ANALISIS SENTIMEN MASA NYATA TWEET BAHASA MELAYU TERHADAP VAKSIN

WONG WAI JIAN

SABRINA BINTI TIUN

Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Pada masa kini, dengan peningkatan teknologi rangkaian yang pesat, banyak orang dapat dengan mudah menyatakan dan meninggalkan komen, pendapat dan juga perasaan mereka mengenai beberapa topik hangat tertentu secara terbuka di media sosial seperti Twitter, dan Facebook. Disebabkan wujudkan wabak COVID-19 yang berlaku di seluruh dunia yang juga termasuk di Malaysia dari tahun lalu hingga sekarang, hal ini telah menwujudkan keadaan di mana semakin banyak tinjauan atau pendapat tentang vaksin (yang dapat membantu mencegah jangkitan dari COVID-19) dibincangkan oleh warganegara Malaysia melalui Twitter. Analisis sentimen merupakan proses mengumpulkan data subjektif dan mengkategorikannya mengikut kekutuban yang positif, negatif, atau neutral. Dalam projek ini, objektifnya adalah untuk mengumpulkan semua *Tweet* bahasa Melayu mengenai vaksin sebagai data dari Twitter dan membina model untuk mengklasifikasikan data (yang telah dilakukan proses pembersihan data dan pra-pemprosesan) kepada kategori positif, negatif, atau neutral dalam masa nyata. Selain itu, teknik pembelajaran mesin juga akan digunakan untuk melatih dan menguji model analisis sentimen berdasarkan prestasi ketetapan. Kemudian, satu atau beberapa graf akan dipaparkan untuk keseluruhan sentimen *Tweet* bahasa Melayu tentang vaksin melalui laman web dengan reka bentuk yang mudah. Akhirnya, projek ini telah berjaya membangunkan pelaksanaan dan pembinaan model pengelas bagi projek yang bertajuk Analisis Sentimen *Tweet* Melayu Masa Nyata mengenai vaksin dan juga dapat menunjukkan hasilannya di laman web. Tambahan pula, projek ini telah menepati objektif-objektif kajian projek, iaitu menyediakan set data analisis sentimen *Tweet* bahasa Melayu untuk domain vaksin dan juga membina model analisis sentimen untuk teks *Tweet* dalam masa nyata menggunakan kaedah pembelajaran mesin dengan pengelas yang paling sesuai.

1 PENGENALAN

Projek yang bertajuk Analisis Sentimen Twitter Melayu Masa Nyata mengenai vaksin adalah menggunakan kaedah analisis sentimen untuk menganalisis semua kandungan *Tweet* dalam bahasa Melayu yang berkaitan dengan topik tentang vaksin

COVID-19 secara masa nyata, dan mengkategorikan *Tweet* tersebut sama ada dalam kumpulan Sentimen positif, negatif atau neutral.

Media sosial ialah teknologi komunikasi untuk orang ramai mencipta dan berkongsi pendapat, ideal dan maklumat melalui aplikasi dan rangkaian. Pada masa kini, terdapat pelbagai jenis platform media sosial yang dicipta seperti Facebook, Twitter, Instagram, dan LinkedIn. Oleh itu, dengan peningkatan pesat teknologi rangkaian hari ini di seluruh dunia, lebih ramai orang dapat menyiarkan cerita, komen dan pendapat kehidupan sebenar mereka terhadap beberapa topik tertentu melalui media sosial dengan mudah. Oleh itu, semakin banyak jenis perasaan manusia, emosi juga diluahkan sepanjang menyiarkan pendapat dan meninggalkan komen di media sosial. Pada masa yang sama, beberapa laman media sosial sanggup mendedahkan antara *application programming interfaces* mereka, juga boleh dipanggil sebagai API yang membolehkan penyelidik dan pelajar mengumpul dan menganalisis data seperti ulasan dan ulasan produk di Twitter. Twitter ialah salah satu laman web rangkaian sosial dan aplikasi media sosial yang popular di mana pengguna boleh berkomunikasi dengan pesanan ringkas yang dikenali sebagai *Tweet* dengan pengguna lain. Pengguna juga boleh menyiarkan dan berkongsi mesej mereka secara terbuka kepada sesiapa sahaja yang telah mengikuti pengguna tersebut di Twitter. Oleh itu, pengguna Twitter boleh menyiarkan dan memberikan pendapat atau perbincangan peribadi mereka mengenai topik hangat khusus semasa dengan mudah seperti topik COVID-19, vaksin dan lain-lain. Selain itu, Twitter mempunyai had aksara pada *Tweet* yang mana setiap pengguna hanya boleh menulis 280 aksara setiap *Tweet*. Ini adalah untuk mengelakkan sekumpulan pengguna daripada menjejalkan fikiran mereka kepada pengguna lain.

Sentimen ialah perasaan atau pendapat, terutamanya berdasarkan emosi (*Oxford Learner Dictionary*). Analisis sentimen, juga dikenali sebagai perlombongan pendapat, ialah cara menganalisis data subjektif seperti pendapat manusia, dan komen di media sosial untuk mengenal pasti emosi manusia dan mengklasifikasikan data ini mengikut polariti iaitu Sentimen positif, negatif atau neutral. Dalam beberapa bulan kebelakangan ini, wabak COVID-19 yang merupakan wabak penyakit berjangkit terbesar di dunia berlaku di seluruh dunia termasuk Malaysia. Disebabkan situasi sebegini, semakin ramai rakyat Malaysia mula berbincang dan menyiarkan pendapat mereka sendiri mengenai jenis vaksin yang digunakan untuk membantu mencegah jangkitan COVID-19 dan kesannya kepada tubuh manusia selepas mendapat suntikan vaksin melalui Twitter.

Oleh itu, tujuan projek ini adalah untuk mencadangkan kaedah bagaimana menggunakan API Twitter untuk mengumpul sekumpulan *Tweet* dalam bahasa Melayu dari Malaysia dalam masa nyata sebagai set data untuk model analisis sentimen, dan cara mengklasifikasikan set data ke dalam skor sentimen selepas proses pra-pemprosesan dan pembersihan data. Dalam erti kata lain, projek ini mungkin membantu untuk mengetahui secara keseluruhan bagaimana perasaan, emosi dan sentimen rakyat Malaysia terhadap vaksin COVID-19 melalui Twitter.

2 PENYATAAN MASALAH

Dalam kajian analisis sentimen ini, data akan dikumpul daripada salah satu platform media sosial iaitu Twitter bagi membantu membina model bagi mengkategorikan *Tweet* pengguna dalam masa nyata mengikut polariti iaitu positif, negatif atau neutral. Salah satu masalah yang dihadapi dalam kajian ini ialah bagaimana untuk mengetahui sentimen pengguna Twitter terhadap topik vaksin daripada kandungan *Tweet* pengguna.

Dengan had aksara dalam menulis *Tweet*, pengguna Twitter akan cuba memendekkan panjang teks mereka dengan menggunakan cara lain seperti emotikon dan hashtag. Dan masalah seterusnya ialah bagaimana menangani atribut tambahan selain perkataan daripada kandungan *Tweet* seperti emotikon, emoji dan hashtag yang kerap digunakan oleh pengguna dalam media sosial. Perkataan dalam bentuk pendek juga merupakan salah satu masalah dalam kajian analisis sentimen. Bentuk pendek agak kerap digunakan pada *Tweet* kerana had aksara.

Akhirnya, masalah yang juga dihadapi dalam kajian ini ialah masalah mencari model pengelas yang paling sesuai untuk model analisis sentimen tersebut. Hal ini disebabkan dalam pembelajaran mesin, ia terdapat pelbagai pengelas yang boleh digunakan untuk pelbagai jenis tujuan. Sebagai contoh, dalam kajian sentimen analisis, pengelas-pengelas yang biasanya digunakan ialah *Logistic Regression*, *Naïve Bayes*, *Support Vector Machine*, *K Nearest Neighbor* dan sebagainya.

3 OBJEKTIF KAJIAN

Projek ini bertujuan menyediakan set data analisis sentimen *Tweet* bahasa Melayu untuk domain vaksin dan membina model analisis sentimen untuk teks *Tweet* dalam

masa nyata menggunakan kaedah pembelajaran mesin dengan pengelas yang paling sesuai.

4 METOD KAJIAN

Penggunaan metodologi menunjukkan langkah-langkah kritikal yang mesti diambil untuk menjamin projek disiapkan dengan berjaya dan mengikut peringkat yang dirancang. Untuk memastikan projek berjalan lancar dan pengeluaran produk kerja berkualiti tinggi, ia adalah penting untuk menggunakan model pembangunan yang betul. Rajah 1 tersebut merupakan kitaran hayat pembelajaran mesin, ia menunjukkan aktiviti yang akan terbabit dalam projek ini, terdapat tujuh langkah utama yang perlu dijalankan dalam kajian analisis sentimen ini.

Rajah 1: Langkah proses kajian analisis sentimen Twitter mengenai vaksin dari Javatpoint, Machine learning Life cycle.

4.1 Fasa Pengumpulan Data

Pengumpulan data merupakan langkah yang pertama dalam projek ini, *'Tweet'* dalam bahasa Melayu akan dikumpulkan melalui API Twitter sebagai set data untuk model analisis sentimen. Pada peringkat ini, pengumpulan data akan dilakukan mengikut objektif dan keperluan seperti data mestilah dalam bahasa Melayu, berkaitan dengan

topik vaksin COVID-19, dan ditulis oleh pengguna Twitter dari Malaysia. Tarikh bagi set data berikut adalah dari 13 April 2022 sampai 25 May 2022.

```
tweet_text = pd.DataFrame(data=data,
                          columns=["id", "user", "time", "text"])
tweet_text
```

	id	user	time	text
0	359329138	dindamedina	2022-04-21 13:53:05	Tadi pagi baru vaksin yang ketiga dan nggak ad...
1	446174007	Indra_zill	2022-04-21 13:53:05	@popienastiti1 Gaji udah dalam bentuk makanan ...
2	1451457302628765703	nb1957_	2022-04-21 13:53:05	RT @deehandhanin: ruu tpks disahkan. vaksin ka...
3	1021087094	NugroDwi	2022-04-21 13:52:58	RT @GratisTerbaik: "Jika masih ada pemberian v...
4	300233479	pulchritudenj	2022-04-21 13:52:44	@tubirfess I KNOW RIGHT?!!! Greget banget baca...
...	...	...	...	...
9995	1085708885044879365	AannasNoah	2022-04-21 02:18:41	RT @KKMPutrajaya: Hari Raya bakal menjelang. R...
9996	1131901132928737280	pinkgurlsonmode	2022-04-21 02:18:41	orang2 anti vaksin kaya gituuu!!! https://t.co...
9997	3249537901	giraffehannie	2022-04-21 02:18:36	RT @tubirfess: 2beer! Tak kira semua org tau p...
9998	1396331301959921664	GinanjarTrisno	2022-04-21 02:18:36	Vaksin Booster dan Prokes Ketat Strategi Ampuh...
9999	1266986419974619137	BiruBrigade	2022-04-21 02:18:33	Mudik Dengan Vaksin Booster, Mudik Dengan Seha...

10000 rows × 4 columns

Rajah 2: Contoh Set Data Yang Dikumpulkan

4.2 Fasa Penyediaan Data

Fasa penyediaan data perlu dilakukan supaya data menjadi lebih sesuai untuk digunakan dalam latihan dan ujian pembelajaran mesin. Langkah pra-pemprosesan data akan dijalankan dengan menggunakan perpustakaan Natural Language Toolkit (NLTK), Pandas dan 'regular expression' (re). Hal ini disebabkan *Tweet* yang tanpa menjalankan fasa pra-pemprosesan data adalah sangat tidak berstruktur dan mengandungi maklumat yang berlebihan seperti simbol-simbol, emotikon, sifat-sifat pengguna dan juga pautan laman web. Data tersebut boleh mempengaruhi ketepatan dan prestasi klasifikasi bagi model analisis sentimen. Oleh itu, data yang dikumpul akan ditapis untuk menghapuskan atribut yang tidak berguna.

4.3 Fasa Pembersihan Data

Dalam peringkat ini, tokenisasi digunakan untuk memotong perkataan daripada teks dan perenggan dalam data yang dikumpul mengikut ruang. Normalisasi ejaan juga akan digunakan untuk membetulkan perkataan dalam bentuk pendek atau perkataan yang salah ejaan kembali kepada perkataan asal. Dalam pengekodan bagi proses pembetulan ejaan, ia mempunyai sekitar 1,500 perkataan untuk mengesankan sama ada data mempunyai sebarang perkataan yang salah berdasarkan kajian Ariffin dan Tiun (2020). Disebabkan dalam bahasa Melayu juga terdapat kewujudan pelbagai perkataan henti

semasa menulis ayat, justeru perkataan henti seperti “mereka”, “yang” dan “dia” akan dikeluarkan daripada teks data.

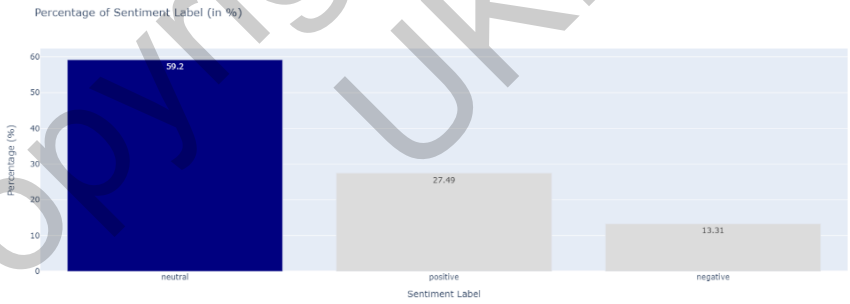
4.4 Fasa Analisis Data

Dalam fasa ini, selepas menjalankan fasa pengumpulan data dan fasa pra-pemprosesan data, pelabelan data secara manual bagi semua set data adalah dijalankan. Model analisis sentimen bagi bahasa Melayu juga dibina untuk menganalisis dan mengklasifikasikan setiap ayat daripada data pada polariti iaitu Sentimen positif, negatif atau neutral mengikut skor sentimen.

	text	label
0	tidak satupun menyesal menolak vaksin sebalik menyesal divaksin	negative
1	vaksin sinovac mengalami pitak rontok hebat menerus	positive
2	indonesia sebentar terdampak wabah cacar api monkeypox baca efek samping vaksin	neutral
3	tenaga kesehatan nakes rampung menerima vaksin	neutral
4	habis vaksin lengan sakit kasi beban	neutral
5	alhamdulillah sertifikat vaksin indonesia diakui negara asean indonesia maju	positive
6	vaksin dianterin tegar vaksin dianterin nabila vaksin tegar muter mulu	neutral
7	pusat veteriner farma pusvetma surabaya memproduksi vaksin mencegah penularan penyakit mulut	negative
8	habis vaksin moderna tangan kek pengen jotosi biar kemeng	neutral
9	untung vaksin kebal	positive
10	vaksin berplatform rna vaksin kuala lumpur kandunga sedikitpun tidak isi adl	neutral
11	menunggu distribusi vaksin kementerian pertanian kementan vaksin kabupaten kota	neutral

Rajah 3: Data yang Selepas Dilabelkan Secara Manual

Seterusnya, bilangan jumlah bagi setiap label dalam set data tersebut adalah didapati sangat ketidakseimbangan seperti Rajah 4.



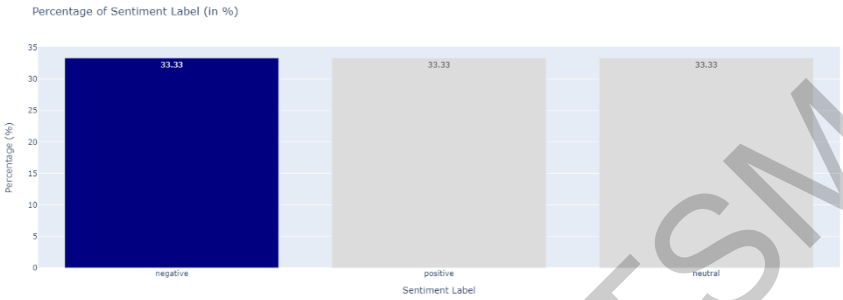
Rajah 4: Graf Bar bagi Pengiraan Jumlah Setiap Label

Oleh itu, teknik pensampelan semula (*resampling*) digunakan dengan berdasarkan Jason Brownlee (2018). Dalam penyelidikan analisis sentimen ini, teknik persampelan terkurang telah dipilih untuk digunakan pada set data. Dengan ini, jumlah label yang terendah digunakan untuk menentukan jumlah data yang perlu dipilih secara rawak daripada label majoriti. Dalam set data tersebut, label negatif mempunyai jumlah data terendah iaitu 3537, oleh itu, jumlah 3537 data akan dipilih secara rawak daripada label neutral dan positif dengan menggunakan *df.sample()*.

```
tweet_text['label'].value_counts()

negative    3537
positive    3537
neutral     3537
Name: label, dtype: int64
```

Rajah 5: Pengiraan Jumlah Setiap Label Selepas Persampelan Terkurang



Rajah 6: Graf Bar bagi Pengiraan Jumlah Setiap Label Selepas Persampelan Terkurang

4.5 Fasa Model Latihan dan Ujian

Set data dikategori kepada set data latihan dan set data ujian dengan menggunakan perpustakaan *sklearn.model_selection.train_test_split()* untuk latihan model dan ujian pada model. Dalam proses tersebut, set data latihan terdapat 75 peratus data dan set data ujian terdapat 25 peratus data daripada set data. Teknik TF-IDF juga digunakan selepas data telah dikategori sebagai set data latihan dan set data ujian supaya mengira kekerapan perkataan untuk menentukan sejauh mana perkataan tersebut berkaitan dengan dokumen tertentu.

Dalam penyelidikan analisis sentimen ini, terdapat 4 pengelas pembelajaran mesin digunakan untuk model analisis sentimen iaitu, *Support Vector Machine*, *Logistics Regression*, *Naive Bayes* dan *K-Nearest Neighbors*. Selepas melakukan latihan dan ujian model, model terlatih yang mempunyai prestasi yang terbaik akan digunakan untuk melakukan analisis sentimen *Tweet* dalam masa nyata. Prestasi bagi setiap klasifikasi adalah berdasarkan keputusan ketepatan, kejituan dan skor F1 dari proses pengujian model.

4.6 Fasa Penyebaran Model

Model pengelas yang terbaik akan disimpankan untuk menjalankan analisis sentimen *Tweet* dalam masa nyata dengan dimuatkan supaya dapat digunakan untuk meramalkan sentimen bagi data teks yang terkini.

4.6 Fasa Keputusan

Akhirnya, hasilan daripada analisis sentimen *Tweet* dalam masa nyata yang diramalkan oleh model pengelas yang terbaik akan ditunjukkan dalam graf melalui antara muka laman web yang ringkas.

5 HASIL KAJIAN

5.1 Hasil Penilaian Model Pengelas

Dalam fasa ini, hasilan dari *sklearn.metrics.classification_report()* bagi setiap model pengelas adalah merupakan Rajah 7, Rajah 8, Rajah 9, dan Rajah 10 di bawah. Dalam rajah tersebut, *precision* merujuk kepada ketepatan, *recall* merujuk kepada kejituan, *f1-score* merujuk kepada Skor F1, dan *support* merujuk sokongan.

```
LR Accuracy with TFIDF: 0.7689408217112702
Percentage of LR Accuracy with TFIDF: 76.89408217112702 %

              precision    recall  f1-score   support

    negative       0.80      0.76      0.78       862
     neutral       0.70      0.83      0.76       902
    positive       0.83      0.71      0.77       889

    accuracy                           0.77      2653
   macro avg       0.78      0.77      0.77      2653
weighted avg       0.78      0.77      0.77      2653
```

Rajah 7: Penilaian Metrik Model Logistics Regression

```
NB Accuracy with TFIDF: 0.4519411986430456
Percentage of NB Accuracy with TFIDF: 45.19411986430456 %

              precision    recall  f1-score   support

    negative       0.55      0.38      0.45       862
     neutral       0.49      0.25      0.33       902
    positive       0.40      0.73      0.52       889

    accuracy                           0.45      2653
   macro avg       0.48      0.45      0.43      2653
weighted avg       0.48      0.45      0.43      2653
```

Rajah 8: Penilaian Metrik Model Naïve Bayes

```
SVM Accuracy with TFIDF: 0.7723332076894082
Percentage of SVM Accuracy with TFIDF: 77.23332076894081 %

              precision    recall  f1-score   support

    negative       0.81      0.75      0.78       862
     neutral       0.69      0.87      0.77       902
    positive       0.86      0.69      0.77       889

    accuracy                           0.77      2653
   macro avg       0.79      0.77      0.77      2653
weighted avg       0.79      0.77      0.77      2653
```

Rajah 9: Penilaian Metrik Model Support Vector Machine

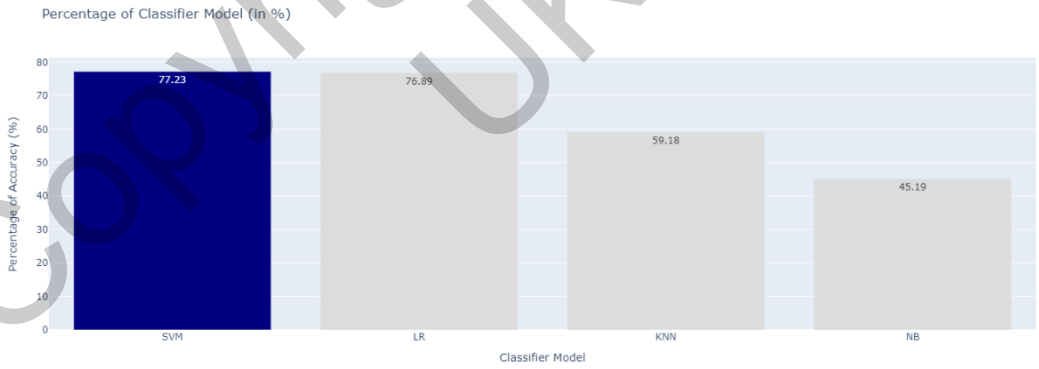
```
KNN Accuracy with TFIDF: 0.5917828872973991
Percentage of KNN Accuracy with TFIDF: 59.17828872973991 %

              precision    recall  f1-score   support

    negative       0.59      0.66      0.63       862
     neutral       0.53      0.58      0.55       902
    positive       0.68      0.54      0.60       889

    accuracy                           0.59      2653
   macro avg       0.60      0.59      0.59      2653
weighted avg       0.60      0.59      0.59      2653
```

Rajah 10: Penilaian Metrik Model K Nearest Neighbor



Rajah 11: Graf Bar Ketepatan bagi Setiap Model Pengelas

Jadual 1: Hasilan Penilaian Bagi Setiap Model Pengelas

Penilaian Metrik	**Model Pengelas**			
	Logistic Regression (%)	Naïve Bayes (%)	Support Vector Machine (%)	K Nearest Neighbor (%)
Purata Makro (*Macro Average*)				
Ketetapan	78	48	79	60
Kejituan	77	45	77	59
Skor F1	77	43	77	59
Purata Wajaran (*Weighted Average*)				
Ketetapan	78	48	79	60
Kejituan	77	45	77	59
Skor F1	77	43	77	59

Jadual 1 merupakan markah penilaian metrik bagi setiap model pengelas. Purata makro adalah dikira dengan mengambil purata bagi setiap markah metrik. Manakala, purata wajaran dikira dengan mengambil purata markah yang mempertimbangkan sokongan setiap kelas. Berdasarkan Jadual 1, ia menunjukkan bahawa model *Support Vector Machine* (SVM) mendapat prestasi yang terbaik berbanding model pengelas lain. Model *Logistic Regression* (LR) juga mempunyai prestasi yang baik dan ketetapannya hampir sama dengan model SVM. Dengan ini, ia menunjukkan model SVM dan LR adalah lebih sesuai digunakan untuk ramalan sentimen bagi data teks bahasa Melayu berbanding model *Naïve Bayes* (NB) dan *K Nearest Neighbor* (KNN). Walaupun keputusan bagi penilaian metrik model SVM dan LR adalah hampir sama, tetapi, bagi peratusan ketetapan, model SVM adalah lebih tinggi 1% daripada model LR. Oleh itu, model SVM dipilih untuk digunakan dalam fasa seterusnya, iaitu fasa pengujian model pengelas dengan data yang masa nyata.

Berdasarkan Jadual 1, ia menunjukkan bahawa prestasi model SVM dan LR adalah lebih tinggi berbanding pengelas yang lain. Hal ini mungkin disebabkan model SVM dan LR boleh menyokong penyelesaian linear (*linear solution*), tetapi model NB dan KNN tidak boleh. Oleh itu, set data analisis sentimen yang digunakan dalam projek lebih sesuai untuk diselesaikan dengan penyelesaian linear. Selain itu, model SVM adalah lebih berkuasa untuk membuat generalisasi dengan baik dalam ruang dimensi tinggi seperti yang sepadan dengan teks dengan menghapuskan keperluan untuk

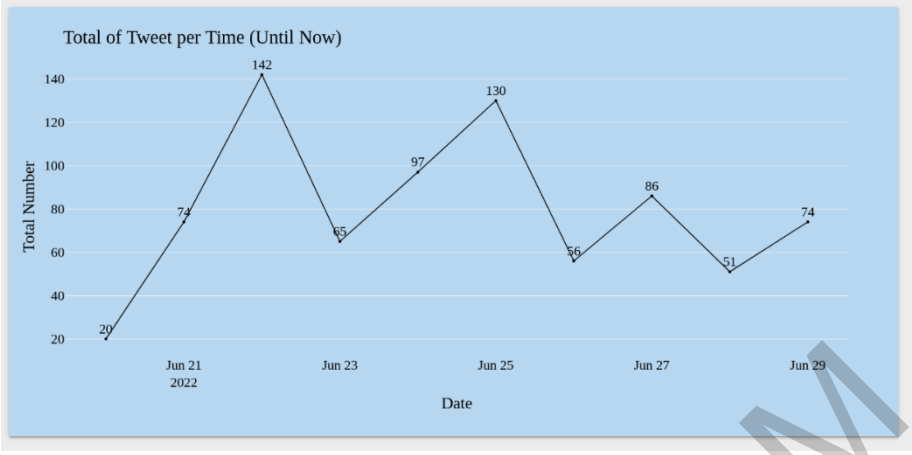
pemilihan ciri, menjadikan aplikasi pengkategorian teks lebih mudah. Bagi model NB, ia mungkin disebabkan ciri berlebihan teks dalam teks, anggaran parameter kasar dan NB menjangkakan semua ciri adalah bebas, oleh itu, hal ini mungkin menyebabkan ia mempunyai prestasi rendah dalam set data ini. Seterusnya, sebab model KNN mempunyai prestasi rendah adalah mungkin disebabkan model KNN tidak mempunyai proses latihan, dan akibatnya, ia tidak cuba untuk mengoptimumkan sebarang ukuran keberkesanan.

5.2 HASILAN PENGUJIAN MODEL PENGELAS DENGAN DATA YANG MASA NYATA

Fasa ini adalah menunjukkan hasilan pengujian model pengelas dengan data yang masa nyata di laman web melalui platform Anvil. Maklumat-maklumat data seperti bilangan data masa nyata yang dikumpul dan juga graf-graf adalah dibuat dan dibina dalam *Google Colaboratory* dengan Python dan menggunakan platform Anvil untuk menghantar maklumat tersebut kepada laman web.

Rajah 12: Hasilan 1 Pengujian Model Pengelas Dengan Data Masa Nyata

Dalam Rajah 12, laman web menunjukkan tajuk project ini, iaitu *Real Time Sentiment Analysis on Vaccine*. Tarikh-tarikh data yang telah dikumpul juga ditunjukkan bawah perkataan "*Overview*" supaya mengetahui tarikh-tarikh pengumpulan data terkini. Seterusnya, ia menunjukkan bilangan jumlah data yang dikumpul dan bilangan bagi setiap label. Sebagai contohnya, berdasarkan rajah tersebut, tarikh dari 20 Jun 2022 sehingga 29 Jun 2022 terdapat 795 *Tweet* berkenaan vaksin yang dipaparkan di Malaysia. Dalam 795 *Tweet* tersebut, terdapat 321 *Tweet* adalah teks yang positif, 397 *Tweet* yang neutral, dan 77 *Tweet* adalah yang negatif. Perlabelan bagi *Tweet* tersebut adalah diramalkan oleh model pengelas yang mempunyai prestasi yang terbaik dari 5.1 HASILAN PENILAIAN MODEL PENGELAS, iaitu model pengelas *Support Vector Machine* yang dibina dan dilatih sebelum ini.



Rajah 13: Hasilan 2 Pengujian Model Pengelas Dengan Data Masa Nyata

Rajah 13 adalah menunjukkan jumlah *Tweet* berkenaan vaksin dalam bahasa Melayu yang dipaparkan bagi setiap hari dari 20 Jun 2022 sehingga 29 Jun 2022. Berdasarkan graf tersebut, ia menunjukkan 25 Jun 2022 mempunyai 130 *Tweet* yang dipaparkan dan juga jumlah yang tertinggi berbanding dengan hari lain.

Rajah 14: Hasilan 3 Pengujian Model Pengelas Dengan Data Masa Nyata

Seterusnya, dalam laman web ini, Rajah 14 adalah menunjukkan peratusan bagi setiap label, iaitu positif, negatif dan neutral melalui graf supaya mudah membuat perbandingan bagi bilangan setiap label dan mengetahui teks dalam *Tweet* berkenaan vaksin dalam bahasa Melayu adalah lebih kepada positif, negatif atau neutral dalam 7 hari ini. Oleh itu, berdasarkan graf dalam Rajah 14 telah menunjukkan *Tweet* yang berlabel positif mempunyai 40.38%, neutral mempunyai 49.94% dan akhirnya, negatif mempunyai peratusan yang terkurang iaitu, 9.69%.

Raw Tweet Data

Index	Time	Text	Location
0	2022-06-27 12:38:29	pelalian vaksin influenza..ada plhiv yg ambik x..	malaysia
1	2022-06-27 11:52:25	tgh syiok tido tiber mama bwk pegi vaksin...🤒 #ginger #mycat #mybelovedcat #catlover @ cat buddy veterinary clinic https://t.co/avqytpsvjy	kuala lumpur, malaysia
2	2022-06-27 11:41:50	@hirumi13 @abckicapp @officialjohor bukan tak bayar gaji . daniel thing tak vaksin . so kena jual lah dri kl city	ggmu 🌈
3	2022-06-27 11:03:42	maaf ya pak..tidak akan terungkap klo itu dampak dr vaksin..mohon jgn memaksa rakyat untuk vaksin kli kalian saja t... https://t.co/awfjm3brii	hulu langat malaysia

Rajah 15: Hasilan 4 Pengujian Model Pengelas Dengan Data Masa Nyata

Rajah 15 tersebut adalah menunjukkan jadual bagi teks *Tweet* dan maklumatnya seperti masa dan lokasi. Jadual ini merupakan data yang dikumpulkan sebelum pra-pemprosesan dan ramalan sentimen.

Tweet Data After Data Pre-Processing & Labeling

Index	Text	Label
70	logik antivax rukun beriman qada qadar habis cucuk vaksin	neutral
71	patut tukar nama pim pum perjuangkan isu vaksin konspirasi isu coklat cikedis	negative
72	djokovic tidak mengulangi pendirian vaksinasi terbuka hilang novak djokovic	negative
73	bawa anak pergi vaksin pneumococcal tidur rumah haih tangguh	negative
74	fokus penyampaian maklumat kepentingan penggalak vaksin tris penilaian kendiri ok	neutral

« ‹ › »

Rajah 16: Hasilan 5 Pengujian Model Pengelas Dengan Data Masa Nyata

Rajah 16 merupakan jadual yang menunjukkan keputusan bagi *Tweet* yang selepas pra-pemprosesan dan juga ramalan sentimen oleh model pengelas. Dengan Rajah 15 dan 16, ia boleh membuat perbandingan bagi sebelum dan selepas pra-pemprosesan dan ramalan sentimen. Jadual bagi Rajah 15 dan 16 juga dibuat dan dibina melalui Google Colaboratory dengan Python serta menggunakan platform Anvil untuk menghantar jadual tersebut ke laman web.

6 KESIMPULAN

Dalam projek ini, tajuknya merupakan Analisis Sentimen *Tweet* Bahasa Melayu Masa Nyata mengenai vaksin. Beberapa model analisis sentimen dengan pengelas-pengelas telah dibina dan dibandingkan untuk memperoleh model pengelas yang paling sesuai untuk menganalisis semua kandungan 'Tweet' dalam bahasa Melayu yang berkaitan dengan topik vaksin COVID-19 secara masa nyata, dan juga mengkategorikan data sama ada dalam kumpulan sentimen positif, negatif atau neutral. Sehingga bab ini, pelaksanaan projek ini telah merangkumi pengenalan projek, sorotan susastera, metodologi, spesifikasi reka bentuk, dan pembangunan serta pengujian sistem. Setiap proses tersebut telah memainkan peranan yang penting dalam pelaksaan projek ini dan proses-proses ini juga mempunyai hubungan kait antara satu sama lain. Akhirnya, model *Support Vector Machine (SVM)* telah dipilih sebagai model pengelas bagi analisis sentimen masa nyata *Tweet* bahasa Melayu terhadap vaksin. Hal ini demikian kerana model pengelas SVM mempunyai prestasi yang terbaik berbanding dengan model pengelas lain dengan hasilan penilaian bagi setiap model pengelas. Kesimpulannya, pelaksanaan dan pembinaan model pengelas bagi projek yang bertajuk Analisis Sentimen *Tweet* Melayu Masa Nyata mengenai vaksin telah berjaya dibangunkan dan dapat menepati objektif kajian projek tersebut. Kajian ini diharapkan dapat memberi informasi yang berguna tentang pandangan rakyat Malaysia terhadap vaksin pada baru-baru ini.

7 RUJUKAN

Oxford University Press. Oxford Learner Dictionaries (Sentiment). https://www.oxfordlearnersdictionaries.com/

Sejal Dua. 2021. Sentiment Analysis of COVID-19 Vaccine *'Tweet'*. https://towardsdatascience.com/sentiment-analysis-of-covid-19-vaccine-*'Tweet'*-dc6f41a5e1af

Abhishek Darekar. 2020. Live Twitter Sentiment Analysis (With Deployment). https://medium.com/analytics-vidhya/live-twitter-sentiment-analysis-with-deployment-e4a84f826b92

Earth Lab. 2020. Automate Getting Twitter Data in Python Using Tweepy and API Access. https://www.earthdatascience.org/courses/use-data-open-source-python/intro-to-apis/twitter-data-in-python/

Javatpoint. Machine learning Life cycle. https://www.javatpoint.com/machine-learning-life-cycle

Homage team. Your Guide To COVID-19 Vaccinations In Malaysia. https://www.homage.com.my/resources/covid-19-vaccine-malaysia/

Muriel Kosaka. 2020. Cleaning & Preprocessing Text Data for Sentiment Analysis. https://towardsdatascience.com/cleaning-preprocessing-text-data-for-sentiment-analysis-382a41f150d6

Muhammad Iqbal Aditama1 , Rizqeya Irfan Pratama, Kevin Hafizzana Untoro Wiwaha and Nur Aini Rakhmawati. 2020. Analisis Klasifikasi Sentimen Pengguna Media Sosial Twitter Terhadap Pengadaan Vaksin COVID-19. https://journal.unesa.ac.id/index.php/jieet/article/view/11018/pdf

Manish Todi. 2019. Sentiment Analysis using the Vader library. https://medium.com/analytics-vidhya/sentiment-analysis-using-the-vader-library-a91a888e4afd

Padmaja Savaram dan Sameen Fatima S. 2013. Opinion Mining and Sentiment Analysis - An Assessment of Peoples' Belief: A Survey.

https://www.researchgate.net/publication/276196657_Opinion_Mining_and_Sentiment_Analysis_-_An_Assessment_of_Peoples'_Belief_A_Survey

Aishwarya Mundalik. 2018. Aspect Based Sentiment Analysis Using Data Mining Techniques Within Irish Airline Industry MSc Research Project Data Analytics. https://www.researchgate.net/publication/334416954_Aspect_Based_Sentiment_Analysis_Using_Data_Mining_Techniques_Within_Irish_Airline_Industry_MSc_Research_Project_Data_Analytics

S. Rani and Parteek Kumar. 2017. A Sentiment Analysis System to Improve Teaching and Learning. https://www.semanticscholar.org/paper/A-Sentiment-Analysis-System-to-Improve-Teaching-and-Rani-Kumar/812964ed8ee979a8d1f8cd7021e1b77b532802af#paper-header

P.G. Preethi, V. Uma and Ajit Kumar. 2015. Temporal Sentiment Analysis and Causal Rules Extraction from *'Tweet'* for Event Prediction. https://www.researchgate.net/figure/Block-Diagram-of-proposed-work_fig1_277949851

Mohit Mertiya and Ashima Singh. 2016. Combining naive bayes and adjective analysis for sentiment detection on Twitter. https://www.semanticscholar.org/paper/Combining-naive-bayes-and-adjective-analysis-for-on-Mertiya-Singh/3fffb2eda13b46c064f9e2162e7a0fb0d64fbe67

Nazlia Binti Omar. 2021. Lecture 8 Sentiment Analysis.pptx. https://ukmfolio.ukm.my/pluginfile.php/2012621/mod_resource/content/1/Lecture%208%20Sentiment%20Analysis.pdf

Jawad Khan and Aftab Alam. 2016. Sentiment Analysis at Sentence Level for Heterogeneous Datasets. https://www.researchgate.net/publication/313248083_Sentiment_Analysis_at_Sentence_Level_for_Heterogeneous_Datasets

Manav Masrani and G. Poornalatha. 2017. Twitter Sentiment Analysis Using a Modified Naïve Bayes Algorithm. https://www.semanticscholar.org/paper/Twitter-Sentiment-Analysis-Using-a-Modified-Na%C3%AFve-Masrani-Poornalatha/80efdbc378fcda3b116664cd6b5b66ad4eb380e4

Laura O'Mahony. 2021. Getting Started with Data Collection Using Twitter API v2 in Less than an Hour. https://towardsdatascience.com/getting-started-with-data-collection-using-twitter-api-v2-in-less-than-an-hour-600fbd5b5558

Norlela Samsudin, Mazidah Puteh, Abdul Razak Hamdan and Mohd Zakree Ahmad Nazri. 2013. Normalization of Noisy Texts in Malaysian Online Reviews. https://www.researchgate.net/publication/287050449_Normalization_of_noisy_texts_in_Malaysian_online_reviews

Sajidah Ibrahim, Nor Zairah Ab Rahim, Fajar Ibnu Fatihan, Nur Azaliah Abu Bakar. 2021. Covid-19 Sentiment Analysis on Facebook Comments http://www.ijmtss.com/PDF/IJMTSS-2021-17-09-01.pdf

Buse Yaren Tekin. 2021. Sentiment Analysis with Logistic Regression. https://towardsai.net/p/nlp/sentiment-analysis-with-logistic-regression#:~:text=%F0%9F%93%8C%20Logistic%20Regression%20is%20a,models%20with%20a%20double%20situation

Sunil Ray. 2017. Analytics Vidhya-Naive Bayes Algorithm. https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

Stopwords (stopwords.net). 2021. Stopwords in Malay (MS) Language. https://stopwords.net/malay-ms/

Aman Khakharia, Vruddhi Shah, dan Pragya Gupta. 2021. Sentiment Analysis of COVID-19 Vaccine Tweets Using Machine Learning. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3869531

Meylan Wongkar, dan Apriandy Angdresey. 2019. Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter. https://www.researchgate.net/publication/339176682_Sentiment_Analysis_Using_Naive_Bayes_Algorithm_Of_The_Data_Crawler_Twitter

Kunnal Jaluthria, dan Sumedha Seniaray. 2021. Sentiment Analysis of Twitter Data Using Machine Learning Algorithm. https://www.semanticscholar.org/paper/Sentiment-Analysis-of-Twitter-Data-Using-Machine-Jaluthria/08af3af7e57706c186ab1102e1213a0032b9de9d

Siti Noor Allia Noor Ariffin and Sabrina Tiun. 2020. “Rule-based Text Normalization for Malay Social Media Texts” International Journal of Advanced Computer Science and Applications(IJACSA), 11(10). http://dx.doi.org/10.14569/IJACSA.2020.0111021 https://thesai.org/Publications/ViewPaper?Volume=11&Issue=10&Code=IJACSA&SerialNo=21

FreeMapTools Developers. 2016. FreeMapTools. https://www.freemaptools.com/

Jason Brownlee. 15 Januari 2020. Random Oversampling and Undersampling for Imbalanced Classification. https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/

Plotly Developers. Graph Objects in Python. https://plotly.com/python/graph-objects/

Ryan. Turning a Google Colab Notebook into a Web App. https://anvil.works/learn/tutorials/google-colab-to-web-app

Scikit-Learn Developers. 2007 – 2022. sklearn.metrics.classification_report. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html

Aman Kharwal. 7 Julai 2021. Classification Report in Machine Learning. https://thecleverprogrammer.com/2021/07/07/classification-report-in-machine-learning/

Kenneth Leung. 4 Januari 2022. Micro, Macro & Weighted Averages of F1 Score, Clearly Explained. https://towardsdatascience.com/micro-macro-weighted-averages-of-f1-score-clearly-explained-b603420b292f

Sarang Narkhede. 9 Mei 2018. Understanding Confusion Matrix https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

Wong Wai Jian (A175985)
Sabrina Binti Tiun
Fakulti Teknologi & Sains Maklumat,
Universiti Kebangsaan Malaysia