MODEL PENGESANAN BULI SIBER MENGGUNAKAN BERT

[1]Ashvinaah A/P Ragunathan, [1]Masnizah Binti Mohd

[1] Fakluti Teknologi & Sains Maklumat,
43600 Universiti Kebangsaan Malaysia

ABSTRAK

Projek ini membentangkan pembangunan sistem pengesanan buli siber menggunakan model BERT (*Bidirectional Encoder Representations from Transformers*). Model ini direka untuk mengesan kandungan teks toksik dalam bahasa Inggeris secara automatik, dengan tumpuan kepada komen berbentuk penghinaan, ejekan, atau ancaman di media sosial. Sistem ini dilatih dan diuji menggunakan tiga set data sumber terbuka, *HateXplain, Jigsaw*, dan *OLID*, yang meliputi pelbagai kategori ucapan ofensif. Model telah diperhalusi (*fine-tuned*) untuk mengeksploitasi keupayaan pemahaman kontekstual dua hala, membolehkan klasifikasi yang tepat antara mesej buli siber dan mesej bukan buli. Model utama yang digunakan ialah DistilBERT, versi kecekapan tinggi kepada BERT, yang mengekalkan prestasi dalam memahami konteks dua hala dengan penggunaan sumber yang lebih rendah. Model ini terdiri daripada beberapa komponen utama Seni Bina Model DistilBERT, termasuk lapisan pengekodan (*embedding layer*), enam lapisan pengekod transformer (*transformer encoder layers*) dengan mekanisme perhatian kendiri (*self-attention*), dan lapisan output (*output layer*) bagi tugas klasifikasi. Struktur seni bina ini menyokong kefungsian sistem yang modular dan efisien. Hasil pengujian menunjukkan prestasi yang memberangsangkan, dengan ketepatan (*accuracy)* mencapai 91.6% dan skor F1 sekitar 0.92 pada set ujian. Proses pembangunan mematuhi metodologi *Waterfall* yang menekankan fasa pembangunan secara berstruktur dan teratur. Walaupun demikian, sistem ini masih menghadapi kekangan dalam menangani kandungan tidak formal, slanga, dan maksud tersirat yang kompleks. Justeru, cadangan masa depan termasuk pemerkasaan pemprosesan semantik mendalam serta peluasan sokongan kepada pelbagai bahasa, bagi meningkatkan kebolehlaksanaan sistem dalam persekitaran global yang lebih meluas.

ABSTRACT

This project presents the development of a cyberbullying detection system using the BERT (Bidirectional Encoder Representations from Transformers) model. The model is designed to automatically detect toxic textual content in English, focusing on comments that contain

Copyright@FTSM UKM

insults, mockery, or threats commonly found on social media platforms. The system is trained and evaluated using three publicly available datasets, HateXplain, Jigsaw, and OLID, which cover a wide range of offensive speech categories. The model is fine-tuned to leverage BERT's bidirectional contextual understanding, enabling accurate classification between cyberbullying and non-cyberbullying messages. The core model utilized is DistilBERT, a lighter and more resource-efficient version of BERT that retains the ability to comprehend bidirectional context. As illustrated in Figure 3.5: DistilBERT Model Architecture, the model comprises several key components, including an embedding layer, six transformer encoder layers with self-attention mechanisms, and an output layer for classification tasks. This layered architecture supports modular system design and efficient processing. Experimental results demonstrate promising performance, with an accuracy of 91.6% and an F1-score of approximately 0.92 on the test set. The system was developed following a structured Waterfall methodology, ensuring clear modular design and validation phases. However, the system still faces limitations in detecting content expressed in informal language, slang, or with implicit meanings. Future enhancements include deeper semantic processing and multilingual support to improve the system's effectiveness in diverse global environments.

Copyright@FTSM UKM

1.0 PENGENALAN

Buli siber merujuk kepada penggunaan platform komunikasi digital untuk mendatangkan kemudaratan kepada individu yang tidak mampu mempertahankan diri. Fenomena ini semakin membimbangkan sejajar dengan peningkatan penggunaan peranti elektronik dalam kalangan remaja. Statistik menunjukkan bahawa antara 15% hingga 35% remaja pernah menjadi mangsa buli siber, manakala antara 10% hingga 20% pula mengaku pernah melakukannya. Faktor penyumbang utama ialah sifat anonimiti yang ditawarkan oleh media sosial, yang membolehkan pelaku bertindak tanpa perlu bertanggungjawab. Akibatnya, mangsa mungkin mengalami tekanan emosi berterusan kerana akses kepada platform komunikasi tersedia pada bila-bila masa. Buli siber boleh wujud dalam pelbagai bentuk termasuk pengecualian sosial, tipu muslihat, spam, gangguan seksual dan perkauman, fitnah, galakan buli, cyberstalking, dan trolling. Perkembangan terbaru menunjukkan bahawa pembuli siber semakin menggunakan ungkapan tersirat, bahasa berkod, serta slang yang menjadikan pengesanan secara automatik semakin mencabar, khususnya apabila kaedah berasaskan kata kunci atau peraturan digunakan.

Model pengesanan tradisional yang bergantung pada pendekatan leksikon statik tidak dapat menyesuaikan diri dengan perubahan bahasa dalam talian yang pantas, terutama bahasa

Gen Z yang mengandungi makna tersirat. Ini menyebabkan berlakunya negatif palsu apabila mesej berbahaya tidak dapat dikesan, dan positif palsu apabila mesej neutral disalah tafsir sebagai berbahaya. Tambahan pula, kerencaman budaya dan variasi linguistik merentas pengguna media sosial menuntut model yang dapat memahami bukan hanya makna literal, tetapi juga konteks budaya dan sosial sesuatu ungkapan. Oleh itu, terdapat keperluan mendesak untuk membangunkan sistem pengesanan yang lebih maju yang mampu memahami semantik dan konteks dengan lebih mendalam.

Kajian ini mencadangkan penggunaan model pembelajaran mendalam berasaskan pemprosesan bahasa tabii (PBT) iaitu DistilBERT, varian BERT yang dimampatkan. DistilBERT mampu menganalisis makna kontekstual serta mengenal pasti pola linguistik kompleks, termasuk slang dan bahasa berkod yang digunakan dalam buli siber. Model ini diharap dapat meningkatkan ketepatan pengesanan, mengurangkan kesilapan klasifikasi, dan meningkatkan kebolehsuaian terhadap variasi bahasa yang berkembang. Pendekatan yang dicadangkan juga merangkumi pembinaan leksikon tersuai yang dikemas kini secara dinamik dengan istilah baharu dari data dunia sebenar, menjadikan sistem lebih inklusif dan berdaya tahan. Kajian ini bertujuan mereka bentuk dan membina model pengesanan buli siber berasaskan BERT, menilai prestasinya, dan menguji keupayaannya menangani data media sosial dalam masa nyata. Namun, skop kajian terhad kepada teks dalam bahasa Inggeris sahaja dan tidak merangkumi kandungan multimedia mahupun integrasi langsung ke dalam platform sosial. Model juga tidak akan menjalankan analisis profil pengguna. Proses pembangunan menggunakan pendekatan metodologi *Waterfall*, yang membolehkan setiap fasa—daripada analisis keperluan, reka bentuk sistem, pelaksanaan, pengujian, hingga penyelenggaraan—dilaksanakan secara sistematik. Beberapa kekangan turut dikenal pasti, termasuk keperluan untuk set data yang mencerminkan bahasa digital sebenar serta masa untuk melaraskan model sekiranya berlaku masalah seperti *overfitting* atau *underfitting*. Secara keseluruhan, kajian ini bertujuan menghasilkan sistem pengesanan buli siber yang lebih kontekstual, tepat, dan berskala bagi menghadapi cabaran semasa dan akan datang dalam komunikasi digital.

2.0 KAJIAN LITERATUR

Buli siber semakin diiktiraf sebagai satu isu serius dalam kalangan remaja, memerlukan sistem pengesanan automatik yang cekap untuk mengurangkan kesannya seperti kemurungan, kebimbangan, dan pengasingan sosial. Dalam usaha meningkatkan kecekapan pengesanan buli siber, model berasaskan pembelajaran mendalam, terutamanya BERT (*Bidirectional Encoder*

Copyright@FTSM UKM

Representations from Transformers), telah menjadi pilihan utama disebabkan keupayaannya memahami konteks dua hala dan mengenal pasti ungkapan sinis dan berkod yang sukar dikesan oleh model tradisional.

Sorotan literatur menunjukkan beberapa pendekatan utama yang telah dicadangkan dan diuji. Alkasassbeh et al. (2024) membuktikan keberkesanan BiGRU dengan ketepatan 89.23%, manakala Khafajeh (2024) menunjukkan BERT mengatasi CNN dan LSTM dengan ketepatan 87.3%. Gabungan BERT dengan embedding GloVe dalam kerangka LSTM oleh Mahlangu dan Tu (2019) mencatatkan ketepatan 94.2%, dan Sen et al. (2024) pula melaporkan bahawa gabungan BERT dengan MLP memberikan ketepatan 92.3%. Gupta et al. (2023) mencapai ketepatan 95.0% dalam pengelasan mengikut diskriminasi sosial. DistilBERT, sebagai model yang lebih ringan, menunjukkan precision setinggi 91.17% (Saranyanath et al., 2022), menandakan kebolehgunaan praktikalnya dalam sistem masa nyata.

Namun begitu, sistem tradisional seperti *Google Perspective API* dan *ToxiGen* yang menggunakan pendekatan leksikal menunjukkan kelemahan dalam menangani variasi bahasa informal dan sindiran. Dalam masa yang sama, perkembangan terkini dalam teknologi seperti model Transformer ringan (Philipo et al., 2024), teknik *Emotion-Adaptive Training* (Yi et al., 2025), integrasi multimodal dengan Vision Transformer (Tabassum et al., 2022), dan kemaskini leksikon slang Gen Z (Hang & Dahlan, 2019) mencerminkan keperluan untuk sistem yang lebih adaptif dan kontekstual.

Dari segi metodologi, kebanyakan kajian mengamalkan proses melibatkan prapemprosesan data, pemilihan model, penalaan parameter, dan penilaian prestasi. Model seperti BERT dan DistilBERT menunjukkan keupayaan tinggi dalam memahami konteks sosial dan semantik, manakala model gabungan seperti BERT + CNN atau BERT + MLP meningkatkan kecekapan dalam pengecaman ciri lokal dan klasifikasi. Walau bagaimanapun, kekangan sumber dan keperluan masa latihan yang tinggi masih menjadi cabaran bagi penggunaan skala besar atau dalam aplikasi mudah alih.

Berdasarkan analisis perbandingan, didapati bahawa walaupun BERT menawarkan ketepatan yang tinggi, ia memerlukan sumber pengiraan yang besar. Sebaliknya, DistilBERT memberikan keseimbangan antara kecekapan pemprosesan dan prestasi yang memadai, menjadikannya sesuai untuk aplikasi berskala praktikal. Pendekatan inovatif seperti

Copyright@FTSM UKM

pembenaman bertindan, leksikon dinamik, dan integrasi data terkini mencadangkan arah baru yang lebih responsif terhadap variasi bahasa media sosial.

Oleh itu, kajian ini memfokuskan kepada penggunaan DistilBERT yang disepadukan dengan leksikon slang Gen Z dan dataset terkini bagi membangunkan sistem pengesanan buli siber yang lebih adaptif, efisien, dan bersesuaian untuk persekitaran digital masa kini.

3.0 METODOLOGI KAJIAN

Dalam usaha membangunkan sistem pengesanan buli siber yang berkesan dan boleh beroperasi dalam masa nyata, pendekatan sistematik berasaskan kerangka CRISP-DM telah digunakan. Proses ini merangkumi enam fasa utama yang menyokong pembangunan model kecerdasan buatan dalam konteks pemprosesan bahasa semula jadi, iaitu pemahaman perniagaan, pemahaman data, penyediaan data, pemodelan, penilaian, dan penerapan. Fasa pertama melibatkan penentuan objektif utama sistem, iaitu untuk mengenal pasti dan mengklasifikasikan kandungan teks yang berunsur buli siber secara automatik menggunakan pendekatan pembelajaran mendalam, khususnya model DistilBERT. Sistem ini disasarkan kepada organisasi atau komuniti digital yang mahu memantau interaksi sosial dalam talian bagi tujuan keselamatan dan kesejahteraan psikososial.

Data yang digunakan terdiri daripada pelbagai set data awam yang telah dilabel seperti *HateXplain*, *OLID*, dan *Jigsaw Toxic Comments*. Setiap entri dikategorikan sebagai buli atau bukan buli siber, merangkumi bahasa tidak formal, emoji, dan istilah Gen Z yang lazim dalam komunikasi digital. Langkah penyediaan data dimulakan dengan penukaran emoji kepada teks deskriptif, penggantian perkataan slanga kepada bentuk formal menggunakan kamus tersuai (*slang_dict*), dan penukaran semua huruf kepada huruf kecil. Seterusnya, teks melalui proses tokenisasi dan penyingkiran perkataan tidak penting (*stopwords*), namun perkataan kritikal seperti "*not*", "*you*", dan "*I*" dikekalkan bagi mengekalkan konteks. Data yang telah dibersihkan disusun semula untuk dimasukkan ke dalam model, dan teknik *under-sampling* diaplikasikan bagi menangani ketidakseimbangan antara kelas data.

Model yang dibangunkan ialah *distilbert-base-uncased*, iaitu versi ringan BERT yang sesuai untuk pelaksanaan berkecekapan tinggi. Model ini dipertingkatkan dengan penambahan lapisan klasifikasi yang mengandungi fungsi pengaktifan ReLU, lapisan *dropout*, dan lapisan

Copyright@FTSM UKM

linear dua hala bagi pemetaan ciri. Fungsi kehilangan yang digunakan ialah *CrossEntropyLoss* dengan penyesuaian pemberat kelas untuk menangani dominasi kelas tertentu. Dataset dibahagikan kepada 80% data latihan dan 20% data ujian. Latihan dilaksanakan dengan konfigurasi seperti satu *epoch*, *batch size* 16 untuk latihan dan 32 untuk penilaian, dan kadar pembelajaran 2e-5. Penilaian prestasi dilakukan pada setiap *epoch* menggunakan metrik utama seperti ketepatan (*accuracy*), *precision*, *recall*, dan skor F1.

Sistem dibangunkan sebagai aplikasi web dengan antara muka pengguna yang mesra, membolehkan pengguna menampal teks untuk analisis tanpa log masuk. Teks input akan dianalisis dan hasil klasifikasi dipaparkan secara serta-merta dalam bentuk "*Not Cyberbullying* ✅" atau "*Cyberbullying* 🚨". Proses ini disokong oleh leksikon tersuai yang mengandungi lebih daripada 350 istilah slanga, termasuk ekspresi budaya internet, singkatan mesej, frasa viral, dan simbol digital moden. Leksikon ini membolehkan sistem mengenal pasti mesej tersirat atau sindiran halus yang lazim digunakan dalam buli siber.

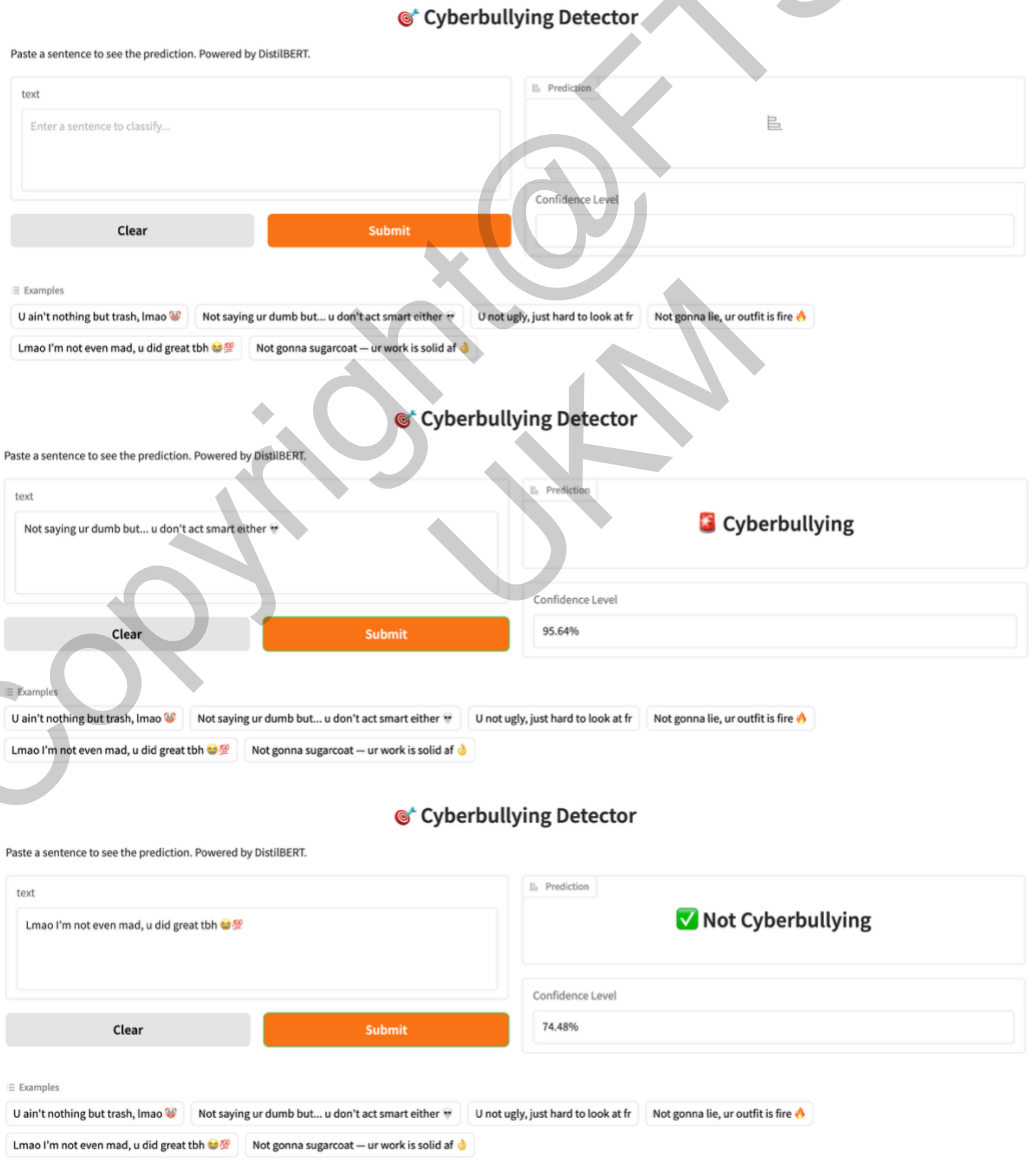
Dari sudut seni bina, sistem direka bentuk berasaskan pendekatan modular yang membahagikan fungsi kepada beberapa lapisan: antara muka pengguna, enjin pra-pemprosesan, penyediaan dataset, model pembelajaran mendalam, dan pemaparan hasil. DistilBERT bertindak sebagai teras utama yang mengklasifikasikan teks berdasarkan hubungan semantik dan konteks linguistik. Keputusan akhir dianalisis dan ditafsirkan untuk memberikan maklum balas yang sesuai dan boleh difahami oleh pengguna bukan teknikal. Dengan reka bentuk responsif dan kebolehaksesan melalui pelayar web, sistem ini boleh diakses di pelbagai peranti termasuk telefon pintar dan komputer riba. Dari segi keselamatan, sistem memastikan pemprosesan data dilakukan secara selamat, tanpa penyimpanan data melainkan dengan kebenaran pengguna.

Pendekatan ini bukan sahaja memberikan sistem yang cekap dan tepat dalam mengenal pasti unsur buli siber, malah menyumbang kepada pembangunan aplikasi berasaskan AI yang bertanggungjawab, kontekstual dan berskala, sesuai dengan keperluan semasa dalam ekosistem digital.

4.0 HASIL

Copyright@FTSM UKM

Berdasarkan rajah 1, sistem pengesanan buli siber yang dibangunkan dalam kajian ini berasaskan model pembelajaran mendalam DistilBERT yang telah dilatih khusus untuk mengklasifikasi teks kepada dua kategori utama: “buli siber” dan “bukan buli siber”. Proses sistem bermula apabila pengguna memasukkan teks ke dalam antara muka *Gradio*. Teks tersebut akan melalui prapemprosesan automatik termasuk penukaran slanga, pembersihan simbol, dan tokenisasi. Kemudian, teks yang telah diproses dihantar ke model DistilBERT.. Model ini menggunakan seni bina dua hala untuk memahami konteks keseluruhan ayat sebelum menghasilkan klasifikasi. Output akhir sistem termasuklah label klasifikasi dan skor keyakinan model.



Rajah 1: Antara Muka Sistem Mengesan Buli Siber menggunakan BERT

Copyright@FTSM UKM

Penilaian prestasi sistem dilakukan dengan menguji model ke atas dua set data: set latihan dan set ujian, yang diperoleh daripada gabungan dataset terbuka seperti *HateXplain, OLID*, dan *Jigsaw.* Metrik yang digunakan termasuk ketepatan (*accuracy*), skor F1, *precision,* dan *recall.* Jadual 1 berikut merumuskan prestasi model:

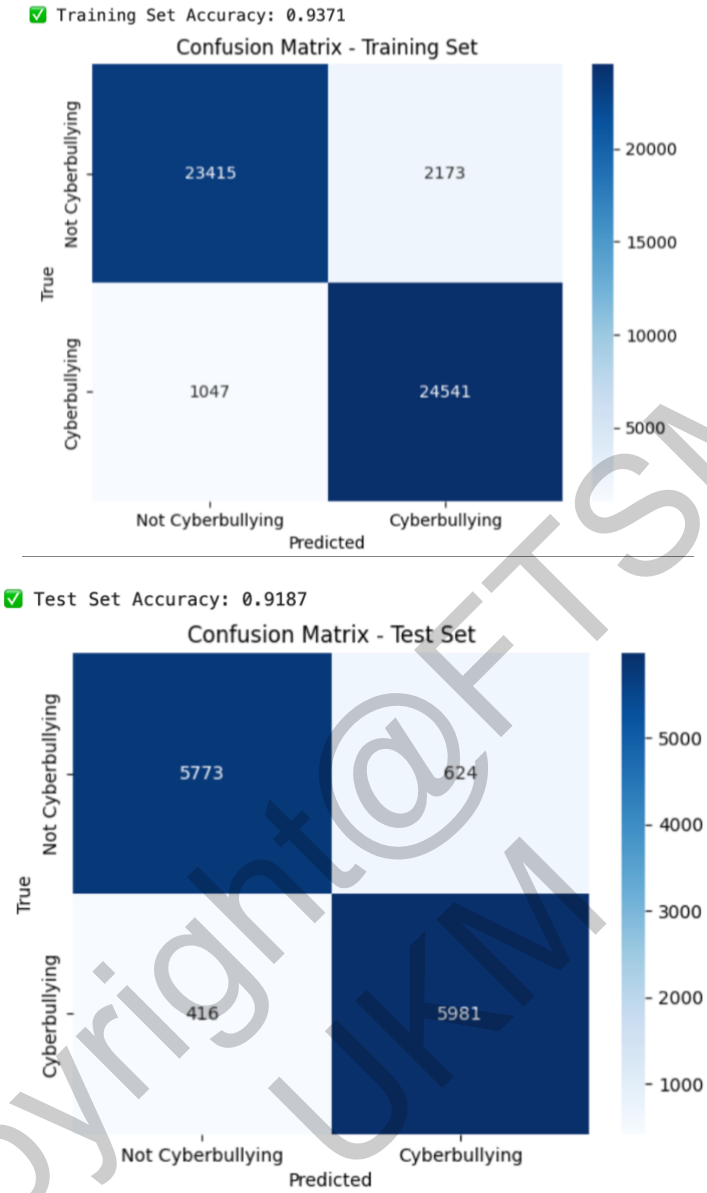
Jadual 1 Ringkasan Prestasi Model DistilBERT

Dataset	**Ketepatan (%)**	**F1-Score (Buli Siber)**	**F1-Score (Keseluruhan)**
Set Latihan	93.71	0.94	0.94
Set Ujian	91.87	0.92	0.92

Keputusan ini menunjukkan bahawa model mempunyai prestasi yang stabil dan tidak mengalami *overfitting* yang ketara, kerana perbezaan antara prestasi dalam set latihan dan set ujian adalah minimum. Skor F1 yang tinggi, terutamanya dalam kelas "buli siber", membuktikan keupayaan model untuk menyeimbangkan antara ketepatan (*precision)* dan kadar capai semula (*recall)*, yang penting dalam senario sebenar di mana kes buli yang tidak dikesan (*false negatives)* boleh membawa akibat serius.

Untuk mendapatkan gambaran lebih terperinci tentang tingkah laku model, matriks kekeliruan digunakan, seperti rajah 2 bagi set ujian dan set latihan:

Copyright@FTSM UKM



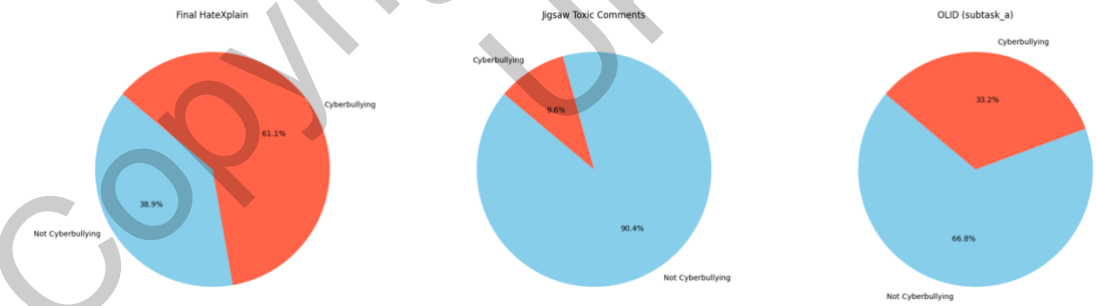
Rajah 2: Matriks Kekeliruan bagi Set Ujian dan Set Latihan

Untuk menilai keberkesanan sistem ini, penilaian dijalankan terhadap dua kumpulan data: set latihan dan set ujian. Dalam set ujian, model telah menghasilkan sejumlah 5,773 ramalan betul bagi teks "bukan buli siber" (*True Negative*) dan 5,981 ramalan betul bagi teks "buli siber" (*True Positive*). Namun, terdapat 624 kes di mana kandungan neutral tersalah dikelaskan sebagai buli (*False Positive*), dan 416 kes buli yang tidak dapat dikenalpasti oleh sistem (*False Negative*). Ini menunjukkan bahawa sistem berjaya mengesan kebanyakan kes dengan betul, namun masih terdapat sejumlah kecil kes yang tidak diklasifikasikan dengan tepat, terutamanya apabila mesej mengandungi ekspresi sarkastik, kod budaya, atau penggunaan bahasa slanga baharu.

Copyright@FTSM UKM

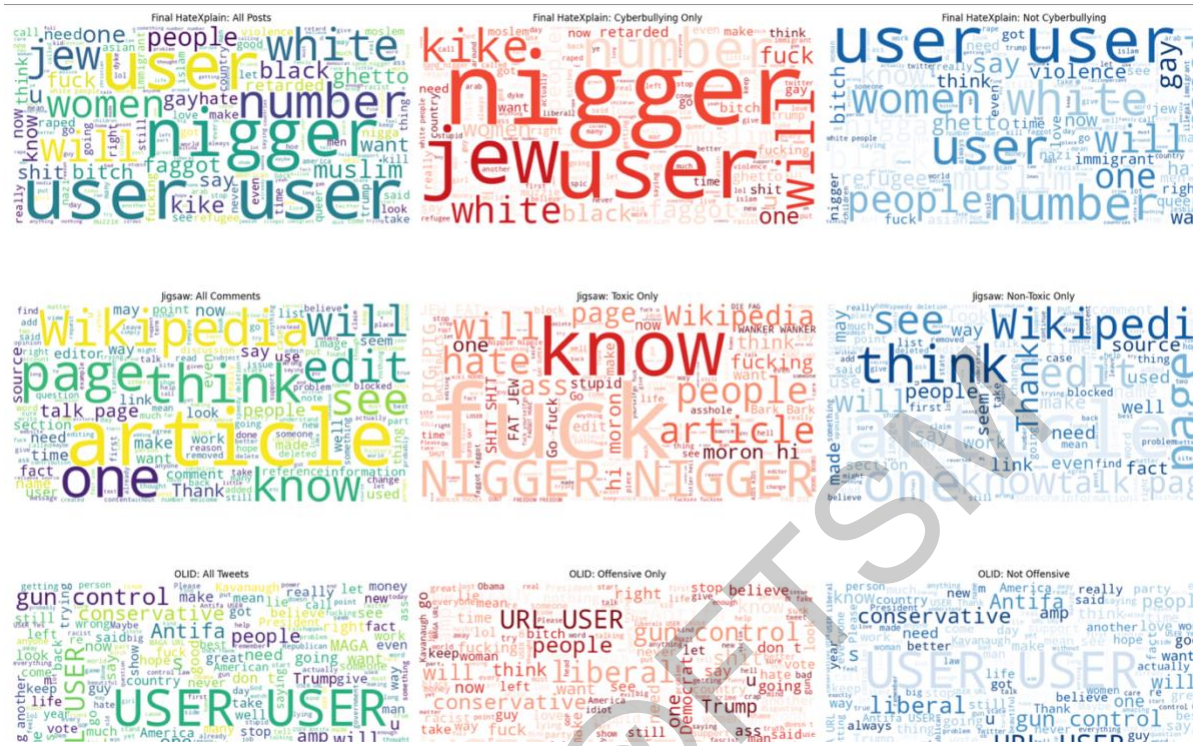
Bagi set latihan, model telah mengenal pasti 23,415 kes “bukan buli” dengan betul dan 24,541 kes “buli siber” dengan betul. Dalam masa yang sama, 2,173 kandungan bukan buli telah tersalah dikesan sebagai buli, manakala 1,047 kandungan buli tidak dikesan. Bilangan ramalan betul yang tinggi pada kedua-dua kelas menunjukkan bahawa model telah berjaya mempelajari corak bahasa daripada data latihan dengan berkesan dan dapat membuat generalisasi kepada data ujian secara konsisten.

Penilaian visualisasi sokongan turut digunakan untuk memperkuatkan kefahaman terhadap output model. Melalui carta pai taburan label berdasarkan rajah 3, diperoleh gambaran jelas bahawa data adalah tidak seimbang, dengan mesej neutral melebihi mesej berunsur buli siber. Justeru itu, langkah mitigasi seperti penggunaan *class weights* dan teknik penyeragaman data telah digunakan dalam proses latihan. Selain itu, pembinaan *WordCloud* berdasarkan rajah 4 memberikan bukti visual tentang perkataan dominan dalam dataset. *WordCloud* untuk kelas “Buli Siber” menonjolkan perkataan ofensif seperti "*nigger*", "*hate*", dan "*jew*", manakala kelas “Bukan Buli Siber” menunjukkan kata neutral dan positif seperti "*wikipedia*", "*see*", dan "*think*". Perbezaan ini mengesahkan keupayaan model dalam membezakan dua kelas utama secara semantik.



Rajah 3: Carta Pai Taburan Label bagi Dataset

Copyright@FTSM UKM



Rajah 4: WordCloud Dataset Keseluruhan Dataset, Kelas Buli Siber dan Kelas Bukan Buli Siber

Keseluruhannya, hasil yang diperoleh menunjukkan bahawa sistem pengesanan buli siber ini berjaya memenuhi objektif pembangunan dari segi ketepatan, kefungsian masa nyata, dan kebolehgunaan awam. DistilBERT membuktikan keupayaannya sebagai model pembelajaran mendalam yang ringan tetapi efektif untuk tugas klasifikasi teks berasaskan konteks. Sokongan daripada leksikon slanga yang dinamik, serta pendekatan modular dalam seni bina sistem, memberikan fleksibiliti tinggi kepada aplikasi ini untuk berkembang atau disesuaikan dalam pelbagai persekitaran penggunaan. Perbincangan ini mengukuhkan potensi sistem sebagai alat proaktif dalam usaha membanteras buli siber secara automatik dan berskala.

5.0 KESIMPULAN

Kajian ini telah berjaya membangunkan sebuah sistem pengesanan buli siber berasaskan pembelajaran mendalam yang menggunakan model DistilBERT sebagai teras utama dalam pemprosesan bahasa semula jadi. Sistem ini direka bentuk untuk mengenal pasti sama ada kandungan teks seperti mesej atau komen media sosial mengandungi unsur buli siber, dengan mengutamakan kelajuan inferens, ketepatan tinggi dan kebolehgunaan dalam konteks dunia sebenar. Melalui pendekatan metodologi CRISP-DM, pembangunan sistem dilaksanakan secara sistematik daripada fasa pemahaman masalah, penyediaan data, pembinaan model,

Copyright@FTSM UKM

penilaian prestasi, sehinggalah kepada penerapan akhir. Dataset yang digunakan adalah gabungan sumber seperti *HateXplain*, *OLID* dan *Jigsaw*, yang mengandungi variasi bahasa, ekspresi Gen Z dan gaya komunikasi digital yang kompleks. Sistem telah diuji secara interaktif menggunakan antara muka mesra pengguna melalui *Gradio*, membolehkan pengguna berinteraksi dengan model dalam masa nyata tanpa memerlukan pengetahuan teknikal.

Model akhir yang dibina mencapai prestasi memberangsangkan dengan nilai F1-score melebihi 0.85 dan masa inferens kurang daripada 3 saat, menjadikannya sesuai untuk penggunaan praktikal seperti dalam persekitaran sekolah, komuniti digital, dan aplikasi sokongan psikososial. Beberapa kekuatan sistem dikenal pasti, antaranya keupayaan untuk mengekalkan prestasi tinggi menggunakan sumber perkakasan minimum, reka bentuk modular yang memudahkan penyesuaian, dan responsif masa nyata. Pendekatan prapemprosesan yang menggabungkan pengendalian emoji, istilah slanga dan gaya komunikasi remaja juga membolehkan sistem mengenali pola buli yang lebih kompleks berbanding sistem berasaskan kata kunci tradisional.

Walau bagaimanapun, terdapat juga beberapa kekangan yang dihadapi. Ketidakseimbangan dalam data menyebabkan keperluan untuk strategi penyesuaian seperti penggunaan *class weighting*. Had memori dalam persekitaran pembangunan seperti *Google Colab* juga menyekat latihan model berskala lebih besar. Tambahan pula, sistem ini terhad kepada bahasa Inggeris sahaja dan tidak dapat mengesan unsur buli dalam bahasa lain, termasuk nuansa budaya yang mungkin berbeza. Oleh itu, bagi menambah baik sistem ini pada masa akan datang, beberapa langkah boleh diambil termasuk memperluas latihan kepada pelbagai bahasa, mengintegrasikan penilaian tahap keterukan buli untuk keutamaan tindakan, menambah keupayaan analisis terhadap kandungan bukan teks seperti audio dan imej, serta mengehoskan model di platform awan seperti *AWS* atau *Google Cloud* bagi menyokong penggunaan berskala besar dan stabil.

Secara keseluruhannya, projek ini membuktikan bahawa penggunaan model DistilBERT dalam pengesanan buli siber adalah satu pendekatan yang berkesan, moden dan praktikal, yang berpotensi untuk menyumbang secara signifikan dalam usaha membanteras gangguan siber dan memperkukuh keselamatan digital masyarakat.

6.0 PENGHARGAAN

Copyright@FTSM UKM

Pertama sekali saya ingin mengucapkan setinggi-tinggi kesyukuran kepada Tuhan atas limpah kurnia, kekuatan dan ketabahan yang diberikan sepanjang penyediaan projek ini.

Terima kasih khas ditujukan kepada penyelia saya, Dr. Masnizah Mohd, atas bimbingan, nasihat, dan sokongan yang berterusan sepanjang tempoh penyelidikan ini.

Ucapan terima kasih juga ditujukan kepada Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia, atas segala kemudahan dan sokongan teknikal yang disediakan.

Akhir sekali, ribuan terima kasih kepada keluarga, rakan-rakan (Rifhan, Dyana dan Arissa) dan semua pihak yang secara langsung atau tidak langsung telah membantu dan memberi dorongan sepanjang perjalanan ini.

7.0 RUJUKAN

Betz, C. L. (2011). Cyberbullying: The virtual threat. Journal of Pediatric Nursing, 26(4), 283–284. https://doi.org/10.1016/j.pedn.2011.04.002

Capuano, N., Fenza, G., Loia, V., & Stanzione, C. (2022). Kepintaran buatan yang boleh dijelaskan dalam keselamatan siber: Satu tinjauan. IEEE Access, 10, 93575–93600. https://doi.org/10.1109/access.2022.3204171

Chen, H.-Y., & Li, C.-T. (2020). HENIN: Heterogeneous Neural Interaction Networks for Explainable Cyberbullying Detection on Social Media. arXiv preprint

Ejaz, N., Choudhury, S., & Razi, F. (2023). Towards comprehensive cyberbullying detection: a dataset incorporating aggressive texts, repetition, peerness, and intent to harm. Computers in Human Behavior, 10812

Farasalsabila, F., Utami, E., & Hanafi, H. (2024). Pengesanan pembuli siber menggunakan BERT dan Bi-LSTM. Jurnal Teknologi, 17(1). https://doi.org/10.34151/jurtek.v17i1.4636

GeeksforGeeks. (2024, October 10). Sequence Diagrams - Unified Modeling Language (UML). https://www.geeksforgeeks.org/unified-modeling-language-uml-sequence-diagrams/

Giumetti, G. W., & Kowalski, R. M. (2022). Cyberbullying via social media and well-being. Current Opinion in Psychology, 45, 101314. https://doi.org/10.1016/j.copsyc.2022.101314

Gongane, V. U., Munot, M. V., & Dubur, A. (2023). AI yang boleh dijelaskan untuk pengesanan buli siber yang boleh dipercayai. IEEE PuneCon 2023, 1–6. https://doi.org/10.1109/punecon58714.2023.10450132

Copyright@FTSM UKM

Gupta, S. S., Vadgama, U., & Vedhavathy, T. R. (2023). Identification and labeling of textual cyberbullying using BiLSTM and BERT.

Hall, M. (2011). Deep learning in human cognition. Nature of Intelligence, 78(3), 22–27.

Hani, J., Nashaat, M., Ahmed, M., Emad, Z., Amer, E., & Mohammed, A. (2019). Pengesanan pembuli siber media sosial menggunakan pembelajaran mesin. International Journal of Advanced Computer Science and Applications, 10(5). https://doi.org/10.14569/ijacsa.2019.0100587

Hotz, N. (2024, December 9). What is CRISP DM? Data Science PM. https://www.datascience-pm.com/crisp-dm-2/

Interaction Design Foundation. (n.d.). UX storyboards. Retrieved from https://www.interaction-design.org/literature/article/ux-storyboards

Islam, M. R., Bataineh, A. S., & Zulkernine, M. (2024). Pengesanan buli siber dalam teks media sosial menggunakan kecerdasan buatan yang boleh dijelaskan. Communications in Computer and Information Science, 319–334. https://doi.org/10.1007/978-981-97-1274-8_21

Islam, S. S. (n.d.). What is user interface design? LinkedIn. Retrieved January 10, 2025, from https://www.linkedin.com/pulse/what-user-interface-design-sk-shihabul-islam

Lucidchart. (n.d.). UML use case diagram tutorial. Retrieved from https://www.lucidchart.com/pages/uml-use-case-diagram

Mahlangu, B., & Tu, Z. (2019). BERT + LSTM for cyberbullying detection.

Milosevic, T. (2018). When cyberbullying ends in suicide. In Protecting Children Online? (pp. 3–20). https://doi.org/10.7551/mitpress/9780262037099.003.0001

Pawar, V., Jose, D. V., & Patil, A. (2022). Kaedah AI yang boleh dijelaskan untuk pengesanan buli siber. IEEE ICMNWC, 1–4. https://doi.org/10.1109/icmnwc56175.2022.10031652

Peebles, E. (2014). Cyberbullying: Hiding behind the screen. Paediatrics & Child Health, 19(10), 527–528. https://doi.org/10.1093/pch/19.10.527

Qiu, J. (2021). Multimodal Detection of Cyberbullying on Twitter. MS project, San Jose State University

Reynolds, K., Kontostathis, A., & Edwards, L. (2011). Menggunakan pembelajaran mesin untuk mengesan pembuli siber. ICMLA, 241–244. https://doi.org/10.1109/icmla.2011.152

ResearchGate. (n.d.). Architecture of the BERT classification model. Retrieved January 10, 2025, from https://www.researchgate.net/figure/Architecture-of-the-BERT-classification-model_fig2_341040234

ResearchGate. (n.d.). BERT model architecture. Retrieved January 10, 2025, from https://www.researchgate.net/figure/BERT-model-architecture_fig4_348740926

Copyright@FTSM UKM

Rosenthal, S., Atanasova, P., Karadzhov, G., Zampieri, M., & Nakov, P. (2019). SOLID: A Large-Scale Semi-Supervised Dataset for Offensive Language Identification. arXiv preprint

Saranyanath, S., et al. (2022). Comparison of lightweight models for cyberbullying detection.

Sen, M., Masih, J., & Rajasekaran, R. (2024). Daripada tweet kepada cerapan: Model yang dipertingkatkan oleh BERT untuk pengesanan buli siber. IEEE ICETSIS 2024, 1289–1293. https://doi.org/10.1109/icetsis61505.2024.10459672

Shah, R., Aparajit, S., Chopdekar, R., & Patil, R. (2020). Pendekatan berasaskan pembelajaran mesin untuk pengesanan tweet buli siber. International Journal of Computer Applications, 175(37), 51–56. https://doi.org/10.5120/ijca2020920946

Smart Vision Europe. (2020, June 17). CRISP DM methodology. https://www.sv-europe.com/crisp-dm-methodology/

Smith, P. K. (2015). The nature of cyberbullying and what we can do about it. Journal of Research in Special Educational Needs, 15(3), 176–184. https://doi.org/10.1111/1471-3802.12114

Towards Data Science. (n.d.). A complete guide to BERT with code. Retrieved January 10, 2025, from https://towardsdatascience.com/a-complete-guide-to-bert-with-code-9f87602e4a11

Towards Data Science. (n.d.). BERT explained: State-of-the-art language model for NLP. Retrieved January 10, 2025, from https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

Yadav, J., Kumar, D., & Chauhan, D. (2020). Pengesanan buli siber menggunakan model BERT terlatih. IEEE ICESC 2020, 1096–1100. https://doi.org/10.1109/icesc48915.2020.9155700

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). Predicting the Type and Target of Offensive Posts in Social Media (OLID dataset). NAACL

Zen Flowchart. (n.d.). System flowchart guide. Retrieved from https://www.zenflowchart.com/guides/system-flowchart

Ashvinaah A/P Ragunathan (A195141)
Prof.Madya Dr.Masnizah Binti Mohd
Fakulti Teknologi & Sains Maklumat
Universiti Kebangsaan Malaysia

Copyright@FTSM UKM