PERSEPSI MASYARAKAT TERHADAP KESIHATAN MENTAL DENGAN MENGGUNAKAN ANALISIS SENTIMEN

TASSLIM BIN MANSOOR ALI
LAILATUL QADRI ZAKARIA

Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Kesihatan mental merujuk kepada kesejahteraan kognitif, tingkah laku, dan emosi seseorang. Ini merangkumi bagaimana seseorang berfikir, merasa, dan berkelakuan. Isu kesihatan mental bukanlah suatu isu yang asing. Persepsi dan penerimaan masyarakat terhadap kesihatan mental berkait rapat dengan tahap kefahaman dan pengetahuan mereka dalam isu ini. Apabila seseorang mempunyai masalah kesihatan mental, ada sebahagian daripada mereka menggunakan media sosial seperti Twitter sebagai platform untuk meluahkan masalah mereka. Walaupun begitu, ada sesetengah dalam kalangan masyarakat yang tidak bertanggungjawab menggunakan masalah mangsa kesihatan mental sebagai bahan lucu dan menyendakan perasaan mereka. Sebaliknya, terdapat segelintir masyarakat yang menyokong dan menghulurkan bantuan kepada mangsa. Objektif utama untuk projek ini ialah untuk melakukan analisa untuk mengenal pasti persepsi masyarakat terhadap kesihatan mental berdasarkan ciapan yang ditulis oleh pengguna di dalam aplikasi Twitter. Projek ini akan menggunakan teknik Pemprosesan Bahasa Tabii (NLP), pengelasan, Pembelajaran Mesin (ML) dan analisis sentimen. Twitter API merupakan satu platform untuk mengumpul ciapan pengguna daripada Twitter. Selain itu, teknik pembersihan data seperti tokenisasi, normalisasi dan penyinkiran emotikon dan hashtag akan dilakukan. Selepas itu, pengekstrakan ciri dalam teks seperti teknik TF-IDF dan Word2Vec dilaksanakan dan teknik pembelajaran mesin seperti Naïve Bayes, *Random Forest* dan *K-Nearest Neighbour* digunakan untuk mengenal pasti kategori yang dibincang oleh pengguna. Hasil kajian seperti pengkategorian ciapan kepada kategori emosi, persekitaran dan sokongan sosial divisualisasikan dalam bentuk carta pai dan bar. Hasil kajian mendapati bahawa model Naïve Bayes mempunyai prestasi yang baik berbanding *Random Forest* dan *K-Nearest Neighbour*. Kajian ini juga menunjukkan teknik TF-IDF mempunyai prestasi baik berbanding Word2Vec. Hal ini kerana Word2Vec memerlukan set data besar berbanding TF-IDF yang mampu memberikan prestasi baik dengan set data kecil. Hasil analisis pengelasan kategori mendapati bahawa model mengkategorikan ciapan kesihatan mental kepada kategori emosi berbanding kategori persekitaran dan sokongan sosial. Kebanyakan pengguna membincangkan isu kesihatan mental lebih cenderung untuk mengongsikan emosi mereka dalam platform ini. Akhirnya, kaedah analisis sentimen seperti VADER dan TextBlob digunakan untuk mengenal pasti sentimen pengguna terhadap setiap kategori yang ditetapkan. Perkara yang diharapkan untuk projek ini adalah supaya dapat menentukan persepsi masyarakat terhadap isu kesihatan mental dan menjadi rujukan untuk pakar kesihatan mental dan komuniti sains data pada masa depan untuk meningkatkan kesedaran orang ramai tentang isu masalah kesihatan mental.

1 PENGENALAN

Kesihatan mental merupakan komponen kesihatan yang amat penting dalam kehidupan seseorang manusia. Perlembagaan Pertubuhan Kesihatan Sedunia (WHO) menyifatkan kesihatan mental sebagai keadaan kesejahteraan fizikal, mental dan sosial yang lengkap dan bukan sekadar sihat daripada mana-mana penyakit atau kelemahan. Definisi yang diberi oleh WHO memberikan satu gambaran bahawa kesihatan mental adalah lebih daripada sekadar ketiadaan masalah mental atau kecacatan. Kesihatan mental merupakan satu keadaan kesejahteraan di mana seseorang individu menyedari kebolehan untuk menjalankan sesebuah

tugasan dengan sendiri, dapat mengatasi tekanan hidup yang normal, dapat bekerja dengan produktif dan menyumbangkan jasa kepada masyarakat. Kesihatan mental adalah asas kepada keupayaan kolektif seseorang manusia untuk berfikir, beremosi, berinteraksi antara satu sama lain, mencari rezeki dan menikmati kehidupan.

Masalah kesihatan mental adalah perkara yang biasa, mempengaruhi puluhan ribu manusia di serata dunia. Walau bagaimanapun, ada sesetengah golongan masyarakat yang mempunyai persepsi yang negatif terhadap masalah kesihatan mental. Mereka yang mempunyai masalah kesihatan mental kerap mengalami diskriminasi dalam kehidupan seharian. Stigma dan diskriminasi menjadikan masalah kesihatan mental seseorang menjadi semakin teruk. Ia boleh wujud dari masyarakat, majikan, media sosial, malah rakan dan keluarga kita sendiri. Terdapat beberapa sebab mangsa kesihatan mental mengalami diskriminasi daripada masyarakat. Antararanya ialah stereotaip. Sebilangan orang percaya bahawa orang yang mempunyai masalah kesihatan mental adalah berbahaya kepada diri mereka. Sebenarnya, mereka yang mempunyai masalah kesihatan mental mempunyai risiko yang lebih tinggi untuk diserang atau mencederakan diri sendiri daripada orang lain. Selain itu, media massa turut mendorong masyarakat untuk mendiskriminasikan mangsa kesihatan mental secara tidak langsung. Laporan media massa sering menghubungkan mangsa kesihatan mental dengan kekerasan atau menggambarkan mangsa sebagai merbahaya, jenayah, cacat dan tidak dapat menjalankan kehidupan yang normal.

Analisis sentimen adalah proses untuk mengesan sentimen positif atau negatif dalam sesebuah teks. Ia sering digunakan oleh perniagaan dan syarikat untuk mengesan sentimen dalam data sosial, mengukur reputasi jenama, dan memahami pengguna dengan lebih mendalam. Oleh kerana pengguna meluahkan perasaan dan pendapat mereka secara lebih terbuka di media sosial, analisis sentimen menjadi alat penting untuk memantau dan memahami sentimen seseorang. Terdapat dua teknik asas untuk analisis sentimen. Antaranya ialah analisis sentimen berasaskan peraturan. Teknik pertama adalah berasaskan peraturan dan menggunakan kamus perkataan yang dilabelkan oleh sentimen untuk menentukan sentimen ayat. Sesebuah pendapat yang diberi oleh seseorang boleh dikategorikan mengikut polariti positif, negatif atau neutral. Teknik kedua ialah analisis sentimen berasaskan Pembelajaran Mesin (ML). Teknik ini berasaskan model pembelajaran mesin untuk mengenali sentimen berdasarkan kata-kata dan berdasarkan set latihan sentimen. Teknik ini banyak bergantung pada jenis algoritma dan kualiti data latihan yang digunakan.

Kajian ini akan mengkaji secara terpeinci tentang persepsi masyarakat terhadap kesihatan mental dengan menggunakan analisis sentimen. Masyarakat umum mempunyai

pendapat yang berbeza tentang kesihatan mental. Ada sebahagian masyarakat mempunyai persepsi yang positif terhadap kesihatan mental manakala ada golongan masyarakat yang mempunyai pandangan yang negatif terhadap kesihatan mental. Antara sebab memilih kedua-dua persepsi ini adalah kerana emosi orang ramai terhadap topik ini adalah sangat berbeza dan ini akan membantu untuk mendapatkan hasil analisis sentimen yang menarik.

2 PENYATAAN MASALAH

Media sosial seperti Twitter telah menjadi sumber utama untuk kebanyakan pengkaji untuk mengumpul data dan maklumat. Hal ini demikian kerana aplikasi Twitter merupakan antara laman web yang digunakan oleh orang ramai untuk berkongsi pendapat, perasaan dan aktiviti harian mereka. Twitter juga membantu pengkaji untuk membina model pembelajaran mesin untuk mengklasifikasi ciapan kepada sentimen positif, negatif dan neutral. Walaupun begitu, terdapat beberapa masalah telah dikenal pasti. Antaranya ialah kekurangan kajian terhadap pengkelasan persepsi masyarakat terhadap kesihatan mental di Malaysia dan tiada set data yang tersedia yang boleh digunakan. Dalam pada itu, ketiadaan analisis sentimen terhadap persepsi masyarakat terhadap kesihatan mental yang boleh digunakan sebagai rujukan untuk kajian ini. Oleh sebab itu, wujudnya keperluan untuk mencari, mengumpul dan menganalisa data yang berkaitan dengan kesihatan mental daripada Twitter.

Selain itu, kebanyakkan pengguna media sosial terutamanya Twitter lebih cenderung untuk menggunakan *sarcasm*, emotikon, *hashtag* dan aksara khas dalam penulisan mereka. Sebagaimana yang kita sedia maklum, Twitter mempunyai had jumlah perkataan untuk setiap ciapan iaitu sebanyak 280 aksara. Sekatan saiz ciapan ini telah menjadikan pengguna untuk bersikap kreatif dalam ciapan mereka. Dengan keadaan sedemikian, mesej yang cuba dibawa oleh pengguna kadang-kadang senang ataupun sukar untuk dibaca konteksnya oleh orang ramai.

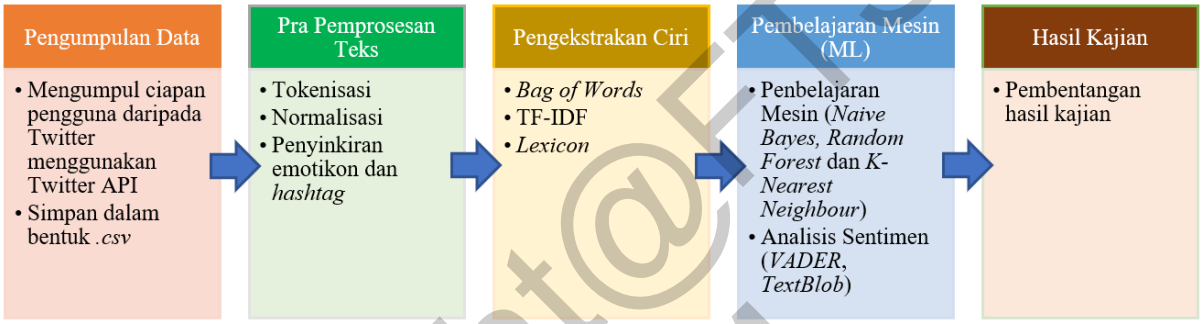
3 OBJEKTIF KAJIAN

i. Mengumpul data dan mengenal pasti persepsi masyarakat terhadap kesihatan mental berdasarkan ciapan (*tweets*) yang ditulis oleh pengguna di Twitter.
ii. Mengkelas persepsi yang berkaitan dengan kesihatan mental menerusi pendekatan pembelajaran mesin iaitu pengelasan Naïve Bayes, *Random Forest* dan *K-Nearest Neighbour*.

iii. Meramal sentimen persepsi masyarakat terhadap kesihatan mental di Malaysia dengan menggunakan kaedah analisis sentimen.

4 METOD KAJIAN

Kajian ini akan menggunakan Model Klasifikasi Sentimen (*Sentiment Classification Model*) untuk pembangunan algoritma ini. Model metodologi ini adalah mudah untuk difahami oleh umum dan akan digunakan dalam proses menganalisa sentimen masyarakat terhadap kesihatan mental. Rajah 1 di bawah menunjukkan Model Klasifikasi Sentimen untuk kajian ini.



Rajah 1 Model Klasifikasi Sentimen

4.1 Fasa Perancangan

Fasa ini membabitkan proses memperoleh dan mengenal pasti objektif, skop dan kekangan kajian. Sorotan kajian kesusasteraan yang melibatkan pengumpulan maklumat, pencarian dan perbandingan jurnal dan kajian lepas dilaksanakan untuk mendapatkan gambaran dan hala tuju kajian yang dilakukan. Antara topik yang dikaji adalah berkaitan dengan persepsi masyarakat terhadap kesihatan mental. Set data yang berkaitan dengan kajian dikumpul daripada media sosial Twitter.

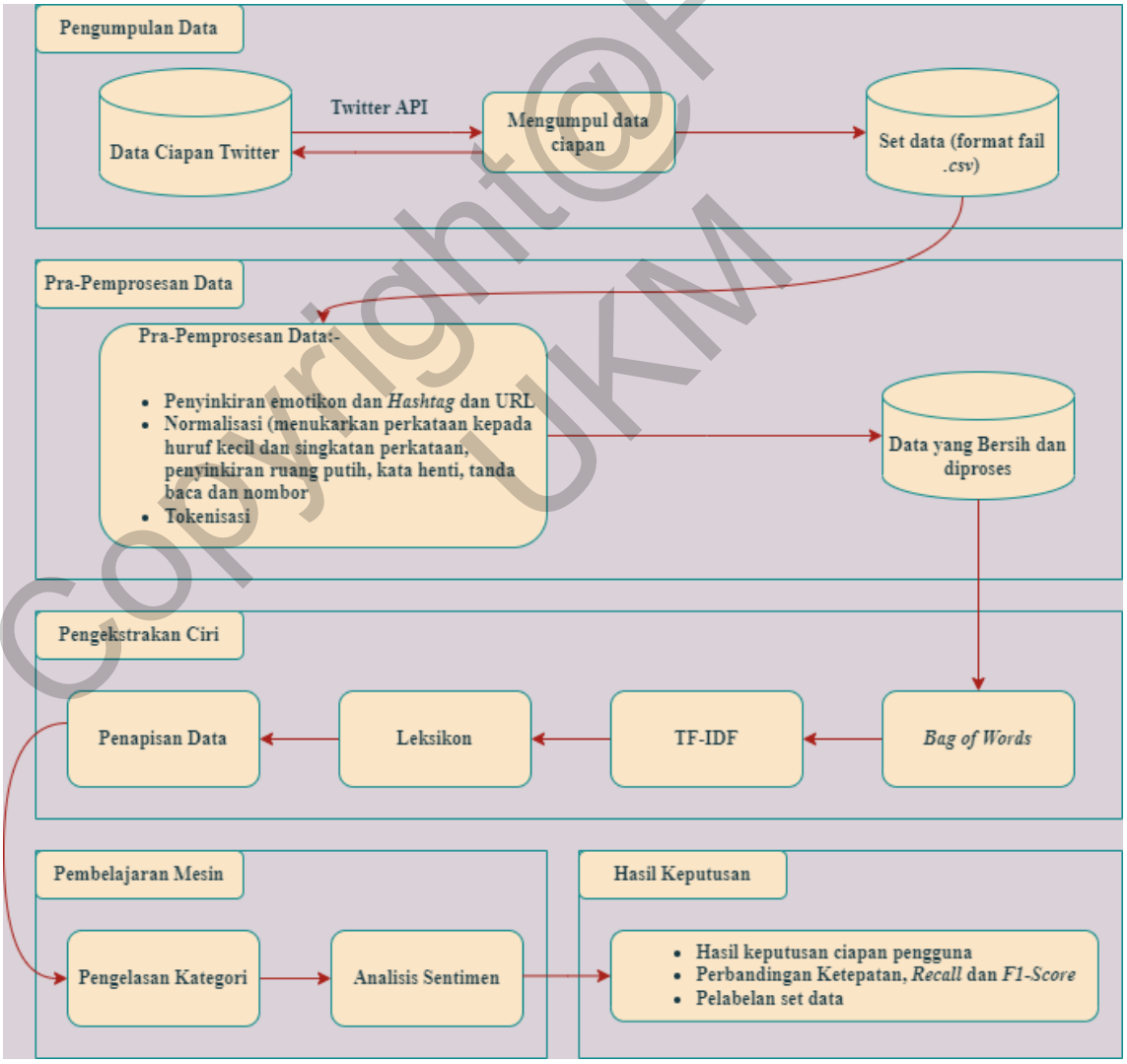
4.2 Fasa Analisis

Fasa ini melibatkan analisis dan tafsiran terhadap set data yang dikumpul daripada fasa perancangan. Tujuan fasa ini dilaksanakan adalah untuk memastikan data yang dikumpul adalah sesuai dan relevan dengan topik kajian. Selepas kesemua data ditapis kepada ciapan yang berkaitan dengan kesihatan dan Bahasa Inggeris, set data menjalani proses pra-pemprosesan data dan pengekstrakan ciri. Sebanyak 1530 ciapan data Twitter berjaya dikumpul dan dilabelkan secara manual kepada tiga kategori iaitu emosi, persekitatan dan sokongan sosial (Fahad Riaz Choudhry, Vasudevan Mani, Long Chiau Ming & Tahir

Mehmood Khan 2016). Data ciapan dikategorikan untuk memudahkan data untuk dianalisa. Analisis data dilaksanakan untuk memastikan set data konsisten supaya hasil keputusan yang diperoleh adalah berkualiti.

4.3 Fasa Reka Bentuk

Fasa ini memfokuskan reka bentuk seni bina yang boleh memberikan gambaran keseluruhan algoritma yang dibangunkan. Rajah 2 di bawah menunjukkan reka bentuk seni bina persepsi masyarakat terhadap kesihatan mental dengan menggunakan analisis sentimen dalam bentuk ilustrasi. Reka bentuk seni bina untuk sistem ini terbahagi kepada enam fasa utama iaitu pengumpulan data, pra-pemprosesan data, pengekstrakan ciri, pembelajaran mesin yang terdiri daripada pengelasan kategori dan analisis sentimen dan hasil keputusan.



Rajah 2 Reka Bentuk Seni Bina Analisis Sentimen Persepsi Masyarakat Terhadap Kesihatan Mental

4.4 Fasa Implementasi

Fasa ini melibatkan implementasi setiap fasa utama yang dinyatakan dalam fasa reka bentuk. Kesemua fasa utama yang dilaksanakan menghasilkan satu algoritma yang mencapai objektif kajian. Cara pelaksanaan bagi setiap fasa utama kajian ini diperincikan di bawah.

a) Fasa Pengumpulan Data

Pengumpulan sesebuah set data atau *corpus* diambil daripada aplikasi Twitter. Twitter API merupakan satu sistem untuk pengkaji untuk mengumpul ciapan pengguna atau data. Antara kata kunci yang releven dengan kesihatan mental di Malaysia ialah *mental health*, *depressed*, *stress*, *struggle*, *toxic*, *supportive* dan sebagainya. Kata kunci ini dipilih kerana Twitter API mampu mengumpul kesemua ciapan pengguna yang menggunakan perkataan yang berkaitan dengan kesihatan mental di Malaysia. Kesemua data yang dikumpul daripada Twitter akan disimpan dalam bentuk *.csv* agar mudah untuk digunakan dalam fasa seterusnya. Data yang dikumpul ini adalah data yang bersifat mentah dan mempunyai jumlah yang besar.

	A	B	C	D	E	F	G	H
1	user_name	user_location	user_description	user_verif	date	text	hashtags	source
2	National Resource Directory	Washington, D.	The NRD is a resou	TRUE	########	If you are struggling with a #mentalhealth	['mentalhealth']	Sprout Social
3	One Patient Wolf	Washington, DC	Fortyish. Writer. P	FALSE	########	Your doomscrolling involves the news. My		Twitter for Android
4	Ben Farrell	Milton Keynes,	Business account (	FALSE	########	Looking after no.1 #mentalhealth	['mentalhealth', 'cc	Twitter for iPhone
5	Tish Nicole		#IncHimInAllThatY	FALSE	########	Deliverance Counseling Services, LLC		Twitter for iPhone
6	Dr. Sadaqat Ali	Lahore, Pakistal	Dr. Sadaqat Ali is a	FALSE	########	ð–ðiðcðœðì ðð«ð®ð ð– ð«ðšð©ðcðð¥ð²		Twitter Web App
7	Wormhole Manifest	Missouri	Wormhole Manife	FALSE	########	Have ya heard the latest episode yet? Check	['podcast', 'podcast	Twitter for Android
8	Bridges Healthcare Inc.	Milford, CT	Bridges is where t	FALSE	########	COVID-19 has left children in crisis; CT	['mentalhealth']	Twitter Web App
9	Hey Diddle Diddle	New York, NY	Critically-acclaime	FALSE	########	"Hey Diddle Diddle - The Film" DOWNLOAD (	['social']	Dynamic Tweets
10	Jewish Family & Children's Service	New Jersey, US	Based on Jewish tl	FALSE	########	Take a #Vacation! Vacations are great for	['Vacation', 'stress'	Hootsuite Inc.
11	Kooth_Michelle Smith	London and the	Engagement lead	FALSE	########	A handy guide to #kooth for parents. #suppo	['kooth', 'support',	Twitter for iPhone
12	Thereâ€™s A Chance We Are Similar		You may feel alon	FALSE	########	The lesser of two evils is still not okay. Too		Twitter for iPhone
13	Aishu		sup	FALSE	########	we'll are constantly trying to prove		Twitter Web App
14	Kelly Anderson	New York, USA	#HealthIT News &	FALSE	########	How Telehealth Interoperability with EHR	['EMRfinder']	Twitter Web App
15	JamieðŸŒ»		Unconditional Sup	FALSE	########	When you think everything is going great and then you walk int		Twitter for iPhone
16	PurrfectlyPlanted	Your gfâ€™s pla	Fareeha-ÙØ±ÛŒØ	FALSE	########	Happy Mental Health Monday purrties, this		Twitter for iPhone
17	Caroline Beagan		volunteer for FFU	FALSE	########	I am proud of myself & everyone else v	['Fibromyalgia', 'ch	Twitter for iPhone
18	Ty Harrison	Oakland, CA	On the up and up.	FALSE	########	If you make yourself better you inherently n	['WordsOfWisdom	Twitter for iPhone
19	BipolarBandit	North Carolina	I'm a mental	FALSE	########	I'm not telling you it's going to be easy. I'm	['determination']	Twitter for Android
20	ðŸ¦• My Lochness Unicorn ðŸ¦„	London, Englan	Iâ€™m a man who	FALSE	########	A great morning at @Zonegym_wdgreen . Pe	['weights', 'fitness'	Twitter for iPhone

Jadual 1 Contoh Fail *CSV* Yang Mengandungi Data Yang Dikumpul

b) Fasa Pra-Pemprosesan Data

Pra-pemprosesan data dilakukan untuk menjadikan data yang dikumpul itu mudah dibaca dan diproses oleh sistem. Antara kaedah pra-pemprosesan data ialah menyinkirkan *hashtag*, emotikon, ruang kosong, perkataan yang tiada makna dan URL. Normalisasi data seperti penukaran perkataan kepada huruf kecil, penukaran singkatan perkataan dan penyinkiran ruang putih, kata henti, tanda baca dan nombor dilakukan agar lebih mudah dibaca dan dianalisa. Seterusnya, tokenisasi akan dilaksanakan di mana setiap perkataan di dalam data dipecahkan kepada perkataan yang tunggal untuk pembersihan data. Akhirnya, data yang menjalani fasa ini akan disimpan untuk menjalani fasa seterusnya.

Text	clean_text
I felt like i wanna sell some of my "good projects" nft bcuz of financial struggle but its kinda hard for me to let go as im one	felt like wan na sell good project nft bcuz financial struggle kinda hard let go im one among people hold value m
@notjayus People assume things about people because it's easier then understanding and that fact gives me depression	people assume thing people easier understanding fact give depression
Ã¢â‚¬â€ Like I already feel very self conscious and uneasy talking to cis men in general and having to bare my trauma and r	like already feel self conscious uneasy talk cis men general bare trauma mental health one feel terrify use pron
Not that i dont have dips in my mental health where i rely way too much on social media for validation, but im glad its not	dont dip mental health rely way much social medium validationbut im glad constant anymore mean im get good
I donÃ¢â‚¬â„¢t think thereÃ¢â‚¬â„¢s a point in bringing up what PH/PN/PN+BN didnÃ¢â‚¬â„¢t do. LetÃ¢â‚¬â„¢s face it. The	dont think there point bring ph pn pn bn didnt let face past year sham whereby power struggle rob u proper poli
Zaitul Akma Muhd Zin, who has two children suffering from a neurodegenerative disorder, tells of her struggle and what k	zaitul akma muhd zinwho two child suffer neurodegenerative disordertells struggle keep go
I'm back I took a break cause I've been having a rough time with my mental health i missed you guys	back take break cause rough time mental health miss guy
So, never take your mental health lightly and take care of yourself.	sonever take mental health lightly take care
Taking care of my mental health Ã°Å¸â€™â„¢c	take care mental health
@PupAmp Your mental health is what is most important. Enjoy your night off and hey, it's an excuse to be able to watch a	mental health important enjoy night heyit excuse able watch guilty pleasure halloween movie
Damnnnn the struggle is real. Semoga tix kita lepas @idontknowwho112 @thevirginafie	damnnnn struggle real semoga tix kita lepas
@MRibnek Praying for your mental health here too	pray mental health
i took a little break from twitter for my mental health* idk what happened lmao	take little break twitter mental health idk happen lmao

Jadual 2 Data Ciapan Twitter Sebelum Dan Selepas Pra-Pemprosesan Data

c) Fasa Pengekstrakan Ciri

Fasa pengekstrakan ciri dibuat sebaik data yang bersih diperoleh daripada pra-pemprosesan data dan akan digunakan untuk mengenal pasti ciri unik yang terdapat dalam data tersebut. Ciri unik dalam konteks ini ialah kandungan yang terdapat dalam ciapan pengguna yang mengandungi kata kunci yang berkaitan dengan kesihatan mental dan kategori topik yang dibincangkan. Dalam kajian ini, data yang dikumpul hanya memfokuskan persepsi masyarakat di Malaysia sahaja. Ciapan yang dikumpul juga mengandungi bahasa-bahasa yang lain seperti Bahasa Melayu, Sepanyol, Tagalog dan lain-lain. Penapisan ciapan dari segi bahasa amat diperlukan dalam kajian ini untuk memastikan ciapan yang diproses hanya mengandungi Bahasa Inggeris sahaja. Jadual 3 dibawah menunjukkan jumlah data mentah sebelum dan selepas proses penapisan. Selain itu, Frekuensi Istilah-Frekuensi Dokumen Terbalik (TF-IDF) adalah ukuran statistik yang menilai sejauh mana sesuatu perkataan itu releven dengan dokumen tersebut dalam koleksi dokemen. Seterusnya, Word2Vec merupakan sebuah teknik pengekstrakan ciri yang digunakan selain TF-IDF untuk menukar data teks kepada angka dengan menggunakan Rangkaian Neural (Neural Network). Word2Vec menukarkan sesebuah perkataan dalam perwakilan ruang vektor. Secara teknikalnya, Word2Vec menggunakan hubungan semantik antara perkataan untuk ditempatkan dalam perwakilan vektor.

	Jumlah Data Mentah Yang Dikumpulkan	
	Sebelum Proses Penapisan	**Selepas Proses Penapisan**
Jumlah Keseluruhan Data	3479	1530

Jadual 3 Jumlah Data Mentah Sebelum Dan Selepas Proses Penapisan

d) Fasa Pengkelasan Kategori

Fasa pengelasan kategori adalah fasa yang dilaksanakan dengan menggunakan teknik pembelajaran mesin iaitu *Naïve Bayes, Random Forest* dan *K-Nearest Neighbour*. Tujuan teknik ini digunakan adalah untuk mengkategorikan topik yang dibincangkan oleh pengguna berdasarkan kategori yang dinyatakan dalam Bab 2. Data yang diperoleh akan dibahagikan kepada set percubaan (*training set*) dan set pengujian (*testing set*) untuk memastikan model pembelajaran mesin yang dibangunkan mampu untuk mengelas data. Ketepatan model pembelajaran mesin direkodkan dan dibandingkan untuk menilai keupayaan model pembelajaran mesin untuk mengenal pasti kategori dengan tepat.

e) Fasa Analisis Sentimen

Fasa analisis sentimen merupakan fasa yang menentukan sentimen sama ada data tersebut mempunyai kandungan positif atau negatif. Dalam kajian ini, *TextBlob* digunakan untuk menganalisa sentimen bagi setiap input yang dimasukkan. Penganalisis sentimen *TextBlob* menghasilkan dua sifat untuk setiap input yang diberikan. Antaranya ialah polariti di mana -1 merujuk kepada sentimen negatif manakala +1 merujuk kepada sentimen positif. Selain itu, subjektiviti juga mempunyai julat [0,1] di mana ayat subjektif kebiasaanya merujuk kepada emosi dan persepsi atau pendapat peribadi.

f) Fasa Hasil Keputusan

Fasa hasil keputusan memaparkan hasil keputusan analisis sentimen sama ada ciapan yang dianalisa terkandung dalam sentimen positif atau negatif. Selain itu, sistem juga memaparkan pengelasan kategori yang terdiri daripada emosi, persekitaran dan sokongan sosial. Akhirnya, hasil keputusan divisualisasikan dalam bentuk graf dan *Word Cloud* juga digunakan untuk mengenal pasti kekerapan perkataan yang digunakan dalam setiap kategori dalam bentuk yang menarik.

4.5 Fasa Pengujian

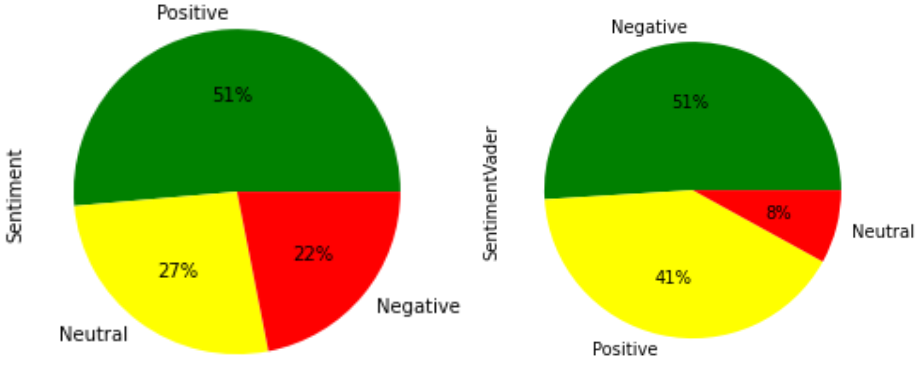
Fasa pengujian dijalankan untuk memilih model pembelajaran mesin dan teknik analisis sentimen yang terbaik bagi projek ini. Antara model pembelajaran mesin yang digunakan untuk mengelas persepsi ialah Naïve Bayes. Selain pengelas Naïve Bayes, kaedah *Random Forest* juga digunakan untuk mengelas dan memperoleh ketepatan untuk kajian ini. Kedua-dua model ini dibangunkan untuk mengelaskan setiap kategori persepsi pengguna Twitter terhadap kesihatan mental kepada kategori emosi, persekitaran dan sokongan sosial. Daripada

keseluruhan set data, 80% daripada data digunakan untuk melatih model pembelajaran mesin manakala 20% daripada data digunakan untuk menguji model pembelajaran mesin yang dilatih. Selepas pelabelan dan pembahagian data, kaedah TF-IDF dan Word2Vec digunakan sebagai pengekstrakan ciri utama untuk tujuan vektorisasi. Ketepatan bagi kedua-dua model pembelajaran mesin direkodkan ketepatan mengikut kaedah pengekstrakan ciri yang digunakan. Jadual 4 menunjukkan pengujian ketepatan Naïve Bayes dengan TF-IDF dan Word2Vec.

	TF-IDF	**Word2Vec**
Naïve Bayes	0.86	0.37
Random Forest	0.70	0.51

Jadual 4 Nilai Ketepatan Bagi Setiap Teknik Pengelas Beserta Pengekstrakan Ciri

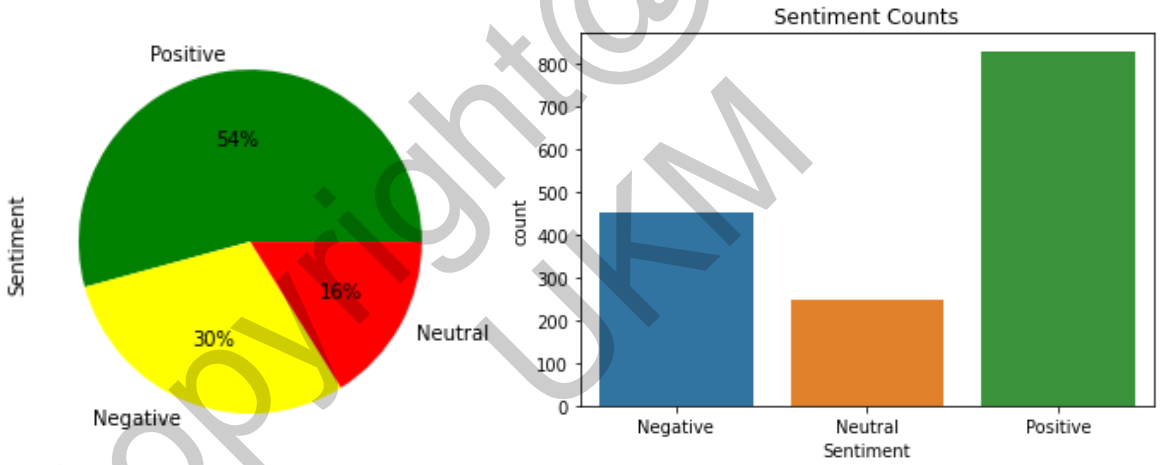
Berdasarkan prestasi kedua-dua teknik pengelas dengan pengekstrakan ciri yang digunakan, teknik Naïve Bayes dengan TF-IDF mempunyai nilai ketepatan yang lebih tinggi iaitu 0.87 peratus berbanding teknik *Random Forest* yang mempunyai ketepatan 0.70 peratus manakala kedua-dua teknik Naïve Bayes dan *Random Forest* dengan Word2Vec mengandungi nilai ketepatan yang rendah. Antara sebab kaedah Word2Vec menghasilkan ketepatan yang rendah ialah kekurangan data di mana Rangkaian Neural (*Neural Network*) memerlukan set data yang besar untuk menghasilkan ketepatan yang baik. Teknik pengelas Naïve Bayes dengan TF-IDF didapati mampu mengkategorikan persepsi masyarakat terhadap kesihatan mental dengan lebih tepat dalam kajian ini. Selain itu, teknik analisis sentimen TextBlob dan VADER telah digunakan untuk mengenal pasti sentimen ciapan pengguna terhadap kesihatan mental berdasarkan kategori yang dikelaskan. Rajah 3 di bawah menunjukkan perbandingan carta pai bagi TextBlob dan VADER.



Rajah 3 Perbandingan Carta Pai TextBlob dan VADER

5 HASIL KAJIAN

Hasil kajian bagi projek ini memaparkan hasil keputusan analisis sentimen sama ada ciapan yang dianalisa terkandung dalam sentimen positif atau negatif. Hasil keputusan divisualisasikan dalam bentuk graf dan Awan Perkataan (*WordCloud*) juga digunakan untuk mengenal pasti kekerapan perkataan yang digunakan dalam setiap kategori dalam bentuk yang menarik. Selain itu, sistem ini juga memaparkan pengelasan kategori topik perbincangan pengguna Twitter yang terdiri daripada emosi, persekitaran dan sokongan sosial. Set data yang bersih, diproses dengan ciri yang diekstrak juga dibentangkan sebagai salah sebuah hasil keputusan. Rajah 4 menunjukkan carta pai dan carta bar bagi sentimen ciapan yang dikelas bagi keseluruhan set data. Berdasarkan rajah di bawah, sentimen set data yang diperoleh daripada ciapan pengguna Twitter terdiri daripada 54% ciapan positif (830 ciapan), 30% ciapan negatif (453 ciapan) dan 16% ciapan neutral (247 ciapan).



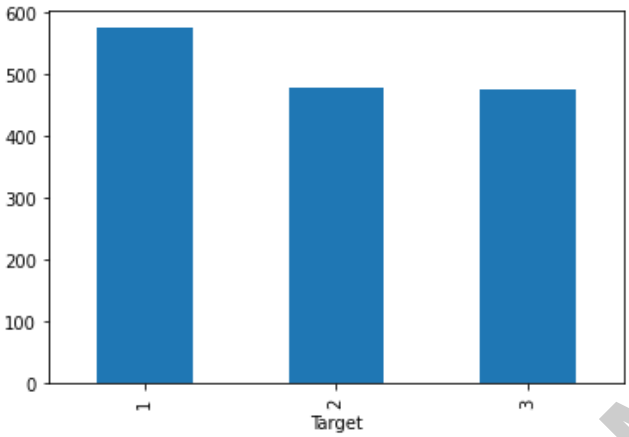
Rajah 4 Jumlah Ciapan, Carta Pai Dan Carta Bar Bagi Sentimen Ciapan Yang Dikelas

Rajah 5 di bawah menunjukkan jumlah ciapan yang dikelas mengikut setiap kategori dalam bentuk carta bar. Persepsi kategori yang telah ditetapkan pada awal kajian ini telah dikelaskan secara manual terdahulu dengan nombor untuk memudahkan proses klasifikasi ciapan pengguna dengan menggunakan kaedah pembelajaran mesin. Berikut merupakan nombor yang diberikan untuk setiap kategori:-

“Emosi” = 1

“Persekitaran” = 2

“Sokongan Sosial” = 3



Rajah 5 Jumlah Ciapan Yang Dikelas Mengikut Setiap Kategori Dalam Bentuk Carta Bar

Rajah 6 di bawah menunjukkan contoh ciapan yang telah dikenal pasti sebagai kategori emosi. Merujuk kepada Rajah 6, terdapat beberapa perkataan yang menggambarkan sifat unik seperti "bad", "loss", "neglect", "suck" dan sebagainya dapat dikaitkan dengan kategori emosi. Antara perkataan unik lain yang wujud dalam ciapan lain ialah "depressed", "stressed", "struggling", "hate" dan seumpamanya.

Text	Persepsi Kategori
thought back tooold memory stuff wish could speak someone kinda thing great mental health service suck probably mention mindfulness b im seriously verge something	1
butyou guy must neglect agency governmentwhose voice heard sadly drown struggle	1
help guy deal stress feeling stress agitate past day cant seem shake	1
today exceptionally bad mental health day idk might call loss try tomorrow work	1
someone gon na tell depression lazy imma smack	1
neglect people say think might affect mental health goodmorningggggg	1

Rajah 6 Contoh Ciapan Persepsi Kategori "Emosi"

Rajah 7 di bawah menunjukkan contoh ciapan yang telah dikenalpasti sebagai kategori persekitaran. Merujuk kepada Rajah 7, terdapat beberapa perkataan yang menggambarkan sifat unik seperti "relationship", "abuse", "troll", "manipulate" dan sebagainya dapat dikaitkan dengan kategori persekitaran. Antara perkataan unik lain yang wujud dalam ciapan lain ialah "toxic", "environment", "society", "family" dan seumpamanya.

Text	Persepsi Kategori
already separate way try fix save relationship many way toxic one handle attitude anymore love know im hurt already time leg go love till meet	2
imagine manipulate mentally abuse somebody go back pretend innocent preach ab mental health lol	2
acknowledge would absolutely ignorant selfish equate struggleshowever difficult might beto others broken family well life support far bad past	2
toxic guy like u never change necessary mention player guy wrong u know guy abuse troll indian player consider anti nationalist	2
comment return handle toxic family member toxic family member really number u childhood even adulthood never late take action	2
parent didnt raise u way allow u prioritize mental healthand really hope generation change	2

Rajah 7 Contoh Ciapan Persepsi Kategori "Persekitaran"

Rajah 8 di bawah menunjukkan contoh ciapan yang telah dikenalpasti sebagai kategori sokongan sosial. Merujuk kepada Rajah 8, terdapat beberapa perkataan yang menggambarkan

sifat unik seperti “appreaciate”, “support”, “thankful”, “understanding” dan sebagainya dapat dikaitkan dengan kategori persekitaran. Antara perkataan unik lain yang wujud dalam ciapan lain ialah “supportive”, “thankful”, “talented”, “blessed” dan seumpamanya.

Text	Persepsi Kategori
alhamdullah supportive community friend lot good use year ago	3
people back need supportive understanding	3
thanks reply everyone personally close family supportive im content fly radar otherwise reason ask saw friend work colleague buy book think sweet maybe quite unusual curious	3
awthanks appreciate support always one first people like postsand want thank supportive sothank	3
tonight best friend wedding get catch people see year everyone nothing supportive proud career step make hard year life lately friends always step make sure stay continue	3
besides really enjoy hear everything finethat important everything fineyou work enoughyou reach goal also need restthat kinda stuff say cozy supportive	3

Rajah 8 Contoh Ciapan Persepsi Kategori “Sokongan Sosial”

6 KESIMPULAN

Kesihatan mental merupakan satu perkara yang yangat penting dalam setiap peringkat kehidupan seseorang, dari kecil dan remaja, hingga ke dewasa. Hal ini demikian kerana kesihatan mental merupakan asas daya fikiran, pembelajaran, komunikasi, daya tahan dan harga diri manusia. Setiap orang di dunia ini mestilah menjaga kesihatan mental dengan baik agar dapat menjalankan kehidupan seharian dengan sempurna. Apabila seseorang mempunyai masalah kesihatan mental, sokongan sosial memainkan peranan penting dalam pemulihan individu tersebut. Oleh sebab itu, setiap pengguna media sosial diingatkan supaya sentiasa berhemah dalam penulisan mereka supaya tidak menyakiti perasaan mangsa.

Objektif kajian ini ialah untuk mengumpul data dan mengenal pasti persepsi masyarakat terhadap kesihatan mental berdasarkan ciapan yang dtulis oleh pengguna di Twitter. Selain itu, kajian ini juga mengambil kira tiga kategori utama yang dibincangkan pengguna Twitter iaitu emosi dan fizikal, persekitaran dan sokongan sosial. Oleh kerana kurangnya kajian terhadap persepsi masyarakat terhadap kesihatan mental di Malaysia dan ketiadaan set data yang tersedia untuk digunakan, wujudnya keperluan pengkaji untuk mencari, mengumpul dan menganalisa data Twitter untuk mengenal pasti persepsi masyarakat terhadap kesihatan mental.

Bagi menyelesaikan masalah yang wujud dan mencapai objektif yang ditetapkan, data Twitter dikumpulkan dengan menggunakan Twitter API. Data yang dikumpul disimpan dalam bentuk *CSV* supaya mudah untuk diproses. Selain itu, pra-pemprosesan data dilakukan di mana teks data menjalani proses penyinkiran perkataan yang tidak membawa makna dan tokenisasi. Di samping itu, data yang diproses menjalani fasa pengekstrakan ciri dan pembelajaran mesin untuk tujuan pengelasan kategori dan analisis sentimen. Kaedah pengelas Naïve Bayes dan *Random Forest* digunakan untuk mengelas teks kepada tiga

kategori yang sesuai. Bagi proses analisis sentimen pula, kaedah TextBlob dan VADER digunakan untuk mengenal pasti ciapan yang bernada positif atau negatif. Akhirnya, hasil keputusan iaitu output analisis sentimen dan pengelasan kategori diharapkan dapat memenuhi objektif kajian ini dan divisualkan dalam bentuk graf.

Kajian ini diharap dapat membantu pihak yang bertanggungjawab dalam isu kesihatan mental seperti Kementerian Kesihatan Malaysia (KKM) dan pakar kesihatan mental dalam mencari cara yang terbaik untuk mendidik pengguna media sosial tentang isu kesihatan mental. Pembangunan sistem ini juga diharap dapat meningkatkan ilmu pengetahuan dalam Pembelajaran Mesin (ML) dan bahasa pengaturcaraan Python yang semakin popular dalam kalangan komuniti Sains Komputer, khususnya Sains Data.

7 RUJUKAN

Haji Talib, N. 'Aina F., & Abdullah @ Mohd. Nor, H. (2020). Persepsi Masyarakat dan Pesakit Terhadap Kesihatan Mental. *Jurnal Wacana Sarjana*, *4*(1),1-13. Retrieved from http://spaj.ukm.my/jws/index.php/jws/article/view/261

Morrow, N. (2021). *What is mental illness?* SANE Australia. https://www.sane.org/information-stories/facts-and-guides/what-is-mental-illness

World Health Organization (WHO). (2018). *Mental health: strengthening our response*. https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response

Mental Health Foundation. (2021). *Stigma and discrimination*. https://www.mentalhealth.org.uk/a-to-z/s/stigma-and-discrimination

Malaysian Psychiatric Association *Buku Panduan Kesihatan Mental*

Thoits, P. A. (2010). Stress and health major findings and policy implications. Journal of Health and Social Behavior, 51(1), 41–53.

Luke, F. (2021). *Kesihatan Mental: Apakah Punca Kes Ini Semakin Meningkat?* DoctorOnCall. Retrieved from: https://www.doctoroncall.com.my/health-centre/kesihatan-mental/masalah-kesihatan-mental

Hassan, M. F. bin, Hassan, N. M., Kassim, E. S., & Hamzah, M. I. (2018). Issues and Challenges of Mental Health in Malaysia. International Journal of Academic Research in Business and Social Sciences, 8(12), 1685–1696.

Lee, M. F., & Lai, C. S. (2017). Exploring Learners' Mental Health Profile: A study in Universiti Tun Hussein Onn Malaysia. In IOP Conference Series: Materials Science and Engineering (Vol. 226, p. 12194). IOP Publishing.

Health, I. for P. (2011). National Health and Morbidity Survey 2011 (NHMS 2011). Vol. II: Non-Communicable Disease. Ministry of Health Malaysia Kuala Lumpur.

Rose, D., Thornicroft, G., Pinfold, V., & Kassam, A. (2007). 250 labels used to stigmatise people with mental illness. *BMC health services research*, *7*, 97. https://doi.org/10.1186/1472-6963-7-97

Choudhry, F. R., Mani, V., Ming, L. C., & Khan, T. M. (2016). Beliefs and perception about mental health issues: a meta-synthesis. *Neuropsychiatric disease and treatment*, *12*, 2807–2818. https://doi.org/10.2147/NDT.S111543

Kurama, V. (2021). *Document Classification | Nanonets Document Classifier*. AI & Machine Learning Blog. https://nanonets.com/blog/document-classification/

Thanusha K. R. (2021). *Analisis Sentimen Filem Dan Drama Bersiri Di Netflix Menggunakan Pembelajaran Mesin.* Bangi: Universiti Kebangsaan Malaysia

Shivanandhan, M. (2020). What is Sentiment Analysis? A Complete Guide for Beginners. Retrieved from freeCodeCamp: https://www.freecodecamp.org/news/what-is-sentiment-analysis-a-completeguide-to-for-beginners/

Pradeep Kumar Tiwari et al (2021) IOP Conf. Ser.: Mater. Sci. Eng. 1099 012043 https://iopscience.iop.org/article/10.1088/1757-899X/1099/1/012043

Chen L, Wang L, Qiu XH, Yang XX, Qiao ZX, Yang YJ, et al. (2013) Correction: Depression among Chinese University Students: Prevalence and Socio-Demographic Correlates. PLoS ONE 8(11): 10.1371/annotation/e6648eb3-37d6-44d7-8052-979af14fa921. https://doi.org/10.1371/annotation/e6648eb3-37d6-44d7-8052-979af14fa921

Shivani, Pravin, M.K Nivangune (2019) Predicting Depression Level using Social Media Sites. International Journal of Research in Engineering, Science and Management

Gamon, Michael & Choudhury, Munmun & Counts, Scott & Horvitz, Eric. (2013). Predicting Depression via Social Media. Association for the Advancement of Artificial Intelligence. https://www.researchgate.net/publication/259948193_Predicting_Depression_via_Social_Media

Schwartz, H. & Eichstaedt, Johannes & Kern, Margaret & Park, Gregory & Sap, Maarten & Stillwell, David & Kosinski, Michal & Ungar, Lyle. (2014). Towards Assessing Changes in Degree of Depression through Facebook. 10.3115/v1/W14-3214. https://www.researchgate.net/publication/283270503_Towards_Assessing_Changes_in_Degree_of_Depression_through_Facebook

Sayali Shashikant Kale (2013) Tracking Mental Disorders Across Twitter Users. University of Mumbai

Tay Fui Kien. (2019). Analisis Sentimen Twitter Mengenai Peristiwa-Persitiwa penting yang Berlaku di Sekitar UKM, Bangi. Bangi: Universiti Kebangsaan Malaysia.

Nik Najwa Binti Nik A. (2021). *Analisis Sentimen Kesan Pembelajaran Atas Talian Terhadap Pelajar Ipt Berikutan Pandemi Covid-19 Di Malaysia Dengan Menggunakan Kaedah Naïve Bayes.* Bangi: Universiti Kebangsaan Malaysia

Roul, A. (2021). *Sentiment Analysis- Lexicon Models vs Machine Learning*. Medium. https://medium.com/nerd-for-tech/sentiment-analysis-lexicon-models-vs-machine-learning-b6e3af8fe746

Shahul, E. (2021). *Sentiment Analysis in Python: TextBlob vs Vader Sentiment vs Flair vs Building It From Scratch*. Neptune.Ai. https://neptune.ai/blog/sentiment-analysis-python-textblob-vs-vader-vs-flair

Pai, A. (2021). *What is Tokenization | Tokenization In NLP*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/05/what-is-tokenization-nlp/

Lum Choi K. (2019). *Analisis Sentimen Dalam Bitcoin Tweets.* Bangi: Universiti Kebangsaan Malaysia.

A. (2020). *What is TF IDF and why it's important for SEO*. Digital Marketing Chef. http://www.digitalmarketingchef.org/what-is-tfidf-and-why-does-it-matter-for-seo/

Shah P (2020) My absolute go-to for sentiment analysis - textblob. https://towardsdatascience.com/my-absolute-go-to-for-sentiment-analysis-textblob-3ac3a11d524.

Arnal JR (2020) Introduction to NLP: Sentiment analysis and Wordclouds ⋆ Quantdare. In: Quantdare. https://quantdare.com/introduction-to-nlp-sentiment-analysis-and-wordclouds/.

Javatpoint. (2022). *Machine Learning Random Forest Algorithm* . www.javatpoint.com. Retrieved from: https://www.javatpoint.com/machine-learning-random-forest-algorithm

Kumar, P. (2020). *Word embeddings with word2vec tutorial: All you need to know*. H2kinfosys Blog. Retrieved from: https://www.h2kinfosys.com/blog/word-embeddings-with-word2vec-tutorial-all-you-need-to-know/

Twitter. https://twitter.com/

Twitter Developer. https://developer.twitter.com/en

Tasslim Bin Mansoor Ali (A177398)
Lailatul Qadri Zakaria
Fakulti Teknologi & Sains Maklumat,
Universiti Kebangsaan Malaysia