PENERAPAN MACHINE LEARNING TERHADAP ANALISIS OCEAN MENGENAI OPINI MASYARAKAT MENGENAI PPKM DI MEDIA SOSIAL

3.1.1 Pendahuluan

Pandemi COVID-19 di Indonesia masih belum teratasi secara keseluruhan. Upaya mengurangi jumlah penyebaran sudah dilakukan oleh pemerintah dalam mengurangi penyebaran COVID-19. Meningkatnya jumlah penyebaran di awal tahun 2021 disebabkan karena banyak terjadi pelanggaran protokol kesehatan. Untuk menurunkan penularan virus, pemerintah melakukan pembatasan kegiatan masyarakat yang disebut dengan aturan PPKM (Pemberlakuan Pembatasan Kegiatan Masyarakat).

PPKM berdasar pada Instruksi Menteri Dalam Negeri Nomor 1 dan 2 tahun 2021 tentang PPKM berskala mikro sebagai upaya mengurangi penyebaran COVID-19. Peraturan ini diberlakukan khususnya dalam lingkup Pulau Jawa dan Bali. Penerapan aturan ini ditentukan oleh pemerintah daerah menyesuaikan kondisi dan urgensi.

Namun penerapan PPKM menuai berbagai respons dari masyarakat umum. Penerapan PPKM yang berkepanjangan dan memiliki banyak level menggiring opini publik kearah yang tidak pasti. Dengan maraknya penggunaan media sosial masyarakat lebih mudah untuk menyuarakan pikirannya di internet. Berdasarkan data yang didapatkan dari 3 *platform* media social yaitu *twitter, instagram,* dan *youtube* terdapat 2.506.835 mention untuk topik PPKM pada periode Juli – Oktober 2021. Hal ini menunjukkan derasnya opini masyarakat yang mengalir terhadap PPKM di media social.

OCEAN adalah metode untuk mengukur sifat atau kepribadian sesorang. OCEAN dapat dijabarkan menjadi *openness* (O) yaitu keterbukaan pada pengalaman, *conscientiousness* (C) yaitu berhati-hati, *extraversion* (E) yaitu ekstraversi, *agreeableness* (A) yaitu keramahan, dan *neuroticism* (N).

Analisis OCEAN terhadap opini masyarakat terkait PPKM diperlukan untuk mengelompokan sifat atau kepribadian dari setiap orang yang mengutarakan pendapatnya ke media sosial. Analisis ini menggunakan metode *machine learning* untuk membuat model sehingga model tersebut bisa digunakan sebagai alat pendeteksi OCEAN secara otomatis untuk efesiensi waktu, tenaga, dan biaya yang juga akurat.

Berdasarkan latar belakang yang telah dipaparkan, maka tujuan dalam penelitian ini adalah untuk mengklasifikasikan opini masyarakat mengenai PPKM terhadap 5 kelas utama didalam OCEAN dan membuat model yang dapat mengklasifikasikan OCEAN secara akurat. Selain itu diharapkan hasil penelitian ini dapat menjadi bahan pertimbangan dan gambaran bagi brand atau instansi dalam mengambil kebijakan dan pemanfaatan media social.

3.1.2 Metodologi Penelitian

3.1.2.1 Data

Data yang digunakan dalam penyusunan laporan ini adalah data sekunder yang berasal dari aplikasi Ripple10 (new.ripple10.com). Data yang digunakan merupakan data jumlah mention pada isu nasional terkait PPKM, dimulai pada periode 1 Juli 2021 sampai 30 Oktober 2021. Periode tersebut memiliki 2.506.835 data mention harian.

Data memiliki 6 jenis label yaitu *Openness*, *Extraversion*, *Agreeableness*, *Neuroticism*, *Conscientiousness*, dan *Escape*. Setiap kelas mengandung satu karakteristik dalam analisis OCEAN kecuali *escape*. *Escape* digunakan sebagai kelas untuk data text yang tidak mengandung unsur OCEAN di dalamnya.

Data kemudian direduksi dengan cara disample untuk efesiensi waktu, tenaga, dan biaya saat proses pelabelan secara manual. Data disample sebanyak 800 amatan secara random. Data lalu dibagi kedalam dua bagian yaitu data latih untuk pemodelan dan data test untuk validasi model. Data latih terdiri dari 640 amatan (80%), sedangkan data validasi sebanyak 160 amatan (20%).

3.1.2.2 OCEAN

OCEAN adalah metode untuk mengukur sifat kepribadian dengan cara penerapan model lima besar sifat kepribadian (*big five personality traits model*). Model lima besar sifat kepribadian itu sendiri terdiri dari *openness* yaitu keterbukaan pada pengalaman, *conscientiousness* yaitu berhatihati, *extraversion* yaitu ekstraversi, *agreeableness* yaitu keramahan, dan *neuroticism* yaitu neurotisme, yang disingkat menjadi OCEAN.

Openness adalah sifat dimana seseorang menunjukkan sifat kreativitas, keingintahuan, kecerdasan, dan mampu menerima perubahan, menunjukkan nilai positif terhadap lingkungan, lebih suka memperluas wawasan mereka secara mandiri, dan cenderung memiliki sikap positif terhadap pembelajaran. *Extraversion* adalah sifat yang menunjukkan seseorang mampu bersosialisasi dan berinteraksi pada lingkungannya, tegas dalam mengambil keputusan, memiliki emosi positif, serta antusias terhadap dunia sosial dan material. *Agreeableness* adalah seseorang dengan sifat yang adil, toleransi, fleksibilitas, tingkat interaksi interpersonal yang tinggi, tidak mementingkan diri sendiri, dan berusaha untuk menangani konflik secara kooperatif dan kolaboratif. *Neuroticism* adalah sifat di mana tingkat emosional seseorang tidak stabil, mudah cemas, memiliki motivasi kerja yang buruk, termasuk melakukan penetapan tujuan, karena ketidakstabilan emosi dan kerentanan terhadap stress dan kecemasan yang ada dalam dirinya. *Conscientiousness* merupakan sifat keteraturan, ketekunan, kedisiplinan, orientasi pada pencapaian dan tanggung jawab

3.1.2.3 Text Mining

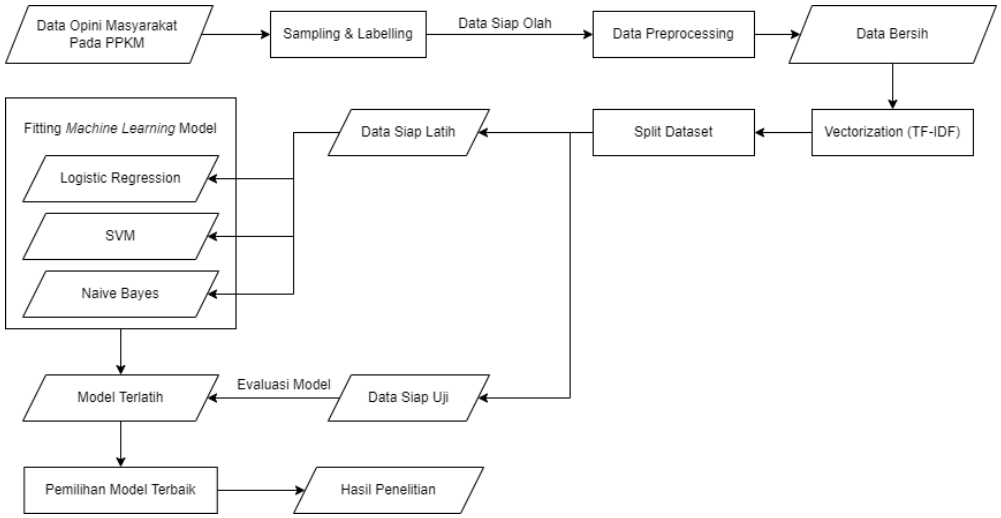
Text Mining adalah proses yang mencoba mengekstrak informasi berguna dari teks. Hal itu dapat diartikan sebagai proses menganalisa teks untuk mengekstrak informasi yang berguna untuk tujuan tertentu (Lokesh & Parul, 2013). Fungsi dari text mining cukup banyak karena cakupannya adalah teks, maka hal apapun yang ingin dianalisis dari teks dapat dilakukan dengan teknik Text Mining. Text Mining dapat digunakan untuk mencari pola – pola dalam teks yang diperlukan untuk proses klasifikasi.

3.1.2.4 Machine Learning

Machine Learning (ML) adalah disiplin ilmu yang mencakup perancangan dan pengembangan algoritma yang memungkinkan komputer untuk mengembangkan perilaku yang didasarkan kepada data empiris. *Machine Learning* juga dapat diartikan sebagai kemampuan komputer untuk melakukan pembelajaran dari pengalaman terhadap tugas yang dibebankan dengan kinerja yang terukur (Mitchell, 1997). *Machine Learning* dapat dikelompokan menjadi 3 kategori utama yaitu :

1. Pembelajaran terarah (*supervised learning*), merupakan suatu pembelajaran yang terawasi dimana jika output (label) yang diharapkan telah diketahui sebelumnya. *Sepervised learning* dibagi menjadi 2 bagian yaitu klasifikasi dan regresi.
2. Pembelajaran tidak terarah (*unsupervised learning*), merupakan pembelajaran yang tidak terawasi atau tidak memerlukan output (label) saat pembelajaran dilakukan. *Unsupervised learning* dibagi menjadi 2 bagian yaitu assosiasi dan *clustering*.
3. *Reinforcement learning* merupakan pembelajaran yang bertujuan untuk menggunakan pengamatan dan mengumpulkan data melalui interaksi langsung dengan lingkungan untuk mengambil tindakan yang akan memaksimalkan *reward* dan meminimalkan resiko. Pada *reinforcement learning* mesin dilatih untuk membuat keputusan yang spesifik sehingga akan didapatkan keputusan yang akurat.

Tahapan analisis data yang dilakukan menggunakan metode *machine learning* dapat dilihat pada began berikut ini.



Gambar 3.1 Bagan Tahapan Analisis Data

Dalam pengaplikasiannya terdapat banyak model *machine learning* yang digunakan dalam kasus pengklasifikasian teks seperti *Logistic Regression, Support Vector Machine (SVM),* dan *Naïve Bayes (NB).*

3.1.2.5 Model Machine Learning

1. *Multinomial Logistic Regression*

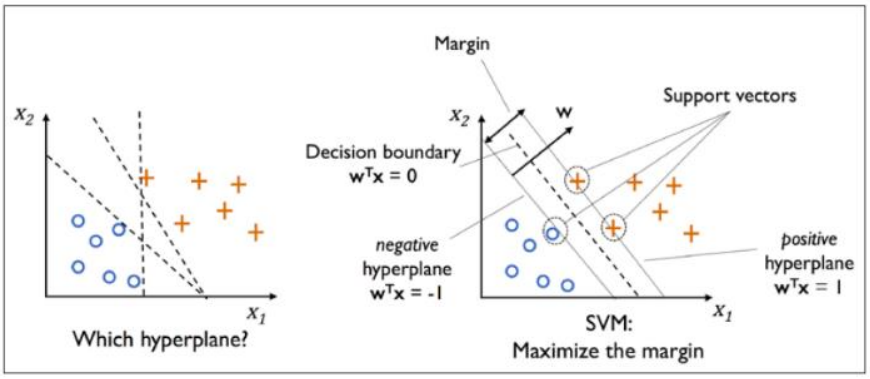
   Regresi Logistik Multinomial merupakan regresi logistik yang digunakan saat variabel dependen mempunyai skala yang bersifat polichotomous atau multinomial (Yudisasanta A. & Ratna M. 2012). Skala multinomial adalah suatu pengukuran yang dikategorikan menjadi lebih dari dua kategori. Menurut Hosmer dan Lemeshow bentuk umum model regresi logistic multinomial adalah sebagai berikut

   $$\pi(x) = \frac{e^{g(x)}}{1+e^{g(x)}} \quad (3.1)$$

   dengan $g(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$

2. *Support Vector Machine (SVM)*

   Support Vector Machine (SVM) merupakan metode klasifikasi yang kini banyak dikembangkan dan diterapkan. SVM bekerja dengan mencari *hyperplane* terbaik dengan memaksimalkan jarak antar kelas. *Hyperplane* adalah sebuah fungsi yang dapat digunakan untuk pemisah antar kelas.



Gambar 3.2 *Hyperplane* pada SVM

Hyperplane posisinya berada ditengah-tengah antar kelas, artinya jarak antara *hyperplane* dengan objek-objek data berbeda dengan kelas yang berdekatan (terluar). Dalam SVM objek data terluar yang paling dekat dengan hyperplane disebut support vector. Objek yang disebut support vector paling sulit diklasifikasikan dikarenakan posisi yang hampir tumpang tindih (overlap) dengan kelas lain. Mengingat sifatnya yang kritis, hanya support vector inilah yang diperhitungkan untuk menemukan hyperplane yang paling optimal oleh SVM.

3. *Naïve Bayes (NB)*

Naïve Bayes adalah metode pengklasifikasian dengan menggunakan metode probabilitas yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu Teorema Bayes . Prediksi Bayes yang berdasarkan pada Teorema Bayes memiliki rumus umum seperti pada persamaan berikut:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} \quad (3.3)$$

Dengan,

$P(H|E)$ = Probabilitas akhir bersyarat suatu hipotesis H terjadi jika diberikan bukti E terjadi.

$P(E|H)$ = Probabilitas sebuah bukti E terjadi akan mempengaruhi hipotesis H

$P(H)$ = Probabilitas awal hipotesis H terjadi tanpa memandang bukti apapun

$P(E)$ = Probabilitas awal bukti E terjadi tanpa memandang hipotesis/bukti yang lain

3.1.2.6 Term Frequency – Inverse Document Frequency (TF-IDF)

Untuk dapat dipahami oleh komputer maka perlu dibuat sebuah vektor yang merepresentasikan token-token hasil preprocessing dengan memanfaatkan kemunculan kata-kata atau istilah dalam dokumen. (Miner, et.al., 2012). Proses ini dinamakan sebagai proses pembobotan. Pembobotan TF-IDF adalah suatu pengukuran statistik untuk mengukur seberapa penting sebuah kata dalam kumpulan dokumen. Tingkat kepentingan meningkat ketika sebuah kata muncul beberapa kali dalam sebuah dokumen tetapi diimbangi dengan frekuensi kemunculan kata tersebut dalam kumpulan dokumen.

Metode TF-IDF akan menghitung nilai bobot Wi dalam dokumen d yaitu melalui frekuensi kemunculan suatu term atau istilah di tiap dokumen (TF (t, d)) dan frekuensi kemunculan term pada beberapa dokumen. Oleh karena itu, dihitung terlebih dahulu Term Frequency (TF) nya. Selanjutnya adalah menghitung nilai IDF (Inverse Document Frequency), yaitu nilai bobot suatu term dihitung dari seringnya suatu term muncul di beberapa dokumen. Semakin sering suatu term muncul di banyak dokumen, maka nilai IDF nya semakin kecil. DF (t) adalah jumlah dokumen yang berisi kata t. Sehingga Inverse Document Frequency (IDF) dapat dihitung dari jumlah dokumen |D| dibagi dengan banyaknya dokumen yang mengandung kata t.

Dalam metode TF-IDF, Wi adalah bobot suatu kemunculan term semakin besar jika term tersebut sering muncul dalam suatu dokumen dan semakin kecil jika term tersebut muncul dalam banyak dokumen. Skema normalisasi pembobotan TF-IDF dapat dihitung menggunakan rumus matematis sebagai berikut

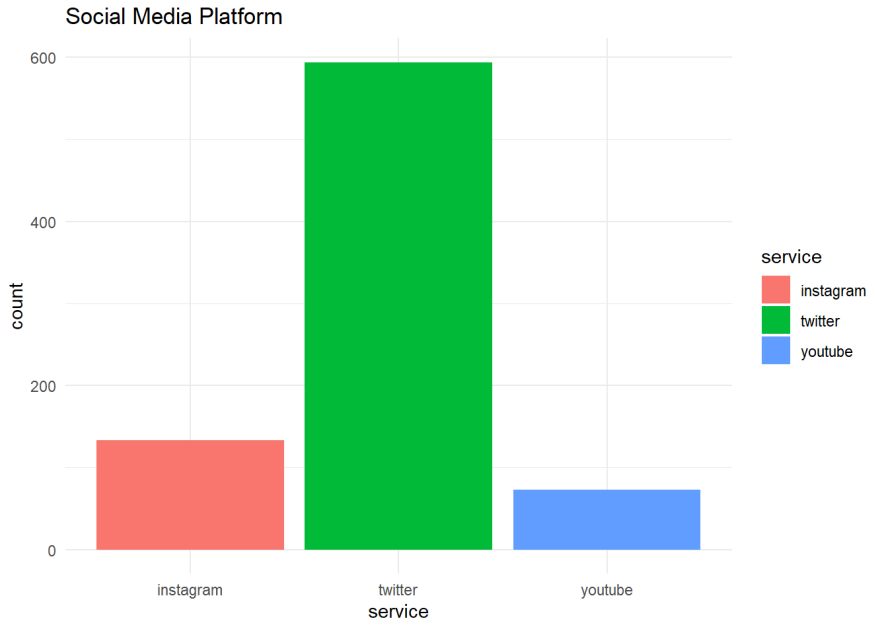
$$W_i = TF(t,d) \times IDF(t) \tag{3.4}$$

3.1.3 Hasil dan Pembahasan

Pada tahap awal penelitian penulis melakukan *sampling* terhadap data dengan tujuan efesiensi waktu dan tenaga untuk melakukan pelabelan secara manual karena

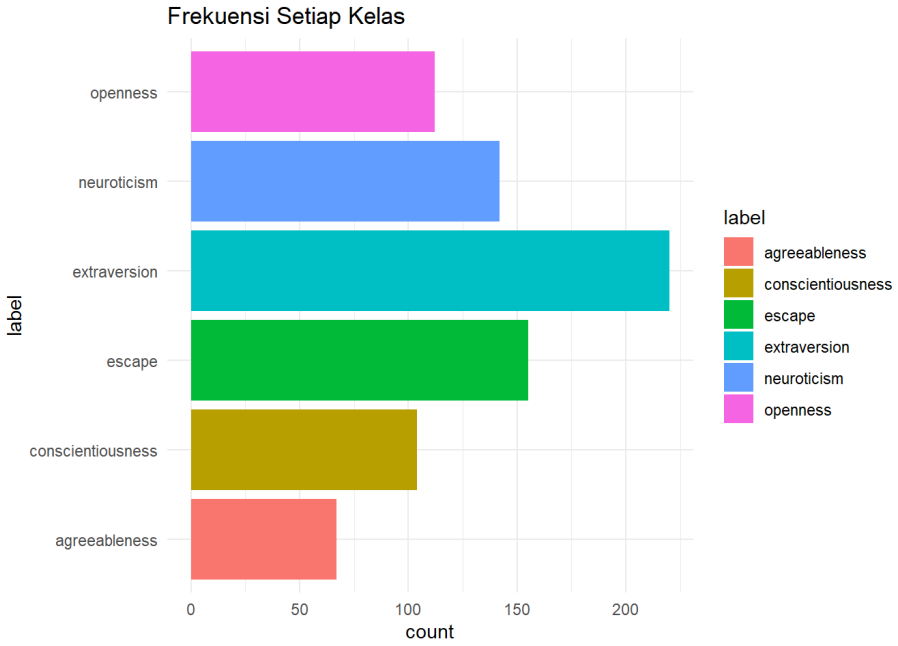
data yang didapatkan tidak memiliki variable target yang berisi kelas – kelas OCEAN di dalamnya. Proses *sampling* dilakukan untuk mereduksi data sebanyak 2.506.835 menjadi 3000. Selanjutnya data yang telah disample akan dilakukan filterisasi dan pelabelan secara manual dengan mengecek setiap metion satu – persatu. Setelah proses tersebut didapatkan 800 data yang layak untuk diolah dan diteruskan.

Selanjutnya peneliti melakukan ekporasi terhadap data untuk melihat sebaran frekuensi media sosial dengan menggunakan grafik. Berikut adalah plot untuk frekuensi sebaran mention pada platform social media



Gambar 3.3 Plot frekuensi sebaran mention pada platform social media

Berdasarkan Gambar 3.3 dapat dilihat social media *twitter* memiliki jumlah mention terbanyak dibandingkan dengan platform social media lainnya. Peneliti juga melakukan ekporasi terhadap data untuk melihat sebaran frekuensi kelas (OCEAN). Berikut adalah plot untuk frekuensi sebaran kelas (OCEAN)



Gambar 3.4 Plot frekuensi sebaran kelas (OCEAN)

Dari Gambar 3.4 dapat dilihat mention yang berlabel *extraversion* memiliki frekuensi terbanyak dibanding kelas lainnya menandakan mayoritas masyarakat memiliki sifat mampu bersosialisasi dan berinteraksi pada lingkungannya, tegas dalam mengambil keputusan, memiliki emosi positif, serta antusias terhadap dunia sosial dan material.

Selanjutnya adalah melakukan preprocessing terhadap data. *Preprocessing* dilakukan untuk membersihkan data text serta dari elemen – elemen yang tidak diperlukan atau mengganggu saat fitting model. Berikut adalah langkah – langkah yang dilakukan dalam *preprocessing*:

1. Mengganti Emoji dan Menghapus tag HTML, proses menghapus emoticon yang biasa digunakan pada setiap mention di sosial media dan menghapus tag/code html yang terdapat di dalam teks
2. Menghapus URL, proses menghapus link atau URL yang terdapat di dalam teks
3. Mengganti kata slang menjadi kata formal, proses mengganti kata – kata slang yang sering digunakan dalam mention di social media menjadi kata baku (formal)

4. Stemming, proses penguraian kata menjadi bentuk dasarnya (kata tanpa imbuhhan)
5. Menghapus Stopwords, proses menghapus kata umum yang tidak memiliki makna seperti “yang”, “dan”, dll
6. Normalisasi Teks, proses normalisasi data teks dengan cara *undercase*, menghapus *newline*, menghapus *punctuation*, menghandle huruf yang berulang disatu kata, dan menghapus *whitespace*.

Sebelum melakukan pemodelan penulis melakukan plotting terhadap data teks yang telah di *preprocess* dengan menggunakan *wordcloud* untuk melihat kata dengan frekuensi kemunculan terbesar. Berikut adalah plot untuk *wordcloud*

Gambar 3.5 *wordcloud*

Dari Gambar 3.5 dapat dilihat kata “orang” dan “ppkm” menjadi kata dengan tingkat kemunculan tertinggi atau kata yang paling sering digunakan.

Selanjutnya data yang telah bersih akan dilakukan pembobotan dengan menggunakan TF-IDF. Proses pembobotan mengubah setiap kata yang terdapat di dalam data menjadi sebuah variable yang bernilai atau bertipe numerik sehingga data

dapat di *fit* kedalam model. Dari hasil pembobotan dihasilkan 3062 variable baru yang dimana setiap variable tersebut adalah representasi dari sebuah kata. Setelah data dilakukan pembobotan data di*split* menjadi 2 bagian yaitu data latih dan data test. Data latih digunakan untuk proses *fitting* model (pelatihan model) sdangkan data *test* digunakan untuk evaluasi model.

Pada tahap pemodelan digunakan metode *Multinomial Logistic Regression, Support Vector Machine*, dan *Naïve Bayes* untuk mengkalsifikasikan data. Data latih akan di fit pada model – model tersebut dan data test akan digunakan untuk pengevaluasian.

Untuk mengukur performa model digunakan metric akurasi. Akurasi didefinisikan sebagai jumlah item yang diidentifikasi dengan benar sebagai benar-benar positif atau benar-benar negatif dari total jumlah item. Untuk jumlah kelas = m, di mana i, j = 1,2,...,m. Jika nilai akurasi semakin mendekati 100 %, maka performa klasifikator semakin tinggi. Adapun rumus umum dari akurasi adalah sebagai berikut

$$akurasi = \frac{\sum_i^m c_{ii}}{\sum_i^m \sum_j^m cij} \tag{3.4}$$

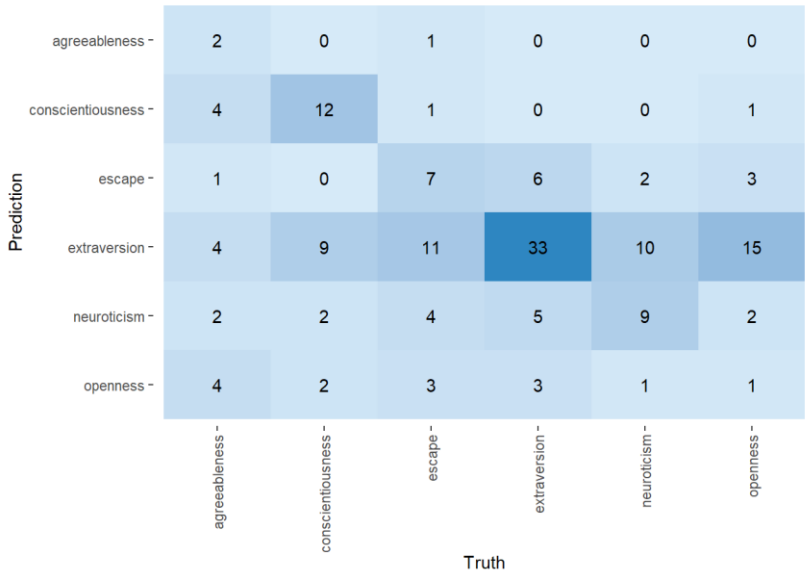
Dari hasil evaluasi model terhadap data test didapatkan hasil sebagai berikut

Tabel 3.1 Akurasi Model

Model	Akurasi
Multinomial Logistic Regression	0.4000
Support Vector Machine	0.3000
Naïve Bayes	0.29375

Dari hasil tersebut dapat dilihat bahwa *Multinomial Logistic Regression* memiliki nilai akurasi terbesar dibanding model lainnya. Sehingga *Multinomial Logistic Regression* adalah model terpilih untuk melakukan pengklasifikasian OCEAN

terhadap mention yang mengandung opini masyarakat tentang PPKM. Berikut adalah *confusion matrix* dari model *Multinomial Logistic Regression*



Gambar 3.6 *confusion matrix* model Multinomial Logistic Regression

Dari *confusion matrix* dapat dilihat bahwa kelas *extraversion* adalah kelas yang paling dipahami oleh model karena memiliki *truth rate* terbesar, Sedangkan kelas *openness* dan *agreeableness* merupakan kelas yang kurang dipahami oleh model karena memiliki *truth rate* yang kecil dan banyak terjadi *misclassification*.

3.1.4 Kesimpulan

Berdasarkan hasil analisis yang telah dilakukan maka kesimpulan yang dapat diambil adalah:

1. Dalam penelitian ini, metode yang didapat dalam mengklasifikasikan opini masyarakat terhadap PPKM kedalam OCEAN adalah *Multinomial Logistic Regression.*
2. *Multinomial Logistic Regression* memiliki tingkat akurasi terbesar dibandingkan dengan metode lainnya yaitu sebesar 0.4, Namun nilai tersebut masih jauh dari predikat baik dan masih dapat dikembangkan lagi.

3. *Extraversion* merupakan kelas dengan frekuensi terbanyak, menandakan mayoritas masyarakat memiliki sifat mampu bersosialisasi dan berinteraksi pada lingkungannya, tegas dalam mengambil keputusan, memiliki emosi positif, serta antusias terhadap dunia sosial dan material.