PENGEMBANGAN MODEL KLASIFIKASI TOKSISITAS MULTIMODAL PADA PLATFORM SOSIAL MEDIA MENGGUNAKAN *LARGE LANGUAGE MODEL* (LLM) DENGAN KOMBINASI JENIS MEDIA TEKS, GAMBAR DAN VIDEO

BAB I
PENDAHULUAN

1.1. Latar Belakang

Perkembangan teknologi internet dan aplikasinya telah berkembang sangat pesat dan memberikan dampak yang cukup hebat. Salah satu teknologi yang memberi dampak tersebut adalah layanan sosial media, sosial media merupakan salah satu wadah yang disediakan untuk berbagi konten dan berinteraksi sosial untuk mengekspresikan pemikiran, ide-ide, foto dan video. Sosial media memiliki banyak aspek positif, salah satunya adalah rasa kebersamaan yang diberikan kepada masyarakat. Orang-orang dari semua lapisan masyarakat di seluruh dunia dapat terhubung dengan individu yang tepat dan membangun jaringan yang saling menguntungkan. Menurut (Akhsi dan Nitin, 2021) Meningkatnya ketersediaan layanan data yang wajar dan kehadiran sosial media telah memberikan dampak tanpa hambatan di mana pengguna online telah menemukan cara-cara yang salah dan melanggar hukum untuk menyakiti dan mempermalukan individu melalui komentar kebencian di *platform* atau aplikasi *online*. Menurut (Regiolina et all, 2023) Terlepas dari kenyataan bahwa sosial media menawarkan banyak hal baik bagi dunia, sosial media juga memiliki sejumlah aspek negatif. Terkadang komentar dan diskusi terbuka bisa memicu perdebatan, bisa karena perbedaan pendapat atau karena kesal dengan konten yang disajikan. Namun seringkali perdebatan yang terjadi muncul hal-hal yang tidak baik dan menggunakan cara-cara yang kotor untuk berdebat.

Pada umumnya penggunaan sosial media dibuat untuk akun pribadi, untuk masyarakat biasa ataupun artis hingga menjadi sebuah bisnis. Banyak pengguna sosial media yang belum memahami etika-etika dalam bersosialisasi pada dunia maya. Toksisitas atau perilaku beracun di sosial media sudah menjadi hal yang biasa, namun hal ini semakin tidak dapat ditoleransi. Toksisitas dalam lingkup sosial dapat digambarkan sebagai penyebaran hal-hal negatif atau kebencian yang tidak perlu yang pada akhirnya berdampak negatif pada orang-orang yang mengalaminya. Toksisitas di

sosial media ini berupaya menyebarkan ujaran kebencian dan melecehkan orang lain dalam sebuah diskusi. Toksisitas sering kali menyebar dalam bentuk teks. Namun, Internet dan sosial media memungkinkan penggunaan berbagai cara yang dapat membuat toksisitas menjadi lebih parah dan berdampak, misalnya pada sebuah meme dalam bentuk gambar atau video yang lebih mudah dikonsumsi dan menarik lebih banyak perhatian. Menurut (Revati & Meetkumar, 2018) klasifikasi komentar beracun online dapat dibagi tingkat toksisitasnya menjadi 6 label yang sudah disediakan oleh kumpulan data di platform Kaggle, yaitu *toxic*, *severe-toxic*, pelecehan, ancaman, penghinaan dan ujaran kebencian terhadap identitas.

Penelitian yang dilakukan oleh (Akhsi & Nitin, 2021) Menyajikan model CNN untuk mendeteksi *cyberbullying* dalam tiga modalitas data sosial yang berbeda, yaitu tekstual, visual dan infografis (teks yang disematkan bersama gambar). Penelitian ini menggunakan arsitektur CapsNet-ConvNet, terdiri dari jaringan saraf dalam jaringan Capsule (CapsNet) dengan perutean dinamis untuk memprediksi konten intimidasi tekstual dan jaringan saraf konvolusi (ConvNet) untuk memprediksi konten intimidasi visual. Konten infografis didiskritisasi dengan memisahkan teks dari gambar menggunakan Google Lens dari Aplikasi Google Foto. Evaluasi eksperimental dilakukan pada kumpulan data modal campuran yang berisi 10.000 komentar dan postingan yang diambil dari YouTube, Instagram, dan Twitter. Model yang diusulkan mencapai kinerja superlatif dengan AUC-ROC sebesar 0,98.

Metode *deep learning* terbukti berguna dan memperoleh hasil canggih untuk berbagai tugas bahasa alami dengan pelatihan ujung ke ujung dan kemampuan pembelajaran representasi (Tom et all, 2017). Studi terkait melaporkan penggunaan model *deep learning* seperti CNN, RNN, dan fitur gambar semantik untuk mendeteksi konten intimidasi dengan menganalisis fitur tekstual, berbasis gambar, dan pengguna (Akhsi dan Nitin, 2021). *Deep learning* sering digunakan untuk berbagai aplikasi seperti pengenalan wajah, deteksi objek, pemrosesan bahasa alami, dan banyak lagi. Perkembangan *Deep learning* sudah semakin pesat, salah satunya mengenai multimodal *learning*. Dimana model dapat menangani berbagai jenis data (misalnya teks, gambar, suara dan video) secara bersamaan.

Menurut (Firoj et all, 2022) Deteksi disinformasi multimodal yang mencakup berbagai kombinasi modalitas: teks, gambar, ucapan, video, struktur jaringan media sosial, dan informasi temporal. Kecanggihan deteksi disinformasi multimodal berdasarkan penelitian sebelumnya mengenai berbagai modalitas, dengan fokus pada disinformasi, yaitu informasi yang salah dan bertujuan untuk merugikan. Survei ini menghadirkan beberapa tantangan penelitian yang menarik untuk deteksi disinformasi multimodal, seperti menggabungkan berbagai modalitas, yang seringkali tidak selaras dan berada dalam representasi yang berbeda, misalnya teks vs gambar atau teks vs video dll. Menurut (Anastasia et all, 2020) penggabungan fitur dari berbagai komponen efektif untuk pendeteksian berita palsu dan menggabungkan fitur dari berbagai gambar lebih efektif daripada menggunakan fitur visual hanya dari satu gambar.

Multimodal juga dapat dilakukan dengan menggunakan *Large Language Models* (LLM). *Large Language Models* merupakan model kecerdasan buatan yang dirancang untuk menangani dan memproses lebih dari satu jenis data atau modalitas, seperti teks, gambar, video, dan audio. *Large Language Models* (LLM) merupakan kemajuan luar biasa dalam pemrosesan bahasa alami dan penelitian kecerdasan buatan (M. Usman et all, 2023). Model-model ini telah meningkatkan kemampuan mesin secara signifikan untuk memahami dan menghasilkan bahasa seperti manusia (Jie & Kevin, 2023). Dengan memanfaatkan teknik *Deep learning* dan kumpulan data yang luas, LLM telah menunjukkan kemahirannya dalam berbagai tugas yang berhubungan dengan bahasa, termasuk pembuatan teks, penerjemahan, peringkasan, menjawab pertanyaan, dan analisis sentimen. Berbeda dengan LLM tradisional yang fokus hanya pada teks, multimodal LLM dapat memahami dan mengintegrasikan berbagai jenis informasi untuk meningkatkan pemahaman dan kinerja dalam berbagai tugas. Menurut (Andrei K et all, 2024) Meskipun kinerja LLM dalam pemrosesan dan pembuatan teks sudah mengesankan, ada potensi keuntungan tambahan dalam mengintegrasikan LLM dengan jaringan syaraf lainnya. *Large Language Models* (LLM) juga dapat sangat efektif untuk melakukan tugas klasifikasi, yang merupakan salah satu aplikasi utama dalam pemrosesan bahasa alami (NLP).

Beberapa penelitian dengan topik LLM pernah dilakukan oleh beberapa peneliti,

diantaranya yaitu penelitian yang dilakukan oleh (Chenyang et all, 2023) Mengusulkan MACAW-LLM yaitu LLM multimodal baru yang mengintegrasikan informasi visual, audio dan tekstual. Pada MACAW-LLM terdiri dari tiga komponen utama, yaitu modul modalitas untuk mengodekan data multimodal, modul kognitif untuk memanfaatkan LLM yang telah dilatih sebelumnya dan juga modul penyelarasan untuk menyelaraskan berbagai representasi. Set data instruksi multimodal terdiri dari 69K contoh gambar dan 50K contoh video.

Penelitian yang dilakukan (Feilong at all, 2023) Mengusulkan X-LLM yang dapat mengubah multimodal (gambar, suara dan video) kedalam Bahasa asing yang menggunakan antarmuka X2L dan memasukkannya kedalam model LLM (ChatGLM). Pelatihan X-LLM terdiri dari 3 tahap yaitu (1) Mengkonversi informasi multimodal, yaitu melatih setiap antarmuka X2L agar selaras dengan encoder masing-masing secara terpisah untuk mengkonversi informasi multimodal ke dalam Bahasa. (2) Menyelaraskan representasi X2L dengan LLM, yaitu encoder modal Tunggal diselaraskan dengan LLM melalui antarmuka X2L secara independent. (3) Mengintegrasikan beberapa modalitas, yaitu semua encoder modal tunggal diselaraskan dengan LLM melalui antarmuka X2L untuk mengintegrasikan kapabilitas multimodal ke dalam LLM. Penelitian ini menghasilkan skor relative 84.5%.

Klasifikasi toksisitas adalah proses mengidentifikasi dan mengklasifikasikan konten atau perilaku yang dianggap merugikan, berbahaya, atau mengandung kebencian di lingkungan online. Pada platform sosial media menunjukkan bahwa algoritma klasifikasi dapat digunakan di platform-platform seperti Twitter, Instagram dan Tiktok dimana konten yang dibagikan oleh pengguna sangat beragam dan dapat berupa teks, gambar, atau video. Pengembangan model klasifikasi toksisitas pada platform sosial media dengan kombinasi multimodal yang digunakan merujuk pada upaya untuk menciptakan atau meningkatkan suatu model atau algoritma yang dapat mengidentifikasi dan mengklasifikasikan konten yang bersifat *toxic* atau merugikan di platform-platform sosial media. Dalam konteks ini, sebuah algoritma klasifikasi adalah serangkaian prosedur atau aturan yang digunakan untuk mengidentifikasi dan memisahkan konten menjadi kategori yang relevan, seperti *toxic*, netral atau *non-toxic*.

Toksisitas juga dapat mengacu pada tingkat bahaya atau ketidakamanan konten, seperti kebencian, pelecehan, atau ancaman.

Penelitian dengan model klasifikasi sudah dilakukan oleh (Khairul et all, 2023), Melakukan klasifikasi batik tanah liat Sumatera Barat menggunakan metode CNN. Data yang digunakan penelitian ini adalah 400 citra batik dan dibagi menjadi 4 kelas, ditentukan 320 citra sebagai data latih dan 80 citra sebagai data uji. Hasil pengujian dan pelatihan menggunakan CNN didapat nilai akurasi batik tanah liat Sumatera Barat sebesar 98.75% pada data latih dan 62.5% pada data uji. Tingkat akurasi ini cukup baik sebagai rujukan dalam membangun *real application* pengenalan motif batik secara umum. Hasil ini menunjukkan metode CNN dapat diterapkan untuk mengklasifikasi batik tanah liat Sumatera Barat.

Penelitian oleh (Peiyu & Shuangtao, 2019), Melakukan klasifikasi sentimen multimodal untuk sebuah tweet yang berisi teks dan gambar. Metode yang digunakan adalah model Bidirectional-LSTM untuk mengekstraksi modalitas teks dan model VGG-16 digunakan untuk mengekstraksi fitur modalitas gambar. Menggunakan algoritma fusi digunakan untuk menyelesaikan fitur teks dan gambar. Metode fusi pertama adalah sum, yang berarti fitur global adalah jumlah fitur teks ftext dan fitur gambar fimage, menggunakan metode fusi ini skor mikro f1 adalah 79,6%. Metode fusi kedua adalah *concatena*te, yang berarti fitur global adalah penggabungan fitur teks ftext dan fitur gambar fimage, dan skor mikro f1 adalah 82,3%. Metode fusi ketiga adalah metode fusi berbasis perhatian yang diusulkan dalam makalah ini, yang mencapai skor mikro f1 tertinggi hingga 84,2%

Penelitian juga dilakukan oleh (Hong Fan, 2021) Untuk mendeteksi toksisitas dengan mengadopsi model BERT untuk pengklasifikasian komentar beracun dari data pada media sosial Twitter. Hasil evaluasi menunjukkan bahwa BERT memiliki kemampuan klasifikasi dan memprediksi komentar beracun dengan tingkat akurasi yang tinggi. Selain itu penelitian ini juga membandingkan model berbasis BERT dengan 3 model lainnya yaitu Multilingual BERT, RoBERTa dan DistilBERT. Model berbasis BERT mengunguli semua model yang dibandingkan dan mencapai hasil terbaik.

Pada penelitian ini akan dilakukan pengembangan klasifikasi toksisitas

multimodal pada platform sosial media menggunakan *Large Language Model* (LLM) dengan kombinasi jenis media teks, gambar dan video. Hasil dari klasifikasi toksisitas nantinya berupa label yang menunjukkan apakah konten atau perilaku tersebut dianggap *toxic*, *non-toxic*, dan netral. Informasi ini dapat digunakan untuk memicu tindakan seperti penyaringan konten, menghapus konten yang melanggar kebijakan, atau memberikan peringatan kepada pengguna yang melanggar aturan. Dengan demikian, klasifikasi toksisitas merupakan komponen penting dari upaya untuk menjaga lingkungan *online* yang aman, positif, dan inklusif.

1.2. Rumusan Masalah

Rumusan masalah pada penelitian ini difokuskan pada :

1. Bagaimana mengembangkan algoritma untuk merepresentasikan teks, gambar dan video dari berbagai platform sosial media?
2. Bagaimana mengembangkan model klasifikasi toksisitas multimodal dengan metode *Large Language Model* (LLM) dengan jenis media teks, gambar dan video?

1.4 Tujuan Penelitian

Tujuan dalam penelitian ini adalah melakukan pengembangan model klasifikasi toksisitas multimodal dengan kombinasi jenis media teks, gambar dan video pada platform sosial media menggunakan *Large Language Model* (LLM). Pengembangan ini dilakukan untuk meningkatkan kemampuan deteksi dan pemahaman konten berbahaya secara menyeluruh di platform sosial media. Secara khusus dapat dijabarkan sebagai berikut :

1. Mengembangkan algoritma untuk merepresentasikan teks, gambar dan video dari berbagai platform sosial media seperti Twitter, Instagram dan Tiktok.
2. Mengembangkan model klasifikasi toksisitas menggunakan multimodal dengan metode *Large Language Model* (LLM) dengan jenis media teks, gambar dan video.

Pengembangan menggunakan *Large Language Model* (LLM) ini mampu menjembatani modalitas dengan berbagai jenis media yang digunakan seperti teks, gambar, video dan suara. Penelitian yang sudah dilakukan (Feilong, 2023) dapat mengubah multimodal (gambar, suara dan video) kedalam bahasa Asing. Penelitian yang dilakukan (Junnan, 2023) menggunakan model BLIP-2, dimana BLIP-2 ini mampu menjembatani modalitas dengan transformator kueri yang sudah dilatih. Penelitian yang dilakukan (Chenyang, 2023) Mampu mengintegrasikan informasi visual, audio dan tekstual menggunakan MACAW-LLM. Meskipun kinerja LLM dalam pemrosesan dan

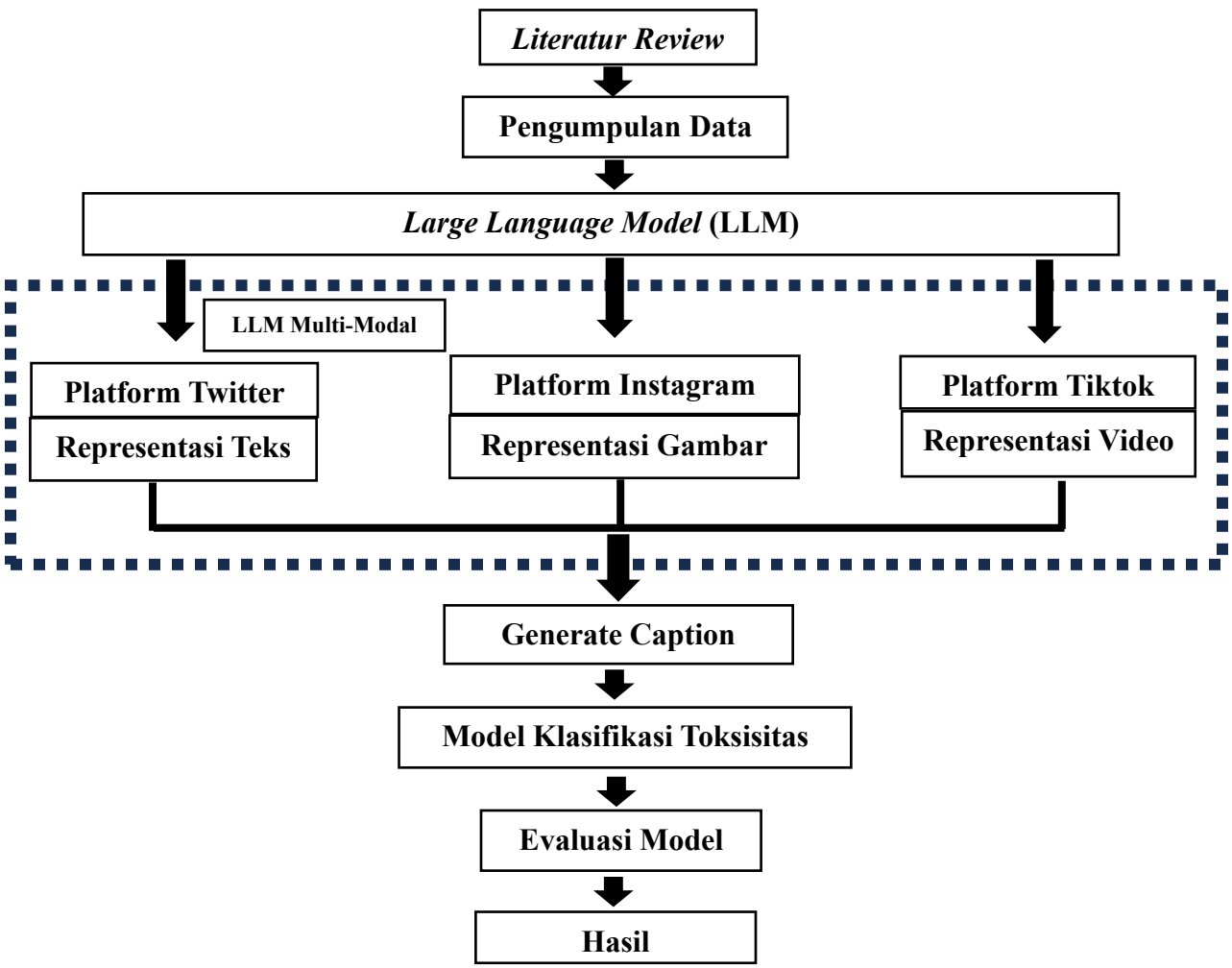
pembuatan teks sudah mengesankan, ada potensi keuntungan tambahan dalam mengintegrasikan LLM dengan jaringan syaraf lainnya. Menurut (Hong Fan, 2021) Melakukan penelitian untuk mendeteksi toksisitas menggunakan BERT. Hasil evaluasi dalam melakukan klasifikasi menggunakan BERT menunjukkan bahwa BERTmemiliki kemampuan klasifikasi dan memprediksi komentar beracun dengan tingkat akurasi yang tinggi. Dengan metode ini, peneliti ingin melakukan pengembangan klasifikasi multimodal dengan jenis media teks, image dan video menggunakan LLM agar nantinya dapat lebih efektif dalam menangani toksisitas di lingkup sosial media.

BAB III
METODE PENELITIAN

Bab ini akan menjelaskan tentang metodologi penelitian yang digunakan sebagai gambaran dari langkah-langkah yang akan dilakukan untuk menyelesaikan penelitian ini.

3.1 Tahapan Penelitian

Penelitian ini melakukan pengembangan model klasifikasi toksisitas pada *platform* sosial media. Tahapan penelitian yang digunakan dapat dilihat pada gambar 3.1.



Gambar 3.1 Tahapan Metode Penelitian

Tahapan metode penelitian pada gambar 3.1 terdiri dari beberapa langkah, yaitu :

1. **Tahap *Literature Review***

   Pada tahap ini dimulai dengan melakukan kajian dari berbagai sumber tertulis dalam bentuk buku, artikel dan jurnal serta penelitian-peneltiian terkait guna memahami dan mengidentifikasi kesenjangan dalam topik penelitian serta menemukan kelemahan dan kelebihan dalam penelitian. Selain itu juga untuk menentukan dan membandingkan metode serta algoritma yang sudah digunakan pada penelitian sebelumnya, yang nantinya akan mengembangkan atau menciptakan suatu metode atau algoritma terbaru.

2. **Tahap Pengumpulan Data**

   Pada tahap ini dilakukan pengumpulan data yang akan digunakan untuk melatih dan menguji model. Data ini dapat berupa konten-konten pada sosial media yang akan dikategorikan ke dalam 3 kategori toksisitas yaitu *toxic*, *non-toxic*, dan netral. Data tersebut harus mencakup berbagai jenis media, seperti teks, gambar dan video, untuk memungkinkan model mengenali toksisitas dari berbagai jenis konten yang ada pada platform sosial media.

3. ***Large Language Model* (LLM)**

   Large Language Model merupakan jenis model kecerdasan buatan (*Artificial Intelligence*) yang dilatih untuk memahami, menghasilkan dan memproses bahasa alami (*Natural Languange*) dalam skala besar. *Large Language Model* dilatih menggunakan dataset yang besar, terdiri dari teks yang diambil dari berbagai sumber seperti artikel, buku, situs web dan lainnya.

4. ***Large Language Model* (LLM) Multimodal**

   Large Language Model pada penelitian yang dilakukan untuk memproses tidak hanya dalam bentuk jenis media teks, melainkan gambar dan juga video. *Large Language Model* (LLM) untuk klasifikasi multimodal melibatkan teks, gambar dan juga video. Meskipun LLM berfokus pada teks, model ini dapat diadaptasi atau dikombinasikan dengan model lain yang mendukung modalitas non-teks seperti gambar dan video, melalui pendekatan yang disebut dengan model multimodal. Pada

tahap ini dilakukan *pre-processing* dari masing-masing jenis media yang digunakan.

- Untuk representasi teks menggunakan teknik-teknik pemrosesan bahasa alami seperti tokenisasi, vektorisasi data (*word embedding*) dan penggunaan model bahasa *pre-trained* seperti BERT untuk mewakili teks dalam bentuk vektor numerik yang dapat dimengerti oleh model.

7. Untuk representasi gambar dan video menggunakan teknik-teknik pemrosesan gambar seperti ekstraksi fitur dengan *convolutional neural networks* (CNN) atau menggunakan model *pre-trained* seperti ResNet atau VGG untuk mewakili gambar dalam bentuk vektor numerik.

5. Generate Caption

Pada tahapan ini menggunakan model LLM, seperti BLIP atau Flamingo untuk menggabungkan kemampuan visual dan bahasa dalam menghasilkan teks/*captioning* dari representasi gambar dan video.

6. Model Klasifikasi Toksisitas

Pada tahapan ini dilakukan pengembangan model dari hasil penggabungan ketiga representasi tersebut, dengan menggunakan teknik *fusion*, seperti *concatenation* atau *attention mechanism* untuk menghasilkan hasil klasifikasi akhir. Model klasifikasi yang digunakan adalah *Convolutional Neural Network* (CNN).

7. Evaluasi Model

Pada tahapan ini dilakukan evaluasi untuk mengetahui kinerja terhadap model yang dikembangkan dengan menggunakan pengukuran akurasi, seperti *precision*, *recall* dan juga *F1-score* untuk klasifikasi teks, dan mengukur akurasi dengan *confusion matrix* untuk gambar dan video.

8. Hasil

Tahapan ini menghasilkan klasifikasi sesuai dengan label yang sudah dikategorikan ke dalam 3 kategori toksisitas yaitu *toxic*, *non-toxic*, dan netral.