

ANALISIS SENTIMEN MASA NYATA TENTANG COVID-19 MALAYSIA MELALUI TWEET BAHASA MELAYU

TEO JIE XI

Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Aspek utama projek ini adalah fokus pada penyediaan set data analisis sentimen dan pembinaan model analisis sentimen dengan penggunaan teknik pembelajaran mesin. Masalah yang dihadapi adalah pembinaan set data semasa permulaan projek ini, di mana terdapat perkataan yang tidak formal dan campuran bahasa dalam data. Ketidakseimbangan data selepas dilabel juga merupakan antara satu masalah yang dihadapi dalam projek ini. Masalah lain yang dihadapi adalah membina model analisis sentimen yang terbaik daripada lima algoritma pembelajaran mesin. Bagi menyelesaikan masalah-masalah ini, data dijalankan pembersihan kepada satu set data yang bebas daripada perkataan tidak formal dan bahasa campuran. Perkataan bahasa Inggeris akan dikeluarkan daripada tweet. Teknik persampelan berlebihan digunakan untuk melaraskan taburan kelas bagi set data latihan. Pelbagai jenis algoritma mesin pembelajaran yang sesuai akan digunakan untuk membina model analisis sentimen teks tweet bahasa Melayu bertema COVID-19 Malaysia. Hanya model yang terbaik berbanding model lain akan dipilih sebagai model penyebaran (deployment model). Metodologi proses pembangunan yang digunakan dalam projek ini ialah 'Machine Learning Life Cycle'. Kaedah ini adalah proses kitaran di mana satu langkah demi satu langkah perlu diikuti sebelum sampai akhir projek. Dalam kaedah ini, selepas beberapa latihan dan ujian model, model pembelajaran mesin akan digunakan dengan data yang telah diproses. Dengan menggunakan kaedah ini, penyelidik perlu jelas tentang masalah projek dan menyelesaikan masalah melalui keseluruhan kitaran hayat. Kaedah ini dipilih kerana algoritma pembelajaran mesin yang berbeza digunakan dalam pembinaan model sebelum model dilatih dan diuji dengan data diproses. Demi analisis sentimen masa nyata ini berfungsi semasa implementasi, para-pemprosesan data dan teknik pembelajaran mesin perlu digunakan untuk memastikan model analisis sentimen yang dibina boleh berfungsi secara fleksibel dengan ketepatan tinggi. Pada akhir projek ini, satu sistem analisis sentimen yang mampu menganalisis tweet masa nyata berkaitan COVID-19 dengan kejutuan yang baik.

1 PENGENALAN

Pada tahun 2020, sesuatu penyakit berjangkit yang tidak pernah didengari merebak tanpa amaran dalam kalangan rakyat Malaysia. Penyakit tersebut semakin menjadi satu perkara yang ditakuti dan dirisau oleh orang ramai kerana kebanyakan orang yang dijangkiti penyakit tersebut meninggal dunia selepas beberapa hari. Penyakit yang menakutkan orang tersebut ialah Coronavirus (COVID-19).

COVID-19 ialah sejenis penyakit berjangkit baharu yang berpunca daripada virus corona yang tidak pernah dijumpa atau dikaji oleh saintis biologi. Mengapakah COVID-19 akan membawa kematian orang sebegini pantas? Hal ini kerana orang dijangkiti COVID-19 tanpa amaran. Simptom-simptom COVID-19 ini adalah serupa

dengan selesema yang biasa iaitu demam, batuk kering, sakit kepala dan keletihan. Simptom-simptom ini akan menjadi semakin serius apabila tidak ditemui dan dirawat dengan awal. Virus COVID-19 disebarkan melalui titisan cecair atau 'droplet' yang terhasil daripada percikan air liur ketika bercakap atau batuk. (BH Online, Mac 2020) Jangkitan ini juga mudah berlaku apabila orang berada di persekitaran tanpa pengudaraan yang baik atau berinteraksi dengan orang dijangkiti pada jarak yang dekat tanpa memakai topeng muka. (Kementerian Kesihatan Malaysia, May 2021) Oleh itu, rakyat Malaysia terpaksa duduk di rumah dan mengurangkan masa untuk keluar dari rumah kerana virus sedang merebak di kalangan masyarakat. Dengan ini, kebanyakan orang meluahkan perasaan dan meluangkan lebih banyak masa dalam penggunaan media sosial seperti Facebook, Instagram dan Twitter.

Sejak penciptaan media sosial, orang ramai bermula kerap menggunakan media sosial kerana ciri-cirinya yang menarik di mana orang ramai boleh menyiarkan informasi pada awam dan dapat mengenali ramai rakan maya yang mempunyai minat yang sama. Dalam erti kata lain, media sosial telah menjadi bahagian utama dalam kehidupan orang ramai; di mana masyarakat sering bergantung pada media sosial untuk mendapatkan berita dan mesej terbaru daripada rakan-rakan di alam maya. Pemprosesan analisis sentimen pada 'tweet' Canada berhubung dengan penjarakan sosial akan memberi pandangan yang berbeza tentang bagaimana sentimen ini boleh mempengaruhi tindakan orang lain serta membantu kerajaan dalam merangka tindakan muktamad dalam membentuk dasar dan prosedur awam yang berkaitan dengan penjarakan sosial dalam memerangi COVID-19, sambil mengekalkan kestabilan mental individu (Carol Shofiya dan Samina Abidi, 2021). Perasaan dan pemikiran rakyat Malaysia, terutamanya semasa wabak yang mendadak ini memainkan peranan penting dalam membantu keputusan berterusan kerajaan dan sentiasa mengetahui mentaliti rakyat. Oleh itu, analisis sentimen masa nyata Melayu akan digunakan untuk menganalisis dan mengetahui sentimen rakyat Malaysia terhadap COVID-19 kerana media sosial adalah platform yang sesuai untuk memudahkan pengekstrakan data.

Kini, era dunia sedang menuju ke era teknologi yang canggih dan moden, begitu juga negara Malaysia. Analisis sentimen masa nyata telah digunakan secara meluas dalam kalangan saintis data dan penganalisis data untuk memproses teks dalam pelbagai sektor. Analisis sentimen ialah teknik perlombongan teks yang menggunakan

pemrosesan bahasa tabii (NLP) dan algoritma pembelajaran mesin untuk mengekstrak dan menganalisis teks yang mengandungi sentimen. Analisis sentimen masa nyata untuk korpus bahasa Inggeris biasa dilihat tetapi tidak begitu kerap bagi korpus bahasa Melayu. Tambahan lagi, proses untuk analisis teks Melayu akan menjadi lebih rumit kerana kurang kamus atau perpustakaan dalam talian tersedia. Walau bagaimanapun, analisis sentimen akan dijalankan dengan teks Melayu untuk menentukan dan lebih memahami sentimen rakyat Malaysia terhadap COVID-19.

Membangunkan model teguh yang mampu menganalisis sentimen dalam masa nyata bagi isu COVID-19 di Malaysia boleh dilihat sebagai salah satu sumbangan dalam membantu kerajaan Malaysia dalam membuat dasar dan polisi mengenai COVID-19 dengan mengambilkira pandangan dan pendapat rakyat Malaysia. Projek ini juga boleh digunakan oleh Kementerian Kesihatan Malaysia (KKM) sebagai salah satu fitur dalam aplikasi laman web COVID-19 Malaysia – COVIDNOW (<https://covidnow.moh.gov.my>).

2 PENYATAAN MASALAH

Masalah-masalah yang dihadapi untuk membina sistem analisis sentimen bahasa Melayu untuk isu COVID-19 di kalangan pengguna Twitter di Malaysia adalah: (i)Pembinaan set data semasa permulaan projek ini di mana data tweet yang diperolehi belum dijalani pembersihan, dan terdapat perkataan yang tidak formal dalam data. Campuran bahasa Melayu dan bahasa Inggeris dalam penulisan tweet data juga merupakan masalah yang dihadapi. Selain daripada masalah bahasa, data tweet yang diperolehi adalah tidak seimbang. Ketidakseimbangan data selepas dilabel adalah di mana data bagi label neutral terlalu banyak berbanding label positif dan negatif. (ii)Masalah utama yang lain adalah membina model analisis sentimen yang terbaik daripada lima algoritma pembelajaran mesin yang terpilih.

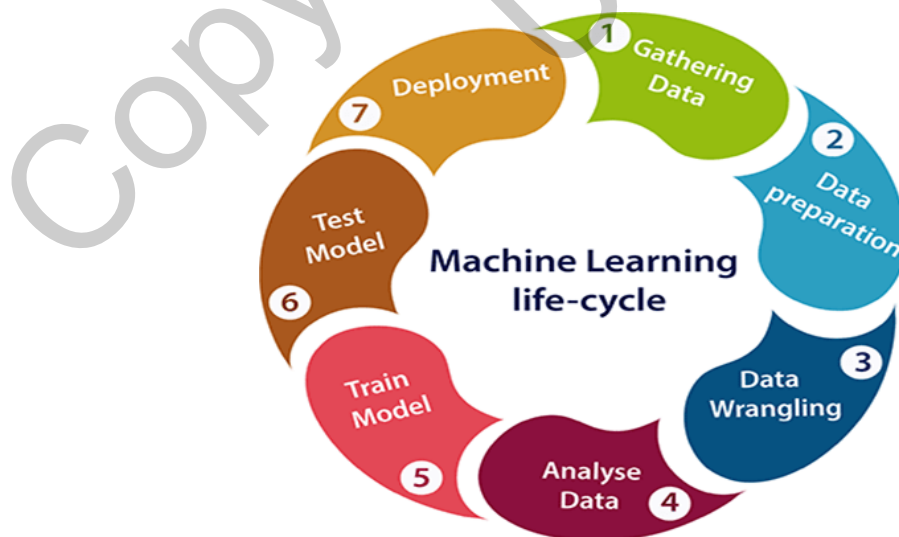
3 OBJEKTIF KAJIAN

Objektif kajian ini adalah: -

- Menyediakan set data analisis sentimen tweet bahasa Melayu untuk domain COVID- 19 Malaysia
- Membina model analisis sentimen tentang COVID-19 Malaysia pada tweet bahasa Melayu dengan menggunakan teknik pembelajaran mesin.

4 METOD KAJIAN

Metodologi proses pembangunan yang digunakan dalam projek ini ialah ‘Machine Learning Life Cycle’. Kaedah ini adalah proses kitaran di mana satu langkah demi satu langkah perlu diikuti sebelum sampai akhir projek. Dalam kaedah ini, selepas beberapa latihan dan ujian model menggunakan data yang telah diproses, model pembelajaran mesin berprestasi terbaik akan dipilih sebagai model penyebaran (deployment model). Dengan menggunakan kaedah ini, penyelidik perlu jelas tentang masalah projek dan menyelesaikan masalah melalui keseluruhan kitaran hayat. Kaedah ini dipilih kerana algoritma pembelajaran mesin yang berbeza diguna dalam pembinaan model sebelum model dilatih dan diuji dengan set data. Untuk memastikan analisis sentimen masa nyata ini dalam projek ini berfungsi semasa implementasi, pemprosesan data dan teknik pembelajaran mesin perlu digunakan untuk memastikan model analisis sentimen yang dibina boleh berfungsi tanpa ralat dengan ketepatan ramalan yang tinggi.



Rajah 4.1: Gambar rajah tentang Machine Learning Life Cycle dari Sonoo Jaiswal (2021)

4.1 Fasa Data Kumpulan, Penyediaan Data dan Pra-Pemrosesan Data

Semua data yang diekstrak daripada Twitter tweet adalah mentah, data tersebut belum menjalani sebarang proses teks. Teks pra-pemrosesan termasuk pembersihan, tokenisasi, penormalan dan penyingkiran kata henti. Pembuangan tanda baca juga dilakukan untuk memastikan seluruh data adalah bersih untuk meneruskan analisis sentimen. Bahasa dalam data adalah bahasa campuran. Oleh itu, pembuangan perkataan bahasa Inggeris juga dilakukan selain daripada pra-pemrosesan teks. Selepas pra-pemrosesan, data teks yang akhir akan bebas daripada URL, sebutan pengguna, slanga, singkatan, nombor dan tanda baca. Data akan dilabel untuk digunakan semasa latihan dan ujian model.

4.2 Fasa Analisis Data

Data yang telah dilabel (positif, negatif dan neutral) akan dibahagikan kepada data latihan dan data ujian. Data latihan akan digunakan untuk model latihan manakala data ujian akan digunakan untuk model ujian. Pengekstrakan ciri dengan Kekerapan Jangka - Kekerapan Dokumen Songsang (TFIDF) diguna dalam proses pembahagian data kepada data latihan dan ujian. Jumlah data label bagi set data adalah tidak seimbang. Teknik persampelan berlebihan digunakan supaya jumlah bagi ketiga-tiga label sama dan seimbang. Persampelan berlebihan ialah teknik yang digunakan untuk menyesuaikan taburan kelas bagi set data latihan. Teknik persampelan berlebihan dipilih untuk menyeimbangkan label set data kerana teknik ini dapat mengandakan data dalam label set data yang berkurangan jumlahnya supaya jumlah setiap label seimbang. Kegunaan teknik ini dapat mengelakkan data tidak mencukupi selepas data diseimbangkan.

4.3 Fasa Latihan Model

Dalam fasa ini, model akan dilatih dengan data terlatih untuk menambahbaik prestasi model ini. Dalam langkah ini, satu model pembelajaran mesin dibina untuk menganalisis data dengan penggunaan pelbagai teknik analisis dan mengkaji hasilnya. 5 algoritma pembelajaran mesin dicadangkan untuk digunakan dalam membina model analisis sentimen COVID-19 Bahasa Melayu, iaitu Mesin Vektor Sokongan (SVM), Regresi Logistik, Naïve Bayes, K-Nearest Neighbors dan Hutan Rawak. Model yang dibina akan diuji dalam langkah kemudian.

4.4 Fasa Ujian Model

Setelah model pembelajaran mesin dilatih, model ini kemudiannya diuji untuk menilai ketepatan model, sama ada model ini boleh mencapai keperluan projek atau tidak.

$$\%Ketepatan = \frac{\text{bilangan klasifikasi yang betul}}{\text{bilangan semua sampel}}$$

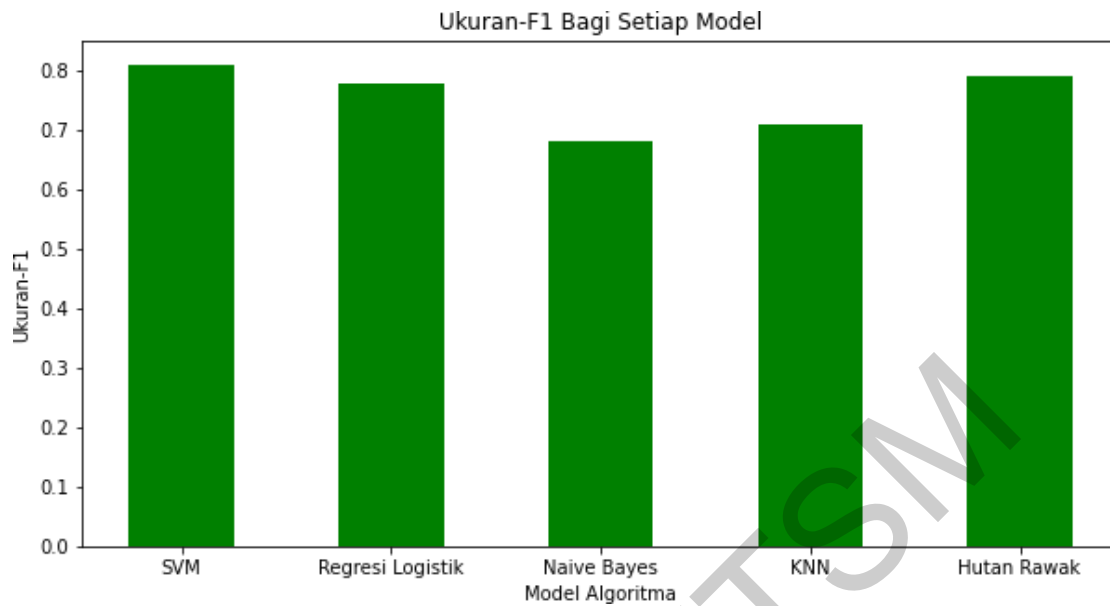
4.5 Fasa Penyebaran Model

Jika model yang disediakan dapat menghasilkan ketepatan yang tinggi dan mencapai keperluan projek, maka model ini dapat digunakan dalam analisis sentimen masa nyata untuk menganalisis sentimen tweet dengan tepat.

5 HASIL KAJIAN

Jadual 5.1 Keputusan Pengujian Setiap Model

Jenis Model	Purata Mikro	Purata Makro			Purata Wajaran		
	Skor-F1	Kepersisan	Ingatan	Skor-F1	Kepersisan	Ingatan	Skor-F1
SVM	0.81	0.62	0.57	0.59	0.79	0.81	0.80
Regresi	0.78	0.58	0.59	0.58	0.79	0.78	0.79
Logistik							
Naïve	0.68	0.52	0.64	0.54	0.80	0.68	0.71
Bayes							
KNN	0.71	0.50	0.56	0.52	0.76	0.71	0.73
Hutan	0.79	0.59	0.50	0.53	0.77	0.79	0.77
Rawak							



Rajah 5.1 Graf Bar Ukuran -F1 (Purata Mikro) Bagi Setiap Model

Jadual 5.2 Perbandingan Ketepatan Model Nisbah 80:20 dan 70:30

Jenis Model	Ketepatan (%)	
	Nisbah 80:20	Nisbah 70:30
SVM	80.61	79.41
Regresi Logistik	78.38	78.87
Naïve Bayes	67.68	65.68
KNN	70.71	70.39
Hutan Rawak	79.40	80.35

Jadual 5.1 menunjukkan nilai kepersisan, ingatan dan skor-F1 setiap model bagi purata mikro, purata makro dan purata wajaran. Purata mikro dinamakan sebagai ketepatan dalam laporan klasifikasi kerana ketepatan membawa definisi yang sama dengan purata mikro, di mana purata mikro mengira bahagian pemerhatian yang betul terkelas daripada semua pemerhatian. Purata mikro dijadikan peranan utama untuk menilai model yang dibina kerana purata mikro sesuai digunakan semasa jumlah set data adalah seimbang dan mudah difahami untuk prestasi keseluruhan model. Model SVM mempunyai skor-F1 yang paling tinggi dalam purata mikro, purata makro dan purata wajaran. Model Naïve Bayes mempunyai ketepatan yang paling rendah bagi ketepatan dan purata wajaran, tetapi bukan yang paling rendah bagi purata makro. Jadual 5.2 menunjukkan ketepatan semasa nisbah 80:20 dan 70:30 digunakan. Ketepatan model SVM,

model SVM. Terdapat positif, negatif dan neutral dalam tweet diekstrak. Tweet dan sentimen diramal dimasukkan ke *DataFrame* supaya keputusan yang ringkas mudah dilihat seperti di Rajah 5.3.

```
tweet_text['label'].value_counts()
neutral      482
positive     56
negative     45
Name: label, dtype: int64
```

Rajah 5.4 Kod Mengira Jumlah Label

Bagi mengetahui jumlah setiap sentimen diramal, 'value_counts' diguna untuk mengirajumlah mereka. Daripada Rajah 5.4, dapat mengetahui bahawa jumlah tweet yang bersentimen neutral melebihi yang bersentimen positif dan negatif. Ini bermaksud kebanyakan tweet yang diekstrak adalah bermaklumat dan tidak membawa apa-apa sentimen.

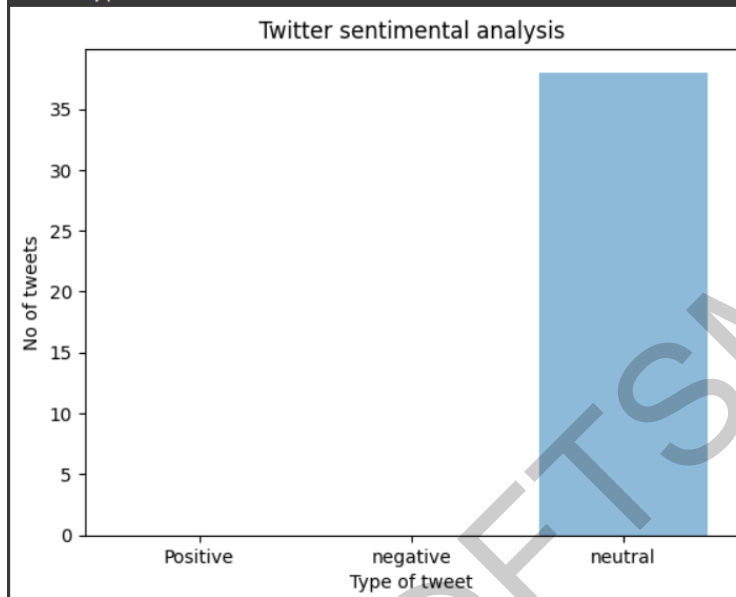
```
what you want to search? : covid19 malaysia

No of total tweets: 38
no of positive tweets: 0
no of negative tweets: 0
no of neutral tweets: 38

Positive tweets percentage: 0.0 %
Negative tweets percentage: 0.0 %
Neutral tweets percentage: 100.0 %
```

For positive tweets press 1
 For negative tweets press 2
 For neutral tweets press 3
 For bar chart press 4

:4



press -1 for exit press any no for repeat

:5

For positive tweets press 1
 For negative tweets press 2
 For neutral tweets press 3
 For bar chart press 4

:3

Neutral tweets:

Info Covid 19 Malaysia 9 Jun 2022 setakat jam 11 59 malam infografik coronavirus covid19 penjarakanfizikal
 Info Covid 19 Malaysia 9 Jun 2022 setakat jam 11 59 malam covid19 kitajagakita jagajarak
 RT Jumaat 10 Jun 2022 Jumlah kes baharu COVID19 yang dilaporkan SEMALAM adalah 1 887 kes menjadikan jumlah kumulatif sebanyak 22 074 kes aktif direkodkan dengan 95 5 daripadanya sedang menjalani kuarantin sendiri MalaysiaNow Mnow
 Jumaat 10 Jun 2022 Jumlah kes baharu COVID19 yang dilaporkan SEMALAM adalah 1 887 kes menjadikan jumlah kumulatif
 RT Khamis 9 Jun 2022 Jumlah kes baharu COVID19 yang dilaporkan SEMALAM adalah 1 518 kes menjadikan jumlah kumulatif sebanyak
 RT Tentera Laut Diraja Malaysia TLDM berjaya kesan dan pintas bot PATI Indonesia ceroboh perairan Johor rasmi
 RT Minggu Sains Negara Keluarga Malaysia IOW KomunikasiKita LindungDiriLindungSemua KitaJagaKita KitaMestiMenang C
 Tentera Laut Diraja Malaysia TLDM berjaya kesan dan pintas bot PATI Indonesia ceroboh perairan Johor rasmi
 21 589 kes aktif direkodkan dengan 95 4 daripadanya sedang menjalani kuarantin sendiri MalaysiaNow Mnow
 RT Langkah Kerajaan menstabilkan harga barangan makanan Menangani isu kenaikan harga barangan Minggu Sains Negara Keluarga
 Khamis 9 Jun 2022 Jumlah kes baharu COVID19 yang dilaporkan SEMALAM adalah 1 518 kes menjadikan jumlah kumulatif
 Minggu Sains Negara Keluarga Malaysia IOW KomunikasiKita LindungDiriLindungSemua KitaJagaKita
 Malaysia Perluas Penggunaan Obat Antiviral Oral Covid 19
 RT 1 364 sembuh COVID 19 ubat antiviral Paxlovid diperluas ke fasiliti swasta AWANInews AWANI745
 RT Protes di HQ AirAsia jalan akhir 8 Jun 2022 MalaysiaNow News FlashNews Malaysia airasia protest airasiayahanoi
 Protes di HQ AirAsia jalan akhir 8 Jun 2022 MalaysiaNow News FlashNews Malaysia airasia protest
 1 364 sembuh COVID 19 ubat antiviral Paxlovid diperluas ke fasiliti swasta AWANInews AWANI745
 Namun pengembara masih tertakluk dengan satu lagi syarat setelah memasuki negara itu SouthKorea Covid19
 RT Rabu 8 Jun 2022 Jumlah kes baharu COVID19 yang dilaporkan SEMALAM adalah 1 128 kes menjadikan jumlah kumulatif sebanyak 4
 RT Empat daripada kes kematian berlaku di luar hospital atau brought in dead BID AWANInews AWANIpagi MalaysiaSihat htt
 Empat daripada kes kematian berlaku di luar hospital atau brought in dead BID AWANInews AWANIpagi
 RT Rabu 8 Jun 2022 Jumlah kes baharu COVID19 yang dilaporkan SEMALAM adalah 1 128 kes menjadikan jumlah kumulatif se
 RT Minggu lepas demam minggu ni Covid APA KOMEN ANDA Via Awani

Langkah Kerajaan menstabilkan harga barangan makanan Menangani isu kenaikan harga barangan Minggu Sains Negara
 RT Kumpulan pertama jemaah haji Malaysia sudah pun tiba di Madinah Rasa bersyukur mengiringi ketibaan itu selain memecah ker
 Kumpulan pertama jemaah haji Malaysia sudah pun tiba di Madinah Rasa bersyukur mengiringi ketibaan itu selain meme
 RT Risaalah Keluarga Malaysia Bendera Negeri amp Wilayah Risaalah Perpaduan dalam Kepelbagaian Risaalah Panduan Pengurus
 RT Menurut PAPMSA kebanyakan industri terutama perladangan serta perusahaan kecil dan sederhana berdepan kekurangan tenaga kerj
 RT Sejumlah 1 733 437 atau 48 9 peratus daripada kumpulan itu menerima sekurang kurangnya satu dos vaksin AWANInews AWANIpag
 RT Selasa 7 Jun 2022 Jumlah kes baharu COVID19 yang dilaporkan SEMALAM adalah 1 330 kes menjadikan jumlah kumulatif sebanyak
 RT WAJARKAH PRU KE 15 DIADAKAN DALAM MASA TERDEKAT INI KATA RAKYAT MALAYSIA rakyatmalaysia malaysia politikmalaysia co
 Sejumlah 1 733 437 atau 48 9 peratus daripada kumpulan itu menerima sekurang kurangnya satu dos vaksin AWANInews
 Menurut PAPMSA kebanyakan industri terutama perladangan serta perusahaan kecil dan sederhana berdepan kekurangan t
 Situasi Terkini COVID 19 di Malaysia 6 Jun 2022 COVID19 LindungiDiriLindungiSemua KeluargaMalaysia
 Sementara itu 93 7 populasi remaja telah lengkap divaksinasi MalaysiaNow MNow MNowNews Malaysia COVID19
 RT Ini menjadikan jumlah keseluruhan kematian disebabkan pandemik berkenaan sebanyak 35 690 AWANInews AWANIpagi MalaysiaSih
 Kadar penggunaan kemudahan ICU di seluruh negara kini sebanyak 57 6 dengan kes Covid 19 merangkumi 4 9

press -1 for exit press any number for repeat

:-1

Rajah 5.5 Pengeluaran bagi Penyebaran Model

6 KESIMPULAN

Kesimpulannya, pembangunan sistem model analisis sentimen masa nyata tentang COVID-19 Malaysia melalui tweet bahasa Melayu telah berjaya dibangunkan. Cadangan penyelesaian untuk masalah-masalah dan objektif dalam projek ini juga dicapai sepanjang pembangunan projek ini.

Analisis sentimen bahasa Melayu terutamanya pada topik COVID-19 masih kurang diselidik dan dikaji oleh penyelidik pakar data kerana topik masih baharu dan kurang analisis sentimen topik ini dalam bahasa Melayu. Cadangan penambahbaikan projek dinyatakan iaitu menambah lebih banyak tinjauan dalam analisis sentimen yang berbahasa lain daripada bahasa Inggeris, menggunakan algoritma pemprosesan bahasa semula jadi untuk mengesan perkataan yang tidak formal atau salah ejaan dalam teks supaya membetulkan perkataan tersebut secara mudah, dan mengumpul set data yang besar tentang COVID-19 dalam bahasa Melayu supaya sebuah model analisis sentiment yang dapat meramal sentiment untuk tweet bahasa Melayu masa nyata dengan lebih tepat. Cadangan-cadangan tersebut diberikan supaya sistem seperti projek ini dapat menjadi kajian yang ditambahbaik dan banyak digunakan pada masa hadapan.

7 RUJUKAN

Abdullah Alsaedi, Mohammad Zubair Khan, 2019. A Study on Sentiment Analysis Techniques of Twitter Data. *International Journal of Advanced Computer Science and Applications* 10(2):361-374
DOI:10.14569/IJACSA.2019.0100248

Catris Code, 2022. Extracting (or Removing) Mentions and Hashtags in Tweets using Python. <https://catriscode.com/2021/03/02/extracting-or-removing-mentions-and-hashtags-in-tweets-using-python/>

Data Science Parichay, 2022. Get data from Twitter API in Python – Step by Step Guide. <https://datascienceparichay.com/article/get-data-from-twitter-api-in-python-step-by-step-guide/>

- Hassan Raza, M. Faizan, Ahsan Hamza, Ahmed Mushtaq, Naeem Akhtar, 2019. Scientific Text Sentiment Analysis using Machine Learning Techniques. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 12, 2019.
- Jaspreet Singh, Gurvinder Singh, Rajinder Singh, 2017. Optimization of Sentiment Analysis Using Machine Learning Classifiers. Singh et al. Hum. Cent. Comput. Inf. Sci. (2017) 7:32 DOI 10.1186/s13673-017-0116-3
- Kementerian Kesehatan Malaysia, 2021. Bolehkah COVID-19 merebak melalui udara?. <https://covid-19.moh.gov.my/semasa-kkm/2021/05/bolehkah-covid-19-merebak-melalui-udara>
- Kurtis Pykes, 2020. Oversampling and Undersampling, Towards Data Science. <https://towardsdatascience.com/oversampling-and-undersampling-5e2bbaf56dcf>
- K. Sentamilselvan, D. Aneri, A. C. Athithiya, P. Kani Kumar, 2020. Twitter Sentiment Analysis using Machine Learning Techniques. International Journal of Engineering and Advanced Technology (IJEAT), ISSN: 2249 – 8958, Volume 9 Issue-3 DOI: 10.35940/ijeat.C6281.029320
- Matt Clarke, 2021. How to save and load machine learning models using Pickle, Practical Data Science. <https://practicaldatascience.co.uk/machine-learning/how-to-save-and-load-machine-learning-models-using-pickle>
- N.Mahendran, T.Mekala, 2018. A Survey: Sentiment Analysis Using Machine Learning Techniques for Social Media Analytics. International Journal of Pure and Applied Mathematics, Volume 118 No.8 2018, 419-423 ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (online version)
- Siti Noor Allia Noor Ariffin, dan Sabrina Tiun 2020. Rule-based Text Normalization for Malay Social Media Texts, International Journal of Advanced Computer Science and Applications. 11(10):156-162 11(10):156-162
- Sonoo Jaiswal, 2021. Machine Learning Life Cycle, javaTpoint. <https://www.javatpoint.com/machine-learning-life-cycle>
- Sourabh-agrawal, 2017. Twitter Sentimental analysis of real time data. Github. <https://github.com/sourabh-agrawal/Twitter-Sentimental-analysis-of-real-time-data.git>