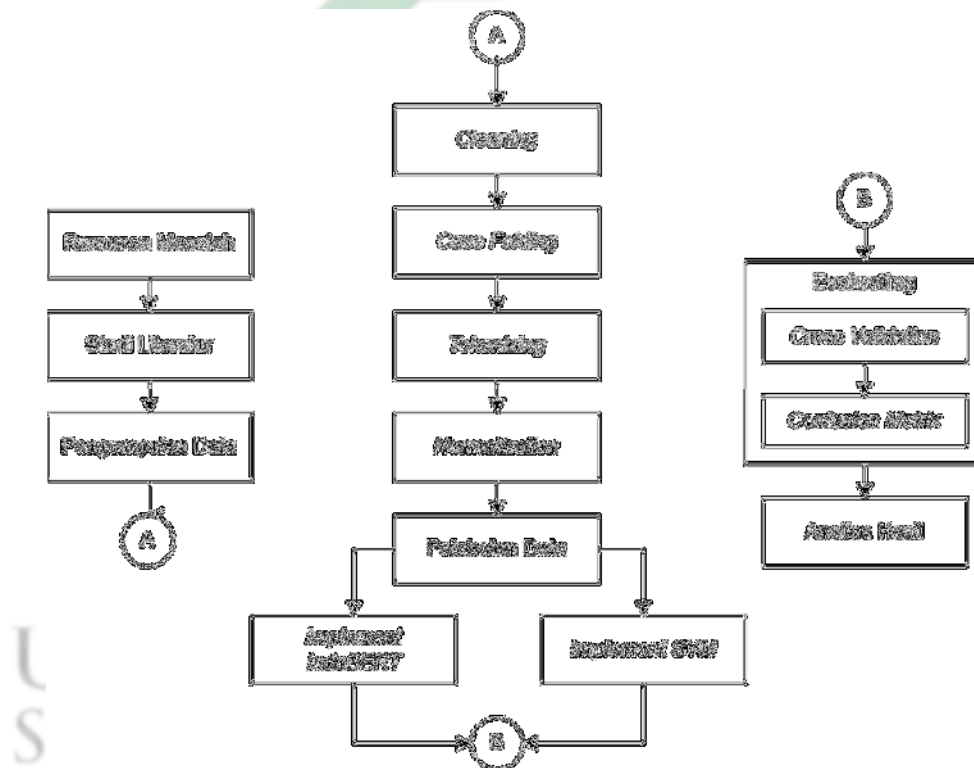
METODE PENELITIAN

3.1. Jenis Data

Jenis data merupakan data primer yang *dicrawling* dari Twitter. Proses *crawling* dilakukan dengan menggunakan Python *programming*.

3.2. Tahapan Penelitian

Alur dalam penelitian ini ditunjukkan pada Gambar 3.1.



Gambar 3.1. Alur Penelitian.

3.2.1. Perumusan Masalah

Dalam penelitian ini, dasar yang digunakan adalah formulasi masalah terkait isu yang terdapat pada latar belakang, yaitu analisis sentimen SARA positif, SARA negative, dan SARA netral. Klasifikasi ini dilakukan untuk membuat prediksi dari data yang dimasukkan. Ini dapat membantu dalam analisa unsur SARA dalam tweet masyarakat Indonesia.

3.2.2. Studi Literatur

Pengumpulan topik penelitian dilakukan dengan melakukan studi literatur dari berbagai sumber yang tersedia. Studi tersebut berfokus pada tiga topik yaitu klasifikasi dengan IndoBERT, kalimat dengan unsur SARA, dan analisis sentimen. Sumber informasi yang digunakan yaitu buku, jurnal atau paper, serta website dan dipilih berdasarkan kredibilitas dan relevansi informasi yang ada.

3.2.3. Pengumpulan Data

Data yang digunakan pada penelitian ini merupakan dataset *tweet* berbahasa Indonesia dengan unsur SARA. Dataset tersebut diambil dengan metode *scraping* menggunakan *snsrape library* dan bahasa pemrograman Python. *Keyword* yang digunakan dalam proses *scraping* yaitu "*buruh*", "*lgbt*", "*feminis*", "*pki*", "*komunis*", "*bule*", "*pribumi*", "*golongan*", "*kaum*", "*etnis*", "*jawa*", "*madura*", "*cina*", "*ambon*", "*arab*", "*suku*", "*agama*", "*islam*", "*kristen*", "*hindu*", "*budha*", "*muslim*". Dataset diambil pada tanggal 1 Februari 2023 hingga tanggal 28 Februari 2023. Dataset memiliki tiga kolom atribut yaitu kolom "Date" yang berisikan data tanggal publikasi setiap *tweet* oleh *user*, kolom "User" yang berisikan data nama setiap *user* yang melakukan *tweet*, dan kolom "Tweet" yang berisikan data *tweet* oleh *user*. Setelah data dilakukan pelabelan, kolom atribut akan bertambah satu dengan nama "Sentiment" yang berisi data label pada setiap *tweet*. Dalam penelitian ini, kolom atribut yang digunakan hanya dua yaitu kolom "Tweet" dan kolom "Sentiment". Contoh sampel dataset ditunjukkan pada Tabel 3.1.

Tabel 3.1. Dataset Tweet.

Date	User	Tweet
@KGenjah	2023-02-01 00:00:15+00: 00	Kelapa orang-orang yang ingkar, pelindung-pelindungnya ialah syaitan, yang mengeluarkan mereka daripada cahaya kepada kegelapan (keingkaran). Mereka itu adalah penghuni Neraka; mereka kekal di dalamnya. (2:257) Bersatu Bela Islam .. #rpHeV
@anon_sia lan	2023-02-01 00:00:18+00: 00	Cawe2 Ijazah yai, Al Qur'an memang tak menjadi hina ketika dibakar para pembenci Islam. Yang menjadi hina dan baper ketika dinyinyirin hanya ormas sampean aja

Date	User	Tweet
@BEBANTEM AN	2023-02-01 00:00:40+00: 00	orang jawa, approved dengan taste bakmie jawa di jakarta ini
@pcpmiipa suruan	2023-02-01 00:00:51+00: 00	PC PMII Ceremony .. KONFERENSI CABANG XXI PERGERAKAN MAHASISWA ISLAM INDONESIA Grand Tema: Meneguhkan spirit gerakan PMII Pasuruan Menyambut Satu Abad Nahdlatul Ulama "Merawat Jagad Membangun Peradaban" 3-5 Februari 2023 Gedung Pacuan Kuda Ki Ageng Astro Joyo
@aminullo 7	2023-02-01 00:00:53+00: 00	, PELACURAN , LGBT, JUDI, MINUMAN KERAS, PENIPU, ITU SEMUA PELAKUNYA DAN BOSNYA DARI GOLONGAN SETAN IBLIS YG KERJANYA MENGAGUNGKAN NAFSU SERAKA DILUAR KEMANUSIAN *****
@SF_GGyoo na	2023-02-01 00:01:11+00: 00	Im menguasai bahasa Korea, Cina (dasar) dan Jepang (dasar)
@dessalsa a	2023-02-01 00:01:37+00: 00	minimal nonton anime di yt, nonton film di netofuri atau klikfilm, drama cina di iqiyi wetv dan yt resmi punya mereka
@AmrizalN inna	2023-02-01 00:02:42+00: 00	Rasul biasa mengerjakan shalat malam hingga kakinya bengkak & beliau tdk senang bila ada orang berjalan2 dibelakangnya" (HR Bukhari-Muslim) Bersatu Bela Islam .. #TlxHx
@BangMaes tro	2023-02-01 00:02:52+00: 00	Provinsi toleran..?? Indonesian Human Rights Monitor (Imparsial) mengungkapkan Jawa Barat (Jabar) merupakan provinsi yang paling banyak menjadi lokasi pelanggaran kebebasan beragama dan keyakinan pada 2022. https://cnnindonesia.com/nasional/20221228201055-20-893331/imparsial-jabar-jadi-provinsi-paling-banyak-kasus-intoleran
@Kemal2sk	2023-02-01 00:04:21+00: 00	Presidennya tokoh muslim berpengaruh didunia, katanya Dijuluki mirip Umar bin Khattab juga Usman bin Affan. Pamer juga waktu jadi imam shalat (walau salah bacaan) Tapi saat umat demo krn Alquran dilecehkan malah dihadang kawat berduri & aparat bersenjata Bener dia Islam juga ?

3.2.4. Preprocessing

Tahap preprocessing data merupakan tahap pengolahan data dari data mentah menjadi data bersih. Dalam prosesnya terdapat lima tahap yaitu.

1. Cleaning

Tahap pertama yaitu *cleaning* yang bertujuan untuk menghilangkan unsur kata yang tidak bermakna didalam kalimat seperti simbol, angka, emoji, *hashtag*, *mention*, dan *url*. Dan juga menghapus *tweets* yang tidak sesuai dengan topik penelitian seperti *tweets* iklan.

2. *Case Folding*

Tahap selanjutnya yaitu *case folding* merupakan tahapan dalam preprocessing dengan tujuan untuk merubah format semua teks menjadi *lowercase*. Contohnya seperti kata “Mantap” menjadi “mantap”. Sehingga proses klasifikasi tidak terganggu.

3. *Tokenizing*

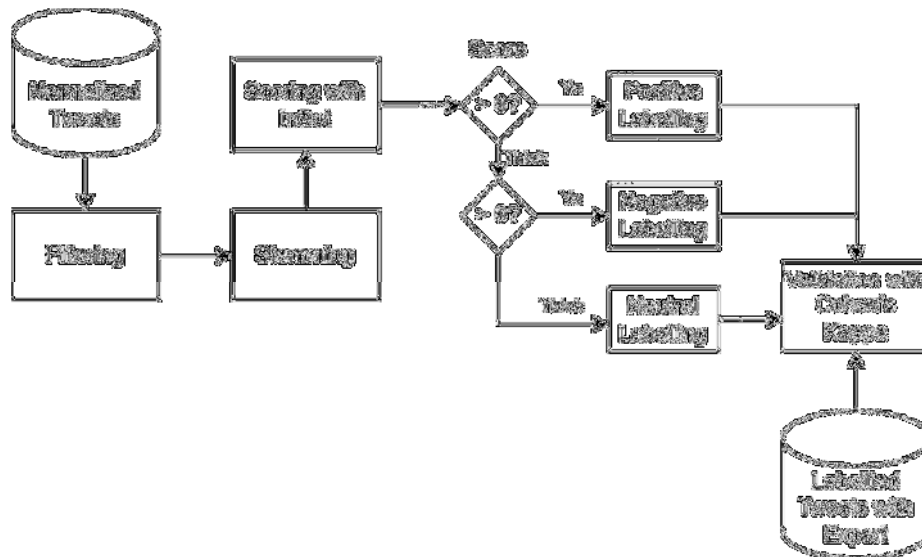
Pada tahapan ini, kalimat pada data *tweets* dipecah menjadi token-token atau potongan kata. Dalam proses *tokenizing*, spasi atau *whitespace* dianggap sebagai pemisah kata. Contohnya seperti kalimat “silat seperti dlm filem cina yg diutamakan” yang dipecah menjadi token seperti berikut [silat, seperti, dlm, filem, cina, yg, diutamakan].

4. *Normalization*

Tahap *normalization* adalah tahap untuk merubah setiap kata alay atau kata gaul menjadi bentuk kata yang sebenarnya. Dalam prosesnya, setiap token kata yang akan dicek dan disesuaikan dengan *corpus* yang telah disediakan. *Corpus* yang digunakan pada tahap ini Colloquial Indonesian Lexicon dengan 15.167 kata alay. (Aliyah Salsabila et al., 2018)

3.2.5. Pelabelan Data

Tahap pelabelan data merupakan tahap dimana setiap kalimat pada data tweet akan diberi label SARA negatif, SARA positif, dan SARA netral sesuai dengan konteks kalimat. Sebelum masuk ke proses pelabelan, data harus dilakukan *preporcessing* lanjutan yaitu *filtering* dan *stemming*. Alur pelabelan data ditunjukkan pada gambar 3.2.



Gambar 3.2. Alur Pelabelan Data.

1. *Filtering*

Filtering dilakukan untuk menghapus kata hubung seperti ‘di’, ‘ke’, ‘yang’, dan ‘itu’. Selain itu, *filtering* juga menghapus *tweets* yang memiliki kata kurang dari 2 huruf dan *tweets* yang kurang dari 2 kata, serta yang tidak berbahasa Indonesia.

2. *Stemming*

Stemming merupakan tahap untuk mengubah kata ke bentuk dasar atau baku. Contohnya seperti kaya “mempunyai” menjadi “punya” dan “memberi” menjadi “beri”.

3. *Labelling*

Labelling dilakukan dengan teknik *lexicon-based labelling* menggunakan InSet Lexicon. InSet Lexicon merupakan leksikon sentimen untuk bahasa Indonesia. Proses *labelling* dilakukan dengan menghitung *score* bobot pada setiap kata didalam *tweets*.

4. *Validation*

Selanjutnya dari dataset diambil data sampel dengan metode *non-probability* yaitu *judgement sampling*. Dari data sampel yang didapat dilakukan pelabelan oleh ahli bahasa untuk kemudian divalidasi menggunakan metode Cohen’s Kappa. Cohen's Kappa merupakan metode statistik untuk mengukur tingkat kesepakatan antara dua orang pelabel atau

dua alat ukur dalam melakukan pelabelan pada objek yang sama. Metode ini menghasilkan nilai antara 0 hingga 1, dengan nilai 0 yang menunjukkan tidak ada kesepakatan antar pelabel dan nilai 1 yang menunjukkan kesepakatan sempurna antara pelabel. Rumus perhitungan ditunjukkan pada persamaan 3.1.

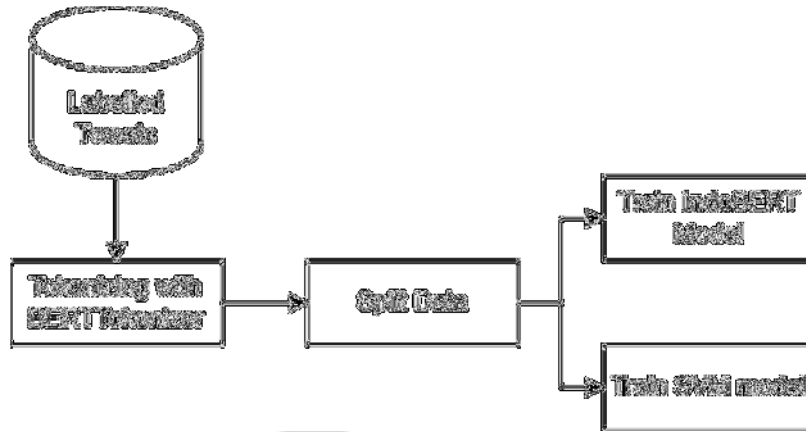
$$k = \frac{P_o - P_e}{1 - P_e} \quad (3.1)$$

Dari persamaan tersebut nilai P_o adalah nilai proporsi kesepakatan antar pelabel, sedangkan P_e adalah nilai proporsi kesepakatan yang diharapkan secara kebetulan. Untuk nilai k yaitu nilai Cohen's Kappa yang menunjukkan tingkat kesepakatan antar pelabel atau alat ukur. (Delgado & Tibau, 2019).

3.2.6. *Modelling*

Di tahap *modelling*, setelah melalui proses *preprocess* dan pelabelan data dilakukan *split data* dan *tokenizing*. *Split data* merupakan tahap dimana data pisah menjadi dua jenis data yaitu data *training*, dan data *testing* dengan persentasi 90% untuk data *training*, 10% untuk data *testing*. Data *training* dibagi kembali menjadi dua untuk mendapatkan data validasi dengan metode *k-fold cross-validation*. Dalam penelitian ini nilai k yang digunakan yaitu 5.

Setelah data di-*split*, *tokenizing* dilakukan kembali dengan *library* yang berbeda dengan sebelumnya yaitu menggunakan BertTokenizer. BertTokenizer merupakan *library* untuk melakukan proses tokenisasi dengan data kamus kata dari IndoBERT. Dalam BERT Tokenizer ditambahkan token spesial seperti [CLS] pada setiap awal kalimat dan [SEP] untuk akhir dari setiap kalimat. Contohnya seperti kalimat “orang jawa approved dengan taste bakmie jawa di jakarta ini” yang dipecah kedalam bentuk token seperti ini ‘[CLS]’, ‘orang’, ‘jawa’, ‘approved’, ‘dengan’, ‘taste’, ‘bakmie’, ‘jawa’, ‘di’, ‘jakarta’, ‘ini’, ‘[SEP]’. Alur *modelling* ditunjukkan pada gambar 3.3.



Gambar 3.3. Alur Modelling.

1. IndoBERT

Dalam model IndoBERT terjadi dua proses yang berjalan secara bersamaan yaitu *pre-training* dan *fine-tuning*. Proses *pre-training* pada IndoBERT menggunakan versi IndoBERT_{base} dengan 12 *layer encoder*, 768 *hidden nodes*, dan 124 juta parameter. Versi tersebut dipilih karena jumlah parameter yang digunakan lebih besar dibandingkan versi IndoBERT_{lite}.

Model IndoBERT hasil *pre-training* diambil sebagai model dasar dan dilakukan proses *fine-tuning*. Proses *fine-tuning* dilakukan dengan mengoptimalkan hyperparameter model IndoBERT yaitu *learning rate*, jumlah *epoch*, dan *batch size*. Dalam penelitian ini digunakan learning rate $5e-5$, jumlah *epoch* 5, 10, dan 20, serta *batch size* 16 dan 32. Nilai tersebut diambil berdasarkan penelitian terdahulu yang memberikan hasil optimal pada kasus analisis sentimen. (Willie et al., 2020) Tabel 3.2 menunjukkan data hyperparameter yang digunakan pada proses *fine-tuning*.

Tabel 3.2. Data Hyperparameter IndoBERT.

Hyperparameter	Nilai
<i>Learning rate</i>	2e-5
<i>Epoch</i>	5 dan 10
<i>Batch size</i>	16 dan 32

2. SVM

Model SVM pada penelitian ini menggunakan kernel *polynomial* dikarenakan fleksibilitasnya dalam mengatur parameter. Parameter yang

digunakan yaitu d (*degree polynomial*), γ (*learning rate*), dan λ (*error control*). Kombinasi parameter tersebut dipilih karena mampu menghasilkan nilai *f1-score* tertinggi. Tabel 3.3 menunjukkan data parameter yang digunakan pada proses *tuning model*. (Karyawati et al., 2022).

Tabel 3.3. Data Parameter SVM.(Karyawati et al., 2022).

Parameter	Nilai
γ	0.1
d	1
λ	3

3.2.7. *Evaluating*

Setelah melakukan tahapan implementasi pada IndoBERT dan SVM, tahap berikutnya adalah memvalidasi model dengan *k-fold cross-validation*. Validasi model dimaksudkan untuk validasi hasil yang keluar dari proses *training* pada masing-masing model. Selanjutnya model diuji dan dievaluasi dengan data uji menggunakan *confusion matrix* untuk mengukur *accuracy* dan *f1-score* pada model IndoBERT dan SVM.

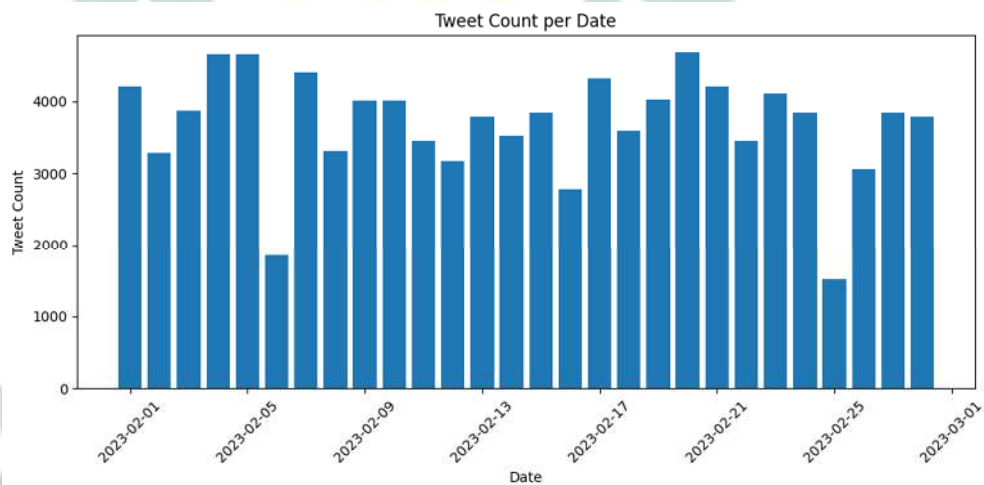
3.2.8. Analisis Hasil

Dalam tahap ini, dilakukan pembahasan terhadap penerapan model yang diusulkan yaitu IndoBERT dengan model pembanding yaitu SVM. Kemudian pembahasan dilanjutkan dengan membandingkan hasil evaluasi model IndoBERT dengan hasil evaluasi model SVM. Dan diakhiri dengan pembahasan analisis hasil sentimen pada masing-masing model dengan menampilkan jumlah sentimen SARA positif, SARA netral, dan SARA negatif.

BAB IV HASIL DAN PEMBAHASAN

4.1 Pengumpulan Data

Data yang dipakai dalam penelitian ini yang diambil dari Twitter. Data yang diambil merupakan *tweets* pada tanggal 1 Februari 2023 hingga 28 Februari 2023 dan memiliki salah satu kata kunci yang mengacu pada unsur SARA yaitu "*buruh*", "*lgbt*", "*feminis*", "*pki*", "*komunis*", "*bule*", "*pribumi*", "*golongan*", "*kaum*", "*etnis*", "*jawa*", "*madura*", "*cina*", "*ambon*", "*arab*", "*suku*", "*agama*", "*islam*", "*kristen*", "*hindu*", "*budha*", "*muslim*". *Keyword* yang dipilih merupakan *keyword* yang sudah divalidasi oleh ahli bahasa bahwa dapat mewakili unsur SARA. Proses pengumpulan data *tweets* memakai *library* *snsrape* dan terkumpul sebanyak 103.180 *tweets*. Grafik jumlah *tweets* setiap harinya dapat dilihat pada Gambar 4.1.



Gambar 4.1. Visualisasi Jumlah Tweet Bulan Februari.

Dataset hasil *scraping* menghasilkan file dengan format *csv*. Dalam dataset terdapat tiga atribut yaitu atribut “Date” yang berisikan data tanggal publikasi setiap *tweet* oleh *user*, atribut “User” yang berisikan data nama setiap *user* yang melakukan *tweet*, dan atribut “Tweet” yang berisikan data *tweet* oleh *user*. Agar dataset dapat digunakan secara optimal, atribut “User” dihilangkan. Sehingga dalam penelitian ini, dataset hanya memiliki dua atribut yaitu atribut “Date” dan “Tweet”. Selanjutnya dataset akan dibagi menjadi dua dengan

perbandingan 70% dan 30%. Dataset dengan proporsi 70% merupakan dataset untuk analisis hasil dengan memprediksi sentimen menggunakan model yang telah dilatih dan diuji. Sedangkan dataset 30% merupakan dataset untuk melatih serta menguji model. Sehingga jumlah data untuk *modelling* yaitu 30.954 *tweets* dan data untuk analisis hasil yaitu 72.226.

4.2 *Preprocessing*

Hasil dari pengumpulan data merupakan data mentah sehingga perlu diolah terlebih dahulu. *Preprocessing* data dilakukan agar data menjadi data yang lebih bersih sehingga tidak terjadi *overfitting* maupun *underfitting* ketika melatih model.

4.2.1. *Cleaning*

Cleaning dilakukan dengan menghapus seluruh kata dalam setiap *tweets* yang tidak memiliki makna. Dan juga menghapus *tweets* yang tidak sesuai dengan topik penelitian seperti iklan menggunakan kata kunci yang berhubungan dengan iklan yaitu "shopee", "shope", "lazada", "baju muslim", "akun premium", "netflix", "spotify", "joki", "recommended", "rekomendasi", "sell", "rent", "pembayaran", "bukhari", "wts", "agent", "slot", "lelang", dan "gacor". Sehingga tidak ada *tweets* yang mengganggu jalannya penelitian. Kata-kata tersebut seperti emoji, *hashtag*, *mention*, dan *url*.

Tabel 4.1. *Tweet Setelah Cleaning.*

Sebelum <i>Cleaning</i>	Setelah <i>Cleaning</i>
Gondangrejo #Arti Mimpi Pindahan Rumah Menurut Islam dan Primbon Jawa - iNews #Gondangrejo #Karanganyar	Gondangrejo Mimpi Pindahan Rumah Menurut Islam dan Primbon Jawa - iNews

4.2.2. *Case Folding*

Penerapan *case folding* digunakan untuk mengubah format *tweets* menjadi huruf kecil. Dengan begitu meminimalisir adanya gangguan selama proses berjalan dan semua *tweets* diperlakukan sama.

Tabel 4.2. *Tweet Setelah Case Folding.*

Sebelum <i>Case Folding</i>	Setelah <i>Case Folding</i>
Gondangrejo Mimpi	gondangrejo mimpi

Pindahan Rumah Menurut Islam dan Primbon Jawa - iNews	pindahan rumah menurut islam dan primbon jawa - inews
---	---

4.2.3. Tokenizing

Pada tahap *tokenization*, *tweets* diubah ke bentuk token-token, seperti kata atau frasa, sehingga pengolahan data dapat dilanjutkan ke proses selanjutnya. Pada proses ini spasi atau *whitespace* dianggap sebagai pemisah.

Tabel 4.3. *Tweet* Setelah *Tokenizing*.

Sebelum <i>Tokenizing</i>	Setelah <i>Tokenizing</i>
gondangrejo mimpi pindahan rumah menurut islam dan primbon jawa - inews	gondangrejo,mimpi,pindahan,rumah,menurut,islam,dan,primbon,jawa,-,inews

4.2.4. Normalization

Tweets dengan bahasa Indonesia cenderung mengandung bahasa yang tidak baku seperti kata-kata singkatan, bahasa gaul, dan bahasa daerah. Sehingga perlu dilakukan normalisasi kata terhadap kata-kata di luar kosakata atau disebut *Out of Vocabulary* (OOV) menggunakan kamus *Colloquial Lexicon*. Selain dapat menormalisasi bahasa gaul sehari-hari juga dapat menangani singkatan.

Tabel 4.4. *Tweet* Setelah *Normalization*.

Sebelum <i>Normalization</i>	Setelah <i>Normalization</i>
cina,aneh,bgt,cok,rasanya,bkn,gaenak,si,cm,ya,manis,aneh,gt,..,heran,bgt,pd,suka	['cina', 'aneh', 'banget', 'cok', 'rasanya', 'bukan', 'tidak enak', 'sih', 'cuma', 'iya', 'manis', 'aneh', 'begitu', '...', 'heran', 'banget', 'pada', 'suka']

4.3 Pelabelan Data

Dalam proses pelabelan data terjadi 4 tahap yaitu *filtering*, *stemming*, *labelling*, *validation*. Beberapa tahap *preprocessing* yang lain seperti *filtering* dan *stemming* diterapkan untuk mempermudah dalam mendeteksi bobot kata. Dalam tahap pelabelan data, jumlah data mengalami penurunan menjadi 23.768 data.

4.3.1. Filtering

Tahap *filtering* dilakukan dengan menghapus seluruh kata-kata yang tidak bermakna seperti kata hubung. Selain itu, tahap *filtering* juga menghapus *tweet* yang hanya memiliki kurang dari 3 kata serta *tweet* yang berbahasa selain bahasa Indonesia.

Tabel 4.5. *Tweet* Setelah *Filtering*.

Sebelum <i>Filtering</i>	Setelah <i>Normalization</i>
['silat', 'seperti', 'dalam', 'film', 'cina', 'yang', 'diutamakan', '.']	['silat', 'film', 'cina', 'diutamakan', '']

4.3.2. Stemming

Pada tahap *stemming*, seluruh kata hasil tahap *filtering* di ubah ke bentuk dasar. Dalam penerapannya digunakan *library* Sastrawi. Tahap ini digunakan agar kata mudah dideteksi ketika *labelling*.

Tabel 4.6. *Tweet* Setelah *Stemming*.

Sebelum <i>Filtering</i>	Setelah <i>Normalization</i>
['silat', 'film', 'cina', 'diutamakan', '']	['silat', 'film', 'cina', 'utama', '']

4.3.3. Labelling

Label sentimen menggunakan teknik *rule based* dengan InSet Lexicon. Label didapat dengan melakukan penjumlahan skor berdasarkan bobot kata pada InSet Lexicon. Label sentimen ditentukan berdasarkan skor "Polarity". Terdapat tiga label sentimen yang didapatkan yaitu positif, netral, dan negatif.

Tabel 4.7. *Tweet* Setelah *Labelling*.

Tweet	Polarity	Sentiment
Adly Abg Zul Amin untuk segala nasihat Moga abg ditempatkan dalam golongan orang yang beriman Orang pertama dapat airborne and low pass with success	0.01250	Positive
silat seperti dlm filem cina yg diutamakan.	0.0	Neutral
ngeluh tapi malu sama warung Madura yang buka 24 jam.	-0.06667	Negative

4.3.4. Validation

Proses validasi labelling dilakukan dengan menggunakan metode manual yang melibatkan pakar dari Lembaga Balai Bahasa. Data yang menjalani proses validasi merupakan sebagian kecil dari keseluruhan data yang sudah dilabeli sebelumnya. Penentuan sampel data dilakukan dengan menggunakan teknik judgement sampling, di mana data sampel dipilih secara sengaja. Jumlah total data yang telah dilabeli menggunakan InSet lexicon adalah sebanyak 23.768 tweets. Untuk validasi pelabelan, diambil sampel sebanyak 1.000 data yang akan dilabeli oleh 3 pakar. Setiap pakar akan melabeli 1.000 data dalam sampel tersebut.

Tabel 4.8. Label Tweet oleh InSet Lexicon dan Pakar.

<i>Tweets</i>	Pelabelan <i>Inset Lexicon</i>	Pelabelan Pakar		
		1	2	3
adly abang zul amin untuk segala nasihatmoga abang ditempatkan dalam golongan orang yang berimanorang pertama dapat airborne and kalau pas dengan sukses	1	1	1	0
silat seperti dalam film cina yang diutamakan .	0	0	-1	0
gondangrejo mimpi pindahan rumah menurut islam dan primbon jawa inews	1	1	1	0
mengeluh tapi malu sama warung madura yang buka jam .	-1	-1	1	-1
perjalanan ini yang setelah diterbitkan sempat menjadi buku kontroversial merekam masa ketika gairah keislaman di negaranegara yang dikunjunginya sedang berkobar disulut oleh kemenangan revolusi islam iran	1	0	0	-1

Proses validasi data menggunakan cohen's kappa. Cohen's kappa dipilih dikarenakan mampu menghitung tingkat kesepakatan antara 2 pelabel secara kebetulan sehingga dapat meminimalisir kesalahan interpretasi akibat kesepakatan yang terjadi secara acak. Selain itu cohen's kappa juga dapat mengatasi masalah ketidakseimbangan kelas dengan cara memperhatikan proporsi setiap kelas sentimen. (Delgado & Tibau, 2019). Dalam penelitian ini, penerapan cohen's kappa menggunakan function `cohen_kappa_score()` dari library sklearn. Dalam cohen's kappa terdapat rentang skor yang menggambarkan seberapa kuat

kesepakatan dari skor kappa. Label tersebut ditunjukkan pada tabel. (Landis & Koch, 1977).

Tabel 4.9. Standar Nilai Kesepakatan. (Landis & Koch, 1977)

<i>Kappa Statistic</i>	<i>Strenght of Agreement</i>
<0.00	<i>Poor</i>
0.00-0.20	<i>Slight</i>
0.21-0.40	<i>Fair</i>
0.41-0.60	<i>Moderate</i>
0.61-0.80	<i>Substantial</i>
0.81-1.00	<i>Almost Perfect</i>

Dari hasil eksekusi function `cohen_kappa_score()` didapat hasil skor sebesar 0,486. Dari skor tersebut dapat dikatakan bahwa label hasil pelabelan InSet lexicon memiliki kesepakatan yang cukup moderat.

4.4 *Modelling*

Pada tahap *modelling*, dataset yang telah dilabeli dilakukan *split data* untuk memisah dataset menjadi subset data *train (train set)*, data *test (test set)*, dan data validasi (*validation set*). *Split data* ini penting untuk membedakan data mana yang digunakan melatih model dan menguji model yang telah dibangun. Data *train* untuk melatih model dan data *test* untuk menguji model. Sedangkan data validasi digunakan untuk melakukan validasi performa model yang dilatih.

Data dibagi dengan proporsi 90% untuk data *train* dan 10% untuk data *test*. Sehingga data *train* memiliki jumlah 21.391 *tweets* dan data *test* memiliki jumlah 2.377 *tweets*. Dari data *train* dibagi kembali menjadi 2 yaitu data *train* dan data validasi. Pembagian tersebut menggunakan *k-fold cross-validation* dengan $k = 5$ sehingga proporsi menjadi 80% data *train* dan 20% data validasi.

Data yang telah di-*split* dilakukan *tokenizing* menggunakan BertTokenizer, BertTokenizer merupakan sebuah *library* yang menggunakan kamus kata dari IndoBERT. *Library* menghasilkan token-token yang merepresentasikan setiap kata dalam teks. Dalam proses *tokenizing* dengan BertTokenizer, peneliti menambahkan token-token spesial seperti [CLS] di awal setiap kalimat dan [SEP] di akhir setiap kalimat. Hal ini dilakukan untuk memberikan informasi penting kepada model BERT tentang awal dan akhir kalimat dalam teks yang akan digunakan dalam proses selanjutnya. Selain itu pembatasan panjang *tweets*

diterapkan dengan menggunakan parameter $\text{max_length} = 128$ yang berarti setiap *tweets* yang memiliki jumlah kata lebih dari 128 kata akan dipotong dan menambah token [PAD] apabila *tweets* memiliki panjang kurang dari 128 kata. *Attention mask* berfungsi sebagai pemberi tanda dalam bentuk nilai padding pada token kata. Token [PAD] memiliki nilai 0 sedangkan token kata diberi nilai 1.

Tabel 4.10. *Tweet Hasil Tokenizing dan Encoding.*

<i>tweets</i>	ngeluh tapi malu sama warung Madura yang buka 24 jam.
<i>token</i>	['[CLS]', 'ngel', '##uh', 'tapi', 'malu', 'sama', 'warung', 'madura', 'yang', 'buka', '24', 'jam', '.', '[SEP]']
<i>encode</i>	[2, 10293, 144, 469, 5535, 500, 6540, 8643, 34, 3121, 1684, 727, 30470, 3]
<i>attention mask</i>	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

4.4.1. IndoBERT

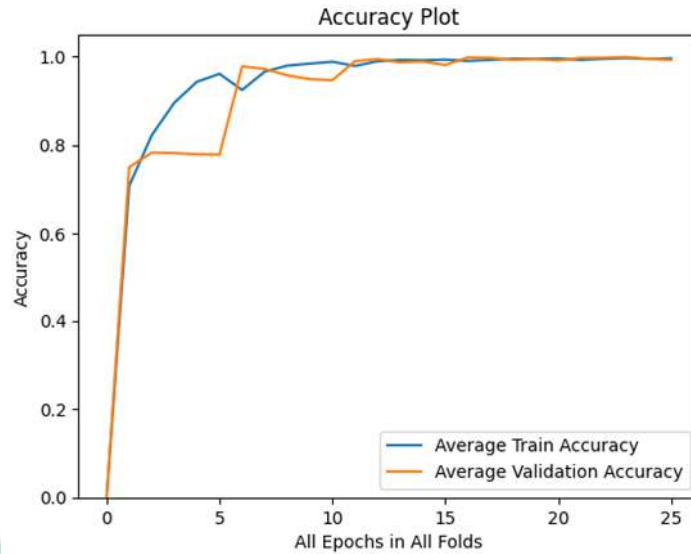
Pada penerapan model IndoBERT, hasil dari proses pengolahan data yang telah dilakukan dilanjutkan ke BERT *network* yang tersusun dari 12 lapisan (*layers*) Transformers Encoder. Setiap lapisan *encoder* terdiri dari 2 sub-*layers*, yaitu *multi-head self-attention* sebagai sub-*layers* pertama, dan *fully connected feed forward network* sebagai sub-*layers* kedua. Setelah melalui semua *encoder*, diperoleh *output* vektor untuk setiap token.

Pada kasus klasifikasi, proses *fine-tuning* IndoBERT terjadi ketika *layer* klasifikasi ditambahkan diujung. Klasifikasi yang dilakukan menggunakan BertForSequenceClassification yaitu salah satu *library* Transformers yang didesain untuk tugas klasifikasi. Dalam BertForSequenceClassification, *ouput* dari *pooler* digunakan untuk menghitung logits. Logits kemudian diinput ke dalam fungsi *softmax* untuk mendapatkan nilai prediksi.. Pada penelitian ini, peneliti menggunakan model IndoBERT_{base} *phase 1* yang memiliki 12 *layers encoder*, 768 *hidden nodes*, dan 124 juta *parameters* dengan *hyperparameter* yang disarankan agar medapat performa model yang optimal.(Wilie et al., 2020).

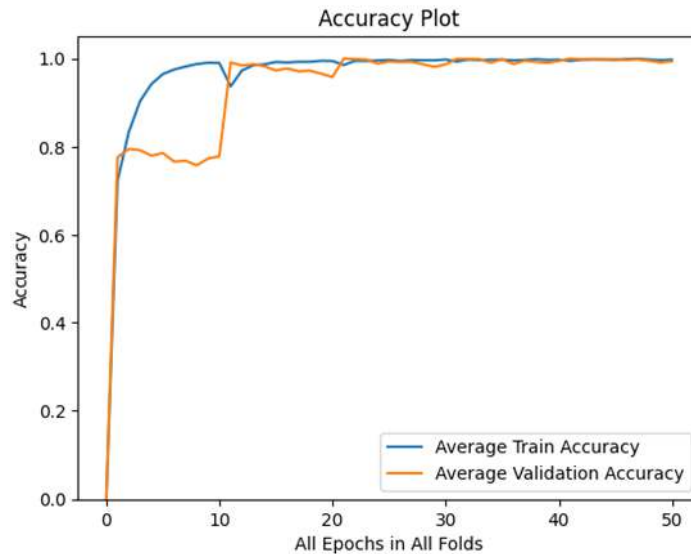
Tabel 4.11. Parameter yang Digunakan untuk *Fine-Tuning* IndoBERT.

Hyperparameter	Nilai
<i>Learning rate</i>	2e-5
<i>Epoch</i>	5 dan 10
<i>Batch size</i>	16 dan 32

Gambaro 4.2, dan Gambar 4.3 menunjukkan grafik komparasi akurasi *training* dan akurasi validasi pada seluruh *fold* hasil *training* model IndoBERT menggunakan *cross-validation* dengan $k=5$ serta *batch size* 16 dengan iterasi 5 *epoch* dan 10 *epoch*.

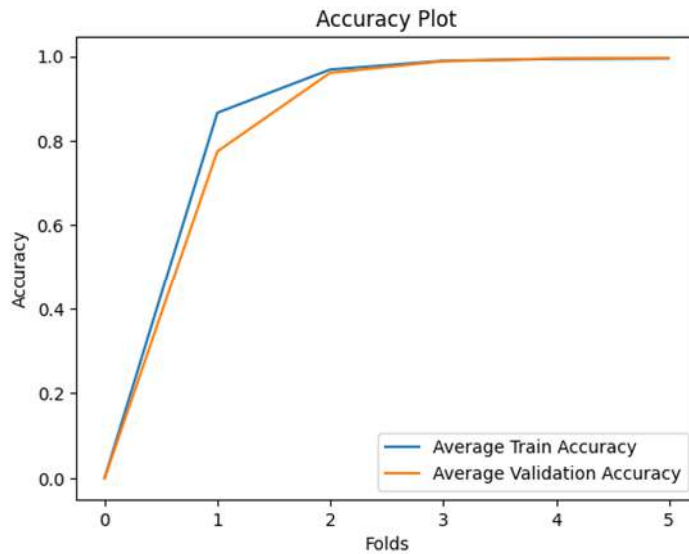


Gambar 4.2. Grafik Akurasi *Train Valid* IndoBERT *Batch Size* 16 (5 *Epoch*).

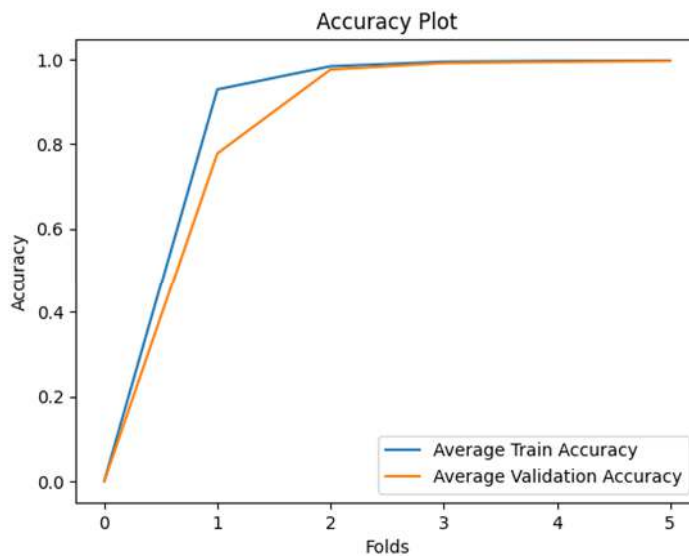


Gambar 4.3. Grafik Akurasi *Train Valid* IndoBERT *Batch Size* 16 (10 *Epoch*).

Gambar 4.4, dan Gambar 4.5 menunjukkan grafik komparasi rata-rata akurasi training dan rata-rata akurasi validasi pada setiap *fold* hasil *training* model IndoBERT menggunakan *cross-validation* dengan $k=5$ serta *batch size* 16 dengan iterasi 5 *epoch* dan 10 *epoch*.



Gambar 4.4. Grafik Akurasi Setiap *Fold* IndoBERT *Batch Size* 16 (5 *Epoch*).



Gambar 4.5. Grafik Akurasi Setiap *Fold* IndoBERT *Batch Size* 16 (10 *Epoch*).

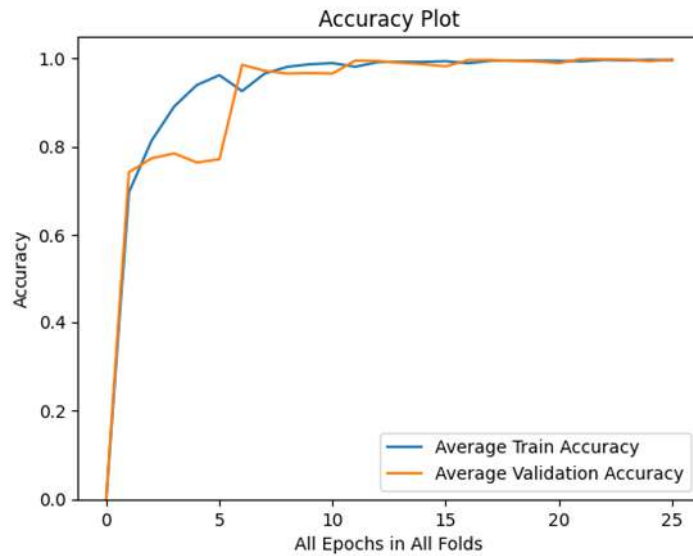
Dari grafik akurasi *training* model dengan *batch size* 16 yang ditunjukkan pada gambar 4.2, gambar 4.3, gambar 4.4, dan gambar 4.5 dapat menjelaskan bahwa akurasi data *training* sangatlah baik pada iterasi 5 *epoch* dan 10 *epoch*. Detail performa IndoBERT dengan *batch size* 16 ditunjukkan pada tabel 4.12.

Tabel 4.12. Data Akurasi *Train Valid* Model IndoBERT dengan *Batch Size* 16.

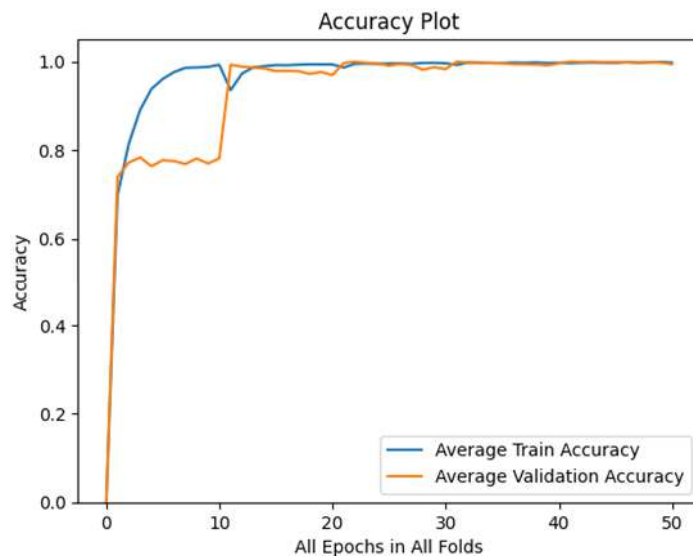
<i>Batch Size</i> = 16		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
5 Epoch	Akurasi <i>Training</i>	86,60%	96,82%	98,88%	99,33%	99,46%
	Akurasi Validasi	77,45%	96,06%	98,78%	99,43%	99,57%
10 Epoch	Akurasi <i>Training</i>	92,55%	98,38%	99,45%	99,65%	99,74%
	Akurasi Validasi	76,87%	97,36%	99,23%	99,40%	99,64%

Dari tabel 4.12 dapat diketahui skor akurasi *training* mendekati skor sempurna yaitu 96.82% di *fold* ke-2 dengan iterasi 5 *epoch* dan 98,38% dengan iterasi 10 *epoch*. Skor tersebut meningkat hingga skor akurasi *training* mencapai 99,33% di *fold* ke-4 dan 99,46% di *fold* ke-5 pada iterasi 5 *epoch*. Untuk iterasi 10 *epoch* skor akurasi meingkat hingga 99,65% di *fold* ke-4 dan 99,74% di *fold* ke-5 . Namun ketika validasi tingkat akurasi dengan iterasi 5 *epoch* dan 10 *epoch* lebih rendah yaitu 77,45% dan 76,87% pada *fold* ke-1 dan meningkat secara signifikan menyusul akurasi *training* dengan skor 96,06% untuk iterasi 5 *epoch* dan 97,36% untuk iterasi 10 *epoch* pada *fold* ke-2. Hal tersebut menandakan adanya *overfitting* pada *fold* ke-1. Dan juga iterasi 10 *epoch* memiliki *overfitting* yang lebih besar dibanding iterasi 5 *epoch*.

Gambar 4.6, dan Gambar 4.7 menunjukkan grafik komparasi akurasi *training* dan akurasi validasi pada seluruh *fold* hasil *training* model IndoBERT menggunakan *cross-validation* dengan k=5 serta batch size 32 dengan iterasi 5 *epoch* dan 10 *epoch*.

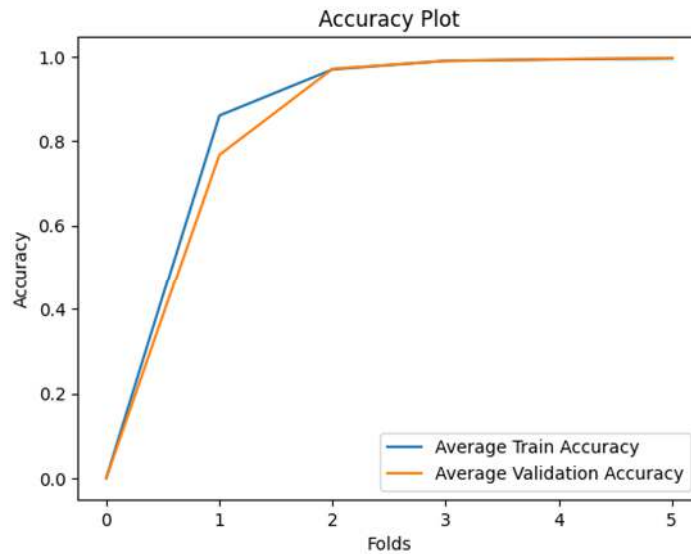


Gambar 4.6. Grafik Akurasi *Train Valid* IndoBERT *Batch Size* 32 (5 *Epoch*).

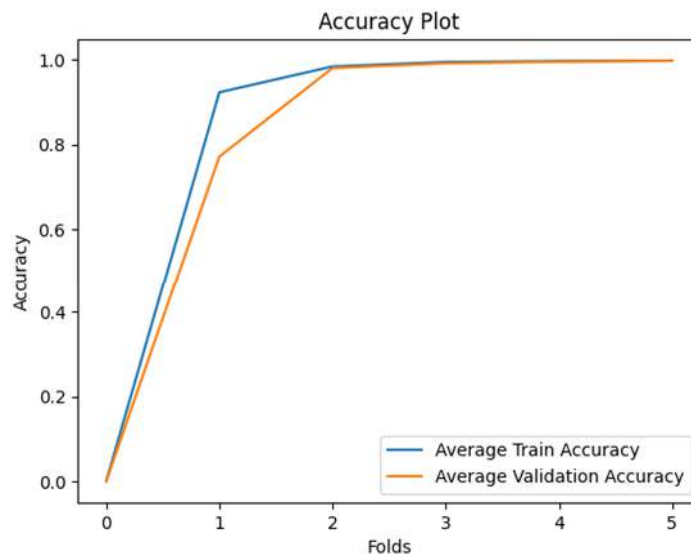


Gambar 4.7. Grafik Akurasi *Train Valid* IndoBERT *Batch Size* 32 (10 *Epoch*).

Gambar 4.8, dan Gambar 4.9 menunjukkan grafik komparasi rata-rata akurasi *training* dan rata-rata akurasi validasi pada setiap *fold* hasil *training* model IndoBERT menggunakan *cross-validation* dengan $k=5$ serta *batch size* 32 dengan iterasi 5 *epoch* dan 10 *epoch*.



Gambar 4.8. Grafik Akurasi Setiap *Fold* IndoBERT *Batch Size* 32 (5 *Epoch*).



Gambar 4.9. Grafik Akurasi Setiap *Fold* IndoBERT *Batch Size* 32 (10 *Epoch*).

Hampir sama seperti grafik sebelumnya, akurasi *training* model dengan *batch size* 32 yang ditunjukkan pada gambar 4.6, gambar 4.7, gambar 4.8, dan gambar 4.9 memberikan hasil yang sama pada iterasi 5 *epoch* dan 10 *epoch*. Yang membuat berbeda yaitu selisih akurasi *training* dan validasi di setiap *epoch* pada *fold* ke-2 lebih kecil dibandingkan ketika model menggunakan *batch size* 16. Detail performa IndoBERT dengan *batch size* 32 ditunjukkan pada tabel 4.13..

Tabel 4.13. Data Akurasi *Train Valid* Model IndoBERT dengan *Batch Size* 32.

<i>Batch Size</i> = 32		<i>Fold 1</i>	<i>Fold 2</i>	<i>Fold 3</i>	<i>Fold 4</i>	<i>Fold 5</i>
5 Epoch	Akurasi <i>Training</i>	86,04%	96,95%	98,99%	99,32%	99,54%
	Akurasi Validasi	76,72%	97,02%	98,93%	99,36%	99,67%
10 Epoch	Akurasi <i>Training</i>	92,31%	98,42%	99,48%	99,68%	99,77%
	Akurasi Validasi	77,11%	98,06%	99,18%	99,56%	99,82%

Pada tabel, skor akurasi *training* mendekati skor sempurna yaitu 96.95% di *fold* ke-2 dengan iterasi 5 *epoch* dan 98,42% dengan iterasi 10 *epoch*. Sama seperti percobaan sebelumnya, skor akurasi *training* mengalami peningkatan hingga skor akurasi mencapai 99,32% di *fold* ke-4 dan 99,46% di *fold* ke-5 pada iterasi 5 *epoch*. Sedangkan pada iterasi 10 *epoch* skor akurasi *training* meningkat hingga 99,68% di *fold* ke-4 dan 99,77% di *fold* ke-5.

Pada validasi model, skor akurasi dengan iterasi 5 *epoch* dan 10 *epoch* lebih rendah yaitu 76,72% dan 77,11% pada *fold* ke-1 dan meningkat secara signifikan menyusul akurasi *training* dengan skor 97,02% untuk iterasi 5 *epoch* dan 98,06% untuk iterasi 10 *epoch* pada *fold* ke-2. Dapat disimpulkan bahwa pada iterasi 10 *epoch* selalu terjadi *overfitting* yang lebih besar dibanding iterasi dengan 5 *epoch* di *fold* ke-1.

4.4.2. SVM

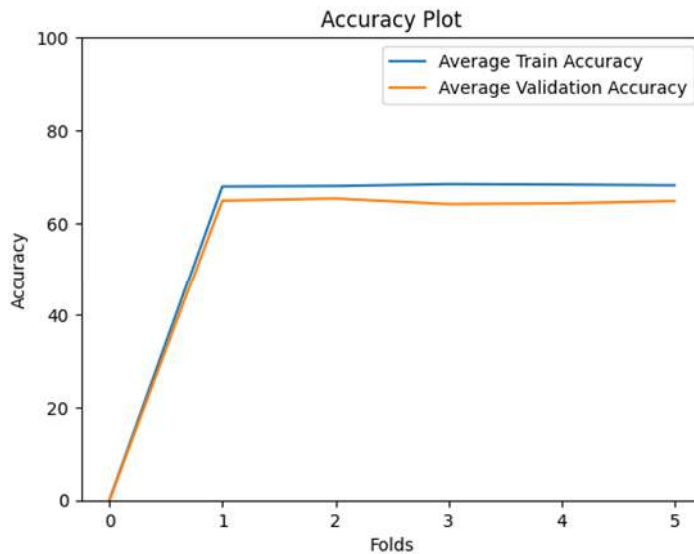
Setelah dilakukan *splitting* dan *tokenizing*, data diteruskan ke *feature extraction* menggunakan teknik *embedding* yaitu WordPiece Embedding. Model WordPiece Embedding yang digunakan yaitu IndoBERT_{base} phase 1 yang memiliki 12 *layers encoder*, 768 *hidden nodes*, dan 124 juta *parameters*. Fitur yang dihasilkan kemudian dilanjutkan ke pelatihan model.

Dalam penelitian ini model SVM menggunakan *kernel polynomial* dan parameter yang ditentukan yang ditunjukkan pada tabel. Selain itu, agar dapat dibandingkan dengan model IndoBERT, SVM menerapkan *k-fold cross-validation* dengan nilai $k = 5$ ketika melatih serta memvalidasi model. Kernel dan parameters dalam pelatihan model dipilih berdasarkan jurnal rujukan dengan kasus yang sama yaitu Sentimen Analisis SARA.

Tabel 4.14. Parameter untuk *Modelling SVM*.

Parameter	Nilai
γ	0.1
d	1
λ	3

Gambar 4.10 menunjukkan grafik akurasi pada setiap *fold* hasil model SVM menggunakan parameter $\gamma = 0,1$, $d = 1$, dan $\lambda = 3$. Dalam gambar tersebut menunjukkan bahwa akurasi naik sebesar 64% pada fold pertama dan mengalami penurunan akurasi pada titik terendah yaitu 61% di fold ke-3. Kemudian akurasi naik kembali menjadi 62% di fold ke-5.



Gambar 4.10. Grafik Akurasi Setiap *Fold SVM*.

Tabel 4.15. Data Akurasi *Training* dan Validasi Model SVM.

	<i>Fold 1</i>	<i>Fold 2</i>	<i>Fold 3</i>	<i>Fold 4</i>	<i>Fold 5</i>
Akurasi Training	67,92%	68,03%	68,44%	68,34%	68,18%
Akurasi Validasi	64,87%	65,33%	64,14%	64,30%	64,79%

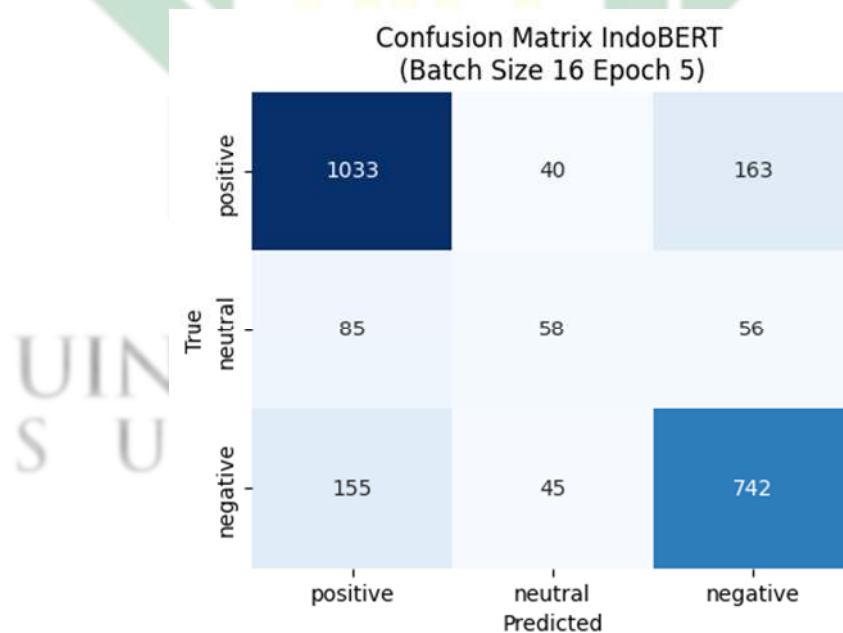
Pada tabel, skor akurasi *training* memiliki skor yang rendah yaitu 67.92% di *fold* ke-2. Skor akurasi *training* mengalami peningkatan hingga mencapai 68,44% di *fold* ke-3 dan terus menurun hingga 68,18% di *fold* ke-5. Sedangkan skor akurasi validasi dimulai dengan 64,87% di *fold* pertama dan naik menjadi

65,33% di *fold* ke-2. Namun, di *fold* ke-3 akurasi turun menjadi 64,14% dan kemudian naik Kembali hingga 64,79% di *fold* ke-5.

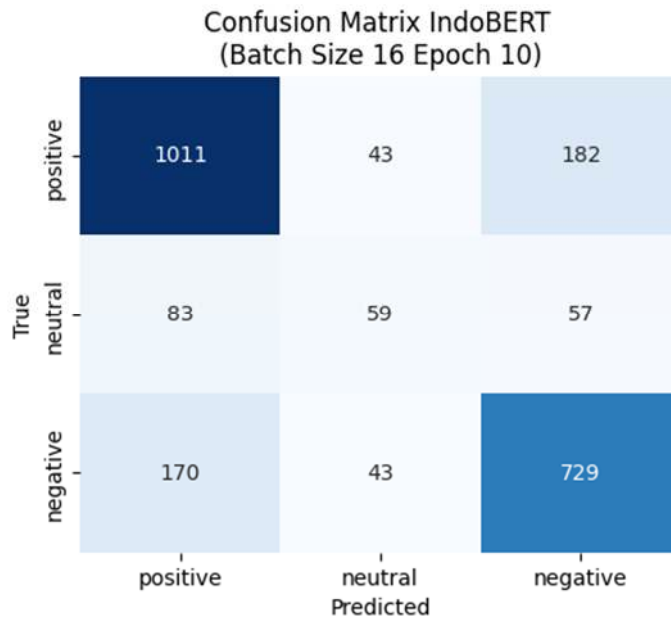
4.5 Evaluating

Pada tahap ini model IndoBERT dan SVM dievaluasi menggunakan *confusion matrix*. Berdasarkan hasil pengujian model terhadap data *testing* didapatkan akurasi sebesar 76% dan 77% dari model IndoBERT menggunakan *batch size* 16 dengan iterasi 5 *epoch* dan 10 *epoch*. Sedangkan pada model IndoBERT dengan *batch size* 32 akurasi yang didapatkan sebesar 78% pada iterasi 5 *epoch* dan 10 *epoch*. Akurasi tersebut jauh lebih baik dibandingkan dengan akurasi pada model SVM dengan *5-fold cross-validation* sebesar 62%.

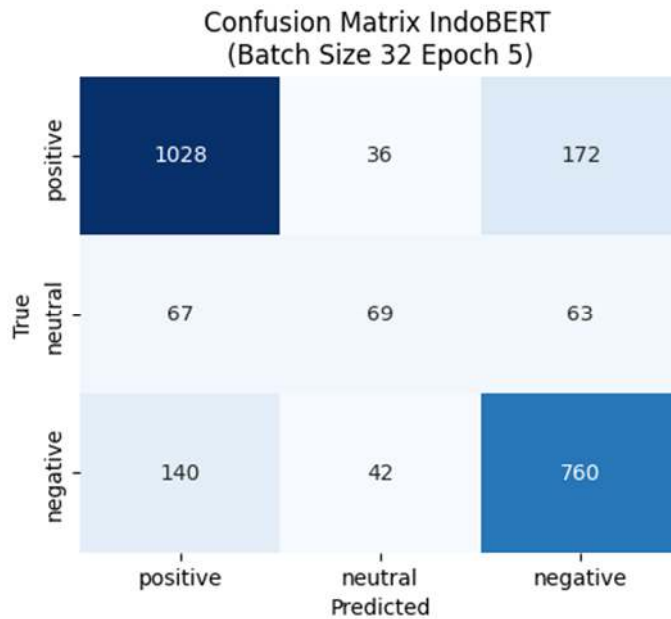
Hasil evaluasi pada model IndoBERT dapat dilihat pada gambar 4.11, gambar 4.12, gambar 4.13, gambar 4.14, gambar 4.15, gambar 4.16, gambar 4.17, dan gambar 4.18. Sedangkan hasil evaluasi pada model SVM ditunjukkan pada gambar 4.19 dan gambar 4.20.



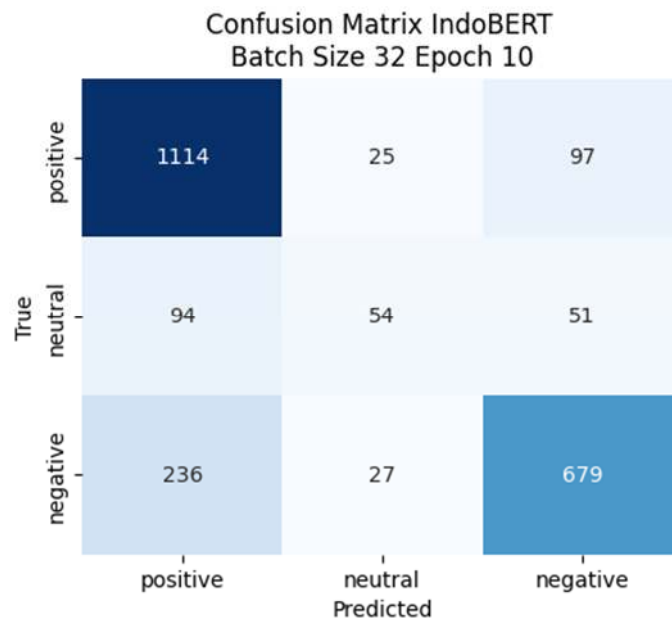
Gambar 4.11. Diagram *Confusion Matrix* IndoBERT *Batch Size* 16 (5 *Epoch*).



Gambar 4.12. Diagram *Confusion Matrix* IndoBERT *Batch Size* 16 (10 *Epoch*).



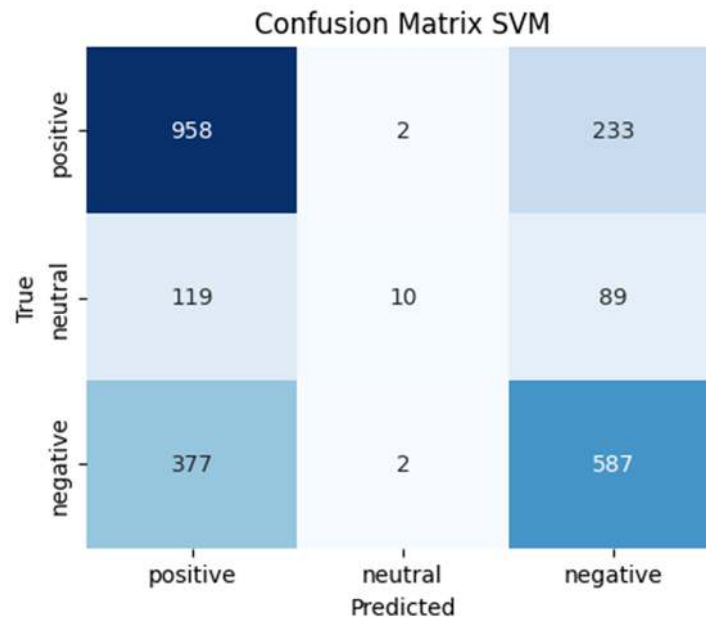
Gambar 4.13. Diagram *Confusion Matrix* IndoBERT *Batch Size* 32 (5 *Epoch*).



Gambar 4.14. Diagram *Confusion Matrix* IndoBERT *Batch Size* 32 (10 *Epoch*).

Dari diagram *confusion matrix* diatas menunjukkan bahwa model IndoBERT memiliki performa yang serupa pada iterasi menggunakan iterasi 5 *epoch* dan 10 *epoch*. Akurasi terbaik yang dihasilkan model dari seluruh percobaan sebesar 78% yaitu pada percobaan model IndoBERT menggunakan *batch size* 32 dengan *epoch* 5 dan 10. Sedangkan akurasi terendah yaitu 76% pada percobaan model IndoBERT menggunakan *batch size* 16 dan *epoch* 10.

Selain itu nilai *precision*, *recall*, dan *f1-score* tertinggi yaitu 77% , 78%, dan 78% yang dimiliki oleh percobaan model IndoBERT dengan *batch size* 32. Sedangkan pada *batch size* 16, nilai *precision*, *recall*, dan *f1-score* lebih rendah yaitu 76%, 77%, dan 77% dengan iterasi 5 *epoch*, serta 75%, 76%, dan 75% dengan iterasi 10 *epoch*. Dapat terlihat perbedaan nilai *batch size* sangat berpengaruh pada akurasi dibandingkan dengan nilai *epoch*.



Gambar 4.15. Diagram *Confusion Matrix* SVM.

Berdasarkan diagram *confusion matrix* hasil evaluasi model SVM dapat dilihat bahwa model SVM memiliki nilai akurasi yang jauh dibawah model IndoBERT yaitu sebesar 65%. Selain itu nilai *precision*, *recall*, dan *f1-score* juga rendah pada bobot rata-rata yaitu 66%, 65%, dan 63%.

Tabel 4.16. Tabel Perbandingan Performa Model

Model	Accuracy	Precision	Recall	F1-Score
IndoBERT (BS 16 & 5 Epoch)	77%	76%	77%	77%
IndoBERT (BS 16 & 10 Epoch)	76%	75%	76%	75%
IndoBERT (BS 32 & 5 Epoch)	78%	77%	78%	78%
IndoBERT (BS 32 & 10 Epoch)	78%	77%	78%	77%
SVM	65%	66%	65%	63%

Dari tabel diatas dapat diketahui bahwa model IndoBERT memiliki hasil yang lebih baik secara keseluruhan. Dari keseluruhan parameter IndoBERT, hanya parameter batch size 32 dan iterasi 5 epoch yang memiliki nilai terbaik secara merata yaitu accuracy 78%, precision 77%, recall 78%, dan f1-score 78%.

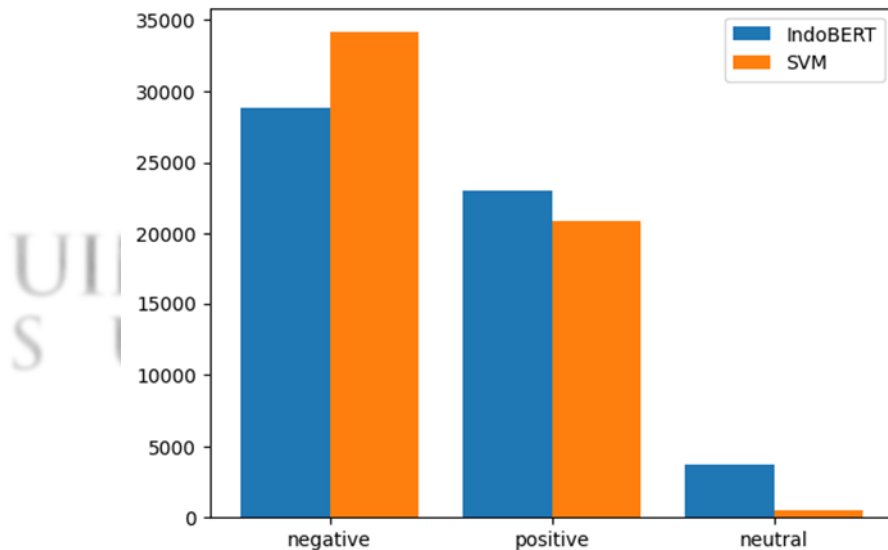
4.6 Analisis Hasil

Dari hasil evaluasi model dapat dilihat bahwa model IndoBERT dengan batch size 32 dan iterasi 5 epoch lebih baik dibandingkan model IndoBERT dengan hyperparameter yang lain. Bahkan model IndoBERT dengan batch size 32

dan iterasi 5 epoch jauh lebih baik dibandingkan model SVM dengan parameter $\gamma = 0,1$, $d = 1$, dan $\lambda = 3$. Hal tersebut terjadi dikarenakan adanya perbedaan pada arsitektur model sehingga proses training juga sangat berbeda dimana IndoBERT menggunakan parameter *epoch* untuk melakukan iterasi saat training model sedangkan SVM hanya melakukan training model dengan 1 kali iterasi. Selain itu, model IndoBERT merupakan model dengan penerapan *transfer learning* dari *corpus* yang besar sehingga model IndoBERT lebih memahami data baru dibandingkan model SVM.

Model yang telah dievaluasi digunakan untuk analisis sentimen dengan memprediksi sentimen data analisis sehingga tujuan dari penelitian ini dapat terpenuhi. Data analisis yang digunakan yaitu 72.226 *tweets*. Namun sebelum diteruskan ke masing-masing model, data perlu dilakukan *preprocessing*. Proses *preprocessing* yang digunakan merupakan proses yang saya yang digunakan sebelumnya. Dari hasil *preprocessing*, terdapat pengurangan data yaitu dari 72.226 *tweets* menjadi 55.601 *tweets*.

Setelah data analisis di *preprocessing*, data diteruskan ke masing-masing model untuk dilakukan prediksi sentimen. Model yang digunakan dalam prediksi sentimen yaitu IndoBERT dengan batch size 32 dan iterasi 5 epoch dan SVM dengan kernel polynomial serta parameter $\gamma = 0,1$, $d = 1$, dan $\lambda = 3$. Hasil dari masing-masing model ditunjukkan pada gambar 4.21 dan 4.22.



Gambar 4.16. Diagram Jumlah *Tweets* Berdasarkan Sentimen.

Berdasarkan gambar 4.21 dapat dilihat bahwa prediksi tweets SARA menggunakan model IndoBERT dapat menghasilkan jumlah sentimen yang lebih

seimbang dibandingkan dengan model SVM. Detail jumlah sentimen yang dihasilkan dari prediksi pada masing-masing model dapat dilihat pada tabel 4.16.

Tabel 4.17. Detail Jumlah *Tweets*.

Sentimen	Jumlah	
	IndoBERT	SVM
Negatif	28.846 <i>tweets</i>	34.147 <i>tweets</i>
Positif	23.038 <i>tweets</i>	20.918 <i>tweets</i>
Netral	3.717 <i>tweets</i>	536 <i>tweets</i>

1. Sentimen Positif

Guna memudahkan analisis *tweets* pada sentimen positif hasil prediksi dengan model IndoBERT dan model SVM, Gambar 4.17 menampilkan perbandingan wordcloud untuk *tweets* berlabel positif. Kata yang ditampilkan dengan ukuran semakin besar berarti semakin tinggi frekuensi kata tersebut digunakan pada *tweets* berlabel positif.

Word Cloud dari Sentimen Positif (IndoBERT)



Word Cloud dari Sentimen Positif (SVM)



Gambar 4.17. *Wordcloud* Sentimen Positif.

Tabel 4.18. Jumlah Kata Sentimen Positif Berdasarkan Frekuensi Kemunculan.

Frekuensi Kata Sentimen Positif			
IndoBERT		SVM	
Kata	Frekuensi	Kata	Frekuensi
islam	6765	islam	5714
jawa	5030	jawa	4780
muslim	2919	muslim	3138
agama	2559	agama	2524
orang	2129	indonesia	2022
indonesia	2042	orang	1812
allah	1846	allah	1769
kaum	1588	arab	1255
barat	1158	timur	1223
umat	1157	barat	1157

Berdasarkan hasil *wordcloud* untuk tweets SARA berlabel positif, beberapa kata yang sering digunakan oleh masyarakat Indonesia dapat diidentifikasi. Kata-kata ini kemudian diurutkan berdasarkan frekuensi tertinggi untuk memberikan gambaran tentang kata-kata apa yang sering digunakan dalam konteks tweets SARA yang memiliki sentimen positif. Salah satu kata dengan frekuensi tertinggi adalah "islam", yang menjadi kata kunci yang muncul paling sering dalam tweets SARA berlabel positif. Ini menunjukkan bahwa unsur agama, khususnya agama Islam, sering digunakan dalam konteks yang positif dalam platform media sosial seperti Twitter.

Selanjutnya, kata "jawa" juga muncul dalam urutan frekuensi tinggi. Penggunaan kata "jawa" dalam konteks *tweets* SARA berlabel positif dapat merujuk pada unsur suku dan wilayah tertentu, yang mengindikasikan bahwa masyarakat Indonesia cenderung menyampaikan sentimen positif terkait dengan identitas suku Jawa atau budaya Jawa secara umum. Penggunaan identitas suku dalam konteks positif dapat mencerminkan rasa bangga akan budaya dan tradisi suku tersebut.

Selain itu, kata "muslim" juga muncul dalam urutan frekuensi tinggi dalam *wordcloud*. Penggunaan kata "Muslim" dalam konteks *tweets* SARA berlabel positif menunjukkan adanya sentimen positif terkait dengan identitas keagamaan suatu kaum atau kelompok. Hal ini menandakan bahwa masyarakat Indonesia cenderung menyampaikan sentimen positif terkait dengan Muslim atau Islam dalam platform media sosial.

Penggunaan unsur-unsur seperti agama, suku, dan identitas keagamaan dalam *tweets* SARA berlabel positif menunjukkan bahwa ada keragaman dan pluralisme dalam opini dan pandangan masyarakat Indonesia di media sosial. Namun, penting untuk memahami bahwa konten positif dalam analisis sentimen tidak selalu berarti bahwa konten tersebut tidak sensitif atau tidak dapat merugikan orang lain. Meskipun kata-kata yang diidentifikasi dalam *wordcloud* memiliki sentimen positif, masih mungkin bagi konten tersebut untuk menyampaikan pesan yang merendahkan atau membagikan informasi yang tidak akurat.

Dalam konteks analisis sentimen SARA pada *tweets*, hasil *wordcloud* yang diurutkan berdasarkan frekuensi tertinggi memperlihatkan beberapa kata yang paling banyak disebutkan, yaitu "agama", "jawa", dan "orang". Kemunculan kata-kata ini dalam frekuensi tinggi dapat menjadi cerminan tentang bagaimana identitas agama dan suku sering menjadi fokus perbincangan dan sentimen negatif di media sosial.

Penggunaan kata "agama" dalam konteks yang negatif mungkin mencerminkan adanya perdebatan, kontroversi, atau perselisihan terkait agama di Indonesia. Isu-isu seperti intoleransi agama, kekerasan yang dilakukan atas nama agama, atau pandangan negatif terhadap suatu agama dapat menjadi faktor yang mempengaruhi sentimen negatif terhadap kata "agama" dalam *tweets* SARA berlabel negatif. Dalam masyarakat multikultural seperti Indonesia, isu-isu agama seringkali menjadi sensitif dan memicu perdebatan yang tajam di media sosial.

Kemunculan kata "jawa" dalam *wordcloud* dengan frekuensi tinggi dapat menunjukkan adanya sentimen negatif terkait dengan suku Jawa. Meskipun suku Jawa merupakan salah satu suku mayoritas di Indonesia dan memiliki warisan budaya yang kaya, tetapi seperti halnya dengan identitas suku lainnya, ada kemungkinan timbulnya sentimen negatif atau prasangka terhadap suku tertentu. Dalam konteks SARA, prasangka terhadap suku atau etnis tertentu dapat memicu konflik dalam masyarakat.

Selain itu, kemunculan kata "orang" dalam frekuensi tinggi dalam *wordcloud* menunjukkan bahwa kata ini seringkali terlibat dalam konteks sentimen negatif di media sosial. Penggunaan kata "orang" dalam sentimen negatif dapat menunjukkan bahwa *tweets* SARA berlabel negatif cenderung terkait dengan perlakuan buruk, stereotip, atau penghinaan terhadap kelompok atau individu tertentu berdasarkan identitas mereka. Ini mencerminkan adanya sikap yang tidak toleran dan kurang menghormati hak asasi manusia dari beberapa pengguna media sosial.

Hasil *wordcloud* ini seharusnya diinterpretasikan dengan hati-hati dan memperhatikan konteks yang lebih luas. Sentimen negatif dalam *tweets* SARA tidak selalu mencerminkan pandangan mayoritas masyarakat, karena pada media sosial, suara minoritas yang keras dan aktif sering kali lebih terdengar. Dalam

analisis sentimen SARA, penting untuk menyadari bahwa *tweet* berlabel negatif tidak selalu mewakili pandangan masyarakat secara keseluruhan, tetapi dapat mencerminkan pandangan kelompok-kelompok ekstrem atau individu yang memiliki agenda tertentu.

Untuk mengatasi analisis sentimen SARA pada *tweet* yang memiliki sentimen negatif, perlu diambil langkah-langkah untuk meningkatkan kesadaran dan edukasi masyarakat tentang pentingnya menghormati perbedaan dan nilai-nilai pluralisme. Kampanye kesadaran yang bertujuan untuk mengajak masyarakat untuk lebih bijaksana dalam berkomunikasi dan berinteraksi di media sosial dapat membantu mengurangi sentimen negatif terhadap suku, agama, ras, dan golongan sosial tertentu. Selain itu, kolaborasi aktif antara perusahaan media sosial, pemerintah, dan masyarakat sipil dapat membantu dalam mengidentifikasi dan menghapus konten yang mengandung unsur SARA atau sentimen negatif yang merugikan.



UIN SUNAN AMPEL
S U R A B A Y A

BAB V

PENUTUP

Bab ini berisi kesimpulan dan saran dari penelitian analisis sentimen SARA pada *tweets* berbahasa Indonesia menggunakan IndoBERT dan SVM.

5.1 Kesimpulan

Dari serangkaian proses penelitian diatas, didapatkan hasil yang telah dijelaskan diatas. Maka dapat disimpulkan bahwa:

1. Model IndoBERT dan SVM bisa digunakan untuk proses analisis sentimen. Namun perlu diperhatikan bahwa input tweets yang diperlukan untuk proses *modelling* hanya sampai pada tahap *normalization* ketika proses *preprocessing*. Hal tersebut dilakukan agar kata hubung dan kata imbungan tidak hilang sehingga model dapat mengetahui konteks pada *tweets*. Maka dari itu tahap *filtering* dan *stemming* hanya dilakukan untuk pelabelan data *tweets* menggunakan InSet Lexicon.
2. Model IndoBERT memberikan performa yang lebih baik dibandingkan dengan model SVM. Dimana model IndoBERT memberikan akurasi sebesar 78% sedangkan model SVM hanya memberikan akurasi sebesar 65%. Sehingga dapat disimpulkan penerapan *deep learning* terbukti lebih baik dalam mengklasifikasikan sentimen dibandingkan dengan *traditional machine learning*.

5.2 Saran

Berdasarkan hasil penelitian yang dijelaskan pada bab analisis sentimen, maka diperlukan penelitian lebih lanjut terkait dengan analisis sentimen terhadap *tweets* SARA. Berikut saran yang dapat mengembangkan penelitian ini:

3. Metode *deep learning* dengan model IndoBERT dapat diganti dengan model yang lain dengan basis arsitektur yang sama yaitu arsitektur *Transformer* dengan model BERT.
4. Proses pelabelan menggunakan InSet *lexicon* dapat diganti menggunakan pelabelan dengan *lexicon* yang lebih besar atau *pseudo labeling* dengan model yang lebih baik.

5. Nilai parameter yang digunakan dalam percobaan yang dilakukan seperti *learning rate* dan *epoch* dapat diganti dengan nilai parameter yang lebih kecil ataupun lebih besar untuk mengetahui performa yang lebih baik.



UIN SUNAN AMPEL
S U R A B A Y A

DAFTAR PUSTAKA

- Aliyah Salsabila, N., Ardhito Winatmoko, Y., Akbar Septiandri, A., & Jamal, A. (2018). Colloquial Indonesian Lexicon. *2018 International Conference on Asian Language Processing (IALP)*, 226–229.
<https://doi.org/10.1109/IALP.2018.8629151>
- Asrif, K. B. M. (2020, April 9). *Salah Kaprah terhadap Frasa “Tidak Mengandung SARA.”* Kemendikbud.
<https://kantorbahasamaluku.kemdikbud.go.id/2020/04/salah-kaprah-terhadap-frasa-tidak-mengandung-sara/>
- Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep Learning for Hate Speech Detection in Tweets. *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*, 759–760.
<https://doi.org/10.1145/3041021.3054223>
- Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). *Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation* (arXiv:1802.02611). arXiv. <http://arxiv.org/abs/1802.02611>
- Christine, P. A. (2018). EFEK POSTINGAN SARA DI MEDIA SOSIAL TERHADAP PERTEMANAN. *KRITIS (Jurnal Ilmu Sosial Dan Ilmu Politik Universitas Hasanuddin)*, 4(1).
- Comm, J., Taylor, D., & Comm, J. (2015). *Twitter power 3.0: How to dominate your market one tweet at a time* (Third edition). Wiley.
- Danukusumo, K. P. (2017). *Implementasi Deep Learning Menggunakan Convolutional Neural Network Untuk Klasifikasi Citra Candi Berbasis Gpu*. UNIVERSITAS ATMA JAYA YOGYAKARTA.
- Delgado, R., & Tibau, X.-A. (2019). Why Cohen’s Kappa should be avoided as performance measure in classification. *PLOS ONE*, 14(9), e0222916.
<https://doi.org/10.1371/journal.pone.0222916>
- Deng, L. (2014). Deep Learning: Methods and Applications. *Foundations and Trends® in Signal Processing*, 7(3–4), 197–387.
<https://doi.org/10.1561/20000000039>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North*, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
- Gholami, R., & Fakhari, N. (2017). Support Vector Machine: Principles, Parameters, and Applications. In *Handbook of Neural Computation* (pp. 515–535). Elsevier. <https://doi.org/10.1016/B978-0-12-811318-9.00027-2>

- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press.
- Grandini, M., Bagli, E., & Visani, G. (2020). *Metrics for Multi-Class Classification: An Overview*. <https://doi.org/10.48550/ARXIV.2008.05756>
- Hagan, M. T., Demuth, H. B., Beale, M. H., & De Jesús, O. (2014). *Neural network design* (2nd edition). Martin T. Hagan.
- Jayadianti, H., Kaswidjanti, W., Utomo, A. T., Saifullah, S., Dwiyanto, F. A., & Drezewski, R. (2022). Sentiment analysis of Indonesian reviews using fine-tuning IndoBERT and R-CNN. *Jurnal Ilmiah ILKOM*, 14(3), 348–354. <https://doi.org/10.33096/ilkom.v14i3.1505.348-354>
- Juanita, S. (2020). Analisis Sentimen Persepsi Masyarakat Terhadap Pemilu 2019 Pada Media Sosial Twitter Menggunakan Naive Bayes. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, 4(3), 552. <https://doi.org/10.30865/mib.v4i3.2140>
- Karyawati, A. E., Utomo, P. A., & Wibawa, I. G. A. (2022). Comparison of SVM and LIWC for Sentiment Analysis of SARA. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, 16(1), 45. <https://doi.org/10.22146/ijccs.69617>
- Kemp, S. (2023, February 9). DIGITAL 2023: INDONESIA. *DATAREPORTAL*. <https://datareportal.com/reports/digital-2023-indonesia>
- Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). *IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP*. <https://doi.org/10.48550/ARXIV.2011.00677>
- Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 159. <https://doi.org/10.2307/2529310>
- Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool.
- Lu, K., & Wu, J. (2019). Sentiment Analysis of Film Review Texts Based on Sentiment Dictionary and SVM. *Proceedings of the 2019 3rd International Conference on Innovation in Artificial Intelligence*, 73–77. <https://doi.org/10.1145/3319921.3319966>
- Marshan, A., Kansouzidou, G., & Ioannou, A. (2021). Sentiment Analysis to Support Marketing Decision Making Process: A Hybrid Model. In K. Arai, S. Kapoor, & R. Bhatia (Eds.), *Proceedings of the Future Technologies Conference (FTC) 2020, Volume 2* (Vol. 1289, pp. 614–626). Springer International Publishing. https://doi.org/10.1007/978-3-030-63089-8_40
- Maulana, M. A., Setyanto, A., & Kurniawan, M. P. (2018). ANALISIS SENTIMEN MEDIA SOSIAL UNIVERSITAS AMIKOMYOGYAKARTA SEBAGAI SARANA PENYEBARAN INFORMASI MENGGUNAKAN

ALGORITMA KLASIFIKASI SVM. *SEMNAS TEKNOLOGI MEDIA ONLINE*
(Seminar Nasional Teknologi Informasi Dan Multimedia), 6(1), 1–2.

Nabilah, G. Z., Prasetyo, S. Y., Izdihar, Z. N., & Girsang, A. S. (2023). BERT base model for toxic comment analysis on Indonesian social media. *Procedia Computer Science*, 216, 714–721. <https://doi.org/10.1016/j.procs.2022.12.188>

Navér, N. (2021). *The past, present or future? : A comparative NLP study of Naive Bayes, LSTM and BERT for classifying Swedish sentences based on their tense* [Uppsala University]. <http://uu.diva-portal.org/smash/record.jsf?pid=diva2%3A1571017&dswid=4465>

Rohmah, T. A. S., & Maharani, W. (2022). Personality Detection on Twitter Social Media Using IndoBERT Method. *Building of Informatics, Technology and Science (BITS)*, 4(2), 448–453. <https://doi.org/10.47065/bits.v4i2.1895>

Saadah, S., Kaenova Mahendra Auditama, Ananda Affan Fattahila, Fendi Irfan Amorokhman, Annisa Aditsania, & Aniq Atiqi Rohmawati. (2022). Implementation of BERT, IndoBERT, and CNN-LSTM in Classifying Public Opinion about COVID-19 Vaccine in Indonesia. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, 6(4), 648–655. <https://doi.org/10.29207/resti.v6i4.4215>

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, 2(3), 160. <https://doi.org/10.1007/s42979-021-00592-x>

Srivastava, R., Bharti, P. K., & Verma, P. (2022). Comparative Analysis of Lexicon and Machine Learning Approach for Sentiment Analysis. *International Journal of Advanced Computer Science and Applications*, 13(3). <https://doi.org/10.14569/IJACSA.2022.0130312>

Twitter. (2023). *Our company and priorities*. About Twitter. <https://about.twitter.com/en>

We Are Social. (2021, October 21). *SOCIAL MEDIA USERS PASS THE 4.5 BILLION MARK*. We Are Social. <https://wearesocial.com/uk/blog/2021/10/social-media-users-pass-the-4-5-billion-mark/>

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., & Purwarianti, A. (2020). *IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding*. <https://doi.org/10.48550/ARXIV.2009.05387>

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., ... Dean, J. (2016). *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. <https://doi.org/10.48550/ARXIV.1609.08144>