BAB II
TINJAUAN PUSTAKA

2.1 Kalimat *Toxic*

Kalimat *toxic* dari segi komunikasi dapat dikatakan sebagai pola komunikasi yang merugikan dan tidak sehat antara individu. Teori ini menekankan pentingnya komunikasi yang efektif, menghormati, dan mendukung dalam hubungan antarmanusia (Arwan, 2018). Dimana Kalimat *toxic* cenderung memiliki efek emosional negatif pada penerima pesan atau orang yang dituju. Paparan berulang terhadap kalimat *toxic* dapat menyebabkan penurunan harga diri, depresi, kecemasan, dan stres kronis. Ini dapat mengganggu hubungan interpersonal, kinerja kerja, dan kesejahteraan secara keseluruhan (Thahir, 2014). Kalimat yang merendahkan dapat menciptakan persepsi negatif tentang kemampuan diri dan kepercayaan diri (Ulfah & Winata, 2021). Beberapa contoh kalimat atau kata *toxic* yang dipaparkan oleh (Alika dkk., 2022) dalam peneltian yang telah dilakukanya tentang "Urgensi Penggunaan Tata Bahasa yang Baik dalam Berkomentar di Aplikasi Media Sosial *Tiktok* Terhadap Kesehatan Mental dan Pembentukan Karakter pada Siswa SMP dan SMA" dan peneltian yang dilakuan oleh (Robbani dkk., 2023) yang melakukan penelitian tentang "Analisis Wacana Digital Penggunaan *Verbal Abuse* dalam Konten *Gameplay Stumble Guys* Luthfi Halimawan" dapat dilhat pada tabel 2.1 berikut:

Tabel 2. 1 Contoh Kalimat *Toxic*

Contoh Kalimat *Toxic*	
anak anjing	tolol
anjing	lonte
bangsat	kayak lonte

Contoh Kalimat *Toxic*	
anjir	alay
jancok	kampungan prik/bocah prik
ngentot	bego
babi	bacot
pantek	gak jelas
brengsek	lebay
kontol	lo ngapain? Keren kaga jelek iya
konthol	gaje lu
koncol	gendut
memek	kaya tante tante
bodoh	jelek
goblok	kurus banget
Oh, ini yang Winnery-Winnery nih ngentot nih. Eh, gua mau menang lagi lawan bot anjing. Jangan dipeluk! Goblok! Jangan peluk gua, anjing! Winnery ngentot gua tandain lu, goblok! Anjing, Winnery. Jangan peluk gua. Jangan peluk, gua lagi lawan bot anjing! Winnery goblok gue pengen menang sekali seumur hidup anjing. ... Aaa! Memek! Awas! Awas ngentot!	Tai ya, Winnery ya, menang lagi dia, bangsat anjing. Gua doain lu dipanggil ibu lu anjing. Awas goblok! Winnery, tolol! Anjing Winnery ni ngentot. Gue doain lu gapunya temen di sekolah, bangsat. Gausah peluk-peluk, anjing! Goblok, tengkorak merah, tolol! Diem dulu, diem

Contoh Kalimat *Toxic*	
Ga usah sliding-sliding memek! Awas anjing! Goblok, bot kontol! Anjing ni orang sumpah... tai anjing! Goblok! Game Stumble, game beta memek! Anying	dulu, ini, ini memang ngentot yak. Ini isinya tu anak kon... ehh. Anjing, kesel gua. ... Nih Winnery aja kek anjing hidupnya ni. Winnery-Winnery kentot, anjing. ... Eh mati kan lu Winnery, goblok! Aaa! Winnery Goblok!!!
Liat, kan. Game beta, goblok. Benerin game lu, goblok! Memek, anjing, kesel banget gua, ngentot! Sss.. Aaa!!! Gamau, keluar, anjing, memek, tolol! Blok! Anjing. Kesel bet gua, anjiiing! Tau ni, nulis review lagi ini ... kontol, ya. Gue nulis (review game) apa ya, kemarin ya? 'Memek Goldplay Button' anjing, gue nulis.	Gausah nempel-nempel ya, anjing! Gausah nempel-nempel! Goblok! Semua aja meluk gua tolol! Kontoolll!!! Ayam ngentot! Weh, udah dong. Anjing, kasih sekali gua menang apa, kontol! Weh, kasih guesekali menang apa, jangan kaya anjingg! ... Lu liat nih, spasinya lepas, anjing! Lu Tempel di keyboard nih, tempel ni keyboard sendiri! Nih! Lu tempel ni keyboard, goblok!
What the fuck, dude. Anjing, udah gitu dia (Istri Luthfi) ga main lagi, ngentot! Sumpah ya, ini yang main peluk-peluk, dua doain pas BO (menyewa pekerja seks komersial)dapet memeknya HIV anjing! Pas lu BO, memek BO-an lu berlalat kontol!	.. A[L] Cupuk mem... wah ngentot, anjing! Fuck, anjing! A[L] Cupuk! Sumpah jangan kaya kontol! Sumpah jangan kaya kontol! Please ini jangan kaya kontol, A[L] Cupuk, anjing! Sumpah jangan... A[L] Cupuk ... kaya kontol! Huw! Let's go! Let's fucking go! ...Let's go, sir!

Contoh Kalimat *Toxic*	
	Didzolimi ngentot, gua ... Mampus A[L] Cupuk! Didzolimi, diperkosa, dikentot rame-rame tapi aku tetap bangkit, anying
angaaan. Undang aja. Eh, gue tuh, orangnya, pengennya bermain bersama kalian guys. Karena, karena menurut gua bermain tanpa kalian tuh, ga seru. ... Gua tu orangnya, solid, bro, sama kalian bro. Ayo buat guys, ayo, masuk-masuk. Masuk, temen-temen. ... Nah ayo guys, 96264 (kode custom room), ayo temen-temen, masuk semuanya! Masuk guys! Gue orangnya solidaritas banget bro, udah gausah... anjing. Udah, bodo amat, ngentot. Lima orang aja yang boleh masuk	Ah, gagal anjing. Kontol! Anjing. Boro-boro crown... Udah-udah keluar, keluar, keluar. ... Uda berapa jam si streaming? Anjing udah 2 jam, cok. Main ginian doang ngentot
Kalau gua lebih ke, gue nih (sebagai) anak-anak juga nih. Lu, anak-anak tongkrongan juga, gitu. Jadi, kaya gua ngomong, 'Anjing lu ya, bangsat. Lu nipu gua ya, tai!' gitu kan, biasanya gue gitu kan ke anak-anak. Maksudnya, anak-anak pada ketawa juga kan. Jadi kaya, apa ya, jokes-jokes-nya itu kaya jokes-jokes anak tongkrongan gitu lho. Jadi kaya mereka curhat ke gua, gua kata-katain balik. Terus gue lagi main, gue diisengin, katain balik. Gitu lho	Maju, maju, maju! Ini siapa ngentot! Ahh! Itu siapa namanya, Affiliator (nama pemain lain), anjing. Goblok! Hiu, anjing! Maju-maju, bangsaat! Winnery jangan kaya kontol lu ya! Ngentot ni ngentot ya! Istri Luthfi (nama pemain lain)! Kontoool! Memek! Aah kontol! Gue ga main-main anjing! Memek!"

2.2 *Instagram*

Instagram adalah *platform* media sosial yang populer dengan lebih dari 1 miliar pengguna aktif bulanan di seluruh dunia. Diluncurkan pada tahun 2010, *Instagram* fokus pada berbagi foto dan video, dan telah berkembang menjadi salah satu *platform* terkemuka dalam industri ini. Pengguna *Instagram* berasal dari berbagai latar belakang dan rentang usia, meskipun mayoritas pengguna berada dalam rentang usia 18-34 tahun. Platform ini menawarkan berbagai fitur yang memungkinkan pengguna untuk mengunggah foto dan video ke akun mereka, serta menambahkan teks, stiker, filter, dan efek lainnya untuk meningkatkan kreativitas. Fitur utama *Instagram* meliputi *Feed*, di mana pengguna dapat melihat *posting*an terbaru dari akun yang diikuti. Selain itu, terdapat *Stories*, yang memungkinkan pengguna membagikan foto dan video yang akan hilang setelah 24 jam. *IGTV* adalah fitur yang memungkinkan pengguna mengunggah video yang lebih panjang, sementara *Direct Messaging* memungkinkan pengguna untuk berkomunikasi secara pribadi dengan pengguna lain. Interaksi dan *engagement* merupakan bagian penting dari *Instagram*. Pengguna dapat menyukai, mengomentari, dan berbagi *posting*an dari pengguna lain. Fitur "*Explore*" memungkinkan pengguna menemukan konten baru yang sesuai dengan minat mereka. *Instagram* juga menjadi *platform* yang populer untuk periklanan. Bisnis dapat menggunakan berbagai opsi iklan yang ditawarkan, termasuk iklan dalam feed, *Stories*, dan *Explore*. Platform ini juga menyediakan profil bisnis dengan informasi kontak dan fitur analitik untuk melacak performa *posting*an dan iklan. Dengan fokus pada konten berkualitas tinggi, *Instagram* memungkinkan pengguna menggunakan *hashtag* untuk meningkatkan visibilitas dan mempermudah temuan oleh pengguna dengan minat yang sama. Secara keseluruhan, *Instagram* menyediakan ekosistem yang luas bagi individu, kreator konten, dan bisnis untuk berinteraksi dan berbagi konten secara kreatif.

2.3 *Artificial Intelligence*

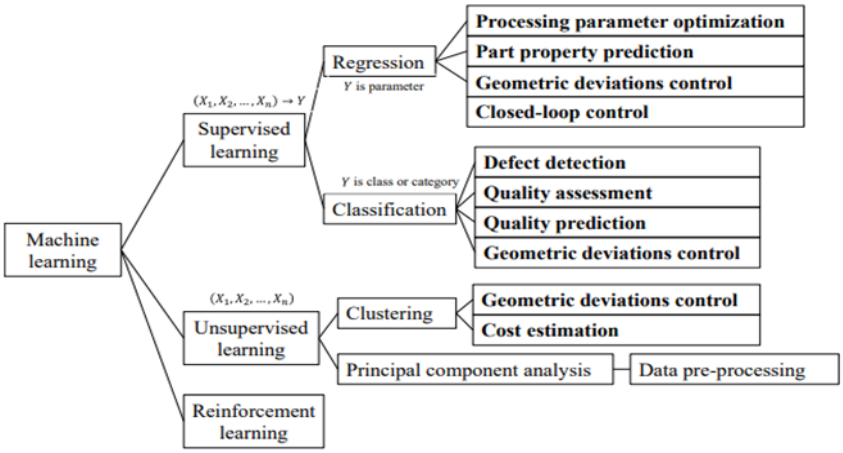
Artificial Intelligence (AI) atau dalam bahasa Indonesia dikenal sebagai Kecerdasan Buatan, adalah suatu bidang ilmu komputer yang bertujuan untuk membuat mesin atau sistem komputer dapat melakukan tugas yang biasanya memerlukan kecerdasan manusia (Novamizanti & Siadari, 2022). *AI* mencakup pengembangan algoritma dan model komputasional yang dirancang untuk memungkinkan mesin untuk "belajar" dari *data*, mengenali pola, mengambil keputusan, dan melakukan tugas-tugas yang cerdas (Markiewicz & Zheng, 2020).

2.3.1 *Machine Learning*

Machine Learning adalah bagian dari Kecerdasan Buatan (*AI*) yang melibatkan penggunaan algoritme dan model statistik untuk memungkinkan komputer atau mesin belajar dari dan membuat prediksi atau keputusan berdasarkan *data* tanpa diprogram secara eksplisit (Carleo dkk., 2019). Dengan kata lain, ini adalah metode atau pendekatan di mana mesin dapat secara otomatis belajar dari pengalaman, meningkat seiring waktu, dan melakukan tugas tanpa diprogram secara eksplisit untuk setiap tugas tertentu.

Pembelajaran Mesin melibatkan penggunaan sejumlah besar *data*, yang dimasukkan ke dalam algoritme dan model untuk melatih mesin. Selama proses pelatihan, mesin mengidentifikasi pola, tren, dan hubungan dalam *data* dan menyesuaikan parameter modelnya untuk mengoptimalkan kinerjanya. Setelah mesin dilatih, ia dapat menggunakan pengetahuan yang dipelajarinya untuk membuat prediksi atau keputusan pada *data* baru yang tidak terlihat (Zhou dkk., 2017).

Menurut (Meng dkk., 2020) *Machine Leraning* dapat dikategorikan menjadi tiga jenis seperti pada gambar 2.1 berikut:



Gambar 2. 1 *Machine Learning Method*

1. *Supervised Learning*

   Dalam jenis Pembelajaran Mesin ini, model dilatih pada *data* berlabel, di mana *data* masukan disertai dengan label keluaran yang sesuai. *Model* belajar dari *data* berlabel untuk membuat prediksi atau mengklasifikasikan *data* baru yang tidak terlihat ke dalam kategori yang telah ditentukan sebelumnya. Beberepa contoh dari jenis pembelajaran mesin *supervised learning* ini adalah deteksi penyakit berdasarkan gejala pada pasien, prediksi harga rumah berdasarkan fitur-fitur seperti ukuran, lokasi, dan jumlah kamar tidur, klasifikasi jenis hewan berdasarkan ciri-ciri fisik, identifikasi jenis tumbuhan berdasarkan karakteristik daun dll.

2. *Unsupervised Learning*

   Dalam jenis Pembelajaran Mesin ini, model dilatih pada *data* yang tidak berlabel, di mana *data* masukan tidak memiliki label keluaran yang menyertainya. *Model* belajar mengidentifikasi pola atau struktur dalam *data*, seperti pengelompokan atau pengurangan dimensi. Beberepa contoh dari jenis pembelajaran mesin *unsupervised learning* ini adalah pengelompokan berita berdasarkan topik yang serupa, segmentasi konsumen

berdasarkan perilaku belanja, reduksi dimensi pada *data* gambar untuk analisis visual dan analisis sentimen pada ulasan produk tanpa label positif/negatif.

3. *Reinforcement Learning*

   Dalam Pembelajaran Mesin jenis ini, model belajar dari interaksi dengan lingkungan dan menerima umpan balik dalam bentuk penghargaan atau hukuman. *Model* belajar mengambil tindakan di lingkungan untuk memaksimalkan imbalan kumulatif dari waktu ke waktu. Beberepa contoh dari jenis pembelajaran mesin reinforcement *learning* ini adalah pengembangan agen permainan video yang belajar bermain game secara optimal, pengaturan pengiriman inventaris oleh robot dalam gudang, pengembangan sistem perdagangan saham yang memutuskan kapan harus membeli atau menjual dan pengembangan agen cerdas untuk bermain permainan catur atau *Go*.

2.3.2 Natural Language Processing

NLP adalah singkatan dari Natural Language Processing, yang merupakan subbidang Kecerdasan Buatan (*AI*) yang berfokus pada memungkinkan komputer atau mesin untuk memahami, menafsirkan, dan berinteraksi dengan bahasa manusia dengan cara yang bermakna dan bermanfaat. *NLP* melibatkan pengembangan algoritma dan model yang dapat memproses dan menganalisis bahasa manusia, termasuk teks dan ucapan, untuk mengekstrak makna, mengidentifikasi pola, dan menghasilkan respons (Markiewicz & Zheng, 2020).

2.4 *Decision Tree*

Algoritma *Decision Tree* (*DT*), yang termasuk dalam kelas algoritma pembelajaran yang diawasi sebagian besar lebih disukai untuk memecahkan masalah klasifikasi tetapi bagaimanapun juga, itu dapat digunakan dalam

mengklasifikasikan serta dalam kasus regresi. Ini terdiri dari simpul dalam yang mewakili struktur cabang, kumpulan *data*, yang mewakili keputusan yang diberikan oleh algoritma, dan setiap simpul daun mewakili hasil. Ada dua simpul, pertama adalah simpul keputusan, yang digunakan untuk mengambil keputusan dan memiliki berbagai cabang; dan yang kedua adalah *leaf node*, yang merupakan output dari decision *node* dan tidak memiliki cabang lagi. Itu berutang namanya pada pohon karena kesamaan bentuknya. Simpul akar adalah titik awal yang selanjutnya berkembang ke berbagai cabang menjadikannya struktur seperti pohon. Pohon keputusan hanya membagi pohon menjadi sub-pohon berdasarkan jawaban atas pertanyaan, yaitu apakah ya atau tidak (Bansal dkk., 2022). Berikut merupakan persamaan yang digunakan pada algoritma decision tree

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} Entropy\ (Si) \quad \textbf{2.1}$$

$$Entropy\ (S) = \sum_{i=1}^{n} - pi * \log_2 pi \quad \textbf{2.2}$$

Berdasarkan persammaan diatas maka dapat di jelaskan sebagai berikut:

1. Menyiapkan *data training*.
2. Menentukan akar dari pohon.
3. Hitung *Gain* Untuk memilih atribut sebagai akar, didasarkan pada nilai *Gain* tertinggi dari atribut-atribut yang ada. Untuk menghitung *Gain* dapat menggunakan persamaan 2.1 diatas.
4. Ulangi langkah kedua hingga setiap cabang terpenuhi. Sementara itu, untuk penghitungan nilai *Entropy* dapat menggunakan persamaan 2.2
5. Proses partisipasi pohon keputusan akan berhenti saat semua cabang dalam *node* N mendapat kelas yang sama.

2.5 Confusion Matrix

Confusion Matrix adalah sebuah *matrix* yang memvisualisasikan kinerja algoritma klasifikasi menggunakan *data matrix*, dengan membandingkan

klasifikasi yang di prediksi dengan klasifikasi aktual dalam bentuk *false positive, true positive, false negative* dan *true negative* (Awad & Khanna, 2015).

Confusion Matrix menggunakan tabel 2.2 seperti *matrix* di bawah ini:

Tabel 2. 2 *Confusion Matrix*

		Predicted	
		Positive	*Negative*
Actual	*Positive*	*TP*	*FN*
	Negative	*FP*	*TN*

Keterangan:

TP : Total kasus *Positive* yang dikategorikan sebagai *Positive*

FP : Total kasus *Negative* yang dikategorikan sebagai *Positive*

TN : Total kasus *Negative* yang dikategorikan sebagai *Negative*

FN : Total kasus *Positive* yang dikategorikan sebagai *Negative*

Berdasarkan nilai *True Negative* (*TN*), *Values Positive* (*FP*), *Values Negative* (*FN*), dan *True Positive* (*TP*) dapat diperoleh nilai akurasi, presisi dan *Recall* menggunakan persamaan berikut

$$Akurasi = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad \textbf{2.3}$$

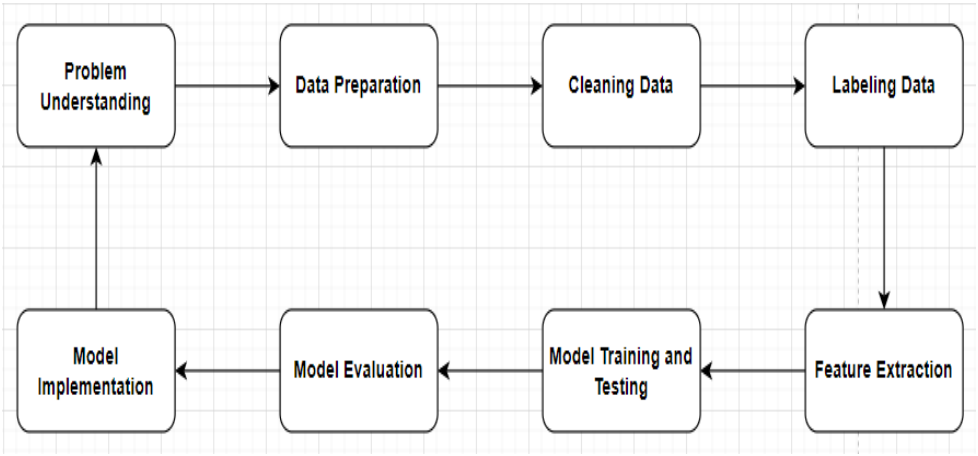
$$Precision = \frac{TP}{(TP + FP)} \quad \textbf{2.4}$$

$$Recall = \frac{TP}{(TP + FN)} \quad \textbf{2.5}$$

2.6 Metode Pengembangan *Machine Leraning Model*

Berdasarkan studi literatur yang dilakuan oleh peneliti maka, dapat diberikan analisa bahwa dalam penelitian yang akan dilakukan ini terdapat kerangka pemikiran yang dijadikan untung menunajang keberlangsungannya penelitian ini yang disebut dengan *Machine Learning Life Cycle*.

Machine Learning Life Cycle merupakan salah satu model pengembangan model machine *learning* pada kecerdasan buatan (*AI*). Tujuan dari machine *learning life* cycle adalah untuk menemukan solusi untuk masalah yang diberikan dengan menerapkan model *Machine Learning*. *Machine Learning Life Cycle* dibagi menjadi delapan tahap utama seperti pada gambar 2.2 dibawah ini:



Gambar 2. 2 *Machine Learning Life Cycle*

Pada tahapan pembuatan model klasifikasi *Machine Learning* menggunakan *Machine Learning Life Cycle* memiliki tahapan-tahapan sebagai berikut:

1. *Problem Understanding*

   Tahap ini melibatkan pemahaman yang komprehensif tentang masalah yang akan dipecahkan, termasuk identifikasi jenis masalah, pemahaman konteks penggunaan model, serta penetapan batasan dan tujuan yang jelas.

2. *Data Preparation*

   Pada tahap ini dilakukan persiapan *data* yang akan digunakan dalam pelatihan dan pengujian model, termasuk pengumpulan *data* yang relevan, penggabungan *data* dari sumber yang berbeda, pengaturan *data* dalam format yang sesuai, dan pembagian *data* menjadi *data* latih (*data training*) dan *data* uji (*data testing*) (Hameed & Naumann, 2020).

3. *Data Cleaning*

   Tahap *data cleaning* ini melibatkan pembersihan *data* untuk memastikan *data* yang digunakan dalam model adalah *data* yang berkualitas dan valid, termasuk penghapusan *data* yang tidak relevan, penggantian atau penghapusan *data* yang hilang atau rusak, serta penanganan duplikasi atau *outliers* dalam *data*(Chu dkk., 2016).

4. *Data Labeling*

   Tahap *data labeling* merupakan tahap pemberian label pada *data* yang akan digunakan sebagai *data* latih model, yang mengindikasikan kategori atau kelas yang ingin diprediksi oleh model, baik secara manual atau menggunakan teknik pemrosesan *data* atau pemrosesan bahasa alami (*NLP*) (Fredriksson dkk., 2020).

5. *Feature Extraction*

   Feature extraction merupakan tahap mengubah *data* menjadi representasi numerik yang dapat digunakan sebagai fitur dalam model (Salahat & Qasaimeh, 2017), termasuk pengubahan *data* teks menjadi vektor kata atau vektor dokumen, penggunaan metode seperti *bag-of-words* atau *word embeddings*, serta ekstraksi fitur-fitur lainnya yang relevan untuk masalah yang dipecahkan..

6. *Training* dan *Testing Model*

   Yang akan dilakukan pada tahap ini yaitu melatih model menggunakan algoritma atau teknik machine *learning* yang sesuai, dengan membagi *data* latih dan *data* uji, mengfitkan model pada

data latih, dan menguji performa model pada *data* uji untuk membuat prediksi atau klasifikasi *data* (Spjuth dkk., 2021).

7. *Model Evaluation*

   Tahap ini dilakukan sebagai pengukuran kinerja model menggunakan metrik evaluasi yang sesuai (Spjuth dkk., 2021), seperti akurasi, presisi, *recall*, atau *F1-score*, untuk mengevaluasi seberapa baik model dalam melakukan prediksi atau klasifikasi *data* serta memvalidasi model dengan menggunakan teknik *k-fold cross validation*.

8. *Implementation*

   Dalam tahap ini dilakuan dengan mengimplementasikan model dalam lingkungan produksi atau aplikasi yang diinginkan, termasuk integrasi model ke dalam sistem produksi, pengujian ulang model di lingkungan produksi, pengelolaan versi model, serta pemeliharaan model untuk memastikan kinerjanya tetap optimal seiring waktu (Spjuth dkk., 2021).

2.7 Penelitian Terkait

Tabel 2. 3 Penelitian Terkait

Judul, Penulis dan Tahun	**Jumlah dan Atribut *Data*set**	**Algoritma**	*Preprocessing*	**Feature Selection**	**Validasi**	**Open Source *Data*set**	**Akurasi**
Klasifikasi Ujaran Kebencian pada Cuitan dalam Bahasa Indonesia (Antariksa dkk., 2019)	20.305 *data* dengan dua atribut	Bernoulli Naïve Bayes, Multinomial Naïve Bayes, Logistic Regression dan Support	Menghapus karakter spesial, menghapus huruf acak, menhapus *data* yang tidak lengkap	*TF -IDF*, n-grams, dan word2vec	*Confusion matrix*	-	Bernoulli Naïve Bayes + *TF-IDF* 94%, Bernoulli Naïve Bayes + N-gram Char Level 80%, Multinomial Naïve Bayes + Word2Vec 98%, Logistic Regression + *TF-IDF* 98%, Logistic Regression + Ngram Char Level 98%, Logistic

Judul, Penulis dan Tahun	**Jumlah dan Atribut *Data*set**	**Algoritma**	*Preprocessing*	**Feature Selection**	**Validasi**	**Open Source *Data*set**	**Akurasi**
		Vector *Machine*					Regression + Word2Vec 98%, *SVM* + *TF-IDF* 98%, *SVM* + N-gram Word Level 98% dan *SVM* + Word2Vec 98%
Analisis Sentimen Pola Pikir Masyarakat Indonesia Terkait Virus Covid-19 Dalam Media Sosial *Twitter* Menggunakan	22.856 *data* dengan tiga atribut	Metode Rule Based Leksikon	*Data cleaning*, lowering *case*, emoticon conversion, *stemming*, and *stopwords data*	*Tokenizing* kata unigram (1 suku kata) dan biagram (2 suku kata)	*Confusion matrix*	-	*Accuracy* sebesar 81 %, nilai precision sebesar 93 %, nilai *recall* sebesar 95 %, dan nilai f1score sebesar 83%.

Judul, Penulis dan Tahun	**Jumlah dan Atribut *Data*set**	**Algoritma**	*Preprocessing*	**Feature Selection**	**Validasi**	**Open Source *Data*set**	**Akurasi**
Metode Rule Based Leksikon (Riskiyanti, 2022)							
Deteksi Cyberbullying pada Facebook Menggunakan Algoritma K-Nearest Neighbor (Detect Cyberbullying on Facebook Using K-Nearest Neighbor	3.000 *data* dengan tiga atribut	KNN	Convert *Emoticon* dan *Punctuation*, *Case Folding*, *Tokenizing*, Filtering, *Stemming* dan Weighting Labelling	Normalisasi	*Confusion Matrix*, *cross validation*	-	1-NN 70.80%, influence punctuation 71.40%, lower*case* dan upper*case* influencce 71.43%, 3-NN 65.93%

Judul, Penulis dan Tahun	**Jumlah dan Atribut *Data*set**	**Algoritma**	*Preprocessing*	**Feature Selection**	**Validasi**	**Open Source *Data*set**	**Akurasi**
Algorithm) (Hasan & Wati, 2021)							
Analisa Tanggapan Terhadap Psbb Di Indonesia Dengan Algoritma *Decision Tree* Pada *Twitter* (Aditya Quantano Surbakti dkk., 2021)	2.439 *data* dengan dua atribut	*Decision Tree*, *Naïve Bayes* dan *K-NN*	*Cleansing*, *Stemming*, Tokenisasi dan *Stopword*	-	*Confusion Matrix*, *cross validation*	-	*Decision Tree* 84.78%, Naive Bayes 84.73% dan K-NN 77.75%
PENDETEKSIAN *HATE SPEECH* PADA SOSIAL	13.169 dengan	*Support Vector Machine*	*Case Folding*, *Remove Special Character*,	*Text Vectorization*	*Confusion Matrix*	-	*Decision Tree* 79%, *Support Vector Machine*

Judul, Penulis dan Tahun	**Jumlah dan Atribut *Data*set**	**Algoritma**	*Preprocessing*	**Feature Selection**	**Validasi**	**Open Source *Data*set**	**Akurasi**
MEDIA INDONESIA DENGAN ALGORITMA *SUPPORT VECTOR MACHINE (SVM)* DAN *DECISION TREE* (Kusuma & Pamungkas, 2023)	dua atribut	*(SVM), BaggingCla ssifier* dan *Decision Tree*	*Remove Username, Remove Url Remove White Space Dan Remove Number*				83% dan *BaggingClassifier* 82%.
Deteksi Komentar Cyberbullying pada Media Sosial	170 data dengan	*Random Forest*	*Cleaning, Case Folding Tokenization &*	*Word Embedding FastText*	*Confusio n Matrix*	-	*Random Forest* sebesar 84%

Judul, Penulis dan Tahun	**Jumlah dan Atribut *Data*set**	**Algoritma**	*Preprocessing*	**Feature Selection**	**Validasi**	**Open Source *Data*set**	**Akurasi**
Instagram Menggunakan Algoritma *Random Forest* (Santoso dkk., 2023)	dua atribut		*Normalization, Stopwords Removal, Stemming*				
Big Five Personality Detection on Twitter Users Using Gradient Boosted Decision Tree Method	275 data dengan lima atribut	*Gradient Boosted Decision Tree*	*Data Cleaning, Case Folding, Tokenization, Stopwords Removal* dan *Stemming*	*TF-IDF* dan *SMOTE*	*Confusion Matrix*	-	Hasil pada skenario pertama menunjukkan akurasi tertinggi pada uji fitur sentimen sebesar 41,82%, dengan presisi dan recall yang sama pada angka 41,82%, serta nilai f1-score sebesar 41,61%.

Judul, Penulis dan Tahun	**Jumlah dan Atribut *Data*set**	**Algoritma**	*Preprocessing*	**Feature Selection**	**Validasi**	**Open Source *Data*set**	**Akurasi**
(Sugiono & Maharani, 2023)							Pada skenario kedua dengan penerapan SMOTE. Akurasi tertinggi pada uji fitur emosi meningkat menjadi 60,36% (+27,63%), presisi meningkat menjadi 59,63% (+33,64%), recall meningkat menjadi 60,36% (+27,63%), dan f1-score meningkat menjadi 60,78% (+27,61%).

Judul, Penulis dan Tahun	**Jumlah dan Atribut *Data*set**	**Algoritma**	*Preprocessing*	**Feature Selection**	**Validasi**	**Open Source *Data*set**	**Akurasi**
Studi Komparasi Metode Analisis *Sentimen Naïve Bayes, SVM,* dan *Logistic Regression* Pada Piala Dunia 2022 (Anbari & Sugiantoro, 2023)	10.680 data dengan dua atribut	*Naïve Bayes, SVM,* dan *Logistic Regression*	*Remove Username, Remve Url, Remove Emoticon, Remove Special Character*	*TF-IDF*	*Confusion Matrix*	-	Bernaouli Naïve Bayes menghasilkan nilai parameter presisi adalah 71%, parameter recall 99%, dan akurasi 76%. Sedangkan metode Support Vector Classifier menghasilkan nilai parameter presisi adalah 94%, parameter recall 93%, dan akurasi 92%. Adapun metode Logistic Regression menghasilkan nilai parameter presisi

Judul, Penulis dan Tahun	**Jumlah dan Atribut *Data*set**	**Algoritma**	*Preprocessing*	**Feature Selection**	**Validasi**	**Open Source *Data*set**	**Akurasi**
							adalah 93%, parameter recall 93%, dan akurasi 92%.
Multi-Tier Sentiment Analysis of Social Media Text Using Supervised Machine Learning (Rahman dkk., 2023)	156.060 data dengan lima atribut	*SVM, Decision tree* dan *Naïve Bayes*	*Remove Stopword, Remove Special Character, Lemmatization* dan *Tokenization*	*TF-IDF*	*Confusion Matrix*	-	SVM Multi-tier model memperoleh akurasi sebesar 0.52%, Decision tree Multi-tier model 0.52% dan Naïve Bayes Multi-tier model 0.56% sedangkan untuk SVM Single-tier model memperoleh akurasi sebesar 0.51%, Decision

Judul, Penulis dan Tahun	**Jumlah dan Atribut *Data*set**	**Algoritma**	*Preprocessing*	**Feature Selection**	**Validasi**	**Open Source *Data*set**	**Akurasi**
							tree Single-tier model 0.51% dan Naïve Bayes Single-tier model memperoleh akurasi sebesar 0.55%
PERBANDINGAN ALGORITMA *SUPPORT VECTOR MACHINE* DENGAN *DECISION TREE* PADA APLIKASI	1.500 data dengan dua atribut	*SVM* dan *Decision Tree*	*Case Folding, Tokenization, Stemming, Filtering*	*TF-IDF*	*Confusion Matrix*	-	SVM melakukan perhitungan probabilitas pada data dengan pembobotan kata menggunakan TF-IDF mendapatkan hasil nilai akurasi sebesar 84.2%, sedangkan pada algoritma

Judul, Penulis dan Tahun	**Jumlah dan Atribut *Data*set**	**Algoritma**	*Preprocessing*	**Feature Selection**	**Validasi**	**Open Source *Data*set**	**Akurasi**
RUANG GURU (Hassanah dkk., 2023)							Decision Tree mendapatkan nilai akurasi sebesar 70%.

Secara umum, penelitian-penelitian di atas menunjukkan bahwa penerapan algoritma *Machine Learning* membantu dalam mendeteksi kalimat *toxic* pada media sosial. Namun, terdapat variasi akurasi yang dihasilkan tergantung pada berbagai faktor seperti ukuran *dataset*, atribut, algoritma, preprocessing, dan *feature selection* yang digunakan. Oleh karena itu beberapa aspek yang memberikan nilai tambah atau kebaruan dalam penelitian ini. Pertama, penelitian ini memiliki fokus pada penentuan ukuran *dataset* yang sesuai, memungkinkan pemilihan sampel yang lebih representatif untuk analisis. Kedua, dengan berfokus pada algoritma *Decision Tree,* pendekatan yang diambil dapat menghasilkan solusi yang berbeda dan lebih akurat dibandingkan dengan menggunakan beberapa algoritma sekaligus seperti dalam penelitian terdahulu. Ketiga, penggunaan fitur ekstraksi *TF-IDF* memberikan kontribusi pada peningkatan kualitas *dataset*. Keempat, dengan membatasi penelitian pada *platform Instagram*, peneliti dapat menghasilkan hasil yang lebih terfokus dan khusus pada media sosial tersebut.