PENGELASAN SENTIMEN TERHADAP VAKSINASI COVID-19 DENGAN MENGGUNAKAN PEMBELAJARAN MEISN

LAU YONG JIE
AZURALIZA ABU BAKAR

Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Pandemik COVID-19, yang biasanya dikenali sebagai pandemi coronavirus 2019, ia adalah pandemi di seluruh dunia. Wabak ini dikenalpasti di Wuhan, China pada bulan November 2019 dan telah menyebar dengan cepat ke seluruh dunia sebagai akibat dari kegagalan membendung wabah ini. Pandemik COVID-19 menyebabkan sebilangan besar orang meninggal dunia dan memberi kesan kepada kehidupan seharian kita. Walau bagaimanapun, penemuan vaksinasi Covid-19 dapat membantu mengurangkan kes COVID-19 yang aktif di Asia kerana kebanyakan orang telah menyelesaikan vaksinasi, dengan pengecualian anti-vaxxer, yang merupakan orang yang tidak setuju dengan penggunaan vaksin untuk pelbagai alasan dan terus menyebarkan bahaya vaksinasi di media sosial, seperti Twitter, dan mempengaruhi orang lain untuk menyingkirkannya. Dalam projek ini, sistem akan dikembangkan untuk menganalisis sentimen vaksinasi di negara Asia. Orang pada masa kini memilih untuk mengekspresikan diri menggunakan platform media sosial seperti Twitter, yang merupakan salah satu platform yang paling banyak digunakan di seluruh dunia. Sebahagian daripada mereka akan menyebarkan berita yang mengelirukan dan menimbulkan masalah bagi orang lain yang tidak dapat menentukan kebenaran bahan tersebut. Sebelum data dapat dianalisis, data mesti diperoleh melalui Sambungan Twitter dan Pengesahan melalui API. Setelah pengumpulan data, pembersihan data dilakukan untuk menghilangkan kebisingan dan data yang tidak dapat digunakan. Algoritma Pembelajaran Mesin yang akan digunakan dalam model untuk klasifikasi, seperti Naïve Bayes, Mesin Vektor Sokongan, Random Forest, dan Regresi Logistik untuk membandingkan ketepatan setiap algoritma.

1 PENGENALAN

Asal usul COVID-19 dikatakan bermula pada Disember 2019 dan berasal dari Wuhan, Hubei, China. Hal ini kerana pada masa yang sama, beberapa pesakit dari Wuhan, Hubei dapat melaporkan jangkitan penafasan yang amat teruk dan mereka mempunyai titik persamaan, iaitu mereka pernah bekerja dan berada di pasar borong ikan dan makanna laut, juga dikenali sebagai pasar basah. Pada Januari 2020, pasar terpaksa ditutup sepenuhnya dengan tujuan membersihkan pasar dengan sepenuhnya supaya dapat menghapuskan virus yang menyebabkan jangkita penafasan yang teruk. Pada 7 Januari 2020, para penyelidik dapat menyelidik novel coronavirus yang dikenali sebagai SARS-CoV-2 atau 2019-nCoV. Pada mulanya, iaitu 11 Januari 2020, Pertubuhan Kesihatan Sedunia (WHO) menafikan bahawa kemungkinan penularan COVID-19 dari manusia ke manusia. Walau bagaimanapun, kes semakin meningkat dengan capat dan pada 30 Januari 2020, Pertubuhan Kesihatan Sedunia (WHO) akhirnya mengisytiharkan pandemik COVID-19 ini sebagai Kecemasan Kesihatan Awam Kebimbangan Antarabangsa (PHEIC).

Media sosial merupakan sebahagian daripada kehidupan masyarakat yang berperanan sebagai sebuah platform untuk berkomunikasi antara satu sama lain tanpa batasan dan juga boleh berkongsi pendapat sendiri kepada orang lain. Perkataan media sosial ialah gabungan daripada media dan sosial. Media merupakan alat atau perantara komunikasi dalam perhubungan manakala sosial pula adalah berkaitan dengan persahabatan, pergaulan dan aktiviti masa lapang (Kamus Dewan dan Pustaka Edisi Keempat 2014). Pelbagai aplikasi seperti Twitter, Facebook, Instagram dan lain-lain lagi adalah sebahagian Media Sosial. Pada era globalisasi ini, semakin ramai pengguna media sosial menggunakan media sosial sebagai satu platform untuk meluahkan pendapat dan juga perasaan sendiri.

Disebabkan Vaksinasi COVID-19 dihasilkan, kadar kes COVID-19 di dunia semakin menurun. Sehingga hari ini, kadar melengkapkan Vaksinasi COVID-19 semakin meningkat dan ini bermaksud semakin ramai orang telah melengkapkan Vaksin COVID-19. Anti-Vaxxer, dan juga dikenali sebagai orang tidak bersetuju dengan penggunaan vaksin dan mereka cuba menyebarkan bahaya-bahaya Vaksin di dalam media sosial supaya dapat memberikan ramai orang anti vaksin bersama-sama dengan pelbagai alasan, tetapi juga ada sekumpulan orang yang menyokong Vaksinasi COVID-19 juga menyebarkan perasaan mereka terhadap kebaikan vaksinasi COVID-19 di dalam media sosial. Oleh itu, ini akan menyebabkan pertambahan sentimen terhadap Vaksinasi COVID-19 yang ada di dalam media sosial.

Kesimpulannya, kajian ini dicadangkan kerana cabaran yang dihadapi oleh pengkaji untuk menganalisis sentimen terhadap Vaksinasi COVID-19 dalam era pandemic COVID-19 ini. Pendekatan secara analisis sentimen dicadangkan untuk mendapat data, iaitu tweets yang berkaitan dengan Vaksinasi COVID -19 di Twitter dan menganalisis tweet tersebut bagi mengkaji pandangan awam terhadap perkara tentang Vaksinasi COVID-19. Sentimen-sentimen ini akan dibahagikan kepada beberapa kumpulan, iaitu sentimen positif, semtimen neutral dan juga sentimen negatif. Hasil akhir kepada kajian ini dapat membantu pihak yang berkenaan dapat mengetahui bahawa sentimen terhadap pandangan awam dan dapat membuat keputusan untuk mengubahkan sudut pandangan terhadap Anti-Vaxxer supaya semua orang dapat melengkapkan vaksinasi COVID-19 dan mengharapkan kehidupan dapat memulih secara biasa.

2 PENYATAAN MASALAH

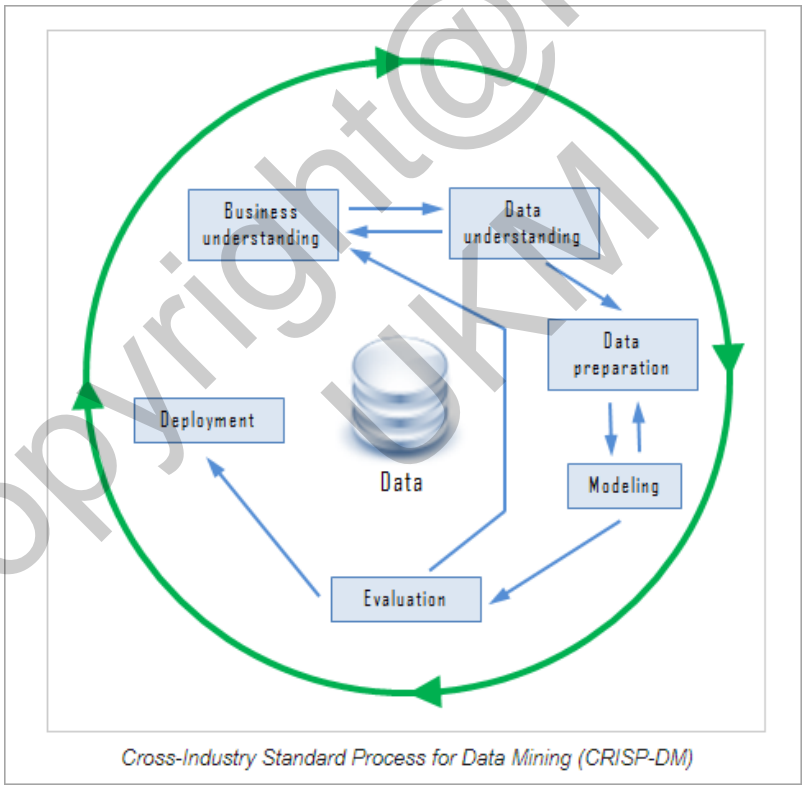
Melalui pemerhatian yang dijalankan sebelum melakukan kajian tersebut, perkara terhadap Vaksinasi COVID-19 amat popular di dalam media sosial Twitter. Hal ini kerana mempunyai orang yang bersetuju dan juga tidak bersetuju bahawa pemvaksinan vaksinasi COVID-19 ini. Orang yang bersetuju terhadap vaksinasi COVID-19 akan menyebarkan kebaikan vaksinasi di media sosial manakala Anti-vaxxer, yang merupakan orang yang tidak bersetuju dengan penggunaan vaksin cuba menyebarkan bahaya vaksinasi di media sosial, seperti Twitter. Oleh itu, ini akan menyebabkan peningkatan jumlah tweet sentimen yang amat banyak seperti sentimen positive, sentimen negative dan juga neutral. Perkara ini telah mendatangkan cabaran kepada semua orang kerana tidak dapat menganalisasi bahawa kebaikan atau keburukan kepada Vaksinasi COVID-19 dan juga memdatangkan cabaran kepada seseorang pengkaji yang ingin menganalisisi sentimen terhadap vaksinasi COVID-19 ini. Disebabkan data kepada sentimen vaksinasi sukar untuk dianalisai dan ini akan menyebabkan pihak berkepentingan kepada semua negara sukar untuk mendapat gambaran keseluruhan kepada perkara tersebut.

3 OBJEKTIF KAJIAN

Objektif kajian kepada tajuk Pengelasan Sentimen terhadap Vaksinasi COVID-19 dengan menggunakan Pembelajaran Mesin adalah dapat membangunkan model untuk mengelaskan sentimen terhadap Vaksinasi COVID-19 yang terdapat di dalam media sosial dengan menggunakan algoritma pembelajaran mesin dan dapat membangunkan satu sistem visualisasi pengelasan sentimen terhadap Vaksinasi COVID-19.

4 METOD KAJIAN

Metodologi merupakan proses yang amat penting untuk memastikan sesuatu kajian dapat berjalan dengan lancar dengan mengikut fasa-fasa yang telah ditetapkan. Untuk kajian Pengelasan Sentimen terhadap Vaksinasi COVID-19 dengan Menggunakan Pembelajaran Mesin, metodologi yang digunakan sepanjang tempoh dalam kajian tersebut ialah *Cross Industry Standard Process for Data Mining (CRISP-DM). CRISP-DM* ialah model proses dengan eman fasa yang secara semula jadi dengan menerangkan kitaran hayat sains data dan *CRISP-DM* juga dapat membantu untuk merancang, mengatur dan melaksanakan projek sains data atau pembelajaran mesin. Terdapat enam fasa di dalam *CRISP-DM* ini ialah Pemahaman Perniagaan (*Business Understanding*), Pengetahuan Data (*Data Understanding*), Penyediaan Data (*Data Preparation*), Pemodelan (*Modelling*), Penilaian (*Evaluation*) dan Penggunaan (*Deployment*) seperti yang ditunjukkan pada Rajah 4.1.



Rajah 4.1: 6 Fasa CRISP-DM

4.1 Fasa Pemahaman Perniagaan (*Business Understanding*)

Fasa Pemahaman Perniagaan (Business Understanding) adalah fasa untuk memahami tujuan dan kebutuhan dari sudut pandangan perniagaan dan mengenalpasti objektif utama kepada kajian. Dalam kajian ini, objektif utamanya adalah untuk membangunkan sebuah modal pengelasan sentimen tentang Vaksinasi COVID-19 melalui media sosial dan juga membina

sebuah sistem visualisasi pengelasan dengan menggunakan keputusan yang telah dianalisis. Sebelum memulakan kajian tersebut, pelbagai laporan kajian, artikel, jurnal dan bulletin perlu dikumpulkan supaya dapat mengetahui bahawa situasi semasa tentang Vaksinasi COVID-19. Selepas mengetahui situasi-situasi semasa, kajian ini akan dicadangkan dengan menggunakan pelbagai algoritma pembelajaran mesin seperti Naïve Bayes, Support Vector Machine (SVM), Random Forest dan Logistic Regression untuk menganalisis sentimen terhadap Vaksinasi COVID-19. Dalam fasa ini, kriteria yang akan digunakan untuk menentukan sama ada projek ini Berjaya dari sudut perniagaan perlu dikenalpasti. Jika data-data yang telah dianalisis dapat dikelaskan mengukut sentimen masing-masing, iaitu sentimen positif, sentimen neutral dan sentimen negatif, kajian tersebut akan dianggapka berjaya dan dapat diterima oleh pengguna.

4.2 Fasa Pengetahuan Data (*Data Understanding*)

Fasa Pengetahuan Data (*Data Understanding)* adalah fasa untuk mengumpulkan data yang perlu digunakan untuk kajian Pengelasan Sentimen terhadap Vaksinasi COVID-19. Media Sosial yang dipilih adalah Twitter dan *tweets* yang berkaitan tentang perkara vaksinasi COVID-19 perlu didapatkan melalui media sosial Twitter. Dalam kajian ini, data-data sentimen perlu diekstrak dengan menggunakan sambungan Twitter dangan pengesahan melalui API. Fasa ini juga memerlukan data untuk diterangkan dari segi format, kuantiti dan ciri permukaan lain yang ditemui. Selain itu, memeriksaan kepada semua data adalah diperlukan kerana perlu memastikan bahawa data yang dikumpul adalah lengkap.

Sebagai Contoh Tweets adalah:

1. Got cancer? Try some tea tree oil. Hepatitis? Some peppermint will sort you. MS? Rub yourself down with jojoba. https://t.co/V2jh5warIc
2. @BeardedGenius Hepatitis C Trust is a peer led charity raising awareness by going to the heart of the community. There is now a cure for Hep C but sadly alot of those affected are not accessing treatment. The Hep C trust tries to identify those affected and support them into treatment.
3. Hepatitis A outbreak linked to Long Beach steakhouse, health officials say https://t.co/U5hCv37BSM https://t.co/zG9jcSnfp8
4. Prevalence and characteristics of hypoxic hepatitis in the largest single-centre cohort of avian influenza A(H7N9) virus-infected patients with severe liver impairment in the intensive care unit https://t.co/jvvjVfsdIB

5. In response to the outbreak, Philly health officials implemented targeted outreach to at-risk populations, including people who inject drugs, are experiencing homelessness or live in high-risk areas. https://t.co/lX1wQdXFP3

Jadual 4.1: Contoh Atribut Set Data tweets COVID-19

Atribut	Keterangan	Format
tweet	Tweet yang terdapat di Twitter tentang perkara Vaksinasi Covid-19.	string

4.3 Fasa Penyediaan Data (*Data Preparation*)

Fasa Penyediaan Data (Data Preparation) adalah fasa untuk mengemaskan data-data sentimen yang telah dikumpul semasa Fasa Pengetahuan Data. Semasa Fasa Penyediaan Data, data yang dikumpul perlu melalu pra-pemprosesan, seperti pembersihan data, penyepaduan data, transformasi data dan pengurangan data. Proses pra-pemprosesan ini amat penting kerana proses ini ialah proses yang dapat mengubah data mentah kepada format yang boleh difahami oleh mesin dan pra-pemprosesan ini juga boleh membuangkan data-data yang tiada berkaitan tetapi salah dikumpul. Oleh itu, melalui fasa tersebut, ketepatan model akan meningkat.

4.4 Fasa Pemodelan (Modelling)

Fasa Pemodelan (Modelling) adalah fasa untuk membangunkan model yang berkaitan dengan kajian. Dalam kajian Pengelasan Sentimen terhadap Vaksinasi COVID-19 dengan Menggunakan Pembelajaran Mesin, sebuah model akan dibangunkan yang berkait dengan kajian tersebut. Dengan menggunakan Python sebagai platform utama untuk membangunkan model serta disokong dengan algoritma pembelajaran mesin, model ini dapat mengelaskan setiap sentimen yang telah diekstrak mengikut sentimen masing-masing, iaitu sentimen positif, sentimen neutral dan sentimen nagatif. Sebelum model ini debangunkan, ia perlu dilatih dengan menggunakan set latihan. Hal ini kerana model yang dilatih dengan bilangan yang banyak, ketepatan kepada model akan lebih tinggi.

4.5 Fasa Penilaian (Evaluation)

Fasa Penilaian (*Evaluation*) adalah fasa untuk menilai keputusan yang dianalisis melalui model yang telah dibangun. Semasa fasa penilaian, keputusan yang diperolehi dari model yang telah dibangunkan akan direkod dan disimpan. Selepas mendapatkan keputusan tersebut, keputusan ini akan dianalisis untuk menentukan sama ada kajian tersebut dapat

memenuhi objektif yang dinyatakan semasa Fasa Pemahaman Perniagaan iaitu pengelasan sentimen tentang Vaksinasi COVID-19. Selain itu, ulasan projek perlu dilakukan untuk membuat sebuah ringkasan kepada keseluruhan proses dan menyatakan proses yang perlu diulangi semula. Selepas itu, keputusan akan dibuat jika keputusan projek sudah memadai.

4.6 Fasa Penggunaan (Deployment)

Fasa Fasa Penggunaan (Deployment) adalah fasa terakhir kepada Cross Industry Standard Process for Data Mining (CRISP-DM) dan fasa ini akan mengeluarkan sebuah kajian yang telah lengkap. Dalam fasa ini, model kepada Pengelasan Sentimen terhadap Vaksinasi COVID-19 telah dibangunkan dan telah memenuhi objektif yang dinyatakan sebelum ini. Selain itu, laporan akhir yang mengandungi semua hasil sebelumyan, iaitu keputusan keseluruhan yang telah direkod dan telah dianalisis, ringkasan dan mangatur hasilnya yang perlu direkodkan. Di samping itu, kepada pembentangan akhir, membina sebuah sistem visualisasi pengelasan dengan menggunakan keputusan yang telah dianalisis amat diperlukan kerana ini akan memudahkan semua orang faham dan melihatmya. Akhir sekali, pengalaman dokumentasi perlu dihasilkan untuk merekod pengalaman penting yang diperoleh semasa membuat kajian ini.

5 HASIL KAJIAN

Bab ini mengandungi perbincangan mengenai Pengelasan sentimen terhadap vaksinasi Covid-19 dengan menggunakan pembelajaran mesin dan juga mengujikan proses-proses yang dilakukan terhadap projek yang telah dibangunkan. Skop kepada bab ini adalah lebih terperinci berkenaan dengan proses pembangunan projek dari segi pengaturcaraan pemprosesan data dengan menggunakan Python dan juga dapat menerangkan tentang perpustakaan Python yang dapat digunakan dalam projek tersebut.

Selepas implementasi, pengujian juga amat penting dalam bab ini. Bagi pengajian pula, ia juga merupakan satu aktiviti yang dilakukan untuk memeriksa bahawa sistem keputusan adalah sama atau hampir sama dengan keputusan yang dijangkakan. Selain itu, pengujian juga perlu dijalankan supaya dapat memastikan sistem yang telah dibangunkan dalam keadaan yang baik dan tidak mempunyai sebarang masalah. Pengujian ini juga dilakukan berdasarkan komponen perisian atau komponen sistem.

5.1 EKSTRAK DATA

Dalam projek pengelasan sentimen terhadap vaksinasi Covid-19 dengan mengguunakan pembelajaran mesin, set data perlu diekstrak melalui media sosial *Twitter*. Cara untuk mengekstrak tweets dari Twitter sebagai set data adalah melalui sambungan Twitter dengan pengesahan melalui API dan API boleh didapat melalui Platform Pembangunan Twitter. Selepas mendapatkan Twitter API, Python akan digunakan untuk mengekstrak data.

Kepada proses mengekstrak data, import perpustakaan adalah penting kerana perpustakaan mempunyai banyak fungsi dan boleh digunakan dalam proses tersebut seperti perpustakaan *tweepy* dan *csv*.

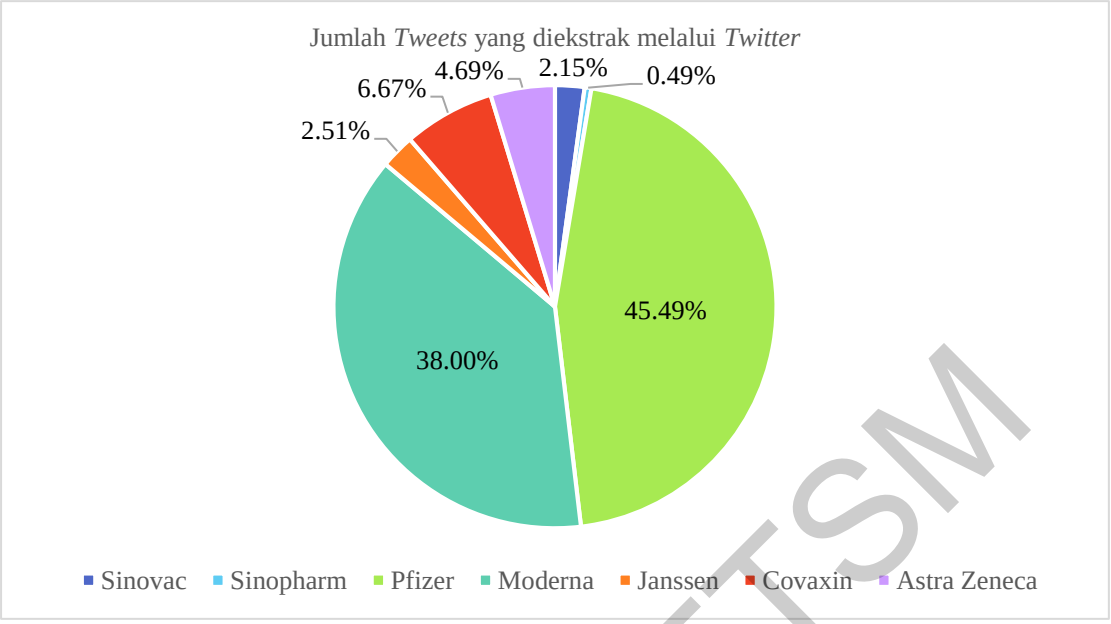
Memeriksa kelayakan *Twitter API* adalah penting kerana *Twitter API* yang tidak layak tidak boleh mengekstrak data. Selepas memeriksa kelayakan *Twitter API*, membuat objek pengesahan dan menghantar maklumat pengesahan adalah penting supaya *Twitter API* boleh digunakan kepada proses seterusnya.

Selepas mendapat kelayakan menggunakan *Twitter API*, *Markdown* akan digunakan untuk mendapatkan set data. *Markdown* mempunyai fungsi yang amat sama dengan *HTML*, tetapi *Markdown* boleh membuat antara muka di dalam *Google Colab* dan berbanding dengan *HTML*, *Markdown* amat ringan.

Apabila set data telah diekstrak melalui Markdown, Dataframe akan digunakan untuk menjadikan set data dalam bentuk *dataframe* dan menyimpan set data dalam jenis *.csv* untuk menjalankan proses seterusnya. Terdapat 8 *dataframe* yang telah disimpan dalam jenis *.csv* iaitu *vaccine.csv, moderna.csv, pfizer.csv, az.csv, sinopharm.csv, sinovac.csv, janssen.csv dan covaxin.csv*. Set data Vaksinasi sebagai set data yang paling besar mempunyai sebanyak 18852 *tweets* digunakan untuk menganalisis sentimen terhadap Vaksinasi, manakala set data mengikut jenis vaksin digunakan untuk menganalisiskan sentimen terhadap jenis vaksin. Jenis vaksin yang berbeza mempunyai perbezaan tahap perbincangan di dalam *tweets*.

Jadual 5.1 : Bilangan Tweets yang didapat mengikut jenis vaksin

Jenis Vaksin	Jumlah Tweets
Sinovac	844
Sinopharm	193
Pfizer	17878
Moderna	14937
Janssen	987
Covaxin	2621
Astra Zeneca	1845



Rajah 5.1: Carta Pie kepada peritus bilangan tweets yang didapat mengikut jenis vaksin

Berdasarkan Jadual 5.1 dan Rajah 5.1 menunjukkan bilangan dan peritus *tweets* yang didapati mengikut jenis vaksin. Vaksin Pfizer telah dapat bilangan *tweets* yang paling banyak, iaitu sebanyak 17878 tweets dan mempunyai sebanyak 45.49% daripada semua jenis vaksin dan seterusnya merupakan vaksin Moderna, iaitu sebanyak 14937 dan 38% daripada semua jenis vaksin. Manakala bilangan *tweets* kepada vaksin Sinopharm adalah paling sedikit, hanya 193 tweets dan 0.49% daripada semua jenis vaksin.

5.2 PRA-PEMPROSESAN DATA

Dalam analisis sentimen, proses pra-pemprosesan data amat penting dan diutamakan kerana ianya mempengaruhi ketepatan dan prestasi pengelas. Hal ini kerana sekiranya proses ini tidak dijalankan secara optimum, maka ianya akan memaparkan keputusan kepada analisis sentimen yang tidak memuaskan.

Sebelum melakukan proses pra-pemprosesan, beberapa perpustakaan yang perlu diimportkan seperti *numpy*, *pandas* dan *re*. Selepas mengimport perpustakaan, set data perlu dibacakan dengan menggunakan fungsi *pandas*, iaitu *pd.read_csv()*.

Selepas membacakan set data yang perlu digunakan, pra-pemprosesan data akan dijalankan seperti menyingkirkan ruang putih dengan menggunakan *strip()* , tukar semua *tweets* dalam huruf kecil dengan menggunaka *lower()*, menyingkirkan *URL*, *hashtag*, *mention*, bukan abjab angka dan tanda baca dengan menggunakan *replace()*. Emoji juga perlu menyingkirkan dengan menggunakan str.encode().

Tambahan pula, menyingkirkan kata henti dan lemmatisasi juga penting semasa menjalankan proses pra-pemprosesan data. Hal ini kerana kata henti ialah perkataan yang tidak mempunyai sebarang emosi dan tiada berguna dalam analysis sentimen. Lemmatisasi juga bertujuuan untuk menghapuskan pengakhiran infleksi sahaja dan mengembalikan bentuk pangkal atau kamus sesuatu perkataan.

Selain itu, fungsi *drop_duplicates()* juga perlu digunakan untuk menyingkirkan data-data yang sama dalam set data. Hal ini kerana terdapat data yang sama dalam set data yang sama akan menyebabkan ketepatan keputusan semasa menjalankan analisis sentimen.

Akhir sekali, selepas menjalankan pra-pemprosesan data, set data yang telah menjalankan pra-pemprosesan perlu disimpan dalam bentuk *.csv*.

Pra-pemprosesan data akan diulangkan dengan menggunakan set data yang berbeza, iaitu *vaccine.csv, moderna.csv, pfizer.csv, az.csv, sinopharm.csv, sinovac.csv, janssen.csv dan covaxin.csv*.

5.3 ANALISIS SENTIMEN

Dengan menggunakan perpustakaan *NLTK* seperti *Textblob* dan juga *VADER*, proses menganalisis sentimen akan menjadi semakin mudah. NLTK ialah rangkaian perpustakaan dan program untuk pemprosesan bahasa semula jadi simbolik dan statistik untuk bahasa Inggeris yang ditulis dalam bahasa pengaturcaraan Python. Hal ini kerana perpustakaan ini sudah tersiap sedia di dalam nltk dan ianya sangat berguna untuk membuat sebarang analisis sentimen yang menggunakan bahasa Inggeris.

Textblob ialah analisis sentimen berasaskan leksikon dan juga adalah perpustakaan yang menyokong analisis dan operasi yang kompleks pada data teks. Untuk pendekatan berdasarkan leksikon, sentimen didefinisikan oleh orientasi semantiknya dan intensiti setiap perkataan dalam ayat. Hasil keputusan kepada TextBlob akan mengembalikan kekutuban dan subjektiviti ayat.

Valence Aware Dictionary and Sentimen Reasoner (*VADER*) ialah analisis sentimen berasaskan leksikon dan *VADER* juga adalah model yang digunakan untuk analisis sentimen yang sensitive terhadap kekutuban iaitu positif, negatif dan intensity emoji kepada sesuatu ayat. Skor sentimen teks dapat diperoleh dengan menjumlahkan intensiti setiap perkataan dalam teks.

5.4 PEMBANGUNAN PENGELASAN

Bagi membangunkan pengelasan sentimen terhadap Vaksinasi Covid-19 dengan menggunakan pembelajaran mesin. Perpustakaan yang perlu digunakan adalah *Numpy* kerana *Numpy* merupakan pakej asa dalam *Python* untuk memproseskan data. Dalam kajian ini, set data vaksinasi dan set data mengikut jenis vaksin, iaitu set data az, set data covaxin, set data janssen, set data moderna, set data pfizer, set data sinopharm dan set data sinovac akan digunakan untuk menganalisis sentimen.

Dalam pembangunan pengelasan sentimen kepada kajian ini, set data telah dibahagikan kepada 2 set iaitu set data latihan dan set data ujian. Set data latihan digunakan untuk memasukkan data yang telah dilabel dengan sentimen ke dalam model bagi tujuan melatih model pengelasan terlebih dahulu. Selain itu, penilaian dibuat oleh pengelas untuk mengklasifikasikan data kepada kelas positif bagi yang memberikan sentimen yang positif, kelas negatif bagi yang memberikan sentimen negatif dan neutral bagi yang tidak memberikan sebarang sentimen.

Dalam projek pengelasan sentimen terhadap Vaksinasi Covid-19 dengan menggunakan pembelajaran mesin. Terdapat 4 jenis algoritma pembelajaran mesin akan digunakan untuk mengklasifikasi tweets iaitu *Naïve Bayes, Support Vector Machine (SVM), Logistic Regression* dan *Random Forest*. Melalui 4 jenis algoritma pembelajaran mesin, setiap algoritma akan memberikan keputusan yang berbeza. Oleh itu, perbandingan antara pengelas perlu dilakukan supaya model mempunyai keputusan yang lebih tepat dan dapat dipilih sebagai model akhir.

Dengan menggunakan perpustakaan sklearn, algoritma *Support Vector Machine (SVM), SVC(), Logistic Regression, LogisticRegression(), Naïve Bayes, MultinomialNB()* dan *Random Forest, RandomForestClassifier()* akan dipanggil dengan mudah. Kepada *Naïve Bayes,* ia mempunyai 3 jenis, iaitu Gaussian, Multinomial dan Bernouilli. Multinomial telah dipilih dalam projek ini kerana ia merupakan model yang paling sesuai dalam projek ini. Untuk melatih model ini, fungsi .fit() digunakan untuk memasukkan data yang telah diasingkan kepada set latihan dan set ujian ke dalam model. Selepas menyelesaikan latihan, model pengelas telah disediakan untuk meramalkan keputusan bagi set ujian. Kepada fungi .predict(), keputusan akan disimpan di dalam prediction. Bagi mengira prestasi untuk model pengelas, fungsi .score() digunakan untuk mendapat ketepatan min pada set ujian dan label yang diberi.

5.5 PENGESANAN PRESTASI ALGORITMA PEMBELAJARAN MESIN

Kepada projek pengelas sentimen terhadap vaksinasi Covid-19 dengan menggunakan pembelajaran mesin. Sentimen tweets dikesan dengan menggunakan positif benar (tp), positif palsu (fp), negatif benar (tn) dan negatif palsu (fn). Dengan menentukan bahawa positif benar merupakan kelas bagi tweets yang mengandungi sentimen positif manakala positif palsu merupakan kelas bagi tweets mempunyai kemungkinan mempunyai sentimen positif tetapi tidak dilabel sebagai positif. Negatif benar merupakan kelas bagi tweets yang mengandungi sentimen negatif manakala negatif palsu merupakan kelas bagi tweets mempunyai kemungkinan mempunyai sentimen negatif tetapi tidak dilabel sebagai negatif. Dalam projek ini, ketepatan kejituan, dapatan dan markah F1 akan digunakan dalam mengira ketepatan analisis data.

Dengan menggunakan classification_report(), hasil keputusan model pengelas dapat dilihat dan juga dapat dianalisi dengan ketepatan, kejituan, dapatan dan markah F1. Pertama sekali ialah ketepatan dan ia dapat dikita dengan menggunakan formula seperti berikut:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Ketepatan ialah ukuran prestasi yang paling intuitif dan ia hanyalah nisbah pemerhatian yang diramalkan dengan betul kepada jumlah pemerhatian. Ketepatan ialah ukuran yang amat bagus tetapi hanya apabila mempunyai set data yang nilai positif palsu dan negatif palsu hampir sama. Oleh itu, parameter lain juga penting untuk menilai prestasi pembelajaran mesin seperti kejituan, dapatan dan juga markah F1.

Tambahan pula adalah kejituan dan ia dapat dikira dengan menggunakan formula seperti berikut:

$$Precision = \frac{TP}{TP + FP}$$

Kejituan ialah nisbah pemerhation positif yang diramalkan dengan betul kepada jumlah pemerhatian positif yang diramalkan. Sebagai contoh, tweets yang dilabelkan positif, berapa banyak tweets yang sebenarnya positif. Kejituan yang tinggi berkaitan dengan kadar positif palsu yang rendah.

Selain itu, dapatan juga dapat dianalisis dan ia dapat dikira dengan menggunakan formula tersebut:

$$Recall = \frac{TP}{TP + FN}$$

Dapatan ialah nisbah pemerhatian positif yang diramalkan dengan betul kepada semua pemerhatian dalam set data yang sebenar. Sebagai contoh, berapa tweets yang telah dilabel adalah betul. Dapatan yang melebihi 0.5 dianggap bagus.

Seterusnya, Markah F1 dapat dianalisis dan ia dapat dikira dengan menggunakan formula tersebut:

$$F1 = 2 * \frac{(Percision * Recall)}{(Percision + Recall)}$$

Markah F1 ialah purate wajaran kejituan dan dapatan. Oleh itu, markah ini mengambil kira positif palsu dan negative palsu. Markah F1 tidak semudah untuk difahami seperti ketetapan, tetapi markah F1 biasanya lebih berguna berbanding dengan ketetapan, terutamanya jika mempunyai pengedaran kelas yang tidak sekata.

Apabila setiap model pengelas mendapat keputusan kepada ketapatan, kejituan, dapatan dan markah F1, maka pengelas-pengelas ini telah bersedia untuk dianalisis bagi mengenalpasti kebaikan kepada model pengelas dan manakah yang akan memberi keputusan yang lebih baik berbanding dengan model pengelas lain.

5.6 PEMILIHAN PAKEJ PERPUSTAKAAN NLTK MENGIKUT PRESTASI

Terdapat dua pakej perpustakaan yang telah digunakan di dalam projek ini. Kedua-dua pakej ini memainkan peranan yang amat penting untuk menentukan sentimen tweets sama ada positif, neutral atau negatif dan ianya adalah pakej perpustakaan *Textblob* dan *VADER*. *TextBlob* adalah perpustakaan yang menyokong analisis dan operasi yang kompleks pada data teks. Untuk pendekatan berdasarkan leksikon, sentimen didefinisikan oleh orientasi semantiknya dan intensiti setiap perkataan dalam ayat. *TextBlob* mengembalikan kekutuban dan subjektiviti ayat. *VADER* adalah model yang digunakan untuk analisis sentimen teks yang sensitif terhadap kekutuban iaitu positif dan negatif, dan intensiti emosi kepada sesuatu ayat. Skor sentimen teks dapat diperoleh dengan menjumlahkan intensiti setiap perkataan dalam teks.

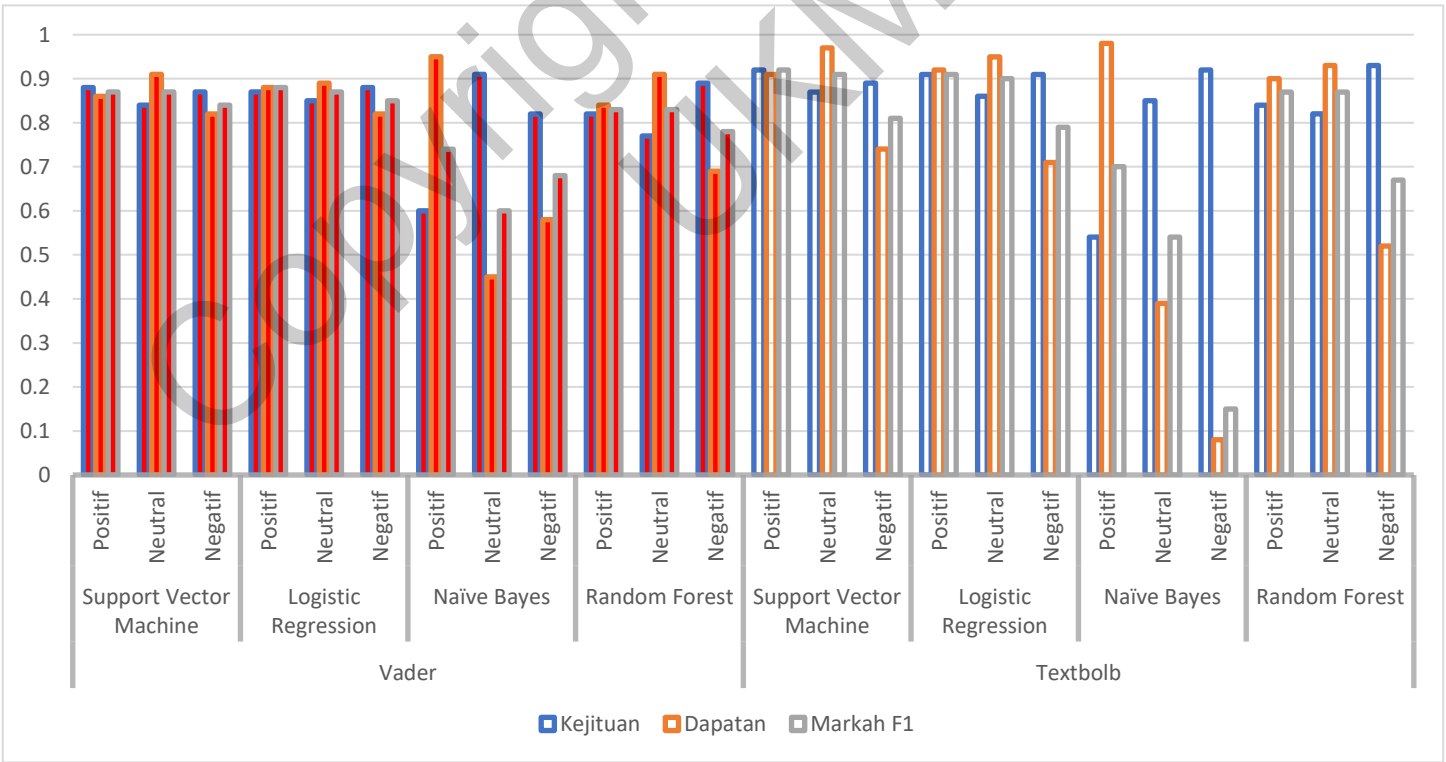
Selepas proses analisis sentimen dijalankan, semua set data yang telah dilabelkan dengan sentimen akan digunakan untuk menjalankan pengelasan dengan menggunakan algoritma *Support Vector Machine (SVM), Logistic Regression, Naïve Bayes* dan *Random Forest*.

Untuk menganalisiskan ketepatan kepada kedua-dua pakej perpustakaan, iaitu *Textblob* dan *Vader*, set data Vaksinasi telah dipilih untuk menjalankan analisis pengelasan

dulu kerana set data Vaksinasi merupakan set data yang terbesar dalam kajian ini dan data yang terdapat dalam Set data Vaksinasi mempunyai sebanyak 18852 *tweets*. Set latihan dan set ujian bagi kajian ini ialah 80% set latihan dan 20% set ujian.

Jadual 5.2: Perbandingan prestasi kepada *VADER* dan *Textblob* (Set Data Vaksinasi)

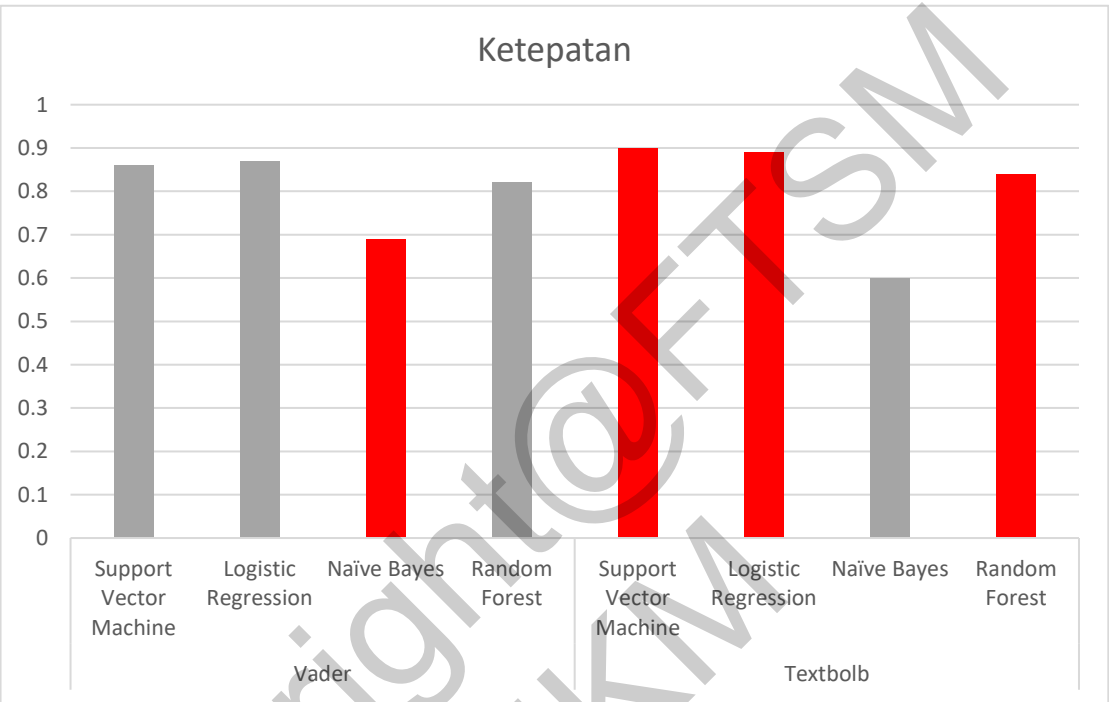
VADER												
	SVM			Logistic Regression			Naïve Bayes			Random Forest		
	Positif	Neutral	Negatif	Positif	Neutral	Negatif	Positif	Neutral	Negatif	Positif	Neutral	Negatif
Kejituan	0.88	0.84	0.87	0.87	0.85	0.88	0.6	0.91	0.82	0.82	0.77	0.89
Dapatan	0.86	0.91	0.82	0.88	0.89	0.82	0.95	0.45	0.58	0.84	0.91	0.69
Markah F1	0.87	0.87	0.84	0.88	0.87	0.85	0.74	0.6	0.68	0.83	0.83	0.78
Textblob												
	SVM			Logistic Regression			Naïve Bayes			Random Forest		
	Positif	Neutral	Negatif	Positif	Neutral	Negatif	Positif	Neutral	Negatif	Positif	Neutral	Negatif
Kejituan	0.92	0.87	0.89	0.91	0.86	0.91	0.54	0.85	0.92	0.84	0.82	0.93
Dapatan	0.91	0.97	0.74	0.92	0.95	0.71	0.98	0.39	0.08	0.9	0.93	0.52
Markah F1	0.92	0.91	0.81	0.91	0.9	0.79	0.7	0.54	0.15	0.87	0.87	0.67



Rajah 5.2: Graf perbandingan prestasi kepada *VADER* dan *Textblob* (Set Data Vaksinasi)

Jadual 5.3: Perbandingan Ketepatan kepada *VADER* dan *Textblob* (Set Data Vaksinasi)

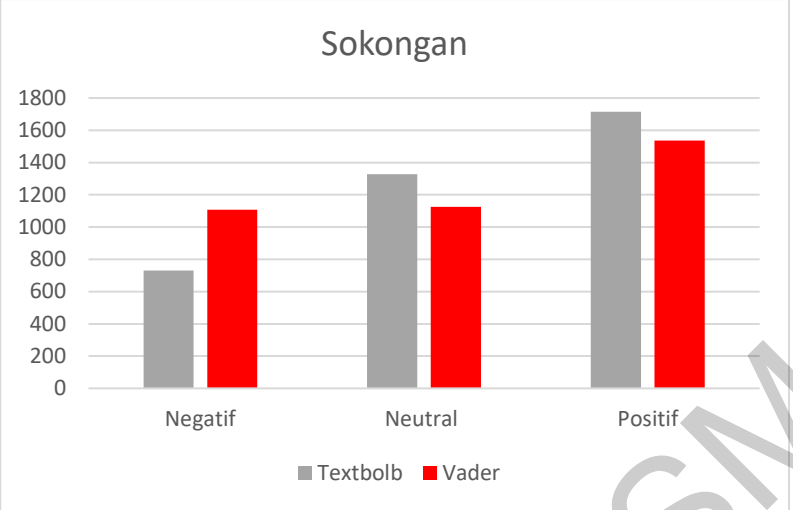
	VADER			
	SVM	Logistic Regression	Naïve Bayes	Random Forest
Ketepatan	0.86	0.87	0.69	0.82
	Textblob			
	SVM	Logistic Regression	Naïve Bayes	Random Forest
Ketepatan	0.9	0.89	0.6	0.84



Rajah 5.3: Graf perbandingan Ketepatan kepada *VADER* dan Textblob (Set Data Vaksinasi)

Jadual 5.4: Perbandingan Sokongan kepada *VADER* dan *Textblob* (Set Data Vaksinasi)

	VADER		
	Positif	Neutral	Negatif
Sokongan	1537	1126	730
	Textblob		
	Positif	Neutral	Negatif
Sokongan	1714	1327	730



Rajah 5.4: Graf perbandingan Sokongan kepada *VADER* dan *Textblob* (Set Data Vaksinasi)

Berdasarkan jadual 5.2 dan jadual 5.3, kepada perbandingan antara pakej perpustakaan *NLTK*, *VADER* dan *Textblob*, hasil keputusan perhitungan kejituan, dapatan, markah F1 dan ketepatan bagi pakej perpustakaan *Textblob* adalah lebih tinggi berbanding dengan pakej perpustakaan *VADER*. Akan tetapi, jika dilihat pada bahagian kelas sentimen bagi positif dan negatif yang ditunjukkan pada jadual 5.4, *VADER* mempunyai perhitungan yang lebih tinggi berbanding dengan *Textblob*. Fungsi yang dibekalkan dalam pakej *VADER* membantu dalam pengiraan skor sentimen. Perkara ini juga berjaya meningkatkan ketepatan keputusan yang diperoleh dan *VADER* akan digunakan bagi menjalankan proses analisis sentimen data yang seterusnya.

5.7 PENGUJIAN PRESTASI MODEL PENGELASAN KEPADA SET DATA VAKSINASI

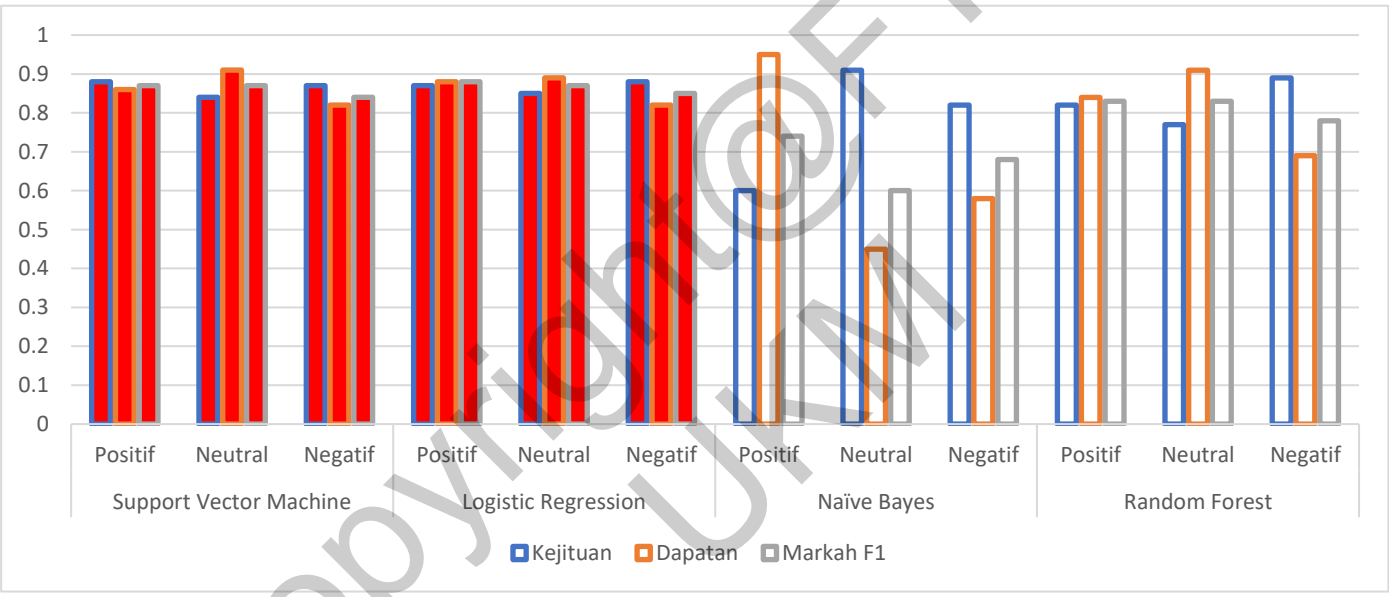
Selepas menguji perhitungan ketepatan kepada pakej perpustakaan *NLTK*, *VADER* akan digunakan untuk menjalankan proses analisis sentimen data yang seterusnya, iaitu menggunakan set data yang telah dilabelkan dengan menggunakan pakej perpustakaan *VADER* dan menguji prestasi model pengelas kepada 4 algoritma pembelajara mesin iaitu *Support Vector Machine (SVM)*, *Logistic Regression*, *Naïve Bayes* dan *Random Forest*.

Set data Vaksinasi yang telah dilabelkan dengan menggunakan pakej perpustakaan *VADER* akan digunakan untuk menganalisis prestasi kepada 4 model pengelas tersebut kerana set data Vaksinasi merupakan set data yang terbesar dalam kajian ini dan ketepatan tersebut akan lebih tinggi daripada set data yang kecil. Selepas menguji prestasi kepada 4 model pengelasan, 2 model yang terbaik akan digunakan untuk menganalisis dengan menggunakan set data seterusnya.

SET DATA VAKSINASI (KEJITUAN, DAPATAN DAN MARKAH F1)

Jadual 5.5: Perbandingan prestasi kepada pengelasan sentimen algoritma pembelajaran mesin (Set Data Vaksinasi)

Set Data Vaksinasi												
	SVM			Logistic Regression			Naïve Bayes			Random Forest		
	Positif	Neutral	Negatif	Positif	Neutral	Negatif	Positif	Neutral	Negatif	Positif	Neutral	Negatif
Kejituan	0.88	0.84	0.87	0.87	0.85	0.88	0.6	0.91	0.82	0.82	0.77	0.89
Dapatan	0.86	0.91	0.82	0.88	0.89	0.82	0.95	0.45	0.58	0.84	0.91	0.69
Markah F1	0.87	0.87	0.84	0.88	0.87	0.85	0.74	0.6	0.68	0.83	0.83	0.78

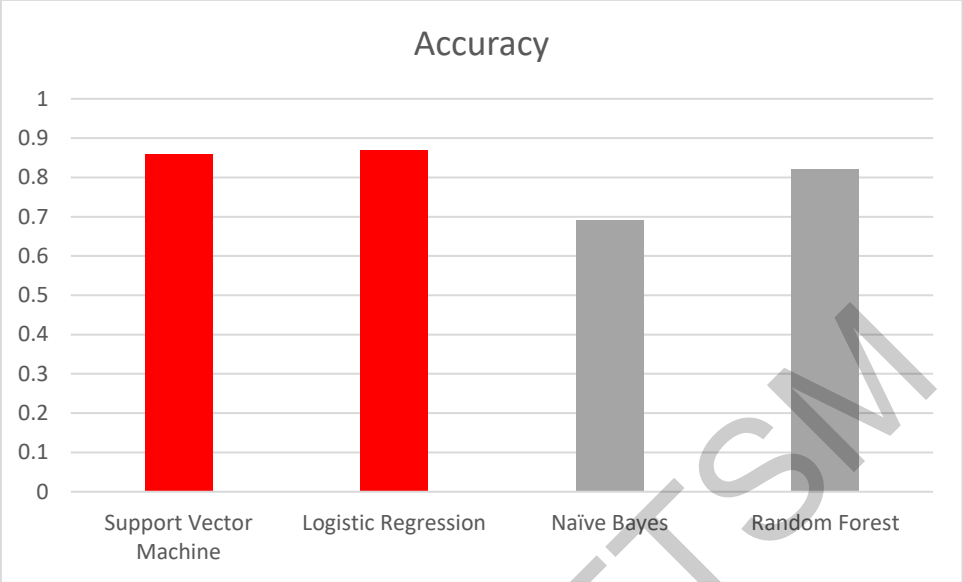


Rajah 5.5: Graf Perbandingan prestasi kepada pengelasan sentimen algoritma pembelajaran mesin (Set Data Vaksinasi)

SET DATA VAKSINASI (KETEPATAN)

Jadual 5.6: Perbandingan Ketepatan kepada pengelasan sentimen algoritma pembelajaran mesin (Set Data Vaksinasi)

	SVM	Logistic Regression	Naïve Bayes	Random Forest
Ketepatan	0.86	0.87	0.69	0.82



Rajah 5.6: Graf Perbandingan Ketepatan kepada pengelasan sentimen algoritma pembelajaran mesin (Set Data Vaksinasi)

Keputusan kepada markah F1 adalah keputusan yang paling penting untuk mengenalpasti model pengelas yang manakah yang akan memberikan keputusan yang lebih baik. Berdasarkan kepada jadual 5.5 dan jadual 5.6, markah F1 kepada *Support Vector Machine (SVM)* dan *Logistic Regression* lebih tinggi daripada *Naïve Bayes* dan *Random Forest*. *Naïve Bayes*. Markah F1 kepada *Support Vector Machine (SVM)* adalah paling tinggi dan seterusnya adalah *Logistic Regression*. Markah F1 kepada Naïve Bayes agak rendah, iaitu antara 0.6 hingga 0.74. Random Forest dapat markah F1 yang amat tinggi, tetapi amat rendah berbanding dengan *Support Vector Machine (SVM)* dan *Logistic Regression*. Oleh itu, *Support Vector Machine (SVM)* dan *Logistic Regression* akan digunakan bagi menjalankan proses analisis sentimen data kepada set data *Sinovac*, *Sinopharm*, *Pfizer*, *Moderna*, *Janssen*, *Covaxin* dan *Astra Zeneca* supaya dapat mengenalpasti algoritma pembelajaran mesin yang paling baik dan dapat memberikan keputusan yang paling bagus.

5.8 PENGUJIAN PRESTASI MODEL PENGELASAN KEPADA SET DATA MENGIKUT JENIS VAKSIN

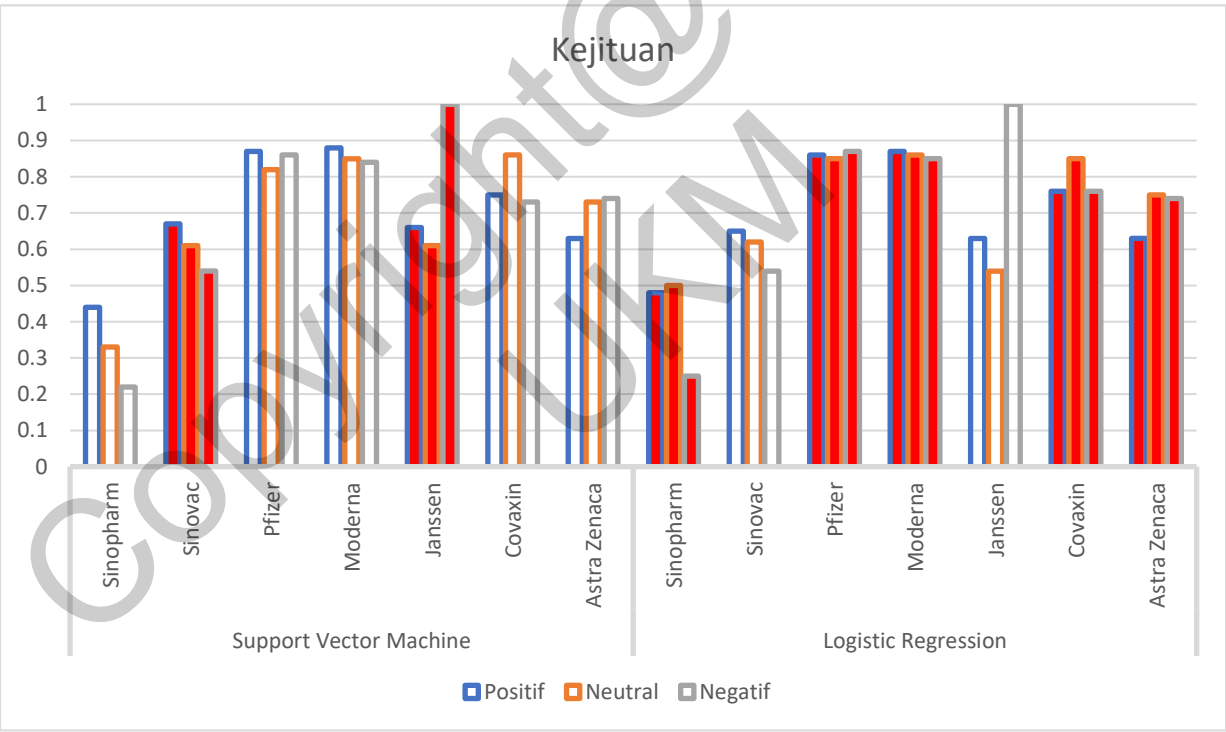
Selepas menjalankan analisis sentimen kepada set data vaksinasi, algoritma yang sesuai untuk digunakan dalam kajian ini adalah *Support Vector Machine (SVM)* dan Logistic Regression. Untuk menganalisis algoritma pembelajaran mesin yang paling terbaik, *Suport Vector Machine (SVM)* dan *Logistic Regression* akan digunakan untuk menganalisis sentimen kepada set data *Sinovac*, *Sinopharm*, *Pfizer*, *Moderna*, *Janssen*, *Covaxin* dan *Astra Zeneca*

supaya dapat mengenalpasti algoritma pembelajaran mesin yang paling baik dan dapat memberikan keputusan yang paling bagus.

SET DATA MENGIKUT JENIS VAKSIN (KEJITUAN)

Jadual 5.7: Perbandingan prestasi Kejituan kepada pengelasan sentimen algoritma pembelajaran mesin

Kejituan	Support Vector Machine						
	Sinopharm	Sinovac	Pfizer	Moderna	Janssen	Covaxin	Astra Zeneca
Positif	0.44	0.67	0.87	0.88	0.66	0.75	0.63
Neutral	0.33	0.61	0.82	0.85	0.61	0.86	0.73
Negatif	0.22	0.54	0.86	0.84	1	0.73	0.74
	Logistic Regression						
	Sinopharm	Sinovac	Pfizer	Moderna	Janssen	Covaxin	Astra Zeneca
Positif	0.48	0.65	0.86	0.87	0.63	0.76	0.63
Neutral	0.5	0.62	0.85	0.86	0.54	0.85	0.75
Negatif	0.25	0.54	0.87	0.85	1	0.76	0.74



Rajah 5.7: Graf Perbandingan prestasi Kejituan kepada pengelasan sentimen algoritma pembelajaran mesin

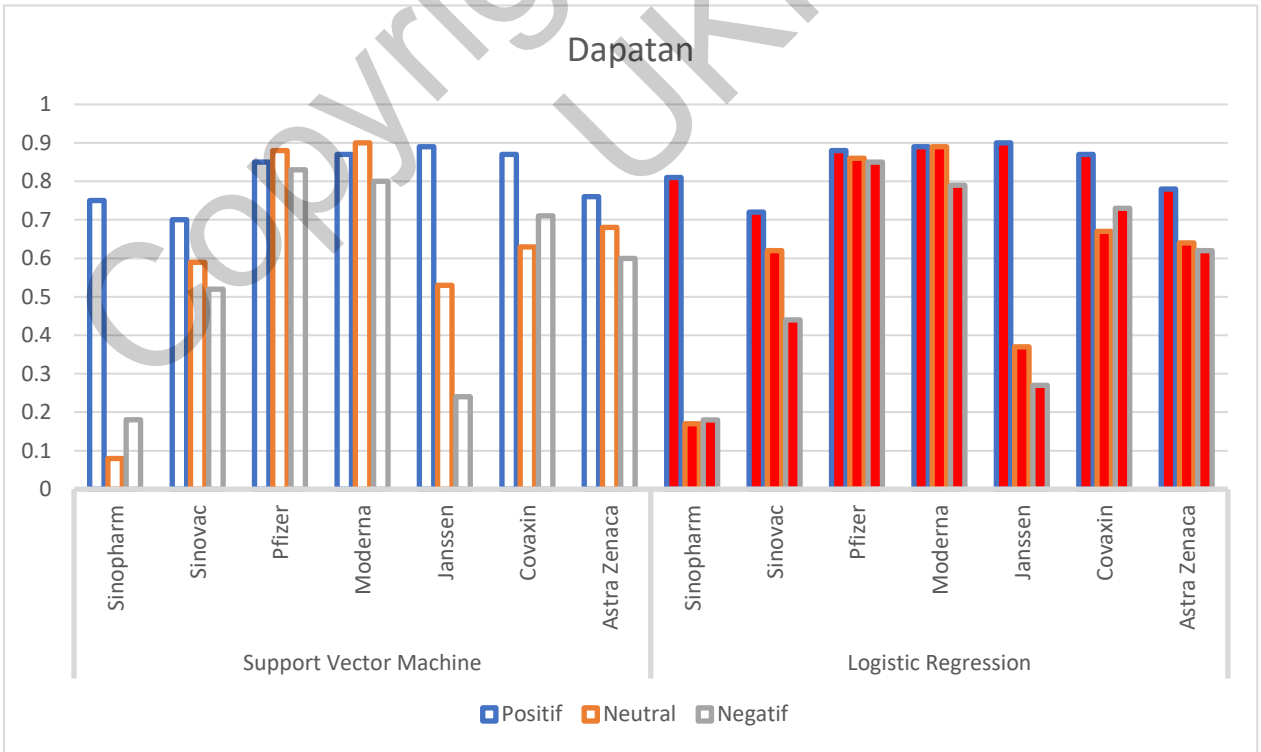
Berdasarkan jadual 5.7 dan rajah 5.7, keputusan kejituan yang diperolehi dengan *Logistic Regression* adalah lebih tinggi berbanding dengan *Support Vector Machine (SVM).* Dalam 7 set data mengikut jenis vaksinasi, keputusan kejituan yang diperolehi dengan Logistic Regression kepada set data *Sinopharm, Pfizer, Moderna, Covaxin* dan *Astra Zeneca* adalah lebih tinggi. Ini bermaksud *tweets* yang dilabeklan dengan sentimen majoriti betul,

Sebagai contoh, kepada set data *Pfizer*, tweets yang dilabelkan positif mempunyai 87% *tweets* yang sebenarnya positif jika menggunakan Logistic Regression, manakala 86% *tweets* yang sebenarnya positif jika menggunakan *Support Vector Machine (SVM)*. Oleh itu, dalam perbandingan antara *algoritma Support Vector Machine (SVM)* dan *Logistic Regression* dengan menggunakan keputusan kejituan, *Logistic Regression* adalah lebih baik berbanding dengan *Support Vector Machine (SVM).*

SET DATA MENGIKUT JENIS VAKSIN (DAPATAN)

Jadual 5.8: Perbandingan prestasi Dapatan kepada pengelasan sentimen algoritma pembelajaran mesin

Dapatan	Support Vector Machine						
	Sinopharm	Sinovac	Pfizer	Moderna	Janssen	Covaxin	Astra Zeneca
Positif	0.75	0.7	0.85	0.87	0.89	0.87	0.76
Neutral	0.08	0.59	0.88	0.9	0.53	0.63	0.68
Negatif	0.18	0.52	0.83	0.8	0.24	0.71	0.6
	Logistic Regression						
	Sinopharm	Sinovac	Pfizer	Moderna	Janssen	Covaxin	Astra Zeneca
Positif	0.81	0.72	0.88	0.89	0.9	0.87	0.78
Neutral	0.17	0.62	0.86	0.89	0.37	0.67	0.64
Negatif	0.18	0.44	0.85	0.79	0.27	0.73	0.62



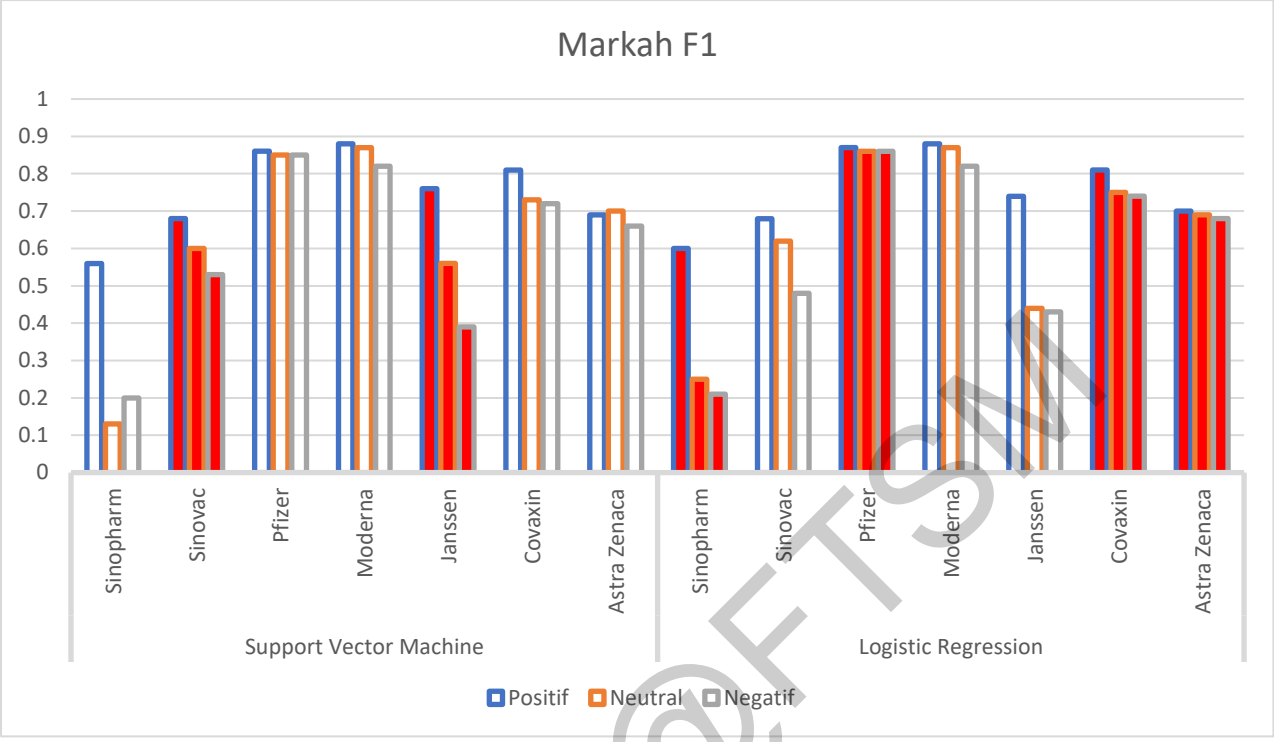
Rajah 5.8: Graf Perbandingan prestasi Dapatan kepada pengelasan sentimen algoritma pembelajaran mesin

Berdasarkan jadual 5.8 dan rajah 5.8, keputusan dapatan yang diperolehi dengan Logistic Regression adalah lebih tinggi berbanding dengan Support Vector Machine (SVM). Keputusan dapatan yang tinggi bermaksud tweets yang telah dilabel majoriti adalah betul. Kepada keputusan yang ditunjukkan di atas, keputusan dapatan Logistic Regression kepada semua set data adalah lebih tinggi daripada Support Vector Machine. Sebagai contoh, keputusan dapatan kepada set data Pfizer dengan menggunakan Logistic Regression mempunyai 88% betul kepada sentimen positif, 86% betul kepada sentimen neutral manakala 85% betul kepada sentimen negatif lebih tinggi daripada keputusan dapatan dengan menggunakan Support Vector Machine (SVM) iaitu 85% betul kepada sentimen positif, 88% betul kepada sentimen neutral dan 83% betul kepada sentimen negatif. Oleh itu, menurut kepada keputusan dapatan, Logistic Regression adalah lebih baik daripada Support Vector Machine (SVM).

SET DATA MENGIKUT JENIS VAKSIN (MARKAH F1)

Jadual 5.9: Perbandingan prestasi Markah F1 kepada pengelasan sentimen algoritma pembelajaran mesin

Markah F1	Support Vector Machine						
	Sinopharm	Sinovac	Pfizer	Moderna	Janssen	Covaxin	Astra Zeneca
Positif	0.56	0.68	0.86	0.88	0.76	0.81	0.69
Neutral	0.13	0.6	0.85	0.87	0.56	0.73	0.7
Negatif	0.2	0.53	0.85	0.82	0.39	0.72	0.66
	Logistic Regression						
	Sinopharm	Sinovac	Pfizer	Moderna	Janssen	Covaxin	Astra Zeneca
Positif	0.6	0.68	0.87	0.88	0.74	0.81	0.7
Neutral	0.25	0.62	0.86	0.87	0.44	0.75	0.69
Negatif	0.21	0.48	0.86	0.82	0.43	0.74	0.68



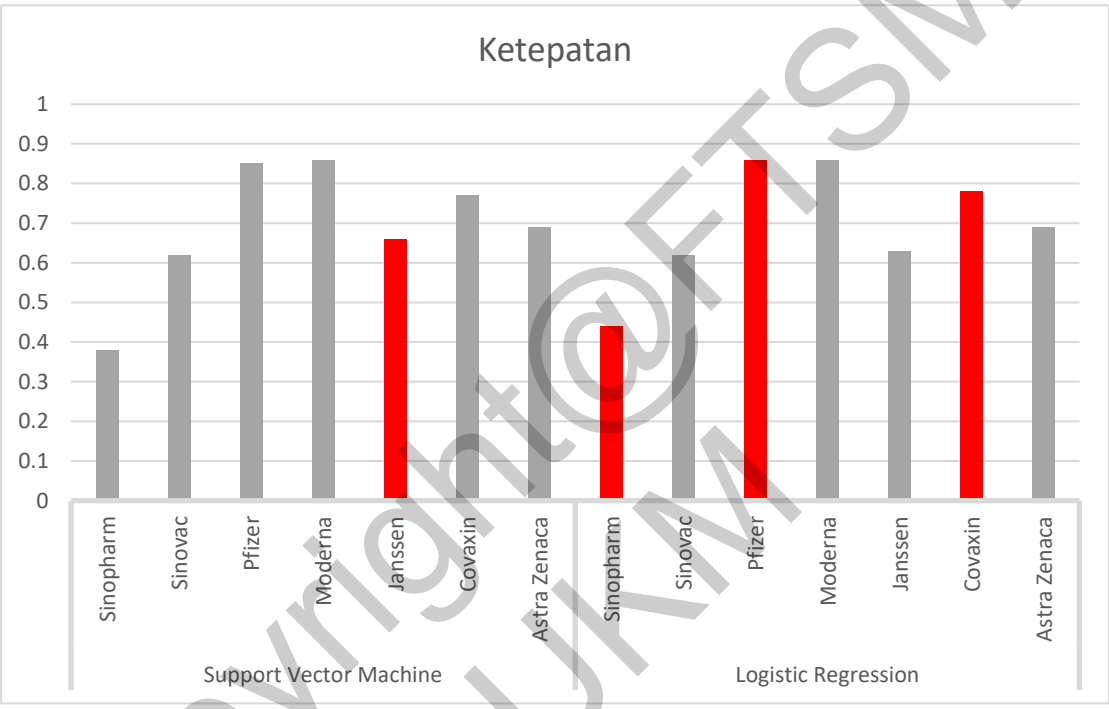
Rajah 5.9: Graf Perbandingan prestasi Markah F1 kepada pengelasan sentimen algoritma pembelajaran mesin

Markah F1 ialah purata wajaran kejituan dan dapatan dan markah ini juga dapat mengambil kira positif palsu dan negatif palsu. Berdasarkan jadual 5.9 dan rajah 5.9, keputusan markah F1 yang diperolehi dengan Logistic Regression adalah hampir sama berbanding daripada Support Vector Machine (SVM). Tetapi keputusan markah F1 yang diperolehi dengan Logistic Regression majoriti lebih tinggi berbanding dengan Support Vector Machine. Kepada set data Sinopharm, Pfizer, Covaxin dan Astra Zeneca markah F1 diperolehi dengan Logistic Regression adalah lebih tinggi berbanding dengan Support Vector Machine (SVM). Markah F1 juga dianggapkan sebagai min harmonik kepada Kejituan dan Dapatan. Oleh itu, markah F1 yang tinggi dapat dianggapkan sebagai algoritma tersebut sesuai digunakan dalam analisis pengelasan sentimen tersebut.

SET DATA MENGIKUT JENIS VAKSIN (KETEPATAN)

Jadual 5.10: Perbandingan prestasi Ketepatan kepada pengelasan sentimen algoritma pembelajaran mesin

	Support Vector Machine						
	Sinopharm	Sinovac	Pfizer	Moderna	Janssen	Covaxin	Astra Zeneca
Ketepatan	0.38	0.62	0.85	0.86	0.66	0.77	0.69
	Logistic Regression						
	Sinopharm	Sinovac	Pfizer	Moderna	Janssen	Covaxin	Astra Zeneca
Ketepatan	0.44	0.62	0.86	0.86	0.63	0.78	0.69



Rajah 5.10: Graf Perbandingan prestasi Ketepatan kepada pengelasan sentimen algoritma pembelajaran mesin

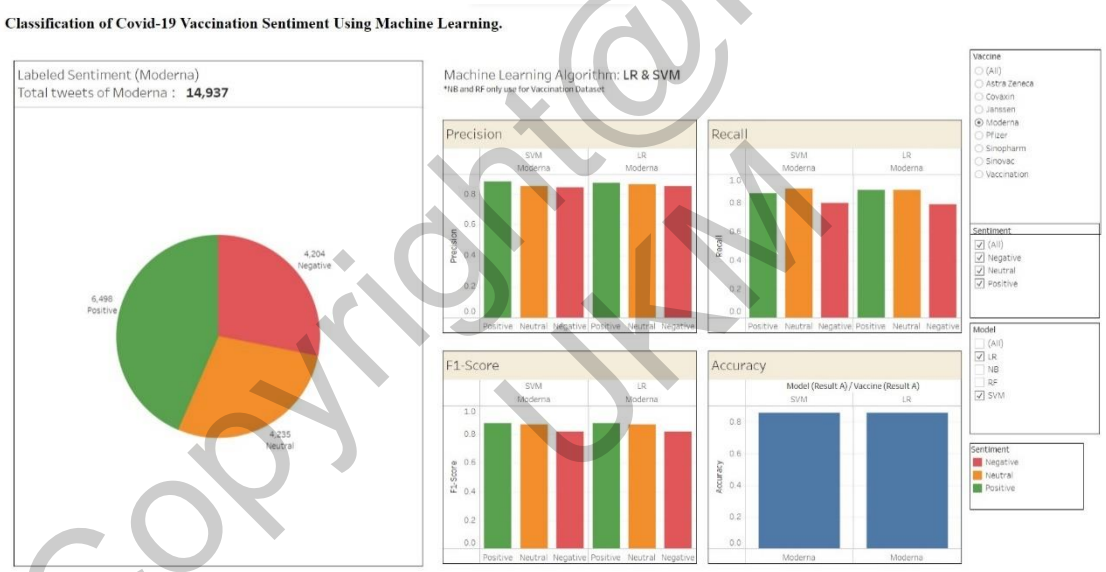
Berdasarkan jadual 5.10 dan rajah 5.10, *Logistic Regression* adalah lebih baik berbanding dengan *Support Vector Machine (SVM)*. Hal ini kerana keputusan ketepatan yang diperolehi dengan Logistic Regression majoriti lebih tinggi berbanding dengan Support Vector Machine seperti set data Sinopharm, Pfizer dan Covaxin. Keputusan kepada ketepatan dapat menganalisiskan sama ada prestasi algoritma pengelasan sentimen tersebut baik atau tidak. Oleh itu, berdasarkan keputusan tersebut. *Logistic Regression* adalah lebih baik berbanding dengan *Support Vector Machine (SVM).*

Berbandingan antara algoritma pengelasan sentimen *Logistic Regression* dan *Support Vector Machine (SVM)* dengan keputusan kejituan, dapatan, markah F1 dan ketepatan. Logistic Regression merupakan algoritma pengelasan sentimen yang paling baik berbanding dengan algoritma lain seperti *Support Vector Machine (SVM)*, *Naïve Bayes* dan *Random Forest* kerana prestasi kepada *Logistic Regression* adalah baik dan mempunyai keputusan

yang amat tinggi. Secara kesimpulannya, *Logistic Regression* adalah algoritma yang paling sesuai untuk digunakan dalam kajian pengelasan sentimen terhadap vaksinasi Covid-19.

5.9 PAPAN PEMUKA

Dalam kajian ini, Tableau telah digunakan untuk menghasilkan visual untuk papan pemuka. Tableau ialah aplikasi visualisasi dan analisis data interaktif yang membolehkan pengguna menjana papan pemuka dengan muda dan hanya dengan memasukkan data. Selepas memasukkan data, Tableau akan menyediakan beberapa pilihan graf visualisasi supaya pengguna dapat memilih graf yang paling sesuai. Papan pemuka dihasilkan supaya dapat memudahkan untuk memahami keputusan yang dihasil dalam kajian. Rajah di bawah menunjukkan papan pemuka kepada kajian ini.



Rajah 5.11: Papan Pemuka yang dihasilkan dengan menggunakan Tableau.

6 KESIMPULAN

Kajian yang bertajuk Pengelasan Sentimen terhadap Vaksinasi COVID-19 dengan menggunakan Pembelajaran Mesin telah berjaya dibangunkan dan sistem yang telah dibangunkan dapat memenuhi objektif kajian yang telah dirangcang dan ditunjukkan semasa fasa awal dahulu. Objektif kepada kajian ini adalah dapat membangunkan model untuk mengelaskan sentimen terhadap Vaksinasi COVID-19 yang terdapat di dalam media sosial dengan menggunakan algoritma pembelajaran mesin dan juga dapat membangunkan satu

sistem visualisasi pengelasan sentimen terhadap Vaksinasi COVID-19. Dengan ini, kedua-dua objektif kajian telah mencapai sepanjang kajian ini dilakukan.

Dalam kajian ini, ekstrak data adalah fasa yang penting dan data perlu diekstrak melalui media sosial Twitter dengan menggunakan Twitter API. Selepas mendapatkan tweets di Twitter, proses pra-pemprosesan dijalankan untuk membersihkan data yang telah diekstrak seperti data yang tiada berkaitan dengan kajian, data yang salah atau rosak, format data yang salah dan tidak lengkap supaya data ini tidak dapat mengaruhkan ketepatan kepada keputusan seterusnya. Di samping itu, penggunaan kepada pakej perpustakaan NLTK, iaitu *VADER* dan *Textblob* supaya dapat melabelkan sentimen kepada semua tweets. Selepas tweets dapat dilabelkan, tweets perlu dimasukkan ke dalam pembelajaran mesin untuk melatihkan pembelajaran mesin dan menentukan prestasi kepada 4 algoritma pembelajaran mesin, iaitu *Support Vector Machine (SVM), Logistic Regression, Naïve Bayes* dan *Random Forest.*

Mengikut hasil keputusan yang diperoleh dari kajian ini, Logistic Regression dapat dipilih sebagai pembelajaran mesin yang terbaik kerana ia menghasilkan ketepatan yang paling tinggi berbanding dengan algoritma lain seperti *Support Vector Machine (SVM), Naïve Bayes* dan *Random Forest* dan keseluruhan keputusan analisis sentimen telah dipaparkan dalam papan pemuka dalam bentuk visualisasi dengan menggunakan Tableau.

7 RUJUKAN

Akash Dutt Dubey, 9 Apr 2020, Twitter Sentimen Analysis during COVID-19 Outbreak, SSRN, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3572023

Bryan White, 27 May 2020, Sentimen Analysis: VADER or Textblob, Towards Data Science, https://towardsdatascience.com/sentimen-analysis-vader-or-textblob-ff25514ac540#:~:text=Both%20libraries%20offer%20a%20host,with%20more%20formal%20language%20usage.

David J Cennimo, 25 Jun 2021, What is COVID-19. Medscape https://www.medscape.com/answers/2500114-197401/what-is-covid-19

Difference Between CSV and XLS, toggle track, https://toggl.com/track/difference-between-csv-xls/

Huseyin Kucukali, 29 May 2021, Vaccine hesitancy and anti-vaccination attitudes during the start of COVID-19 vaccination program: A content analysis on Twitter data, Istanbul Medipol University,

https://www.medrxiv.org/content/10.1101/2021.05.28.21257774v1.full.pdf

Khalid K. Al-jabery, Donald C. Wunsch II, 2020, Data preprocessing, ScienceDirect,

https://www.sciencedirect.com/topics/engineering/data-preprocessing

Keith M. Bower, What is Design of Experiments (DOE), ASQ,

https://asq.org/quality-resources/design-of-experiments

Mampu, Jabatan Perdana Menteri, 25 May 2020, Spesifikasi Keperluan Sistem,

MySQA, https://sqa.mampu.gov.my/index.php/ms/3-10-penyediaan-spesifikasi-keperluan-sistem-f2-6

Manish Shama, 11 May 2022, Sentimen Analysis (An Introduction to Naïve Bayes

Algorithm), Toward Data Science, https://towardsdatascience.com/sentimen-analysis-introduction-to-naive-bayes-algorithm-96831d77ac91

Mohit Sharma, 27 Oct 2020, Data Preprocessing: 6 Necessary Steps for Data

Scientists, Hackermoon, https://hackernoon.com/what-steps-should-one-take-while-doing-data-preprocessing-502c993e1caa

Nagesh Singh Chauhan, 8 Apr 2022, Naïve Bayes Algorithm: Everything You Need

To Know, KDnuggets, https://www.kdnuggets.com/2020/06/naive-bayes-algorithm-everything.html

Nai-Chen Cheng, Feb 2020, Application of Support Vector Machine (SVM) in the

Sentimen Analysis of Twitter Dataset, ResearchGate, https://www.researchgate.net/publication/339112527_Application_of_Support_Vector_Machine_SVM_in_the_Sentimen_Analysis_of_Twitter_DataSet

Nick Hotz, 16 Apr 2022, What is CRISP DM, Data Science Process Alliance

https://www.datascience-pm.com/crisp-dm-2/

Özgür Genç, 16 Apr 2019, The basics of NLP and real time sentimens analysis with

open source tools, Toward Data Science, https://towardsdatascience.com/real-time-sentimen-analysis-on-social-media-with-open-source-tools-f864ca239afe

Perbezaan Antara Keperluan Fungsional dan Tidak Berfungsi, Strephonsays,

https://ms.strephonsays.com/functional-and-non-functional-requirements-3325

Pranjal Pandey, 25 Nov 2019, Data Preprocessing: Concepts, Toward Data Science,

https://towardsdatascience.com/data-preprocessing-concepts-fa946d11c825

Quyen G. To, 12 Apr 2021, Applying Machine Learning to Identify Anti-Vaccination

Tweets during the COVID-19 Pandemic, MDPI, https://www.mdpi.com/1660-4601/18/8/4069

Rohith Gandhi, 8 Jun 2018, Support Vector Machine - Introduction to Machine

Learning Algorithms, Towards Data Science, https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Saishruthi Swaminathan, 15 Mar 2018, Logistic Regression – Detailed Overview,

Towards Data Science, https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

Samira Yousefinaghani, 11 July 2021, An analysis of COVID-19 vaccine sentimens

and opinions on Twitter, ScienceDirect, https://www.sciencedirect.com/science/article/pii/S1201971221004628

Stephen Wai Hang Kwok, Sai Kumar Vadde, Guangjin Wang, 19 May 2021,

Tweet Topics and Sentimens Relating to COVID-19 Vaccination Among Australian Twitter Users: Machine Learning Anaysis, Journal of Medical Internet Research, https://www.jmir.org/2021/5/e26953

Tony You, 12 Jun 2019, Understanding Random Forest, Towards Data Science,

https://towardsdatascience.com/understanding-random-forest-58381e0602d2

Twitter, Twitter API, Twitter Developer Platform,

https://developer.twitter.com/en/docs/twitter-api

World Health Organization, 22 May 2022, COVID-19 vaccines, WHO 2022

https://www.who.int/emergencies/diseases/novel-coronavirus-2019/covid-19-vaccines

Lau Yong Jie (A176689)
Azuraliza Abu Bakar
Fakulti Teknologi & Sains Maklumat,
Universiti Kebangsaan Malaysia