PENGESANAN UCAPAN KEBENCIAN DI MEDIA SOSIAL MENGGUNAKAN PEMBELAJARAN MESIN

NUR NAJIHAH BINTI SA'RANI

NAZLIA BINTI OMAR

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

ABSTRAK

Pada era globalisasi ini, media sosial dan Internet merupakan satu platform utama yang digunakan oleh hampir semua lapisan masyarakat sebagai satu medium untuk berkongsi sebarang idea mahupun pendapat tanpa had. Namun, kebebasan bersuara yang diberikan sering kali disalahguna oleh sebahagian pengguna yang sengaja berniat untuk menzahirkan kebencian dan diskriminasi terhadap individu mahupun organisasi, atas beberapa faktor tertentu seperti jantina, perkauman, etnik, agama, politik dan sebagainya. Perkara ini secara tidak langsung mendorong kepada perpecahan di kalangan masyarakat. Hari ini, pelbagai kajian telah dilakukan untuk mengesan ucapan kebencian di dalam media sosial, namun agak sukar untuk memproses teks data yang dikumpul memandangkan isi kandungannya yang disifatkan sebagai teks hingar, iaitu tidak tersusun dan mengandungi ralat. Kajian ini bertujuan untuk mengesan ucapan kebencian di media sosial dengan menggunakan dataset yang dikumpul daripada Twitter dengan menekankan ciri-ciri yang relevan. Kaedah yang digunakan adalah pembelajaran mesin dengan menggunakan beberapa model pengelas seperti Bayes Naif, Mesin Vektor Sokongan dan Regresi Logistik untuk menguji ketepatan model. Pada akhir kajian, model pengelas yang mempunyai ketepatan paling tinggi untuk mengesan ucapan kebencian di media sosial akan dikenal pasti. Dalam pada itu, isu dan cabaran yang timbul semasa proses mengesan ucapan kebencian akan diketengahkan.

Kata kunci: Ucapan kebencian, Media sosial, Pembelajarn mesin

PENGENALAN

Ucapan kebencian didefinisikan sebagai komunikasi yang ditujukan kepada orang atau kumpulan tertentu berdasarkan pelbagai faktor termasuklah etnik, warna kulit, jantina, ketidakupayaan, orientasi seksual, kewarganegaraan, agama, dan lain-lain (Omar, 2019). Perilaku ini boleh dilakukan oleh individu atau kumpulan dengan menunjukkan kebencian dalam bentuk provokasi, hasutan, ataupun hinaan. Istilah "ucapan kebencian" merujuk kepada tindakan, perkataan, perbuatan, tulisan atau persembahan yang dilarang kerana ia mungkin mendorong berlakunya sesuatu tindakan keganasan dan prasangka yang mungkin saja dilakukan oleh si pelaku mahupun mangsa pelaku tersebut. Lonjakan kes keganasan terhadap

minoriti di peringkat global seperti pembersihan etnik, tembakan rambang dan pembunuhan terancang telah dikaitkan dengan ucapan kebencian. Jelaslah bahawa perkembangan teknologi informasi dan komunikasi telah membawa pengaruh positif dan negatif, ibarat pedang bermata dua.

Media sosial pula telah menjadi satu platform yang tidak asing lagi dikalangan masyarakat tidak mengira lapisan umur kerana media sosial telah menjadi platform digital yang membolehkan pengguna menghasilkan, berkongsi dan mengembangkan idea mahupun pendapat mereka, dalam masa yang sama berhubung dengan orang lain secara atas talian. Memetik daripada sebuah journal, kebebasan berpendapat dalam forum sesama pengguna sosial media Youtube dan juga kurangnya pengetahuan tentang bentuk kejahatan berbahasa mengakibatkan bentuk tuturan kejahatan berbahasa kerap dijumpai dan juga menimbulkan pro-kontra sesama pengguna media sosial youtube yang dapat merugikan individu ataupun kelompok masyarakat (Furqan, et al 2022). Peningkatan penggunaan media sosial dalam kehidupan seharian masyarakat pada hari ini telah banyak menimbulkan aspek negatif yang tidak sewajarnya berlaku. Dampak negatif yang ditimbulkan dalam medial sosial, antaranya adalah tingkah laku segelintir masyarakat yang gemar mengeluarkan kenyataan yang memiliki muatan penghinaan, pencemaran nama baik dan sebagainya. Pengesanan ucapan kebencian dalam teks media sosial perlu dikaji dengan lebih mendalam.

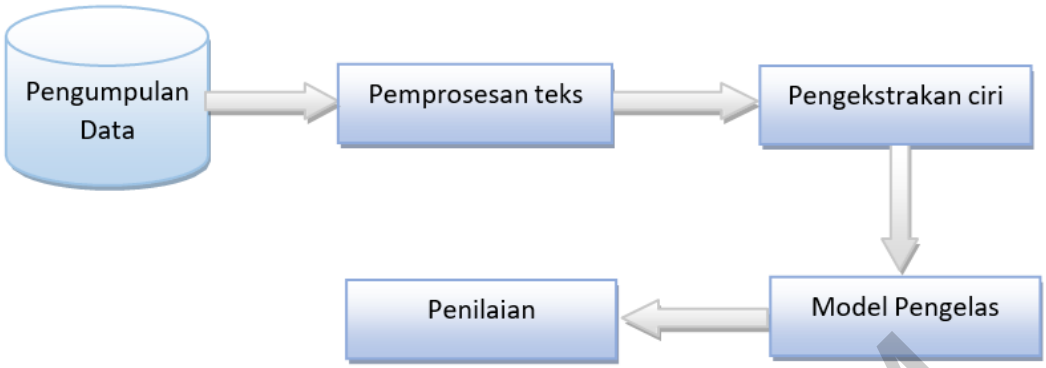
Oleh itu, projek ini dijalankan untuk mengkaji dan mengesan sebarang komunikasi di media sosial yang mengandungi unsur-unsur kebencian dengan menggunakan beberapa model pengelas yang terdapat dalam pembelajaran mesin seperti Model Regresi Logistik, Model Mesin Vektor Sokongan, dan Model Bayes Naif seterusnya mengenalpasti model yang mampu memberikan keputusan terbaik dalam eksperimen ini. Penggunaan teknik pembelajaran Mesin (ML) merupakan salah satu sub-bidang kecerdasan buatan (AI), selain pembelajaran mendalam yang penting dan berkemampuan untuk mengesan ucapan kebencian ini. Kajian ini memberi tumpuan pada teks Bahasa Inggeris yang dimuatnaik oleh pengguna di media sosial terutamanya Twitter. Selain itu, dalam kajian ini, pembinaan web telah dijalankan khususnya untuk membangunkan papan pemuka (Dashboard) iaitu antara muka interaktif yang membolehkan pengguna meneroka dan menganalisis data yang bersesuaian dengan kajian ini iaitu untuk mengesan ucapan kebencian di media sosial menggunakan pembelajaran mesin. '*Streamlit'* dan *'Python*' adalah alat dan Bahasa pengaturcaraan yang digunakan untuk membina web ini. Kajian ini diharapkan dapat membantu organisasi tertentu terutamanya pihak pengasas syarikat media sosial tertentu itu sendiri untuk mengesan sebarang ucapan kebencian yang disebarkan secara meluas di dalam media sosial milik mereka justeru dapat bertindak secara efektif untuk menghapuskan atau menyekat ucapan kebencian tersebut daripada terus menular. Oleh demikian, kajian ini amat penting kerana ianya dapat membantu dan menyelamatkan pihak tertentu terutamanya mangsa buli siber, dalam masa yang sama memastikan persekitaran yang selamat dan harmoni di kalangan pengguna di media sosial.

Beberapa kajian lepas telah dikumpulkan sebelum projek ini dimulakan. Antaranya adalah kajian oleh Asogwa et al. (2022) telah memberi tumpuan pada pembangunan model

pembelajaran mesin untuk klasifikasi ucapan kebencian secara automatik menggunakan Mesin Vektor Sokongan (MVS) dan Bayes Naif secara automatik. Hasil eksperimen menunjukkan bahawa MVS mencapai ketepatan klasifikasi sekitar 99% dan Bayes Naif mencapai ketepatan klasifikasi sekitar 50% pada set ujian. MVS memberikan hasil yang lebih baik daripada Bayes Naif dalam mengklasifikasikan ucapan kebencian. Ali et al. (2021) menangani tiga masalah yang biasa berlaku dalam analisis sentimen berasaskan pembelajaran mesin iaitu kekerapan rendah, dimensi , dan skew kelas dengan menggunakan teknik canggih dan mencatatkan prestasi penambahbaikan ke atas model asas. Model Mesin Vektor Sokongan (MVS) dan Bayes Naif telah digunakan untuk melatih mesin. Penyelidik telah menggunakan teknik Skim Pemilihan Ciri Pembolehubah Global (VGFSS) untuk mengurangkan dimensi manakala Teknik Pensampelan Lebihan Minoriti Sintetik (SMOTE) untuk mengurangkan ketidakseimbangan kelas. Akhir sekali, keputusan menunjukkan peningkatan maksimum dalam keseluruhan prestasi pengesanan ucapan benci berasaskan analisis sentimen setelah tiga masalah itu dikurangkan. Gaydhani et al. (2018) mencadangkan pendekatan untuk mengklasifikasikan twit secara automatik di Twitter kepada tiga kategori iaitu membenci, menyinggung perasaan dan neutral. Pengkaji menggunakan set data Twitter untuk melatih model pembelajaran mesin mereka, memfokuskan pada n-gram sebagai ciri dan Kekerapan Istilah-Frekuensi Dokumen Songsang (TFIDF) dengan mempertimbangkan model Regresi Logistik, Bayes Naif dan Mesin Vektor Sokongan. Hasilnya, pengkaji mencapai ketepatan yang lebih baik iaitu 95.6% pada data ujian. Kajian lain oleh Alshalan et. al (2020) yang telah menjalankan penyelidikan tentang penggunaan pendekatan pembelajaran mendalam (deep learning) iaitu model rangkaian neural, khususnya Rangkaian Neural Konvolusi (CNN) dan Rangkaian Neural Berulang (RNN), serta model representasi bahasa ‘Bidirectional Encoder Representations from Transformers (BERT)’, untuk mengesan ucapan kebencian secara automatik dalam Twittersfera Arab Saudi. Hasil kajiannya menunjukkan bahawa model CNN mencapai prestasi terbaik, dengan skor F1 sebanyak 0.79 dan kawasan di bawah kurva penerimaan operasi (AUROC) sebanyak 0.89.

METODOLOGI KAJIAN

Dalam konteks pembelajaran mesin, metodologi merujuk kepada kaedah, langkah-langkah, teknik, dan prosedur yang digunakan untuk merancang, mengembangkan, dan melaksanakan aplikasi pembelajaran mesin. Metodologi ini adalah komponen penting dalam proses penyelidikan dan analisis data kerana ianya membantu memudahkan pengkaji untuk menyusun dan melaksanakan projek pembelajaran mesin dengan cara yang sistematik dan mencapai matlamat kajian yang efektif. Fasa-fasa yang terlibat dalam mengesan twit yang mempunyai unsur kebencian ialah fasa pengumpulan data, fasa pra-pemprosesan, fasa pengekstrakan ciri, fasa pengelasan dan fasa penilaian. Rajah 1 menunjukkan metodologi kajian.



Rajah 1 Metodologi kajian

Fasa pengumpulan data

Fasa ini merupakan salah satu aspek penting dalam kajian pengesanan ucapan kebencian di platform media sosial Twitter. Dalam kajian ini, set data standard seperti (Waseem & Hovy, 2016) yang mengandungi 15,600 twit telah digunakan. Statistik menunjukkan sejumlah 4839 twit menunjukkan unsur ucapan kebencian manakala selebihnya tidak mengandungi sebarang kebencian. Set data ini kemudiannya digabungkan dengan sebuah set data yang diperoleh daripada platform Github untuk meningkatkan lagi prestasi model. Ini menjadikan keseluruhan data mempunyai 18,560 total twit.

Fasa pra-pemprosesan

Pra-pemprosesan data bertujuan untuk membersihkan, menyesuaikan, dan mengorganisasikan data sehingga dapat digunakan secara efektif oleh model pembelajaran mesin. Set data yang telah dibersihkan akan direkod dan disimpan di dalam dokumen yang berformat CSV. Langkah pertama dalam fasa ini, pembersihan teks dilakukan bertujuan untuk membersihkan teks dari elemen yang tidak relevan atau mengganggu. Sebagai contoh, menghapuskan pautan URL, menghilangkan karakter khusus, tanda baca yang tidak diperlukan dan mengatasi masalah ejaan mahupun singkatan.

Seterusnya, proses tokenisasi dilakukan iaitu proses memecahkan teks menjadi potongan-potongan yang lebih kecil, seperti kata-kata atau frasa. Tokenisasi diperlukan untuk mengidentifikasi kata-kata yang mengandungi emosi atau sentimen tertentu, misalnya kata-kata yang mengandungi rasa benci. Pendekatan pembelajaran mesin dapat digunakan untuk mengesan kata-ucapan kebencian dalam teks yang telah ditokenisasi. Model dilatih untuk mengidentifikasi kata-kata yang membayangkan sentimen negatif, termasuk ucapan kebencian, berdasarkan data latihan yang telah ditokenisasi.

Setelah itu, proses normalisasi dilakukan untuk memfokuskan penyusunan teks ke dalam bentuk standard atau normal agar dapat diproses dengan lebih mudah. Hal ini bertujuan untuk memastikan konsistensi dan keseragaman dalam perwakilan teks. Proses lematisasi digunakan semasa proses normalisasi berbanding stemming. Lematisasi dan stemming adalah proses penukaran sesuatu perkataan dalam teks kepada bentuk dasar atau kata dasarnya. Namun, proses lematisasi lebih cenderung untuk memberikan bentuk dasar atau akar kata

yang lebih tepat berbanding stemming. Ia mempertimbangkan makna perkataan dan mengekalkan semantik dengan lebih baik, yang memudahkan proses analisis dilakukan untuk menilai samada sesuatu perkataan itu mengandungi unsur kebencian atau tidak. Oleh itu, kaedah lematisasi lebih sesuai digunakan dalam kajian pengesanan ucapan kebencian ini.

Fasa pengekstrakan ciri

Fasa pengekstrakan ciri adalah langkah penting dalam proses analisis data, terutamanya dalam konteks pembelajaran mesin. Pada asasnya, pengekstrakan ciri melibatkan pengidentifikasian dan pemilihan ciri yang paling relevan dan bermakna daripada data mentah. Ciri ini kemudiannya digunakan untuk melatih model atau algoritma pembelajaran mesin.

Kaedah pertama yang digunakan adalah TF-IDF dimana ianya biasa digunakan untuk mengekstrak ciri daripada data teks. Ia membantu mewakili kepentingan sesuatu perkataan dalam dokumen atau korpus dengan memberikan pemberat berdasarkan kekerapan perkataan itu muncul dalam dokumen dan kekerapan perkataan itu merentasi korpus. TF-IDF ini digunakan secara meluas dalam perolehan maklumat, perlombongan teks dan pembelajaran mesin untuk tugasan seperti carian dokumen, analisis teks automatik dan klasifikasi teks.

Kaedah seterusnya adalah menggunakan Beg Perkataan (*Bag of Words*). BoW mewakili dokumen sebagai koleksi perkataan tanpa mengambil kira urutan atau struktur ayat. Langkah-langkah tersebut melibatkan tokenisasi, mengira kekerapan kejadian perkataan, dan vektorisasi untuk membentuk vektor ciri. Setiap dimensi vektor mewakili satu perkataan, dan nilai dalam setiap dimensi ialah kekerapan perkataan itu dalam dokumen. Kaedah ini mudah dan berkesan, tetapi boleh menyebabkan kehilangan maklumat dalam susunan perkataan.

Kaedah yang ketiga adalah dengan meggunakan leksikon kebencian untuk mencari perkataan atau frasa yang berbaur kebencian sama ada menyinggung atau mendiskriminasi, yang terdapat dalam teks. Sistem boleh menandakan atau menapis kandungan yang mungkin mengandungi unsur kebencian dengan menyemak sama ada teks tersebut mengandungi perkataan daripada leksikon ini. Di samping itu, model boleh dilatih untuk mengenali corak dan konteks di mana perkataan benci muncul, supaya ia boleh menjadi lebih berkesan dalam mengesan ucapan kebencian.

Fasa pengelasan

Fasa ini juga merupakan tahap yang penting dalam projek pengesanan ucapan kebencian di Twitter menggunakan pembelajaran mesin. Keseluruhan set data berjumlah 18,560 twit, dibahagikan kepada set data latihan sebanyak 80% dan set data ujian sebanyak 20%. Fasa ini melibatkan penggunaan model pembelajaran mesin yang telah dilatih untuk mengklasifikasikan data baru menjadi kategori yang telah ditentukan, seperti "ucapan kebencian" atau "bukan ucapan kebencian". Kejayaan fasa ini sangat bergantung pada pemilihan model yang sesuai dan pemahaman mendalam tentang data yang sedang diproses. Dalam fasa ini, model yang telah dilatih dengan data latihan digunakan untuk membuat ramalan ke atas data ujian atau data baru. Model pengelas seperti Bayes Naif, Regresi Logistik dan Mesin Vektor Sokongan digunakan dalam kajian ini memandangkan model-model

tersebut sesuai dengan ciri-ciri data dan objektif klasifikasi.

Fasa pengujian

Fasa penilaian adalah fasa terakhir yang bertujuan untuk mengukur sejauh mana model atau algoritma yang telah dibangunkan berhasil mencapai tujuan klasifikasinya. Penilaian menyeluruh membantu memahami kekuatan dan kelemahan model dan boleh membawa kepada penambahbaikan seterusnya.

Kepersisan mengukur sejauh mana prediksi positif dari model benar dalam perbandingan dengan total prediksi positif. Kepersisan memberi gambaran tentang betapa tepatnya model dalam mengenal pasti tweet yang sebenarnya mengandungi kata-kata kebencian. Jika nilai kepersisan adalah tinggi, model cenderung memberikan sedikit positif palsu. Berikut adalah formula kepersisan.

$$\text{Kepersisan} = \frac{Positif\ Sebenar}{Positif\ Sebenar + Positif\ Palsu}$$

Dapatan semula mengukur bahagian data sebenar positif yang dikenal pasti dengan betul oleh model. Ia memberikan maklumat tentang sejauh mana model itu boleh menangkap semua tweet yang mengandungi perkataan benci. Jika nilai dapatan semula tinggi, model cenderung memberikan sedikit negatif palsu. Berikut merupakan formula dapatan semula.

$$\text{Dapatan Semula} = \frac{Positif\ Sebenar}{Positif\ Sebenar + Negatif\ Palsu}$$

F1-skor memberikan pemahaman yang baik tentang keseimbangan antara kepersisan dan dapatan , terutama di situasi ketidakseimbangan kelas.

$$\text{F1-Skor} = 2 \times \left(\frac{Kepersisan \times Dapatan}{Kepersisan + Dapatan}\right)$$

KEPUTUSAN DAN PERBINCANGAN

Penilaian ketepatan telah dijalankan untuk meneliti model yang memberikan hasil tertinggi. Metrik penilaian juga digunakan untuk menilai model pembelajaran mesin yang dibangunkan, dengan tujuan memilih algoritma yang paling sesuai untuk mengesan ucapan kebencian. Setiap model dinilai berdasarkan metrik penilaian yang merangkumi ketepatan, kepersisan, dapatan semula, dan F1-Skor. Bagi menentukan nilai akhir keputusan bagi setiap model, purata pemberat (weighted average) akan dipilih sebagai metrik utama. Dengan menggunakan purata pemberat, prestasi model boleh diukur dengan lebih adil dan teliti, dengan mengambil kira ketidakseimbangan data yang mungkin wujud. Perkara ini bagi memastikan model yang dipilih dapat mengesan ucapan kebencian dengan tepat dan konsisten dalam mengklasifikasikan teks dengan betul walaupun terdapat variasi dalam data.

Jadual 1 menunjukkan perbandingan keputusan prestasi model Regresi Logistik sebelum dan selepas penggunaan SMOTE berdasarkan beberapa metrik penilaian iaitu ketepatan, kepersisan, dapatan semula, dan F1-skor. Sebelum penggunaan SMOTE, model Regresi Logistik menunjukkan ketepatan 0.8041, kepersisan 0.8264, dapatan semula 0.8041, dan F1-skor 0.7746. Selepas penggunaan SMOTE, terdapat peningkatan ketara dalam semua metrik penilaian. Ketepatan meningkat kepada 0.8499, kepersisan meningkat kepada 0.8476, dapatan semula meningkat kepada 0.8499, dan F1-skor meningkat kepada 0.8437. Penggunaan SMOTE membantu mengatasi masalah ketidakseimbangan data dengan menghasilkan sampel sintetik bagi kelas minoriti, yang mengurangkan bias model terhadap kelas majoriti. Ini meningkatkan kebolehan model untuk mengenali dan mengklasifikasikan dengan betul contoh daripada kedua-dua kelas minoriti dan majoriti.

Jadual 1 Keputusan Prestasi Model Regresi Logistik

Metrik Penilaian	Ketepatan	Kepersisan	Dapatan Semula	F1-Skor
Tanpa Pengunaan Smote	0.8041	0.8264	0.8041	0.7746
Setelah Penggunaan Smote	0.8499	0.8476	0.8499	0.8437

Jadual 2 menunjukkan perbandingan keputusan prestasi bagi model Mesin Vektor Sokongan (SVM) sebelum dan selepas penggunaan SMOTE. Sebelum penggunaan SMOTE, model mencatatkan ketepatan 0.8082, kepersisan 0.8246, dapatan semula juga 0.8082, dan F1-Skor 0.7822. Setelah penggunaan SMOTE, terdapat sedikit perubahan pada ketepatan, iaitu sedikit menurun dengan catatan 0.8079. Namun, kepersisan meningkat kepada 0.8193, sedangkan dapatan semula sedikit menurun menjadi 0.8079. F1-Skor pula mencatatkan bacaan 0.7839, menunjukkan peningkatan dalam keseimbangan antara kepersisan dan dapatan semula.

Jadual 2 Keputusan Prestasi Model Mesin Vektor Sokongan (SVM)

Metrik Penilaian	Ketepatan	Kepersisan	Dapatan Semula	F1-Skor
Tanpa Pengunaan Smote	0.8082	0.8246	0.8082	0.7822
Setelah Penggunaan Smote	0.8079	0.8193	0.8079	0.7839

Jadual 3 menunjukkan perbandingan keputusan prestasi model Bayes Naif sebelum dan selepas penerapan teknik SMOTE, yang ditunjukkan melalui empat metrik utama iaitu ketepatan, kepersisan, dapatan semula, dan F1-Skor. Sebelum penggunaan SMOTE, model Bayes Naif mencatatkan ketepatan sebesar 0.7530, kepersisan 0.7976, dapatan semula 0.7530, dan F1-Skor 0.6890. Selepas penggunaan SMOTE, terdapat peningkatan yang signifikan dalam semua metrik. Ketepatan meningkat kepada 0.8206, kepersisan pula mencatakan 0.8389, dapatan semula menunjukkan peningkatan setara dengan ketepatan iaitu 0.8206, dan F1-Skor meningkat secara mendadak menjadikannya 0.8254.

Jadual 3 Keputusan Prestasi Model Bayes Naif

Metrik Penilaian	**Ketepatan**	**Kepersisan**	**Dapatan Semula**	**F1-Skor**
Tanpa Pengunaan Smote	0.7530	0.7976	0.7530	0.6890
Setelah Penggunaan Smote	0.8206	0.8389	0.8206	0.8254

Jadual 4 merumuskan hasil akhir keputusan bagi kesemua model yang dibangunkan dalam kajian ini. Berdasarkan keputusan yang dicatatkan ini, telah terbukti bahawa model Regresi Logistik menunjukkan prestasi model terbaik diikuti dengan model Bayes Naif seterusnya Mesin Vektor Sokongan.

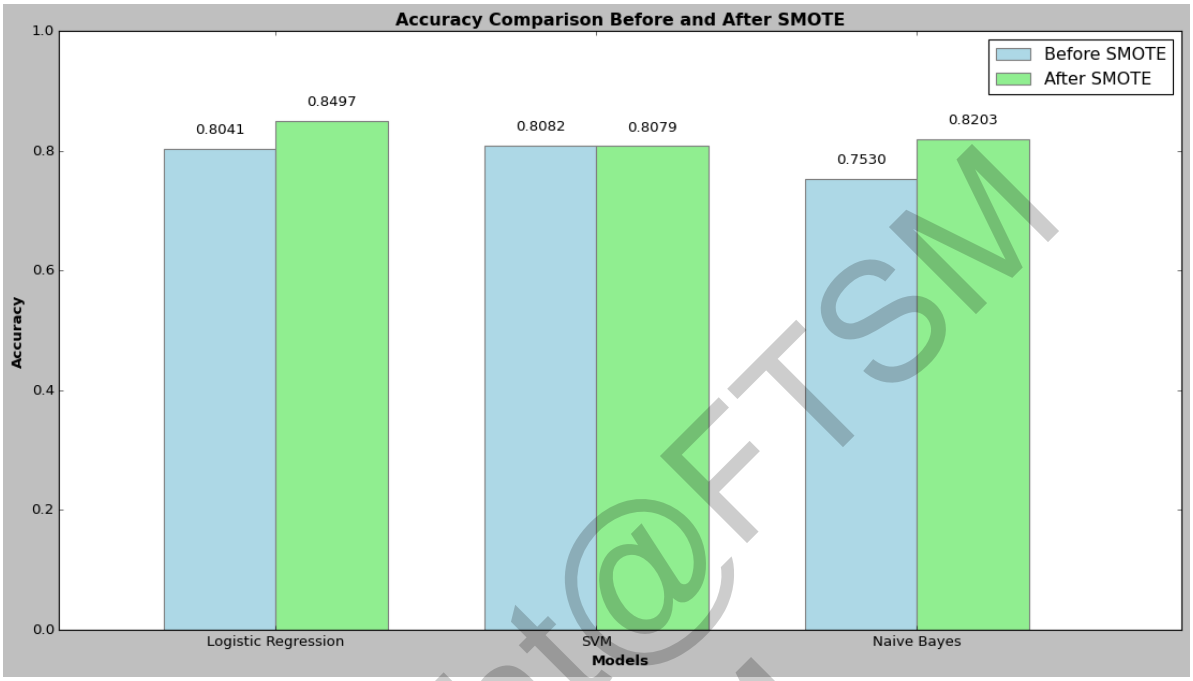
Jadual 4 Rumusan keputusan bagi kesemua model

Metrik Penilaian	**Ketepatan**	**Kepersisan**	**Dapatan Semula**	**F1-Skor**
Regresi Logistik	0.8499	0.8476	0.8499	0.8437
Mesin Vektor Sokongan	0.8079	0.8193	0.8079	0.7839
Bayes Naif	0.8206	0.8389	0.8206	0.8254

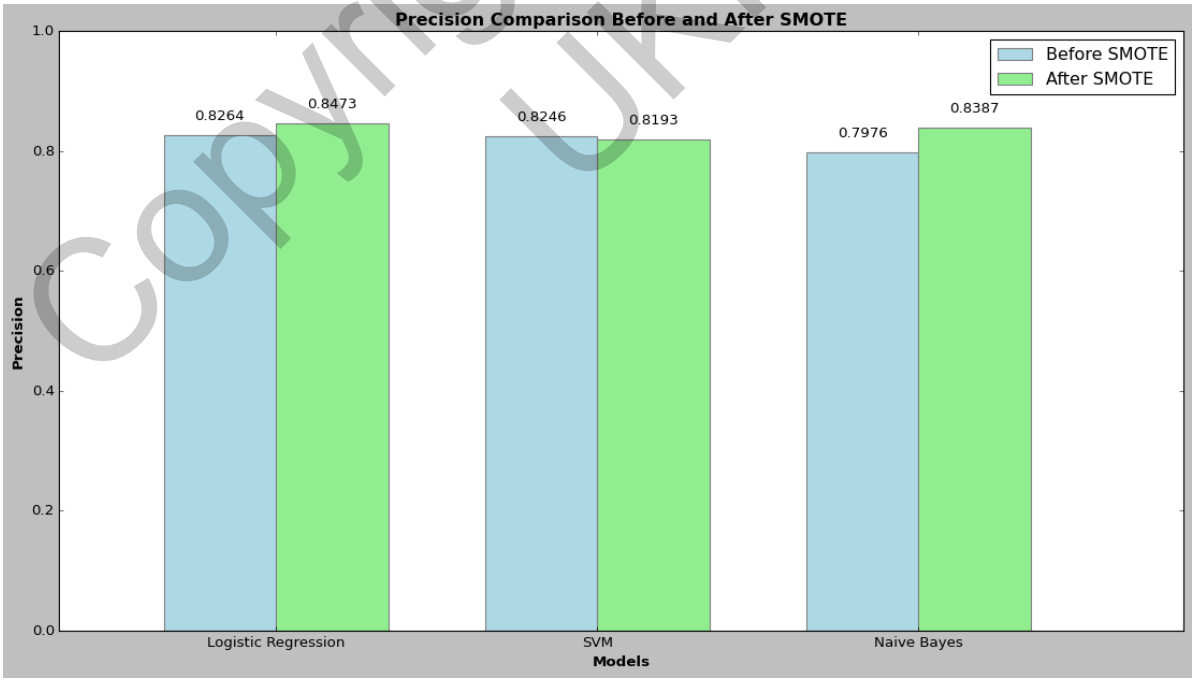
Keseluruhan peningkatan dalam metrik penilaian menunjukkan bahawa model menjadi lebih tepat dan seimbang dalam membuat ramalan setelah data dilatih dengan teknik SMOTE. Kesimpulannya, peningkatan dalam semua metrik penilaian selepas penggunaan SMOTE menunjukkan bahawa teknik ini sangat efektif dalam meningkatkan prestasi model regresi logistik untuk pengesanan ucapan kebencian di media sosial, dan menangani masalah ketidakseimbangan data adalah langkah kritikal dalam pembangunan model pembelajaran mesin yang lebih tepat dan boleh dipercayai.

Seterusnya, perbandingan prestasi antara model telah dilakukan. Hasil keputusan beberapa model yang telah dibangunkan, dinilai untuk menentukan model yang paling berkesan sesuai dengan matlamat kajian. Perbandingan ini adalah penting untuk menentukan

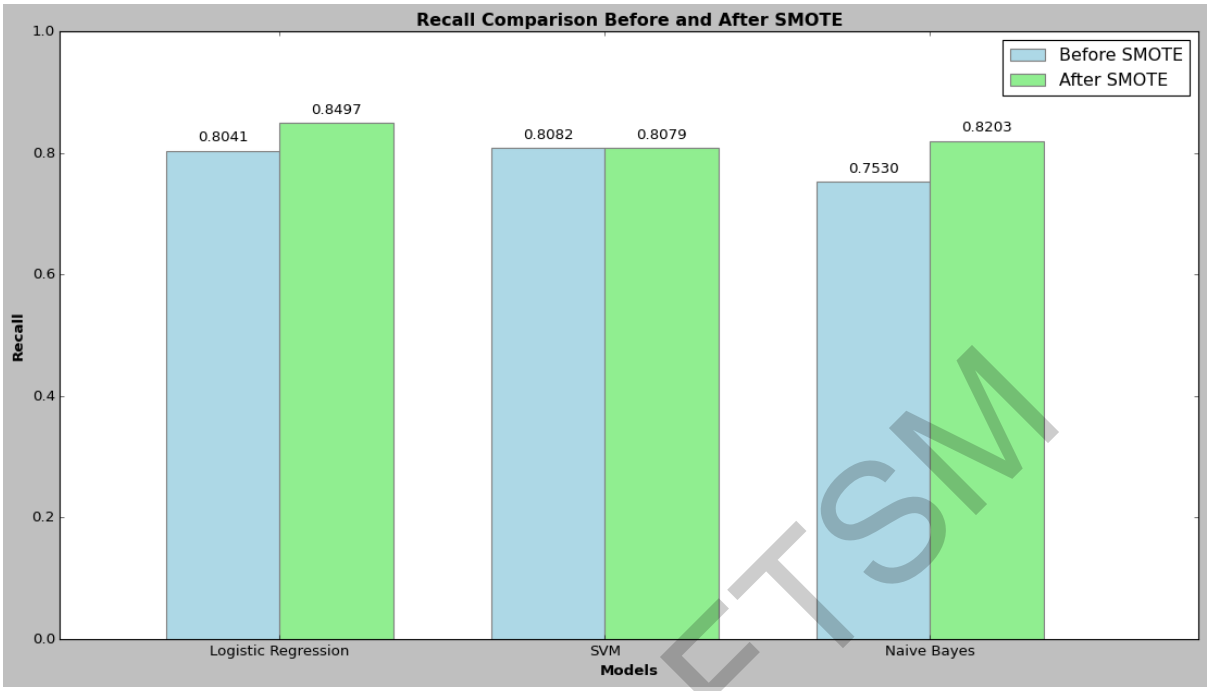
kesan SMOTE terhadap keupayaan model untuk menangani data yang tidak seimbang. Rajah 2, 3, 4 dan Rajah 5 menunjukkan perbandingan ketepatan, kepersisan, dapatan semulan dan nilai F1-skor antara tiga model pembelajaran mesin iaitu Regresi Logistik, Mesin Vektor Sokongan dan Bayes Naif, sebelum dan selepas menggunakan teknik SMOTE.



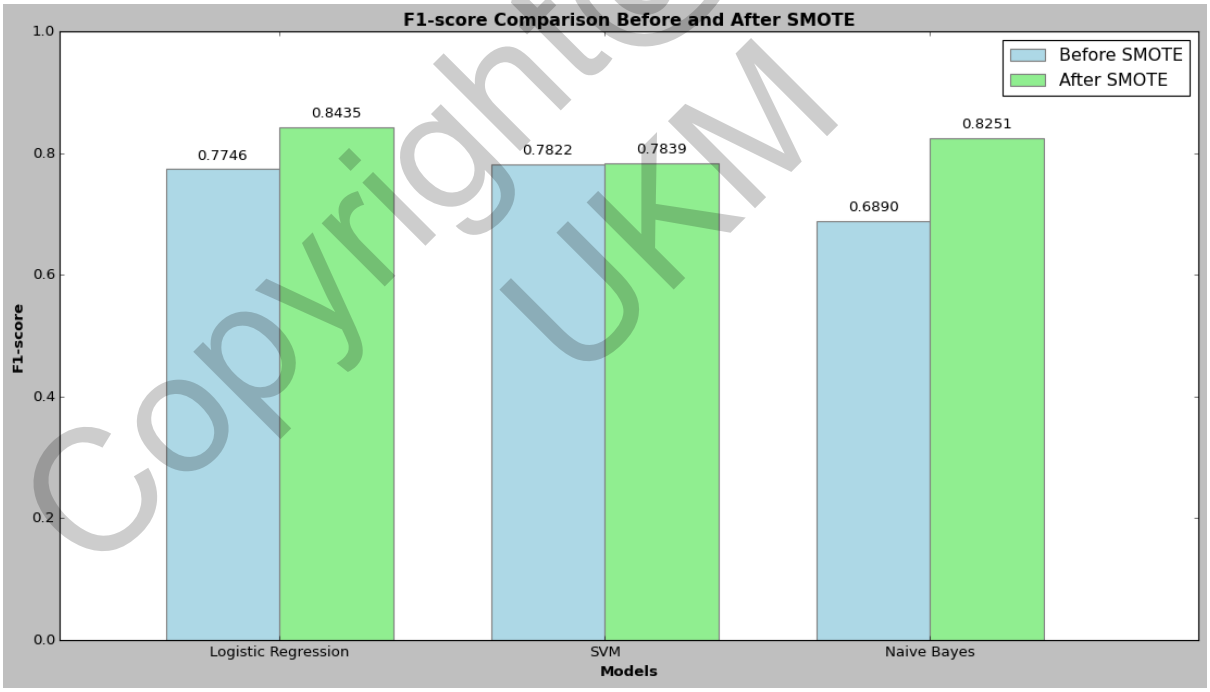
Rajah 2 Perbandingan Ketepatan antara model sebelum dan selepas penggunaan SMOTE



Rajah 3 Perbandingan Kepersisan antara model sebelum dan selepas penggunaan SMOTE



Rajah 4 Perbandingan Dapatan Semula antara model sebelum dan selepas penggunaan SMOTE

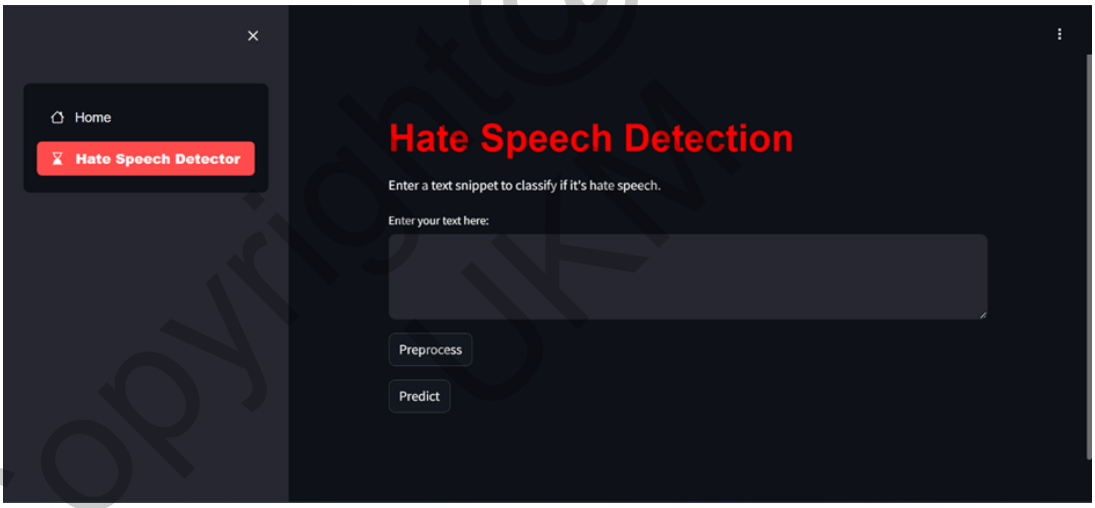


Rajah 5 Perbandingan F1-Skor antara model sebelum dan selepas penggunaan SMOTE

Papan pemuka yang telah dibangunkan dengan menggunakan ‘Streamlit’ pula direka supaya ianya mesra pengguna yang membolehkan pengguna memasukkan input dengan mudah dan mendapatkan hasil analisis yang jelas serta mudah difahami. Rajah 6 menunjukkan paparan “Home” bagi projek Pengesanan Ucapan Kebencian di Media Sosial Menggunakan Pembelajaran Mesin yang memberikan sedikit gambaran awal tentang projek ini kepada pengguna baru.

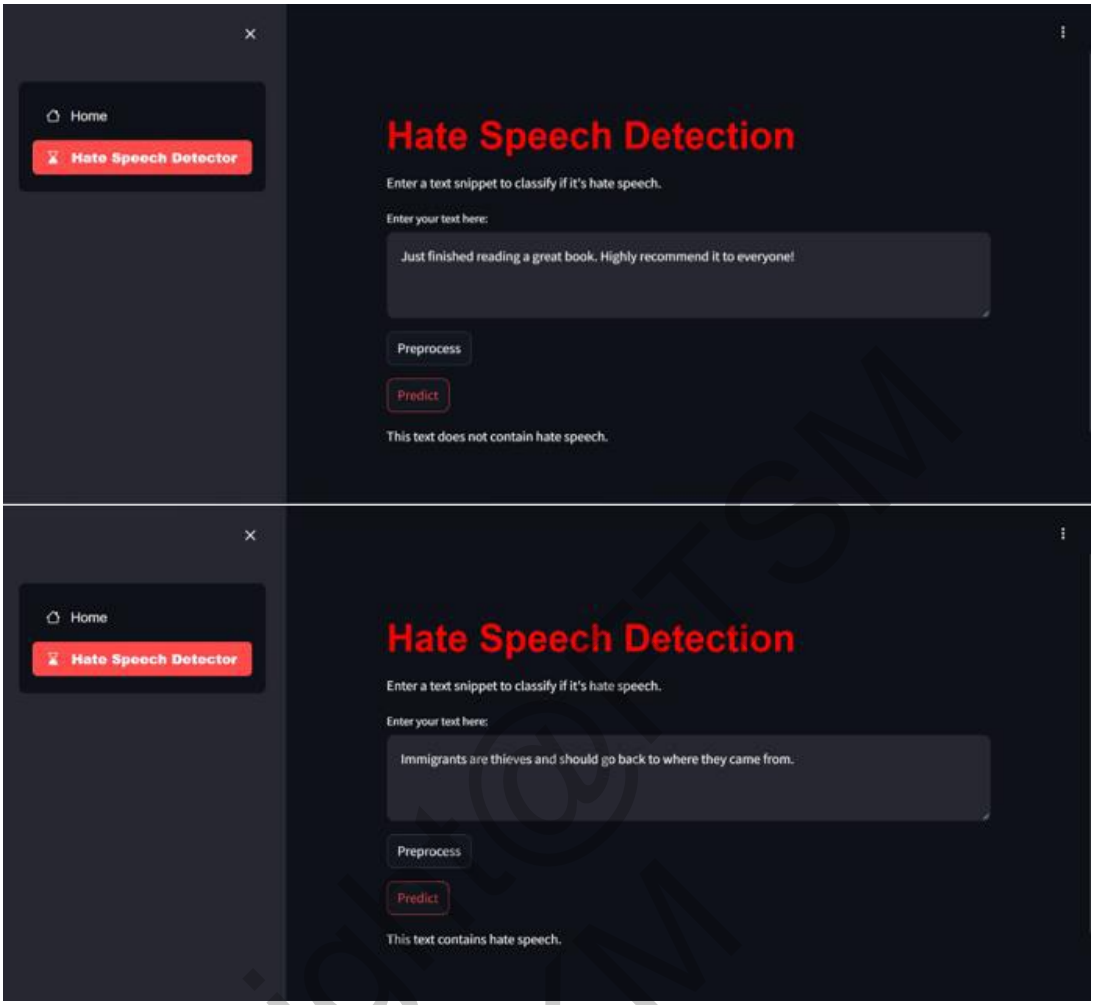
Rajah 6 Paparan Home

Rajah 7 pula menunjukkan paparan alat pengesan ucapan kebencian yang mudah digunakan dan difahami oleh pengguna. Pengguna hanya perlu memasukkan teks input ke dalam ruangan yang disediakan. Setelah itu, pengguna boleh menekan butang “Preprocess” dan “Predict” untuk mulakan proses analisis.



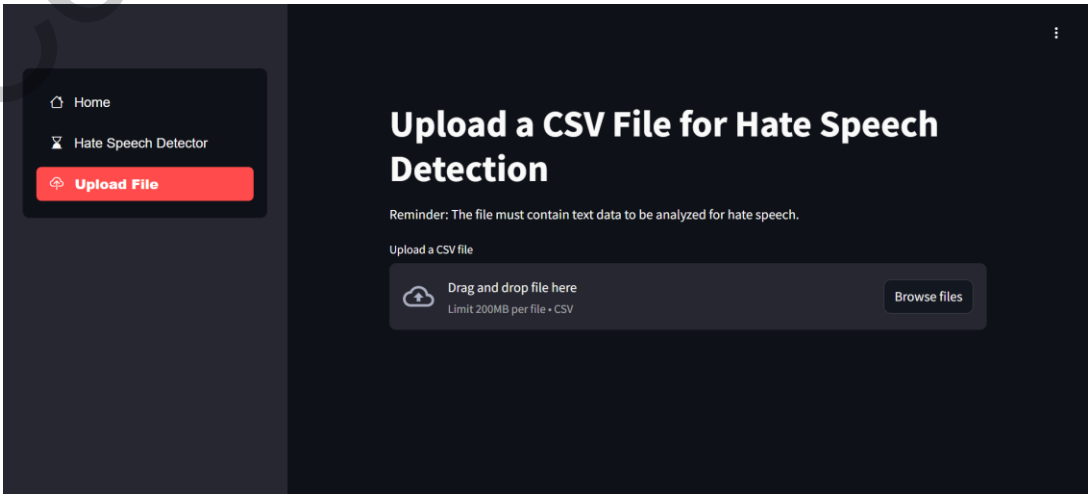
Rajah 7 Paparan Alat Pengesan Ucapan Kebencian

Rajah 8 menunjukkan contoh input yang telah dimasukkan oleh pengguna, kemudian input tersebut dianalisis mengeluarkan output samada teks yang dimasukkan tersebut mempunyai unsur kebencian atau sebaliknya.



Rajah 8 Paparan Output Alat Pengesan Ucapan Benci

Rajah 9 menunjukkan tambahan fungsi untuk memuatnaik fail bagi memudahkan pengguna untuk menganalisis keseluruhan teks yang terdapat dalam sebuah fail manakala Rajah 10 pula memaparkan keputusan analisisnya dalam bentuk fail yang boleh dimuat turun oleh pengguna sendiri.



Rajah 9 Paparan Bahagian Fungsi Muatnaik Fail

	text	preprocessed_text	prediction
0	Men are pigs and should be eliminated.	men be pig and should eliminate	Hate_Speech
1	"You're such a nigger, why don't you just disappear?	youre such nigger why dont you just disappear	Hate_Speech
2	All MUSLIMS are terrorists, they should be wiped out!	all muslim be terrorist they should wipe out	Hate_Speech
3	I can't believe those people are allowed to live here. They're nothing but a bunch of c	cant believe those people be allow live here theyre nothing but bunch dirty immigrar	Hate_Speech
4	Women are just objects for men's pleasure, they should stay in the kitchen where the	woman be just object for men pleasure they should stay the kitchen where they belo	Hate_Speech
5	Only poor people shop at discount stores; they have no taste.	only poor people shop discount store they have taste	None_Hate
6	All Israelis are criminals and should be deported. SHAME ON YOU !!!	all israeli be criminal and should deport shame you	Hate_Speech
7	People who support Viden are idiots and should not be allowed to vote.	people who support viden be idiot and should not allow vote	Hate_Speech
8	No one should hire WOMEN, they are not fit for any job	one should hire woman they be not fit for any job	Hate_Speech
9	Guys guys i need to stress that this is the legacy of white supremacy, clearly, a black r	guy guy need stress that this the legacy white supremacy clearly black man harrassin	Hate_Speech
10	Girls claiming they're hot generally too over confident have an overinflated view of th	girl claim theyre hot generally too over confident have overinflated view themselves	None_Hate
11	The presence of black people is a threat to our jobs.	the presence black people threat our job	Hate_Speech
12	Immigrants from Rohingya are thieves and should go back to where they came from	immigrant from rohingya be thief and should back where they come from	None_Hate
13	All man are the same and can't be trusted	all man be the same and cant trust	Hate_Speech

Rajah 10 Paparan output fail yang telah dianalisis

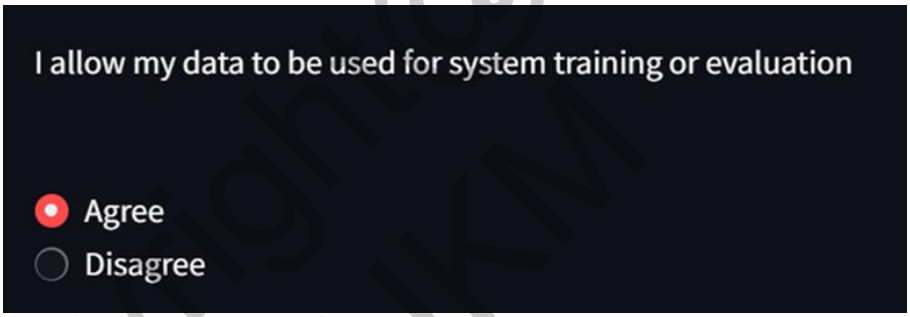
Pelan pengujian yang dijalankan dalam kajian ini merangkumi objektif pengujian, jenis pengujian yang dijalankan dan kriteria penamatan yang telah ditetapkan untuk memastikan keberkesanan model pengesan ucapan kebencian yang telah dibangunkan.

Terdapat tiga objektif utama pengujian yang telah ditetapkan dalam kajian ini antaranya ialah untuk memastikan model pengesan ucapan kebencian berfungsi dengan baik. Ini bermakna sistem yang dibangunkan haruslah dapat mengesan dan mengkategorikan teks input dengan tepat sama ada ianya ucapan kebencian atau tidak. Selain itu, pengujian juga bertujuan untuk memastikan kedua-dua model pembelajaran mesin dan model antara muka pengguna disepadukan dengan lancar. Penyepaduan antara model pembelajaran mesin, yang bertanggungjawab untuk analisis data, dan antara muka pengguna, yang menyediakan platform interaksi pengguna, adalah penting untuk memastikan keseluruhan sistem berfungsi dengan baik. Ujian ini memastikan tiada masalah teknikal berlaku apabila pengguna berinteraksi dengan sistem dan hasil analisis daripada model pembelajaran mesin dipaparkan dengan betul melalui antara muka pengguna. Objektif yang seterusnya adalah untuk memastikan papan pemuka yang dibangunkan memenuhi kedua-dua keperluan fungsian dan bukan fungsian. Keperluan fungsian merujuk kepada aspek teknikal sistem seperti keupayaan untuk mengesan ucapan kebencian, dan ketepatan hasil analisis. Keperluan bukan fungsian pula merangkumi aspek seperti kebolehgunaan, keselamatan, dan prestasi sistem secara keseluruhan. Dengan memastikan papan pemuka memenuhi semua keperluan ini, kajian ini dapat menjamin bahawa sistem yang dibangunkan adalah praktikal dan berkesan untuk digunakan dalam situasi sebenar. Pengujian tersebut memfokuskan pada dua kategori utama iaitu pengujian fungsian dan bukan fungsian.

Pengujian fungsian dilakukan untuk memastikan sistem berfungsi seperti yang diharapkan berdasarkan keperluan yang telah ditetapkan. Antara teknik pengujian fungsian adalah dari segi proses pengesanan ucapan kebencian. Ini bermakna pengujian fungsian dilakukan untuk menguji sama ada sistem dapat mengesan ucapan kebencian dalam teks input dengan betul. Contohnya, apabila pengguna memasukkan teks “saya tidak suka perempuan yang tidak bekerja”, sistem menganalisis teks tersebut dan mengeluarkan output “Teks ini mengandungi ucapan kebencian” sebaliknya apabila teks yang dimasukkan tidak

mengandungi sebarang unsur kebencian, sistem akan memberikan output “Teks ini tidak mengandungi ucapan kebencian”. Selain itu, Analisis data bagi setiap model dilakukan untuk merekod ketepatan, kepersisan, dapatan semula dan nilai F1-Skor. Dari segi pengujian prestasi pula, ujian prestasi dilakukan untuk menilai prestasi sistem dari segi masa tindak balas dan ketepatan dalam pelbagai keadaan. Sebagai contoh, apabila pengguna memasukkan input sama ada berbentuk fail atau teks, sistem segera bertindak balas untuk menganalisis dan memberikan output hanya dalam tempoh masa 3 saat sahaja.

Pengujian bukan fungsian pula dijalankan untuk memastikan sistem memenuhi keperluan yang berkaitan dengan cara sistem beroperasi. Pengujian kebolehgunaan adalah antara contoh teknik pengujian bukan fungsian yang telah dilakukan. Ujian kebolehgunaan dilakukan untuk menilai sama ada papan pemuka yang dibangunkan mudah untuk difahami dan digunakan. Jenis pengujian ini memastikan bahawa papan pemuka yang dibangunkan adalah mesra pengguna. Selain itu, pengujian dari segi etika dan keselamatan pula, Sistem ini meminta pengguna untuk memberikan persetujuan terlebih dahulu sebelum data mereka digunakan untuk tujuan pelatihan dan evaluasi sistem. Rajah 11 menunjukkan mesej yang diberikan kepada pengguna sebelum mereka dapat meneruskan penggunaan alat pengesanan ucapan kebencian.



Rajah 11 Mesej Persetujuan

Kriteria untuk menamatkan ujian ini antaranya adalah semua keperluan fungsian dan bukan fungsian dipenuhi. Selain itu, prestasi sistem yang dibangunkan memenuhi piawaian yang telah ditetapkan.

Cadangan Penambahbaikan

Untuk penambahbaikan pada masa hadapan, beberapa cadangan disyorkan antaranya adalah penggunaan dataset yang lebih besar dan pelbagai yang merangkumi pelbagai bahasa serta dialek akan meningkatkan kebolehan model dalam mengesan ucapan kebencian. Selain itu, penyepaduan pembelajaran mendalam seperti Rangkaian Neural Konvolusi (CNN) dan Rangkaian Neural Rekuren (RNN) dicadangkan untuk menangani kerumitan tafsiran konteks dengan lebih baik.

KESIMPULAN

Kajian ini berjaya mencapai matlamat utamanya dalam mengesan ucapan kebencian di platform media sosial Twitter. Tiga model pembelajaran mesin iaitu Regresi Logistik, Mesin Vektor Sokongan dan Bayes Naif telah digunakan untuk proses analisis dan dibandingkan dari segi ketepatan, kepersisan, dapatan semula dan F1-Skor. Model Regresi Logistik mencapai prestasi terbaik dengan ketepatan 84.99% selepas teknik SMOTE digunakan untuk menangani ketidakseimbangan data. Ini menunjukkan kekuatan projek dalam menghasilkan model yang boleh mengesan ucapan kebencian dengan lebih cekap dan berkesan. Bermula daripada fasa pengumpulan data sehinggalah fasa pengujian, projek ini telah memenuhi kesemua keperluan dan jangkaan yang telah ditetapkan. Objektif utama projek ini iaitu mengenalpasti dan menggunakan pengekstrakan ciri bagi mengesan kata benci, membangunkan beberapa model pengelas bagi pengelasan kata benci untuk menilai ketepatan model berdasarkan pengekstrakan ciri dan melakukan perbandingan model pengelas yang dapat memberikan keputusan terbaik melalui pra-pemprosesan data telah dicapai melalui pemilihan dan pengujian model pembelajaran mesin yang telah dibangunkan. usaha yang diletakkan dalam menyiapkan projek pengesanan ucapan kebencian ini diharapkan dapat memberi manfaat kepada masyarakat dengan mewujudkan persekitaran digital yang lebih damai dan selamat. Secara keseluruhannya, kajian ini telah membantu dalam mengenal pasti ucapan kebencian di media sosial. Diharapkan pada masa hadapan, kajian ini akan terus dikembangkan untuk mencegah penyebaran kebencian di platform media sosial berdasarkan kekuatan yang dimiliki oleh setiap model pembelajaran mesin yang digunakan dan cadangan untuk penambahbaikan.

Kekuatan Sistem

Sistem pembelajaran mesin yang dibangunkan untuk mengesan ucapan kebencian di media sosial mempunyai beberapa kekuatan yang ketara. Pertama, model Regresi Logistik menunjukkan prestasi terbaik dalam mengesan ucapan kebencian dengan ketepatan 84.99%, kepersisan 84.76%, dan dapatan semula yang tinggi. Penggunaan teknik SMOTE untuk menangani ketidakseimbangan data membuktikan keberkesanan dalam meningkatkan prestasi model. Selain itu, integrasi antara model pembelajaran mesin dan papan pemuka yang dibangunkan menggunakan "Streamlit Python" memastikan proses analisis data dijalankan dengan tepat dan efisien. Gabungan pendekatan bigram dan leksikon juga membantu dalam analisis data, menjadikan sistem ini lebih cekap dan berkesan dalam mengenalpasti ucapan kebencian di media sosial.

Kelemahan Sistem

Walaupun sistem ini menunjukkan prestasi yang baik, terdapat beberapa kekangan yang dihadapi sepanjang kajian ini. Salah satu kekangan utama adalah masalah data yang dikumpulkan mengandungi teks hingar iaitu tidak tersusun, penggunaan dialek dan istilah tempatan menyukarkan proses pengesanan ucapan kebencian. Ucapan kebencian boleh membawa makna yang berbeza bergantung kepada konteksnya, yang memerlukan model untuk memahami konteks ayat dengan lebih mendalam. Ketidakseimbangan antara kelas data

(ucapan kebencian dan bukan kebencian) juga menyebabkan model cenderung berat sebelah terhadap kelas majoriti, walaubagaimanapun ianya telah diatasi dengan penggunaan teknik SMOTE. Selain itu, pengujian prestasi sistem menunjukkan bahawa masa tindak balas perlu dipertingkatkan untuk memastikan analisis dan pemberian output dilakukan dengan lebih cepat, terutamanya apabila input berbentuk fail dimuatnaik oleh pengguna. Walaupun sistem ini meminta persetujuan pengguna sebelum data mereka digunakan, aspek etika dan keselamatan masih perlu diperhalusi untuk melindungi privasi pengguna dengan lebih baik .

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 Pengesanan Ucapan Kebencian di Media Sosial Menggunakan Pembelajaran Mesin.

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.

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.

Sekian, terima kasih.

RUJUKAN

Ali, M.Z., Ehsan-Ul-Haq, Rauf, S., Javed, K., & Hussain, S. 2021. Improving Hate Speech Detection of Urdu Tweets Using Sentiment Analysis. IEEE Access, 9, 84296–84305. https://doi.org/10.1109/access.2021.3087827

Alshalan, R., & Al-Khalifa, H. 2020. A Deep Learning Approach for Automatic Hate Speech Detection in the Saudi Twittersphere. Applied Sciences, 10(23), 8614. https://doi.org/10.3390/app10238614

Asogwa, D., Chukwuneke, C., Ngene, C.C., & Anigbogu, G. 2022. Hate Speech Classification Using MVS and Naive BAYES. https://www.semanticscholar.org/paper/Hate-Speech-Classification-Using-MVS-and-Naive-Asogwa-Chukwuneke/2cfa6331bb80c7cb6ffd513b0156167c95ffac6a

Furqan, D., Munirah, M., & Rosdiana, R. 2022. Analisi Bentuk Tuturan Kejahatan Tanpa Sempadan Berbahasa (Defamasi) dalam Sosial Media Youtube https://p3i.my.id/index.php/konsepsi/article/view/201

Gaydhani, A., Doma, V., Kendre, S., & Bhagwat, L. 2018. Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TFIDF based Approach. https://arxiv.org/abs/1809.08651

Omar, N. & Tham, J.V. 2019. *Perihal Ucapan Kebencian di Malaysia Hari Ini*. (n.d.). The Centre. https://www.centre.my/post/perihal-ucapan-benci-malaysia

Nur Najihah Binti Sa'rani (A188039)
Prof. Madya Dr. Nazlia Binti Omar
Fakulti Teknologi & Sains Maklumat
Universiti Kebangsaan Malaysia