SENTIMEN ANALISIS COVID-19 DI TWITTER MENGGUNAKAN PEMBELAJARAN MESIN

Syed Putra Bin Syed Salim

Prof. Madya Dr. Nazlia Binti Omar

ABSTRAK

Isu berkaitan vaksin di dunia semasa pandemik *CoVid-19* menjadi semakin panas apabila terdapat pelbagai isu yang dikaitkan dengan kematian. Ternyata ianya tidak dapat dibuktikan dengan kes siasatan dan fungsi vaksin seperti diperkatakan di media sosial seperti Twitter. Hari semakin hari, jumlah twit yang terkumpul semakin banyak sehingga mencecah ratusan ribu. Hal ini telah menyukarkan sesetengah pihak yang ingin mendapatkan gambaran keseluruhan mengenai sesuatu isu yang diperkatakan. Oleh itu, kajian ini bertujuan untuk mengelaskan twit berasaskan vaksin berdasarkan sentimen seperti positif, neutral dan negatif bagi membantu dan memudahkan pihak berkepentingan atau kerajaan membuat keputusan penting dalam pentadbiran. Kajian ini menggunakan bahasa pengaturcaraan Python dan beberapa jenis model pengelas seperti *Naïve Bayes*, *Random Forest* dan *Support Vector Machine*(SVM*)*. Hasil analisis tersebut memberikan peratusan sentimen keseluruhan bagi setiap topik dan boleh dirujuk apabila sesetengah pihak berkepentingan memerlukan panduan dalam isu-isu yang tersenarai.

1 PENGENALAN

Media sosial merupakan sebahagian daripada kehidupan masyarakat yang berperanan sebagai sebuah platform untuk berkomunikasi antara satu sama lain tanpa batasan. 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). Media sosial meliputi pelbagai aplikasi seperti Reddit, Facebook, Instagram, Twitter dan banyak lagi. Pada era yang semakin berkembang teknologinya, kebanyakan pengguna media sosial menggunakan media sosial sebagai platform untuk meluahkan perasaan dan juga pendapat. Sejajar dengan keadaan semasa, pengguna media sosial sering meluahkan pendapat mereka terhadap isu mengambil vaksin yang semakin terpesong daripada fakta asal di saat pandemik CoVid-19 melanda bukan sahaja di negara kita malah satu dunia. Hal ini telah menyebabkan pertambahan sentimen yang ada di dalam media sosial.

Secara kesimpulannya, kajian ini dicadangkan kerana terdapat cabaran yang dihadapi oleh pengkaji untuk menganalisis sentimen vaksin dalam era pandemik CoVid-19 ini. Pendekatan secara analisis sentimen dicadangkan untuk menganalisis twit di Twitter bagi mengkaji pandangan awam terhadap isu-isu berkaitan vaksin yang berlaku semenjak wabak pandemik melanda di dunia ini. Sentimen-sentimen ini dianalisis dan dikelaskan kepada beberapa kumpulan sentimen yang berbeza iaitu positif, neutral dan negatif. Hasil akhir kajian ini boleh membantu pihak yang berkepentingan atau kerajaan dalam membuat keputusan penting dalam urusan pentadbiran.

2 PENYATAAN MASALAH

Melalui pemerhatian yang dijalankan, terdapat ribuan sentimen twit yang berasaskan vaksin di dalam media sosial *Twitter*. Semenjak pengambilan vaksin diwajibkan di seluruh negara, terdapat beberapa individu yang memandang perkara ini secara remeh dan ada juga individu yang berkempen untuk menolak pengambilan vaksin. Hal ini telah menyebabkan peningkatan jumlah twit sentimen yang sangat banyak. Apabila terdapat begitu banyak pendapat yang dilontarkan oleh orang ramai, sentimen positif, neutral dan juga negatif telah bercampur-aduk. Ini telah mendatangkan cabaran kepada seseorang pengkaji yang ingin menganalisis sentimen vaksin pada era pandemik *CoVid-19* ini. Perkara ini juga telah menyukarkan semua pihak

seperti orang ramai, pihak berkepentingan dan kerajaan dari mendapat gambaran keseluruhan keputusan bagi setiap isu vaksin yang dibincangkan.

3 OBJEKTIF KAJIAN

Objektif utama dalam projek Analisis Sentimen Terhadap Vaksin CoVid-19 di Twitter Menggunakan Pembelajaran Mesin ini ialah:

i. Merekabentuk dan membangunkan sebuah model yang dapat mengelaskan sentimen vaksin CoVid-19 yang terdapat di dalam Twitter menggunakan pendekatan algoritma pembelajaran mesin.
ii. Membangunkan sistem visualisasi pengkelasan sentimen vaksin secara atas talian.

4 METOD KAJIAN

Metodologi menunjukkan proses penting yang berlaku dalam memastikan sesuatu projek yang dijalankan berjalan dengan lancar mengikut fasa yang telah ditetapkan. Rajah 1 menunjukkan langkah-langkah yang terlibat dalam melakukan proses analisis sentimen vaksin CoVid -19 di Twitter.

Rajah 1 Metodologi

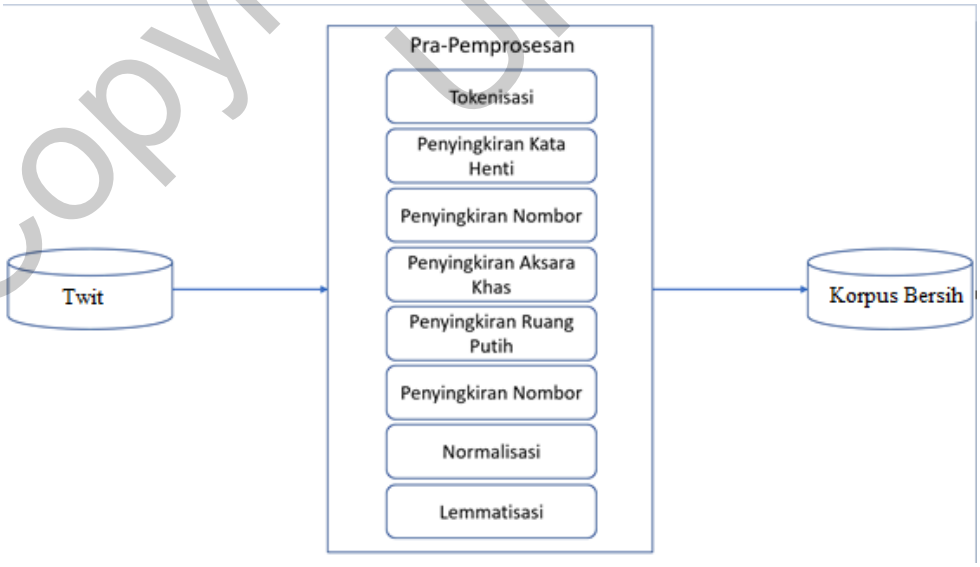
4.1 Fasa Pengumpulan Data

Fasa ini melibatkan cara pengumpulan data dan analisis terhadap data terkumpul. Ini bertujuan untuk memastikan data yang dikumpulkan sesuai dengan model yang dipilih dengan melalui pra-pemprosesan data. Sebanyak 79,650 set data berlabel positif, neutral dan negatif diambil daripada github. Sebanyak 38,017 dilabel positif, 25,368 dilabel neutral dan 16265 dilabel negatif. Data yang digunakan bagi menjalankan kajian ini diambil daripada platform Twitter.

Set data daripada Twitter dikumpulkan melalui aplikasi Twitter *API*. Set data yang dikumpul merupakan data mentah (*raw data*) dan disimpan dalam dokumen berformat CSV. Sebanyak 7016 set data berjaya dirangkak keluar daripada Twitter.

4.2 Fasa Pra-pemprosesan

Proses pra-pemprosesan ini penting kerana ianya proses yang mengubah data mentah kepada format yang boleh difahami oleh mesin dan membantu mengubah bunyi dari ciri dimensi tinggi ke ruang dimensi rendah untuk mendapatkan seberapa banyak maklumat yang tepat dari teks. Teknik ini dapat membantu meningkatkan ketepatan model pengelas. Rajah 2 menunjukkan proses pra-pemprosesan.



Rajah 2 Proses Pra-pemprosesan

4.3 Fasa Pengekstrakan Ciri

Pada fasa ini, set data bersih akan diubah menjadi satu set ciri untuk menjalankan tugas dalam melabel twit yang ditulis oleh pengguna Twitter sama ada positif, negative atau neutral. Antara pengekstrakan ciri yang digunakan dalam kajian ini adalah:

a. ***Bag of Words(BoW)***

Setelah melakukan fasa pra-pemprosesan yang menghasilkan token, proses seterusnya adalah pengekstrakan ciri yang memgubah token menjadi ciri yang digunakan untuk model. Pendekatan berasaskan leksikon menggunakan kata adjektif dan kata sifat untuk menemukan orientasi semantik teks(Neha Gupta, 2020).

b. ***TF-IDF (Frekuensi Istilah – Frekuensi Dokumen Terbalik)***

i. **Frekuensi Istilah(TF)** - merupakan istilah kekerapan perkataan dalam dokumen. Terdapat beberapa cara untuk mengira frekuensi, antaranya ialah jumlah kejadian mentah yang muncul dalam satu dokumen. Kemudian, cara lain ialah menyesuaikan frekuensi dengan panjang dokumen atau dengan frekuensi mentah dari perkataan yang banyak kekerapan dalam suatu dokumen

ii. **Frekuensi Dokumen Terbalik(IDF)** – merupakan kekerapan dokumen terbalik bagi sebilangan dokumen. Ini bermaksud, seberapa umum atau jarang perkataan yang terdapat dalam keseluruhan set dokumen. Semakin dekat dengan 0, semakin umum perkataan. Metrik ini dapat dikira dengan mengambil jumlah dokumen, membahaginya dengan jumlah dokumen yang mengandungi perkataan dan mengira logaritma.

4.4 Fasa Pengelas

Dalam fasa model pengelas, terdapat beberapa proses yang dialami oleh korpus sebelum ianya disimpan di dalam fail format .csv. Korpus yang sudah dibersihkan dimasukkan ke dalam model pengelas iaitu pengelasan menggunakan naif bayes. Model pengelas ini terlebih dahulu dilatih dengan menggunakan set data latihan untuk membenarkan model pengelas belajar data yang yang ingin dikaji. Setelah model pengelas ini dilatih, set data ujian pula akan dimasukkan

untuk mengalami analisis sentimen. Apabila selesai analisis sentimen ke atas korpus, pemarkahan bagi setiap twit akan diberikan.

Pemarkahan sentimen adalah berdasarkan keseluruhan kekuatan sentimen bagi setiap perkataan dalam twit. Setelah setiap twit mempunyai markah yang tersediri, maka proses klasifikasi dalam dijalankan berdasarkan sentimen sesuatu twit itu positif, neutral atau negatif. Pada proses klasifikasi ini, pemeriksaan manual secara rawak dilakukan bagi memastikan analisis sentimen yang dilakukan oleh model berfungsi dengan baik. Setelah semuanya dalam keadaan yang baik, twit yang telah dilabel mengikut sentimen masing-masing akan disimpan di dalam fail format .csv yang baru.

4.5 Fasa Penilaian

Dalam fasa ini, F-Skor bagi model pengelas *Naive Bayes*, *Random Forest*, dan SVM juga dianalisis. Ini kerana F-Skor menunjukkan betapa tepat model dengan memberitahu berapa banyak klasifikasi betul yang dilakukan. Julat nilai F-Skor adalah di antara 0 dan 1. Jadi, semakin tinggi nilai F-Skor, semakin tinggi prestasi model. Model pengelas yang memberi skor tertinggi akan digunakan dalam data pengujian untuk menguji ketepatan sentimen.

5 HASIL KAJIAN

Bahagian ini membincangkan hasil daripada proses pembangunan model Analisis Sentimen Covid-19 Di Twitter Menggunakan Pembelajaran Mesin. Penilaian terhadap prestasi bagi setiap pengelas dibentangkan di dalam jadual dan dibincangkan. Jadual 1 merupakan hasil penilaian dan perbandingan bagi tiga pengelas yang digunakan untuk sentimen vaksin yang diukur dari penilaian Kejituan, Dapatan, dan F1-Skor.

Jadual 1 Hasil Penilaian dan Perbandingan prestasi model pengelas

	Kejituan	***Dapatan***	***Skor-F1***
Naive Bayes	0.81	0.41	0.54
SVM	0.76	0.74	0.76
Random Forest	0.74	0.63	0.69

F1-skor adalah salah satu penilaian metrik yang paling efektif jika dibandingkan dengan ketepatan dalam mendapatkan model pengelas yang terbaik. Ini kerana F1-skor sangat membantu jika taburan kelas bagi set data adalah tidak seimbang. Untuk mendapatkan F1-skor yang tinggi, pemilihan model pengelas amat penting bagi kajian berkaitan. Rajah 4.10 menunjukkan visualisasi prestasi model pengelas berdasarkan nilai F1-skor. SVM memberikan prestasi yang paling baik iaitu dengan nilai 0.76. Kemudian model SVM dan *feature extraction* disimpan menggunakan library *pickle*.

Fasa ini menggunakan data pengujian iaitu data yang diekstrak menggunakan *Tweepy* yang mengandungi 7016 twit yang tidak dilabel. Data pengujian terlebih dahulu memerlukan melalui pra-pemprosesan bagi menghasilkan data bersih. Data pengujian akan dimasukkan ke dalam model yang telah simpan. Jadual 2 menunjukkan hasil analisis kajian ke atas data pengujian dan jadual 3 menunjukkan hasil penilaian dan perbandingan model pengelas ke atas data pengujian.

Jadual 2 Hasil Analisis Kajian ke atas Data Pengujian

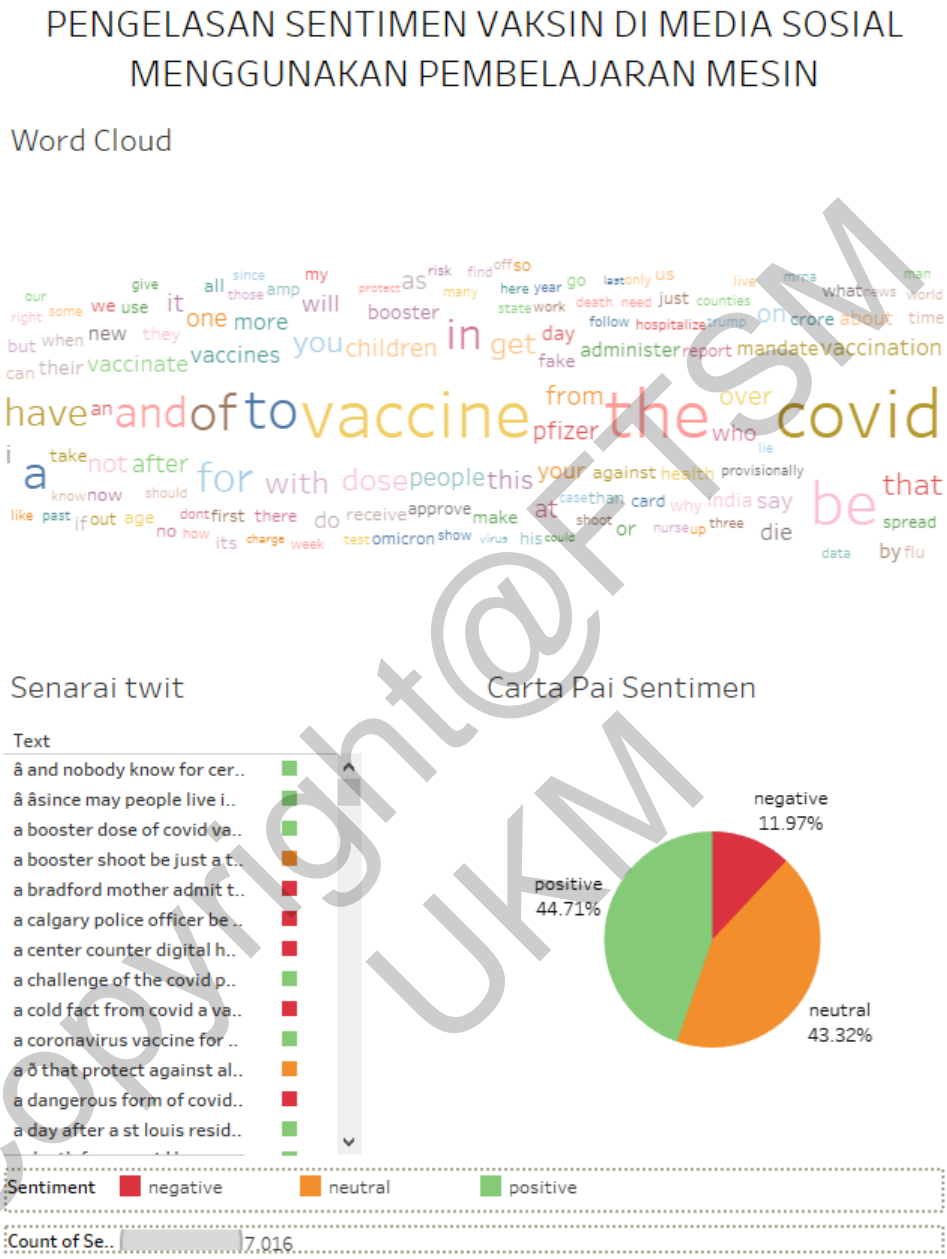
Model/ Kejituan	Naive Bayes	SVM	Random Forest
Kejituan data latihan	0.88	0.99	1.0
Kejituan data pengujian	1.0	1.0	1.0

Jadual 3 Hasil Penilaian dan Perbandingan prestasi model pengelas ke atas Pengujian Data

	Kejituan	***Dapatan***	***Skor-F1***
Naive Bayes	0.83	0.62	0.74
SVM	0.86	0.75	0.82
Random Forest	0.83	0.68	0.77

Dalam kajian ini, *Tableau Public* digunakan bagi menghasilkan visual untuk papan pemuka. *Tableau Public* merupakan satu aplikasi visualisasi dan analisis data interaktif yang membolehkan pengguna menjana papan pemuka dengan mudah hanya dengan memasukkan data. Selepas data dimasukkan, aplikasi ini akan memberi beberapa pilihan visualisasi untuk dipilih oleh pengguna. Papan pemuka dihasilkan bagi memudahkan penganalisis memahami

keputusan yang dihasilkan. Rajah 3 dibawah menunjukkan visual bagi papan pemuka untuk kajian ini.



Rajah 3 Papan pemuka

6 KESIMPULAN

Analisis Sentimen di Twitter Menggunakan Pembelajaran Mesin ini dijangka dapat membantu dan memudahkan pihak berkepentingan atau kerajaan membuat keputusan penting dalam pentadbiran. Hasil daripada analisis sentimen yang dilakukan telah divisualisasikan dalam bentuk graf bagi memudahkan orang ramai dan juga pihak berkepentingan untuk memahami hasil yang diperolehi. Penggunaan perisian *Tableau Public* dalam projek ini telah membantu memudahkan kerja visualisasi hasil yang diperoleh. Fungsi interaktif yang ditawarkan oleh *Tableau Public* telah menjadikan hasil visualisasi semakin cantik dan menarik. Walaupun terdapat kekangan yang dihadapi semasa melakukan kajian ini, ianya telah diselesaikan dengan menggunakan teknik yang sesuai.

7 RUJUKAN

Anonymous. (Dec 2015). Evolusi Media Sosial.

Aswinkrishna21. (2021). COVID19_sentinentanalysis. www.github.com/sydney-machine-learning/COVID19_sentinentanalysis.

Azhar, Muhammad. Hafidz, Noor. Rudianto, Biktra. Gata, Windu. (Sep 2014). Marketplace Sentiment Analysis Using Naïve Bayes and Support Vector Machine. Universiti Islam 45 Bekasi.

Amarja Adgaonkar, Bharti Khemani. (Apr 2021). A Review on Reddit News Headlines with NLTK Tools. SSRN.

Bannister, Kristian. (Oct 2018). Understanding Sentiment Analysis: What It Is & Why It's Used. Brandwatch.

Davydova, Olga. Data Monsters. (Oct 2018). Text Pre-processing in Python: Steps, Tools, and Examples. Medium.

D. A. Nurdeni, I. Budi and A. B. Santoso, Sentiment Analysis on Covid19 Vaccines in Indonesia: From The Perspective of Sinovac and Pfizer, 2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT), Surabaya, Indonesia, 2021, pp. 122-127

Leshuan A/L Sivakumar. (2022). Persepsi Masyarakat Terhadap Filem dan Drama Bersiri Netflix Menggunakan Analisis Sentimen. Universiti Kebangsaan Malaysia.

Lum Choi Kian. (2020). Analisis Sentimen Dalam Bitcoin Tweets. Universiti Kebangsaan Malaysia.

Manasee Godsay. (2015). The Process of Sentiment Analysis. Int J Comp Appl.

Nations, Daniel. (Oct 2018). What Is Social Media?. Lifewire.

Neha Gupta. (2020). Application and techniques of opinion mining. Hybrid Computational Intelligence.

Nikit Periwal. (2021). Twitter Sentiment Analysis for Beginners. www.kaggle.com/code/stoicstatic/twitter-sentiment-analysis-for-beginners/notebook.

Nur Izaty Hana. (2021). Pengesanan Buli Siber Bahasa Melayu di Twitter Menggunakan Pembelajaran Mesin. Universiti Kebangsaan Malaysia.

Nuser Maryam, Alsukhni Emad, Saifan Ahmad, Khasawneh Rama & Ukkaz Dina. (2022). Sentiment analysis of COVID-19 vaccine with deep learning. J Theor Appl Inf Technol, 100(12), 4513-4521.

Rahmansyah, Arief. (Oct 2016). Natural Language Processing. wordpress.

Sentiment Analysis Using Naïve Bayes and Support Vector Machine. Universiti Islam 45 Bekasi.

Tan Jie Mie. (2019). Analisis Sentimen Menggunakan Teks Agresif Dalam Pengesanan Pembulian Siber. Universiti Kebangsaan Malaysia.

Tay Fui Kien. (2019). Analisis Sentimen Twitter Mengenai Peristiwa Penting Yang Berlaku di Sekitar UKM. Universiti Kebangsaan Malaysia.

Yin, H., Song, X., Yang, S. et al. Sentiment analysis and topic modeling for COVID-19 vaccine discussions. World Wide Web 25, 1067–1083 (2022)