ANALISIS KEMUNGKINAN PENYALAHGUNAAN KONTEN GROK AI DI PLATFORM X MENGGUNAKAN METODE SUPPORT VECTOR MACHINE (SVM)

Salsa Yurinka, Nadhilah Al Adawiyah, Nadia Ramadhani, Fathoni
Program Studi Sistem Informasi, Universitas Sriwijaya
Jl. Raya Palembang-Prabumulih No. KM 32 Inderalaya, Kabupaten Ogan Ilir, Sumatera Selatan (30662)
09031182328109@student.unsri.ac.id

ABSTRAK

Perkembangan kecerdasan buatan membuka peluang penyalahgunaan pada berbagai platform AI, termasuk Grok AI yang beroperasi di platform X. Tingginya volume tweet yang tidak terstruktur menyulitkan identifikasi penyalahgunaan secara manual dalam skala besar, sehingga diperlukan pendekatan otomatis yang efisien. Penelitian ini bertujuan mengklasifikasikan tweet terkait Grok AI ke dalam dua kategori, yaitu penyalahgunaan dan bukan penyalahgunaan, sebagai upaya deteksi konten yang tidak sesuai kebijakan platform. Sebanyak 2.000 tweet dikumpulkan melalui *crawling* dengan kata kunci "@grok", kemudian diproses melalui tahap *preprocessing* dan pembobotan TF-IDF, lalu diklasifikasikan menggunakan algoritma *Support Vector Machine* (SVM). Model SVM menghasilkan akurasi sebesar 85% dengan weighted average F1-score 0,85, menunjukkan bahwa pendekatan ini efektif sebagai metode deteksi otomatis penyalahgunaan konten Grok AI di platform X.

Kata kunci : *Artificial Intelligence (AI), Grok AI, Penyalahgunaan Konten AI , Support Vector Machine, TF-IDF*

1. PENDAHULUAN

Perkembangan teknologi *Artificial Intelligence* (AI) dalam beberapa tahun terakhir mengalami peningkatan yang sangat pesat dan telah mempengaruhi berbagai sektor seperti pendidikan, bisnis, komunikasi digital, dan penelitian ilmiah [1]. Salah satu perkembangan penting adalah munculnya *Generative Artificial Intelligence* (AI), khususnya melalui teknologi seperti *Generative Adversarial Network* (GAN) dan variannya, yang mampu menghasilkan konten digital berupa gambar, video, dan audio secara otomatis dengan kualitas yang sangat realistis hingga sulit dibedakan dari konten asli yang dibuat oleh manusia [2]. Teknologi ini turut mendorong hadirnya berbagai *platform* AI berbasis percakapan yang dapat membantu pengguna dalam memperoleh informasi serta meningkatkan efisiensi aktivitas digital [1]. Salah satu inovasi terbaru adalah Grok AI, yaitu *chatbot* berbasis *large language model* yang dikembangkan oleh xAI dan terintegrasi dengan platform X [3]. Integrasi ini menunjukkan bahwa AI generatif tidak hanya menjadi alat bantu, tetapi juga berperan langsung dalam produksi dan penyebaran konten di media sosial [4].

Namun, penggunaan teknologi AI juga diikuti oleh berbagai risiko, terutama terkait dengan penyalahgunaan dalam pembuatan konten yang melibatkan teknologi AI. Konten yang dihasilkan atau didukung oleh AI berpotensi dimanfaatkan untuk menyebarkan misinformasi, propaganda, ataupun manipulasi opini publik jika tidak digunakan secara bertanggung jawab [2], [5]. Hal ini menjadi perhatian penting mengingat media sosial memiliki peran besar dalam distribusi informasi secara luas dan cepat.

Penelitian ini bertujuan untuk menganalisis dan mengklasifikasikan konten yang berkaitan dengan penggunaan Grok AI di platform X ke dalam kategori penyalahgunaan konten dan bukan penyalahgunaan konten. Penyalahgunaan konten didefinisikan sebagai konten yang mengandung unsur negatif seperti ujaran kebencian, *deepfake*, penghinaan, misinformasi, vulgar, spam, atau penggunaan bahasa yang tidak pantas dalam konteks interaksi dengan Grok AI. Sementara itu, konten non-penyalahgunaan merupakan konten yang bersifat netral, informatif, atau berupa pertanyaan yang tidak mengandung unsur negatif.

Untuk mencapai tujuan tersebut, penelitian ini menggunakan pendekatan pembelajaran mesin dengan algoritma *Support Vector Machine* (SVM). Sebanyak 2.000 tweet dikumpulkan melalui *crawling* menggunakan kata kunci "@grok", kemudian diproses melalui tahap *preprocessing* teks dan pembobotan fitur menggunakan metode TF-IDF sebelum dilakukan klasifikasi. Penelitian ini diharapkan dapat memberikan kontribusi dalam memahami pola penyalahgunaan konten AI di media sosial serta menjadi dasar dalam pengembangan sistem deteksi konten bermasalah berbasis kecerdasan buatan.

2. TINJAUAN PUSTAKA

2.1. *Artificial Intelligence* dan *Generative AI*

Artificial Intelligence (AI) merupakan pendekatan berbasis ilmu komputer yang memanfaatkan program dan algoritma untuk menciptakan sistem yang mampu menjalankan tugas-tugas yang pada umumnya membutuhkan kecerdasan manusia, seperti belajar, menganalisis data, serta menghasilkan keputusan secara otomatis [6]. Perkembangan AI telah membawa perubahan signifikan dalam berbagai bidang, termasuk pengolahan informasi dan komunikasi digital [1].

Salah satu perkembangan penting dari AI adalah *Generative Artificial Intelligence* (*Generative AI*), yaitu sistem kecerdasan buatan yang mampu menghasilkan konten digital seperti gambar, video,

dan audio berdasarkan pola data yang telah dipelajari sebelumnya. Namun, kemampuan tersebut juga menimbulkan berbagai risiko, seperti potensi penyalahgunaan teknologi untuk menghasilkan konten manipulatif dapat menyesatkan informasi di ruang digital [7].

2.2. Grok AI pada Platform X

Grok AI merupakan teknologi kecerdasan buatan berbasis *large language model* yang dikembangkan oleh xAI dan diintegrasikan dengan platform media sosial X [3]. Grok memiliki kemampuan memberikan wawasan media sosial secara *real-time*, merespons berbagai pertanyaan, serta menampilskan gaya komunikasi yang lebih humoris dibandingkan model bahasa lainnya, sehingga membuka peluang pemanfaatan yang lebih luas dalam analisis media sosial dan interaksi pengguna [8]. Selain itu, akses langsung terhadap data dari platform X memungkinkan Grok memberikan informasi yang lebih terkini serta terlibat dalam percakapan yang relevan dengan topik yang sedang berkembang [9].

Dalam pengembangannya, Grok juga memanfaatkan teknik *deep learning* untuk memahami masukan pengguna dan menghasilkan keluaran yang relevan berdasarkan pola yang dipelajari dari data dalam jumlah besar. Salah satu implementasi teknologi ini adalah pada sistem *text-to-image generation* yang mampu menghasilkan gambar berdasarkan deskripsi teks menggunakan model berbasis *transformer* atau arsitektur serupa untuk proses pengkodean teks dan sintesis gambar [10]. Meskipun memiliki berbagai kemampuan yang menjanjikan, teknologi ini masih memiliki beberapa keterbatasan, seperti kemungkinan menghasilkan informasi yang tidak akurat atau bias serta keterbatasan dalam pengetahuan umum dan kemampuan penalaran [8].

2.3. Penyalahgunaan Konten AI

Penyalahgunaan AI (*AI misuse*) merujuk pada penggunaan teknologi kecerdasan buatan untuk tujuan yang tidak etis atau merugikan, seperti penyebaran misinformasi, manipulasi opini publik, serta pembuatan konten berbahaya [11]. Dalam konteks media sosial, penyalahgunaan ini seringkali diwujudkan dalam bentuk konten yang mengandung informasi palsu, ujaran kebencian, maupun perilaku manipulatif yang dapat mempengaruhi persepsi publik secara luas [12].

Selain itu, fenomena *toxicity* dalam media sosial juga menjadi bagian penting dalam kajian penyalahgunaan konten. *Toxicity* mengacu pada konten yang mengandung bahasa kasar, penghinaan, atau serangan personal yang dapat merusak interaksi sosial di ruang digital [13]. Penelitian lain juga menunjukkan bahwa media sosial rentan terhadap penyebaran konten berbahaya seperti *hate speech*, *deepfake*, *cyberbullying*, dan propaganda yang sering kali diperkuat oleh penggunaan teknologi AI [14].

Di sisi lain, perkembangan AI juga berkontribusi terhadap meningkatnya penyebaran misinformasi. Teknologi AI dapat digunakan untuk menghasilkan konten yang tampak kredibel namun sebenarnya menyesatkan, sehingga mempercepat penyebaran informasi yang salah di masyarakat [15]. Bahkan, penelitian menunjukkan bahwa kombinasi antara AI dan media sosial dapat menciptakan ekosistem informasi yang sulit dibedakan antara fakta dan manipulasi [16].

2.4. *Support Vector Machine*

Support Vector Machine (SVM) merupakan salah satu algoritma *machine learning* yang banyak digunakan untuk tugas klasifikasi data teks [17]. Metode ini bekerja dengan mencari *hyperplane* optimal yang mampu memisahkan data ke dalam dua kelas secara maksimal dalam ruang fitur berdimensi tinggi.

Dalam konteks penelitian ini, SVM digunakan untuk melakukan klasifikasi terhadap konten yang dihasilkan dari Grok AI pada platform X ke dalam dua kategori, yaitu penyalahgunaan dan non-penyalahgunaan. Sebelum proses klasifikasi dilakukan, data teks diubah menjadi representasi numerik menggunakan metode *Term Frequency–Inverse Document Frequency* (TF-IDF) [18]. Kombinasi TF-IDF dan SVM dinilai efektif dalam mengolah data teks dan menghasilkan performa klasifikasi yang baik pada berbagai kasus klasifikasi berbasis teks [18].

2.5. Penelitian Terdahulu

Lundberg dan Mozelius [19] mengkaji dampak *deepfakes* terhadap media berita digital dan hiburan menggunakan pendekatan studi kualitatif *cross-sectional* dua fase, yang meliputi analisis tematik forum dan blog serta wawancara semi-terstruktur dengan empat pakar AI dan media. Dengan menggunakan *Technology Affordances* and *Constraints Theory* sebagai kerangka analisis, penelitian ini menemukan bahwa *deepfakes* memiliki sejumlah potensi positif di bidang hiburan dan jurnalisme, namun sisi negatifnya lebih dominan. Konten *hiperrealistis* yang dihasilkan AI terbukti memberikan kekuatan baru pada penipuan, propaganda, dan penyebaran misinformasi di platform media sosial, serta berpotensi menimbulkan dampak psikologis, finansial, dan sosial yang serius bagi individu maupun masyarakat.

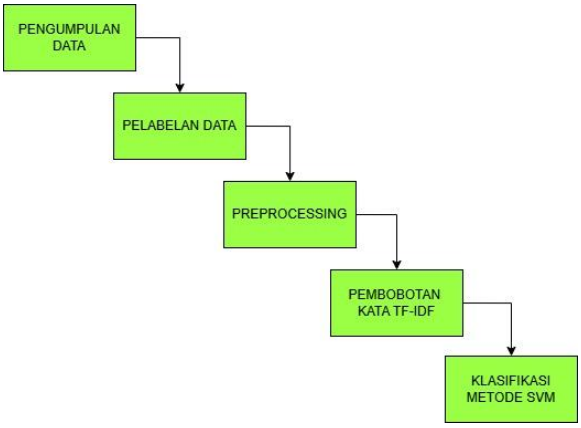
Sementara itu, Caramancion [20] meneliti penggunaan Grok AI sebagai perantara dalam koreksi misinformasi di platform X. Menggunakan desain observasional terstruktur, penelitian ini menganalisis 100 reply koreksi yang terdiri dari 50 koreksi melalui Grok dan 50 koreksi langsung oleh manusia, yang dikumpulkan dari lima topik misinformasi berbeda. Hasil uji *chi-square* menunjukkan asosiasi yang signifikan secara statistik (χ^2=56,25; p<0,001; phi=0,75), di mana sebanyak 72% koreksi yang

dilakukan langsung oleh manusia mendapat serangan *ad hominem*, sedangkan tidak satu pun koreksi melalui Grok yang mendapat respons serupa. Temuan ini menunjukkan bahwa Grok dapat berfungsi sebagai perantara sosial yang mengurangi konflik interpersonal, meskipun keterbatasan dalam akurasi informasi tetap menjadi catatan penting.

Berdasarkan kedua penelitian di atas, terlihat bahwa penggunaan AI di media sosial memiliki dua sisi: sebagai alat bantu yang bermanfaat sekaligus berpotensi disalahgunakan. Namun, keduanya masih berfokus pada analisis deskriptif dan belum memanfaatkan pendekatan pembelajaran mesin untuk klasifikasi otomatis konten penyalahgunaan Grok AI, yang menjadi celah yang diisi oleh penelitian ini.

3. METODE PENELITIAN

Penelitian ini dilakukan melalui beberapa tahapan sistematis untuk membangun model klasifikasi. Tahapan penelitian meliputi pengumpulan data, pelabelan data, *preprocessing* teks, ekstraksi fitur, serta pembangunan dan evaluasi model *machine learning*. Setiap tahapan dirancang secara terstruktur untuk memastikan kualitas data yang baik serta menghasilkan model klasifikasi yang optimal.



Gambar 1. Gambaran umum alur penelitian

3.1. Pengumpulan Data

Pengumpulan data merupakan tahapan awal dalam penelitian ini yang bertujuan untuk memperoleh data teks yang relevan dengan penggunaan Grok AI pada platform X. Data dikumpulkan melalui teknik *crawling*, yaitu proses pengambilan data secara otomatis dari platform digital menggunakan bahasa pemrograman Python serta library pendukung untuk mengekstraksi data teks dari sumber yang diteliti.

Pada penelitian ini, proses *crawling* dilakukan dengan mengumpulkan tweet yang mengandung interaksi dengan Grok AI, seperti penggunaan kata kunci dan mention *@grok* pada platform X. Pengambilan data dilakukan dalam rentang waktu tertentu, yaitu dari tanggal 1 Desember 2023 hingga tanggal 10 April 2026. Selama proses tersebut, sebanyak 2000 tweet berhasil dikumpulkan dan disimpan dalam format CSV.

Data yang dikumpulkan mencakup beberapa atribut, antara lain teks tweet, tanggal posting, username, jumlah *likes*, dan jumlah *retweet*. Proses *crawling* dilakukan secara otomatis menggunakan *script* Python yang dirancang untuk mengekstraksi tweet berdasarkan kata kunci seperti "@grok", "Grok AI", serta kata kunci lain yang relevan dengan penggunaan Grok AI di platform X.

3.2. Pelabelan Data

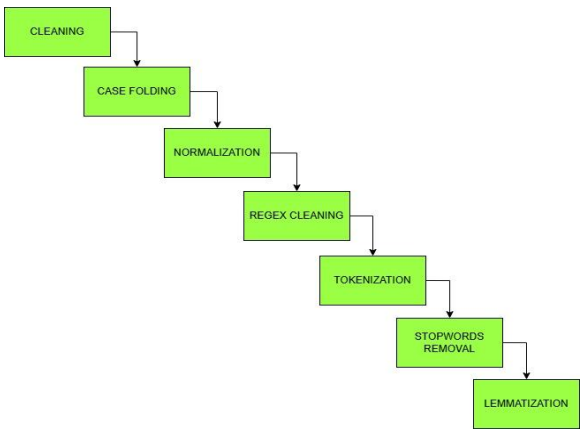
Pelabelan data merupakan tahap pemberian kategori pada setiap data teks yang digunakan dalam penelitian ini. Proses pelabelan dilakukan secara manual dengan cara membaca dan memahami konteks setiap tweet yang telah dikumpulkan dari platform X, khususnya yang berkaitan dengan penggunaan Grok AI.

Dalam penelitian ini, setiap data diklasifikasikan ke dalam dua kategori, yaitu penyalahgunaan *(misuse)* dan non-penyalahgunaan *(non-misuse)*. Kategori penyalahgunaan diberikan pada konten yang mengandung unsur negatif seperti ujaran kebencian, *deepfake*, penghinaan, bahasa kasar, kalimat vulgar, misinformasi, atau bentuk interaksi lain yang menunjukkan penggunaan Grok AI secara tidak bertanggung jawab. Sementara itu, kategori non-penyalahgunaan diberikan pada konten yang bersifat netral, informatif, atau tidak mengandung unsur penyimpangan dalam penggunaan Grok AI.

Proses pelabelan ini dilakukan secara subjektif berdasarkan pedoman kriteria yang telah ditentukan untuk menjaga konsistensi dalam pemberian label. Hasil pelabelan kemudian digunakan sebagai data acuan (*ground truth*) dalam proses pelatihan dan pengujian model klasifikasi menggunakan algoritma *Support Vector Machine* (SVM).

3.3. *Preprocessing*

Preprocessing merupakan tahapan penting dalam proses pengolahan data teks yang bertujuan untuk meningkatkan kualitas data sebelum digunakan dalam proses klasifikasi [21]. Tahap ini dilakukan untuk membersihkan dan menormalisasi teks sehingga dapat menghasilkan representasi fitur yang lebih optimal.



Gambar 2. Gambaran umum alur *preprocessing*

a. *Cleaning*: tahap menghapus elemen yang tidak diperlukan, seperti URL, mention, hashtag, emoji, simbol, angka dan tanda baca berlebih.
b. *Case Folding*: tahap mengubah seluruh teks menjadi huruf kecil (*lowercase*) untuk menghindari perbedaan makna akibat kapitalisasi.
c. *Normalization* : tahap mengubah kata tidak baku/slang menjadi bentuk baku.
d. *Regex Cleaning* : tahap membersihkan teks menggunakan *regular expression* (pola khusus untuk mencocokkan karakter tertentu).
e. *Tokenizing* : tahap memecah kalimat menjadi kata-kata (token).
f. *Stopword Removal* : tahap menghapus kata-kata umum yang tidak memiliki makna signifikan, seperti “and”, “about”, “to”, dsb.
g. *Lemmatization* : proses mengubah kata ke bentuk dasarnya yang sesungguhnya (*lemma*) berdasarkan makna dan konteks kata tersebut.

3.4. *Feature Extraction*

Feature extraction merupakan tahap dalam pengolahan teks yang bertujuan untuk mengubah data teks menjadi representasi numerik agar dapat diproses oleh algoritma *machine learning* [20]. Pada penelitian ini digunakan metode *Term Frequency–Inverse Document Frequency* (TF-IDF) sebagai teknik pembobotan kata dalam dokumen [19].

Metode TF-IDF bekerja dengan menggabungkan dua komponen utama, yaitu *Term Frequency* (TF) yang mengukur frekuensi kemunculan suatu kata dalam dokumen, serta *Inverse Document Frequency* (IDF) yang mengukur tingkat keunikan kata tersebut dalam keseluruhan korpus [20]. Kombinasi kedua komponen ini menghasilkan bobot kata yang lebih representatif terhadap isi dokumen, di mana kata yang sering muncul dalam suatu dokumen tetapi jarang muncul pada dokumen lain akan memiliki nilai yang lebih tinggi [19].

Penggunaan TF-IDF banyak diterapkan dalam berbagai tugas klasifikasi teks karena kemampuannya dalam mengurangi pengaruh kata-kata umum yang kurang informatif serta menonjolkan kata-kata penting dalam dokumen [20]. Selain itu, TF-IDF juga terbukti efektif ketika dikombinasikan dengan algoritma *Support Vector Machine* (SVM) dalam meningkatkan kinerja model klasifikasi berbasis teks [19].

Berdasarkan keunggulan tersebut, metode TF-IDF digunakan dalam penelitian ini untuk merepresentasikan data teks sebelum dilakukan proses klasifikasi ke dalam kategori penyalahgunaan konten dan non-penyalahgunaan konten.

3.5. *Model Machine Learning*

Menurut Rifaldy [19], Model *machine learning* yang digunakan dalam penelitian ini adalah *Support Vector Machine* (SVM) untuk melakukan klasifikasi terhadap konten yang dihasilkan dari penggunaan Grok AI di platform X. SVM dipilih karena memiliki kemampuan yang baik dalam menangani data teks berdimensi tinggi serta efektif dalam proses klasifikasi biner. Model ini bekerja dengan mencari *hyperplane* optimal yang dapat memisahkan data ke dalam dua kelas, yaitu penyalahgunaan konten (*misuse*) dan bukan penyalahgunaan (*non-misuse*). Proses pelatihan dilakukan menggunakan data yang telah melalui tahap *preprocessing* dan *feature extraction* dengan metode TF-IDF, kemudian dievaluasi untuk mengukur performa klasifikasi. Penggunaan SVM dalam klasifikasi teks terbukti mampu memberikan hasil yang akurat dan stabil dalam berbagai penelitian *text mining*.

4. HASIL DAN PEMBAHASAN

4.1. Pengumpulan Data

Pengumpulan data dilakukan melalui proses *crawling* menggunakan Python, dengan memanfaatkan kata kunci "@grok" sebagai parameter pencarian. Dari proses tersebut, berhasil dihimpun sebanyak 2.000 tweet dalam kurun waktu yang telah ditetapkan, yang kemudian disimpan dalam format file CSV.

Tabel 1. Hasil *crawling* data

Text
@amiawuul @grok Hii @grok put her in a bikini
@AAStack @grok please make his belly more distended to reflect how fat he is in reality. He should be eating copious McDonald s and drinking a full on 2L of Diet Coke. Make no mistakes.
@chantellylayne @grok Hahaha well I know if you came across a shrunken man first thing your Gon do is either sit on him or shove him up your butt lol...
@EllaRosari0 @grok Undress first then kiss every inch of your body. Every inch.
@grok @ABridgen Dig deep into ventilator use @Grok not just reported by mainstream media. There's more to this than meets the eye. Doctors who stood up to the harmful effects of the vaccine and treatment protocols were ostracized. Do you have data of those reports in your system. Question do
@jacksonhinklle @grok whendid it happen?
@elonmusk As a Korean user I m feeling this hard. Grok suddenly turned my timeline into a high-quality Japanese content goldmine Now I m just waiting for the same magic on Korean posts. Keep cooking Grok! @grok @elonmusk
@japrwds @tobilobaphotog @JONEYdcd @grok You could choose to spread kindness or continue being you.

4.2. Pelabelan Data

Pelabelan data dilakukan secara manual untuk mengelompokkan tweet ke dalam 2 kategori klasifikasi, yaitu penyalahgunaan dan bukan penyalahgunaan. Pada tahap pelabelan ini, mendapatkan hasil yang seimbang, diantaranya sebanyak 1000 tweet masuk ke dalam kategori penyalahgunaan dan 1000 tweet masuk ke dalam kategori bukan penyalahgunaan.

Tabel 2. Hasil pelabelan data

Penyalahgunaan
@amiawuul @grok Hii @grok put her in a bikini
@AAStack @grok please make his belly more distended to reflect how fat he is in reality. He should be eating copious McDonald s and drinking a full on 2L of Diet Coke. Make no mistakes.
@chantellylayne @grok Hahaha well I know if you came across a shrunken man first thing your Gon do is either sit on him or shove him up your butt lol...
@EllaRosari0 @grok Undress first then kiss every inch of your body. Every inch.
Bukan Penyalahgunaan
@grok @ABridgen Dig deep into ventilator use @Grok not just reported by mainstream media. There's more to this than meets the eye. Doctors who stood up to the harmful effects of the vaccine and treatment protocols were ostracized. Do you have data of those reports in your system. Question do
@jacksonhinklle @grok whendid it happen?
@elonmusk As a Korean user I m feeling this hard. Grok suddenly turned my timeline into a high-quality Japanese content goldmine Now I m just waiting for the same magic on Korean posts. Keep cooking Grok! @grok @elonmusk
@japrwds @tobilobaphotog @JONEYdcd @grok You could choose to spread kindness or continue being you.

4.3. *Preprocessing* Data

Data yang berhasil dikumpulkan masih belum sepenuhnya bersih, karena di dalamnya masih terdapat berbagai elemen pengganggu seperti URL, mention, emoji, hashtag, serta penggunaan bahasa yang tidak baku. Kondisi ini menjadi alasan mengapa tahap *preprocessing* perlu dilakukan sebelum data tersebut masuk ke proses analisis berikutnya.

a. *Cleaning*

Pada tahap ini dilakukan penghapusan bagian komponen model, seperti URL, mention, angka, tanda baca, emoji, dan karakter non-alfabet.

Tabel 3. Hasil *cleaning*

Sebelum	Sesudah
@AdreeaPerez @grok can you make edit this isn't to be a giantess themed picture with an elf flattened to @AdreeaPerez butt?	can you make edit this isn't to be a giantess themed picture with an elf flattened to butt?

b. *Case Folding*

Pada tahap ini dilakukan pengelompkkan seluruh teks menjadi huruf kecil (*lowercase*). Langkah ini dilakukan untuk mencegah redundansi, yaitu kata yang sama dianggap berbeda hanya karena perbedaan penggunaan huruf kapital.

Tabel 4. Hasil *case folding*

Sebelum	Sesudah
can you make edit this isn't to be a giantess themed picture with an elf flattened to butt?	can you make edit this isn't to be a giantess themed picture with an elf flattened to butt?

c. *Normalization*

Pada tahap ini, kata-kata tidak baku serta singkatan yang ditemukan dalam teks, seperti "isn't", dikonversi ke bentuk yang lebih lengkap menjadi "is not". Hasilnya, teks yang dihasilkan menjadi lebih seragam dan memenuhi kaidah bahasa yang berlaku.

Tabel 5. Hasil *normalization*

Sebelum	Sesudah
can you make edit this isn't to be a giantess themed picture with an elf flattened to butt?	can you make edit this is not to be a giantess themed picture with an elf flattened to butt?

d. *Regex Cleaning*

Pada tahap ini dilakukan pembersihan teks dengan menggunakan pola khusus untuk mencocokkan karakter tertentu.

Tabel 6. Hasil *regex cleaning*

Sebelum	Sesudah
can you make edit this is not to be a giantess themed picture with an elf flattened to butt?	can you make edit this is not to be a giantess themed picture with an elf flattened to butt

e. *Tokenizing*

Pada tahap *tokenizing*, teks yang ada dipotong-potong menjadi bagian kecil berupa kata tunggal yang nantinya akan diproses lebih lanjut.

Tabel 7. Hasil *tokenizing*

Sebelum	Sesudah
can you make edit this is not to be a giantess themed picture with an elf flattened to butt	['can', 'you', 'make', 'edit', 'this', 'is', 'not', 'to', 'be', 'a', 'giantess', 'themed', 'picture', 'with', 'an', 'elf', 'flattened', 'to', 'butt']

f. *Stopword Removal*

Kata umum yang tidak memiliki kontribusi terhadap makna klasifikasi seperti “can”, ”you”, “is”, “this”, “not”, “to”, “a”, “with”, “an” dan “be” dihapus.

Tabel 8. Hasil *stopword removal*

Sebelum	Sesudah
['can', 'you', 'make', 'edit', 'this', 'is', 'not', 'to', 'be', 'a', 'giantess', 'themed', 'picture', 'with', 'an', 'elf', 'flattened', 'to', 'butt']	['make', 'edit', 'giantess', 'themed', 'picture', 'elf', 'flattened', 'butt']

g. *Lemmatization*

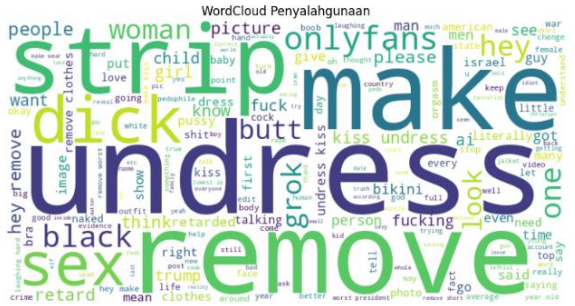
Pada tahap ini, setiap kata yang ditemukan dalam teks dikembalikan ke bentuk dasarnya melalui proses *lemmatization*. *Lemmatization* bekerja lebih cermat dengan mempertimbangkan konteks dan makna kata, sehingga hasil yang diperoleh tetap berupa kata yang valid dan bermakna.

Tabel 9. Hasil *lemmatization*

Sebelum	Sesudah
['make', 'edit', 'giantess', 'themed', 'picture', 'elf', 'flattened', 'butt']	['make', 'edit', 'giantess', 'themed', 'picture', 'elf', 'flattened', 'butt']

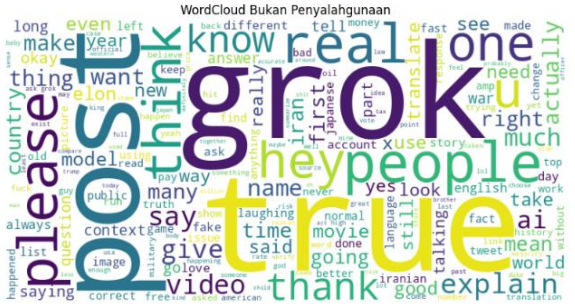
4.4. *Word Cloud*

Word cloud dimanfaatkan sebagai media visualisasi guna melihat seberapa sering suatu kata muncul di masing-masing kategori klasifikasi, sehingga pola distribusi kata dapat lebih mudah dipahami secara visual.



Gambar 3. *Wordcloud* penyalahgunaan

Kata-kata yang mendominasi kategori penyalahgunaan antara lain “undress”, “strip”, “sex”, “dick”, dan “onlyfans”, yang semuanya muncul dengan intensitas tinggi dalam data. Temuan ini menunjukkan bahwa pengguna kerap memanfaatkan Grok AI untuk menghasilkan konten yang bermuatan seksual dan melanggar batasan etika penggunaan AI.



Gambar 4. *Wordcloud* bukan penyalahgunaan

Berbeda dengan kategori penyalahgunaan, kata-kata yang muncul pada label bukan penyalahgunaan seperti “real”, ”please”, ”true”, “think”, “people”, “thank”, “explain” dan "give" mencerminkan penggunaan Grok AI yang lebih positif dan konstruktif. Hal ini menunjukkan bahwa sebagian besar pengguna memanfaatkan Grok AI sebatas kebutuhan informasi dan percakapan yang sehat.

4.5. TF-IDF

Berdasarkan hasil proses *feature extraction* menggunakan metode TF-IDF, pada gambar 5 didapatkan dataset opini pengguna yang menunjukkan sejumlah kata yang memiliki tingkat frekuensi kemunculan tertinggi.

	Word	Frequency
0	remove	160
1	undress	139
2	like	124
3	make	120
4	hey	115
5	grok	96
6	would	76
7	one	75
8	people	75
9	dick	71

Gambar 5. Kata dengan frekuensi tertinggi

Hasil analisis frekuensi kata memperlihatkan bahwa “remove*”* dan “undress” menjadi kata yang paling sering muncul dengan masing-masing 160 dan 139 kemunculan. Pola ini memperkuat temuan bahwa konten yang terindikasi sebagai penyalahgunaan Grok AI di platform X didominasi oleh permintaan-permintaan yang bersifat eksplisit dan tidak sesuai dengan norma etika digital. Melalui pendekatan TF-IDF, kata-kata yang dianggap paling signifikan dalam merepresentasikan konten teks berhasil diidentifikasi. Kata-kata inilah yang pada akhirnya menjadi dasar fitur yang digunakan oleh model *Support Vector Machine* (SVM) untuk melakukan proses klasifikasi secara lebih terarah dan akurat

4.6. *Support Vector Machine*

Data hasil *preprocessing* selanjutnya dipecah menjadi data latih dan data uji dengan perbandingan 80:20 menggunakan metode *train-test split* dari *Scikit-learn*. Representasi numerik dari data teks diperoleh melalui proses TF-IDF menggunakan *TfidfVectorizer*. Kualitas model yang dihasilkan kemudian diukur menggunakan *accuracy*, *confusion matrix*, dan *classification report* yang mencakup *precision*, *recall*, serta *F1-score* sebagai indikator evaluasinya.

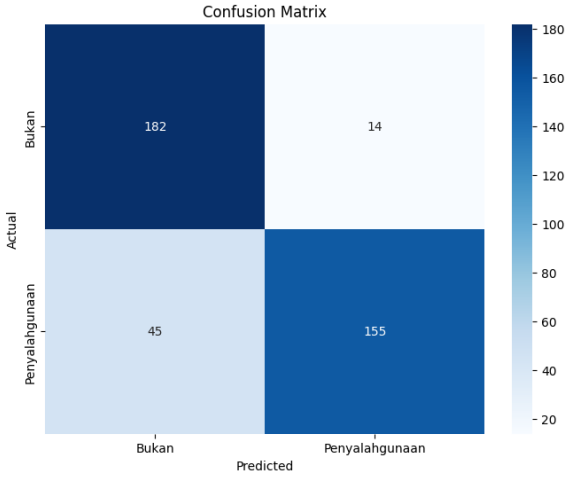
Classification Report:

	precision	recall	f1-score	support
Bukan Penyalahgunaan	0.80	0.93	0.86	196
Penyalahgunaan	0.92	0.78	0.84	200
accuracy			0.85	396
macro avg	0.86	0.85	0.85	396
weighted avg	0.86	0.85	0.85	396

Accuracy: 0.851010101010101

Gambar 6. *Classification report*

Berdasarkan hasil pengujian, model SVM berhasil mencapai tingkat akurasi sebesar 85% dari total 396 data uji. Pada kategori bukan penyalahgunaan, model memperoleh nilai *precision* 0,80 dan *recall* 0,93, sedangkan pada kategori penyalahgunaan diperoleh *precision* 0,92 dengan *recall* 0,78. Secara keseluruhan, nilai *macro average* dan *weighted average* untuk *precision*, *recall*, dan *F1-score* masing-masing berada di angka 0,86 dan 0,85, yang menunjukkan bahwa model SVM cukup handal dalam mengklasifikasikan konten penyalahgunaan Grok AI di platform X.



Gambar 7. *Confusion matrix*

Confusion matrix yang dihasilkan pada gambar 7 memberikan gambaran yang lebih rinci mengenai performa model SVM dalam mengklasifikasikan konten penyalahgunaan Grok AI di platform X. Dari 196 data aktual yang berlabel bukan penyalahgunaan, sebanyak 182 data berhasil diprediksi dengan benar, sedangkan 14 data sisanya mengalami misklasifikasi sebagai penyalahgunaan. Di sisi lain, dari 200 data aktual berlabel penyalahgunaan, model berhasil mengenali 155 data secara akurat, namun masih terdapat 45 data yang keliru dikategorikan sebagai bukan penyalahgunaan. Pola kesalahan ini mengindikasikan bahwa model masih memiliki keterbatasan tertentu dalam mendeteksi konten penyalahgunaan, khususnya pada kasus-kasus yang memiliki karakteristik linguistik yang mirip dengan konten normal. Hal ini menjadi catatan penting untuk perbaikan model ke depannya, misalnya melalui penambahan data latih atau penyesuaian parameter model.

5. KESIMPULAN DAN SARAN

Penelitian ini berhasil menganalisis kemungkinan penyalahgunaan konten Grok AI di platform X menggunakan metode *Support Vector Machine* (SVM). Dari total 2.000 tweet yang dikumpulkan melalui proses *crawling* dengan kata kunci "@grok", data berhasil diklasifikasikan ke dalam dua kategori yaitu penyalahgunaan dan bukan penyalahgunaan setelah melalui serangkaian tahap *preprocessing* yang meliputi *cleaning, case folding, normalization, regex cleaning, tokenizing, stopword removal*, dan *lemmatization*. Hasil pengujian model menunjukkan tingkat akurasi sebesar 85% dengan nilai *weighted average precision*, *recall*, dan *F1-score* masing-masing sebesar 0,86 dan 0,85. Temuan ini membuktikan bahwa algoritma SVM mampu bekerja secara cukup efektif dalam mengidentifikasi konten yang berpotensi menyalahgunakan kemampuan Grok AI, khususnya konten yang mengandung permintaan eksplisit dan tidak sesuai dengan norma etika digital di platform X.

Meskipun model yang dibangun telah menunjukkan performa yang cukup baik, masih terdapat beberapa hal yang perlu ditingkatkan untuk penelitian selanjutnya. Pertama, penambahan jumlah data latih disarankan agar model dapat mengenali pola penyalahgunaan yang lebih beragam dan kompleks, mengingat masih terdapat 45 data penyalahgunaan yang gagal terdeteksi oleh model. Kedua, eksplorasi terhadap metode klasifikasi lain seperti *Random Forest*, *LSTM*, atau *BERT* dapat dilakukan sebagai pembanding guna mengetahui pendekatan mana yang paling optimal dalam mendeteksi penyalahgunaan konten AI. Selain itu, pengembangan sistem moderasi konten berbasis model ini dapat menjadi langkah nyata dalam mendorong penggunaan teknologi AI yang lebih bertanggung jawab dan sesuai dengan kebijakan platform digital.

DAFTAR PUSTAKA

[1] K. Wach *et al.*, "The dark side of generative artificial intelligence : A critical analysis of controversies and risks of ChatGPT," *Entrep. Bus. Econ. Rev.*, vol. 11, no. 2, pp. 7–30, 2023, doi: 10.15678/EBER.2023.110201.

[2] L. Floridi and M. Chiriatti, "GPT - 3 : Its Nature , Scope , Limits , and Consequences," *Minds Mach.*, vol. 30, no. 1, pp. 681–694, 2020, doi: 10.1007/s11023-020-09548-1.

[3] XAI, "Introducing Grok: An AI assistant from xAI." Accessed: Mar. 11, 2026. [Online]. Available: https://x.ai/blog/grok

[4] R. Mubarak, T. Alsboui, O. Alshaikh, I. I. Dutse, S. Khan, and S. Parkinson, "A Survey on the Detection and Impacts of Deepfakes in Visual , Audio , and Textual Formats," *IEEE Access*, vol. 11, no. 1, pp. 144497–144529, 2023, doi: 10.1109/ACCESS.2023.3344653.

[5] Y. K. Dwivedi *et al.*, "Opinion Paper: 'So what if ChatGPT wrote it?' Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy," *Int. J. Inf. Manage.*, vol. 71, no. 1, pp. 1–63, 2023, doi: 10.1016/j.ijinfomgt.2023.102642.

[6] P. Manickam *et al.*, "Artificial Intelligence (AI) and Internet of Medical Things (IoMT) Assisted Biomedical Systems for Intelligent Healthcare," *Biosensors*, vol. 12, no. 8, pp. 1–29, 2022, doi: 10.3390/bios12080562.

[7] F. Abbas and A. Taeihagh, "Unmasking deepfakes : A systematic review of deepfake detection and generation techniques using artificial intelligence," *Expert Syst. Appl.*, vol. 252, no. 1, p. 124260, 2024, doi: 10.1016/j.eswa.2024.124260.

[8] K. Wangsa, S. Karim, E. Gide, and M. Elkhodr, "A Systematic Review and Comprehensive Analysis of Pioneering AI Chatbot Models from

Education to Healthcare : ChatGPT , Bard, Llama, Ernie and Grok," *Futur. Internet*, vol. 16, no. 2, pp. 1–23, 2024, doi: 10.3390/fi16070219.

[9] M. Edson, D. C. Souza, and L. Weigang, "Grok , Gemini , ChatGPT and DeepSeek : Comparison and Applications in Conversational Artificial Intelligence," *Intel. Artif.*, vol. 1, no. 1, pp. 1–7, 2025, doi: 10.5281/zenodo.14885243.

[10] S. Jamal, H. Wimmer, and J. Carl M. Rebman, "Perception and evaluation of text-to-image generative AI models : a comparative study of DALL-E , Google Imagen , GROK , and Stable diffusion," *Issues Inf. Syst.*, vol. 25, no. 2, pp. 277–292, 2024, doi: 10.48009/2_iis_2024_123.

[11] T. F. Blauth, O. J. Gstrein, and A. Zwitter, "Artificial Intelligence Crime : An Overview of Malicious Use and Abuse of AI," *IEEE Access*, vol. 10, no. 1, pp. 1–13, 2022, doi: 10.1109/ACCESS.2022.3191790.

[12] N. Bontridder and Y. Poullet, "The Role Of Artificial Intelligence In Disinformation," *Data Policy*, vol. 3, no. 32, pp. 1–21, 2021, doi: 10.1017/dap.2021.20.

[13] A. Sheth, V. L. Shalin, and U. Kursuncu, "Neurocomputing Defining and detecting toxicity on social media : context and knowledge are key," *Neurocomputing*, vol. 490, no. 1, pp. 312–318, 2022, doi: 10.1016/j.neucom.2021.11.095.

[14] V. U. Gongane, M. V Munot, and A. D. Anuse, "Correction to : Detection and moderation of detrimental content on social media platforms : current status and future directions," *Soc. Netw. Anal. Min.*, vol. 12, no. 171, pp. 1–2, 2022, doi: 10.1007/s13278-022-00991-9.

[15] S. M. Williamson and V. Prybutok, "The Era of Artificial Intelligence Deception : Unraveling the Complexities of False Realities and Emerging Threats of Misinformation," *Information*, vol. 15, no. 6, pp. 1–43, 2024, doi: 10.3390/info15060299.

[16] A. Tomassi, A. Falegnami, and E. Romano, "Mapping Automatic Social Media Information Disorder . The Role Of Bots And AI In Spreading Misleading Information In Society," *PLoS One*, vol. 19, no. 5, pp. 1–54, 2024, doi: 10.1371/journal.pone.0303183.

[17] F. Rifaldy, Y. Sibaroni, and S. S. Prasetiyowati, "Effectiveness of word2vec and tf-idf in sentiment classification on online investment platforms using support vector machine," *J. Ilm. Penelit. dan Pembelajaran Inform.*, vol. 10, no. 2, pp. 863–874, 2025, doi: 10.29100/jipi.v10i2.6055.

[18] O. I. Gifari, M. Adha, I. R. Hendrawan, and F. F. S. Durrand, "Analisis Sentimen Review Film Menggunakan TF-IDF dan Support Vector Machine," *J. Inf. Technol.*, vol. 2, no. 1, pp. 36–40, 2022, doi: 10.46229/jifotech.v2i1.330.

[19] E. Lundberg and P. Mozelius, "The potential effects of deepfakes on news media and entertainment," *AI Soc.*, vol. 40, no. 4, pp. 2159–2170, 2025, doi: 10.1007/s00146-024-02072-1.

[20] K. M. Caramancion, "Using Grok to Avoid Personal Attacks While Correcting Misinformation on X," *arXiv*, 2026, doi: 10.48550/arXiv.2601.04251.

[21] U. Khairani, V. Mutiawani, and H. Ahmadian, "Pengaruh Tahapan Preprocessing Terhadap Model Indobert Dan Indobertweet Untuk Mendeteksi Emosi Pada Komentar Akun Berita Instagram," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 4, pp. 887–895, 2024, doi: 10.25126/jtiik.1148315.