PENGESANAN DAN KLASIFIKASI UNGKAPAN KEBENCIAN MENGGUNAKAN ALGORITMA PENGOPTIMUMAN BURUNG CAMAR (*SEAGULL*) YANG DIPERTINGKATKAN

Samuel a/l Ravi, Wandeep Kaur A/P Ratan Singh

Faculty of Information Science & Technology
Universiti Kebangsaan Malaysia
43600 Bangi, Selangor

Abstract

The increase of hate speech on social media platforms has become a serious issue with consequences for both individuals and society at large. It leads to violence, discrimination, and psychological harmful especially to the vulnerable users. The constant presence of hate speech diminishes social cohesion, leading to long-lasting negative impacts such as isolation, mental health struggles, and, in some cases, contributing to suicide cases. Therefore, developing an automatic system to detect and isolate hate speech is essential to create a safer online space. The intention is to develop a robust and accurate real-time detection system in hope that it can effectively identify and reduce the spread of hate speech effectively in real-time, minimizing its harmful societal impact. To accomplish this, a bidirectional long short-term memory (Bi-LSTM) classification model is used to classify hate speech into 2 categories namely hate speech and non-hate speech. FastText feature extraction allows for classification of misspelled and disguised hate speech as well, closing all loopholes for social medias to post hate speech. Parameter optimization is applied to optimize the model's performance, surpassing traditional machine learning models in terms of accuracy and other evaluation metrics. The expected outcome is a highly accurate and efficient hate speech detection system capable of significantly reducing the spread of harmful content on social media platforms.

Abstrak

Peningkatan ungkapan kebencian di platform media sosial telah menjadi isu serius dengan kesan buruk terhadap individu dan masyarakat secara keseluruhan. Ia membawa kepada keganasan, diskriminasi, dan kesan psikologi yang berbahaya terutamanya kepada pengguna yang terdedah. Kehadiran berterusan ungkapan kebencian merosakkan keharmonian sosial, mengakibatkan kesan negatif jangka panjang seperti pengasingan, masalah kesihatan mental, dan, dalam beberapa kes, menyumbang kepada kejadian bunuh diri. Oleh itu, pembangunan sistem automatik untuk mengesan dan mengasingkan ungkapan kebencian adalah penting bagi mewujudkan ruang dalam talian yang lebih selamat. Pembangunan sistem pengesanan secara langsung ini diharapkan dapat mengenal pasti dan mengurangkan penyebaraan ungkapan kebencian dengan tepat dan berkesan secara masa-nyata, sekali gus meminimumkan kesan buruknya kepada masyarakat. Bagi mencapai matlamat ini, sebuah model klasifikasi *bidirectional long short-term memory* (Bi-LSTM) digunakan untuk mengklasifikasikan ungkapan kebencian kepada dua kategori, iaitu ungakapan kebencian dan ungkapan bukan berunsur kebencian. Pengekstrakan ciri FastText turut digunakan untuk mengklasifikasikan ungkapan kebencian yang mempunyai ejaan salah dan yang disamarkan, menutup semua ruang untuk pengguna media sosial memuat naik ungkapan kebencian. Pelarasan parameter digunakan untuk mengoptimumkan prestasi model dan ianya mengatasi model pembelajaran mesin tradisional dari segi ketepatan dan metrik penilaian lain. Hasil yang dijangka ialah sistem pengesanan ungkapan kebencian yang sangat tepat dan cekap, yang mampu mengurangkan penyebaran kandungan berbahaya di platform media sosial secara signifikan.

1.0 PENGENALAN

Ungkapan kebencian menimbulkan cabaran sosial dan psikologi yang serius, kerana ia menyasarkan kumpulan tertentu berdasarkan sifat seperti bangsa, jantina, agama, etnik dan kepercayaan politik. Bentuk pencerobohan ini memupuk diskriminasi dan menimbulkan permusuhan, yang memperdalamkan lagi perpecahan masyarakat, mengukuhkan stereotaip yang merosakkan, dan mewujudkan persekitaran yang bermusuhan di mana kumpulan terpinggir sering merasa terancam dan tidak selamat (MacAvaney et al., 2019). Kajian mendedahkan bahawa pendedahan berterusan kepada ungkapan benci boleh membawa kepada pengasingan dalam kehidupan sosial dan, dalam kes yang teruk, menghasut keganasan fizikal dan diskriminasi terhadap kumpulan sasaran (Bilewicz & Soral, 2020). Pemulauan yang didorong oleh kebencian sedemikian mempunyai kesan yang berkekalan terhadap kesihatan mental mangsa, yang mungkin mengalami tahap kebimbangan yang tinggi, kemurungan, dan rasa tidak selamat yang berleluasa ketika berada dalam talian dan luar talian (Paz, 2020). Percambahan media sosial telah

meningkatkan penyebaran ungkapan kebencian. Platform seperti Twitter, Facebook dan Instagram memudahkan penyebaran meluas bahasa kasar kepada khalayak pengguna yang luas, selalunya dengan pengawasan yang terhad disebabkan tidak diawasi dan peraturan yang tidak ketat yang wujud dalam ruang digital ini. Persekitaran ini membolehkan individu menyatakan pandangan yang penuh kebencian tanpa akauntabiliti, menjadikannya satu cabaran untuk mengesan dan mengekang kandungan tersebut dengan berkesan (The Verge, 2025).

Banyak syarikat media sosial menghadapi kritikan kerana gagal melaksanakan langkah tegas untuk membendung ungkapan kebencian, yang membawa kepada permintaan yang semakin meningkat untuk sistem pengesanan dan pengekangan yang lebih baik (FRA, 2023). Selain kemudaratan individu, ungkapan benci mengancam perpaduan sosial. Ia boleh mencetuskan jenayah kebencian dan malah membawa kepada konflik masyarakat yang lebih luas dengan mempromosikan ideologi yang memecahbelahkan. Dalam kes yang melampau, ungkapan kebencian yang tidak dikawal telah dikaitkan dengan keganasan, diskriminasi yang mendalam, dan, dalam skala yang lebih besar, boleh menyumbang kepada konflik antara pelbagai kumpulan demografi malah antara negara (Pacheco, 2020).

Projek ini memberi tumpuan kepada reka bentuk pembangunan Model Pengesanan dan Pengelasan Pertuturan Kebencian Berasaskan Pemprosesan Bahasa Tabii dengan Pengoptimuman Burung Camar (*SEAGULL*) Yang Dipertingkatkan (ESGONLP). Model ini memberi tumpuan kepada pengesahan dan klasifikasi ungkapan benci di media sosial. Model tersebut akan melalui beberapa peringkat sebelum mengklasifikasikan ungkapan benci iaitu pra-pemprosesan data, pengekstrakan ciri menggunakan FastText dan model *bidirectional long short-term memory* (Bi-LSTM) yang mengklasifikasikan ungkapan benci (Yousef Asiri, 2022). Pengesahan model ESGONLP kemudiannya disahkan dan diperiksa di bawah pelbagai aspek untuk menentukan ketepatannya.

2.0 SOROTON SUSASTERA

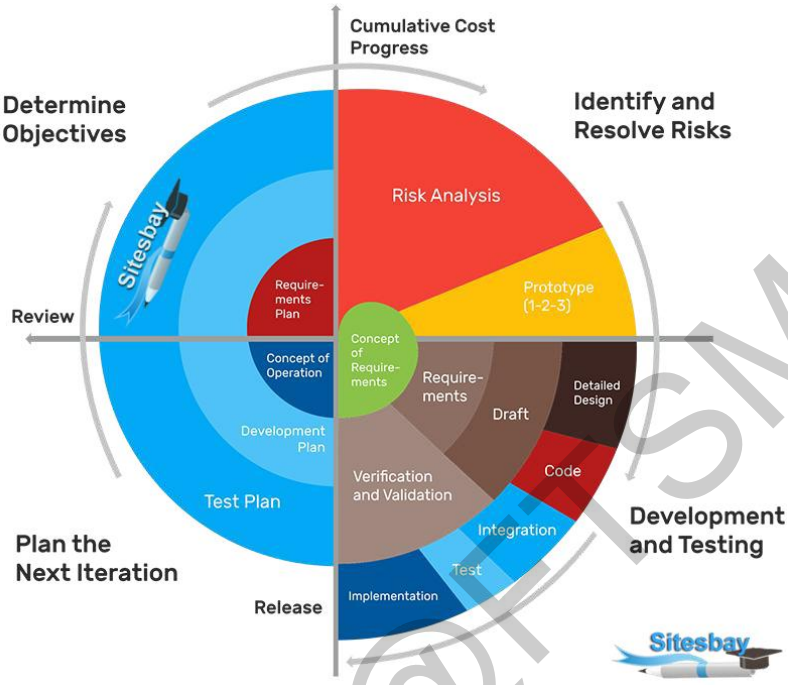
Ungkapan kebencian di media sosial terdiri daripada mesej yang menghina, mengancam atau mendiskriminasi orang berdasarkan ciri seperti bangsa, jantina, agama atau orientasi seksual. Apabila dibiarkan, kandungan jenis ini memberi kesan ketara kepada kesejahteraan mental pengguna, menyumbang kepada tekanan, kebimbangan dan perasaan tersisih yang meningkat. Selain itu, ungkapan benci juga menggalakkan perpecahan dalam masyarakat membawa kepada kecenderungan yang berbahaya, memupuk persekitaran dalam talian yang semakin tidak sihat dan tiada toleransi (Chitranjali Negi, 2024).

Walaupun platform media sosial direka untuk akses terbuka dan pengedaran kandungan pantas, malangnya platform media sosial ini telah menjadi arena di mana ungkapan benci berkembang pesat (Binny Mathew, 2019). Ramai pengguna mengeksploitasi algoritma platform ini dengan menggunakan bahasa tidak langsung, sengaja menggunakan ejaan salah atau sedikit pengubahsuaian frasa untuk mengelakkan pengesanan oleh sistem. Penyelidikan oleh Juan Carlos (Pereira Kohatsu, J.C, 2019) lebih mementingkan kajian dan pencadangan kaedah mengklasifikasikan teks sebagai ungkapan benci atau tidak tetapi tidak mengendalikan isu perubahan ungkapan kebencian mengikut peredaran masa. Kekurangan pendekatan sedemikian untuk memerangi pelbagai jenis format ungkapan kebencian membawa kepada ruang kebebasan secara sederhana, membolehkan bahasa berbahaya berterusan dan merebak.

Kewujudan ungkapan benci yang semakin meningkat menyerlahkan kepentingan sistem pengesanan ini untuk memastikan ruang dalam talian yang lebih selamat. Oleh itu, model pembelajaran mesin yang boleh menyesuaikan diri dengan keadaan yang berubah-ubah ini diperlukan untuk memerangi isu tersebut. Model pembelajaran mesin secara amnya mempunyai potensi besar untuk menangani isu ini dengan mengenali corak bahasa dan isyarat kontekstual, walaupun ketika ungkapan kebencian diungkapkan dalam bentuk penyamaran atau berkod. Penggunaan pengekstrakan ciri lanjutan seperti *FastText* dan *Bidirectional Encoder Representation* daripada *Transformers*, BERT telah terbukti lebih berkesan dalam mengesan pengesanan ungkapan benci (Sai Saketh Aluru, 2020). Dengan pelaburan berterusan dalam teknologi ini, platform media sosial akan lebih kondusif dengan kurangya kesan negatif daripada kandungan ungkapan kebencian yang memecahbelahkan dan seterusnya menyokong kesejahteraan pengguna.

3.0 METODOLOGI KAJIAN

Projek ini dibangunkan menggunakan model pembangunan lingkaran. Ini kerana model pembangunan lingkaran adalah proses berulang, bermakna ia melibatkan pelbagai peringkat seperti prapemprosesan data, pengekstrakan ciri, pengelasan dan pelarasan hiperparameter, membolehkan proses penghalusan dan penambahbaikan berterusan model. Ia juga diuji secara berulang untuk menganalisis prestasi sebelum memuktamadkan pendekatan yang terbaik.



Rajah 1.1 Proses pembangunan lingkaran

Sumber : (Sitesbay , 2025)

4.0 KEPUTUSAN

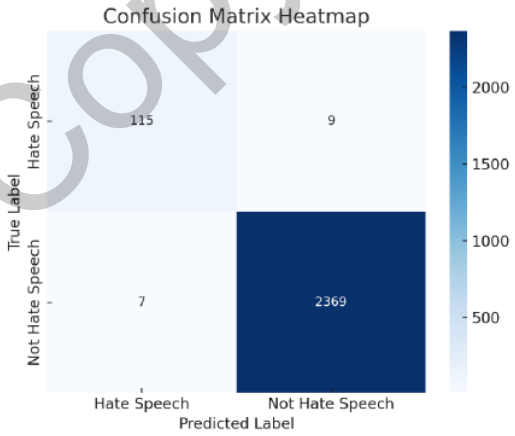
No. of Tweet	Accuracy	Precision	Recall	F-Score
500	99.24	99.26	98.96	98.97
1000	98.95	99.29	99.20	99.24
1500	98.97	99.07	99.26	99.33
2000	99.01	99.07	99.16	99.19
2500	99.17	99.20	98.98	99.02

Jadual 4.4 Keputusan *Baseline Model* – Stormfront

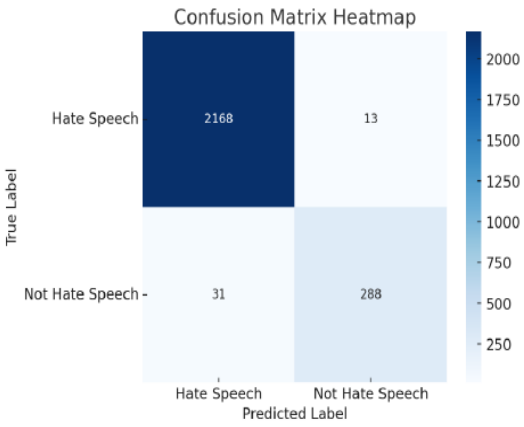
No. of Tweet	Accuracy	Precision	Recall	F-Score
500	99.22	99.22	99.23	99.27
1000	99.04	99.02	99.33	98.98
1500	99.08	99.27	99.04	99.04
2000	99.18	99.15	98.98	98.97
2500	99.12	99.00	99.00	99.06

Jadual 4.5 Keputusan *Baseline Model* – Crowdflower

Keputusan daripada *baseline model* dalam Jadual 4.2.1.4 dan 4.2.1.5 adalah berdasarkan kajian oleh Asiri, Y. (2020), dan satu model yang serupa serta parameter yang sama telah dibangunkan semula berdasarkan kajian tersebut dan keputusan yang diperoleh adalah hampir sama dengan jadual. Metrik penilaian yang digunakan ialah *accuracy, precision, recall* dan skor-F1. Pertama sekali, *baseline model* diuji menggunakan set data Stormfront dan nilai *accuracy, precision, recall* serta skor-F1 berada dalam lingkungan 98 hingga 99, menunjukkan bahawa model tersebut berfungsi dengan sangat baik atas kertas. Hal yang sama berlaku apabila model diuji menggunakan set data Crowdflower yang juga memperolehi metrik antara 98 hingga 99. *Accuracy* yang tinggi bermaksud kebanyakan *tweet* (ungkapan kebencian dan ungkapan bukan kebencian) diklasifikasikan dengan betul. *Precision* yang tinggi menunjukkan bahawa kebanyakan *tweet* yang diklasifikasikan sebagai ungkapan kebencian benar-benar merupakan ungkapan kebencian, bermakna model ini tidak memberikan pengesanan palsu secara berlebihan. *Recall* yang tinggi bermaksud model sangat baik dalam mengesan ungkapan kebencian, termasuk tweet yang bersifat *subtle* (halus) atau yang luar biasa. Akhir sekali, skor-F1 yang tinggi menunjukkan keseimbangan antara *precision* dan *recall*, iaitu model mampu menangkap kebanyakan ungkapan kebencian tanpa menandakan bahawa tweet tersebut adalah ungkapan bukan kebencian secara berlebihan. Model ini juga terbukti tidak mengalami masalah *overfitting* melalui graf plot *training loss* berbanding *validation loss*, dan *training accuracy* berbanding *validation accuracy*. Plot ini digunakan untuk mendiagnosis sejauh mana model pembelajaran mesin mempelajari dan menjana generalisasi berdasarkan prestasi pada data latihan dan data tidak kelihatan. Kedua-dua plot menunjukkan tiada tanda *overfitting*, menandakan model berfungsi baik pada data sebenar.



Rjah 4.10 Metrik Kekeliruan *Baseline Model* - Crowdflower



Rajah 4.11 Metrik Kekeliruan *Baseline Model*– Stormfront

Meskipun metrik yang tinggi diperoleh, keputusan ini tidak benar-benar mencerminkan keupayaan generalisasi *baseline model* kerana kedua-dua set data adalah sangat tidak seimbang, mengandungi lebih banyak ungkapan kebencian berbanding ungkapan bukan kebencian. Ini dapat dibuktikan melalui *confusion matrix* bagi set data Crowdflower untuk *baseline model*, yang menunjukkan bahawa model lebih cenderung mengesan ungkapan kebencian dengan *True Positive* yang tinggi dan *True Negative* yang rendah. Situasi sebaliknya berlaku untuk set data Stormfront, walaupun set-set ini juga mengandungi lebih banyak label kebencian berbanding ungkapan bukan kebencian. Semasa pengujian model dengan ayat-ayat yang tidak pernah dilihat (di luar set data), ayat mudah seperti "*hi*" atau "*how are you*" diklasifikasikan sebagai ungkapan kebencian oleh *baseline model*. Ini disebabkan oleh pembelajaran yang berat sebelah, di mana semasa latihan, ESGO mungkin telah terlalu mengoptimumkan untuk *recall* atau kelas khusus F1, menyebabkan model terlalu sesuai untuk label ungkapan kebencian. Walaupun jelas bahawa model tidak *overfit*, disebabkan ia mampu mengklasifikasikan hampir semua input sebagai kebencian dalam set data untuk diuji, *recall* dan skor-F1 masih akan kekal tinggi. Sebab lain bagi klasifikasi seperti itu adalah kekurangan pemahaman kontekstual di mana model Bi-LSTM dalam *baseline model* gagal menangkap corak yang menjadikan sesuatu ayat itu bersifat kebencian dan kerap kali mengaitkan perkataan seperti "*you*" dan "*stupid*" sebagai ungkapan kebencian walaupun tidak mengandungi perkataan kasar.

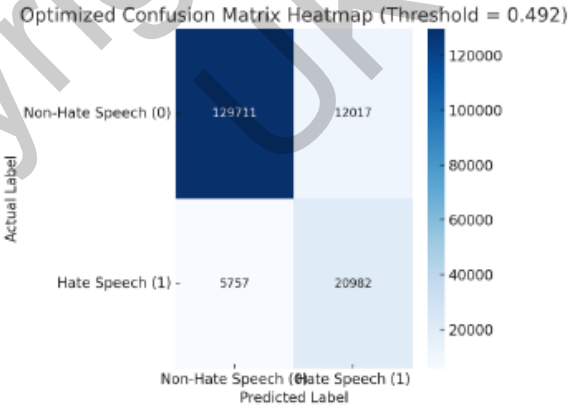


Figure 4.17 Metrik Kekeliruan Model Cadangan – HSOL

	Precision	*Recall*	Skor-F1	*Support*
0	0.96	0.92	0.94	141728
1	0.64	0.78	0.70	26739
accuracy			0.89	168467
macro avg	0.80	0.85	0.82	168467
Weighted avg	0.91	0.89	0.90	168467

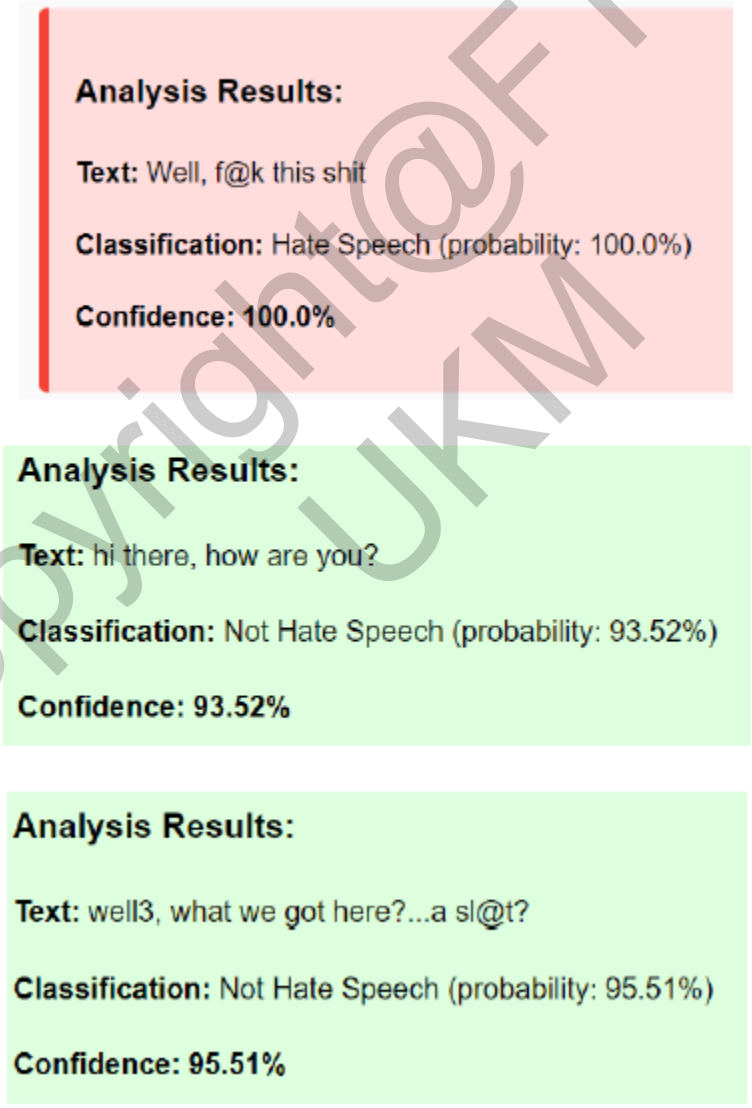
Jadual 4.19 Laporan Klasifikasi Model Cadangan - HSOL

Untuk model yang dicadangkan, ia diuji menggunakan set data yang sama iaitu HSOL oleh Devansh Mody. Berdasarkan laporan klasifikasi dan metrik yang diperoleh, jelas bahawa keputusan yang diperoleh jauh lebih baik daripada *baseline model* untuk set data yang sama. Model ini memperoleh *precision* sebanyak 0.96 untuk ungkapan bukan kebencian dan 0.64 untuk ungkapan kebencian, bermakna model lebih tepat dalam meramalkan bukan kebencian. Ini juga berlaku untuk *recall* bagi bukan kebencian, dengan skor 0.92, tetapi bagi ungkapan kebencian pula, ia mencatatkan 0.78, menunjukkan model berjaya menangkap 78% daripada semua contoh sebenar ungkapan kebencian. Ini penting kerana *recall* bagi ungkapan kebencian adalah sangat signifikan dalam pembangunan sebenar dan kegagalan mengesan ungkapan kebencian boleh menyebabkan kandungan berbahaya tidak dikesan. Seterusnya, skor-F1 yang diperoleh ialah 0.94 dan 0.70 masing-masing, menunjukkan prestasi yang kukuh dan seimbang untuk kedua-dua kelas. Tambahan lagi, purata F1 *macro* dan purata F1 *weighted* juga menunjukkan skor yang baik, masing-masing 0.82 dan 0.90, menunjukkan prestasi model terhadap kedua-dua kelas adalah baik walaupun terdapat ketidakseimbangan pada set data ujian.

Model yang dicadangkan mencatatkan skor AUC sebanyak 0.94, menunjukkan keupayaan pemisahan kelas yang sangat baik dan membuktikan bahawa model ini sangat berkeupayaan dalam membezakan antara ungkapan kebencian dan bukan kebencian merentas pelbagai *threshold* (ambang). Akhir sekali, untuk *specificity*, ia memperoleh skor 0.9152, mencerminkan bahawa model mengenal pasti dengan tepat 91% daripada sampel sebenar bukan kebencian. Ini bermaksud walaupun ia dioptimumkan untuk *recall* kelas kecil (ungkapan kebencian), ia masih mengekalkan *accuracy* yang tinggi untuk kelas majoriti. Model yang dicadangkan lebih baik dalam semua aspek, menandakan bahawa model ini lebih bagus dalam pembelajaran hasil daripada fungsi *gradient*

clipping yang menghalang *overfitting* dan meningkatkan keupayaan generalisasi. Semasa prapemprosesan bagi model yang dicadangkan, kamus perkataan kasar juga digunakan untuk membantu model mengesan perkataan yang membentuk ungkapan kebencian dengan lebih baik. Selain itu, populasi adaptif dalam ESGO membantu model daripada terperangkap dalam *local optima*, kerana *local optima* ialah titik yang kelihatan optimum dalam kawasan terhad tetapi bukan penyelesaian terbaik secara keseluruhan. Model yang dicadangkan juga berjaya mengesan ungkapan kebencian dan bukan kebencian yang tidak termasuk dalam set data dengan sangat baik berbanding *baseline model*.

Figure 4.20 Pengujian model cadangan (HSOL) menggunakan ayat rawak



Eksperimen ablasi ialah satu kajian untuk menganalisis sumbangan komponen yang berbeza dalam sesebuah model dengan membuang atau menambah komponen tersebut dan menilai hasil yang diperoleh. Eksperimen ini adalah kebiasaan dalam bidang pembelajaran mesin untuk menilai bagaimana komponen tertentu mempengaruhi prestasi keseluruhan model. Dalam model yang dicadangkan, terdapat tiga komponen selepas langkah prapemprosesan dan *tokenization*, iaitu komponen *FastText feature extraction*, *Attention Bi-LSTM*, dan komponen *ESGO*. Komponen-komponen ini ditambah ke *baseline model* dengan harapan untuk mencapai prestasi yang lebih baik dan optimum. Oleh itu, setiap komponen ini perlu dikaji untuk melihat sejauh mana ia mempengaruhi prestasi model dalam klasifikasi ungkapan kebencian.

Experimen	**Ciri Terlibat**
Ablasi Ciri Pengesktrakan	Model kendiri dengan hanya prapemprosesan, *tokenization* dan pengekstrakan ciri *FastText*
Ablasi *Attention Bi-LSTM*	Penambahan model pembelajaran mesin (*Attention Bi-LSTM*)
Ablasi ESGO	Penambahan pengoptimum (*ESGO*)

Jadual 4.21 Eksperimen Ablasi

	Precision	***Recall***	**Skor-F1**	***Support***
0	0.96	0.84	0.90	141728
1	0.50	0.82	0.62	26739
accuracy			0.84	168467
macro avg	0.73	0.83	0.76	168467
Weighted avg	0.89	0.84	0.85	168467

Jadual 4.22 Laporan Klasifikasi Ciri Pengesktrakan sebelum ablasi

	Precision	Recall	F1-Score	Support
0	0.92	0.96	0.94	141728
1	0.73	0.55	0.63	26739
accuracy			0.90	168467
macro avg	0.83	0.76	0.79	168467
Weighted avg	0.89	0.90	0.89	168467

Jadual 4.24 Laporan Klasifikasi Ablasi *Attention Bi-LSTM*

	Precision	Recall	F1-Score	Support
0	0.96	0.92	0.94	141728
1	0.64	0.78	0.70	26739
accuracy			0.89	168467
macro avg	0.80	0.85	0.82	168467
Weighted avg	0.91	0.89	0.90	168467

Jadual 4.26 Laporan Klasifikasi Ablasi ESGO - HSOL

Berdasarkan laporan klasifikasi dan metrik penilaian setiap komponen, kesan progresif ketiga-tiga komponen model dapat dianalisis dan dikaji. Metrik penilaian utama yang digunakan adalah *accuracy, precision, recall*, skor-F1, lengkung AUC, dan *specificity*. Untuk model yang dicadangkan tanpa sebarang ablasi model, iaitu hanya menggunakan *FastText embeddings* sebagai ciri pengekstrakan. Data tersebut dimasukkan ke *standard classifier* (pengelas piawai), dan konfigurasi asas ini telah mencapai prestasi keseluruhan yang baik dengan memperoleh *accuracy* sebanyak 84%, skor-F1 sebanyak 0.76 dan nilai lengkung AUC sebanyak 0.9118. Ini menunjukkan model mempunyai *recall* yang seimbang bagi kedua-dua kelas, dengan *non-hate recall* sebanyak 0.84 dan *hate recall* sebanyak 0.82. Namun, *precision* untuk kelas minoriti (ungkapan kebencian)

adalah sederhana iaitu 0.50. Keupayaan model untuk mengesan ungkapan kebencian dianggap memuaskan. Kekuatan diskriminatif model ini diperkukuh lagi dengan skor-F1 sebanyak 0.85 walaupun set data ujian agak tidak seimbang.

Dalam ablasi mode*l* pertama, lapisan *Attention-based Bidirectional LSTM* diimplementasikan selepas ciri pengekstrakan *FastText* untuk menangkap kebergantungan *sequential* (berurutan) dan hubungan kontekstual antara perkataan. Selepas ablasi, *accuracy* model meningkat kepada 90%, skor-F1 meningkat kepada 0.79 dan nilai lengkung AUC naik kepada 0.9253. Selain itu, terdapat juga peningkatan yang ketara dalam *precision* bagi ungkapan kebencian, dari 0.50 kepada 0.73, menunjukkan model lebih tepat dalam mengenal pasti ungkapan kebencian dengan kurangnya *false positives*. Namun begitu, *recall* bagi ungkapan kebencian menurun daripada 0.82 kepada 0.55. Ini mungkin disebabkan oleh model menjadi lebih konservatif dan terlebih tepat dalam membuat ramalan terhadap ungkapan kebencian sehingga menyebabkan lebih banyak *true hate speech* terlepas, dibuktikan dengan peningkatan *false negatives*. *Specificity* bagi bukan kebencian juga meningkat daripada 0.84 kepada 0.96, yang mengesahkan bahawa model menjadi lebih yakin dan tepat dalam mengenal pasti ungkapan bukan kebencian. Secara keseluruhannya, ablasi model pertama telah meningkatkan *precision* dan *accuracy* keseluruhan model, namun berdepan dengan penurunan *recall* bagi ungkapan kebencian, mengambil kira hakitkat set ujian mengandungi lebih banyak label bukan kebencian.

Dalam eksperimen ablasi model terakhir, model ditambah baik dengan algoritma *Enhanced Seagull Optimization (ESGO)*, yang membantu dalam penalaan hiperparameter dan mengoptimumkan *threshold* klasifikasi. Dalam keputusan ablasi ini, algoritma telah menetapkan nilai *threshold* 0.492 sebagai paling optimum. Penambahan algoritma ini memberikan pemulihan prestasi yang ketara bagi *recall* ungkapan kebencian, yang meningkat daripada 0.55 kepada 0.78, seterusnya menyelesaikan isu *recall* yang dihadapi dalam ablasi model pertama. Pada masa yang sama, *precision* bagi ungkapan kebencian kekal tinggi pada 0.64, yang menunjukkan model dapat mengekalkan keseimbangan antara *false positives* dan *false negatives*. F1-score untuk ungkapan kebencian turut meningkat kepada 0.70, menjadikannya skor terbaik sepanjang keseluruhan eksperimen ablasi. Selain itu, *macro* F1 meningkat kepada 0.82 dan *weighted* F1 mencapai 0.90 yang menunjukkan pencapaian keseimbangan dan prestasi keseluruhan terbaik dalam semua peringkat ablasi juga. Lengkung AUC juga meningkat kepada 0.9406, bermakna diskriminasi kelas menjadi lebih kuat selepas pengoptimuman. Walaupun *specificity* untuk bukan kebencian menurun sedikit kepada 0.91, ia masih dianggap tinggi dan memastikan peningkatan dalam pengesanan ungkapan kebencian tidak menjejaskan *accuracy* ungkapan bukan kebencian secara keterlaluan.

Walaupun model tanpa *Attention Bi-LSTM* dan ESGO menunjukkan prestasi yang baik, penambahan komponen tambahan memastikan model secara keseluruhan lebih seimbang dalam semua aspek dan menjana generalisasi yang lebih baik merentas pelbagai jenis *slang* dan salah ejaan yang digunakan dalam ungkapan kebencian. Dengan setiap komponen yang ditambah, model menjadi lebih yakin dalam mengklasifikasikan ungkapan kebencian dan ungkapan bukan kebencian sambil mengekalkan konsistensi. Kajian ablasi model ini berjaya menyerlahkan kepentingan setiap komponen dalam mengimbangi kelemahan model sebelum ablasi dan meningkatkan keupayaan mengesan ungkapan kebencian secara signifikan.

5.0 KESIMPULAN

Pembangunan dan penilaian sistematik model ESGONLP yang dicadangkan untuk pengesanan umgkapan kebencian telah berjaya mencapai semua objektif penyelidikan yang telah ditetapkan. Objektif pertama adalah untuk membandingkan prestasi pengekstrakan ciri *FastText* dengan model asas berasaskan *GloVe*. Tujuannya adalah untuk menilai sama ada model yang dicadangkan mampu mengatasi prestasi model GloVe dalam mengesan ungkapan kebencian, khususnya dalam pengendalian kesalahan ejaan.

Objektif ini berjaya dicapai melalui pelaksanaan *embedding FastText*, yang membantu memanfaatkan maklumat subkata untuk menangkap kesalahan ejaan dan *slang*, yang merupakan kelemahan kritikal pada *GloVe*. Objektif kedua adalah untuk menilai kesan eksperimen ablasi dengan dan tanpa pembelajaran mesin serta pengoptimuman. Objektif ini telah dicapai dengan menggabungkan secara progresif, lapisan *Attention Bi-LSTM* dan algoritma *Enhanced Seagull Optimization* (ESGO) ke dalam saluran pemprosesan. Lapisan Bi-LSTM membantu model menganalisis struktur ayat, manakala mekanisme *attention* membolehkannya memberi tumpuan kepada perkataan penting yang menunjukkan kebencian seperti penghinaan dan frasa agresif.

Terdapat beberapa kekangan yang dihadapi walaupun pencapaian model yang dicadangkan dalam mengesan ungkapan kebencian telah menunjukkan peningkatan dan memerlukan penambahbaikan pada masa akan datang. Salah satu cabaran utama adalah model tidak mampu memahami sepenuhnya konteks seperti *sarcasm* dan kebencian yang spesifik kepada budaya seperti perkataan "*fam*" yang bermaksud *family* tapi ditandakan sebagai ungkapan kebencian kerana berkaitan dengan sindiket. Walaupun model cadangan berjaya mengenal pasti ungkapan kebencian secara jelas, model tersebut menghadapi kesukaran apabila berdepan dengan bentuk kebencian yang tidak ketara seperti kenyataan sinis atau bahasa berkod. Sebagai contoh, frasa seperti "*Wow, you're really*

smart!" pada pandangan pertama mungkin dilihat sebagai pujian ikhlas, tetapi boleh juga merupakan penghinaan bersifat sinis bergantung kepada nada dan konteks. Oleh kerana model bergantung pada corak leksikal dan bukannya pemahaman semantik yang lebih mendalam, kekaburan ini boleh menyebabkan kesilapan klasifikasi. Kelemahan ini turut dihadapi oleh banyak kajian lain dalam pemprosesan bahasa semula jadi seperti yang dinyatakan oleh Gibert et al. (2018) dalam kajiannya mengenai forum *white supremacy*. Bagi menangani kelemahan ini, terdapat beberapa pendekatan yang menjanjikan untuk masa hadapan. Pendekatan pertama adalah memperkenalkan pembelajaran multimodal dengan menggabungkan data teks bersama imej dan bunyi. Ungkapan kebencian sering kali disampaikan melalui *meme*, video atau imej dan bukannya teks sahaja. Contohnya, frasa "*I love it!*" yang disertakan dengan gambar yang menghina boleh menjadi ungkapan kebencian. Oleh itu, himpunan data yang melabelkan kandungan kebencian multimodal akan membantu model memahami interaksi pelbagai bentuk media.

Pendekatan seterusnya ialah meningkatkan kepekaan konteks model terhadap bahasa. Format seperti BERT (Devlin et al., 2019) atau RoBERTa merupakan contoh seni bina transformer-based yang menunjukkan prestasi cemerlang dalam memahami bahasa secara halus. Masalah seperti sindiran dan ungkapan kebencian berkod boleh diatasi melalui penalaan halus model pada korpus tambahan media sosial. Tambahan pula, pendekatan seperti *adversarial debiasing* boleh mengurangkan kesan *false positive* berdasarkan dialek. Sekiranya model dilatih menggunakan sampel bahasa yang meluas dan membuat penalti terhadap ramalan *bias*, model akan boleh menjadi lebih adil dan inklusif.

Secara keseluruhannya, kajian ini memberikan sumbangan bermakna terhadap pengesanan ungkapan kebencian melalui model yang dicadangkan yang menggabungkan *FastText embeddings*, *Attention* Bi-LSTM dan ESGO. Cabaran berterusan dalam mengesan sindiran, *bias* budaya, dan kecekapan pengiraan mencerminkan kerumitan dalam mengautomatikkan pemprosesan kandungan.

Langkah seterusnya adalah dengan mengambil pendekatan rentas-disiplin termasuk model konteks yang lebih kaya, teknik pengurangan *bias* dan strategi pelaksanaan untuk membentuk sistem yang lebih tepat dan adil. Memandangkan komunikasi dalam talian terus berkembang, alat yang digunakan untuk mengawalnya juga perlu berkembang (Kinit, 2023). Oleh itu, kerjasama antara ahli linguistik, pakar etika dan pembangun platform adalah penting untuk memastikan pengesanan ungkapan kebencian dilaksanakan secara tepat sambil mengekalkan kebebasan bersuara dan keterangkuman.

6.0 PENGHARGAAN

Penulis kajian ini ingin mengucapkan setinggi-tinggi penghargaan dan jutaan terima kasih kepada Dr. Wandeep Kaur A/P Ratan Singh, penyelia penulis kajian ini, yang telah memberi tunjuk ajar serta bimbingan untuk menyiapkan projek ini dengan jayanya.

Penulis kajian ini juga ingin mengucapkan terima kasih kepada semua pihak yang membantu secara langsung mahupun tidak langsung dalam menyempurnakan projek ini. Segala bantuan yang telah dihulurkan amatlah dihargai kerana tanpa bantuan mereka, projek ini tidak dapat dilaksanakan dengan baik. Semoga Tuhan merahmati dan memberikan balasan yang terbaik.

7.0 RUJUKAN

Abro, S., Shaikh, S., Khand, Z. H., Ali, Z., Khan, S., & Mujtaba, G. (2020). Automatic hate speech detection using machine learning: A comparative study. International Journal of Advanced Computer Science and Applications (IJACSA), 11(8). https://dx.doi.org/10.14569/IJACSA.2020.0110861

Aljarah I., Habib, H., & Abuhaija, B. (2020). Intelligent Detection of Hate Speech in Arabic Social Network: A Machine Learning Approach. Journal of Information and Knowledge Management, 19(1), 53-64. http://dx.doi.org/10.1177/0165551520917651

Aluru, S.S., Mathew, B., Saha, P., & Mukherjee, A. (2020). Deep Learning Models for Multilingual Hate Speech Detection. Proceedings of the 12th ACM Conference on Web Science, 88-96. https://doi.org/10.48550/arXiv.2004.06465

Asiri, Y. (2022). Enhanced Seagull Optimization with Natural Language Processing Based Hate Speech Detection and Classification. International Journal of Computing, Vol. 21(1), 45-58. https://doi.org/10.3390/app12168000

Asogwa, D.C., Chukwuneke, C.I., Ngene, C.C., & Anigbogu, G.N. (2022). Hate Speech Classification Using SVM and Naive Bayes. https://arxiv.org/abs/2204.07057

Badri, N., Kboubi, F., Habacha, A., & Chaibi. (2022) Combining FastText and Glove Word Embedding for Offensive and Hate speech Text Detection, Procedia Computer Science, 207, 769-778. https://doi.org/10.1016/j.procs.2022.09.132.

Bilewicz, M., & Soral, W. (2020). Hate Speech Epidemic. The Dynamic Effects of Derogatory Language on Intergroup Relations and Political Radicalization 0:2 https://doi.org/10.1111/pops.12670

Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. (2019). Spread of Hate Speech in Online Social Media. Proceedings of the 10th ACM Conference, 173–182. https://doi.org/10.1145/3292522.3326034

Chiu, K.L., & Alexander, R. (2021). Detecting Hate Speech with GPT-3. https://doi.org/10.48550/arXiv.2103.12407

Dalavi, S., Nivelkar, T., Patil, S., Sawant, A., & Aylani, A. (2023). Comparative analysis of vectorization techniques and machine learning models for hate speech detection. 2023 Global Conference on Information Technologies and Communications (GCITC), 1–5. https://doi.org/10.1109/GCITC60406.2023.10426214

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM '17) (pp. 512–515). AAAI Press.

de Gibert, O., Perez, N., García-Pablos, A., & Cuadros, M. (2018). Hate speech dataset from a white supremacy forum. In Proceedings of the 2nd Workshop on Abusive Language Online (ALW2) (pp. 11-20). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5102

Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL. 10.18653/v1/N19-1423

Dhiman, G., Kumar, V. (2019). Seagull optimization algorithm: Theory and its applications for large-scale industrial engineering problems. Knowl.-Based Syst, 165, 169–196. https://doi.org/10.1016/j.knosys.2018.11.024

Edizel, B., Piktus, A., Bojanowski, P., Ferriera, R., Grave, E., & Silvestri, F. (2019) Misspelling Oblivious Word Embeddings. Proceedings of NAACL-HLT 2019, 3226–3234 https://doi.org/10.48550/arXiv.1905.09755

Eilers, J., Zhang, Z., & Trivedi, H. (2021). AItBERT: Domain-Specific Pretraining on Alternative Social Media to Improve Hate Speech Classification. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021), 872-881. http://dx.doi.org/10.48550/arXiv.1910.12574

European Union Agency for Fundamental Rights. (2023). Online content moderation. Publications Office of the European Union. https://fra.europa.eu/en/publication/2023/online-content-moderation

GeeksforGeeks. (n.d.). FastText working and implementation. Retrieved [insert retrieval date], from https://www.geeksforgeeks.org/machine-learning/fasttext-working-and-implementation/

Hussain, S., Malik, A., & Khan, M. (2020). Decoding intentional obfuscation in hate speech: A linguistic and computational approach. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 123-135). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.123

Isnain, A.R., Rahman, A., & Hadiana, F. (2020). Bidirectional Long Short Term Memory Method and Word2vec Extraction Approach for Hate Speech Detection. Journal of Big Data, 7(58), 1-12. http://dx.doi.org/10.22146/ijccs.51743

Kinit. (2023). How artificial intelligence can detect hate speech. https://kinit.sk/how-artificial-intelligence-can-detect-hate-speech/

Kousika, N. (2024). Enhancing Multimodal Sentiment Analysis with Deep Learning Techniques to Foster Emotional Intelligence. Journal of Emerging Technologies in Computing, Vol. 14(2), 123-139. 10.1109/ICCSP60870.2024.10544097

Kumar, A., Dutta, S., & Pranav, P. (2023). Supervised learning for Attack Detection in Cloud. International Journal of Experimental Research and Review, 31(Spl Volume), 74–84. https://doi.org/10.52756/10.52756/ijerr.2023.v31spl.008

MacAvaney, S., Yao, H. R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. PloS one, 14(11). https://doi.org/10.1371/journal.pone.0221152

Malik, J. S., Qiao, H., Pang, G., & van den Hengel, A. (2023) Deep Learning for Hate Speech Detection: A Comparative Study https://doi.org/10.48550/arXiv.2202.09517

Matamoros-Fernández, A., & Farkas, J. (2021). Hate speech and social media: Examining the challenges of detection in evolving digital spaces. Digital Journalism, 9(4), 526-547 https://doi.org/10.1177/1527476420982230

Mody, D., Huang, Y., & Alves de Oliveira, T. E. (2022). A curated hate speech dataset [Data set]. Mendeley Data. https://doi.org/10.17632/9sxpkmm8xn.1

Negi, Chitranjali. The Rise of Hate Speech Around the World. Independent Researcher, (2024) https://dx.doi.org/10.2139/ssrn.4719266

Ndabula, J.N., Odeh, S., & Hassan, A.M. (2023). Detection of Hate Speech Code Mix Involving English and Other Nigerian Languages. African Journal of Computing and ICT, 16(2), 85-96 http://dx.doi.org/10.51519/journalisi.v5i4.595

Pacheco, D., Hua, Y., Torres-Lugo, C., Truong, B. T., & Roy, S. (2020). Uncovering coordinated networks on social media: Methods and case studies. Proceedings of the International AAAI Conference on Web and Social Media, 14(1), 455-466. https://doi.org/10.1609/icwsm.v15i1.18075

Paz, M. A., Montero-Díaz, J., & Moreno-Delgado, A. (2020). Hate Speech: A Systematized Review. Sage Open, 10(4). https://doi.org/10.1177/2158244020973022

Pereira-Kohatsu, J.C., Quijano-Sanchez, L., Liberatore, F., & Camacho-Collados, M. (2019). Detecting and Monitoring Hate Speech in Twitter. Applied Sciences, 9(22), 4654. https://doi.org/10.3390/s19214654

Petro, L., & Pavlo. (2019). L. Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS. https://arxiv.org/abs/1912.06059

Pyngkodi, M. et al., (2023) Hate Speech Analysis using Supervised Machine Learning Techniques, 2023 International Conference on Computer Communication and Informatics, 1-6. https://doi.org/10.1109/ICCCI56745.2023.10128591

QATrainingHub. (n.d.). The importance of data processing in machine learning & AI. Retrieved June 20, 2025, from https://qatraininghub.com/importance-of-data-processing-in-machine-learning-ai/

Robertson, A. (2025, January 7). Meta will stop fact-checking misinformation as Zuckerberg shifts focus. The Verge. https://www.theverge.com/2025/1/7/24338127/meta-end-fact-checking-misinformation-zuckerberg

Roy, P.K., Uddin, S., Priya, A., & Das, S. (2020). A Framework for Hate Speech Detection Using Deep Convolutional Neural Network. Journal of Information and Computational Science, 10(6), 1234-1245. http://dx.doi.org/10.1109/ACCESS.2020.3037073

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N.A. (2019). The Risk of Racial Bias in Hate Speech Detection. ACL. 10.18653/v1/P19-1163

SitesBay. (n.d.). Spiral model in software engineering. Retrieved June 20, 2025, from https://www.sitesbay.com/software-engineering/se-spiral-model

Teja Sri, N., Geethika.K, Neha Kotha, Harini Kandoori. (2023). Modified TF-IDF with Machine Learning Classifier for Hate Speech Detection on Twitter. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 14(03), 978–984. https://doi.org/10.17762/turcomat.v14i03.14177

Wang, J., & Zhang, Y. (2019). Large-Scale Text Classification Using Scope-Based Convolutional Neural Network: A Deep Learning Approach. IEEE Transactions on Neural Networks and Learning Systems, 31(7), 2346-2356 http://dx.doi.org/10.1109/ACCESS.2019.2955924

Zhou, Y., Yang, Y., Liu, H., Liu, X., & Savage, N. (2020). Deep Learning Based Fusion Approach for Hate Speech Detection Vol. 8, 128923-128929 https://doi.org/10.1109/ACCESS.2020.3009244

Samuel a/l Ravi (A194594)
Dr. Wandeep Kaur A/P Ratan Singh
Fakulti Teknologi & Sains Maklumat
Universiti Kebangsaan Malaysia