PERAMALAN JENIS PERSONALITI MYERS-BRIGGS TYPE INDICATOR (MBTI) MELALUI PEMBELAJARAN MESIN

TAN KE YING

NAZLIA BINTI OMAR

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

ABSTRAK

Kajian ini bertujuan membangunkan model pembelajaran mesin bagi meramalkan jenis personaliti MBTI individu berdasarkan data teks yang diperoleh daripada media sosial seperti Twitter. Dengan menggunakan teknik pemprosesan bahasa tabii (PBT), model ini dapat mengenal pasti corak linguistik yang berkaitan dengan keempat-empat dimensi dalam MBTI. Metodologi kajian ini dibahagikan kepada enam fasa utama, iaitu: (1) pengumpulan data, di mana data teks dikumpulkan daripada sumber media sosial; (2) pemprosesan teks, yang melibatkan pembersihan dan penyediaan data; (3) pengayaan dan penyeimbangan data, bagi menambah kepelbagaian dan saiz data dengan kaedah seperti pertukaran sinonim dan penjanaan ayat baharu, di samping menangani isu ketidakseimbangan kelas dalam set data; (4) pengekstrakan ciri, untuk mengenal pasti ciri linguistik yang relevan; (5) pembelajaran, di mana model dilatih untuk memahami pola dalam data dan seterusnya mengklasifikasikan jenis personaliti MBTI; dan (6) penilaian prestasi model menggunakan metrik seperti ketepatan, kepersisan, dapatan dan skor F1. Hasil kajian menunjukkan bahawa model yang dibangunkan mampu mengklasifikasikan jenis personaliti MBTI dengan tahap ketepatan yang tinggi. Ini membuktikan potensi pendekatan analisis teks automatik dalam memahami keperibadian individu melalui data terbuka di media sosial.

PENGENALAN

Myers-Briggs Type Indicator (MBTI) merupakan satu instrumen psikologi yang digunakan secara meluas untuk menguji keperibadian, potensi dan minat individu. Berdasarkan teori psikologi Carl Jung, MBTI mengklasifikasikan personaliti kepada 16 jenis yang berbeza seperti ISTJ, INTJ, ESFJ dan ENFP, berasaskan empat dimensi utama iaitu Ekstrovert (E) atau Introvert (I), Deria (S) atau Intuisi (N), Pemikiran (T) atau Perasaan (F), serta Penilaian (J) atau Persepsi (P). Penilaian MBTI ini membolehkan individu memahami tingkah laku,

pembangunan diri dan potensi kerjaya masing-masing, malah tahap ketepatannya dikatakan mencapai sehingga 90%.

Walau bagaimanapun, kaedah tradisional untuk menentukan MBTI biasanya memerlukan individu untuk menjawab soal selidik yang panjang berdasarkan pengalaman dan pemikiran harian mereka. Kaedah ini kemungkinan menyebabkan hasil penilaian yang tidak konsisten dan kurang tepat akibat faktor luaran seperti keadaan emosi, kejujuran dalam menjawab serta tahap fokus seseorang individu semasa menjawab soalan. Tambahan pula, sebahagian individu mungkin kurang berminat untuk mengambil ujian tersebut kerana memerlukan masa yang lama dan usaha yang tinggi. Dalam masa yang sama, peningkatan sosial media seperti *Twitter*, *Facebook* dan *Instagram* telah membuka ruang untuk pengguna meluahkan perasaan emosi dan tingkah laku melalui teks. Namun begitu, teks data yang berskala besar menjadikan analisis manual tidak praktikal dan memerlukan teknik seperti Pemprosesan Bahasa Tabii (PBT) untuk membantu membuat analisis data secara automatik dan berkesan bagi data yang berskala besar.

Kajian ini dijalankan bagi mengekstrakan dan membandingkan ciri linguistik dalam data teks daripada media sosial untuk meramalkan jenis personaliti MBTI. Sebelum menjalankan proses pengekstrakan ciri linguistik, data teks perlu melalui proses pembersihan dan penyediaan data yang melibatkan penyingkiran unsur yang tidak relevan seperti simbol, URL, emoji, nombor dan kata henti. Seterusnya, proses pengayaan dan penyeimbangan data dijalankan untuk menambah variasi serta memastikan data adalah seimbang antara kelas personaliti. Selain itu, kajian ini juga bertujuan untuk mengenal pasti dan menilai prestasi beberapa model pembelajaran mesin, termasuk model utama yang dipilih, dalam usaha menentukan pendekatan paling berkesan untuk pengelasan jenis personaliti MBTI. Penggunaan PBT dalam kajian ini membolehkan corak linguistik dan ciri-ciri psikologi daripada pos media sosial dikenalpasti secara sistematik dan konsisten yang menyumbang kepada pemahaman lebih mendalam tentang personaliti seseorang individu.

Tuntasnya, kajian ini berhasrat untuk membangunkan sebuah model klasifikasi MBTI yang menggunakan pendekatan pembelajaran mesin d an PBT untuk menjadikan proses ramalan personaliti MBTI dengan lebih efisien, tepat dan automatik. Pendekatan ini mampu menangani kelemahan kaedah traditional dan menawarkan sistem analisis personaliti individu yang lebih canggih dan berkesan. Dengan memanfaatkan data terbuka daripada media sosial, kajian ini membuka ruang baharu dalam penyelidikan teknologi dan pemahaman keperibadian manusia secara berskala besar dan masa nyata.

METODOLOGI KAJIAN

Metodologi yang digunakan dalam pembangunan projek ini melibatkan enam fasa utama iaitu pengumpulan data, pra-pemprosesan data, pengayaan data dan penyeimbangan data, pengekstrakan ciri, pembelajaran mesin, serta penilaian prestasi model. Teknik Pemprosesan Bahasa Tabii (PBT) digunakan untuk membersihkan dan menyediakan data teks yang diperoleh

daripada media sosial, manakala algoritma pembelajaran mesin digunakan bagi melatih model mengenal pasti corak linguistik yang berkaitan dengan personaliti MBTI. Pendekatan ini membolehkan analisis data teks dilakukan secara automatik dan sistematik, sekali gus meningkatkan ketepatan dan keberkesanan ramalan personaliti.

Fasa Pengumpulan Data

Fasa ini melibatkan proses pencarian, pengumpulan dan penggabungan data yang mempunyai label MBTI personaliti untuk tujuan latihan model pembelajaran mesin. Dalam kajian ini, data dikumpulkam daripada repositori *GitHub* penyelidik Data Sains, Syed Muhammad Hamza, yang telah mengumpulkan data teks berdasarkan 16 jenis personaliti MBTI daripada perbincangan pengguna di platform *Reddit*. Set data ini dipilih kerana mengandungi label personaliti yang jelas yang sesuai dalam proses pengelasan teks. Keseluruhan data yang diperoleh mengandungi jumlah entri yang berbeza bagi setiap kelas MBTI. Jadual 1 menunjukkan jumlah entri yang tidak seimbang antara kelas personaliti ayng terdapat dalam set data tersebut.

Jenis MBTI	Jumlah Entri
ISTJ	2,600
INFJ	10,000
INTJ	6,400
ENFJ	6,600
ISTP	9,200
ESFJ	800
INFP	8,600
ESTP	830
ENFP	1,200
ESTP	1,700
ESTJ	700
ENTJ	9,000
INTP	12,000
ISFJ	4,400
ENTP	7,600
ISFP	4,100

Jadual 1 Jumlah Entri Set Data

Fasa Pra Pemprosesan Data

Fasa pra-pemprosesan data bertujuan untuk membersihkan dan menyediakan data teks sebelum proses pembelajaran mesin bagi memastikan set data yang digunakan adalah bersih, konsisten dan berkualiti. Proses ini bermula dengan penyingkiran lajur tidak relevan, penapisan label

berdasarkan 16 jenis personaliti MBTI, serta penyingkiran data duplikasi bagi mengelakkan model dilatih dengan berulang yang boleh menyebabkan berat sebelah atau *overfitting*. Seterusnya, proses ini juga melibatkan beberapa langkah penting seperti penyingkiran URL, tanda baca, nombor, baris baharu, aksara ASCII dan ruang kosong yang berlebihan. Set data ini hanya mengekalkan teks bahasa Inggeris manakala bahasa lain disingkirkan. Kata singkatan dan akronim seperti "*I'll*", "*should've*", "*lol*" dan "*asap*" turut dikembangkan kepada bentuk penuh bagi memastikan keseragaman dan kefahaman teks. Proses tokenisasi juga dilaksanakan untuk memcahkan teks kepada untuk kecil iaitu perkataan untuk memudahkan proses analisis model pembelajaran mesin. Di samping itu, setiap label MBTI ditukar kepada bentuk berangka melalui proses pengekodan. Langkah-langkah ini menggunakan teknik Pemprosesan Bahasa Tabii (PBT) bagi meningkatkan kecekapan dan memudahkan model dalam memahami serta membuat analisis corak linguistik yang terdapat dalam data teks. Jadual 2 merupakan contoh bagi proses yang dijalankan.

Teks asal	"*Good one _____ https://www.youtube.com/watch?v=fHiGbolFFGw||| Of course, to which say I know; that's my blessing and my curse.||| Does being absolutely positive that you and your best friend could be an amazing couple count? If so, than yes.*"
Teks setelah penyingkiran URL, tanda baca, baris dan kata henti	*"Good one course say know thats blessing curse absolutely positive best friend could amazing couple count yes*"
Teks setelah penyeragaman huruf kecil	"*good one course say know thats blessing curse absolutely positive best friend could amazing couple count yes*"
Tokenisasi	"[*'good'*, *'one'*, *'course'*, *'say'*, *'know'*, *'thats'*, *'blessing'*, *'curse'*, *'absolutely'*, *'positive'*, *'best'*, *'friend'*, *'could'*, *'amazing'*, *'couple'*, *'count'*, *'yes'*]"

Jadual 2 Langkah Pra-pemprosesan Teks

Fasa Pengayaan dan Penyeimbangan Data

Fasa ini bertujuan untuk mengatasi masalah ketidakseimbagan data dalam kelas MBTI dengan dua pendekatan utama, iaitu pensampelan rawak secara pengurangan dan pengayaan data (*Data Augmentation*). Pensampelan rawak secara pengurangan melibatkan penyingkiran sebahagian data teks daripada kelas majoriti untuk mencapai keseimbangan dengan kelas majoriti. Seterusnya, pengayaan data secara manual dilaksanakan kepada kelas minoriti iaitu ISFP, ISFJ, ISTJ, ESFP, ESTP, ESFJ, dan ESTJ dengan menambah ayat-ayat baharu berdasarkan ciri-ciri personaliti harian. Tambahan pula, pengayaan automatik juga dijalankan menggunakan model *GPT-2* bagi menjana ayat tambahan yang mempunyai struktur ayat yang serupa dan maksud asalnya. Kedua-dua kaedah ini dapat membantu menghasilkan set data yang lebih seimbang dan sesuai untuk latihan model pembelajaran mesin.

Fasa Pengekstrakan Ciri dan Representasi Vektor

Dalam fasa ini, proses pengekstrakan ciri bertujuan untuk menukarkan set data yang berbentuk teks kepada bentuk vektor yang mudah difahami oleh model pembelajaran mesin. Proses ini juga berfungsi untuk menangkap makna semantik perkataan dalam data teks dan mengekstrak maklumat penting daripada teks. Tiga kaedah pengekstrakan ciri telah digunakan iaitu, Bag-of-Words (BoW), kekerapan jangka-kekerapan dokumen songsang (*Term Frequency-Inverse Document Frequency,* TF-IDF) dan *CountVectorizer* (CV).

Teknik BoW berfungsi mengira kekerapan bagi setiap perkataan yang terdapat dalam teks tanpa mengambil kira urutan, manakala TF-IDF mengira kekerapan perkataan dengan memberi pemberat kepada perkataan yang lebih penting dan bermakna dalam teks. Seterusnya, kaedah *CountVectorizer* ini mempertimbangkan setiap perkataan yang terdapat dalam teks sebagai ciri dan mengira bilangan kemunculan setiap perkataan.

Selain daripada proses pengekstraktan ciri, proses representasi vektor turut dijalankan dengan menggunakan teknik pembenaman kata, *Word2Vec*. Teknik ini mempelajari dan menangkap hubungan kontekstual antara perkataan daripada teks data dan memainkan peranan yang penting dalam menukarkan data yang berbentuk teks kepada bentuk vektor yang berdimensi rendah. Vektor yang dijana merupakan wakil kepada setiap makna semantik perkataan.

Vektor yang diperoleh daripada proses pengekstrakan ciri dan proses representasi vektor digabungkan bagi meningkatkan kualiti representasi ciri agar lebih kukuh dan bermakna. Pendekatan ini mampu memudahkan model pembelajaran mesin untuk mengenal pasti corak linguistik yang tersembunyi dengan lebih berkesan, dan meningkatkan ketepatan ramalan model.

Fasa Pembelajaran Mesin

Fasa pembelajaran mesin bertujuan untuk membangunkan sebuah sistem ramalan jenis personaliti MBTI secara automatik menggunakan dua pendekatan: (1) Peramalan ke atas empat dimensi MBTI secara serentak, dan (2) empat pengelas binari yang mewakili setiap dimensi secara berasingan iaitu: Introvert/Ekstrovert (I/E), Intuisi/Penderiaan (N/S), Perasaan/Pemikiran (F/T) serta Penghakiman/Pemahaman (J/P). Model pembelajaran mesin utama yang digunakan dalam kajian ini ialah *XGBoost* dan model Regresi Logistik untuk proses peramalan empat pengelas binari yang mewakili setiap dimensi secara berasingan. Model *XGBoost* merupakan pengelasan berasaskan pohon pokok keputusan yang cekap dan mampu memberikan prestasi yang tinggi bagi masalah pengelasan kelas serta binari. Manakala, model Regresi Logistik merupakan model pengelasan yang sering digunakan semasa menjalankan klasifikasi teks dalam pembelajaran mesin dan boleh membantu dalam mengelaskan setiap dimensi MBTI berdasarkan ciri teks yang telah diekstrakan.

Set data telah dibahagikan kepada set latihan dan set ujian dalam nisbah 80:20 bagi memastikan proses pengujian model dijalankan secara adil dan tidak berat sebelah. Bagi pendekatan kedua, iaitu peramalan empat dimensi MBTI secara berasingan, kelas positif telah

ditetapkan bagi setiap dimensi untuk memastikan ketepatan semasa latihan dan penilai model klasifikasi binari.

Fasa Penilaian

Dalam fasa penilaian, prestasi model ramalan diukur untuk menilai keberkesanan dan ketepatan model semasa menjalankan analisis dan ramalan personaliti MBTI. Empat metrik utama yang digunakan ialah ketepatan, kepersisan, dapatan dan skor-F1. Ketepatan (*Accuracy*) berfungsi untuk mengukur peratusan ramalan yang betul daripada keseluruhan ramalan. Kepersisan (*Percision*) menilai ramalan positif yang dibuat oleh model adalah tepat. Dapatan (*Recall*) mengukurkan keupayaan model dalam mengenal pasti kes positif sebenar. Akhir sekali, skor-F1 (*F1-Score*) merupakan gabungan kepersisan dan dapatan yang memberikan gambaran prestasi model secara menyeluruh dan metrik ini sesuai digunakan untuk perbandingan semasa menghadapi masalah data tidak seimbang. Rumus bagi empat metrik penilaian di atas adalah seperti:

$$Ketepatan = \frac{Jumlah\ peramalan\ yang\ benar}{Jumlah\ peramalan}$$

$$Kepersisan = \frac{Jumlah\ peramalan\ positif\ yang\ benar}{Jumlah\ peramalan\ positif\ dan\ negatif\ yang\ benar}$$

$$Ingat = \frac{Jumlah\ peramalan\ positif\ yang\ benar}{Jumlah\ peramalan\ positif\ benar\ dan\ negatif\ palsu}$$

$$Skor - F1 = 2 \times \frac{Kepersisan \times Ingat}{Kepersisan + Ingat}$$

KEPUTUSAN DAN PERBINCANGAN

Sistem ramalan jenis personaliti MBTI berdasarkan teks secara automatik telah berjaya dibangunkan. Dua algoritma telah digunakan iaitu *XGBoost* dan Regresi Logistik, dengan pelbagai teknik pengekstrakan ciri seperti *Bag-of-Words*, TF-IDF, dan *CountVectorizer* serta gabungan bersama *Word2Vec*. Semua model dinilai berdasarkan empat set data yang berbeza iaitu data *cleaned*, *balanced*, *augmented*, dan *augmented GPT-2*. Metrik skor-F1 telah dipilih sebagai ukuran utama dalam penilaian model untuk mengelakkan bias dan kesilapan semasa membuat perbandingan prestasi model terhadap set data yang tidak seimbang.

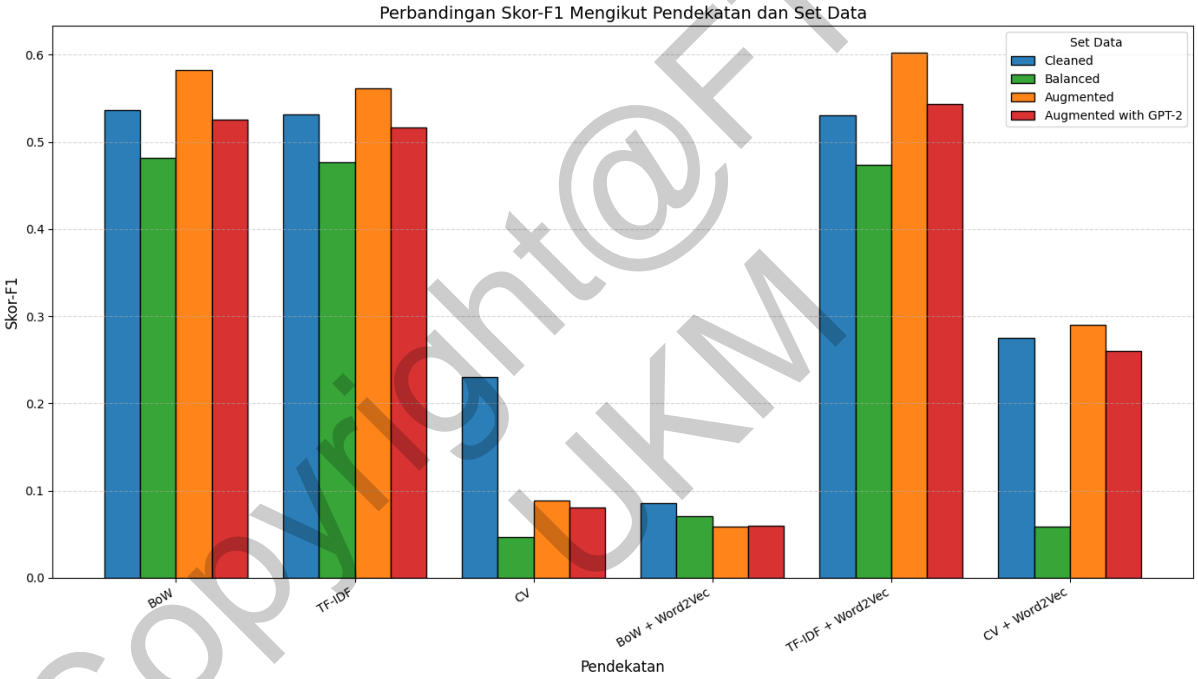
Keputusan Pengujian Model Ramalan Personaliti MBTI

Jadual 3 dan Rajah 1 memaparkan keputusan skor-F1 yang diperoleh daripada model latihan bagi peramalan ke atas empat dimensi MBTI secara serentak, menggunakan pendekatan

pengekstrakan ciri yang berbeza serta gabungannya, iaitu *Bag-of-Words* (BoW), TF-IDF, *CountVectorizer* (CV) serta gabungan masing-masing dengan Word2Vec.

Set Data	**Skor-F1**					
	Pendekatan					
	BoW	**TF-IDF**	**CV**	**BoW + Word2Vec**	**TF-IDF + Word2Vec**	**CV + Word2Vec**
Cleaned	0.5366	0.5313	0. 2299	0.0854	0.5305	0. 2754
Balanced	0.4816	0.4768	0. 0468	0.0709	0.4738	0. 0583
Augmented	0.5823	0.5620	0. 0888	0.0589	0.6029	0. 2906
Augmented with GPT-2	0.5254	0.5170	0. 0802	0.0598	0.5434	0. 2607

Jadual 3 Keputusan skor-F1 bagi peramalan ke atas empat dimensi MBTI secara serentak



Rajah 1 Keputusan skor-F1 bagi peramalan ke atas empat dimensi MBTI secara serentak

Dalam kajian ini, teknik BoW merupakan pendekatan tahap asas yang hanya mengira kekerapan kemunculan perkataan tanpa mengambil kira sususan atau konteks. Pendekatan BoW secara tunggal menunjukkan keputusan analisis yang terbaik berbanding dengan TF-IDF dan *CountVectorizer*. Penggunaan set data *Augmented* bersama teknik BoW mencatatkan skor-F1 sebangyak 0.5823, membuktikan keberkesanan BoW dalam menjalankan analisis peramalan jenis personaliti MBTI secara serentak.

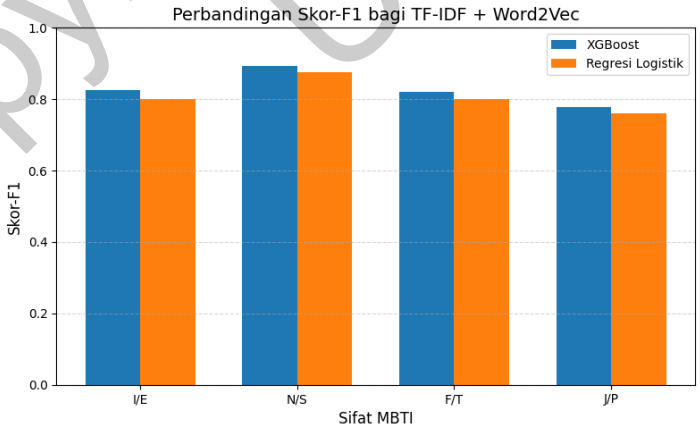
Namun, hasil analisis menunjukkan bahawa gabungan TF-IDF dan *Word2Vec* dengan set data *Augmented* mencatatkan Skor-F1 yang tertinggi iaitu 0.6029. Keputusan ini membuktikan keberkesanan pendekatan gabungan TF-IDF dan *Word2Vec* dalam menangkap makna semantik dan struktur linguistik dalam data teks dan membuat pengelasan jenis personaliti MBTI dengan lebih tepat.

Sebaliknya, gabungan BoW dengan Word2Vec memberikan prestasi paling rendah dalam semua set data, dengan skor-F1 0.0589 bagi set data *Augmented*. Keputusan yang rendah ini disebabkan oleh sifat BoW hanya mengambil kira kekerapan perkataan tanpa menimbangkan makna semantik teks, manakala *Word2Vec* berfungsi untuk memahami konteks semasa menjalankan analisis. Perbandingan ini membuktikan bahawa Word2Vec menyumbang kepada pengekstrakan ciri yang lebih kontekstual, dan meningkatkan keupayaan model dalam mengenalpasti ciri linguistik yang berkaitan dengan personaliti MBTI dalam data teks.

Pendekatan model latihan kedua melibatkan pembangunan empat model pengelasan binari, yang masing-masing mewakili satu daripada empat dimensi personaliti MBTI, iaitu Introvert/Ekstrovert (I/E), Intuisi/Penderiaan (N/S), Perasaan/Pemikiran (F/T), dan Penghakiman/Pemahaman (J/P). Pendekatan ini menggunakan gabungan teknik TF-IDF dan *Word2Vec* dengan dua algoritma pembelajaran mesin: *XGBoost* dan Regresi Logistik. Jadual 4 dan Rajah 2 memaparkan keputusan Skor-F1 yang diperoleh daripada model latihan berdasarkan empat sifat MBTI secara berasingan, menggunakan set data *Augmented* serta gabungan teknik pengekstrakan ciri TF-IDF dan *Word2Vec* dengan dua algoritma, iaitu *XGBoost* dan Regresi Logistik.

Sifat MBTI	**Skor-F1 (TD-IDF + Word2Vec)**	
	Algoritma	
	XGBoost	**Regresi Logistik**
I/E	0. 8250	0. 8003
N/S	0. 8922	0. 8764
F/T	0. 8195	0. 7997
J/P	0. 7784	0. 7597

Jadual 4 Keputusan skor-F1 bagi peramalan berdasarkan empat sifat MBTI secara berasingan



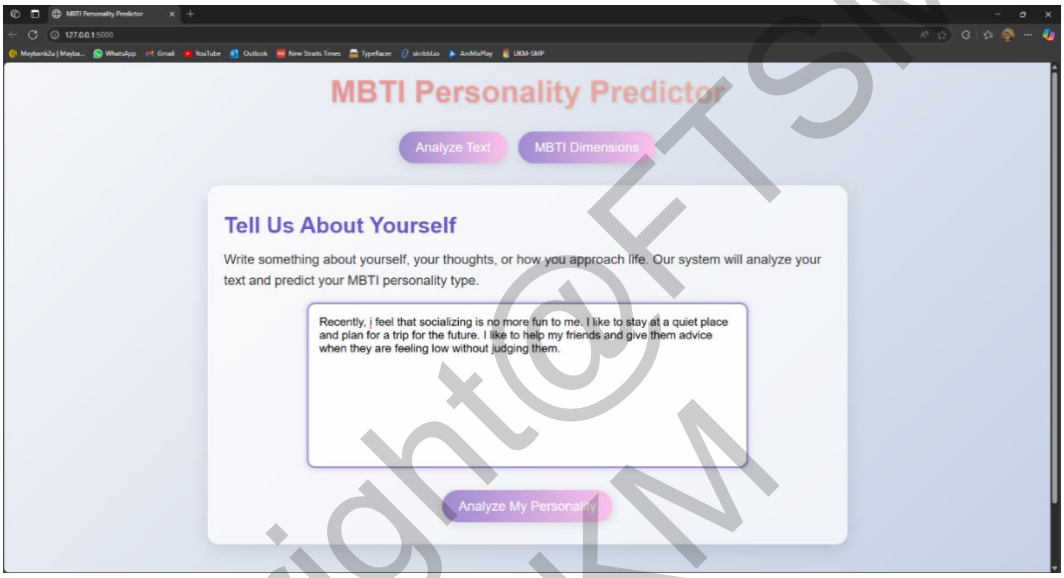
Rajah 2 Keputusan skor-F1 bagi peramalan berdasarkan empat sifat MBTI secara berasingan

Hasil pengujian secara keseluruhannya menunjukkan bahawa model *XGBoost* dengan pendekatan gabungan TF-IDF dan *Word2Vec* memberikan prestasi yang paling tinggi berbanding model Regresi Logistik bagi keempat-empat dimensi, dengan perbezaan Skor-F1 yang konsisten. Dimensi N/S (Intuisi/Penderiaan) mencatatkan Skor-F1 tertinggi iaitu 0.8922 dengan *XGBoost*, diikuti oleh dimensi I/E (0.8250), F/T (0.8195) dan J/P (0.7784). Keputusan

ini menunjukkan bahawa pendekatan pengelasan binari bagi setiap dimensi MBTI lebih berkesan berbanding peramalan serentak ke atas 16 jenis personaliti.

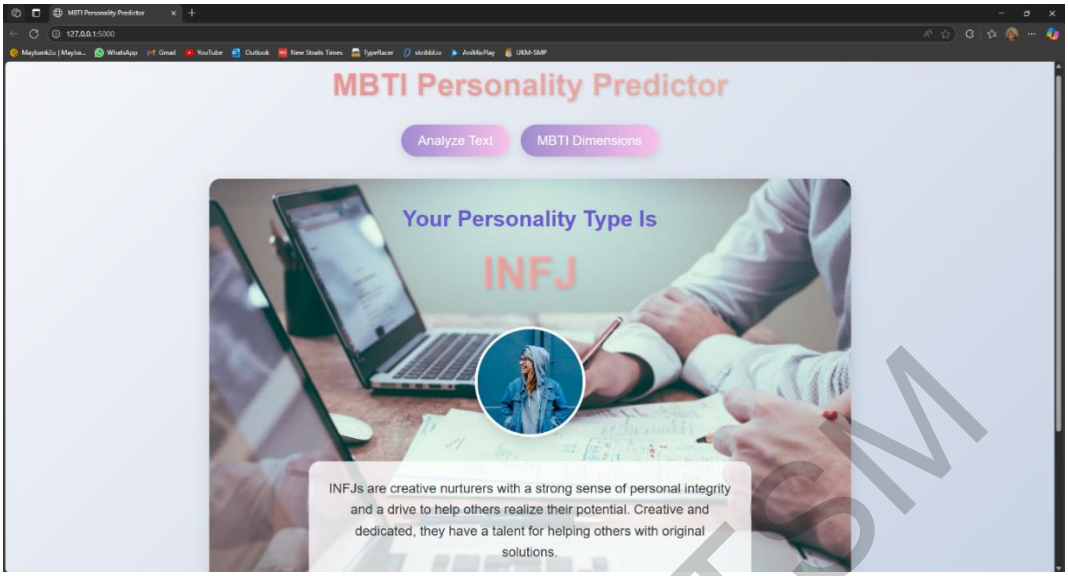
Keputusan Pengujian Sistem Antaramuka

Rajah 3, Rajah 4 dan Rajah 5 menggambarkan antaramuka sistem ramalan jenis personaliti MBTI yang dibangunkan. Antaramuka ini merangkumi halaman utama sistem, paparan keputusan jenis personaliti yang dijana oleh model, serta penerangan ringkas mengenai setiap dimensi personaliti MBTI.



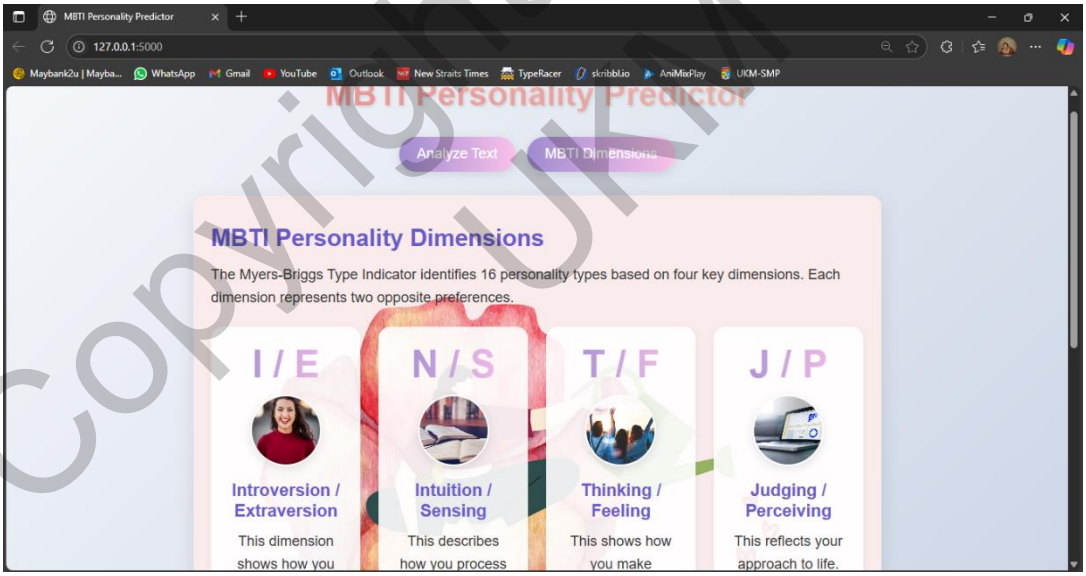
Rajah 3 Antaramuka pengguna memasukkan teks

Berdasarkan antaramuka Rajah 3 merupakan halaman di mana pengguna boleh memasukkan teks untuk dianalisis. Selepas pengguna memasukkan teks dan menekan butang "*Analyze My Personality*", sistem akan menghantar input tersebut kepada model pembelajaran mesin. Model akan membuat analisis corak linguistik berdasarkan teks untuk meramalkan jenis personaliti MBTI yang paling hampir dengan gaya bahasa pengguna. Keputusan ramalan kemudiannya dipaparkan di halaman seterusnya.



Rajah 4 Antaramuka penerangan ringkas hasil ramalan MBTI

Rajah 4 menggambarkan antaramuka selepas proses analisis, di mana jenis personaliti MBTI pengguna dipaparkan dengan saiz besar dan warna terang untuk menarik perhatian. Selain itu, di bawah MBTI mempunyai penerangan ringkas tentang ciri-ciri utama personaliti tersebut yang senang difahami oleh pengguna.



Rajah 5 Antaramuka penerangan empat dimensi MBTI

Rajah 5 menunjukkan antaramuka yang menerangkan empat dimensi utama MBTI: I/E, N/S, T/F dan J/P. Setiap dimensi dijelaskan secara terperinci bagi membantu pengguna memahami maksud dan ciri keperibadian di sebalik setiap huruf. Penerangan merangkumi aspek seperti gaya komunikasi, kekuatan, kelemahan, dan interaksi sosial.

Cadangan Penambahbaikan

Terdapat beberapa masalah yang timbul semasa menyiapkan kajian ini, antaranya ialah set data yang tidak seimbang, yang menyebabkan model cenderung untuk membuat ramalan terhadap kelas majoriti sahaja. Selain itu, teknik pengekstrakan ciri yang digunakan kurang berkemampuan untuk menangkap makna dan hubungan semantik dalam teks secara tepat dan menyeluruh, terutamanya apabila menggunakan pendekatan asas seperti TF-IDF dan *Word2Vec*. Di samping itu, model pembelajaran mesin yang kurang berkemampuan dalam membuat analisis teks yang panjang dan berskala besar, terutamanya dari segi memahami konteks ayat yang kompleks dan hubungan antara perkataan dalam teks.

Langkah utama untuk menambahbaikan sistem ini ialah penambahan data yang berkualiti tinggi. Ini boleh dilaksanakan dengan mengumpulkan data daripada pelbagai media sosial seperti *Twitter*, *Instagram* dan *Facebook* serta memastikan setiap kelas MBTI adalah seimbang. Seterusnya, proses pra-pemprosesan data yang lebih terperinci seperti penapisan data bising dan normalisasi memastikan data adalah bersih.

Langkah seterusnya ialah penggunaan model pemprosesan bahasa tabii (PBT) yang lebih canggih seperti BERT, GPT atau T5. Model-model ini boleh meningkatkan kemampuan model dalam memahami konteks semantik dan hubungan antara setiap perkataan yang terdapat dalam teks. Penggunaan model ini boleh membantu model semasa proses analisis dengan mengurangkan masalah bias dan meningkatkan generalisasi model terhadap pelbagai jenis teks.

Ketiga, penerapan teknik pembelajaran mendalam (*deep learning*) boleh dipertimbangkan untuk memantapkan lagi prestasi sistem. Model seperti Rangkaian Neural Berulang (RNN), Memori Jangka Panjang dan Pendek (LSTM), serta transformer berupaya menangkap struktur linguistik yang kompleks dalam data teks. Model-model ini sesuai digunakan untuk menangani tugas pengelasan pelbagai kelas seperti MBTI yang melibatkan 16 jenis personaliti.

Tuntasnya, langkah-langkah ini berpotensi untuk meningkatkan ketepatan, kebolehpercayaan dan kebersanan sistem dalam meramalkan jenis personaliti MBTI berdasarkan teks pengguna.

KESIMPULAN

Secara keseluruhannya, projek bertajuk *Peramalan Jenis Personaliti Myers-Briggs Type Indicator (MBTI) melalui Pembelajaran Mesin* ini telah berjaya mencapai ketiga-tiga objektif yang ditetapkan. Sistem ramalan jenis personaliti MBTI melalui pembelajaran mesin telah berjaya dibina menggunakan teknik pengekstrakan ciri dan model ramalan yang sesuai. Walaupun menghadapi beberapa kekangan seperti ketidakseimbangan data, masalah tersebut telah berjaya diatasi dengan melaksanankan pendekatan yang sesuai. Projek ini juga membuka jalan untuk kajian-kajian lanjut dalam bidang pemprosesan bahasa tabii, pengelasan personaliti secara automatik, serta pembangunan sistem pintar berasaskan teks.

Kekuatan Sistem

Kekuatan sistem ini mampu meramalkan jenis personaliti MBTI secara tepat dan berkesan dengan penggunaan model *XGBoost* bersama pendekatan TF-IDF dan *Word2Vec*. Pendekatan klasifikasi binari membolehkan pemisahan yang jelas bagi setiap dimensi MBTI, dengan ketepatan sekitar 75% hingga 85%. Antaramuka pengguna yang dibina menggunakan HTML dan CSS juga mesra pengguna dan interaktif. Pengguna hanya perlu memasukkan teks, kemudian, model akan menjana hasil ramalan MBTI yang hampir dengan makna teks dengan penerangan yang mudah difahami.

Kelemahan Sistem

Antara kekangan utama sistem ialah ketidakseimbangan data bagi setiap kelas MBTI, yang menyebabkan masalah bias semasa model latihan di mana model cenderung untuk memihak kepada dimensi tertentu. Selain itu, teknik pengekstrakan ciri seperti TF-IDF yang digunakan masih bersifat asas dan tidak mampu menangkap corak semantik yang kompleks dalam teks. Ini menyebabkan keupayaan model menurun dalam memahami perbezaan antara personaliti yang berbeza.

PENGHARGAAN

Penulis kajian ini ingin merakamkan setinggi-tinggi penghargaan dan jutaan terima kasih kepada Prof. Madya Dr. Nazlia Binti Omar, penyelia projek ini, atas segala tunjuk ajar, bimbingan serta dorongan yang telah diberikan sepanjang tempoh pelaksanaan projek. Komitmen dan kesabaran beliau dalam membimbing penulis amat dihargai dan menjadi pemangkin utama dalam kejayaan menyiapkan kajian ini.

Penulis kajian ini juga ingin mengucapkan terima kasih kepada semua pihak yang telah membantu secara langsung mahupun tidak langsung dalam menyempurnakan projek ini. Tidak dilupakan juga kepada pihak fakulti, rakan-rakan, serta ahli keluarga yang sentiasa memberikan sokongan moral dan teknikal. Segala bantuan yang telah dihulurkan amatlah dihargai dan semoga Tuhan membalas jasa baik semua pihak dengan sebaik-baik ganjaran.

RUJUKAN

Amir, M.S. 2021. *Python Scipy sparse matrices explained.* Sefidian Academy.

Amit, Y. 2024. *A Gentle Introduction to the Bag-of-Words Model*. Medium.

Anon. 2024. *NumPy: numpy.vstack() function*. w3resource.

Ashraf, N, Iqbal, R.S, Bano, S, Azeem, H.M & Naz, S. 2024. *Enhancing MBTI Personality Prediction from Text Data with Advance Word Embedding Techniques.* VFAST Transactions on Software Engineering.

Ashwin, K. 2024. *CountVectorizer and TfidfVectorizer For Beginner*. Medium.

Ashwin, K. 2024. *CountVectorizer vs TfidfVectorizer*. Dev.

Basto, C. 2021. *Extending the Abstraction of Personality Types based on MBTI with Machine Learning and Natural Language Processing*. Academia.edu.

Brussels, Kesatuan Eropah. 2024. *Apakah metrik penilaian?* Institut Persijilan Teknologi Maklumat Eropah - EITCI ASBL.

Carina, C & Donald, H.S. 2017 *Myers-Briggs Type Indicator*. The Science Direct.

Cassie, K. 2020. *Penjelasan paling mudah berkenaan pembelajaran mesin yang bakal dibaca oleh anda.* Terj. Nur Amalina Diyana Suhaimi & Md. Azhar Ibrahim. Medium.

Chakir, O, Rehaimi, A, Sadqi, Y, Alaoui, E.A.A, Krichen, M, Gaba, G.S & Gurtov, A. 2023. *An empirical assessment of ensemble methods and traditional machine learning techniques for web-based attack detection in industry 5.0*. ScienceDirect.

Cherukuru, R.K., Kumar, A, Srivastava, S & Verma, VK. 2022. *Prediction of Personality Trait using Machine Learning on Online Texts*, dalam Proceedings of the 2022 International Conference for Advancement in Technology (ICONAT), IEEE.

Chowanda, A, Suhartono, D, Andangsari, E.W & Zamlim K.Z. 2021. *Machine Learning Algorithms Exploration for Predicting Personality From Text*. Kuantan: Penerbit Universiti Malaysia Pahang.

Cohen, S. 2021. *The basics of machine learning: strategies and techniques*. Dlm. *Artificial Intelligence and Deep Learning in Pathology*, hlm. 13-40. Elsevier.

Cole, S & Jim, H. 2024. *What is NLP (natural language processing)?* IBM TechXchange Conference 2024.

Dylan, S. 2018. *Myers Briggs Personality Tags on Reddit Data Dataset*. Zenodo.

EITCI. 2024. *Apakah metrik penilaian.* Eitica.

Febriyanti, P. 2024. *Perbandingan Penggunaan TFIDFVectorizer, CountVectorizer, dan HashingVectorizer dengan Optimalisasi Parameter pada Machine Learning untuk Analisis Sentimen Pemilu 2024.* Jurnal Mahasiswa Teknik Informatika, JATI 8(4): 7413-7419.

Ghimire, K. 2021. *What is Word2Vec? How does it work? CBOW and Skip-gram* [Video]. YouTube.

Gregorius, R, Pricillia, K & Derwin, S. 2023. *MBTI Personality Prediction Using Machine Learning and SMOTE for Balancing Data Based on Statement Sentences*. Information, 14(4), 217.

Hui, J. 2019. *NLP – Word Embedding & GloVe*. Medium

Jessica, H & Kent, E. 2024. *Word embedding and classification methods and their effects on fake news detection*. Elsevier.

Justine, L, Evariste, NF & Silvia, B. 2024. *Benchmarking feature selection and feature extraction methods to improve the performances of machine-leaning algorithms for patient classification using metabolomics biomedical data.* Elsevier.

Kavita, G. N.d. *Word2Vec: A comparison between CBOW, SkipGram & SkipGramSI*. Kavita Ganesan.

Kendra, C. 2023. *How the Myers-Briggs Type Indicator Works*. Verywell mind.

Khan, A.S, Ahmad, H, Asghar, M.Z, Saddozai, F.K, Arif, A & Khalid, H.A. 2020. *Personality Classification from Online Text using Machine Learning Approach*. *International Journal of Advanced Computer Science and Applications*, 11(3).

Kim, H. 2022. *Predicting Myers-Briggs Personality Types by the Natural Language from Social Media Posts*. ResearchGate.

Manal, M & Nazlia, O. 2020. *Question classification based on Bloom's taxonomy cognitive domain using modified TF-IDF and word2vec*. ResearchGate.

Martelli, A. 2013. *Expanding English language contractions in Python*. Stack Overflow.

Matthew, U(pnyt.). 2024. *Bag-of-Words Model in NLP Explained*. Builtin.

Michal, K. 2021. *Pattern Recognition: Potential Anchoring for Imbalanced Data Classification*. ScienceDirect.

Mitchell, J. 2017. *(MBTI) Myers-Briggs Personality Type Dataset*. Kaggle.

Neri, V.O. 2023. *CountVectorizer Tutorial: How To Easily Turn Text Into Features For Any NLP Task*. Spot Intelligence.

Nisha, K, A, Kulsum, U, Rahman, S, Hossain, M.F, Chakraborty, P & Choudhury, T. 2021. *A Comparative Analysis of Machine Learning Approaches in Personality Prediction Using MBTI*. Computational Intelligence in Pattern Recognition.

Nivesh, T & Rashmi, P. 2023. Dlm. Recent Trends in Multidisciplinary Research. *Advancement in Personality Prediction System: Using The Myers-Briggs Type Indicator (MBTI) Dataset*, hlm 151-160. Red'shine Publication.

NumPy. t.th. *numpy.vstack*. NumPy documentation.

OpenAI. (2024). ChatGPT (Version 4.0) [Large language model]. OpenAI. https://chat.openai.com/

Optuna. 2024. *optuna.trial.Trial — Optuna documentation*. Tokyo: Preferred Networks.

Pankaj, B. 2024. *The Evolution of NLP: From Bag of Words to Large Language Models*. LinkedIn.

Partha, K. 2021. *Exploring Personality and Online Social Engagement: An Investigation of MBTI Users on Twitter*. arXiv.org.

Polipireddy, S & Rahul, K. 2021. *hyOPTXg: OPTUNA hyper-parameter optimization framework for predicting cardiovascular disease using XGBoost*. Elsevier.

Ranto, S, Incheon, P, & Ayato, K. 2021. *Sentence Augmentation for Language Translation Using GPT-2*. Electronics.

Servifyspheresolutions. 2025. *Why Transformers, BERT, and GPT are Shaping the Future of NLP*. Medium.

Shandeep, S. 2022. *CountVectorizer vs TfidfVectorizer*. Medium.

Sheel, S. 2020. *Count Vectorizer vs IF-IDF Vectorizer | Natural Language Processing*. LinkedIn.

Simon, L. 2020. *Comparison of undersampling methods for prediction of casting defects based on process parameters*. Projek Ijazah Sarjana. Sweden: University of Skovde.

Simplilearn. 2025. *What is XGBoost? An Introduction to XGBoost Algorithm in Machine Learning*.

Suzaimah Binti Ramli. 2002. *Reka Bentuk Dan Implement Suatu Penghurai Bahasa Melayu Menggunakan Sistem Logik Selari*. Serdang: Penerbit Universiti Putra Malaysia.

Viso.ai. *Overfitting in Machine: How to Detect and Avoid Overfitting in Computer Version?* eNCORD.

Xingjian, Z & Lu, Z. 2024. *The comparison and analysis of Skip-gram and CBOW in creating financial sentiment dictionary.* ResearchGate

Yu, R. 2024. *An Analysis of Different Machine Learning Algorithms for Personality Type Predictions Based on Social Media Posts.* Scitepress.org.

Zhang, H. 2024. *MBTI Personality Prediction Based on BERT Classification*. ResearchGate.

Tan Ke Ying (A192389)
Prof. Madya Dr. Nazlia Binti Omar
Fakulti Teknologi & Sains Maklumat
Universiti Kebangsaan Malaysia