ANALISIS SENTIMEN BERKENAAN PANDEMIK COVID-19 MALAYSIA PADA MEDIA SOSIAL TWITTER

Looi Hao Shan
Mohd. Juzaiddin Ab Aziz

ABSTRAK

Objektif utama projek ini adalah untuk mengkaji dan menilai post Twitter masyarakat Malaysia semasa perintah kawalan pergerakan pandemik secara analisis sentimen. Pendapat dan sentimen mampu mempengaruhi kecenderungan sosial serta boleh menimbulkan masalah di kalangan masyarakat. Laman sosial Twitter mengandungi koleksi teks yang besar terkumpul daripada pelbagai pengguna sebagai tempat perluahan pendapat, ini menjadikan kumpulan data tersebut salah satu sumber pengumpulan pendapat yang bernilai kepada sesuatu topik. Analisis tweets yang dihantar oleh masyarakat Malaysia semasa perintah kawalan pergerakan pandemik Covid-19 mampu membawa faedah kerana fikiran masyarakat sentiasa berubah. Hasil analisis sentimen seterusnya dijana sebagai model ramalan mampu berguna kepada syarikat yang berminat atau pihak berkenaan untuk mendapatkan peramalan maklumat atau info sentimen daripada sosial media. Pengumpulan data dalam kajian ini akan menggunakan hashtag atau kata kunci dengan bantuan API Twitter dan perpustakaan bahasa Python untuk proses tersebut. Projek ini akan memproses tweets yang menggunakan bahasa Inggeris sahaja. Selepas proses pengumpulan dan pembersihan data bagi menghilangkan kebisingan data dari ruangan kosong, URL serta normalisasi, analisis sentimen akan diimplementasikan. Parameter seperti polariti, subjektiviti, positif, negatif serta parameter neutral diguna ke dalam analisis kajian ini. Algoritma pembelajaran mesin yang digunakan dalam model adalah Naives Bayes, *logistic regression* dan *Support Vector Machine* (SVM) bagi perbandingan dan memilih model paling berkesan.

1 PENGENALAN

Kehidupan harian masyarakat semakin terikat dengan ciptaan-ciptaan teknologi dan Internet. Dengan pembangunan teknologi Web 2.0 pula, terbentuknya penggunaan Internet yang mengalami perubahan besar terutamanya media sosial di mana masyarakat dapat memuatnaik media dan revolusi Web 2.0 telah menyediakan platform untuk perkongsian media sesama penulis dan pembaca. (Siti Ezaleila & Azizah, 2010).

Virus SARS-CoV-2 (COVID-19) telah mula merebak di negara Malaysia pada 25 Januari 2020 dan Malaysia memasuki fasa Perintah Kawalan Pergerakan (PKP) pada 18 Mac 2020 di seluruh negara. Pandemik ini telah membawa banyak perubahan emosi dan mental

masyarakat. Perkataan kunci seperti ‘COVID-19’, ‘pandemic’ menjadi topik hangat di media sosial, terutamanya pada media sosial Twitter. Twitter mempunyai kapasiti pengguna yang besar, mencecah 300 juta pengguna aktif seluruh dunia pada tahun 2019 (Statista Research Department, 2021).

Analisis sentimen ialah satu proses untuk memahami dan menganalisa emosi dalam teks dengan penggunaan teknik-teknik analisis data. Analisis sentimen juga digunakan sebagai suatu cara untuk mempelajari ulasan pelanggan. Kajian ini membincangkan analisis sentimen mengenai emosi masyarakat di tweets sewaktu pandemik COVID-19 dengan tweets apabila pandemik COVID-19 menjadi norma baharu. Maklumat ini penting bagi membantu pihak berkuasa mengendalikan tempoh perintah kawalan pergerakan dan Prosedur Operasi Standard (SOP) semasa pandemik.

2 PENYATAAN MASALAH

Pada masa kini, virus COVID-19 telah menjadi sebahagian daripada komuniti masyarakat. Rakyat terpaksa untuk mengatasi pandemik ini dan meneruskan aktiviti harian mereka. Namun, emosi dan mental selepas mengalami pandemik sudah tentu membawa serba sedikit perubahan kepada sikap dan tingkah laku masyarakat. Kajian dan analisasi perlu dilakukan untuk mengetahui fikiran masyarakat terhadap perintah kawalan pergerakan pandemik supaya pihak berkenaan mampu mengeluarkan tindakan lanjut yang lebih rasional dan berkesan pada masa hadapan.

3 OBJEKTIF KAJIAN

Projek ini bermatlamat untuk melakasanakan analisis dan mengenal pasti keadaan emosi masyarakat mengenai perintah kawalan pergerakan (PKP) COVID-19. Secara umum konsep emosi di sini dibahagi kepada positif, neutral dan juga negatif. Kajian ini serta bermatlamat untuk membangunkan model bagi pengelasan sentimen terhadap PKP pandemik COVID-19 pada media sosial dengan penggunaan pembelajaran mesin.

Kajian ini juga bertujuan untuk menyediakan sebuah analisis dalam format yang berguna, boleh diterima dan digunakan oleh sesebuah organisasi atau orang yang memerlukan analisis data ini.

4 METOD KAJIAN

Pembangunan model yang bersesuaian diperlukan supaya kajian dijalankan dengan lancar dan mendapatkan hasil keputusan yang berkualiti. Model ini digunakan untuk menvisualisasikan tahap nilai sentimen masyarakat terhadap perintah kawalan pergerakan sepanjang PKP pertama. Fasa pembangunan termasuk fasa pengumpulan data, pra-pemprosesan, pengekstrakan ciri, analisis, pemodelan dan juga penilaian.

4.1 Fasa Perancangan

Fasa ini terdiri daripada proses pengenalpastian kekangan kajian, penentuan skop, menganalisis sorotan kajian kesusasteraan dan juga pengumpulan set data yang diperlukan untuk kajian ini. Antara data yang dikutip adalah ditetapkan kata kunci tertentu bagi memenuhui skop kajian set data. Spesifikasi keperluan juga dikenal pasti terlebih dahulu pada fasa ini bagi mengurangkan kesilapan semasa proses kajian. Antara spesifikasi keperluan perisian yang diguna dalam kajian ini adalah seperti di Jadual 1 berikut.

Perisian	**Penerangan**
Windows 10(64-bit)	Sistem pengoperasian yang dipakai dalam kajian ini
Google Chrome	Pelayar web
Google Colaboratory	Untuk menjalankan dan menghasilkan analisis sentimen
Microsoft Excel 365	Penyimpanan data dan maklumat dari Twitter dalam format file .csv
Python	Bahasa pemograman yang diguna untuk seluruh projek

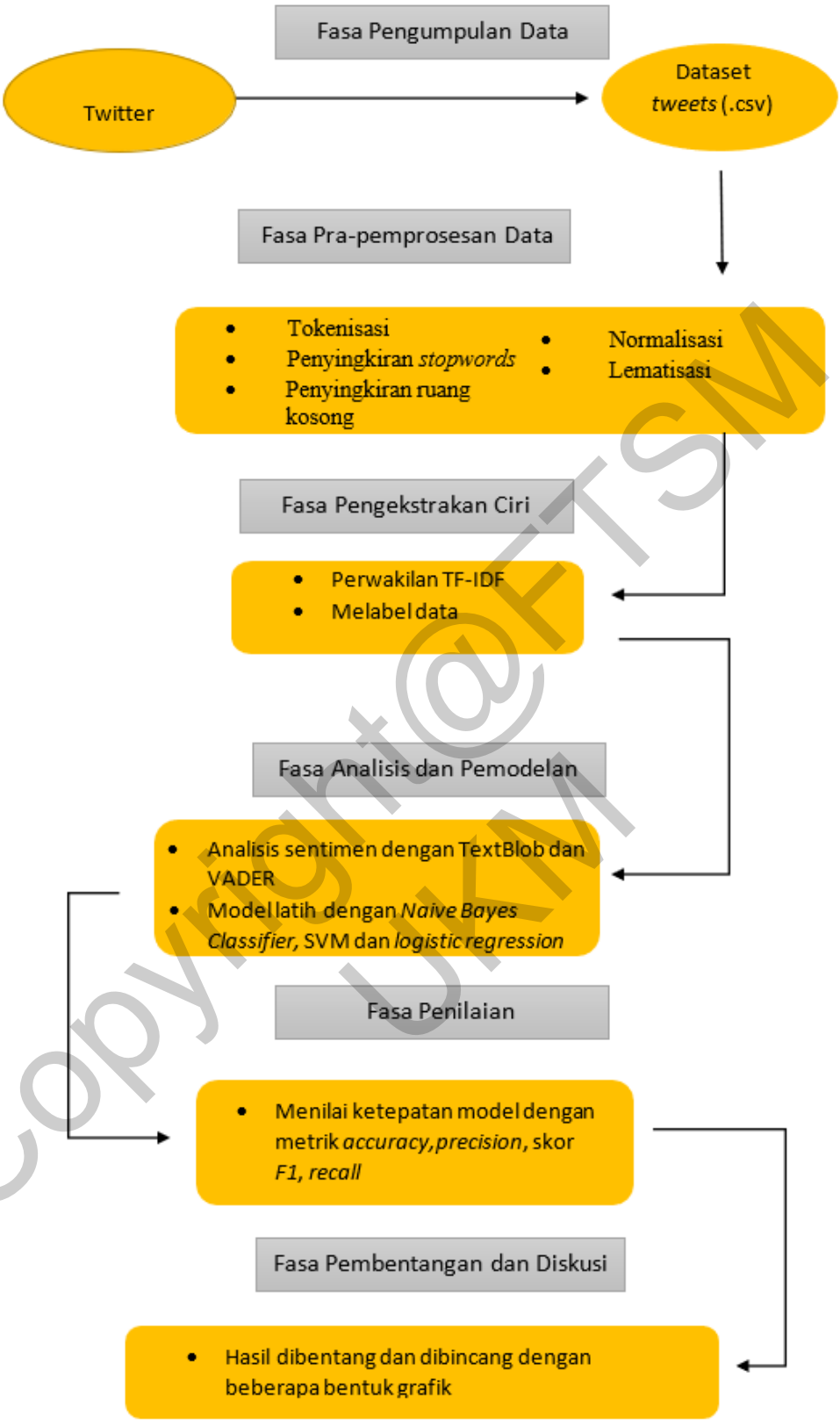
Jadual 1 Spesifikasi Keperluan Perisisan

4.2 Fasa Analisis

Fasa analisis menjalankan tafsiran terhadap data yang dikumpul pada fasa perancangan. Tafsiran terhadap data dijalankan supaya boleh memastikan data yang dikumpul memenuhi syarat dan relevan kepada topik kajian ini. Set data yang telah dikumpul diselaraskan kepada satu jenis bahasa sahaja iaitu Bahasa Inggeris. Selepas itu, data akan ditapis dan dibersihkan dengan proses normalisasi, pra-pemprosesan dan pengekstrakan ciri bagi menyediakan set data bagi fasa-fasa seterusnya. Sebanyak 11406 data telah dikumpul daripada media sosial Twitter. Set data dilabel sentimen dengan menggunakan dua perpustakaan dari NLTK iaitu VADER dan juga TextBlob sama ada positif, neutral atau negatif. Fasa analisis data dijalankan supaya set data yang dikumpul dibersih, dan tidak mempunyai data yang hilang atau berulangan.

4.3 Fasa Reka Bentuk

Dalam fasa ini, gambaran awal bagi pembangunan sistem kajian secara keseluruhan dapat diimejkan. Melalui fasa ini, penghasilan sistem menjadi sistematik dengan bantuan gambaran pembangunan sistem yang jelas. Menurut Rajah 1, reka bentuk bagi kajian ini telah terbahagi kepada enam fasa, iaitu pengumpulan data, pra-pemprosesan data, pengekstrakan ciri, analisis dan pemodelan, penilaian dan juga fasa pembentangan hasil.



Rajah 1 Reka Bentuk Seni Bina Model

i. Pengumpulan data

Dalam fasa ini, pengumpulan data dilakukan dengan menggunakan API Twitter dan juga perpustakaan Snscrape yang mampu mengumpul tweets yang memenuhi kriteria masa pandemik dan juga berkenaan dengan topik COVID-19 daripada Twitter. Untuk mendapatkan keizinan penggunaan API Twitter, permohonan akaun perlu dibuat di laman web rasmi Twitter. Kata kunci yang diguna adalah 'MCO', 'covid19 Malaysia', 'PKP', dan '*Movement Control Order*'. Tarikh pengumpulan data ini adalah pada tempoh perintah kawalan pergerakan pertama, iaitu dari 18 Mac 2020 sehingga 14 April 2020. Selain itu, parameter digunakan bagi mengumpul data berbahasa Inggeris sahaja. Sebanyak 11406 data berjaya dikumpul dan disimpan sebagai fail berformat csv.

ii. Pra-pemprosesan data

Sebagai proses pertama selepas pengumpulan data, fasa pra-pemprosesan data membuat pembersihan data dan menyediakan data untuk menjalankan analisis dan proses selanjutnya. Antaranya pemprosesan normalisasi yang dibuat adalah penukaran perkataan kepada huruf kecil, penyingkiran Pelokasi Sumber Seragam (URLs), penyingkiran penyebut (*mentions*) dan hashtags, penyingkiran ruang putih bagi membersihkan data. Selain itu, penyingkiran perkataan yang bukan *alphanumeric*, penyingkiran emoji, pembuangan baris yang bernilai *null* turut diproses dalam set data. Kolum yang tidak berguna bagi proses seterusnya seperti Unnamed:0, Datetime, Tweet Id, Username dibuang daripada dataframe. Penyingkiran kata henti (stopwords) dan lematisasi terhadap data turut diaplikasikan. Kata henti disingkirkan supaya tidak mengacau analisis kerana kata-kata henti tidak mempunyai emosi dan jarang berguna dalam analisis sentimen. Lematisasi pula bertujuan untuk menukarkan perkataan kembali kepada perkataan asal atau kata dasar bagi meningkatkan prestasi analisis. Daripada 11406 baris data, 10620 baris data tinggal setelah melalui fasa pra-pemprosesan.

iii. Pengekstrakan ciri

Fasa pengekstrakan ciri dijalankan supaya ciri penting dapat diperoleh dalam data. Dalam fasa ini, *Term Frequency-Inverse Document Frequency* (TF-IDF) diaplikasi. TF-IDF merupakan penilaian statistik yang sering diguna untuk melihat tahap kerelevanan sesuatu perkataan dalam dokumen di kalangan kumpulan dokumen. Ia dikira menggunakan hasil darab frekuensi istilah dengan frekuensi dokumen songsang. Frekuensi istilah (TF) ialah kekerapan perkataan muncul pada dokumen manakala frekuensi dokumen songsang (IDF) merupakan kekerapan songsang perkataan itu muncul dalam sekumpulan dokumen.

iv. Analisis dan Pemodelan

Proses menganalisis sentimen terhadap set data dilakukan dengan penggunaan perpustakaan NLTK, iaitu TextBlob dan juga VADER. Dalam kajian ini, tiga label diguna untuk mengenal pasti tahap sentimen dalam data, iaitu label positif, neutral dan juga negatif. Bagi TextBlob, kekutuban atau dikenali sebagai polariti, diaplikasi dalam fasa ini bagi mengetahui tahap positif atau negatif sesebuah tweets dalam set data manakala VADER merupakan sejenis analisis sentimen yang sensitif dan berasaskan leksikon untuk mengesan kekutuban (polariti) dan intensiti (kekuatan) emosi di dalam teks data. Set data akan disimpan secara berasingan bagi set label sentimen TextBlob dan VADER. Setelah mengaplikasikan TextBlob ke atas set data, boleh mendapati bahawa data yang positif mempunyai sebanyak 4405, dan neutral sejumlah 4394 manakala 1822 yang selebihnya adalah data negatif. Dengan analisis sentimen VADER pula, hasil menunjukkan 5075 data *tweets* yang positif, 2147 data yang negatif dan 3399 data *tweets* neutral.

Semasa pemodelan, tiga algoritma akan diguna. Antaranya *Naïve Bayes, Logistic Regression*, dan *Support Vector Machine* (SVM). Ketiga-tiga model pengelasan ini digunakan bersama set data TextBlob dan set data VADER. Bagi melatih model, set data dibahagikan kepada set latihan dan set ujian mengikut nisbah 8:2. Setelah tamatnya latihan, model menjadi bersedia untuk melalui ujian peramalan hasil dengan menggunakan set data ujian.

v. Penilaian

Skor ketepatan dan prestasi setiap model dihitung dengan memanggil fungsi .score() dan juga menggunakan classification_report(). Fungsi .score() digunakan di dalam projek ini untuk mendapatkan nilai ketepatan bagi setiap model di mana nilai ketepatan yang lebih rendah bermakna kesilapan lebih kerap berlaku pada model. Bagi fungsi classification_report() pula, hasil peramalan model dapat diperhati dan dianalisis dari segi *accuracy, precision, F1-score,* dan juga recall.

4.4 Fasa Pengujian

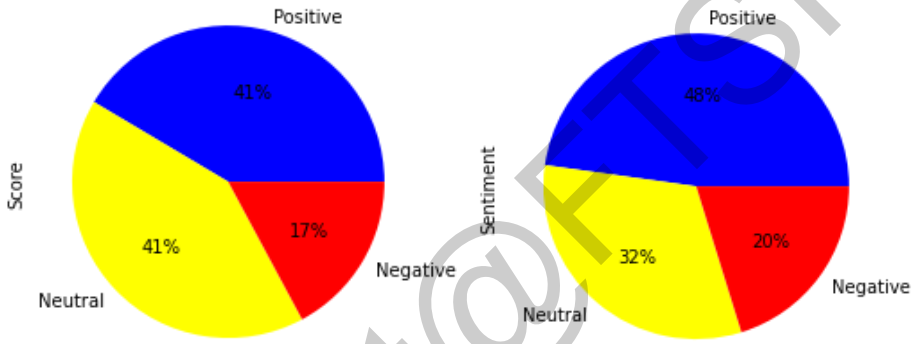
Fasa ini menguji prestasi hasil model pengelasan yang dibangunkan dengan set data kajian ini. Dua perpustakaan yang telah dipakai di dalam projek adalah TextBlob dan VADER. Kedua-dua pakej ini digunakan untuk mengenal pasti sentimen data *tweets* kepada kategori positif, negatif ataupun neutral. Dua pakej perpustakaan ini diguna supaya set data yang banyak ini tidak membebankan daripada proses pelabelan sentimen data *tweets* secara manual. Selepas menggunakan TextBlob pada set data, boleh mendapati bahawa data yang positif merangkumi 41% manakala analisis sentimen VADER pula memberi hasil 48% data *tweets* yang positif,. Pengujian prestasi model dilakukan dengan mendapatkan ketepatan model pengelasan dengan set ujian dan juga menggunakan fungsi classification_report. Di dalam classification_report(), *accuracy* merupakan satu cara pengukuran prestasi yang paling terus dan mudah. Ia dihitung dengan menggunakan formula di bawah, di mana TP mewakili positif sebenar, TF mewakili negatif sebenar, dan FP ialah positif palsu, FN ialah negatif palsu.

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

Untuk *precision,* ia merupakan satu pengiraan untuk peramalan positif yang betul berbandingkan kepada kesemua jumlah peramalan positif yang dibuat. *Precision* yang tinggi membawa maksud sesebuah model meramal label positif secara lebih tepat dengan kadar label positif palsu yang rendah. *Recall* pula merupakan pengiraan bagi peramalan positif yang benar berbandingkan kepada jumlah peramalan positif yang benar dan peramalan negatif yang silap atau palsu. Seterusnya, skor F1 yang terdapat di dalam classification_report() ialah hasil produk *precision* dan *recall* dibahagi dengan jumlah *precision* dan *recall*, didarab dengan 2. Skor F1 mempunyai nilai paling tinggi 1 dan paling rendah bagi skor F1 ialah 0.0.

5 HASIL KAJIAN

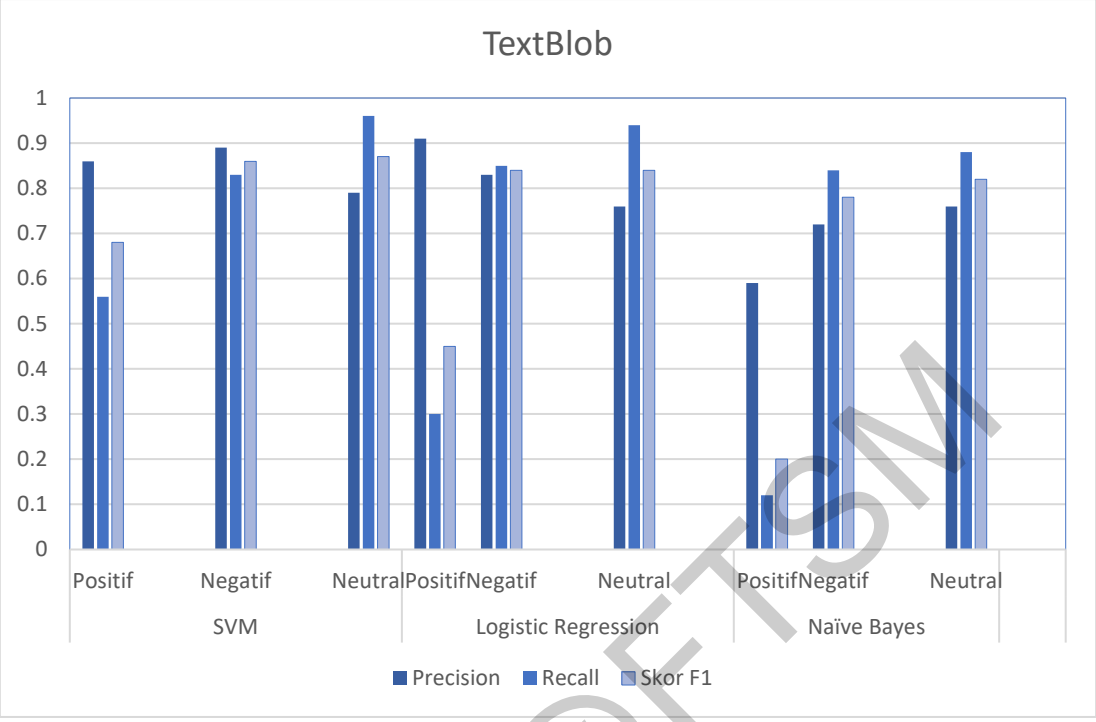
Hasil prestasi bagi setiap model divisualisasi menggunakan bentuk graf supaya pengguna mampu menilai dan membandingkan setiap prestasi dengan lebih mudah. Selepas menggunakan TextBlob pada set data, boleh mendapati bahawa data yang positif merangkumi 41% , dan neutral sejumlah 41% manakala 17% adalah data negatif. Analisis sentimen VADER pula memberi hasil 48% data *tweets* yang positif, 20% data yang negatif dan 32% data *tweets* neutral. Jadual 2 menunjukkan hasil pelabelan sentimen daripada TextBlob dan VADER.



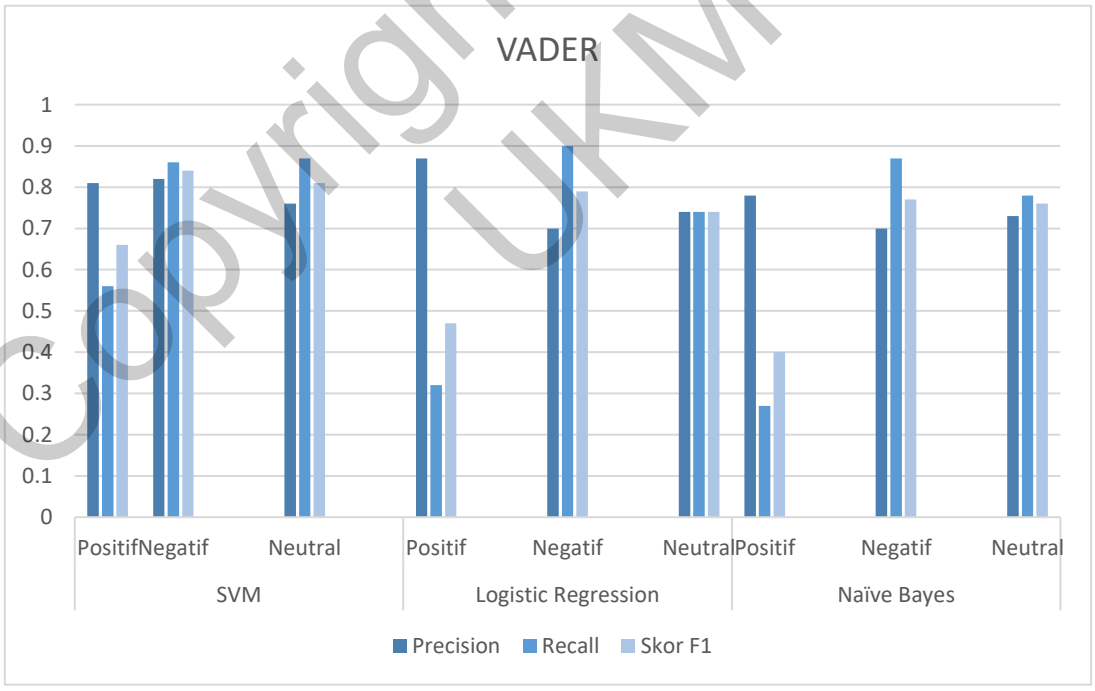
TextBlob VADER

Jadual 2 Carta Pai hasil sentimen TextBlob dan VADER

Bagi penilaian hasil prestasi model pula, model *logistic regression* dengan data sentimen TextBlob membawa *precision* label positif yang paling tinggi di antara semua model dengan nilai 0.91, manakala *precision* label negatif dan juga neutral yang paling tinggi ialah model SVM dengan data TextBlob dengan nilai 0.89 dan 0.79. Untuk *recall* pula, hasil nilai paling tinggi label positif adalah daripada model SVM bagi kedua-dua data VADER dan TextBlob, 0.56, *recall* label negatif paling tinggi ialah model *logistic regression* (VADER) dengan nilai 0.90 dan SVM (TextBlob) mempunyai keputusan *recall* label neutral yang tertinggi, 0.96. Bagi skor F1, model Naïve Bayes masih tidak menunjukkan prestasi yang lebih baik daripada model SVM dan *logistic regression*, skor F1 paling tinggi untuk kesemua label positif, negatif dan neutral ialah model SVM (TextBlob). Menurut Jadual 5.5, model SVM menunjukkan prestasi keseluruhan yang lebih baik berbanding model lain, terutama model SVM yang dilatih dengan set data TextBlob. Jadual 3 merupakan keputusan prestasi model peramalan SVM, *Logistic Regression*, dan Naïve Bayes dengan fungsi classification_report() bagi set data TextBlob dan Jadual 4 untuk set data VADER.

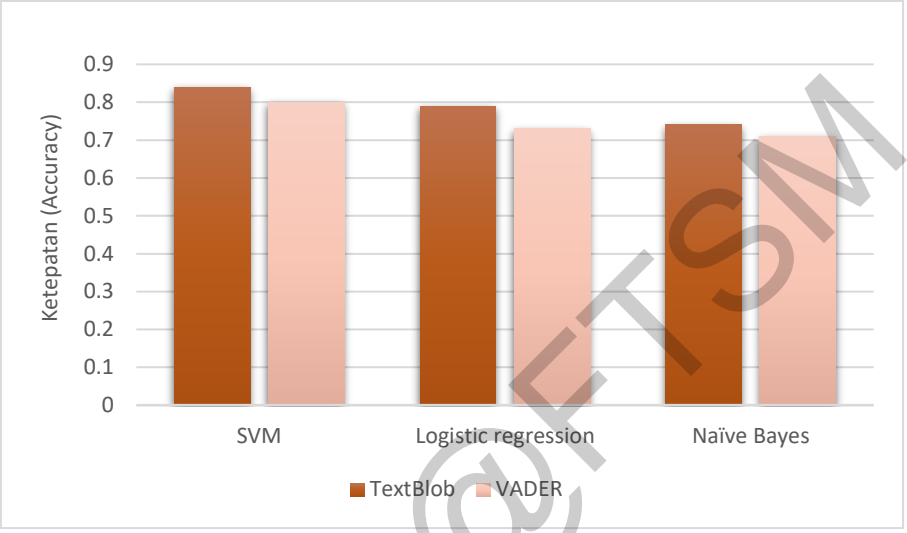


Jadual 3 Keputusan prestasi *precision*, *recall* dan skor F1 bagi data TextBlob pada model SVM, *Logistic Regression*, dan Naïve Bayes.



Jadual 4 Keputusan prestasi *precision*, *recall* dan skor F1 bagi data VADER pada model SVM, *Logistic Regression*, dan Naïve Bayes.

Ketepatan memainkan peranan penting dalam perbandingan keputusan model kerana ketepatan bermaksud jumlah data yang berjaya diramalkan berdasarkan kesemua data. Jadual 5.6 ialah keputusan ketepatan daripada semua model. Jadual 5 merupakan perbandingan ketepatan antara model menggunakan carta bar.



Jadual 5 Carta bar perbandingan ketepatan antara model.

6 KESIMPULAN

Secara kesimpulan, kajian analisis sentimen berkenaan pandemik COVID-19 ini menggunakan algoritma pembelajaran mesin bagi mengelas dan pembangunan model. Beberapa kekangan yang dihadapi semasa proses kajian ialah kebisingan data dengan pencampuran bahasa. Hal ini tidak dapat dielakkan walaupun mempunyai fungsi *filter* semasa pengumpulan data kerana *tweets* masyarakat Malaysia suka mencampurkan serba sedikit bahasa Melayu atau bahasa-bahasa lain semasa menghantar *tweets.* Selain itu, *tweets* yang menggunakan singkatan dan juga cara menyindir meningkatkan kesukaran proses analisis kepada pembelajaran mesin.

Fasa pra pemprosesan, analisis sentimen terdapat peranan masing-masing yang penting dalam pembangunan model pengelas dan pengujian model analisis sentimen ini. Proses mengenal pasti segmen penting diperlukan supaya memastikan kelancaran pembangunan model. Proses pengujian telah menyediakan satu fasa supaya keberkesanan model pengelas

dapat diuji dan dibandingkan dari beberapa segi bagi mendapatkan model yang lebih sesuai untuk data. Daripada keputusan pengujian, model yang lebih tepat dan sesuai dapat dipilih dan dikenal pasti. Bagi kajian set data projek ini, model SVM (TextBlob) menunjukkan prestasi yang paling cemerlang dengan ketepatan, skor F1, *recall* dan *precision* yang lebih stabil dan tinggi.

7 RUJUKAN

Bakhtazad, A., Garmabi, B., & Joghataei, T. M. 2021. Neurological manifestations of coronavirus infections, before and after COVID-19: a review of animal studies. *Nature Public Health Emergency Collection*, 1-21.

Datagy. 2022. *Support Vector Machines (SVM) in Python with Sklearn.* Didapatkan dari datagy: https://datagy.io/python-support-vector-machines/ [25 Februari 2022]

Dubey, A. D. 2020. *Twitter Sentiment Analysis during COVID-19 Outbreak.* Lucknow: Jaipuria Institute of Management.

Kohli, S. 2019. *Understanding a Classification Report For Your Machine Learning Model.* Didapatkan dari Medium: https://medium.com/@kohlishivam5522/understanding-a-classification-report-for-your-machine-learning-model-88815e2ce397#:~:text=Precision%20is%20the%20ability%20of,%3D%20TP%2F(TP%20%2B%20FP) [18 November 2019]

Liu, B. 2012. *Sentiment Analysis and Opinion Mining.* Morgan & Claypool Publishers.

M., T., Suresh, P., & G., N. 2017. Sentiment Analysis on Twitter Using Streaming API. *IEEE 7th International Advance Computing Conference (IACC)*, (hlm. 915-919).

Marrese-Taylor, E., Velásquez , J., Bravo-Marquez, F., & Matsuo, Y. 2013. Identifying Customer Preferences About Tourism Products Using An Aspect-Based Opinion Mining Approach. *Procedia Comput Sci 22*, 182-191.

Pandey, P. 2019. *Data Preprocessing: Concepts.* Didapatkan dari Towards Data Science: https://towardsdatascience.com/data-preprocessing-concepts-fa946d11c825 [25 November 2019]

Ray, S. 2017,. *Learn Naive Bayes Algorithm.* Didapatkan dari Analytics Vidhya: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/ [11 September 2017]

Siti Ezaleila, M., & Azizah, H. 2010. Media Sosial: Tinjauan Terhadap Laman Jaringan Sosial Dalam Talian Tempatan. *Jurnal Pengajian Media Malaysia Jilid 12*, 37-52.

Statista Research Department. 2021. *Twitter: number of monthly active users 2010-2019.* Didapatkan dari Statista: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/#main-content [27 Januari 2021]

Stephen Wai Hang Kwok, S. K. 2021. *Tweet Topics and Sentiments Relating to COVID-19 Vaccination Among Australian Twitter Users: Machine Learning Analysis.* J Med Internet Res.

Velavan, T. P., & Meyer, G. C. 2020. The COVID-19 epidemic. *Wiley Public Health Emergency Collection*, 278-280.

Weller, K., Bruns, A., Burgess, J., Mahrt, M., & Puschmann, C. 2013. Twitter and Society: An Introduction. *Twitter and Society*, 29-37.