

Mengesan Perasaan Melalui Teks: Kenali Asas Analisis Sentimen

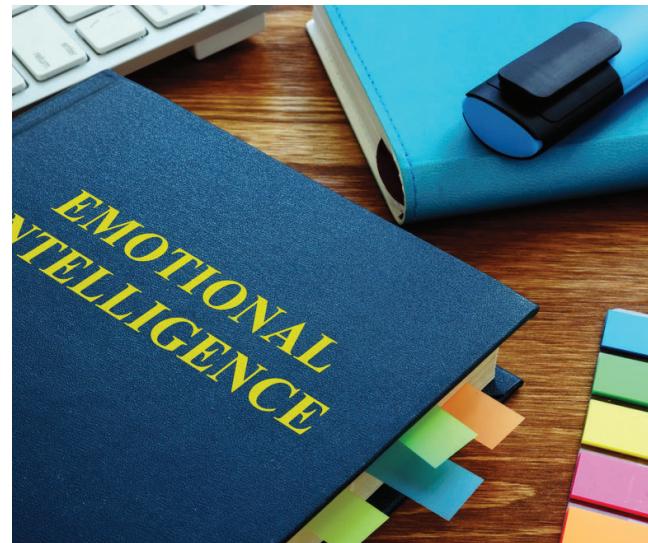
Rozianiwati Binti Yusof & Nor Azlina Aziz Fadzillah

Fakulti Sains Komputer dan Matematik, UiTM Kampus Seremban

Perkembangan teknologi pada masa kini memberi pendedahan kepada orang ramai untuk menggunakan media sosial dalam kehidupan seharian. Mereka menjadi keyboard warrior dan lebih berani tampil di alam maya berbanding alam yang nyata. Selain itu, mereka turut meluahkan dan memberi pendapat di media sosial dengan lebih telus dan tanpa memikirkan kesan ke atas persekitaran mereka. Media sosial pada era pandemik ini telah menjadi medan untuk berdebat dan menyebarkan sesuatu isu. Aplikasi seperti Facebook, Twitter dan Instagram merupakan antara media sosial yang agak popular di Malaysia. Menurut laman web Statista (2022), terdapat 2910 juta pengguna Facebook, 436 juta Twitter dan 1478 juta pengguna Instagram sehingga tahun 2022. Kesan pertambahan pengguna yang ramai bagi media sosial telah menghasilkan lonjakan data yang tidak terduga dan dipanggil sebagai data raya. Penghasilan data yang besar tanpa penggunaan yang maksimum adalah amat merugikan. Oleh itu, analisis sentimen amat penting bagi memastikan data yang terhasil dapat digunakan sepenuhnya.

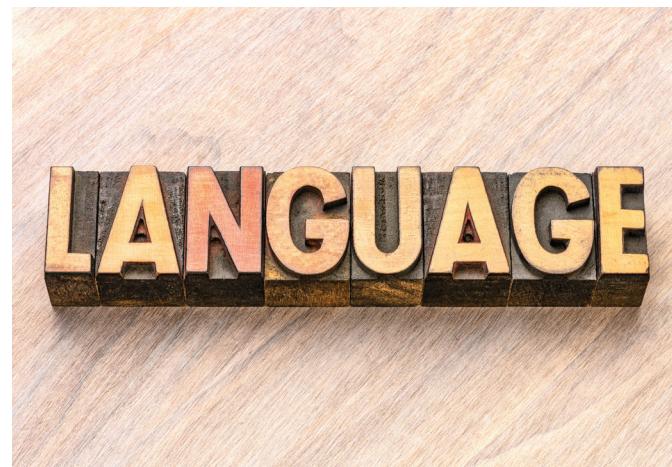
Analisis sentimen yang merupakan sebahagian daripada bidang *Natural Language Processing* (NLP) yang merupakan satu bidang yang agak popular selaras dengan perkembangan teknologi sekarang ini.

Bidang ini diperlukan untuk memproses data bagi mengenal pasti perasaan atau emosi seseorang yang dikeluarkan melalui bahasa atau perkataan samada melalui percakapan atau penulisan. Apabila seseorang berkomunikasi, jarang sekali mereka akan berterus terang. Sebaliknya mereka menganggap orang lain faham mengenai maksud yang ingin disampaikan. Namun begitu, terdapat juga kes yang melibatkan sarcasm atau bahasa yang digunakan untuk memperli seseorang. Maka, analisis sentimen merupakan satu bidang yang akan menganalisis maksud di sebalik perkataan atau percakapan yang dilontarkan. Walaupun manusia menggunakan pelbagai bahasa sebagai perantara komunikasi, tetapi setiap bahasa mempunyai sentimen yang boleh menentukan perasaan seseorang, samada ia bersifat positif, negatif atau neutral.



Bidang analisis sentimen ini agak meluas dan boleh dikaji melalui pelbagai perspektif, antaranya penggunaan bahasa yang berbeza seperti Bahasa Inggeris, Bahasa Melayu, Arab dan Mandarin. Selain itu juga, ia boleh memfokuskan terhadap percakapan dan juga penulisan. Artikel ini membincangkan berkenaan analisis sentimen dalam bentuk penulisan iaitu teks.

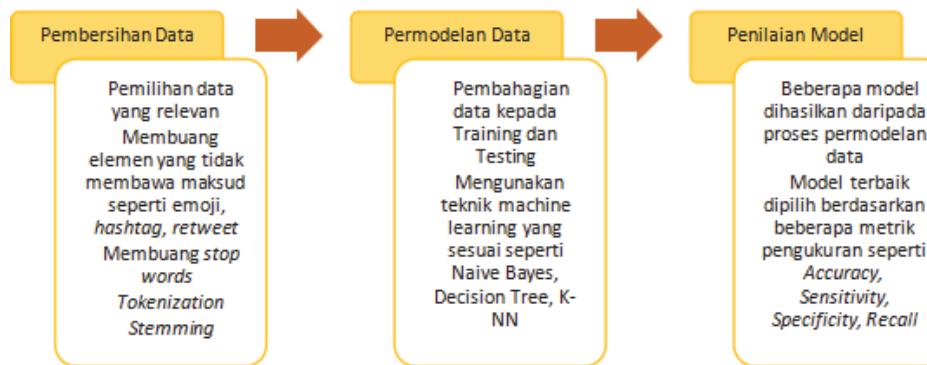
Pada era digital ini, penggunaan internet yang berleluasa terutamanya di media sosial, e-dagang dan pembelajaran atas talian memberi peluang kepada pengguna meluahkan pendapat mereka secara terus dalam bentuk teks. Sebagai contoh, mereka dapat berdebat mengenai sesuatu isu yang tular di media sosial, memberi testimoni terhadap perkhidmatan atau produk yang digunakan serta meluahkan pendapat di sebuah platform. Sebagai contoh, web e-dagang yang popular seperti Shopee menyediakan ruang kepada pengguna untuk memberi komen terhadap sesuatu produk. Pengguna akan memberi komen samada positif, negatif atau sekadar menyatakan pendapat yang neutral iaitu tidak menyokong atau mengkritik sesuatu produk. Walau bagaimanapun, komen yang diberi dapat menghasilkan sesuatu yang bermaklumat jika dianalisis dengan cara yang betul. Sebagai contohnya, daripada komen-komen yang diberi, penjual dapat mengenal pasti faktor yang menyebabkan pengguna berminat dengan produk atau mengkritik produk tersebut. Dengan in, pengguna mungkin dapat menghasilkan produk baharu atau memperbaiki produk sedia ada. Selain itu, penjual juga dapat menawarkan produk lain yang seakan-



akan dengan produk yang pernah dibeli sebelum ini dan mensasarkan jualan terhadap pengguna tertentu sahaja. Oleh itu, analisa sentimen adalah amat penting dalam menganalisis data bagi memberi keputusan kepada penjual terhadap produk yang sedia ada. Ini kerana, keupayaan manusia adalah amat terhad untuk membuat kesimpulan atau mencari corak (pattern) berdasarkan data-data yang banyak dan dihasilkan setiap hari.

Bagi menganalisis sentimen, terdapat beberapa cabaran yang perlu dihadapi. Antaranya ialah penggunaan bahasa dalam komunikasi. Setiap orang mempunyai gaya bahasa yang tersendiri dalam komunikasi samada melalui percakapan atau penulisan. Oleh itu, agak sukar untuk sesuatu mesin seperti komputer memproses dan mentafsir maksud komunikasi tersebut berbanding dengan manusia. Selain itu, setiap bahasa juga mempunyai sintak, semantik dan perkataan yang berbeza-beza. Maka, cara pengolahan analisis sentimen adalah berbeza mengikut jenis bahasa yang dianalisis.

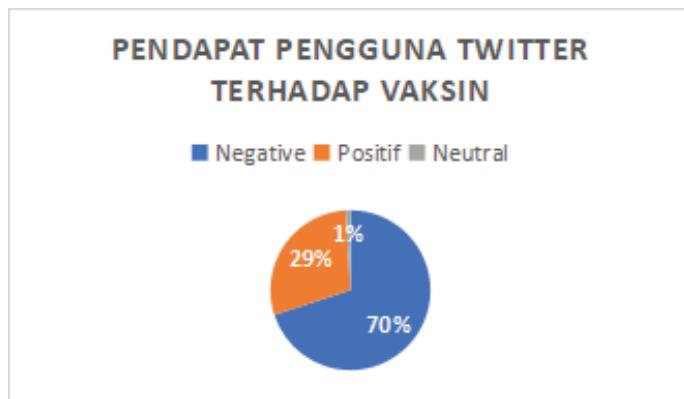
Walau bagaimanapun, terdapat pelbagai perisian yang dapat membantu dalam menjalankan analisis sentimen. Antaranya, penggunaan pustaka program yang terdapat dalam pengaturcaraan Python dan R, perisian Rapid Miner dan Weka. Di samping itu juga, terdapat beberapa langkah yang perlu dilakukan bagi menghasilkan suatu model analisis sentimen atau dikenali sebagai model pengelasan untuk mengenal pasti pendapat-pendapat pengguna melalui teks atau mesej. Sebagai contoh, analisis sentimen telah dilakukan menggunakan perisian Rapid Miner bagi mengesan pendapat orang ramai terhadap vaksin yang diterima oleh kerajaan Malaysia iaitu Astra Zeneca, Pfizer dan Sinovac. Sebanyak 2001 data dikumpul daripada aplikasi Twitter menggunakan Twitter API yang terdapat dalam Rapid Miner. Rajah 1 di bawah menunjukkan langkah-langkah penting dalam menghasilkan model analisis sentimen.



Rajah 1: Proses Penghasilan Model Analisis Sentimen

Tiga langkah utama iaitu pembersihan data, permodelan data dan penilaian model akan melibatkan penggunaan beberapa teknik yang telah disenaraikan di dalam perisian Rapid Miner. Langkah pertama adalah pembersihan data. Ia merupakan proses yang amat penting bagi memastikan model atau keputusan yang dihasilkan adalah tepat. Kebanyakan data-data yang diambil daripada media sosial perlu dibersihkan terlebih dahulu. Ini kerana penggunaan ayat di media sosial adalah tidak mengikut tatabahasa dan ejaan yang betul. Terdapat juga perkataan yang perlu dibuang kerana bertindak sebagai penguat ayat sahaja. Emoji-emoji di dalam ayat juga mengganggu proses permodelan data, yang mana ia perlu dibuang atau ditafsirkan jika memberi sesuatu makna di dalam ayat. Walau bagaimanapun, kaedah pembersihan data ini bergantung kepada jenis bahasa yang digunakan.

Manakala langkah kedua iaitu permodelan data dihasilkan dengan menggunakan teknik-teknik daripada bidang machine learning seperti decision tree, naïve bayes, neural network dan sebagainya. Teknik-teknik yang digunakan bertujuan supaya komputer atau mesin dapat mempelajari corak atau kesimpulan yang dihasilkan daripada data yang dikumpul. Corak yang dihasilkan akan digunakan untuk mengenal pasti jenis pendapat yang diberikan oleh pengguna. Contohnya, berdasarkan mesej atau perbincangan daripada Twitter terhadap vaksin-vaksin, model yang dihasilkan mendapati ramai yang memberi reaksi negatif berbanding positif terhadap vaksin yang digunakan (Rajah 2). Reaksi atau pendapat ini dikesan melalui perkataan-perkataan yang digunakan di dalam teks. Corak ayat-ayat positif, negatif atau neutral dikenalpasti dengan menggunakan teknik daripada machine learning.



Rajah 2: Pendapat pengguna-pengguna Twitter terhadap vaksin

Pengukuran juga dapat dilakukan mengikut kategori vaksin di mana vaksin berjenis AstraZeneca mendapat banyak reaksi negatif. Manakala vaksin Pfizer mendapat banyak reaksi positif dan ramai yang sekadar memberi fakta terhadap vaksin Sinovac. Ketepatan suatu model yang dihasilkan dapat diukur menggunakan ukuran Accuracy, Sensitivity, Specificity dan Recall. Ini adalah kerana, ketepatan suatu model yang dihasilkan adalah amat penting bagi memastikan keputusan yang diperoleh adalah tepat. Berdasarkan contoh yang dibincangkan, model yang dihasilkan ini dapat meramal reaksi pengguna terhadap vaksin dengan menggunakan teks atau mesej dengan peratus ketepatan mencapai 81.78%. Model yang dihasilkan ini boleh dikatakan agak tepat. Walau bagaimana pun, nilai ketepatan sesuatu model juga bergantung kepada domain data yang digunakan.

Oleh itu, data-data yang dihasilkan daripada penggunaan teknologi yang semakin canggih haruslah digunakan secara maksimum. Ini adalah kerana sejumlah data yang besar dapat memberi impak yang ketara jika dianalisis dengan cara yang betul. Penghasilan data raya terutamanya teks-teks daripada media sosial dapat membantu mengembangkan lagi sektor-sektor yang lain dengan penghasilan model analisis sentimen yang dapat membantu membuat sesuatu keputusan.

Rujukan

Laman Web Statistica. (April 24, 2022).<https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>.