KLASIFIKASI KOMEN TOKSIK MENGGUNAKAN PEMBELAJARAN MENDALAM

NUR SABRINA BINTI MOHD HILMAN RAJANG PROF. MADYA DR. NAZLIA BINTI OMAR

Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia, 43600 UKM Bangi,

Selangor Darul Ehsan, Malaysia

ABSTRAK

Komen toksik boleh memberikan kesan negatif yang mendalam terhadap komuniti atas talian, dengan mewujudkan persekitaran yang tidak selamat dan tidak mesra. Dalam konteks pengelasan komen toksik ini ingin memberi tumpuan kepada mengenal pasti dan mengkategorikan komen yang toksik atau mengganggu dalam perbincangan atas talian. Walaupun banyak kajian telah dijalankan dalam kajian pengelasan komen toksik ini, beberapa cabaran masih wujud seperti pengesanan ketoksikan berbilang bahasa, ketidakseimbangan data, dan kebolehfahaman model untuk memahami bahasa sumber terhad seperti Bahasa Melayu.Matlamat utama projek ini adalah untuk mengklasifikasikan komen toksik dalam Bahasa Melayu dengan menggunakan model transformer,.Projek ini menggunakan dua model utama iaitu *DeBERTa* dan *TinyLlama. DeBERTa (Decoding-enhanced BERT with Disentangled Attention)* menambah baik pemahaman bahasa dengan teknik penyahsulitan dan perhatian terpisah, meningkatkan pemahaman konteks dalam teks yang kompleks. *TinyLlama* pula ialah versi ringan dan berskala kecil daripada keluarga model *LLaMA (Large Language Model Meta AI)* yang direka untuk kecekapan dan kelajuan inferens, namun tetap mengekalkan prestasi yang kompetitif dalam pelbagai tugas pemprosesan bahasa semulajadi *(NLP).* Dengan saiz parameter yang lebih kecil dan keperluan sumber yang rendah, *TinyLlama* sangat sesuai digunakan dalam persekitaran dengan kekangan perkakasan seperti sistem masa nyata dan aplikasi web. Kesimpulannya, projek ini diharapkan dapat memberi manfaat yang signifikan kepada masyarakat dengan menyediakan alat yang canggih untuk merangkum komen dalam Bahasa Melayu, serta memajukan bidang pengesanan bahasa semulajadi untuk bahasa ini.

Kata kunci: Komen toksik, Bahasa Melayu, DeBERTa, TinyLlama, Transformer

PENGENALAN

Dalam era yang semakin moden ini, Internet memainkan peranan yang tidak dapat digantikan dalam kehidupan seharian, di mana sejumlah besar pertukaran maklumat dilaksanakan melalui media sosial dalam talian. Klasifikasi Komen Toksik Multibahasa merupakan satu cabaran utama dalam bidang Pemprosesan Bahasa Semula Jadi (PBT) dan pembelajaran mesin. Ia adalah tugas untuk secara automatik mengenal pasti dan mengklasifikasikan komen atau kandungan toksik daripada platform dalam talian merentasi pelbagai bahasa, kerana dunia internet melangkaui sempadan linguistik. (Giridhar Shambharkar et al., 2023) menyatakan kebangkitan media sosial dan komunikasi dalam talian telah menyediakan tahap hubungan yang tidak pernah berlaku sebelum ini bagi individu di seluruh dunia. Walau bagaimanapun, dengan peningkatan komunikasi ini datanglah cabaran dalam menangani aspek negatif perbincangan dalam talian, termasuk buli siber, ucapan kebencian, dan bentuk komen toksik lain. Klasifikasi komen toksik, iaitu proses mengenal pasti dan menandakan kandungan yang berpotensi berbahaya atau tidak wajar, telah menjadi alat yang semakin penting bagi platform dalam talian untuk mempromosikan persekitaran yang selamat dan mesra untuk pengguna.

Komen toksik merujuk kepada sebarang ungkapan yang kasar, menyakitkan hati, atau berniat untuk menimbulkan kebencian dan konflik dalam interaksi atas talian. Ia berbeza daripada sindiran, yang secara halus menggunakan maksud bertentangan, sarkasme, yang menyindir atau mengejek secara tajam, serta ujaran kebencian, yang secara khusus mensasarkan individu atau kumpulan berdasarkan kaum, agama atau jantina. Komen toksik lebih meluas kerana ia boleh merangkumi kesemua unsur ini dan turut mencetuskan suasana yang tidak sihat dan tidak selamat dalam komuniti digital.

Selain itu, media sosial telah menjadi sebahagian yang tidak terpisahkan dalam kehidupan harian, rapat dengan banyak aspek rutin harian setiap individu. Ini juga berlaku di negara-negara yang sedang membangun seperti Malaysia, di mana peningkatan 24% dalam jumlah pengguna media sosial telah diperhatikan dalam populasi Malaysia dari tahun 2016 hingga 2021 (Nurhayati-Wolff, H., 2021). Pada masa ini, 86% daripada populasi Malaysia merupakan pengguna aktif media sosial. Tidak dapat dinafikan, media sosial telah membawa banyak manfaat dan kemudahan kepada penggunanya kerana aksesibiliti dan kemudahan penggunaannya. Malangnya, terdapat sisi negatif kepada media sosial yang memerlukan perhatian lanjut. Salah satu contohnya adalah penyalahgunaan perkauman terhadap seorang pemain bola sepak profesional di media sosial berikutan prestasi yang buruk dalam perlawanan bola sepak (Sports, S., 2021).

Berdasarkan kajian oleh Maity et al. (2023), laporan terkini oleh Pusat Jurnalistik Bebas Malaysia (CIJ) menunjukkan peningkatan ketara dalam ucapan kebencian dalam talian yang menyasarkan kumpulan terpinggir seperti pelarian Rohingya dan komuniti biseksual serta transgender. Menurut The Malaysian Reserve (2020), antara Januari dan Jun 2020, Suruhanjaya Komunikasi dan Multimedia Malaysia (MCMC) telah menerima sebanyak 11,235 aduan mengenai pelbagai kesalahan siber seperti gangguan, buli siber, berita palsu, dan komen toksik. Statistik ini menekankan keperluan mendesak untuk menangani masalah komen toksik dalam Bahasa Melayu. Kekurangan langkah yang mencukupi untuk mengesan dan mengurangkan komen toksik dalam Bahasa Melayu menyebabkan cabaran besar dan memerlukan penyelidikan serta pembangunan teknik pengesanan komen toksik dalam bahasa yang mempunyai sumber terhad ini, bagi memerangi ancaman komen toksik dalam talian yang semakin berkembang di Malaysia.

Oleh itu, projek ini bertujuan untuk mengkaji dan mengesan sebarang komunikasi di media sosial yang mengandungi komen-komen toksik menggunakan kaedah pembelajaran mendalam.

METODOLOGI KAJIAN

Metodologi ini akan dibahagikan kepada lima peringkat utama yang akan dijalankan dalam kajian ini. Kajian ini melibatkan tujuh fasa utama, iaitu Fasa Penyediaan Data, Fasa Prapemprosesan Data, Fasa Penambahan Data, Fasa Peyeimbangan Data, Fasa Latihan Model, Fasa Pengujian dan Fasa Pembangunan Prototaip. Melalui metodologi ini, ia dapat membantu untuk memastikan bahawa setiap perkara dan kajian yang dijalankan adalah sistematik, berfokus, dan berdasarkan prinsip-prinsip yang kukuh. Melalui perjalanan tujuh fasa ini, pembangunan sistem klasifikasi komen toksik menggunakan pembelajaran mendalam dapat dijalankan secara sistematik, memastikan setiap langkah membawa sumbangan yang signifikan untuk keseluruhan proses pengembangan sistem.

Fasa Penyediaan Data

Komponen penting dalam fasa penyediaan data ini, iaitu pengumpulan data. Pengumpulan data melibatkan pemilihan sumber yang relevan dan sahih, manakala pra-pemprosesan awal pula melibatkan penyusunan format data agar sesuai untuk latihan model pembelajaran mendalam. Objektif utama fasa ini adalah untuk memastikan data yang dikumpul bersifat bersih, seragam, dan mempunyai kualiti tinggi bagi membolehkan latihan model dijalankan dengan cekap dan berkesan. Dalam pembangunan sistem pengelasan komen toksik Bahasa Melayu, pemilihan set data yang sesuai amat penting bagi menjamin keberkesanan dan prestasi model klasifikasi. Set data utama yang digunakan dalam kajian ini ialah *HateMalay Dataset*, iaitu satu set data yang dibina khusus untuk mengelas komen toksik dalam Bahasa Melayu.

HateMalay Dataset merupakanset data beranotasi yang mengandungi 4,892 twit yang diperoleh daripada platform media sosial seperti X (*Twitter*). Set data ini telah dilabel sama ada sebagai komen toksik atau bukan toksik, menjadikannya sesuai untuk tugas klasifikasi dwikelas (*binary classification*). Ia dibangunkan sebagai salah satu inisiatif awal untuk menangani kekurangan sumber berbahasa Melayu dalam bidang pemprosesan bahasa semula jadi (*NLP*). Set data ini diperkenalkan oleh Krishanu Maity dan rakan-rakannya melalui projek sumber terbuka yang diterbitkan di *GitHub* (Maity et al., 2023). Set data *HateMalay* direka bentuk untuk menyokong pembangunan model pembelajaran mendalam bagi tujuan pengesanan ketoksikan dalam pelbagai bentuk ayat atau komen yang ditulis secara spontan oleh pengguna atas talian. Komen dalam set dataini merangkumi pelbagai isu sosial, agama, dan peribadi yang sering digunakan dalam penyebaran ucapan kebencian, sekali gus memberikan konteks yang realistik untuk latihan model klasifikasi.

Pemilihan set data ini adalah signifikan kerana ia memperkasakan pembangunan sistem dalam Bahasa Melayu, satu bahasa yang sebelum ini kurang mendapat perhatian dalam penyelidikan NLP berbanding Bahasa Inggeris. Dengan menggunakan data sebenar yang telah dianotasi secara manual, sistem dapat dilatih dengan konteks yang relevan dan mampu melakukan klasifikasi yang lebih tepat dalam situasi dunia sebenar.

Fasa Prapemprosesan Data

Fasa prapemprosesan data merupakan salah satu komponen penting dalam pembangunan sistem pengelasan komen toksik, kerana ia memainkan peranan utama dalam memastikan data teks yang digunakan berada dalam bentuk yang bersih, seragam, dan sesuai untuk diproses oleh model pembelajaran mendalam. Oleh itu, sebelum data digunakan dalam fasa latihan dan pengujian, beberapa langkah penting telah dilaksanakan untuk meningkatkan kualiti input kepada model.Pertama sekali, penukaran huruf ke huruf kecil dilaksanakan dengan menukarkan semua huruf besar kepada huruf kecil bagi memastikan keseragaman data. Sebagai contoh, ayat asal “Toksik Sangat!” ditukar menjadi “toksik sangat!”. Seterusnya, unsur yang tidak relevan seperti emoji, URL, tanda pagar (#hashtag), dan nama pengguna (@user) dibuang kerana elemen-elemen ini tidak menyumbang kepada konteks sebenar komen.

Kemudian, proses penyingkiran aksara bukan ASCII dilaksanakan dengan membuang simbol atau karakter asing yang tidak boleh diproses seperti ñ, €, atau emoji tertentu. Setelah itu, tokenisasi dilakukan untuk memecahkan ayat kepada token individu, contohnya ayat “Dia tu gila!” ditukar menjadi ["dia", "tu", "gila"]. Langkah seterusnya ialah pembuangan stopword, iaitu perkataan yang kerap digunakan tetapi tidak membawa makna penting dalam klasifikasi seperti “dan”, “itu”, dan “adalah”. Selain itu, huruf berulang dalam perkataan dikurangkan untuk mengekalkan bentuk asal

perkataan, contohnya “baikkkkkkk” menjadi “baik”. Setelah itu, proses lemmatisasi dilakukan untuk menukarkan perkataan kepada bentuk dasar, seperti “berlari” kepada “lari”. Semua token yang telah diproses kemudian digabungkan semula menjadi satu ayat yang lengkap. Seterusnya, komen yang terlalu pendek atau tidak bermakna seperti “ok”, “hmm”, atau “la” akan ditapis keluar dalam proses penapisan teks pendek. Akhir sekali, penukaran label dilaksanakan bagi menyeragamkan output model di mana 0 merujuk kepada komen bukan toksik dan 1 kepada komen toksik, agar bersesuaian dengan keperluan model klasifikasi binari.

Justeru, langkah-langkah prapemprosesan ini bukan sahaja menyumbang kepada peningkatan kecekapan model, malah ia juga memastikan input yang diberikan kepada model berada dalam keadaan optimum dari segi struktur dan kandungan. Jadual 1.0 di bawah menunjukkan contoh praktikal hasil daripada proses prapemprosesan yang dijalankan ke atas satu komen.

Jadual 1.0 Langkah Pra-pemprosesan Teks

Proses Prapemprosesan	Komen
Teks Asal	Bodoh ke apa aku dah kata duit tu mmg aku nak pakai pegi Perlis hari lagi jam nak gerak boleh pm takleh nak bayar full eh mmg aku pegi Perlis stay sana terus ke kau nak pulang separuh
Selepas Pengecilan Huruf	bodoh ke apa aku dah kata duit tu mmg aku nak pakai pegi perlis hari lagi jam nak gerak boleh pm takleh nak bayar full eh mmg aku pegi perlis stay sana terus ke kau nak pulang separuh
Selepas Buang Unsur Tidak Relevan	bodoh ke apa aku dah kata duit tu mmg aku nak pakai pegi perlis hari lagi jam nak gerak boleh pm takleh nak bayar full eh mmg aku pegi perlis stay sana terus ke kau nak pulang separuh
Selepas Tokenisasi	[“bodoh”, “ke”, “apa”, “aku”, “dah”, “kata”, “duit”, “tu”, “mmg”, “aku”, “nak”, “pakai”, “pegi”, “perlis”, “hari”, “lagi”, “jam”, “nak”, “gerak”, “boleh”, “pm”, “takleh”, “nak”, “bayar”, “full”, “eh”, “mmg”, “aku”, “pegi”, “perlis”, “stay”, “sana”, “terus”, “ke”, “kau”, “nak”, “pulang”, “separuh”]
Selepas Buang Stopword	[“bodoh”, “kata”, “duit”, “mmg”, “nak”, “pakai”, “pegi”, “perlis”, “hari”, “jam”, “gerak”, “boleh”, “pm”, “takleh”, “bayar”, “full”, “eh”, “stay”, “sana”, “terus”, “kau”, “pulang”, “separuh”]
Selepas Lemmatisasi	[“bodoh”, “kata”, “duit”, “mampu”, “pakai”, “pergi”, “perlis”, “hari”, “jam”, “gerak”, “boleh”, “mesej”, “tidak”, “bayar”, “penuh”, “eh”, “tinggal”, “sana”, “terus”, “awak”, “pulang”, “separuh”]
Selepas Gabungan Semula	bodoh kata duit mampu pakai pergi perlis hari jam gerak boleh mesej tidak bayar penuh eh tinggal sana terus awak pulang separuh

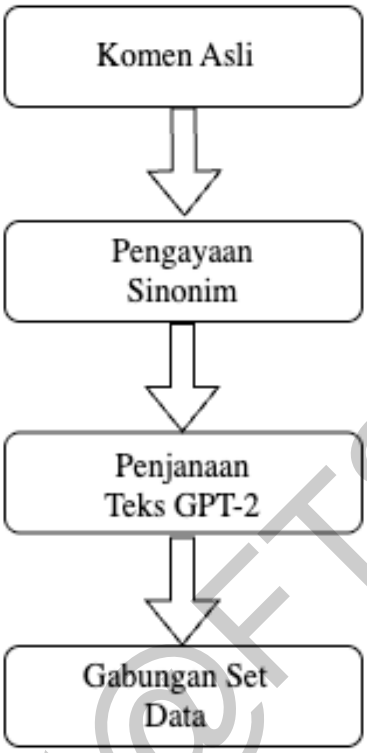
Fasa Penambahan Data

Fasa penambahan data merupakan salah satu komponen penting dalam pembangunan model klasifikasi komen toksik, terutamanya apabila berdepan dengan isu ketidakseimbangan kelas yang lazim berlaku dalam data dunia sebenar. Penambahan data, atau data augmentation, didefinisikan sebagai satu proses yang memperluaskan saiz dan kepelbagaian dataset dengan mencipta data baharu melalui transformasi ataupun penjanaan secara automatik daripada data sedia ada. Teknik ini terbukti mampu meningkatkan prestasi model, memperbaiki keupayaan generalisasi, serta mengurangkan risiko overfitting (Shorten & Khoshgoftaar, 2019). Dalam kajian ini, dua pendekatan utama telah digunakan untuk proses augmentasi, iaitu pengayaan sinonim dan penjanaan teks menggunakan model generatif GPT-2. Pendekatan pertama, iaitu pengayaan sinonim, dilaksanakan dengan menggunakan pustaka nlpaug yang digabungkan bersama kamus sinonim daripada WordNet. Melalui teknik ini, setiap perkataan dalam komen toksik ditukar kepada sinonim yang sesuai dalam konteks, dan

proses ini diulang sebanyak lima kali bagi setiap komen untuk menjana variasi yang mencukupi. Hasilnya telah disimpan dalam fail synonym_augmented_cleaned.csv selepas melalui pembersihan data.

Seterusnya, teknik penjanaan teks dilaksanakan menggunakan model DistilGPT2, iaitu versi ringan dan dipercepatkan daripada model GPT-2 asal. Dalam pelaksanaan ini, sebanyak 2,000 komen toksik asal digunakan sebagai seed, dan bagi setiap komen, lima ayat baharu dijana menggunakan model ini. Proses ini membolehkan sistem menghasilkan gaya bahasa toksik yang lebih pelbagai, sekali gus melatih model untuk mengenal pasti komen yang berbahaya dalam pelbagai bentuk. Data hasil penjanaan ini kemudian dibersihkan dan disimpan dalam fail gpt2_augmented_cleaned.csv.

Akhirnya, ketiga-tiga komponen iaitu data asal, hasil augmentasi sinonim, dan hasil penjanaan GPT-2 digabungkan untuk membentuk satu dataset akhir yang lengkap, besar, dan seimbang. Gabungan ini dinamakan sebagai merge_HateMalay_expanded.csv, dan digunakan dalam proses latihan akhir model. Proses gabungan ini bukan sahaja berjaya meningkatkan bilangan komen toksik melebihi 10,000 entri, tetapi turut memperkukuh prestasi model dalam mengenal pasti komen toksik yang mungkin tidak begitu jelas dari segi struktur atau bahasa. Secara keseluruhan, fasa ini menyumbang kepada ketahanan model dalam tugas klasifikasi dengan meningkatkan kepelbagaian linguistik, mengurangkan bias terhadap data asal, serta menyediakan data mencukupi untuk proses pembelajaran mendalam yang lebih mantap.



Rajah 1.0 Langkah Penambahan Data (Augmentation)

Fasa Pembahagiaan & Penyeimbagan Data

Fasa ini merangkumi dua komponen penting, iaitu proses pembahagian data kepada subset tertentu dan penyeimbangan semula taburan kelas dalam dataset. Kedua-dua langkah ini amat penting dalam memastikan prestasi model klasifikasi dapat dinilai secara adil dan tepat. Dalam konteks pembelajaran mendalam, proses pembahagian data merujuk kepada tindakan memecah dataset kepada tiga subset yang berbeza, iaitu set latihan, set validasi dan set ujian. Tujuan utama pembahagian ini adalah untuk membolehkan model mempelajari corak daripada data latihan, menyesuaikan parameter melalui proses penalaan semasa validasi, dan akhirnya diuji menggunakan data yang tidak pernah dilihat untuk menilai prestasi sebenar model dalam menghasilkan generalisasi yang baik (Brownlee, 2020).

Walau bagaimanapun, hasil analisis awal mendapati bahawa dataset asal bagi kajian ini mengalami isu ketidakseimbangan kelas, di mana bilangan komen bukan toksik adalah jauh lebih banyak berbanding komen toksik. Keadaan ini menimbulkan kebimbangan kerana model berpotensi untuk berat sebelah kepada kelas majoriti dan gagal mengesan ciri-ciri penting dalam kelas minoriti, iaitu komen toksik. Justeru, bagi menangani masalah ini, kaedah Synthetic Minority Over-sampling Technique (SMOTE) telah diterapkan. SMOTE ialah satu pendekatan pensampelan berasaskan ciri yang berfungsi untuk menjana contoh sintetik baharu bagi kelas minoriti dengan

menginterpolasi antara titik data sebenar, dan bukannya menggandakan data yang sedia ada (Chawla et al., 2002).

Dalam fasa ini, dua jenis dataset telah digunakan iaitu dataset asal yang belum diperluas dan kedua, dataset yang telah dipertingkatkan melalui proses augmentasi. Kedua-dua dataset ini telah melalui proses vektorisasi menggunakan pendekatan Term Frequency-Inverse Document Frequency (TF-IDF) bagi menukar teks kepada bentuk numerik. Selepas itu, hasil vektorisasi ini ditukar ke dalam format padat (dense array) sebelum diterapkan kaedah SMOTE untuk menyeimbangkan taburan kelas. Kaedah ini telah berjaya menghasilkan dataset yang seimbang antara kelas 0 (bukan toksik) dan kelas 1 (toksik).

Visualisasi hasil proses penyeimbangan dapat dilihat dalam Rajah 3.13 dan Rajah 3.14. Rajah 3.13 menunjukkan taburan data selepas SMOTE untuk dataset asal yang menghasilkan jumlah 5,962 contoh bagi setiap kelas. Manakala Rajah 3.14 pula memaparkan hasil penyeimbangan bagi dataset augmentasi, di mana kedua-dua kelas mencapai 42,358 contoh. Visualisasi ini dengan jelas menggambarkan kejayaan proses SMOTE dalam mengatasi ketidakseimbangan kelas secara berkesan.

Secara keseluruhannya, hasil daripada fasa ini menunjukkan bahawa penyeimbangan data memainkan peranan yang signifikan dalam meningkatkan keupayaan model untuk mengesan komen toksik. Dengan taburan data yang seimbang, model dapat memberikan prestasi yang lebih stabil dan adil, selain mengurangkan risiko berat sebelah terhadap kelas yang lebih dominan. Justeru, fasa ini membentuk asas yang kukuh untuk latihan model yang lebih efektif dalam fasa seterusnya. Jadual 1.1 di bawah menunjukkan penyeimbangan data sebelum dan selepas menggunakan SMOTE.

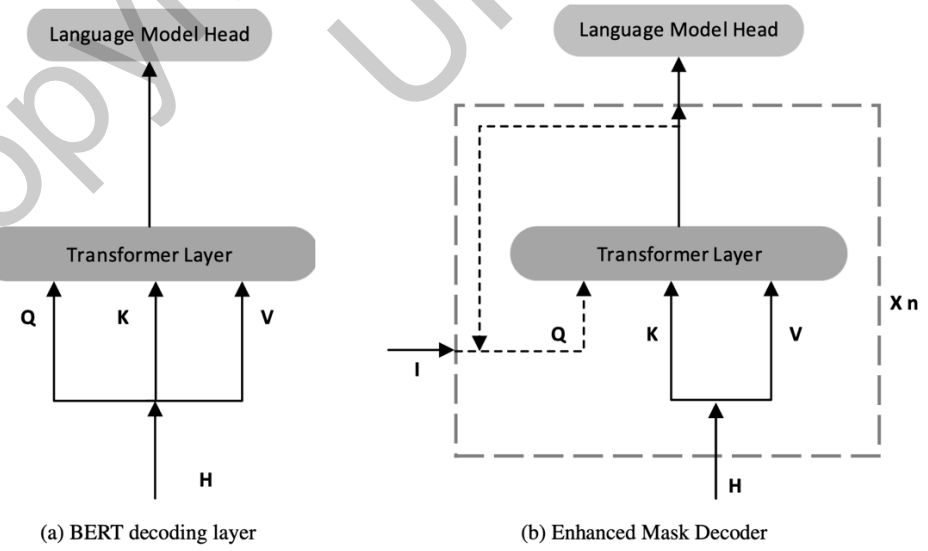
Jadual 1.1 Taburan Kelas Sebelum dan Selepas Smote

Set Data	Label	Sebelum SMOTE	Selepas SMOTE
Asal	Bukan Toksik	3002	5962
Asal	Toksik	1890	5962
Penambahan	Bukan Toksik	20000	42358
Penambahan	Toksik	4160	42358

Fasa Latihan Model

Fasa ini memberi tumpuan kepada proses melatih dua buah model pembelajaran mendalam, iaitu DeBERTa-v3-base dan TinyLlama, untuk melaksanakan tugas klasifikasi komen toksik dalam Bahasa Melayu. Latihan dilakukan dengan menggunakan dua jenis dataset yang berbeza, iaitu dataset asal (HateMalay.csv) dan dataset yang telah diperluas melalui teknik augmentasi. Pembahagian data bagi kedua-dua jenis dataset dilakukan secara stratifikasi dengan nisbah 80:20, di mana sebanyak 4,769 data digunakan untuk latihan dan 1,193 data untuk ujian dalam dataset asal, manakala dataset augmentasi mengandungi 33,886 data latihan dan 8,472 data ujian.

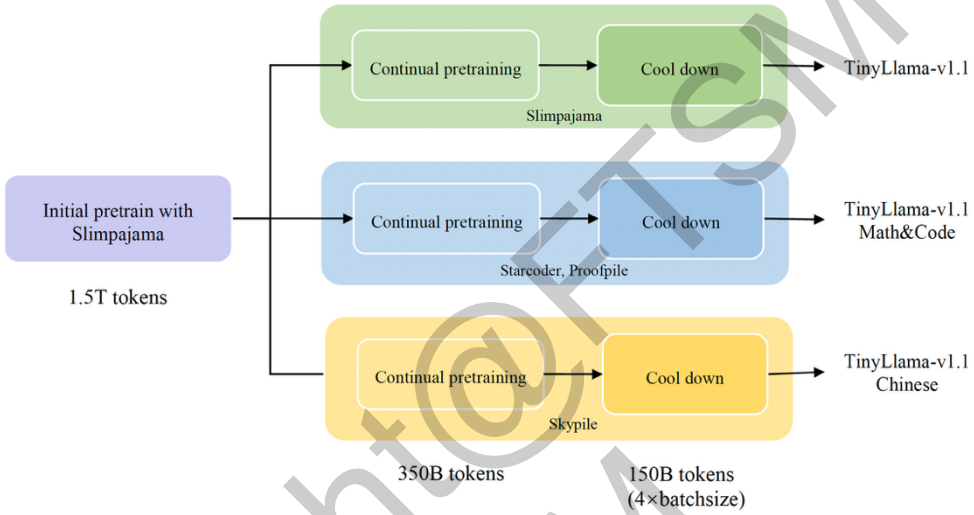
Model pertama yang digunakan dalam kajian ini ialah DeBERTa-v3-base, iaitu model transformer berprestasi tinggi yang dibangunkan oleh Microsoft. Keistimewaan model ini terletak pada mekanisme perhatian yang dipisahkan (disentangled attention), yang membolehkannya mengasingkan maklumat kedudukan dan kandungan perkataan untuk menghasilkan pemahaman konteks yang lebih mendalam dan tepat (Liu et al., 2020). Dalam kajian ini, DeBERTa dilatih selama lima epoch dengan menggunakan teknik fine-tuning terhadap model terlatih sedia ada, serta menggunakan Adam optimizer dan mekanisme EarlyStoppingCallback untuk mengelakkan masalah overfitting. Model ini sangat sesuai untuk menangani komen-komen toksik yang panjang, berlapis makna dan kompleks dari segi struktur bahasa. Rajah 3.10 menunjukkan seni bina model DeBERTa-v3-base.



Rajah1.1 Seni Bina DeBerta-v3-base

Model kedua pula ialah TinyLlama, iaitu model transformer ringan dan efisien yang dibangunkan berdasarkan keluarga model LLaMA oleh Meta AI. TinyLlama dirancang untuk memberikan inferens yang cepat dan menjimatkan memori,

menjadikannya ideal bagi penggunaan dalam sistem masa nyata atau pada peranti berkeupayaan rendah (Hugging Face, 2024; Zhai, 2022). Latihan model ini dijalankan selama tiga epoch sahaja bagi mengekalkan kecekapan tanpa mengorbankan ketepatan. TinyLlama menggunakan pendekatan pretraining yang dioptimumkan serta digabungkan dengan teknik tokenisasi pintar, padding, truncation dan pelaksanaan latihan menggunakan format pemprosesan ketepatan bercampur (fp16) untuk mempercepatkan proses latihan. Rajah 3.11 menunjukkan seni bina model TinyLlama.



Rajah 1.2 Seni Bina Tiny Llama

Latihan kedua-dua model dilakukan dengan sokongan pustaka Hugging Face Transformers dan teknik klasifikasi yang konsisten. Selain itu, kedua-dua model dan tokenizer masing-masing disimpan ke dalam direktori khas selepas latihan untuk membolehkan proses inferens dilakukan pada masa hadapan tanpa perlu melatih semula. Lokasi penyimpanan hasil latihan untuk TinyLlama ialah *results_Original_TinyLlama* dan *results_Merged_Expanded_TinyLlama*, manakala untuk *DeBERTa* ialah *results_Original_deberta* dan *results_Merged_Expanded_deberta.* Secara keseluruhannya, latihan model dalam kajian ini telah berjaya menghasilkan model klasifikasi yang berupaya mengenal pasti komen toksik dengan tahap ketepatan yang tinggi, sama ada menggunakan dataset asal mahupun dataset yang telah diperkaya melalui proses augmentasi. Prestasi yang memberangsangkan ini turut membuktikan keberkesanan pemilihan model dan strategi latihan yang digunakan.

Fasa Pengujian

Fasa pengujian adalah langkah penting dalam memastikan keberkesanan dan prestasi model pengesanan komen toksik yang telah dilatih. Fasa ini melibatkan ujian terhadap model menggunakan set data ujian yang berasingan daripada data latihan. Tujuannya

adalah untuk menilai prestasi model dalam menghasilkan klasifikasi komen toksik dengan tepat tanpa bias terhadap data latihan yang telah diproses.

Metrik Pengujian

Untuk menilai keberkesanan model, beberapa metrik utama digunakan, termasuk ketepatan (accuracy), precision, recall, dan F1-score. Berikut adalah penjelasan dan formula untuk setiap metrik:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Samples}}$$

Rajah 1.3 Formula Ketepatan

i. Ketepatan (Accuracy):

Ketepatan mengukur peratusan klasifikasi yang betul daripada jumlah keseluruhan data yang diuji. Formula ketepatan adalah:

Di mana:

a. **True Positives (TP):** Jumlah komen toksik yang diklasifikasikan dengan betul.

b. **True Negatives (TN):** Jumlah komen bukan toksik yang diklasifikasikan dengan betul.

c. **Total Samples:** Jumlah keseluruhan data ujian.

ii. Kepersisan:

Kepersisan mengukur berapa tepat model dalam mengklasifikasikan komen sebagai toksik, mengurangkan bilangan positif palsu (false positives). Formula precision adalah:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Rajah 1.4 Formula Kepersisan

iii. Recall (Dapatan Semula):

Recall mengukur kemampuan model dalam menangkap semua komen toksik yang wujud dalam data ujian, mengurangkan bilangan negatif palsu (false negatives). Formula dapatan semula adalah:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Rajah1.5 Formula Dapatan Semula

iv. Skor-F1:

Skor-F1 adalah purata harmonik antara precision dan recall, memberikan gambaran yang lebih seimbang tentang prestasi model, terutamanya dalam situasi ketidakseimbangan kelas. Formula F1-score adalah:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Rajah1.6 Formula Skor F1

Rumusan-rumusan ini memberikan ukuran kuantitatif bagi prestasi model dalam mengklasifikasikan komen toksik. Penting untuk diingat bahawa walaupun ketepatan, precision, recall, dan F1-score adalah metrik yang berharga untuk menilai kualiti klasifikasi, ia mempunyai kelemahan, seperti tidak sepenuhnya menggambarkan nuansa semantik atau konteks komen dan hanya bergantung pada hasil positif sebenar dan positif palsu. Dalam pemilihan metrik ini, skor evaluasi diberikan sebagai angka, di mana skor yang lebih tinggi menunjukkan prestasi yang lebih baik. Pemilihan metrik dapat bergantung pada keperluan dan karakteristik spesifik dari tugas pengesanan komen toksik yang dihadapi. Metrik ini sering digunakan untuk memberikan gambaran yang lebih komprehensif tentang kualiti pengesanan komen toksik yang dihasilkan oleh model. Evaluasi yang baik membantu mengidentifikasi kelemahan model dan membimbing penambahbaikan atau pengembangan lebih lanjut.

Fasa Pembangunan Prototaip

Fasa pembangunan prototaip merupakan peringkat terakhir dalam metodologi kajian ini, yang bertujuan untuk mengintegrasikan model pembelajaran mendalam yang telah dibangunkan ke dalam antara muka pengguna berbentuk aplikasi web. Prototaip ini

direka bentuk bukan sahaja untuk memudahkan pengguna akhir dalam membuat klasifikasi komen toksik secara interaktif, malah juga untuk membolehkan ujian fungsi model dijalankan secara langsung dalam persekitaran yang mesra pengguna.

Dalam pelaksanaan prototaip ini, kerangka kerja Flask telah digunakan sebagai backend utama kerana ia bersifat ringan dan sangat sesuai untuk pembangunan aplikasi web berasaskan Python. Flask mengendalikan logik pelayan, termasuk memuatkan model DeBERTa dan TinyLLaMA, menerima input komen daripada pengguna, serta menjalankan proses inferens model bagi menghasilkan keputusan klasifikasi. Di samping itu, SQLite digunakan sebagai pangkalan data untuk menyimpan log ramalan komen secara automatik, termasuk maklumat seperti teks asal, label keputusan, dan cap masa (timestamp).

Antara muka pengguna pula direka menggunakan HTML, CSS, dan Bootstrap bagi memastikan ia responsif serta boleh diakses dengan baik menerusi pelbagai peranti, sama ada komputer meja mahupun peranti mudah alih. Paparan utama antaramuka membolehkan pengguna memasukkan komen dalam Bahasa Melayu, memilih model pembelajaran mendalam yang ingin digunakan, dan menetapkan ambang ketoksikan (threshold) untuk klasifikasi. Setelah pengguna menekan butang "Klasifikasikan", keputusan akan dipaparkan dalam bentuk kad interaktif yang menunjukkan sama ada komen tersebut tergolong dalam kategori Toksik atau Tidak Toksik.

Tambahan pula, satu fungsi log sistem turut disediakan melalui halaman logs.html, di mana pengguna atau penyelidik boleh menyemak semula sejarah klasifikasi yang telah dijalankan. Fungsi ini amat berguna bagi tujuan pemantauan dan penambahbaikan sistem dari semasa ke semasa. Keseluruhan aplikasi ini telah diuji secara tempatan melalui pelayan localhost (127.0.0.1) bagi memastikan kestabilan, kebolehgunaan, dan kefungsian setiap komponen.

Secara keseluruhannya, fasa pembangunan prototaip ini bukan sahaja menjadikan sistem pengesanan komen toksik lebih praktikal dan boleh diakses, malah ia turut memperlihatkan potensi penggunaan model pembelajaran mendalam dalam aplikasi dunia sebenar. Prototaip yang dibangunkan ini berfungsi sebagai bukti konsep (proof-of-concept) bahawa sistem klasifikasi komen toksik dalam Bahasa Melayu mampu dilaksanakan secara menyeluruh dan berskala kecil untuk tujuan penyelidikan atau aplikasi pendidikan.

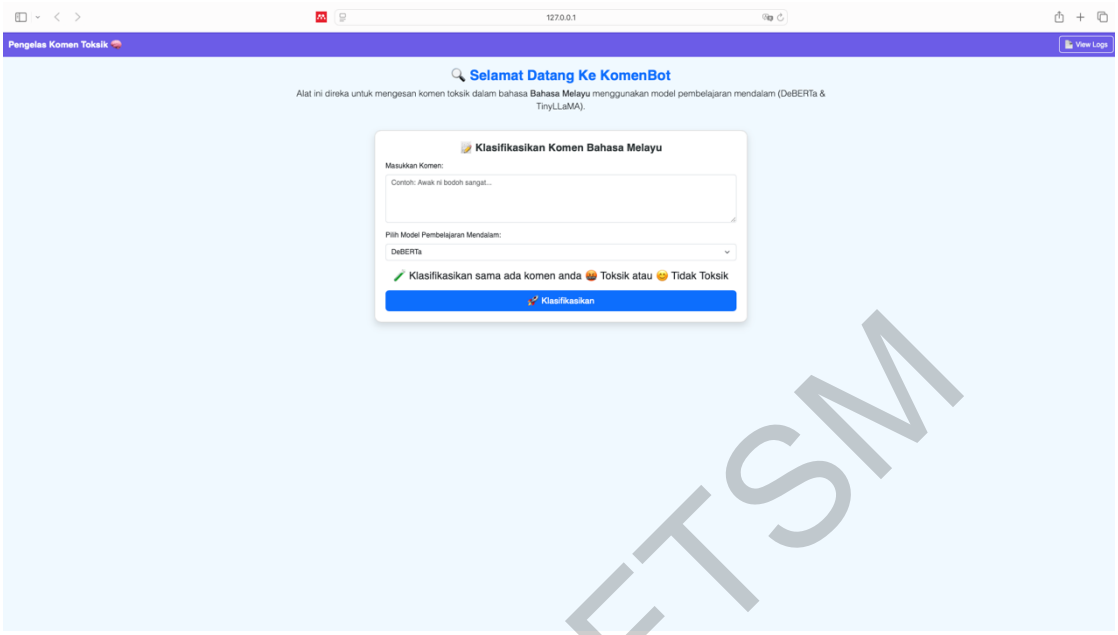
Antara muka sistem ini membolehkan pengguna untuk berinteraksi secara langsung dengan sistem klasifikasi komen toksik. Melalui kotak input yang disediakan, pengguna boleh memasukkan teks komen secara manual untuk dianalisis. Di samping itu, pengguna juga diberi pilihan untuk memilih model pembelajaran mendalam yang diingini, iaitu sama ada DeBERTa atau TinyLLaMA. Pemilihan model ini dilaksanakan

melalui menu lungsur yang terletak di bawah kotak input. Setelah komen dimasukkan dan model dipilih, pengguna boleh menekan butang “Klasifikasikan” yang berwarna biru terang. Butang ini akan memulakan proses inferens menggunakan model yang telah dilatih untuk menentukan sama ada komen yang dimasukkan bersifat toksik atau tidak. Keputusan klasifikasi akan dipaparkan serta-merta dalam bentuk teks ringkas yang menunjukkan label “Toksik” atau “Tidak Toksik”. Paparan ini disusun dengan kemas di bawah kotak input, menjadikan pengalaman pengguna lebih jelas dan mudah difahami.

Selain itu, antara muka ini turut menyediakan beberapa ciri tambahan yang menambah nilai kepada pengalaman pengguna. Sebagai contoh, navigasi bar di bahagian atas halaman mengandungi pautan ke halaman “Log” yang membolehkan pengguna melihat sejarah klasifikasi komen yang telah dilakukan. Ini bukan sahaja memberikan kefahaman terhadap prestasi sistem, tetapi juga membolehkan pengguna menilai semula keputusan sebelumnya. Reka bentuk navigasi ini disusun dalam warna ungu gelap yang kontras dengan latar halaman, seterusnya meningkatkan keterlihatan dan aksesibiliti elemen-elemen utama.

Antara muka ini juga disusun menggunakan susun atur kad Bootstrap dan grid yang kemas. Setiap komponen seperti label, input dan butang diletakkan dengan margin dan padding yang konsisten bagi memudahkan interaksi dan pembacaan. Warna ikon dan teks digunakan dengan harmoni untuk menyerlahkan perbezaan antara tindakan utama dan sekunder, seperti butang “Kosongkan” yang disediakan dalam warna kelabu lembut sebagai fungsi pilihan untuk menetapkan semula kandungan borang.

Secara keseluruhannya, reka bentuk antara muka sistem ini bukan sahaja memenuhi keperluan fungsi, tetapi juga menekankan prinsip rekabentuk berpusatkan pengguna. Ia menyokong objektif utama sistem untuk menyediakan platform pengesanan komen toksik dalam Bahasa Melayu yang mudah diakses, intuitif dan boleh digunakan oleh pelbagai lapisan pengguna. Rajah 1.7 menunjukkan Reka Bentuk Antara Muka Sistem KomenBot.



Rajah1.7 Reka Bentuk Antara Muka Sistem KomenBot

KEPUTUSAN DAN PERBINCANGAN

Dalam pembangunan sistem klasifikasi komen toksik dalam Bahasa Melayu ini, pengujian prestasi model merupakan satu aspek penting untuk memastikan kebolehgunaan sistem dalam konteks sebenar. Bagi tujuan ini, dua set data yang berbeza telah digunakan, iaitu Set Data 1 yang terdiri daripada dataset asal yang telah dibersihkan dan diseimbangkan (Cleaned + Balanced HateMalay.csv), serta Set Data 2 yang merupakan hasil daripada proses augmentasi lanjutan yang melibatkan teknik penggantian sinonim dan penjanaan komen menggunakan model GPT-2 (Cleaned + Balanced + Synonym Augmented + GPT-2 Augmented). Kedua-dua set data ini digunakan untuk melatih serta menilai prestasi dua model transformer utama iaitu DeBERTa-v3-base dan TinyLLaMA. Penilaian dilakukan menggunakan metrik standard pembelajaran mesin yang merangkumi ketepatan (accuracy), keperisian (precision), dapatan (recall) dan skor-F1 (F1-score). Dapatan daripada ujian prestasi ini bukan sahaja membantu menentukan keberkesanan model dalam mengenal pasti komen toksik, malah memberikan gambaran tentang kesesuaian model untuk digunakan dalam situasi sebenar.

Prestasi Model Berdasarkan Set Data 1

Set Data 1 digunakan sebagai asas pengujian prestasi model dalam keadaan tanpa augmentasi. Data dalam set ini telah dibersihkan dan diseimbangkan bagi memastikan taburan antara label komen toksik dan bukan toksik adalah seimbang. Matlamat utama pengujian ini adalah untuk menilai keupayaan asas model dalam mengklasifikasikan komen tanpa bantuan variasi linguistik tambahan. Berdasarkan Jadual 4.2, prestasi model TinyLLaMA didapati lebih tinggi berbanding model DeBERTa-v3-base dalam kesemua metrik. Model TinyLLaMA mencatatkan ketepatan sebanyak 0.69, keperisian 0.68, dapatan 0.69, dan skor-F1 sebanyak 0.67, manakala model DeBERTa-v3-base mencatatkan skor yang lebih rendah dengan ketepatan, keperisian, dan dapatan masing-masing sebanyak 0.61, serta skor-F1 hanya 0.47. Skor-F1 yang rendah ini menunjukkan bahawa model DeBERTa mengalami kesukaran dalam mengenal pasti komen toksik apabila hanya dilatih dengan data asas. Sebaliknya, prestasi yang ditunjukkan oleh TinyLLaMA memperlihatkan bahawa model bersaiz lebih kecil tetapi ringan ini mampu untuk belajar dan mengekstrak corak linguistik yang relevan dengan lebih konsisten dalam keadaan data yang terhad. Hal ini membuktikan bahawa dalam situasi dengan sumber data yang terhad, model berskala ringan seperti TinyLLaMA adalah lebih sesuai untuk digunakan, terutamanya dalam aplikasi masa nyata atau sistem yang perlu dijalankan pada perkakasan yang mempunyai sumber terhad. Secara keseluruhan, dapatan ini memberikan gambaran awal bahawa pemilihan model yang sesuai sangat bergantung kepada saiz dan kerumitan data yang digunakan semasa latihan.

Jadual 1.1 Prestasi Model Berdasarkan Set Data 1

Metrik Penilaian	Ketepatan	Kepersisan	Dapatan Semula	F1-Skor
Deberta-v3-base	0.61	0.61	0.61	0.47
tinyllama	0.69	0.68	0.69	0.67

Prestasi Model Berdasarkan Set Data 2

Set Data 2 dibangunkan melalui proses augmentasi data yang merangkumi dua kaedah utama iaitu penggantian sinonim dan penjanaan teks baharu menggunakan model GPT-2. Pendekatan ini dilaksanakan bagi memperkaya struktur linguistik dalam dataset, seterusnya meningkatkan kebolehan model untuk membuat generalisasi ke atas pelbagai bentuk komen toksik yang berpotensi. Hasil daripada pengujian yang ditunjukkan dalam Jadual 4.3 mendapati kedua-dua model mengalami peningkatan prestasi yang ketara berbanding Set Data 1. Model DeBERTa-v3-base mencapai ketepatan sebanyak 0.94, keperisian 0.94, dapatan 0.94, dan skor-F1 0.94, manakala model TinyLLaMA mencatatkan ketepatan dan dapatan yang sama iaitu 0.94, tetapi sedikit lebih tinggi dari segi keperisian (0.95) dan skor-F1 (0.94). Ini menunjukkan bahawa penambahan data melalui teknik augmentasi berjaya membantu kedua-dua model menghasilkan keputusan klasifikasi yang lebih seimbang dan tepat. Di samping itu, dapatan ini memperlihatkan kebergantungan model DeBERTa terhadap kuantiti dan kepelbagaian data semasa latihan. Model ini hanya mampu menunjukkan prestasi tinggi apabila diberi input latihan yang mencukupi dan pelbagai. Sebaliknya, walaupun bersaiz lebih kecil, TinyLLaMA mampu mencapai prestasi yang hampir setanding apabila dilatih dengan data yang telah ditambah baik. Rajah 4.2 dan Rajah 4.3 turut menyokong dapatan ini melalui visualisasi matriks kekeliruan dan skor-F1 yang menunjukkan jumlah kesilapan klasifikasi yang sangat rendah dan peningkatan metrik prestasi selepas augmentasi.

Secara keseluruhannya, keputusan daripada Set Data 2 membuktikan bahawa strategi augmentasi data merupakan kaedah yang efektif dalam meningkatkan prestasi model pembelajaran mendalam untuk tugas klasifikasi komen toksik dalam Bahasa Melayu. TinyLLaMA bukan sahaja menunjukkan prestasi cemerlang, malah lebih efisien dari segi penggunaan sumber dan masa inferens, menjadikannya model yang sesuai untuk pelaksanaan sebenar dalam aplikasi pengesanan komen toksik yang bersifat ringan dan berasaskan web.

Jadual 1.1 Prestasi Model Berdasarkan Set Data 2

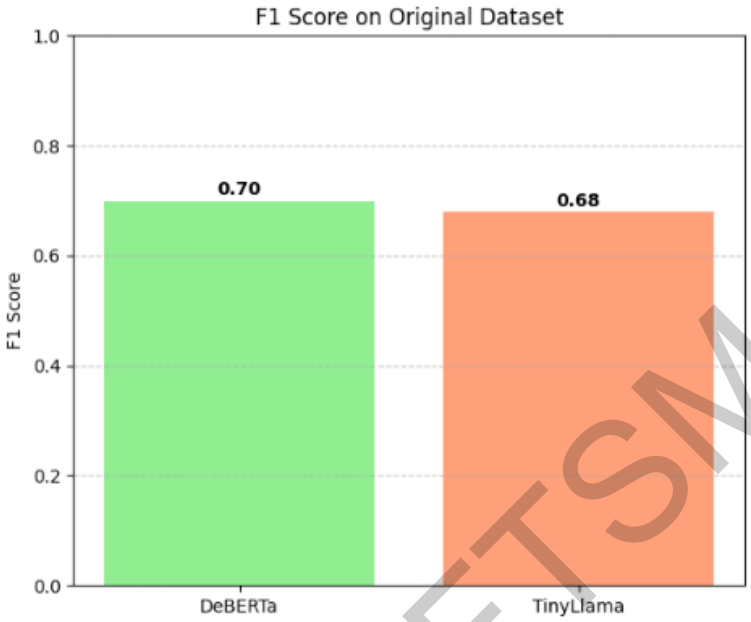
Metrik Penilaian	Ketepatan	Kepersisan	Dapatan Semula	F1-Skor
deberta-v3-base	0.94	0.94	0.94	0.94
tinyllama	0.94	0.95	0.94	0.94

PERBANDINGAN HASIL KEPUTUSAN PENGUJIAN MODEL

Bagi menilai keupayaan dan keberkesanan setiap model yang dibangunkan, perbandingan prestasi telah dilakukan antara dua model utama iaitu DeBERTa dan TinyLLaMA. Perbandingan ini dilakukan berdasarkan dua set data yang berbeza, iaitu Set Data 1 (tanpa augmentasi) dan Set Data 2 (dengan augmentasi sinonim dan GPT-2). Beberapa aspek dianalisis termasuklah skor-F1, matriks kekeliruan, trend skor-F1 mengikut epoch, nilai kerugian (loss) semasa latihan dan penilaian, serta saiz model dan masa inferens. Bagi memperkukuh penemuan, setiap bahagian disertakan dengan visualisasi dalam bentuk graf dan matriks kekeliruan.

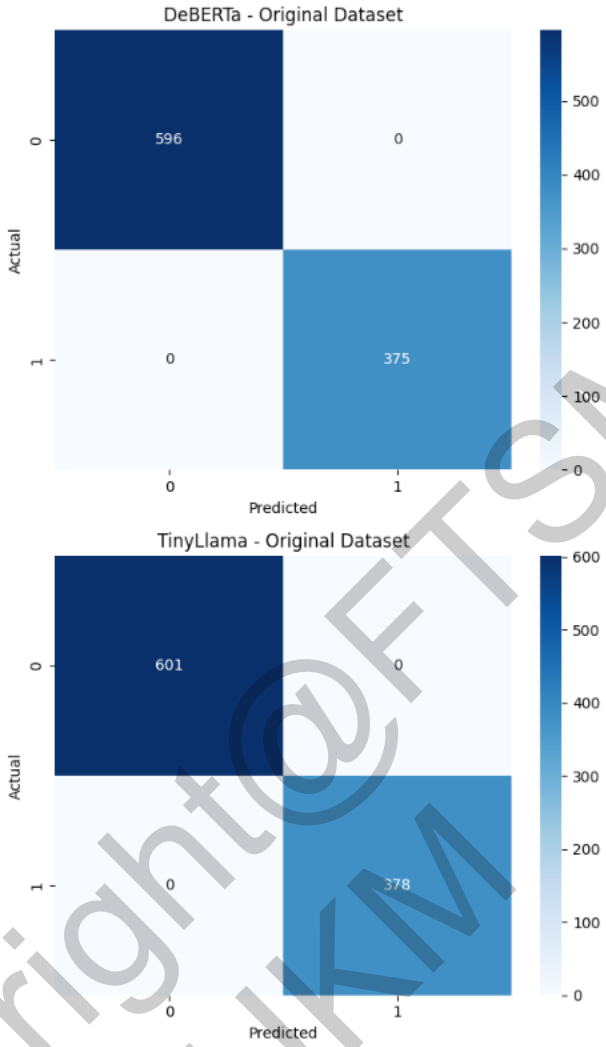
Keputusan Berdasarkan Set Data 1

Berdasarkan Rajah 1.8, graf bar menunjukkan skor-F1 yang dicapai oleh kedua-dua model apabila diuji menggunakan Set Data 1. Model DeBERTa mencatatkan skor-F1 sebanyak 0.70 manakala TinyLLaMA pula memperoleh nilai sedikit lebih rendah iaitu 0.68. Perbezaan ini menggambarkan bahawa walaupun DeBERTa adalah model berskala besar, prestasinya hanya sedikit mengatasi TinyLLaMA apabila dilatih dengan set data asal tanpa sebarang augmentasi. Ini menunjukkan bahawa tanpa sokongan data tambahan, kelebihan saiz dan parameter pada DeBERTa tidak memberi impak yang signifikan terhadap prestasi klasifikasi.



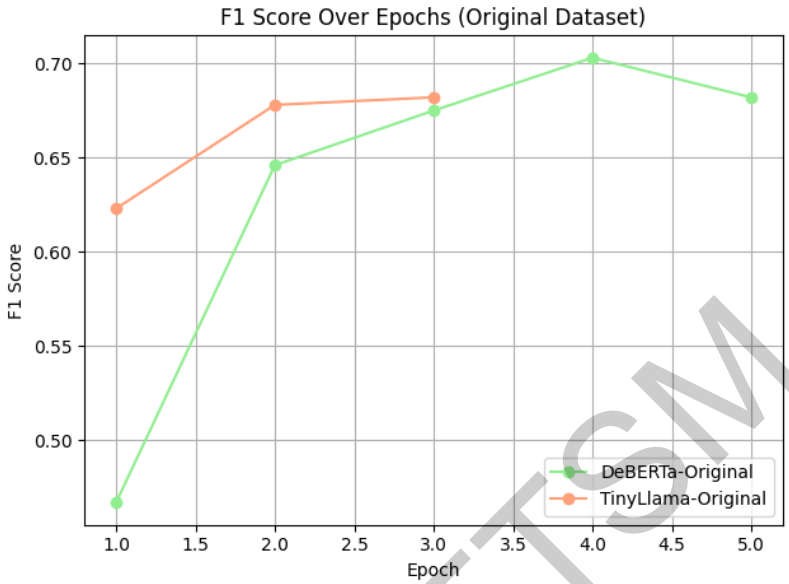
Rajah1.8 Skor F1 Set Data Asal

Selanjutnya, Rajah 1.9 memaparkan matriks kekeliruan bagi kedua-dua model menggunakan set data yang sama. Bagi DeBERTa, model berjaya mengklasifikasikan 596 komen bukan toksik dan 375 komen toksik dengan tepat. TinyLLaMA pula menunjukkan keputusan hampir setara, dengan 601 komen bukan toksik dan 378 komen toksik berjaya diklasifikasikan. Tiada kesilapan klasifikasi yang besar dikesan, namun prestasi sedikit menurun berbanding apabila data telah ditambah secara buatan.



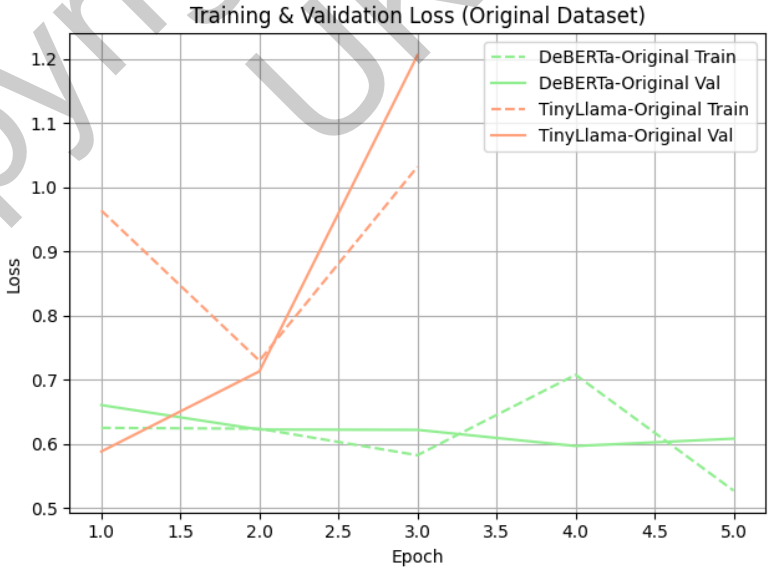
Rajah1.9 Matriks Kekeliruan Set Data Asal

Dalam Rajah 1.10, trend perubahan skor-F1 mengikut epoch bagi kedua-dua model ditunjukkan secara visual. Graf ini jelas menunjukkan peningkatan skor-F1 sepanjang proses latihan, di mana DeBERTa memperlihatkan pertumbuhan yang lebih signifikan daripada epoch pertama sehingga epoch keempat, mencapai puncak sebelum sedikit menurun. TinyLLaMA pula memperlihatkan kestabilan prestasi dengan pertumbuhan yang lebih linear dan terkawal.



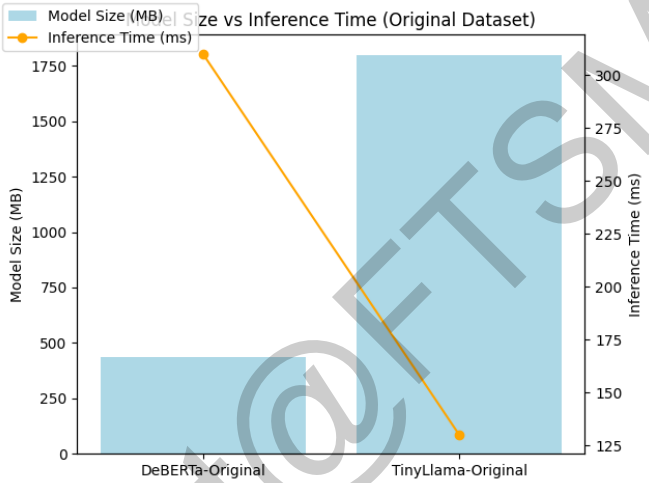
Rajah1.10 Skor F1 Mengikut Epoch Set Data Asal

Seterusnya, Rajah 1.11 menggambarkan nilai kerugian semasa latihan dan penilaian untuk Set Data 1. Dapatan menunjukkan bahawa DeBERTa mencatatkan nilai kerugian yang lebih stabil dan rendah berbanding TinyLLaMA, yang mengalami peningkatan loss pada akhir epoch. Ini menunjukkan bahawa model DeBERTa lebih baik dalam generalisasi terhadap data ujian, walaupun kelebihannya masih sederhana.



Rajah1.11 Nilai Kerugian Semasa Latihan dan Penilaian Set Data Asal

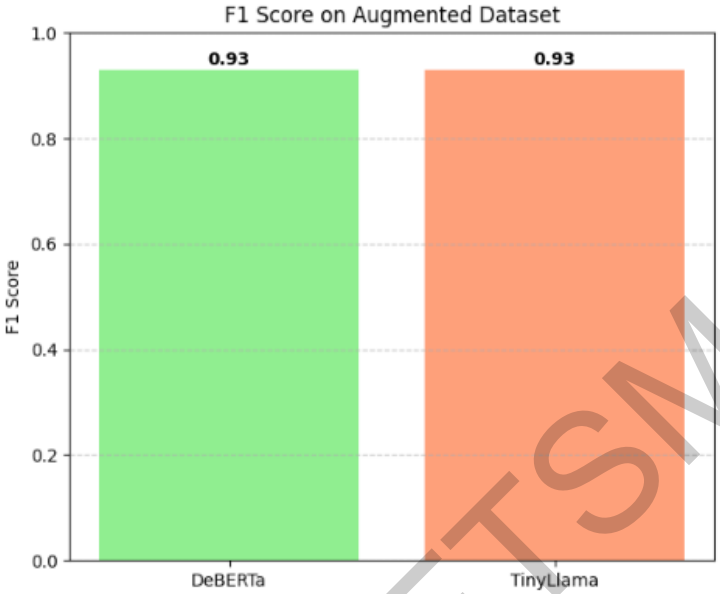
Akhir sekali bagi set data ini, Rajah 1.12 menunjukkan perbandingan saiz model dan masa inferens antara kedua-dua model. DeBERTa dengan saiz model lebih kecil (~435MB) mengambil masa lebih lama untuk inferens (~310ms), manakala TinyLLaMA walaupun bersaiz lebih besar (~1.8GB), mempunyai masa inferens yang jauh lebih pantas (~130ms). Hal ini menandakan bahawa walaupun DeBERTa lebih ringan, TinyLLaMA lebih efisien dalam menjana keputusan klasifikasi.



Rajah1.12 Perbandingan Saiz Model dan Masa Inferens Set Data Asal

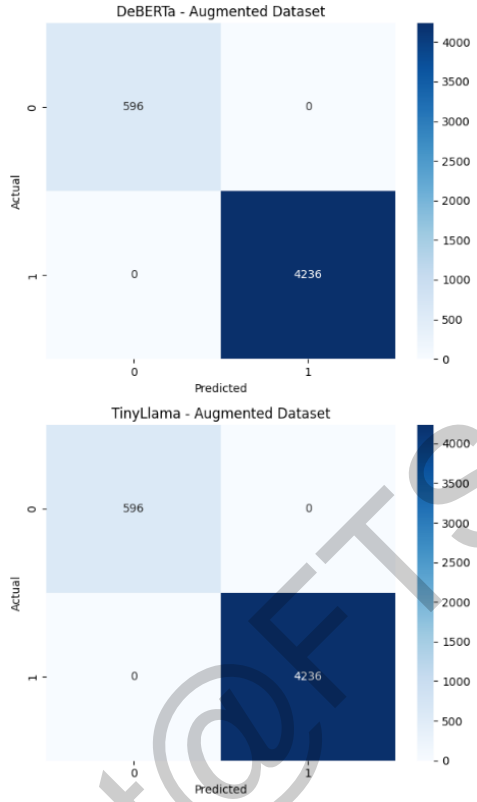
Keputusan Berdasarkan Set Data 2

Bagi Set Data 2 yang telah diperluas menggunakan teknik augmentasi sinonim dan GPT-2, prestasi kedua-dua model menunjukkan peningkatan yang ketara. Rajah 1.13 memperlihatkan bahawa kedua-dua DeBERTa dan TinyLLaMA mencatatkan skor-F1 yang sama tinggi iaitu 0.94. Ini menunjukkan bahawa penggunaan data yang diperbanyak dan diperkaya memberi kesan positif terhadap keberkesanan model, khususnya dalam mengenal pasti komen toksik dalam Bahasa Melayu.



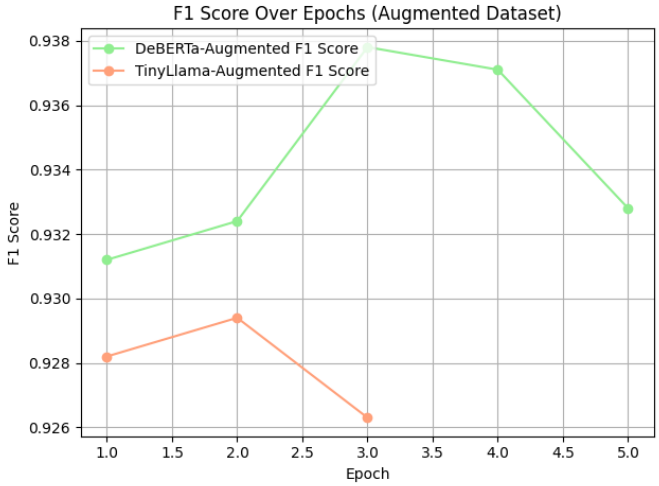
Rajah1.13 Skor-F1 Set Data Augmentasi

Penemuan ini turut disokong oleh Rajah 1.14 yang menunjukkan matriks kekeliruan bagi kedua-dua model apabila diuji dengan set data augmentasi. Kedua-dua model menunjukkan ketepatan mutlak dalam klasifikasi, di mana semua komen toksik dan bukan toksik berjaya dikelaskan dengan tepat (596 bukan toksik dan 4236 toksik bagi setiap model). Ketiadaan kesilapan klasifikasi dalam matriks ini menggambarkan kebolehan kedua-dua model dalam memahami struktur dan pola bahasa toksik yang kompleks apabila dilatih dengan data yang mencukupi dan pelbagai.



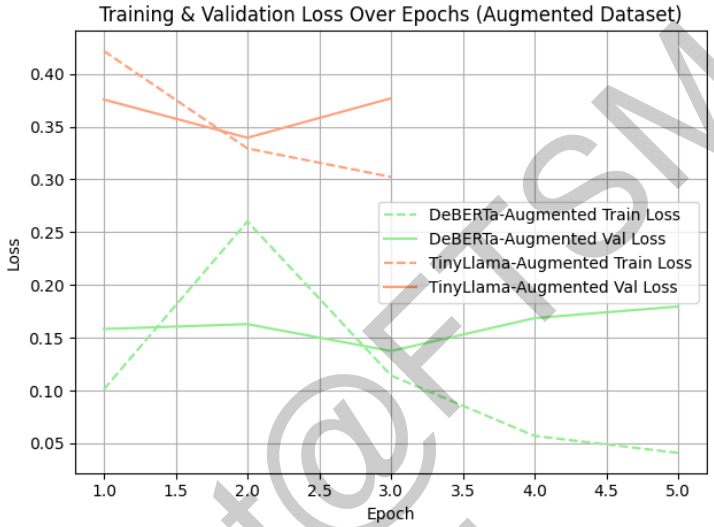
Rajah1.14 Matriks Kekeliruan Set Data Augmentasi

Seterusnya, Rajah 1.15 menggambarkan perubahan skor-F1 mengikut epoch bagi set data yang telah diaugmentasi. Kedua-dua model menunjukkan peningkatan skor-F1 sepanjang sesi latihan, dengan DeBERTa mencapai puncak tertinggi pada epoch ke-3 sebelum sedikit menurun. TinyLLaMA pula menunjukkan peningkatan pada epoch ke-2 namun mengalami penurunan kecil pada epoch berikutnya. Ini memperlihatkan kestabilan model dalam mengadaptasi variasi linguistik dalam data latihan.



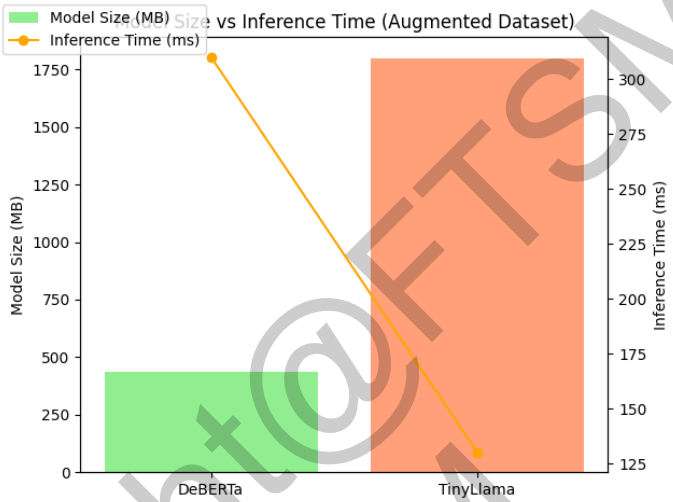
Rajah 1.15 Skor F1 Mengikut Epoch Set Data Augmentasi

Dalam Rajah 1.16, graf kerugian latihan dan penilaian bagi Set Data 2 menunjukkan bahawa DeBERTa mencatatkan nilai loss yang lebih rendah dan stabil sepanjang epoch berbanding TinyLLaMA. Ini menandakan bahawa DeBERTa mampu mengadaptasi lebih baik terhadap data yang kompleks dan mempersembahkan prestasi konsisten.



Rajah 1.16 Nilai Kerugian Semasa Latihan dan Penilaian Set Data Augmentasi

Akhir sekali, Rajah 1.17 menunjukkan bahawa walaupun DeBERTa masih lebih ringan dari segi saiz model, masa inferensnya kekal lebih lama berbanding TinyLLaMA. Walaupun TinyLLaMA mempunyai saiz model hampir empat kali ganda, ia mampu memberikan keputusan dengan masa lebih pantas. Maka, TinyLLaMA lebih sesuai diaplikasikan dalam sistem masa nyata yang memerlukan inferens pantas, manakala DeBERTa pula lebih sesuai dalam aplikasi berasaskan ketepatan klasifikasi tinggi dalam sistem berskala besar.



Rajah 1.17 Perbandingan Saiz Model dan Masa Inferens Set Data Augmentasi

Cadangan Penambahbaikan

Bagi meningkatkan keberkesanan sistem pada masa hadapan, beberapa penambahbaikan telah dikenalpasti berdasarkan kekangan semasa dan potensi peluasan kajian. Pertama sekali, usaha pengumpulan dan pembinaan set data komen Bahasa Melayu yang lebih besar, pelbagai, dan dilabel secara manual daripada sumber tempatan seperti media sosial, forum, serta komen berita adalah sangat digalakkan. Langkah ini bertujuan untuk memperkayakan korpus latihan dan meningkatkan kebolehan model dalam memahami variasi linguistik dalam pelbagai konteks penggunaan sebenar. Kedua, penggunaan kaedah pembelajaran pemindahan (transfer learning) dan penalaan halus (fine-tuning) terhadap model transformer sedia ada boleh dipertingkatkan bagi membolehkan model mengenal pasti struktur ayat dalam Bahasa Melayu dengan lebih berkesan.

Seterusnya, teknik augmentasi data boleh diperluaskan dengan menambah kaedah back translation, iaitu menterjemah ayat Bahasa Melayu ke bahasa lain seperti Bahasa Inggeris dan kemudian menterjemahkannya semula ke Bahasa Melayu.

Pendekatan ini dapat memperkaya struktur linguistik tanpa menjejaskan maksud asal dan seterusnya meningkatkan kebolehan generalisasi model. Di samping itu, kaedah kejuruteraan ciri (feature engineering) juga boleh diterokai sebagai pendekatan hibrid bagi menggabungkan maklumat linguistik seperti bilangan kata kesat, kehadiran emoji, panjang ayat, dan skor sentimen sebagai penambahan kepada input model. Ini berpotensi memperkukuhkan prestasi model dalam mengenal pasti komen toksik secara lebih kontekstual dan tepat.

Selain itu, penggunaan teknik pengoptimuman memori seperti gradient checkpointing dan model quantization perlu dipertimbangkan untuk membolehkan proses latihan model dijalankan dalam persekitaran komputasi yang lebih ringan, terutamanya dalam konteks akademik atau organisasi yang terhad sumber perkakasannya. Seterusnya, pembangunan sistem pengesanan berbasis peraturan (rule-based) yang menyokong model pembelajaran mendalam juga disyorkan sebagai lapisan awal penapisan. Pendekatan ini boleh meningkatkan kecekapan sistem dalam mengenal pasti komen toksik secara eksplisit tanpa perlu melakukan inferens penuh.

Akhir sekali, aspek antara muka pengguna juga boleh ditambah baik dari segi fungsi dan reka bentuk. Fungsi-fungsi tambahan seperti analisis statistik komen, sejarah klasifikasi, visualisasi data dalam bentuk graf interaktif, dan tetapan ambang (threshold) klasifikasi yang boleh dilaras oleh pengguna boleh menjadikan sistem ini lebih fleksibel dan mesra pengguna. Kesemua cadangan ini dapat memperluaskan skop aplikasi sistem ke persekitaran sebenar, sekali gus menyumbang kepada pembangunan teknologi bahasa tempatan yang lebih canggih dan inklusif.

KESIMPULAN

Kesimpulannya, projek ini telah berjaya menghasilkan satu sistem pengesanan komen toksik dalam Bahasa Melayu yang berasaskan pendekatan pembelajaran mendalam dan dapat dijalankan secara dalam talian melalui antara muka web yang interaktif. Model transformer yang digunakan, iaitu DeBERTa dan TinyLLaMA, telah menunjukkan keupayaan yang tinggi dalam tugasan klasifikasi teks terutamanya apabila dilatih dengan set data yang telah diperkaya melalui teknik augmentasi. Walaupun terdapat kekangan dari aspek data dan sumber komputasi, projek ini telah membuktikan bahawa pembelajaran mendalam boleh digunakan secara berkesan untuk menganalisis kandungan teks dalam Bahasa Melayu. Kejayaan ini membuka peluang kepada kajian lanjutan dalam bidang pemprosesan bahasa semula jadi tempatan dan pembangunan sistem kecerdasan buatan yang lebih kontekstual dan inklusif terhadap bahasa ibunda. Sistem ini juga berpotensi untuk digunakan dalam pelbagai aplikasi sebenar seperti

penapisan kandungan media sosial, pemantauan komen dalam talian, dan pendidikan digital bagi memupuk persekitaran komunikasi yang sihat dan bertanggungjawab.

PTA-FTSM-2025-A192047

Kekuatan Sistem

Sistem yang dibangunkan menunjukkan beberapa kekuatan utama yang menyumbang kepada keberkesanan dan kebolehgunaannya dalam kalangan pengguna. Pertama, sistem ini berupaya menghasilkan ringkasan yang padat dan menyeluruh dengan mengekstrak idea-idea utama daripada teks asal. Keupayaan ini sangat penting dalam membantu pengguna memperoleh maklumat penting dengan cepat tanpa perlu membaca keseluruhan kandungan.

Kedua, antara muka pengguna yang mesra dan intuitif turut meningkatkan pengalaman pengguna secara keseluruhan. Ciri fleksibiliti dalam menyesuaikan panjang ringkasan mengikut keperluan pengguna memberikan nilai tambah yang signifikan, menjadikan sistem ini lebih mudah diakses oleh pelbagai peringkat pengguna.

Ketiga, sistem ini juga mempunyai potensi besar dalam memajukan bidang Pemprosesan Bahasa Semulajadi (PBT) khususnya bagi Bahasa Melayu. Ia menyediakan alat yang canggih dan bermanfaat bukan sahaja untuk tujuan akademik, malah juga sesuai digunakan dalam konteks profesional dan harian. Oleh itu, sistem ini dilihat sebagai satu inisiatif yang menyokong pembangunan teknologi bahasa tempatan.

Kelemahan Sistem

Walau bagaimanapun, sistem ini turut menghadapi beberapa kelemahan yang perlu diberi perhatian. Antara cabaran utama ialah kekurangan dataset Bahasa Melayu yang besar dan berkualiti tinggi. Kekurangan ini boleh menjejaskan ketepatan serta kualiti ringkasan yang dihasilkan oleh sistem.

Selain itu, penggunaan dataset Bahasa Inggeris berskala besar bagi tujuan latihan model telah menyebabkan peningkatan masa latihan dan penggunaan memori yang tinggi. Keadaan ini boleh membawa kepada isu seperti sistem terhempas serta peningkatan kos dari segi sumber pengkomputeran.

Kualiti ringkasan yang dihasilkan juga masih memerlukan penambahbaikan, terutamanya dari segi ketepatan dan kerelevanan kandungan. Walaupun teknik seperti pembelajaran pemindahan (transfer learning) dan penggandaan data (data augmentation) digunakan, hasil yang diperoleh tidak sentiasa konsisten dan sangat bergantung kepada kualiti data yang digunakan.

Tambahan pula, penggunaan model yang dilatih secara dominan dengan dataset Bahasa Inggeris boleh menjejaskan keupayaan sistem dalam menjana ringkasan yang tepat dalam Bahasa Melayu. Keadaan ini menuntut penggunaan proses tambahan seperti terjemahan teks, yang secara tidak langsung boleh mengurangkan ketepatan hasil ringkasan.

PTA-FTSM-2025-A192047

PENGHARGAAN

Pertama sekali, saya ingin memanjatkan kesyukuran ke hadrat Allah SWT atas rahmat dan limpah kurnia-Nya yang telah memberikan saya kekuatan dan kesabaran, dalam menyelesaikan projek akhir tahun saya yang bertajuk Pengelasan Komen Toksik Menggunakan Pembelajaran Mendalam.

Ucapan terima kasih yang tidak terhingga saya tujukan khas kepada penyelia saya yang dihormati, Prof. Madya Dr. Nazlia Binti Omar, atas segala bimbingan, nasihat dan dorongan berterusan yang sangat berharga sepanjang projek ini dijalankan. Beliau telah memberi impak yang besar dalam memastikan kejayaan projek ini.

Saya juga ingin mengambil kesempatan ini untuk merakamkan setinggi-tinggi penghargaan kepada Fakulti Teknologi dan Sains Maklumat (FTSM) atas segala kemudahan yang disediakan. Terima kasih juga diucapkan kepada semua pensyarah yang telah mendidik saya sepanjang saya bergelar mahasiswi di Universiti Kebangsaan Malaysia (UKM) serta kakitangan fakulti yang telah memberikan sokongan teknikal dan moral. Tanpa sokongan daripada mereka, adalah mustahil untuk saya mencapai hasil yang memuaskan di tahap ini. Tidak boleh dilupakan juga sokongan moral, semangat, dan doa yang dicurahkan oleh kedua-dua ibu bapa saya tercinta Mohd Hilman Rajang Bin Abdullah dan Norijah Binti Mohamad Yassin. Dorongan adik-beradik yang sentiasa memberikan bantuan dan menadah telinga mendengar keluhan dan lelahan hati sejak hari pertama menjejakkan kaki di UKM sehingga sepanjang penulisan laporan ilmiah ini.

Akhir sekali, ucapan jutaan terima kasih saya tujukan kepada semua pihak yang terlibat secara langsung mahupun tidak langsung dalam menjayakan projek ini, terutamanya keluarga, rakan-rakan, dan individu-individu yang sentiasa memberikan sokongan moral dan dorongan sepanjang tempoh projek ini dijalankan.

PTA-FTSM-2025-A192047

RUJUKAN

Adewumi, T., Sabry, S. S., Abid, N., Liwicki, F., & Liwicki, M. (2023). ("Liwicki, Foteini (0000-0002-6756-0147)") T5 for Hate Speech, Augmented Data, and Ensemble. Sci, 5(4). https://doi.org/10.3390/sci5040037

Ardi, N., Ahmad, A., Daud, N., & Ismail, N. (2021). Speech Act of Flaming in Twitter Status. Asian Journal of University Education, 16(4), 109. https://doi.org/10.24191/ajue.v16i4.11961

Asrafulsyifaa. (2023). Malay Dataset [Repositori GitHub]. GitHub. https://github.com/asrafulsyifaa/Malay-Dataset/tree/master

Biere, S. (2018). Hate Speecsh Detection Using Natural Language Processing Techniques. Vrije Universiteit Amsterdam, 30.

Brownlee, J. (2020). Train/Test Split for Evaluating Machine Learning Algorithms. Machine Learning Mastery.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357.

Cocchieri, A. (2023). Toxic Comment Classification using Machine Learning [Notebook Jupyter]. GitHub. https://github.com/alessiococchieri/toxic-comment-classification/blob/main/Toxic_Comment_Classification.ipynb

Dcunha, M., & Bhangale, N. (2024). Toxic Comment Classification 1 TOXIC COMMENT CLASSSIFICATION. 12(5), 227–232.

Hashmat02. (2023). Fine-Tuning LLaMA 2 for Toxicity Classification [Notebook Jupyter]. GitHub. https://github.com/Hashmat02/Fine-Tuning-LLaMA-2-for-Toxicity-Classification/blob/main/Notebook-1.1.ipynb

Hugging Face (2024). TinyLlama Model Documentation.

Hugging Face. (2024). Chapter 3: Tokenizer – Hugging Face LLM Course. https://huggingface.co/learn/llm-course/en/chapter3/3?fw=pt

Hugging Face. (2024). Trainer class – Hugging Face Transformers documentation. https://huggingface.co/docs/transformers/main/en/main_classes/trainer

Hugging Face. (2024). XLM-RoBERTa model documentation – Hugging Face Transformers. https://huggingface.co/docs/transformers/en/model_doc/xlm-roberta

Hugging Face. TinyLlama Model Documentation. 2024. [Cited 2025 Jul 9]. Available from: https://huggingface.co/TinyLlama

Ismail, N., Losada, D. E., & Ahmad, R. (2024). A Test Dataset of Offensive Malay Language by a Cyberbullying Detection Model on Instagram Using Support Vector Machine.

PTA-FTSM-2025-A192047

Communications in Computer and Information Science, 2001 CCIS(October), 182–192. https://doi.org/10.1007/978-981-99-9589-9_14

Julian. (2019). Jigsaw Multilingual Toxic Comment Classification Dataset [Dataset]. Kaggle. https://www.kaggle.com/datasets/julian3833/jigsaw-multilingual-toxic-comment-classification/data

Liu et al. (2020). DeBERTa: Decoding-enhanced BERT with disentangled attention.

Liu P, Zhou J, He Y, Chen W, Zhao J, Tang J. DeBERTa: Decoding-enhanced BERT with disentangled attention. arXiv preprint arXiv:2006.03654. 2020.

Maity, K. (2023). A Deep Learning Framework for the Detection of Malay Hate Speech [Repositori GitHub]. GitHub. https://github.com/MaityKrishanu/Hate_Malay/blob/master/README.md#a-deep-learning-framework-for-the-detection-of-malay-hate-speech

Maity, K. 2021. A Deep Learning Framework for the Detection of Malay Hate Speech. GitHub. https://github.com/MaityKrishanu/Hate_Malay

Maity, K., Bhattacharya, S., Saha, S., & Seera, M. (2023). A Deep Learning Framework for the Detection of Malay Hate Speech. IEEE Access, 11(June), 79542–79552. https://doi.org/10.1109/ACCESS.2023.3298808

Malaya. (2024). Dokumentasi rasmi Malaya versi 5.1. https://malaya.readthedocs.io/en/5.1/

Mesolitica. (2024). malaya/translation.py [Kod Python]. GitHub. https://github.com/mesolitica/malaya/blob/master/malaya/translation.py

Microsoft. (2024). deberta-v3-base – Hugging Face. https://huggingface.co/microsoft/deberta-v3-base

Mohiuddin, K., Welke, P., Alam, M. A., Martin, M., Alam, M. M., Lehmann, J., & Vahdati, S. (2023). Retention Is All You Need. International Conference on Information and Knowledge Management, Proceedings, Nips, 4752–4758. https://doi.org/10.1145/3583780.3615497

Naseeba, B., Sai, P. H. R., Karthik, B. V. P., Chitteti, C., Sai, K., & Avanija, J. (2023). Toxic Comment Classification. Lecture Notes in Networks and Systems, 647 LNNS(1), 872–880. https://doi.org/10.1007/978-3-031-27409-1_80

Schmidt, A., & Wiegand, M. (2017). A Survey on Hate Speech Detection using Natural Language Processing. Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, 1–10. https://aclanthology.org/W17-1101

Sen, T., Das, A., & Sen, M. (2024). HateTinyLLM : Hate Speech Detection Using Tiny Large Language Models. 1–5. http://arxiv.org/abs/2405.01577

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6(1), 1–48.

PTA-FTSM-2025-A192047

Suresh, S., Yadav, B., Kumari, S., Choudhary, A., Krishika, R., & Mahesh, T. R. (2023). Performance Analysis of Comment Toxicity Detection Using Machine Learning. 2023 International Conference on Computer Science and Emerging Technologies, CSET 2023, 1–6. https://doi.org/10.1109/CSET58993.2023.10346832

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is All You Need. In: Advances in Neural Information Processing Systems (NeurIPS); 2017.

Vaswani et al. (2017). Attention is All You Need.

Venugopal, N. L. V. (2024). Multilingual Toxic Comment Classification using Deep Learning. 2024 2nd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS), 752–757. https://doi.org/10.1109/ICSSAS64001.2024.10760913

Wang, K., Yang, J., & Wu, H. (2021). A Survey of Toxic Comment Classification Methods. December. http://arxiv.org/abs/2112.06412

Wei, J., & Zou, K. (2019). EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. arXiv preprint arXiv:1901.11196.

Wolf, T., et al. (2020). Transformers: State-of-the-art Natural Language Processing. Proceedings of EMNLP: System Demonstrations, 38–45.

Zhai S. TinyLlama Pretraining Strategy. 2022.

Zhai, Z. (2022). Rating the Severity of Toxic Comments Using BERT-Based Deep Learning Method. 2022 IEEE 5th International Conference on Electronics Technology, ICET 2022, 1283–1288.

Zhang, J. (2023). TinyLlama: Open Reproduction of LLaMA on Smaller Scale [Repositori GitHub]. GitHub. https://github.com/jzhang38/TinyLlama

Nur Sabrina Binti Mohd Hilman Rajang (A192047)
PROF. MADYA DR. NAZLIA BINTI OMAR
Fakulti Teknologi & Sains Maklumat
Universiti Kebangsaan Malaysia