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

**Witta Listiya Ningrum**

**99217021**

PROGRAM DOKTOR TEKNOLOGI INFORMASI

UNIVERSITAS GUNADARMA

2024

# DAFTAR ISI

# DAFTAR GAMBAR

# DAFTAR TABEL

# 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?

## 1.3. Batasan Masalah

Penelitian ini merupakan pengembangan model klasifikasi yang di fokuskan pada :

1. Model klasifikasi dibagi menjadi 3 kategori, yaitu *toxic*, *non-toxic*, dan netral.
2. Tingkat toksisitas dari kategori *toxic* dibagi menjadi 2, yaitu ujaran kebencian dan *bullying.*
3. Data set didapatkan dari platform sosial media yang terdiri dari Twitter, Instagram dan Tiktok dalam Bahasa Indonesia.
4. Jenis media yang digunakan adalah teks, gambar dan video.
5. Model LLM yang digunakan untuk representasi teks yaitu BERT, untuk representasi gambar dan video yaitu BLIP atau Flamingo.
6. Model yang digunakan dalam klasifikasi adalah *Convolutional Neural Network*

(CNN).

## 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.

## 1.3 Manfaat Dan Kontribusi Penelitian

Penelitian ini diharapkan memberi kontribusi pengembangan pada bidang-bidang berikut :

1. Pada bidang teknologi menghasilkan metode atau model klasifikasi yang dapat digunakan untuk mendeteksi toksisitas pada platform sosial media.
2. Pada bidang ilmu pengetahuan dapat memberikan kontribusi berupa pemahaman mengenai proses dan pengembangan serta pembentukan klasifikasi multimodal menggunakan *Large Language Model* (LLM) dengan kombinasi jenis media teks, gambar dan video.
3. Pada bidang akademik, penelitian ini dapat dijadikan referensi bagi peneliti yang berminat pada bidang *Information Retrieval*, *Artificial Intelligence*, *Expert System* dan *Decision Support System*.
4. Bagi masyarakat dengan menggunakan klasifikasi toksisitas dapat memberikan pencegahan penyebaran konten berbahaya, dimana dengan

klasifikasi yang lebih akurat dan luas, platform sosial media dapat lebih cepat dan efektif mengidentifikasi serta memoderasi konten *toxic* sebelum menyebar luas sehingga membantu mengurangi dampak negatif terhadap pengguna dan komunitas *online.*

# BAB II

# TELAAH PUSTAKA

## 2.1 Sosial Media

Van Dijk (dalam Rully, 2015) menyatakan bahwa sosial media adalah *platform* media yang memfokuskan pada eksistensi pengguna yang memfasilitasi mereka dalam beraktifitas maupun berkolaborasi. Karena itu media social dapat dilihat sebagai medium (fasilitator) online yang menguatkan hubungan antar pengguna sekaligus sebuah ikatan sosial.

Meike dan Young (dalam Rully, 2015) mengartikan kata sosial media sebagai konvergensi antara komunikasi personal dalam arti saling berbagi diantara individu (*to be share one-to-one*) dan media publik untuk berbagi kepada siapa saja tanpa ada kekhususan individu.

Menurut Boyd (dalam Rully, 2015) sosial media sebagai kumpulan perangkat lunak yang memungkinkan individu maupun komunitas untuk berkumpul, berbagi, berkomunikasi, dan dalam kasus tertentu saling berkolaborasi atau bermain. Media sosial memiliki kekuatan pada *user-generated content* (UGC)dimana konten dihasilkan oleh pengguna, bukan oleh editor sebagaimana di instansi media massa.

## 2.2 Toksisitas

Sentimen *toxic* mengacu pada komentar atau perilaku dalam konteks online yang bersifat merendahkan, menghina atau merugikan individu atau kelompok lain. Ini ada adalah bentuk dari perilaku yang tidak pantas atau tidak menghormati dalam interaksi online, dan dapat muncul dalam berbagai bentuk dan konteks. Sentimen *toxic* dapat muncul dalam berbagai konteks online, termasuk dalam komentar di sosial media, forum diskusi, pesan email atau platform daring lainnya.

Menurut (Reinert, 2020) Toksisitas dapat berisi kata-kata ancaman, cabul, penghinaan atau kebencian terhadap identitas, sehingga akan menimbulkan pelecehan di

sosial media, atau biasa disebut pelecehan *online*. Akibat dari tindakan pelecehan tersebut, sebagian orang akan berhenti memberikan pendapat atau berusaha menghindari perdebatan di media sosial yang berujung pada diskusi yang tidak sehat dan tidak adil.

Menurut (Suresh et.all, 2019) dengan menggunakan CNN dapat mengembangkan model multi-label yang mengklasifikasikan sebuah komentar berdasarkan tingkat toksisitasnya menjadi 6 kategori, yaitu *toxic, severe-toxic,* pelecehan, ancaman, penghinaan dan ujaran kebencian terhadap identitas.

Menurut (Nurul et.all, 2021) analisis sentimen untuk komentar beracun menggunakan data multi-label yang dibagi kedalam 4 label yaitu pencemaran nama baik, pornigrafi, radikalisme, dan SARA.

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.

### *2.3 Machine Learning*

*Machine learning* dilakukan melalui 2 fase yaitu fase training dan fase *application*. Fase training adalah proses pemodelan dari algoritma yang digunakan akan dipelajari oleh sistem melalui training data, sedangkan fase *application* adalah proses pemodelan yang telah dipelajari sistem melalui fase training akan digunakan untuk menghasilkan sebuah keputusan tertentu, dengan menggunakan testing data.

*Machine learning* dapat dilakukan dengan dua cara, yaitu *supervised learning* dan *unsupervised learning*. *Unsupervised learning* adalah pemrosesan sample data dilakukan tanpa mewajibkan hasil akhir memiliki bentuk yang sesuai dengan bentuk tertentu, dengan meggunakan beberapa sample data sekaligus. Penerapan *unsupervised learning* dapat ditemukan pada proses visualisasi, atau eksplorasi data. *Supervised learning* adalah pemrosesan sample data x akan diproses sedemikian rupa, sehingga menghasilkan output

yang sesuai dengan hasil akhir. *Supervised learning* dapat diterapkan pada proses klasifikasi.

Menurut (Jiawei et.al, 2012) *machine learning* dapat didefinisikan sebagai metode komputasi berdasarkan pengalaman untuk meningkatkan performa atau membuat prediksi yang akurat. Definisi pengalaman disini ialah informasi sebelumnya yang telah tersedia dan bisa dijadikan data pembelajar. Dalam pembelajaran machine learning, terdapat skenario-skenario seperti :

- *Supervised Learning*

  Penggunaan skenario *supervised learning,* pembelajaran menggunakan masukan data pembelajaran yang diberi label. Setelah itu membuat prediksidari data yang telah diberi label.

- *Unsupervised Learning*

  Penggunaan skenario *unsupervised learning*, pembelajaran menggunakan masukan data pembelajaran yang tidak diberi label. Setelah itu mencoba untuk mengelompokkan data berdasarkan karakteristik-karakteristik yang ditemui.

- *Semi-Supervised Learning*

  Pada skenario *semi-supervised learning*, teknik *machine learning* yang memanfaatkan contoh berlabel dan tidak berlabel saat mempelajari suatu model.

## *2.4 Deep Learning*

*Deep learning* merupakan salah satu cabang dari *machine learning* yang memanfaatkan jaringan syaraf tiruan untuk implementasi permasalahan dataset yang besar. Teknik *deep learning* memberikan arsitektur yang sangat kuat untuk supervised learning. Dengan menambahkan lebih banyak lapisan model pembelajaran tersebut bisa mewakili data citra berlabel dengan lebih baik. Pada *machine learning* terdapat teknik untuk menggunakan ekstraksi fitur dari data pelatihan dan algoritma pembelajaran

khusus untuk mengklasifikasi citra maupun untuk mengenali suara. Namun, metode ini masih memiliki beberapa kekurangan baik dalam hal kecepatan dan akurasi. Apalagi konsep jaringan syaraf tiruan yang dalam (banyak lapisan) dapat ditangguhkan pada algoritma *machine learning* yang sudah ada sehinggakomputer sekarang bisa belajar dengan kecepatan, akurasi, dan skala yang besar. Prinsip ini terus berkembang hingga deep learning semakin sering digunakan pada komunitas riset dan industri untuk membantu memecahkan banyak masalah data besar seperti *Computer Vision*, *Speech Recognition*, dan *Natural Languange Processing*. *Feature Engineering* juga merupakan teknik yang paling penting untuk mencapai hasil yang baik pada tugsa prediksi. Namun, sulit untuk dipelajari dan dikuasai karena kumpulan data dan jenis data yang berbeda memerlukan teknik pendekatan yang berbeda pula.

Menurut pendapat (Tom et all, 2018) Metode *Deep Learning* terbukti berguna dan memperoleh hasil canggih untuk berbagai tugas bahasa alami dengan pelatihan ujung ke ujung dan kemampuan pembelajaran representasi. 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, 2019).

## 2.5 *Large Language Models* (LLM)

*Large Language Models* adalah jenis model bahasa yang diterapkan untuk pemahaman dan pembuatan bahasa dengan tujuan umum menggunakan sejumlah besar data untuk mempelajari miliaran parameter selama fase pelatihan. Model ini dirancang untuk mengatasi keterbatasan chatbot tradisional. LLM adalah model generatif yang dibangun dari distribusi probabilitas serangkaian kata dari bahasa alami untuk memprediksi kata berikutnya dalam konteks narasi interaktif.

LLM yang paling sukses disebut dengan transformer, yang merupakan jenis jaringan saraf tiruan yang dilatih menggunakan teknik pembelajaran mesin dari media teks yang dapat menghasilkan konten seperti teks, ucapan, video, musik dan keluaran lainnya dalam beberapa domain masal (Min et all, 2024). berada, berdasarkan pada data

latih, dimana tiap anggota data latih tersebut telah diketahui kategorinya. Kategori ini tentunya bersifat diskrit, dimana urutan tidak mempengaruhi (Jiawei et all, 2012). Contohnya seperti: positif, negatif, dan netral; baik dan buruk. Salah satu contoh arsitektur LLM adalah GPT (*Generative Pre-Trained Transformer*) dari OpenAI yang digunakan dalam ChatGPT, sementara PaLM (*Pathways Languange Model*) dari Google adalah LLM yang mendukung Bard. LLM berhasil digunakan dalam pengenalan ucapan, penerjemahan bahasa, pembuatan bahasa alami, peringkasan teks, pengenalan tulisan tangan dan pengambilan informasi, di antara tugas-tugas lainnya.

## 2.6 Teknik Klasifikasi

Teknik klasifikasi bisa disimpulkan sebagai cara memprediksi suatu data baru sehingga bisa ditentukan pada kategori apakah ia berada, berdasarkan pada data latih, dimana tiap anggota data latih tersebut telah diketahui kategorinya. Kategori ini tentunya bersifat diskrit, dimana urutan tidak mempengaruhi (Jiawei et all, 2012). Contohnya seperti: positif, negatif, dan netral; baik dan buruk.

Proses Klasifikasi dalam teknik klasifikasi ada dua proses utama yaitu proses pembangunan model dan penerapan model (Jiawei et all, 2012). Proses pembangunan model melibatkan tahapan sebagai berikut:

1. Menentukan kategori/kelas/label terlebih dahulu. Misal: positif, negatif, dan netral.
2. Dari sekumpulan data yang diperoleh, tentukan kategori untuk tiap datanya.
3. Sekumpulan data yang telah dikategorisasikan ini disebut dengan data latih yang akan digunakan sebagai model.
4. Model ini bisa digambarkan sebagai aturan klasifikasi, pohon keputusan atau formula matematika.
5. Algoritma berdasarkan model di atas untuk mengklasifikasi disebut dengan classifier (pengklasifikasi).

Proses ini dapat disebut juga sebagai *supervised learning* (pelatihan terawasi). Disebut terawasi karena tiap datanya sudah diberikan label. Proses yang kedua adalah proses penerapan model atau proses klasifikasi. Proses ini melibatkan tahap:

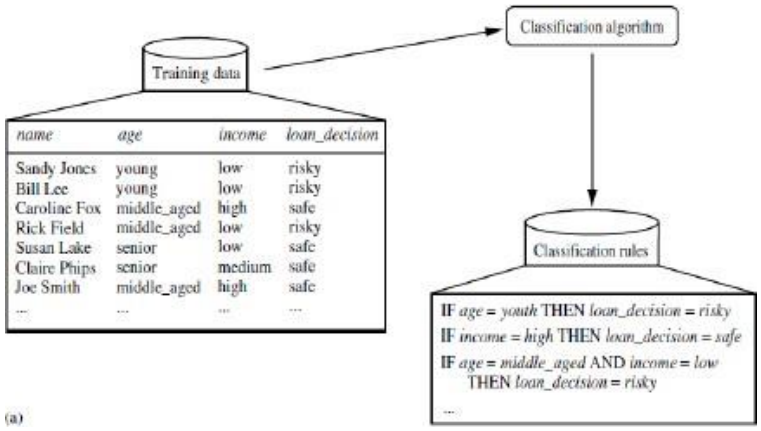
1. Tentukan sekumpulan data untuk diuji.

2. Sekumpulan data uji ini tiap datanya telah diberikan kategori/kelas/label.

3. Dilakukan proses pemetaan dengan menggunakan *classifier* di atas. Data uji ini akan ditentukan kategorinya berdasarkan model di atas dan kemudian hasilnya dibandingkan dengan kategori yang telah diberikan. Misal: pada data uji, dinyatakan bahwa data X adalah positif. Setelah dilakukan proses klasifikasi dengan menggunakan data latih ternyata data X bernilai negatif.

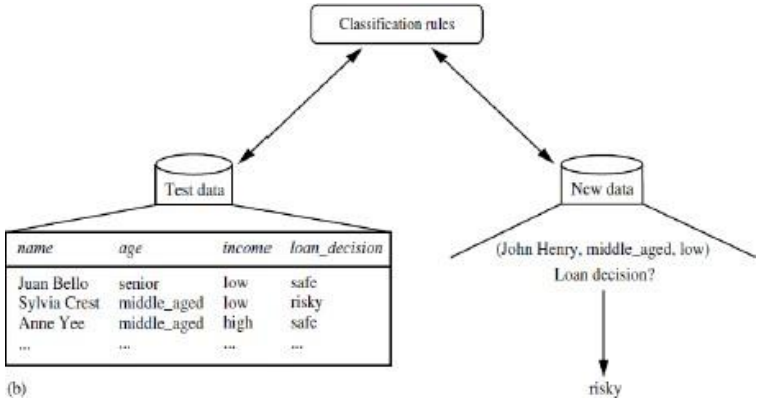
4. Kemudian ditentukan akurasi model di atas dengan menghitung seberapa banyak kategori yang dihasilkan bernilai sama dengan kategori yang telah ditentukan pada data uji diawal.

5. Jika rasio akurasi memuaskan (memenuhi batas minimal yang ditentukan), maka *classifier* tersebut dapat digunakan untuk data baru. Untuk lebih jelasnya gambar di bawah ini bisa menjelaskan proses tersebut diatas.




**Gambar 2.1** Proses Pembangunan Model (Jiawei et al, 2006:287)

*Training data* atau data latih, dengan algoritma klasifikasi dihasilkan *classification rules* atau aturan klasifikasi yang disebut dengan *classifier* (pengklasifikasi). Pada Gambar 2.1 di atas, *loan_decision* adalah label atau kategori yang telah ditentukan.




**Gambar 2.2** Proses Penerapan Model (Jiawei et al, 2006:287)

Dengan *classifier* yang telah dihasilkan, diterapkan pada test data atau datauji untuk diukur keakuratannya. Hasil akurasi adalah perbandingan dari jumlah total hasil klasifikasi menggunakan classifier pada data uji yang nilainya sama dengan nilai *loan_decision* pada data uji.

## *2.7 Pre-Processing*

Menurut (Alexander, 2003) Tahap *text pre-processing* merupakan proses untuk mempersiapkan data mentah sebelum dilakukan proses lain. Pada umumnya, *pre-processing* data dilakukan dengan cara mengeliminasi data yang tidak sesuai atau mengubah data menjadi bentuk yang lebih mudah diproses oleh sistem. Tahap *Pre-processing* adalah tahapan dimana aplikasi melakukan seleksi data yang akan diproses pada setiap dokumen. Proses *pre-processing* ini meliputi : *case folding, tokenizing, filtering,* dan *stemming*. Dimana penjelasan dari tahap-tahap tersebut adalah sebagai berikut:

a. *Case folding*

   *Case folding* merupakan salah satu bentuk *text preprocessing* yang paling sederhana dan efektif meskipun sering diabaikan. Tujuan dari *case folding* untuk mengubah semua huruf dalam dokumen menjadi huruf kecil (Rudy et all, 2020).

b. *Tokenizing*

   *Tokenizing* merupakan tahap pemotongan teks input menjadi kata, istilah,

*symbol*, tanda baca, atau elemen lain yang memiliki arti yang disebut token. Pada proses token yang merupakan tanda baca yang dianggap tidak perlu seperti titik (.), koma (,), tanda seru (!), dan lain-lain akan dihapus (Rudy et all, 2020).

c. *Filtering*

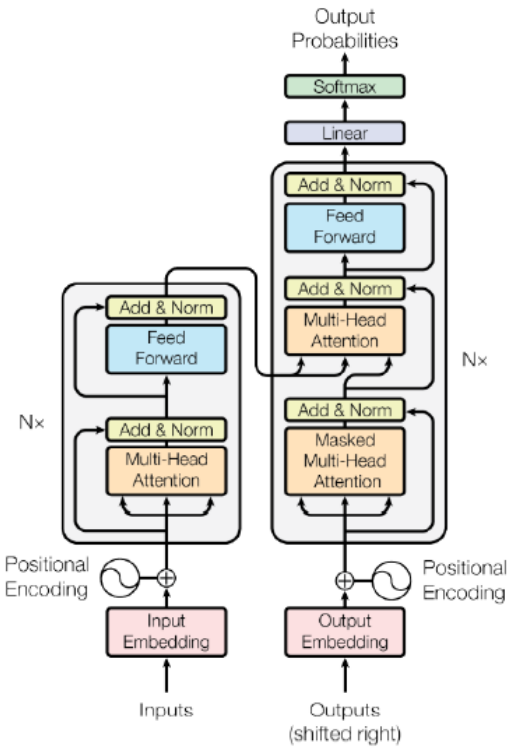
*Filtering* merupakan proses yang dilakukan pada tahap ini yaitu menghapus*stop-word. Stop-word* adalah kata yang bukan merupakan kata unik dalam suatu artikel atau kata-kata umum yang biasanya selalu ada dalam suatu artikel kecil (Rudy et all, 2020).

d. *Stemming*

*Stemming* digunakan untuk mendapatkan kata dasar dari suatu kata. Hal ini dilakukan untuk menormalisasi kata kecil (Rudy et all, 2020). Sebagai contoh, kata membawa, dibawa, membawakan jika dilakukan *stemming* maka kata dasarnya yaitu "bawa". Proses *stemming* pada kata bahasa Indonesia berbeda dengan *stemming* pada kata dalam bahasa Inggris. Pada bahasa Inggris proses yang diperlukan adalah menghilangkan sufiks saja, sedangkan pada bahasa Indonesia selain menghilangkan sufiks juga menghilangkan prefix dan juga konfiks.

## *2.8 Transformer*

Menurut (Vasmani et all, 2017) dalam judul penelitian "*Attention is All You Need*", Transformer adalah arsitektur yang dapat digunakan untuk mengubah satu urutan ke urutan lain dengan bantuan dua bagian, yaitu *encorder* dan *decorder*. Arsitektur ini menggunakan *self-attention* mechanism dan telah menjadi *state-of-the-art* pada bidang NLP (*Natural Language Processing*). Transformer mengakselerasi proses pelatihan dan menggunakan *self-attention* untuk menarik dependensi global antara masukkan yang berbeda. Perkembangan ini membuat model image captioning mulai mengadopsi arsitektur transformer (Umar et all, 2021).




**Gambar 2.3** Model Arsitektur *Transformer* (Vaswani et all, 2017)

*Transformer* adalah model transduksi pertama yang sepenuhnya mengandalkan *self-attention* untuk menghitung representasi input dan output tanpa menggunakan RNN atau konvolusi. *Transformer* menggunakan lapisan *self-attention* dan *point-wise* yang ditumbuk, *fully connected* untuk encoder dan decorder yang masing-masing pada bagian kiri dan kanan pada Gambar 3.2

### *2.8.1 Bidirectional Encoder Representations for Transformers* (BERT)

*Bidirectional Encoder Representations for Transformers* (BERT) adalah model pemrosesan bahasa yang diperkenalkan oleh tim Google AI pada tahun 2018. BERT menghadirkan terobosan dengan mengintegrasikan perhatian dua arah dalam pemrosesan teks, memungkinkan model untuk memahami konteks sebelum dan sesudah suatu kata dalam kalimat. Arsitektur BERT didasarkan pada Transformer, serangkaian blok dengan lapisan *self-attention* dan lapisan *feedforward*, memberikan kemampuan model untuk memahami hubungan antar kata secara simultan. Proses pelatihan BERT dimulai dengan

memberikan representasi vektor untuk setiap kata dalam kalimat, termasuk vektor kata, posisi kata, dan segmen kalimat (Jacob et all, 2018).

Model ini memiliki beberapat tahapan utama dalam proses pelatihannya, yaitu : tahap pertama yaitu tahap *pre-training*, BERT menerapkan dua tugas utama: *Masked Language Models* (MLM), di mana beberapa kata dalam kalimat disamarkan dan model harus memprediksi kata yang tersembunyi berdasarkan konteks, dan *Next Sentence Prediction* (NSP), di mana model harus memprediksi apakah sebuah kalimat adalah kelanjutan langsung dari kalimat sebelumnya. Tugas MLM dan NSPmembantu BERT mengembangkan representasi kata yang lebih kontekstual dan pemahaman hubungan antar kalimat (Jacob et all, 2018).

Tahap kedua yaitu tahap *fine-tuning* untuk menyesuaikan model dengan tugas spesifik seperti *Question Answering* atau *Natural Language Inference*. Proses ini memungkinkan transfer pengetahuan dari tahap *pre- training* ke tugas yang lebih spesifik, memanfaatkan pemahaman kontekstual dan nuansa bahasa alami yang telah dikembangkan sebelumnya. Dengan *fine-tuning*, BERT mencapai kinerja optimal dalam memahami konteks dan memberikan hasil akurat dalam berbagai tugas pemrosesan Bahasa (Jacob et all, 2018).

### *2.9 Convolutional Neural Network* (CNN)

*Convolutional Neural Network* (CNN) adalah pengembangan dari *Multilayer Perceptron* (MLP) yang didesain untuk mengolah data dua dimensi. CNN termasuk dalam jenis Deep Neural Network karena kedalaman jaringan yang tinggi dan banyak diaplikasikan pada data citra. Pada kasus klasifikasi citra, MLP kurang sesuai untuk digunakan karena tidak menyimpan informasi spasial dari datacitra dan menganggap setiap piksel adalah fitur yang independen sehingga menghasilkan hasil yang kurang baik (I Wayan, 2016).

Cara kerja CNN memiliki kesamaan pada MLP, namun dalam CNN setiap neuron dipresentasikan dalam bentuk dua dimensi, tidak seperti MLP yang setiap neuron hanya berukuran satu dimensi (I Wayan, 2016).

Menurut (Regiolina et all, 2023) Pemodelan CNN dibangun dengan arsitektur layer dimulai menggunakan fungsi *Sequential* dimana terdapat satu input tensor dan satu output tensor. Input tensor merepresentasikan matriks sebagai inputan data yang dapat dimasukkan nilai dari hasil pembobotan word2vec sebelumnya. Output tensor merepresentasikan hasil keluaran yang dalam penelitian adalah klasifikasi. Dropout layer dibuat untuk mengurangi terjadinya data overfitting, Setelah menerima output dari lapisan embedding. selanjutnya, Lapisan konvolusi akan mengekstrak fitur dari data input. Dalam pemodelan CNN menggunakan filter 64 dan 2 unit. *Convolution layer* akan membentuk *vektor feature map* sebanyak filter yang digunakan dalam proses konvolusi.

## 2.10 *Feature Fusion*

Menurut (Neelesh, 2023) Teknik *feature fusion* dalam *deep learning* sangat penting bagi keberhasilan sebuah model karena memungkinkan adanya kombinasi atau penggabungan berbagai fitur yang diekstrak dari beberapa lapisan atau sumber untuk menghasilkan sebuah representasi yang lebih informatif dan diskriminatif.

Menurut (Neelesh, 2023) Kemajuan dalam *deep learning* telah memunculkan teknik *feature fusion* berbasis *attention*. Metode ini memanfaatkan attention mechanism, yang memungkinkan model untuk selektif berfokus pada fitur yang relevan dan mengabaikan fitur yang kurang penting. Teknik *feature fusion berbasis attention mechanism* telah menunjukkan keberhasilan yang luar biasa dalam berbagai domain, termasuk teks gambar, terjemahan mesin dan tanya jawab visual.

## 2.11 Penelitian Terkait

Penelitian yang terkait dengan penelitian yang akan dilakukan penulis, mengacu pada penelitian sebelumnya sebagai acuan untuk penelitian selanjutnya. Referensi jurnal-jurnal yang digunakan memiliki kemiripan dari penulisan yang dilakukan oleh penulis, oleh karena itu jurnal-jurnal yang digunakan menjadi pertimbangan bagi penulis dalam melakukan penelitian dan membuat penulisan ini. Berikut adalah jurnal-jurnal terkait yang telah dikaji oleh penulis.

**Tabel 2.1** Penelitian Terkait

| No | Peneliti | Judul | Metodologi | Hasil Penelitian |
|---|---|---|---|---|
| 1. | Revati S, Meetkumar P (2018) | Toxic Comment Classification Using Neural Networks and Machine Learning | - *Deep Learning*<br>- *Convolutional Neural Network*<br>- *Long-Short Term Memory* | Penelitian ini berhasil menemukan solusi optimal terbaik untuk klasifikasi *toxic comment*. Penelitian ini juga mengklasifikasikan komentar *toxic* ke dalam 6 label berdasarkan tingkat toksisitasnya yang sudah disediakan oleh kumpulan data di platform Kaggle, yaitu *toxic*, *severe-toxic*, pelecehan, ancaman, penghinaan dan ujaran kebencian terhadap identitas. Hasil penelitian ini metode LSTM menunjukkan bahwa LSTM berkinerja lebih baik daripada CNN dalam hal akurasi, sedangkan CNN memiliki hasil kinerja waktu yang lebih baik daripada LSTM. |
| 2. | Peiyu Z, Shuangtao Y (2019) | Multimodal Tweet Sentiment Classification Algorithm Based on Attention Mechanism | - Metode fusi<br>- Model Bidirectional-LSTM<br>- Model VGG-16 | Penelitian ini berhasil melakukan klasifikasi sentimen tweet multimodal yang berisi teks dan gambar. Jumlah fitur teks ftext dan fitur gambar fimage, dengan menggunakan metode fusi ini skor mikro f1 adalah 79,6%. Metode fusi kedua adalah *concatenate*, 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%. |

| | | | | |
|---|---|---|---|---|
| 3. | Suresh Mestry, Vishal Bisht, Roshan Chauhan, Kaushik Tiwari, Hargun Singh, (2019) | Multi-Label Classification Of Toxic Comments Using Fast-Text and CNN | - *Convolutional Neural Network* (CNN) | Penelitian ini berhasil mengklasifikasi berbagai jenis toksisitas pada komentar beracun di media social, seperti Wikipedia, Youtube, Twitter, Facebook, dll. Dengan menerapkan metode CNN dan fast-text didapatkan hasil yang lebih efisien daripada model Word2Vec dan GLOVE. |
| 4.. | A. Akshith Sagar, J.Sai Kiran (2020) | Toxic Comment Classification using Natural Language Processing | - Model LSTM,<br>- *NaiveBayes*<br>- *Support Vector Machine,*<br>- *Fasttext and Convolutional Neural Network* | Penelitian ini berhasil menganalisis berbagai pendekatan untuk menyelesaikan masalah klasifikasi komentar beracun secara online. Hasil dari penelitian ini bahwa model CNN bekerja lebih baik daripada model LSTM dan model NB-SVM dengan tingkat akurasi sebesar 98,15%. |
| 5. | Renaldy Permana S, Budi Arif D, Yuyum Umaidah, (2020) | Sentimen Analisis Komentar *Toxic* pada GrupFacebook *GameOnline* Menggunakan Klasifikasi Naïve Bayes | - *Naïve Bayes*<br>- Seleksi fitur TF-IDF<br>- *Information Gain* | Penelitian ini menghasilkan sentimen analisis pada komentar beracun (*toxic comment*) pada grup komunitas *games online* dengan menggunakan 2 seleksi fitur TF-IDF dan Information Gain, menggunakan rasio pembagian data training dan data testing 80:20.<br>Tingkat akurasi 75%, *precision* 63%, *recall* 67%, dan *F-measure* 64%. |

| | | | | |
|---|---|---|---|---|
| 6. | Amit Sheth, Valerie L. Shalin, Ugur Kursuncu. (2021) | Defining and Detecting Toxicity on Social Media: Context and Knowledge are Key | - Multi-level Analisis | Penelitian ini dilakukan untuk mengidentifikasi berbagai pengaruh terhadap deteksi komentar beracun dan membagi kedalam 5 dimensi yaitu konten seksual, rasial, penampilan, inlektual dan konten politik. Untuk menentukan toksisitas dilihat dari multilevel analisis data seperti konten, individu dan juga komunitas. Framework ini dibuat berdasarkan teori sosial dan perilaku. |
| 7. | Akhsi Kumar, Nitin Sachdeva (2021) | Multimodal Cyberbullying Detection UsingCapsule Network with Dynamic Routing and Deep Convolutional Neural Network | - Model CapsNet-ConvNet | Penelitian ini dilakukan untuk mendeteksi *cyberbullying* dalam tiga modalitas yang berbeda yaitu tekstual, visual dan infografis (Teks yang disematkan dengan gambar). Dataset yang disiapkan untuk eksperimen berisi 10.000 komentar dan postingan (teks, gambar, dan infografis) yang disiapkan menggunakan tiga situs media sosial YouTube, Instagram, dan Twitter. Modalitas dalam kumpulan data adalah 60% tekstual, 20% visual, dan 20% infografis. Hasilnya telah dievaluasi dan dibandingkan dengan berbagai baseline dan diamati bahwa model CapsNet-ConvNet yang diusulkan memberikan akurasi kinerja yang superlatif. |

| | | | | |
|---|---|---|---|---|
| 8. | P.Vidyullatha, Satya Narayan Padhy, Javvaji Geetha Priya, Kakarlapudi Srija,Sri Satyanjani Koppisetti (2021) | Identification and Classification of Toxic Comment Using Machine Learning Methods | - Metode relevansi biner dengan multinomial *Naïve Bayes* | Penelitian ini berhasil melakukan klasifikasi komentar beracun yang terbagi menjadi beberapa kategori seperti komentar beracun, pencabulan, penghinaan, sangat beracun, ujaran kebencian dan ancaman. Komentar beracun memiliki jumlah tertinggi dibanding dengan kategori lainnya. Metode Relevansi Biner dengan Multinomial Naïve Bayes adalah sebuah algoritma yang efisien untuk melakukan penelitian ini. |
| 9. | Hong Fan, Wu Du, Abdelghani Dahou, Ahmed A. Ewees, Dalia Yousri, Mohamed Abd Elaziz, Ammar H. Elsheikh, Laith Abualigah, Mohammed A. A. Al-qaness. (2021) | Social Media Toxicity Classification Using Deep Learning: Real-World Application UK Brexit | - *BiDirectional Encoder Representations from Transformers* (BERT),<br>- Multilingual BERT<br>- RoBERTa<br>- DistilBERT. | Penelitian ini dilakukan 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. |
| 10. | Kristian & Joan, 2021 | Multilabel Text Classification | - *Support Vector Machine*<br>- *Doc2Vec* | Penelitian ini mampu mengidentifikasi dan klasifikasi dokumen-dokumen berita dalam bahasa Indonesia ke dalam terbagi |

| | | | | |
|---|---|---|---|---|
| | | Menggunakan SVM dan Doc2Vec Classification pada Dokumen Berita Bahasa Indonesia | | dalam 5 kategori berita utama, yaitu: Politik, Ekonomi, Teknologi, Olahraga, dan Hiburan. Sistem dibuat dengan menggunakan *Python*, memanfaatkan *Doc2Vec* untuk mengambil fitur dataset, dan SVM untuk melakukan klasifikasi terhadap banyak kelas. Hasil pada penelitian ini hanya menargetkan akurasi sebesar 70% saja. Namun dari hasil ujicoba, akurasi yang diperoleh melebihi 90%. |
| 11. | Sreyan Ghosh, Samden Lepcha, S Sakshi, Rajiv Ratn Shah, S. Umesh (2022) | DeToxy: A Large-Scale Multimodal Dataset for Toxicity Classification inSpoken Utterances | - Speech Toxicity Classification (STC) and End-to-End (E2E) | Pada penelitian ini memperkenalkan sebuah kumpulan data baru DeToxy. DeToxy adalah kumpulan data multimodal berskala besar dengan isyarat ucapan dan teks, dianotasi secara manual untuk deteksi toksisitas dalam ucapan lisan. Hasil dari penelitian ini menunjukkan bahwa dalam kasus ucapan lisan, sebuah pendekatan berbasis teks sangat bergantung pada transkrip yang diberianotasi manusia. |
| 12. | Muhammad Taleb, Alami Hamzah, Mohamed Zouitni, Nabil Burmani, Ujar Lafkiar, Noureddine En- Nahnahi (2022) | Detection of toxicity in social media based on Natural Language Processing methods | - Model LSTM<br>- *(LongShort-term Memory)* | Pada penelitian ini mendeteksi bagian beracun dalam komentar. Model LSTM dengan representasi Glove dan LSTM dengan FastText mampu menghasilkan F1 dan akurasi yang lebih tinggi dibandingkan model lain. Untuk mendeteksi komentar beracun diterapkan sebuah metode tanpa pengawasan, pendekatan ini menerapkan algoritma LIME berdasarkan |

| | | | | |
|---|---|---|---|---|
| | | | | pengklasifikasi LSTM dengan Glove yang memperoleh akurasi 98%. |
| 13. | Regiolina Hayami, Sofhia Mohnica, Soni (2023) | Klasifikasi multilabel komentar *toxic* pada sosial media twitter menggunakan *convolutional neural network* (CNN) | - Metode *Convolutional Neural Network* (CNN)<br>- Algoritma Word2Vec | Penelitian ini berhasil membuktikan sebuah metode *Convolutional Neural Network* (CNN) dapat diterapkan untuk klasifikasi multilabel *toxic comment* pada sosial media twitter dengan menggunakan Word2Vec sebagai pembobotan kata. Nilai performa yang didapat dari pengujian model mesin pembelajaran CNN dengan menggunakan optimizer adam menghasilkan akurasi tertinggi sebesar 99,76%, presisi 100%, recall 99% dan F1-Score 99%. Banyaknya data, Jumlah Filter, dan pengunaan optimizer yang berbeda mempengaruhi tingkat *accuracy* dan *loss.* |
| 14. | Chenyang L, MinghaoWu, LongyueWang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi & Zhaopeng Tu (2023) | MACAW-LLM: Multi-Modal Languange Modeling with Image, Audio, Video and Text Integration | - *Large Language Models*<br>- Model MACAW-LLM | Penelitian ini mengintegrasikan informasi visual, audio dan tekstual menggunakan MACAW-LLM. 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. |

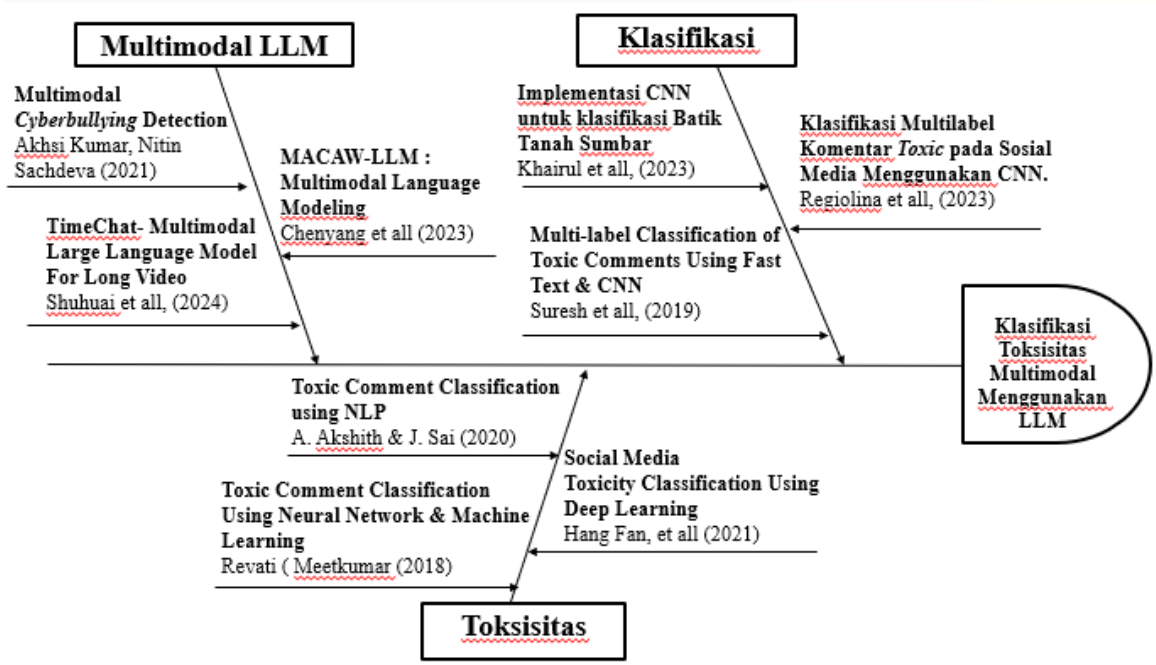
| | | | | |
|---|---|---|---|---|
| 15. | Khairul et all, 2023 | Implementasi *Convolutional Neural Network* (CNN) Untuk Klasifikasi Batik Tanah Liat Sumatera Barat | - *Convolutional Neural Network* | Penelitian ini berhasil melakukan klasifikasi pada batik tanah liat Sumatera Barat menggunakan metode 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. |
| 16. | Junnan Li, Dongxu Li, Silvio Savarese, Steven Hoi (2023) | BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models | - *Large Language Models*<br>- Model BLIP-2 | Penelitian ini berhasil melakukan *pre-training* menggunakan BLIP-2, dimana BLIP-2 ini mampu menjembatani modalitas dengan transformator kueri yang sudah dilatih. Hasil penelitian ini mencapai kinerja yang mengungguli Flamingo80B sebesar 8.7% pada VQAv2 zero-shot dengan 54x lebih sedikit parameter yang dapat dilatih. |
| 17. | Feilong Chen, Minglun Han, Haozhi Zhao, | X-LLM: Bootstrapping Advanced | - Model Multimodal<br>- X2L Interfaces | Penelitian ini mengusulkan menggunakan model X-LLM yang dapat mengubah multimodal (gambar, suara dan video) |

| | | | | |
|---|---|---|---|---|
| | Qingyang Zhang, Jing Shi, Shuang Xu and Bo Xu (2023) | Large Language Models by Treating Multi-Modalities as Foreign Languages | | kedalam bahasa Asing yang menggunakan antarmuka X2L dan memasukkannya kedalam model LLM (ChatGLM). Penelitian ini menghasilkan skor relatif 84,5%. |
| 18. | Shuhuai R, Linli Y, Shicheng L, Xu S, Lu H (2024) | TimeChat: A Time-sensitive Multimodal Large Language Model for Long Video Understanding | - *Large Language Models*<br>- Arsitektur Encorder<br>- Q-Former | Penelitian ini mengusulkan TimeChat untuk model LLM dalam tugas pemahaman video yang panjang. Penelitian ini menggabungkan 2 kontribusi arsitektur yaitu encoder bingkai dan Q-Former. Data set mencakup 5 tugas dengan total 125ribu instans. Hasil eksperimen dalam tugas pemahaman video, seperti teks padat, landasan temporal, dan deteksi sorotan, menunjukkan kemampuan lokalisasi temporal zero-shot dan penalaran yang kuat dari TimeChat. |

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.

Pada **gambar 2.4** terdapat Diagram *Fishbone* yang merupakan sebab-akibat penulis dalam mengambil tema penelitian.



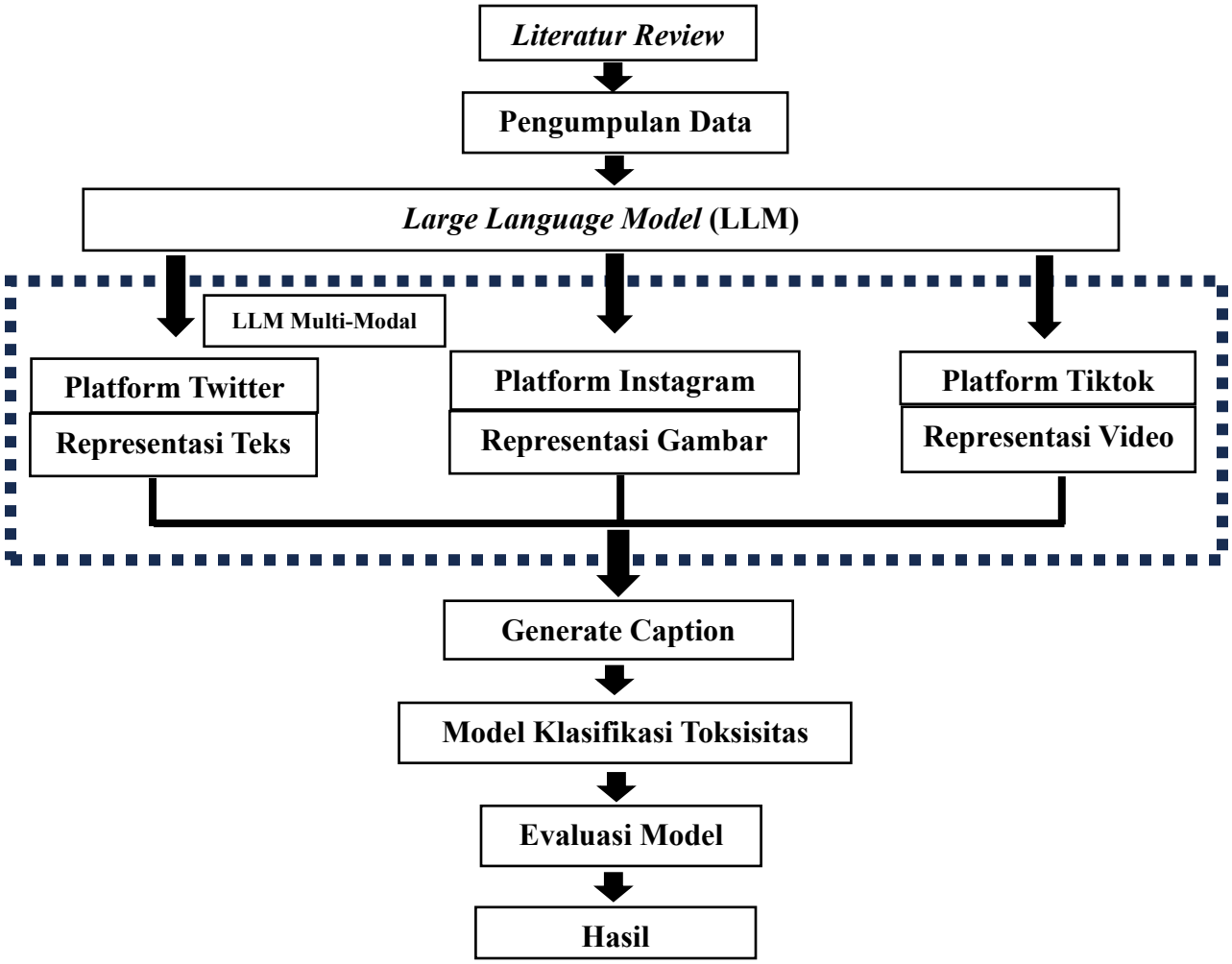

**Gambar 2.4** Diagram *FishBone*

# 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.

# DAFTAR PUSTAKA


1. Akhsi, K., Nitin, S., (2019). Cyberbullying detection on social multimedia using soft computing techniques: a meta-analysis. Multimedia Tools and Applications. 78(17), 23973-24010.

2. Akhsi, K., Nitin, S., (2021). Multimodal Cyberbullying Detection Using Capsule Network with Dynamic Routing and Deep Convolutional Neural Network. Springer.Alexander, C.

3. Anastasia, G., Guobiai, Z., & Paolo, R. (2020). Multimodal Multi-image Fake News Detection. IEEE Computer Society. 647-654. https://doi.org/10.1109/DSAA49011.2020.00091.

4. Andrei, K., Octave, P., Valentine, M., Alain, M., & Vincent, L. (2024). Large Language Models in Cybersecurity. Springer. https://doi.org/10.1007/978-3-031-54827-7.

5. A. Vaswani., Noam, S., Nikim P., Jakob, U., Llion, J., Aidan, N. G., Lukasz, K. (2017). Attention Is All You Need. 31st Conference on Neural Information Processing Systems. pp. 5999–6009.

6. Chenyang, L., Minghao, W., Longyue, W., Xinting, H., Bingshuai, L., Zefeng, D., Shuming, S., & Zhaopeng, T. (2023). MACAW-LLM : Multi-modal Languange Modelling with Image, Audio, Video, and Text Integration. https://www.researchgate.net/publication/371540959.

7. Feilong, C., Minglun, H., Haozhi, Z., Qingyang, Z., Jing, S., Shuang, X., & Bo, Xu. (2023). X-LLM: Bootstrapping Advanced Large Language Models by Treating Multi-Modalities as Foreign Languages. arXiv preprint arXiv:2305.04160v3.

8. Firoj, A., Stefano, C., Tanmoy, C., Fabrizio, S., Dimitar, D., Giovanni, D. S. M., Shaden, Shaar., Hamed, F., & Preslav, N., (2022). A Survey on Multimodal

Disinformation Detection.

9. I. Wayan, S.E.P., Arya, Y. W., Rully, S. (2016). Klasifikasi Citra Menggunakan Convolutional Neural Network (Cnn) pada Caltech 101. JURNAL TEKNIK ITS. Vol.5, No.1.
10. Jacob, D., Ming, W. C., Kenton, L., & Kristina, T., (2018). BERT: Pre- training of Deep Bidirectional Transformers for Language Understanding.
11. Jiawei, H., Micheline, K., & Jian, P., (2012). Data mining: concepts and techniques. Third Edition. San Francisco: Morgan Kaufmann.
12. Jie, H., & Kevin, C. (2023). Towards Reasoning in Large Language Models: A Survey.arXiv:2212.10403v2.
13. Jinhua, Z., Xiaohua, Z., Jiaxin, F., & Hongyun, M. (2024). 3D-CNN and semi-supervised based network forhyperspectral unmixing.INTERNATIONAL JOURNAL OF REMOTE SENSING.VOL. 45, NO. 1, 168–191. https://doi.org/10.1080/01431161.2023.2290998.
14. Junnan, L., Dongxu, L., Silvio, S., & Steven, H. (2023). BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv preprint arXiv:2301.12597v3.
15. Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heintz, I., & Roth, D. (2023). Recent advances in natural language processing via large pre-trained language models: A survey. ACM Computing Surveys, 56(2), 1–40. https://doi.org/10.1145/3605943.
16. Muhamad, R. F., Ridwan, I., Fatan, K., (2020). Klasifikasi Kalimat Ilmiah Menggunakan Recurrent Neural Network. Prosiding The 11th Industrial Research Workshop and National Seminar. hal 488-495.
17. M. Usman, H., Qasem, A., Rizwan, Q., Abbas, S., Amgad, M., M. Irfan., Anas, Z., M. Bilal, S., Naveed, A., Jia, W., & Seyedali, M. (2023). A Survey on Large Languange Models : Application , Challenges, Limitations and Practical Usage. https://doi.org/10.36227/techrxiv.23589741.v1.
18. Neelash, M. (2023). Adaptive Feature Fusion : Enhancing Generalization in Deep Learning Models. Available arXiv:2304.03290v1.

19. Nurul, D. K., Said, A. F., & Mahendra, D. P. (2021). Analisis Sentimen Komentar Beracun pada Media Sosial Menggunakan World2Vec dan Support Vectore Machine (SVM). E-Proceeding of Engineering, Vol. 8, No. 5. Page 10038-10050.
20. Peiyu, Z., & Shuangtao, Y. (2019). Multimodal Tweet Sentiment Classification Algorithm Based on Attention Mechanism. Springer. Pp, 68-79. https://doi.org/10.1007/978-3-030-14880-5_6.
21. Reinert, Y. R., (2023). Multilabel Classifcation for Toxic Comments inIndonesian. *Jurnal Emacs*. Hal 29-34.
22. Rudy, C., Ariesta, D., & Dede, A., (2020). *Recurrent Neural Network* (RNN) Dengan *Long Short Term Memory* (LSTM) Untuk Analisis Sentimen Data Instagram. *Jurnal Informatika dan Komputer (JIKO)*. Vol. 5, No. 1.
23. Rully, N., (2015). Media Sosial; Persfektif Komunikasi, Budaya, dan Sosioteknologi. *Bandung : Simbiosa Rekatama Media*.
24. Renaldy, P. S., Budy, A. D., & Yuyun, U., "Sentimen Analisis Komentar*Toxic* pada Grup Facebook *Game Online* Menggunakan Klasifikasi Naïve Bayes. *Jurnal Informatika Universitas Pamulang*, Vol. 5, 2020.
25. Regiolina, H., Sofhia, M., & Soni., (2023). Klasifikasi multilabel komentartoxic pada sosial media twitter menggunakan convolutional neural network (CNN). *Jurnal Computer Science and Information Technology (CoSciTech)*. Vol. 4, No. 1, hal. 1-6.
26. Revati, S., Meetkumat, P., (2018). Toxic Comment Classification Using Neural Network and Machine Learning. International Advanced Research Journal in Science, Engineering and Technology. Vo. 5, Issue 9, pp 47-52.
27. Tom, Y., Devamanyu, H., D., Soujanya, P., & Erik, C., (2018). Recent Trends in Deep Learning Based Natural Language Processing. *IEEE Computational Intelligence magazine,* vol. 13, no. 3, pp. 55-75.
28. Umar, A. A., & Dhomas, H. F. (2021). Implementasi Arsitektur Transformer pada Image Captioning dengan Bahasa Indonesia. IEEE.
29. Shuhuar, R., Linli, Y., Shiceng, L., Xu, Sun., Lu, H., (2024). TimeChat : A Time Sensitive Multimodal Large Language Model for Long Video Understanding.

Available : arXiv:2312.02051v2.

30. Suresh, M., Vishal, B., Roshanm C., Kaushik, T., Hargun, S., (2019). Multi-Label Classification of Toxic Comment Using Fast-Text and CNN.*International Journal of Advance in Computer Science and Cloud Computing*. vol.7, no.1, pp. 18-21.