ANALISIS SENTIMEN MASYARAKAT TERHADAP KEBIJAKAN PPKM PADA MEDIA SOSIAL *TWITTER* MENGGUNAKAN METODE *NAIVE BAYES CLASSIFIER* (NBC)

Skripsi

Disusun untuk memenuhi salah satu syarat memperoleh gelar Sarjana Statistika

Naufal Zhafran Albaqi

1314618035

PROGRAM STUDI STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS NEGERI JAKARTA

2022

LEMBAR PENGESAHAN

Dengan ini saya mahasiswa Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Jakarta

Nama : Naufal Zhafran Albaqi

No. Registrasi : 1314618035

Jurusan : Statistika

Judul : Analisis Sentimen Masyarakat Terhadap Kebijakan PPKM Pada Media Sosial *Twitter* Menggunakan Metode *Naive Bayes Classifier* (NBC)

Menyatakan bahwa skripsi ini telah siap diajukan untuk sidang skripsi.

Menyetujui,

Dosen Pembimbing I	Dosen Pembimbing II
Prof., Dr. Suyono, M.Si	**Dania Siregar, S.Stat, M.Si**
NIDN. 0018126704	NIDN. 8840600016

Mengetahui,

Ketua Program Studi Statistika

Dr., Ir. Bagus Sumargo, M.Si.

NIP. 196309221986011001

LEMBAR PERNYATAAN KEASLIAN SKRIPSI

Saya menyatakan dengan sesungguhnya bahwa skripsi dengan judul "Analisis Sentimen Masyarakat Terhadap Kebijakan PPKM Pada Media Sosial Twitter Menggunakan Metode *Naive Bayes Classifier* (NBC)" yang disusun sebagai syarat untuk memperoleh gelar Sarjana Statistika dari Program Studi Statistika Universitas Negeri Jakarta adalah karya ilmiah saya dengan arahan dari dosen pembimbing.

Sumber informasi yang diperoleh dari penulis lain yang telah dipublikasikan dan disebutkan dalam teks skripsi ini telah dicantumkan dalam Daftar Pustaka sesuai dengan norma, kaidah, dan etika penulisan ilmiah.

Jika di kemudian hari ditemukan Sebagian besar skripsi ini asli bukan hasil karya saya sendiri dalam bagian-bagian tertentu, saya bersedia menerima sanksi pencabutan gelar akademik yang saya sanding dan sanksi-sanksi lainnya sesuai dengan peraturan perundang-undangan yang berlaku.

Jakarta, 9 Agustus 2022

Naufal Zhafran Albaqi

ABSTRAK

NAUFAL ZHAFRAN ALBAQI. Analisis Sentimen Masyarakat Tehadap Kebijakan PPKM Pada Media Sosial Twitter Menggunakan Metode Naive Bayes Classifier (NBC). Skripsi. Program Studi Statistika, Fakultas Matematika dan Ilmu PengetahuanAlam, Universitas Negeri Jakarta. Agustus 2022.

Penelitian ini dilakukan untuk menerapkan dan mengetahui bagaimana proses metode *Naive Bayes Classifier* dalam melakukan analisis sentimen terhadap *tweet* tentang PPKM, serta seberapa baik metode bekerja. Selain itu, pada penelitian ini juga akan dilihat topik apa saja yang sering menjadi perbincangan pada setiap sentimen. Penelitian ini menggunakan data sekunder dari PT.Ivonesia Solusi Data (Ivosight) tahun 2021. Hasil analisis didapatkan bahwa algoritma *Naive Bayes Classifier* mampu mendapatkan akurasi pada data latih berkisar antara 0,68 hingga 0,71. Sementara itu, tingkat akurasi pada data uji sebesar 0,714. Hasil ini menunjukan bahwa algoritma *Naive Bayes Classifier* sudah bekerja dengan baik. Pada sentimen negatif masyarakat menyoroti topik Perpanjangan PPKM, keritik terhadap penamaan PPKM berlevel, penutupan jalan, pembatasan waktu makan. Serta, penerapan sistem kerja dari rumah. Sementara itu, pada sentimen positif topik yang sering dibahas antara lain, penerapan protokol kesehatan serta yang semakin baik, vaksinasi, penurunan level PPKM serta kasus konfirmasi Covid-19.

Kata Kunci: twitter, sentimen masyarakat, ppkm, analisis sentimen, naive bayes classifier,

ABSTRACT

NAUFAL ZHAFRAN ALBAQI. Analysis of Public Sentiment Against PPKM Policy on Social Media Twitter Using Naive Bayes Classifier (NBC) Method. Thesis. Statistics Study Program, Faculty of Mathematics and Natural Sciences, State University of Jakarta. August 2022

This research is to apply and find out how the process of the Naive Bayes Classifier method in conducting sentiment analysis on tweets about PPKM and how well the method works. In addition, this study will also look at what main topics in every sentiment. This study uses secondary data from PT. Ivonesia Solusi Data (Ivosight) in 2021. The analysis results show that the Naive Bayes Classifier algorithm can get accuracy on training data. Ranged from 0,68 to 0,71. Meanwhile, the level of accuracy on the test data is 0.714. These results indicate that the Naive Bayes Classifier algorithm has worked well. On the negative sentiment, the public highlighted the topic of PPKM Extension, criticism of the naming of PPKM levels, road closures, and restrictions on mealtimes. Also, implementing a work from home system. Meanwhile, on positive sentiment, topics that are often discussed include the implementation of better health protocols, vaccinations, decreasing PPKM levels, and confirmed cases of Covid-19.

Keywords: twitter, public sentiment, ppkm, sentiment analysis, naive bayes classifier,

KATA PENGANTAR

Alhamdulillahi Rabbil' Alamin dengan memanjatkan puji syukur kepada Allah SWT. atas segala limpahan rahmat-Nya sehingga penulis dapat menyelesaikan skripsi ini dengan baik. Skripsi yang berjudul "Analisis Sentimen Masyarakat Tehadap Kebijakan PPKM Pada Media Sosial Twitter Menggunakan Metode Naive Bayes Classifier (NBC)" disusun untuk melengkapi salah satu syarat meraih gelar sarjana statistika pada Fakultas Matematika dan Ilmu Pengetahuan Alam di Universitas Negeri Jakarta.

Penulis menyadari dalam penyusunan skripsi ini penulis telah mendapat banyak bantuan, bimbingan dan dorongan dari berbagai pihak. Oleh karena itu, penulis ingin menyampaikan ucapan rasa terima kasih yang sebesar-besarnya kepada

1. Allah SWT. karena berkat segala limpahan rahmat-Nya telah memberi penulis niat, kesehatan dan kesempatan untuk menyelesaikan skripsi ini.
2. Keluarga penulis, Bapak dan Ibu, Adik, serta Bibi penulis yang selalu memberikan perhatian, semangat, dukungan, kasih sayang dan doa kepada penulis.
3. Bapak Dr. Ir. Bagus Sumargo, M.Si., selaku Koordinator Program Studi Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Negeri Jakarta.
4. Bapak Prof Dr. Suyono M.Si., selaku dosen pembimbing I dan Ibu Dania Siregar S.Stat, M.Si., selaku dosen pembimbing II yang telah meluangkan waktu dan tenaga untuk dapat memberikan bimbingan, pengetahuan, ide, kritik, saran, dan masukan sehingga skripsi ini dapat terselesaikan dengan baik.
5. Seluruh Bapak/Ibu dosen pengajar di lingkungan Program Studi Statistika dan seluruh staf administrasi di lingkungan Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Jakarta, atas segala dukungan,

bimbingan, dan petunjuk selama pelaksanaan pendidikan dan penyusunan skripsi.

6. Rekan-rekan mahasiswa/i Program Studi Statistika Universitas Negeri Jakarta angkatan 2018 yang telah menjadi teman keluh kesah membantu, bekerjasama dan memberikan motivasi, serta semangat kepada penulis selama ini.
7. Rekan-rekan mahasiswa/i Program Studi Statistika Universitas Negeri Jakarta yang telah memberikan motivasi, semangat dan kerjasama selama ini.
8. Semua pihak yang tidak dapat penulis sebutkan satu persatu yang telah membantu penulis dalam penyelesaian skripsi ini.
9. Terima Kasih kepada diri sendiri, atas segala pemikiran, waktu, tenaga serta finansial yang telah dikeluarkan. Terima kasih telah bertahan dan menyelesaikan ini semua, sehingga membuat cerita ini berakhir dengan indah dan berkesan.

Penulis menyadari bahwa masih banyak keterbatasan pengetahuan dan kemampuan yang penulis miliki. Penulis dengan senang hati menerima kritik dan saran yang membangun untuk menyempurnakan skripsi ini. Penulis berharap skripsi ini dapatberguna dan bermanfaat bagi semua yang membaca

Jakarta, Agustus 2022

Naufal Zhafran Albaqi

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR LAMPIRAN

BAB I

PENDAHULUAN

1.1 Latar Belakang

Pada tahun 2019, telah terjadi penyebaran Coronavirus atau yang lebih umum dikenal sebagai Covid-19. Adapun virus penyebab Covid-19 ini dikenal dengan nama SARS-CoV-2. Sejak Maret 2020, penyebaran Covid-19 telah dinyatakan sebagai pandemi global, hal ini dikarenakan Covid-19 telah menginfeksi jutaan warga di seluruh dunia. Berdasarkan laman resmi WHO, dilaporkan bahwa virus ini telah menginfeksi lebih dari 247 juta masyarakat dan menyebabkan 5 juta kematian di dunia (WHO, 2021). Indonesia merupakan salah satu negara yang terdampak dari penyebaran virus Covid-19, tercatat bahwa kasus pertama Covid-19 ditemukan pada 2 Maret 2020. Kasus Covid-19 di Indonesia berkembang cepat. Hingga tanggal 4 November 2021, telah lebih dari 4 juta kasus positif dan lebih dari 143 ribu kematian yang disebabkan oleh Covid-19 (Satgas Covid-19, 2021). Berbagai upaya telah dilakukan pemerintah Indonesia untuk mengantisipasi meningkatnya penyebaran Covid-19. Salah satunya, dengan menyosialisasikan gerakan 3M dan 3T. Gerakan 3M merupakan ajakan kepada masyarakat untuk selalu memakai masker, menjaga jarak dan mencuci tangan. Sedangkan 3T, merupakan salah satu program pemerintah dalam penanganan dan pendeteksian dini terhadap masyarakat yang terinfeksi Covid-19(Yulima et al., 2021). Selain itu, pemerintah juga memberlakukan aturan pembatasan kegiatan masyarakat untuk menekan angka penyebaran Covid-19(Rizal et al., 2021).

Penerapan pembatasan kegiatan masyarakat ini diatur melalui Peraturan Pemerintah Nomor 21 Tahun 2020 dan diresmikan oleh Presiden Jokowi per tanggal 31 Maret 2020. Aturan pembatasan ini diberi nama Pembatasan Sosial Berskala Besar atau lebih dikenal dengan sebutan PSBB. Namun, sejak Januari 2021 aturan tersebut diubah namanya menjadi Pemberlakuan Pembatasan Kegiatan Masyarakat atau PPKM dan mulai berlaku sejak tanggal 11 Januari - 25 Januari 2020. Kegiatan ini pun diperpanjang sebanyak 2 kali, dikarenakan angka penyebaran Covid-19 yang masih cukup tinggi saat itu. Karena dianggap kurang efektif, per 9 Februari 2021 - 20 Juli 2021 pemerintah menerapkan PPKM mikro..

Pada bulan Juli 2021 penyebaran Covid-19 varian delta di Indonesia sangatlah cepat, sehingga membuat pemerintah memberlakukan PPKM Darurat. PPKM darurat merupakan bentuk respon pemerintah terhadap melonjaknya kasus Covid-19 pada bulan Juli 2021. Setelah itu, pemerintah kembali mengganti kebijakan PPKM darurat menjadi PPKM level 1 hingga 4, yang masih berlangsung hingga saat ini. Kebijakan PPKM yang terus berubah-ubah dan telah berlangsung dalam jangka waktu yang lama, menimbulkan banyak persepsi di masyarakat. Menurut survei yang dilakukan *Saiful Mujani Research and Consulting* (SMRC) pada bulan Februari hingga Maret 2021 menyatakan bahwa sebanyak 64% responden mengetahui tentang kebijakan PPKM dan 36% tidak mengetahui tentang kebijakan PPKM, Sementara itu dari 64% yang mengetahui tentang kebijakan PPKM terdapat 55% orang yang setuju dengan kebijakan ini, sedangkan sisanya berharap kebijakan ini dihentikan karena mengganggu aktivitas ekonomi mereka (SMRC, 2021). Hal itu juga disampaikan oleh Rizal et al (2021) yang menyatakan bahwa kebijakan PPKM berdampak signifikan pada sektor UMKM karena warga harus menghindari kerumunan dan menyebabkan menurunnya pendapatan.

Perbedaan pendapat masyarakat tentang penerapan kebijakan PPKM juga terjadi di media sosial. Media sosial merupakan salah satu wadah bagi seseorang untuk menyampaikan pendapat atau pandangan yang efeknya cukup berdampak. Selain itu, dalam media sosial pengguna dapat merepresentasikan dirinya, mencari informasi, serta berbagi informasi dan pemikirannya kepada orang lain dengan jangkauan yang lebih luas dan cara yang lebih mudah. Salah satu media sosial yang paling banyak digunakan di Indonesia adalah Twitter, dengan jumlah pengguna sebesar 59% dari total pengguna media sosial (Hootsuite, 2020). Hal ini membuat Twitter banyak digunakan masyarakat untuk menyikapi atau menanggapi suatu topik. Salah satunya adalah topik penerapan kebijakan PPKM oleh pemerintah Indonesia.

Pendapat, opini, atau pandangan masyarakat yang terdapat di dalam Twitter dapat dimanfaatkan menjadi sebuah informasi yang bermanfaat. Hal ini dikarenakan, setiap *tweet* yang ada di Twitter dapat mewakili pendapat atau pandangan seseorang terhadap suatu kejadian atau topik. Salah satu pemanfaatan

data Twitter yang sering dilakukan adalah analisis sentimen. Analisis sentimen atau *opinion mining* menurut Sutoyo dan Almaarif, (2020) adalah salah satu bidang *natural language processing* (NLP) yang menggunakan analisis teks untuk mendapatkan berbagai sumber informasi dari internet dan platform media sosial. Analisis sentimen banyak digunakan untuk kepentingan ekonomi dan politik sebagai pertimbangan pengambilan atau pembaharuan suatu kebijakan, dalam prosesnya, analisis sentimen merupakan salah satu pengaplikasian klasifikasi dalam data teks. Menurut Sabrani, Alif dan Bimantoro (2020) klasifikasi teks adalah proses pengelompokan teks berdasarkan kata, frase, atau kombinasinya untuk menentukan kategori yang telah ditetapkan sebelumnya. Dalam proses pengklasifikasian data teks pada analisis sentimen, setiap pesan yang terkandung di dalam *tweet* nantinya akan diubah ke dalam bentuk matriks pembobotan kata. Setelah itu, matriks tersebut akan diklasifikasikan ke dalam tiga kategori sentimen yaitu positif, negatif dan netral. Hasil dari pengklasifikasian dan performa model yang digunakan nantinya akan menjadi hasil dari analisis sentimen.

Pengklasifikasian data teks dapat dilakukan dengan berbagai metode, antara lain *logistic regression, support vector machine, K-nearest neighbors*, dan *Naive Bayes Classifier* (NBC), masing-masing metode memiliki kelebihan dan kekurangan masing-masing. Sutoyo dan Almaarif, (2020) melakukan penelitian tentang sentimen masyarakat di media sosial Twitter dengan membandingkan algoritma *Naïve Bayes Classifier , logistic regression, support vector machine,* dan *K-nearest neighbors* hasilnya algoritma *Naive Bayes Classifier* (NBC) mampu menghasilkan tingkat akurasi sebesar 91,65%, Penelitian lain tentang analisis sentimen pada pemilihan presiden 2019, menunjukkan bahwa metode NBC yang dikombinasikan oleh *K-Means* mampu memberikan keakuratan yang cukup tinggi dengan akurasi sebesar 93.35% dan *error rate* rata-rata sebesar 6.66% (Kurniawan & Susanto, 2019) Selain itu, Yulita dkk (2021) juga melakukan penelitian opini Masyarakat Tentang Vaksin Covid-19 dengan menggunakan algoritma NBC, hasilnya algoritma NBC mampu melakukan pengklasifikasian opini masyarakat tentang vaksin dengan akurasi sebesar 93%.

Pada penelitian kali ini, metode yang akan digunakan adalah pengklasifikasian teks NBC untuk melakukan analisis sentimen pada data Twitter

dengan topik penerapan kebijakan PPKM di Indonesia. Adapun berdasarkan penelitian sebelumnya dihasilkan metode NBC mampu melakukan pengklasifikasian teks dan memiliki performa akurasi yang cukup baik. Selain itu, metode ini juga cenderung lebih cepat dan sangat baik digunakan untuk pengklasifikasian teks dengan jumlah kata yang banyak (Jurafsky, 2021). NBC adalah metode klasifikasi yang didasarkan pada penerapan *teorema bayes* untuk melakukan perhitungan probabilitas setiap kelas, dengan asumsi bahwa setiap fitur saling bebas. Namun, Murphy (2022) menyatakan jika asumsi pada NBC tidak terpenuhi, maka model dapat tetap digunakan dan menghasilkan hasil yang baik. Penelitian ini dilakukan dengan tujuan untuk mengetahui bagaimana dan seberapa baik performa algoritma NBC bekerja. Selain itu, dalam penelitian ini juga akan dilihat bagaimana respon masyarakat terhadap penerapan kebijakan PPKM yang diterapkan oleh pemerintah.

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas dapat diperoleh beberapa rumusan masalah sebagai berikut:

1. Bagaimana metode Naive Bayes Classifier dapat melakukan analisis sentimen terhadap *tweet* tentang penerapan PPKM oleh pemerintah Indonesia? Dan seberapa baik metode bekerja?
2. Bagaimana gambaran sentimen masyarakat pada media sosial Twitter terhadap penerapan kebijakan PPKM?

1.3 Batasan Masalah

Penelitian ini membatasi pokok permasalahan dengan beberapa batasan sebagai berikut:

1. Data yang digunakan merupakan data *tweet* milik PT. Ivonesia Solusi Data yang berhubungan dengan PPKM dalam jangka waktu 1 Juli 2021 hingga 31 Oktober 2021, dengan kata kunci pencarian yang digunakan adalah *"PPKM"* dan *"Pemberlakuan Pembatasan Kegiatan Masyarakat"*

2. Jumlah sampel acak yang digunakan sebanyak 5000 *tweet* dengan metode pengambilan sampel yang digunakan adalah *simple random sampling*
3. Jenis pembobotan kata yang digunakan dalam penelitian ini adalah *Term Frequency-Inverse Document Frequency* (TF-IDF)
4. Populasi data diasumsikan homogen yaitu pengguna media sosial Twitter dan terdampak penerapan kebijakan PPKM

1.4 Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah sebagai berikut:

1. Menerapkan dan mengetahui bagaimana proses metode *Naive Bayes Classifier* dalam melakukan analisis sentimen terhadap *tweet* tentang PPKM, serta seberapa baik metode bekerja.
2. Mengetahui gambaran sentimen masyarakat pada media sosial Twitter terhadap penerapan kebijakan PPKM.

1.5 Manfaat Penelitian

Dengan dilakukannya penelitian ini, peneliti diharapkan mampu mempelajari dan menerapkan proses analisis sentimen dengan metode *Naive Bayes Classifier*. Selain itu, bagi pembaca dan masyarakat dapat mengetahui proses analisis sentimen dengan metode *Naive Bayes Classifier,* serta gambaran sentimen yang terbentuk dalam penerapan kebijakan PPKM oleh pemerintah. Bagi pemerintah, penelitian diharapkan dapat dijadikan sumber informasi dalam penerapan atau pembaruan kebijakan serupa.

BAB II

LANDASAN TEORI

2.1 Twitter

Twitter merupakan salah satu media sosial yang ada saat ini, Twitter didirikan oleh Jack Dorsey pada Maret 2006 dan diresmikan pada bulan Juli 2006. Pada media sosial Twitter, pengguna dapat mengirim dan memperbarui status mereka ketika sedang memikirkan atau melakukan sesuatu. Sehingga saat status diperbarui, dapat memudahkan pengguna lain untuk mengakses informasi tersebut. Twitter masuik kedalam media sosial bertipe *micro-blogging* (blog berukuran kecil) Status yang dikirimkan pada Twitter disebut *tweet*, pengguna kata dalam *tweet* dibatasi hingga 140 kata dalam setiap pengiriman, dengan mengirim *tweet* maka pengguna setuju bahwa informasi atau status mereka dapat dilihat oleh pengguna lain (Kurniawan et al., 2017).

Gambar 2. 1 Tampilan Twitter

2.2 Analisis Sentimen

Analisis sentimen menurut Muzaki dan Witanti (2021) merupakan salah satu bidang studi yang memiliki fungsi menganalisis opini, sentimen, penilaian, perilaku dan emosi dalam masyarakat melalui suatu objek seperti produk, layanan umum, organisasi, individual, isu, kejadian atau sebuah topik. Analisis sentimen menggunakan data teks sebagai bahan baku utama untuk dianalisis. Data teks tersebut diharapkan mampu mewakili emosi, pandangan atau pendapat dari orang yang mengirimkan teks tersebut. Selanjutnya, data teks akan diubah menjadi bentuk matriks angka. Tujuanya adalah untuk memudahkan proses analisis dan pemodelan. Pada umumnya, hasil pengelompokan analisis sentimen terbagi dalam tiga kategori emosi yaitu positif, negatif dan netral. Proses penentuan apakah sebuah kalimat termasuk kedalam sentiment positif, negatif dan netral bisa disebut dengan pengklasifikasian(Fahrur Rozi et al., 2012)

2.3 PPKM

Pada bulan Januari 2021 pemerintah mengganti kebijakan pembatasan sosial berskala besar (PSBB), menjadi pemberlakuan pembatasan kegiatan masyarakat (PPKM). PPKM merupakan respon pemerintah dalam mencegah menyebarnya Covid-19 dengan pembatasan kegiatan masyarakat. Adapun beberapa hal yang diatur dalam penerapan PPKM menurut Rizal et al (2021) adalah sebagai berikut:

1. Membatasi kegiatan pada tempat kerja/perkantoran dengan menerapkan kerja dari rumah (WFH) dan memberlakukan protokol kesehatan secara lebih ketat.
2. Melaksanakan kegiatan belajar mengajar secara daring.
3. Pengaturan jam operasional, kapasitas pada sektor esensial dan fasilitas publik.

Adapun ringkasan pelaksanaan PPKM dapat dilihat dari tabel dibawah ini.

Tabel 2.1 Ringkasan penerapan kebijakan PPKM

Kebijakan	Tahap	Mulai	Hingga	Wilayah
PPKM	1	11 Januari 2021	25 Januari 2021	7 provinsi
	2	26 Januari 2021	8 Februari 2021	7 provinsi
PPKM Mikro	1	9 Februari 2021	22 Februari 2021	7 provinsi
	2	23 Februari 2021	8 Maret 2021	7 provinsi
	3	9 Maret 2021	22 Maret 2021	10 provinsi
	4	23 Maret 2021	5 April 2021	15 provinsi
	5	6 April 2021	19 April 2021	20 provinsi
	6	20 April 2021	3 Mei 2021	25 provinsi
	7	4 Mei 2021	17 Mei 2021	30 provinsi
	8	18 Mei 2021	31 Mei 2021	30 provinsi
	9	1 Juni 2021	14 Juni 2021	Nasional
	10	15 Juni 2021	28 Juni 2021	Nasional
	11	22 Juni 2021	5 Juli 2021	Nasional
	12	6 Juli 2021	20 Juli 2021	Nasional
PPKM Darurat	-	3 Juli 2021	20 Juli 2021	Jawa dan Bali
	-	12 Juli 2021	20 Juli 2021	15 Wilayah luar Jawa dan Bali
PPKM Level 1 hingga 4	-	21 Juli 2021	25 Juli 2021	Sejumlah Provinsi
	-	26 Juli 2021	2 Agustus 2021	Sejumlah Provinsi
	-	3 Agustus 2021	Sekarang	Disesuaikan

2.4 *Text Mining*

Text Mining merupakan teori tentang pengumpulan teks dengan tujuan untuk mengetahui dan mengambil informasi yang bermanfaat (Sabrani, dkk, 2020). *Text mining* dapat digunakan dalam mengoptimalkan pengambilan dan pembaharuan keputusan, pencarian sebuah teks dan analisis sentimen. Dalam proses *text mining*, akan dilakukan pencarian kata-kata yang dapat mewakili teks, sehingga dapat dilakukan pengolahan kata dan analisis lebih lanjut. Pada proses *text mining* pada umumnya data teks masih berbentuk mentah, sehingga perlu

dilakukan tahap *pre-processing* data, yang bertujuan untuk menyamaratakan struktur teks agar teks dapat dinilai, dianalisis, dan di klasifikasikan (I. Kurniawan & Susanto, 2019). Adapun tahap *pre-processing* yang dilakukan pada *text mining* adalah sebagai berikut:

1. *Cleansing*, adalah proses yang dilakukan untuk membersihkan fitur-fitur yang tidak diperlukan dalam pengambilan data yang ada pada Twitter, seperti URL, Username, dan lain-lain.
2. *Case Folding*, adalah proses perubahan seluruh huruf kapital (*uppercase*) dikembalikan menjadi huruf kecil (*lowercase*) agar seragam.
3. *Tokenizing*, adalah proses untuk memenggal kalimat menjadi bentuk satuan kata dengan *spasi* dan *enter* sebagai pemisah dari setiap kata.
4. *Filtering*, merupakan proses yang dilakukan dengan cara menghilangkan kata yang tidak diperlukan. Seperti kata hubung dan kata sambung.
5. *Stemming*, adalah proses yang memiliki tujuan untuk menghilangkan imbuhan-imbuhan yang terdapat pada sebuah kata. Proses ini mengubah kembali semua kata menjadi kata dasar.
6. *Normalization* adalah proses penyamaan kata yang memiliki makna sama, namun penulisan berbeda.

2.5 *Machine Learning* dan Klasifikasi

Menurut Murphy (2022), *machine learning* adalah program komputer yang dapat secara otomatis mendeteksi pola dalam data, dan kemudian menggunakan pola yang tidak terungkap untuk memprediksi masa depan. *Machine learning* merupakan cabang dari kecerdasan buatan *(artificial intelligent), machine learning* mampu membentuk sistem atau model yang secara otomatis dapat mempelajari dan meningkatkan kemampuan, dengan cara mengakses dan mempelajari data. Nantinya, *machine learning* akan menemukan pola dalam data untuk pengambilan sebuah keputusan secara optimal. Tujuan utamanya adalah untuk menyederhanakan pekerjaan manusia dan mengurangi kemungkinan kesalahan manusia. Menurut Kyriakos Chatzidimitriou et al (2013) berdasarkan tugasnya machine *learning* terbagi menjadi dua yaitu, *unsupervised learning* dan *supervised learning*

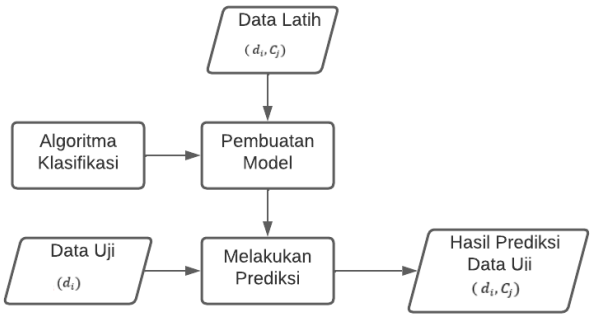
1. *Unsupervised learning*

   Unsupervised learning merupakan jenis *machine learning* yang akan mencoba memahami, membentuk pola, dan menemukan struktur berdasarkan data *input* (x) tanpa ada contoh *output* (y) yang di berikan sebelumnya, Contoh penggunaan *Unsupervised learning* adalah *clustering* dan *dimensionality reduction*

2. *Supervised learning*

   Supervised learning merupakan jenis *machine learning* yang akan mencoba memetakan data *input* $x \epsilon X$ terhadap *output* $y \epsilon Y$, pemetaan dapat diartikan sebagai fungsi dimana setiap *input* (x) akan memiliki nilai *output* (y) masing masing, tujuan *Supervised learning* adalah untuk memperkirakan fungsi pemetaan agar dapat memprediksi nilai *output* (y) ketika memiliki data *input* (x) yang baru. Contoh penggunaan *supervised learning* adalah regresi dan klasifikasi.

Klasifikasi merupakan proses pengelompokan *input* (x) dimana data tersebut telah memiliki *output* (y) berupa label kategorik $Y = \{1,2, \dots, c\}$ proses yang terjadi dalam klasifikasi dikenal dengan pengenalan pola atau *pattern recognition* (Murphy, 2022). Klasifikasi tidak hanya dapat diterapkan pada data numerik tetapi juga dalam pengklasifikasian teks, pada data teks dokumen/kalimat akan diklasifikasikan ke dalam sebuah kategori, misalkan jika $d_i \epsilon D$, dimana D adalah seluruh himpunan dokumen/kalimat dan $\{C_1, C_2, \dots, C_n\}$ merupakan kategori, maka pengkalsifikasian teks akan memasangkan setiap C_j terhadap d_i. Proses klasifikasi dapat dilihat pada gambar 2.1



Gambar 2. 2 Proses Klasifikasi

Pada gambar 2.1 data latih (C_j, d_i) merupakan data yang digunakan untuk membangun model klasifikasi, C_j merupakan *output* dari *input* dokumen d_i pada data latih, kemudian akan dilakukan pemetaan pada data latih (C_j, d_i) dengan menggunakan algoritma klasifikasi, hasil pemetaan akan membentuk model sesuai dengan algoritma klasifikasi yang digunakan, model yang terbentuk dalam pemetaan data latih akan digunakan untuk proses pengklasifikasian pada data uji yang hanya memiliki *input* dokumen d_i (belum memiliki *output* kategorik C_j) hasil pengkasifikasian akan menghasilan *output* katagotik C_j pada data uji.

2.6 *Simple Random Sampling*

Simple random sampling adalah sebuah metode penarikan sampel sejumlah k sampel dari sebuah populasi berjumlah N secara sederhana. Dimana nilai $k \leq N$, dan $k \epsilon N$. Dalam *simple random sampling*, semua anggota populasi memiliki peluang yang sama untuk terpilih menjadi sampel (Scheaffer & Mendenhall, 2012). Untuk menentukan jumlah minimum sampel yang terpilih, dengan jumlah populasi yang besar dan diketahui nilainya, dapat menggunakan formula berikut (Israel, 1992).

$$n = \frac{N}{1+N(e)^2} \tag{2.1}$$

Dimana,

n = Jumlah sampel

N = Jumlah populasi

e = *Standard error sample*

2.7 *Term Frequency-Inverse Document Frequency (TF-IDF)*

*Term Frequency-Inverse Document Frequency (*TF-IDF) merupakan salah satu metode dalam *Term Weighting. Term Weighting* adalah sebuah proses untuk menentukan bobot pada suatu kata. TF-IDF merupakan kombinasi dari *term frequency (TF)* dan *inverse document frequency (IDF),*

1. *Term Frequency (TF)*

   Term Frequency (TF) merupakan perhitungan berapa kali kata *t* muncul dalam sebuah dokumen/kalimat, adapun perhitunganya adalah sebagai berikut:

$$TF_{(t,d)} = \frac{f_{(t,d)}}{\sum f_{(t,d)}} \quad (2.2)$$

   Dimana,

   $TF_{(t,d)}$ = *Term Frequency (TF)*

   $f_{(t,d)}$ = jumlah kata *t* yang ada dalam dokumen *d*

   $\sum f_{(t,d)}$ = jumlah seluruh kata yang ada dalam dokumen *d*

   Pada perhitungan TF nilai frekuensi kata yang lebih tinggi dianggap lebih penting dari pada kata yang memiliki frekuensi lebih rendah, namun penggunaan TF sangat terbatas, karena hanya terpaku pada sebuah dokumen. Sehingga, digunakan *inverse document frequency* (IDF) untuk melengkapinya.

2. *Inverse document frequency* (IDF)

   Inverse document frequency (IDF) merupakan perhitungan apakah sebuah kata umum atau jarang digunakan dalam dokumen. Ini diperoleh dengan membagi jumlah kalimat dalam kumpulan kalimat dengan jumlah kalimat yang berisi istilah, dan kemudian mengambil logaritma hasil bagi.

$$IDF_{(t)} = log\frac{N}{DF_t} + 1 \quad (2.3)$$

   Dimana,

   $IDF_{(t)}$ = *Inverse document frequency* (IDF)

   N = Jumlah dokumen pada data latih

   DF_t = Jumlah dokumen pada data latih yang muncul kata *t*

TF akan menunjukkan betapa pentingnya kata *t* dalam sebuah dokumen *d*, sedangkan IDF menunjukkan seberapa umum kata *t* digunakan dalam kumpulan dokumen/kalimat. semakin kata *t* sering mucul dalam sebuah dokumen dan semakin jarang kata *t* muncul pada dokumen lain, akan membuat bobot kata tersebut semakin besar. Maka formula dari TF-IDF adalah

$$TFIDF_{(t,d)} = TF_{(t,d)} \times IDF_{(t)} \quad (2.4)$$

Sutoyo dan Almaarif (2020) menyatakan pembobotan TF-IDF sering digunakan karena metode ini merupakan salah satu jenis metode pembobotan yang efisien, mudah dan memiliki hasil yang akurat, karena dapat membobotkan setiap kata secara proporsional dalam sebuah dokumen. Selain itu, Rahman dan Doewes, (2017) dalam penelitiannya menyimpulkan bahwa penggunaan TF-IDF yang dikombinasikan dengan *Naive Bayes Classifier* mampu meningkatkan rata-rata akurasi dari model.

2.7.1 Contoh Perhitungan TF-IDF

Berikut ini merupakan contoh perhitungan TF-IDF pada lima dokumen.

Tabel 2. 2 Contoh Data Dokumen

Kalimat	Kode Dokumen
ppkm, percaya, sesal	d1
cegah, sebar, covid, kapolsek, selbar, bagi, masker, gratis, masyarakat, pasar, surabrata	d2
vaksinasi, jaga, sehat	d3
tka, china, bandara, sultan, hasanuddin, ppkm, darurat	d4
ppkm, level, panjang, jokowi	d5

4.1 *Term Frequency* (TF)

Pada tabel TF akan dihitung bobot keberadaan kata pada setiap dokumen, misalkan akan dihitung nilai TF untuk kata *PPKM* pada *d1* maka dapat dihitung bedasarkan formula (2.2) sebagai berikut

$$TF_{(ppkm,d1)} = \frac{f_{(ppkm,d1)}}{\sum f_{(t,d1)}} = \frac{1}{3}$$

secara keseluruhan untuk setiap kata pada setiap dokumen akan didapatkan hasil pada tabel berikut.

Tabel 2.3 Contoh Hasil Perhitungan TF

Kata	d1	d2	d3	d4	d5
ppkm	1/3	0	0	1/7	1/4
percaya	1/3	0	0	0	0

Kata	d1	d2	d3	d4	d5
sesal	1/3	0	0	0	0
cegah	0	1/11	0	0	0
sebar	0	1/11	0	0	0
covid	0	1/11	0	0	0
kapolsek	0	1/11	0	0	0
selbar	0	1/11	0	0	0
bagi	0	1/11	0	0	0
masker	0	1/11	0	0	0
gratis	0	1/11	0	0	0
masyarakat	0	1/11	0	0	0
pasar	0	1/11	0	0	0
surabrata	0	1/11	0	0	0
vaksinasi	0	0	1/3	0	0
jaga	0	0	1/3	0	0
sehat	0	0	1/3	0	0
tka	0	0	0	1/7	0
china	0	0	0	1/7	0
bandara	0	0	0	1/7	0
sultan	0	0	0	1/7	0
hasanuddin	0	0	0	1/7	0
darurat	0	0	0	1/7	0
level	0	0	0	0	1/4
panjang	0	0	0	0	1/4
jokowi	0	0	0	0	1/4

4.2 *Inverse document frequency* (IDF)

Pada perhitungan IDF akan dilihat bobot keberadaan kata dalam keseluruhan kumpulan dokumen, misalkan akan dihitung nilai IDF untuk kata PPKM pada keseluruhan dokumen berdasarkan formula (2.3) sebagai berikut.

$$IDF_{(PPKM)} = log \frac{N}{DF_t} + 1$$

$$= log\frac{5}{3} + 1$$

$$= 1{,}22$$

Dan bobot IDF setiap kata didapatkan hasil pada tabel berikut.

Tabel 2.4 Contoh Perhitungan IDF

Kata	**TF**	**IDF**
ppkm	3	1,221
percaya	1	1,698
sesal	1	1,698
cegah	1	1,698
sebar	1	1,698
covid	1	1,698
kapolsek	1	1,698
selbar	1	1,698
bagi	1	1,698
masker	1	1,698
gratis	1	1,698
masyarakat	1	1,698
pasar	1	1,698
surabrata	1	1,698
vaksinasi	1	1,698
jaga	1	1,602
sehat	1	1,602
tka	1	1,602
china	1	1,602
bandara	1	1,602
sultan	1	1,602
hasanuddin	1	1,602
level	1	1,602
panjang	1	1,602
jokow	1	1,602

Sehingga, perhitungan bobot TF-IDF kata PPKM pada D1 berdasarkan formula (2.4) adalah

$$TFIDF_{(ppkm,d1)} = TF_{(ppkm,D1)} \times IDF_{(ppkm)}$$
$$= (1/3)(\ 1{,}221)$$
$$= \ 0{,}433$$

Bobot TF-IDF dari setiap kata pada setiap dokumen akan menghasilkan matriks sebagai berikut.

Tabel 2. 5 *Contoh* Hasil Matriks TF-IDF

Kata	**Matriks TF-IDF**				
	d1	**d2**	**d3**	**d4**	**d5**
ppkm	0,407	0	0	0,174	0,305
percaya	0,566	0	0	0	0
sesal	0,566	0	0	0	0
cegah	0	0,154	0	0	0
sebar	0	0,154	0	0	0
covid	0	0,154	0	0	0
kapolsek	0	0,154	0	0	0
selbar	0	0,154	0	0	0
bagi	0	0,154	0	0	0
masker	0	0,154	0	0	0
gratis	0	0,154	0	0	0
masyarakat	0	0,154	0	0	0
pasar	0	0,154	0	0	0
surabrata	0	0,154	0	0	0
vaksinasi	0	0	0,566	0	0
jaga	0	0	0,566	0	0
sehat	0	0	0,566	0	0
tka	0	0	0	0,242	0
china	0	0	0	0,242	0
bandara	0	0	0	0,242	0
sultan	0	0	0	0,242	0
hasanuddin	0	0	0	0,242	0
darurat	0	0	0	0,242	0
level	0	0	0	0	0,424
panjang	0	0	0	0	0,424
jokow	0	0	0	0	0,424

Matriks pembobotan kata TF-IDF seperti pada tabel 2.5 nantinya yang akan digunakan menjadi *input* model klasifikasi.

2.8 *Naive Bayes*

Teorema bayes adalah metode digunakan untuk menghitung probabilitas terjadinya suatu peristiwaberdasarkan pengaruh observasi peristiwa sebelumya. *Teorema bayes* menyempurnakan penggunaan teorema peluang bersyarat yang hanya terbatas pada dua kejadian, jika terdapat lebih dari dua kejadian atau n kejadian maka menurut William Bolstad (2007) berlaku aturan berikut:

- Perpaduan kejadian $Y_1 \cup Y_2 \cup \ldots \cup Y_n = U$ atau semesta
- Setiap pasangan kejadian yang berlainan adalah saling lepas, $Y_\mathrm{i} \cap Y_\mathrm{j} = \emptyset$ untuk $i = 1, 2, \ldots, n \;\; j = 1, 2, \ldots, n$ dan $i \neq j$

Karena himpunan kejadian $Y_1, Y_2, \ldots, Y_n$ membagi semesta, kejadian X dapat diperoleh dari gabungan beberapa kejadian $X = (X \cup Y_1) \cup (X \cup Y_2) \cup \ldots \cup (X \cup Y_n)$ dimana, $(X \cup Y_\mathrm{i})$ dan $\left(X \cup Y_j\right)$ saling lepas, karena Y_i dan Y_j saling lepas, sehingga

$$P(X) = \sum_{j=1}^{n} P(X \cap Y_j) \tag{2.5}$$

Persamaan (2.4) adalah hukum probabilitas, dimana total suatu peristiwa X adalah jumlah peluang dari bagian-bagiannya yang saling lepas, dengan menggunakan aturan perkalian pada setiap peluang diperoleh.

$$P(X) = \sum_{j=1}^{n} P(X \cap Y_j) \;\; X \; P(Y_j) \tag{2.6}$$

Maka, dapat dirumuskan peluang bersyarat $P(Y_i|X)$ untuk $i = 1,2,..,n$ adalah sebagai berikut

$$P(Y_i|X) = \frac{P(X \cap Y_i)}{P(X)}$$

$$= \frac{P(X|Y_i) \quad P(Y_i)}{\sum_{j=1}^{n} \sum P(X|\,Y_j) \quad P(Y_j)}$$

Dimana,

$P(Y_i|X)$ = probabilitas Y_i terjadi jika X terjadi
$P\left(X|Y_j\right)$ = probabilitas munculnya X jika Y_j terjadi
$P(Y_i)$ = probabilitas Y_i
$P\left(Y_j\right)$ = probabilitas Y_j

2.8.1 *Naive Bayes Classifier* (NBC)

Naive Bayes Classifier merupakan salah satu metode *Supervised learning* dalam *machine learning* yang dapat efektif digunakan dalam pengolahan dan pemanfaatan data teks dengan jumlah yang besar dan memiliki respons kategorik (Myatt, 2007). Menurut Agustin, dkk (2019) penerapan metode klasifikasi ini ini didasarkan pada penerapan teorema bayes, dalam *Naive Bayes Classifier* hubungan antar variabel bebas yang mempengaruhi variabel respon di asumsikan bersifat saling bebas *(independent)*, Metode ini disebut *"Naive"* karena variabel bebas pada kenyataanya sering kali tidak saling bebas, namun Murphy (2022) menyatakan jika asumsi *independent* pada *Naive Bayes* tidak terpenuhi, maka model dapat tetap digunakan dan menghasilkan hasil yang baik. Secara umum, klasifikasi *Naive Bayes* dengan menerapkan teorema bayes dapat dinyatakan dengan rumus sebagai berikut,

$$P(V_j|x_1, x_2, \dots, x_n) = \frac{P(x_1,x_2,\dots,x_n|V_j)P(V_j)}{P(x_1,x_2,\dots,x_n)} \tag{2.7}$$

$P(V_j,\ x_1, x_2, \dots, x_n)$ merupakan peluang gabungan, maka dengan menggunakan aturan rantai pada persamaan bayes diperoleh

$$\begin{aligned} P(V_j, x_1, x_2, \dots, x_n) &= P(x_1, x_2, \dots, x_n, V_j) \\ &= P(x_1|x_2, \dots, x_n, V_j)\ P(x_2, \dots, x_n, V_j) \\ &= P(x_1|x_2, \dots, x_n, V_j)\ P(x_2|x_3, \dots, x_n, V_j) P(x_3, \dots, x_n, V_j) \dots \\ &= P(x_1|x_2, \dots, x_n, V_j) P(x_2|x_3, \dots, x_n, V_j) P(x_{n-1}|x_n, V_j) P(V_j) \end{aligned}$$

Berdasarkan penjabaran diatas maka terbentuk suatu persamaan baru:

$$P(x_i|x_{i+1}, \dots, x_n, V_j) = P(x_i|V_j) \tag{2.8}$$

Untuk setiap $i = 1,2,3,\dots,n$ dan $j = 1,2,3,\dots,n$

Jika k kejadian $A_1,\ A_2, \dots, A_k$ dikatakan *independent* atau *mutually independent* jika untuk setiap $j = 2, 3, \dots, k$ dan setiap subset pada indeks yang berbeda i_1, $i_2, \dots, i_k$ (Bain, 1992) sehingga

$$P(A_{i1} \cap A_{i2} \cap \dots \cap) = P(A_{i1})P(A_{i2}) \dots P(A_{i1}) \quad (2.9)$$

Maka, dengan menerapkan persamaan (2.8) dan pernyataan pada persamaan (2.9) diatas, persamaan (2.7) dapat disederhanakan menjadi.

$$P(V_j|x_1, x_2, \dots, x_n) = \frac{P(V_j) \prod P(x_i|V_j)}{P(x_1, x_2, \dots, x_n)} \quad (2.10)$$

Karena $P(x_1, x_2, x_3, \dots, x_n)$ bernilai konstan pada setiap v_j, maka persamaan (2.10) dapat ditulis menjadi,

$$P(V_j|x_1, x_2, \dots, x_n) \; \alpha \; P(V_j) \prod P(x_i|V_j) \quad (2.11)$$

Berdasarkan persamaan (2.10) $P(V_j|x_1, x_2, \dots, x_n)$ proporsional atau sebanding (**α**) dengan $P(V_j) \prod P(x_i|V_j)$ Dengan terminologi probabilitas *bayesian* maka persamaan (2.11) dapat ditulis sebagai berikut.

$$Posterior \; \alpha \; prior \times likelihood$$

Pada pengklasifikasian dengan menggunakan algoritma *Naive Bayes Classifier* (NBC), pengambilan keputusan didasari oleh nilai tertinggi dari setiap kategori dari kelompok label *(Maksimum a posterior).* Kategori atau label yang memiliki nilai tertinggi tersebut akan menjadi hasil akhir pengklasifikasian, sehingga formula pengklasifikasian dengan algoritma *Naïve Bayes Classifier* (NBC) jika diterapkan pada persamaan (2.11) adalah sebagai berikut:

$$\hat{V}_{map} = \underset{v_{jeV}}{armax} \hat{P}(V_j) \prod \hat{P}(x_i|V_j) \quad (2.12)$$

Pada persamaan (2.11) terdapat terlalu banyak perkalian pada $\prod \hat{P}(x_i|V_j)$ sehingga menyebabkan *Floating Point Underflow*, yaitu kondisi nilai akan menjadi sangat kecil. Untuk mengatasi itu digunakan aturan penjumlahan dalam logaritma, dimana $log\,(XY) = log\,(X) + log\,(Y)$ sehingga persamaan (2.12) berubah menjadi

$$\hat{V}_{map} = \underset{v_{jeV}}{armax} \left[log \hat{P}(V_j) + \sum log \, \hat{P}(x_i|V_j) \right] \quad (2.13)$$

dimana

$\hat{V}_{map}$ = Nilai Posterior atau hasil pengklasifikasian

$\hat{P}(V_j)$ = Nilai Probabilitas *Prior* pada data latih

$\prod \hat{P}(x_i|V_j)$ = perkalian bersyarat term x_i pada katagorik V_j (fungsi likelihood)

Perhitungan *Prior Information* dan *Sample Information (likelihood function)* adalah sebagai berikut

1. *Prior Information*

   Stone (2013) dalam Agustin, dkk (2019) menyatakan untuk menghitung nilai *prior information* yang tidak diketahui sebaran nilai parameternya, maka seluruh anggota parameter dianggap memiliki peluang yang sama atau distribusi seragam diskrit. Nilai peubah acak X merupakan distribusi seragam diskrit untuk $x_1, x_2, \ldots, x_n$ maka distribusi seragam sebagai berikut:

   $$F(x;n) = \frac{1}{n}, x = x_1, x_2, \ldots, x_n \quad (2.14)$$

   $F(x;n)$ dipakai sebagai pengganti $F(x)$ yang menunjukan bahwa distribusi seragam bergantung pada parameter n untuk menghitung nilai rataan distribusi seragam diskret $F(x;n)$ dapat menggunakan

   $$\begin{aligned} \mu &= E(X) \\ &= \sum X_i \, F(x;n) \\ &= \sum \frac{X_i}{n} \\ &= \frac{\sum X_i}{n} \end{aligned}$$

   Sehingga, perhitungan *prior information* menurut dokumentasi *scikit-learn developers* (2007) dapat digunakan menggunakan formula (2.13)

   $$\hat{P}(V_j) = \frac{N_j}{N} \quad (2.15)$$

   dimana,

   $P(V_j)$ = peluang munculnya sentimen j pada data *train*

   N_j = jumlah dokumen sentimen j pada data *train*

   N = Jumlah seluruh dokumen

2. *Sample Information (likelihood function)*

   Perhitungan *sample information* akan dihitung dengan fungsi *likelihood.* Fungsi *likelihood* adalah fungsi densitas bersama dari n variable acak $Y_1, Y_2, \dots, Y_n$ yang dinyatakan dalam bentuk $f\,(y_1, y_1, \dots, y_n; \theta)$ jika $y_1, y_1, \dots, y_n$ (Bain, 1992), fungsi *likelihood* adalah fungsi dari parameter θ atau $L(\theta)$. Jika $Y_1, Y_2, \dots, Y_n$ menyatakan suatu sampel acak dari $f(y, \theta)$ maka

$$L(\theta) = f(y_1; \theta) f(y_2; \theta) \dots f(y_n; \theta)$$
$$= \prod_{i=1}^{n} f(y_n; \theta)$$

   Dalam penggunaan Algoritma *Naïve Bayes Classifier* pada analisis sentimen perhitungan nilai *likelihood* diestimasikan berdistribusi *multinomial* karena pembobotan kata yang digunakan *TF-IDF* yaitu bobot setiap kata x_i dalam dokumen d_i sehingga dokumen d_i dapat diwakili dengan vector $\overrightarrow{x_i}$ = $\langle x_1, x_2, \dots, x_n \rangle$. fungsi *likelihood* parameter dari kata yang muncul dalam dokumen yang diestimasikan berdistribusi *multinomial* adalah

$$\hat{P}\left(\overrightarrow{x_i} | V_j\right) = P(|d_i|)|d_i|! \, \frac{P(x_i|V_j)}{x_n!} \qquad (2.16)$$

   Atau, bedasarkan dokumentasi *sklearn.multinomial* menggunakan pembobotan *TF-IDF* perhitungan nilai *sample information* dapat menggunakan persamaan berikut.

$$P\left(x_i | V_j\right) = \frac{W_i + 1}{\sum W_j + B'} \qquad (2.17)$$

   W_i adalah nilai pembobotan *TF-IDF* kata i pada sentimen j, $\sum W_j$ merupakan jumlah seluruh bobot kata yang ada pada sentiment j dan B' merupakan jumlah bobot dari kata unik (tidak dikalikan oleh frekuensi). Nilai 1 pada persamaan (2.15) merupakan nilai *laplace smoothing*, *laplace smoothing* adalah metode yang digunakan untuk mencegah nilai nol pada algoritma *Naïve Bayes Classifier*, metode ini merupakan metode sederhana karena hanya menambahkan angka 1. Namun penggunaan metode memiliki performa yang cukup baik jika dibandingkan dengan metode *smoothing* lainya (Rahman dan Doewes, 2017)

2.8.2 Contoh Perhitungan *Naive Bayes Classifier* (NBC)

Misalkan akan dilakukan pengklasifikasian pada dokumen pada tabel 2.5 dimana dokumen d1 hingga d4 digunakan untuk membangun model, sedangkan dokumen d6 sebagai dokumen yang akan akan diklasifikasikan

Tabel 2.6 Contoh Kata Klasifikasi

Kalimat	Sentimen	Kode Dokumen
ppkm, percaya, sesal	Negatif	d1
cegah, sebar, covid, kapolsek, selbar, bagi, masker, gratis, masyarakat, pasar, surabrata	Positif	d2
vaksinasi, jaga, sehat	Positif	d3
tka, china, bandara, sultan, hasanuddin, ppkm, darurat	Negatif	d4
ppkm, level, panjang, jokowi	Netral	d5
ppkm, kaya, panjang	Akan di klasifikasikan	d6

1. Menghitung probabilitas sentimen

   Langkah awal adalah menghitung probabilitas setiap sentimen berdasarkan formula (2.15) adalah sebagai berikut.

$$P(V_{positif}) = \frac{docs_{positif}}{data\ latih} = \frac{2}{5} = 0{,}4$$

$$P(V_{netral}) = \frac{docs_{netral}}{data\ latih} = \frac{1}{5} = 0{,}2$$

$$P(V_{negatif}) = \frac{docs_{negatif}}{data\ latih} = \frac{2}{5} = 0{,}4$$

2. Menghitung probabilitas kata pada setiap sentimen

   Langkah selanjutnya adalah menghitung probabilitas setiap kata pada setiap sentimen berdasarkan formula (2.17) dan nilai pembobotan kata pada tabel 2.4 adalah sebagai berikut.

$$P(x_{ppkm}|V_{positif}) = \frac{W_i + 1}{\sum W_j + B'} = \frac{0 + 1}{3{,}39 + 41{,}01} = \frac{1}{44{,}40} = 0{,}022$$

$$P(x_{ppkm}|V_{netral}) = \frac{W_i + 1}{\sum W_j + B'} = \frac{0{,}305 + 1}{1{,}57 + 41{,}01} = \frac{1{,}305}{42{,}59} = 0{,}030$$

$$P(x_{ppkm}|V_{negatif}) = \frac{W_i + 1}{\sum W_j + B'} = \frac{0{,}618 + 1}{3{,}16 + 41{,}01} = \frac{1{,}581}{44{,}17} = 0{,}035$$

Maka, nilai probabilitas setiap kata pada setiap sentiment dapat dilihat pada tabel berikut.

Tabel 2.7 Perhitungan Probabilitas Kata Pada Sentimen

Kata	Sentimen		
	Positif	**Netral**	**Negatif**
ppkm	0,022	0,030	0,035
percaya	0,022	0,023	0,035
sesal	0,022	0,023	0,035
cegah	0,025	0,023	0,022
sebar	0,025	0,023	0,022
covid	0,025	0,023	0,022
kapolsek	0,025	0,023	0,022
selbar	0,025	0,023	0,022
bagi	0,025	0,023	0,022
masker	0,025	0,023	0,022
gratis	0,025	0,023	0,022
masyarakat	0,025	0,023	0,022
pasar	0,025	0,023	0,022
surabrata	0,025	0,023	0,022
vaksinasi	0,035	0,023	0,022
jaga	0,035	0,023	0,022
sehat	0,035	0,023	0,022
tka	0,022	0,023	0,028
china	0,022	0,023	0,028
bandara	0,022	0,023	0,028
sultan	0,022	0,023	0,028
hasanuddin	0,022	0,023	0,028
darurat	0,022	0,023	0,028
level	0,022	0,033	0,022
panjang	0,022	0,033	0,022
jokowi	0,022	0,033	0,022

Nilai *probabilitas* kata yang didapatkan pada setiap sentimen diatas, akan digunakan sebagai model *probabilistic.* Selanjutnya, akan digunakan sebagai acuan untuk melakukan prediksi sentimen pada data uji.

3. Perhitungan *probabilitas* sentimen pada data uji

   Langkah terakhir yang dilakukan adalah melakukan perhitungan $\hat{V}_{map}$, dengan menggunakan formula (2.13) sebagai berikut

$$P(d6|V_{positif}) = log\hat{P}(V_{positif}) + \sum log\,\hat{P}(x_i|V_{positif})$$

$$= log(0{,}4) + log(0{,}022) + log(0{,}022) + log(\,0{,}022)$$

$$= \text{-}5{,}37$$

$$P(d6|V_{positif}) = log\hat{P}(V_{netral}) + \sum log\,\hat{P}(x_i|V_{netral})$$

$$= log(0{,}2) + log(0{,}030) + log(0{,}023) + log(\,0{,}023)$$

$$= \text{-}5{,}49$$

$$P(d6|V_{negatif}) = log\hat{P}(V_j) + \sum log\,\hat{P}(x_i|V_j)$$

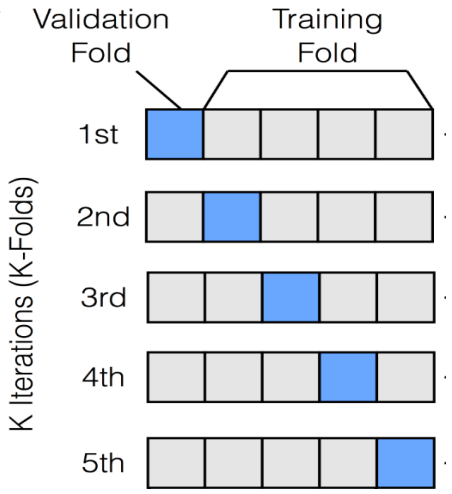
$$= log(0{,}4) + log(0{,}035) + log(0{,}041) + log(\,0{,}041)$$

$$= \text{-}5{,}16$$

Maka berdasarkan perhitungan $\hat{V}_{map}$, diketahui bahwa nilai pada sentimen negatif (-5,16) memiliki nilai yang lebih tinggi jika dibandingkan sentimen positif (-5,37) dan sentimen netral (-5,49). Maka, dapat disimpulkan bahwa hasil pengklasifikasian dengan metode *Naive Bayes Classifier* dari dokumen d6 adalah sentiment negative.

2.9 *K-Fold Cross Validation*

Cross Validation merupakan sebuah prosedur untuk mengevaluasi dan memvalidasi kemampuan generatif dari algoritma yang dilatih. *K-Fold Cross Validation* merupakan salah satu metode *Cross Validation* yang sering digunakan. K-*Fold Cross Validation* adalah sebuah metode penngelompokan data menjadi *K* kelompok data. Selanjutnya, setiap kelompok yang terbentuk akan diuji dengan model yang digunakan. Sehingga sebuah model dapat teruji dalam input data yang beragam dari proses yang dilakukan. Adapaun kerangka kerja dari proses *K-Fold Cross Validation* dapat dilihat pada gambar 2.3



Gambar 2. 3 Kerangka Kerja *K-Fold Cross Validation*

Sumber: *(https://androidkt.com/pytorch-k-fold-cross-validation-using-dataloader-and-sklearn/)*

Pada gambar 2.2 dapat di lihat, langkah pertama adalah menentukan banyaknya *fold* (k) yang akan digunakan, selanjutnya pada setiap *fold* akan ditentukan kelompok data latih berjumlah $(k-1)$ yang digunakan untuk membangun model dan satu kelompok sisanya digunakan untuk melakukan validasi dari model yang telah dibangun, hasil evaluasi dari setiap *fold* nantinya akan dikombinasikan dan ditentukan nilai rata-rata gabungan sebagai hasil akhir dari performa model.

2.10 *Confusion Matriks*

Untuk mengukur performa dari sebuah model, akan digunakan pengukuran evaluasi performa (PEM) atau *confusion matriks. Confusion Matriks* digunakan pada data latih dan data uji untuk mengetahui dan mengevaluasi performa dari model yang sudah di buat. Confusion Matriks merupakan sebuah tabel dengan yang menampilkan sejumlah angka yang dapat dijadikan acuan dalam melihat performa model (Agustin dkk., 2019).

Tabel 2. 8 Tabel *Confusion Matriks*

Aktual	Prediksi			Total
	Positif	**Netral**	**Negatif**	
Positif	TP	Error	Error	Kelas Pos
Netral	Error	TP	Error	Kelas Net
Negatif	Error	Error	TP	Kelas Neg
Total	Prediksi Pos	Prediksi Net	Prediksi Neg	

Acuration atau akurasi adalah ketepatan sebuah model/alat dalam melakukan prediksi jika di bandingkan acuan lain. Akurasi merupakan salah satu perhitungan yang ada di dalam *confusion matriks* Adapun formula yang digunakan untuk menghitung akurasi sebagai berikut:

$$acc = \frac{\sum TP}{\sum Total\ Kelas} \quad (2.18)$$

2.11 Penelitian Terdahulu

Telah banyak penelitian yang dilakukan untuk mengetahui sentimen masyarakat bedasarkan media sosial. Berikut ini adalah beberapa penelitian tentang analisis sentimen masyarakan di media sosial yang penulis jadikan rujukan.

Tabel 2. 9 Penelitian Terdahulu

No	Nama (Tahun Penelitian)	Judul Penelitian	Variabel Penelitian	Kesimpulan
1	Ali Imron (2019)	Analisis Sentimen Terhadap Tempat Wisata di Kabupaten Rembang Menggunakan Metode *Naïve Bayes Classier*	Komentar dari situs Tripadvisor dan aplikasi Facebook	• Hasilnya adalah algoritma Naive Bayes Classifier terbukti algoritma yang akurat nilai akurasi sebesar 0.828 atau 82.8%..
2	Balya (2019)	Analisis Sentimen Pengguna Youtube di	Commentar di youtube pada *review* konten	• Preprocessing pada komentar terbukti efektif

No	Nama (Tahun Penelitian)	Judul Penelitian	Variabel Penelitian	Kesimpulan
		Indonesia Pada Review Smartphone Menggunakan Naïve bayes	*smartphone*	menghasilkan kalimat yang penting terhadap proses analisis sentimen. • Pengujian analisis sentimen pada komentar menggunakan *Gaussian Naïve Bayes* menghasilkan akurasi sebesar 73% sementara dengan menggunakan *Multinomial Naïve Bayes* menghasilkan akurasi sebesar 81%.
3	Akhmad Muzaki , Arita Witanti (2021)	Sentimen Analisis Masyarakat di Twitter Terhadap Pilkada 2020 Ditengah Pandemic Covid-19 Dengan Metode *Naïve Bayes Classifier*	Percakapan di Twitter yang berhubungan dengan pilkada 2020	• analisis sentimen dapat digunakan untuk mengetahui sentimen masyarakat khususnya Netizen Twitter terhadap pelaksanaan pilkada 2020 ditengah pandemik COVID-19.

BAB III

METODOLOGI PENELITIAN

3.1 Data

Data yang digunakan pada penelitian ini adalah data sekunder yang diperoleh dari PT. Ivonesia Solusi Data (Ivosight), data tersebut merupakan data *tweet* dari masyarakat yang membahas tentang penerapan kebijakan PPKM pada media sosial *Twitter* yang diambil dengan metode *Crawling Data* dalam jangka *waktu* 1 Juli 2021 hingga 31 Oktober 2021. Dasar pengambilan waktu ini dikarenakan pada bulan Juli merupakan puncak peningkatan kasus konfirmasi positif di Indonesia serta penerapan PPKM darurat. Total data yang tersedia berjumlah 1.725.627 *tweet*

3.2 Prosedur dan Analisis Data

Berikut ini adalah Langkah-langkah analisis data yang dilakukan:

1) Menampilkan deskripsi dari data yang digunakan menggunakan diagram atau tabel.
2) Mengambil sampel dari data yang tersedia, melakukan proses *cleansing* dan pemberian label sentimen pada data yang layak digunakan.
    a) Proses pengambilan sampel dari populasi menggunakan metode *simple random sampling,* Dengan penentuan jumlah sampel minimal menggunakan formula berdasarkan persamaan (2.1) dengan tingkat *error* sampling 5% berjumlah 400 sampel. Sedangkan jumlah sampel yang diambil sebanyak 5000 sampel.
    b) *Data cleansing* adalah proses pembersihan data dari sampel yang terpilih, sebuah *tweet* akan dikatakan layak untuk dianalisis jika pembahahan tersebut sesuai dengan topik PPKM, sementara itu jika tidak sesuai data *tweet* tersebut akan dihapus dari sampel.
    c) Proses pemberian label awal sentimen tweet pada data dilakukan secara objektif berdasarkan amatan dari peneliti. Adapun batasan umum pemberian label sentimen menurut Samsir et al (2021) adalah sebagai berikut:

- Sentimen negatif, merupakan label yang diberikan pada *tweet* yang menyatakan keluhan, kalimat sindiran, kritik, dan cerminan emosi negatif seperti amarah, kesal, dan kecewa.
- Sentimen positif, merupakan label yang diberikan kepada *tweet* yang menyatakan pujian, saran, masukan, dan cerminan emosi positif seperti puas, senang, dan bahagia
- Sentimen netral, merupakan label yang di berikan kepada *tweet* yang berisi informasi, berita atau respon ketidakberpihakan terhadap topik PPKM.

Tabel 3.1 Contoh hasil dari proses *labelling*

Tweet	**Label**
Hasil putusan MA sah dan PPKM Pulihkan Bangsa https://t.co/1Fjgi9V6xi	Positif
@sutanmangarahrp Dulu sebelum pandemi aku juga ada 2 cabang dagang ayam crispi, namun karena kondisi perekonomian yg tjd akhirnya hanya mampu bertahan 1 lapak, itupun duaÂ²nya sempat mati selama beberapa bulan, & kini dg bermodal beberapa lembar 100rb aku coba bangkitkan kembali, mohon doanya ðŸ¤²	Negatif
Silahkan PPKM mau level 2024 juga gpp,\nYg penting rakyat butuh makan,\nButuh kecukupan,\nHarus bisa menuhin kebutuhan semua rakyat,\nTerutama rakyat bawah yg susah\n\nJika tak mampu memenuhi kebutuhan rakyat, sebaiknya\n#2021MakzulkanPresiden #2021MakzulkanPresiden\nW	Negatif
RT @geloraco: 20 TKA Asal China Tiba di Bandara Sultan Hasanuddin saat PPKM Darurat\nhttps://t.co/hl9zA7eLcw	Netral
Menteri Dalam Negeri Tito Karnavian kembali menerbitkan instruksi terbaru tentang pemberlakuan pembatasan kegiatan masyarakat (PPKM) untuk wilayah Jawa dan Bali yakni Inmendagri Nomor 42/2021.\n\njatim.antaranews.com/berita/524509/â€¦	Netral
udah lah yg kmaren sok sok an ribut nyalahin ina itu..ngebahas lokdown, psbb, dll dan mmg bukan ahli di bidang nya... sekarang mending ambil sabun trus ke kamar mandi... ðŸ¤£	Positif

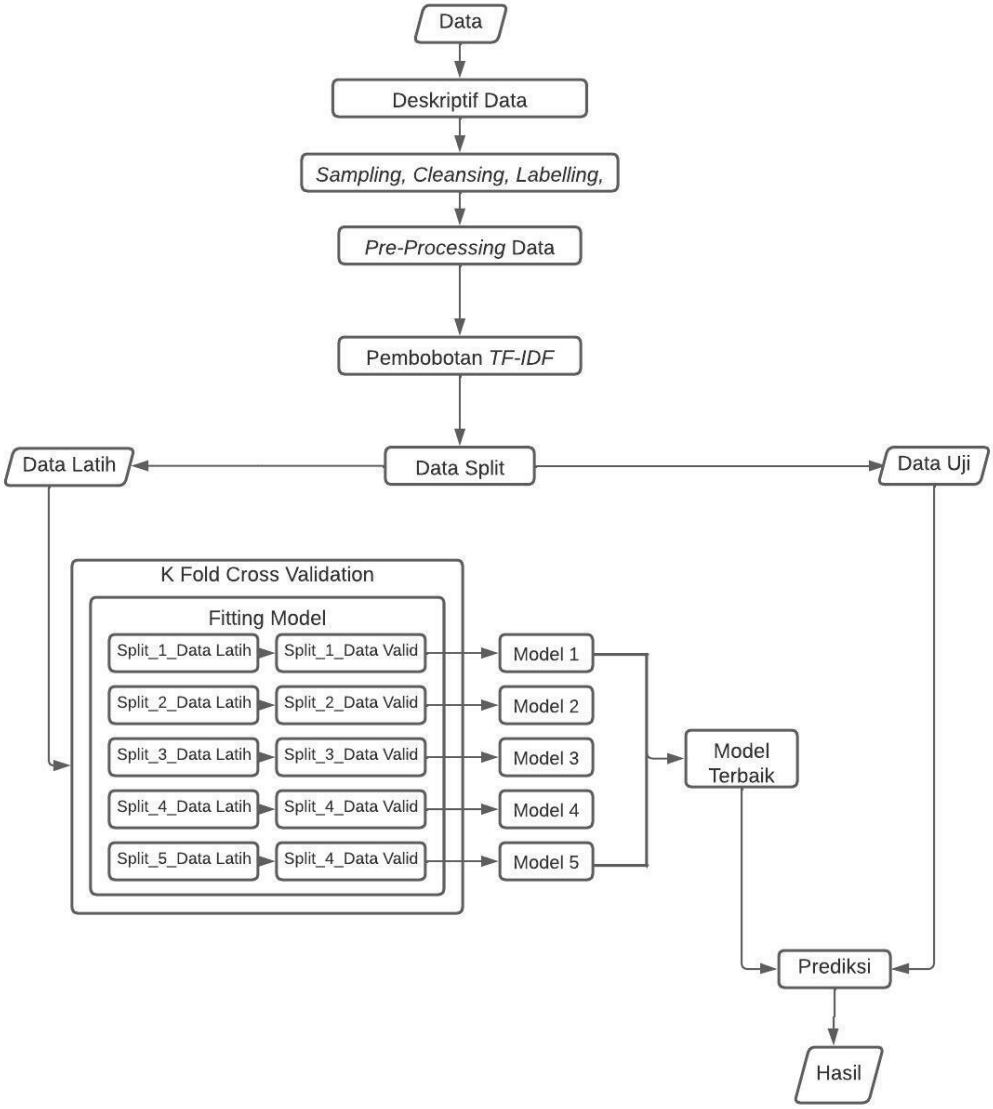
3) Melakukan Preprocessing data teks, adapun proses yang dilakukan sebagai berikut:

a) *Cleansing Text,* yaitu proses pembersihan data Twitter yang akan digunakan
b) *Case Folding,* yaitu proses merubah text menjadi suatu bentuk yang standar
c) *Tokenezing,* yaitu proses merubah kalimat menjadi bentuk satuan kata dengan *spasi* dan *enter* sebagai pemisah dari setiap kata
d) *Stemming* merupakan sebuah proses pengambilan kata dasar dari setiap kata yang akan digunakan dalam penelitian
e) *Normalization* adalah proses penyamaan kata yang memiliki makna sama, namun penulisan berbeda,

4) Merubah teks menjadi pembobotan kata, dengan menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF) pada formula (2.4)
5) Membagi data menjadi dua bagian. Data latih untuk membangun model dan data uji untuk melakukan validitas atas model yang dibagun. Pembagian data menggunakan persentasi 80% untuk data latih dan 20% untuk data uji. Hal ini disesuaikan dengan jumlah data yang digunakan.
6) Membangun validitas dataset menggunakan *K-Fold Cross validation,* menurut sabrani, dkk (2020) jumlah *K* yang di gunakan adalah validitas dataset adalah 5 *fold* (lima dataset)
7) Membangun model menggunakan data latih pada setiap dataset dengan menggunakan *Multinomial Naive Bayes*, menurut Sabrani, dkk (2020) Langkah yang dilakukan adalah
   a) Menghitung *prior* probability berdasarkan formula (2.14)
   b) Menghitung probability kata *i* dalam sentimen *j (likelihood)* bedasarkan formula (2.14)
8) Melakukan penglkasifikasian teks pada data uji bedasarkan data latih, dengan menggunakan formula (2.15)
9) Melakukan perhitungan evaluasi model *Multinomial Naive Bayes* dengan langkah berikut.
   a) Menghitung nilai akurasi dari setiap model dengan menggunakan formula (2.16)

b) Melakukan proses menentukan model terbaik pada dataset untuk mendaptkan hasil final dari evaluasi model yang dilakukan

10) Menampilkan *wordcloud, wordcloud* adalah representasi visual dari frekuensi kata. Semakin umum istilah yang muncul dalam teks yang dianalisis, maka semakin besar kata yang muncul dalam visualisasi(Singh Carter & Atenstaedt Rob, 2012)



Gambar 3.1 *Flowchart* Analsis Data

BAB IV

HASIL DAN PEMBAHASAN

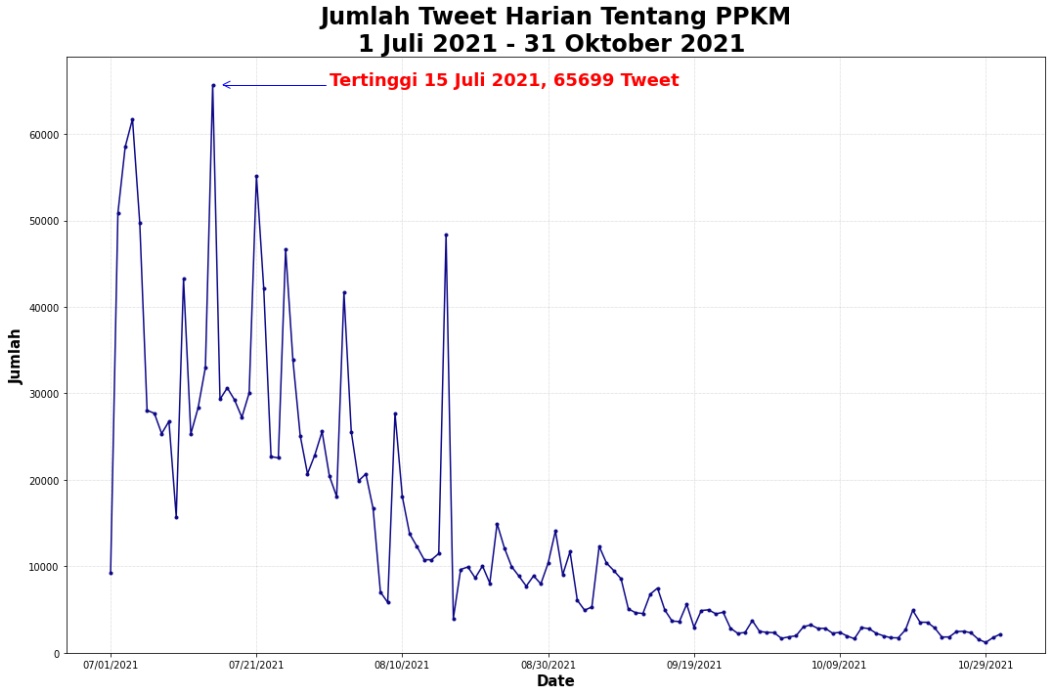
4.1 Deskripsi Data *Tweet* PPKM

Pemberlakukan pembatasan kegiatan masyarakat atau PPKM merupakan sebuah kebijakan publik yang diterapkan oleh pemerintah Indonesia. Penerapan PPKM merupakan bentuk respon pemerintah Indonesia untuk mencegah penyebaran pandemi Covid-19. PPKM pertama kali diterapkan pada bulan Januari 2021. Kebijakan ini diterapkan untuk mengganti kebijakan sebelumnya yaitu Pembatasan Sosial Berskala Besar (PSBB) yang dianggap kurang berhasil. Penerapan kebijakan PPKM mendapatkan respon beragam dari masyarakat. Banyak masyarakat yang setuju dengan kebijakan PPKM ini dikarenakan dianggap mampu menekan laju pertumbuhan Covid-19. Sementara itu, disisi lain terdapat banyak pula masyarakat yang tidak setuju dengan kebijakan ini, karena dianggap mengganggu kegiatan dan perekonomian masyarakat.

Realsi masyarakat terhadap penerapan kebijakan PPKM juga terjadi di media sosial. Twitter merupakan media sosial yang paling banyak digunakan masyarakat Indonesia untuk menyampaikan respon mereka terhadap penerapan kebijakan PPKM.

Tabel 4.1 Statistik Deskriptif Interaksi Harian Masyarakat Pada Media Sosial Twitter Terkait Penerapan PPKM

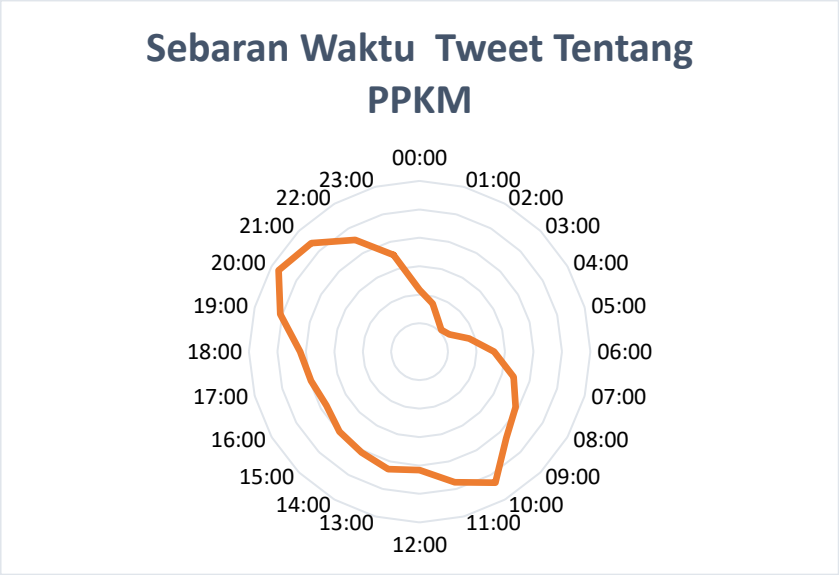
Statistik	Interaksi Masyarakat (Dalam Hari)
Minimum	1.195
Maksimum	65.699
Rata-Rata	14.030
Jumlah Hari	123
Total Interaksi	1.725.627



Gambar 4.1 Grafik Jumlah *Tweet* Harian Tentang PPKM

Berdasarkan Tabel 4.1 dan Gambar 4.1 dapat dilihat tercatat dalam jangka waktu 1 Juli 2021 hingga 31 Oktober 2021 atau selama 123 hari terdapat 1.725.627 *tweet* interaksi masyarakat tentang penerapan kebijakan PPKM pada media sosial Twitter. Diketahui bahwa rata rata setiap harinya terdapat 14000 *tweet* yang membahas tentang kebijakan ini. Sementara itu, puncak interaksi terjadi pada tanggal 15 Juli 2021 dengan 65000 *tweet*, hal ini dikarenakan terjadi lonjakan kasus konfirmasi positif covid-19 di Indonesia. Sedangkan yang terendah, terjadi pada tanggal 29 Oktober 2021. Dari gambar tersebut juga dapat disimpulkan bahwa pembahasan penerapan kebijakan PPKM mengalami *trend* yang menurun sejalan dengan menurunnya kasus covid-19 di Indonesia. Pada bulan Juli 2021 menjadi puncak interaksi masyarakat terhadap penerapan kebijakan PPKM. Hal ini dikarenakan pada bulan tersebut pemerintah menerapkan PPKM darurat sebagai respon terhadap meningkatnya penyebaran covid-19 varian Delta.

Pada gambar 4.2 di bawah ini, dapat dilihat seberan waktu interaksi tentang PPKM pada media sosial twitter, berdasarkan gambar diatas dapat disimpulkan bahwa puncak interaksi terjadi pada jam 19.00 WIB hingga 22.00 WIB dan jam 10.00 WIB hingga 11.00 WIB.



Gambar 4.2 Sebaran Waktu dalam WIB Interaksi Harian Masyarakat Pada Media Sosial Twitter Terkait Penerapan PPKM

4.2 *Sampling*, *Cleansing* dan *Labelling* Data *Tweet* PPKM

Jumlah data yang besar membuat peneliti memutuskan untuk mengambil sampel dari total data yang dimiliki. Metode pengambilan sampel yang digunakan oleh peneliti adalah *Simple Random Sampling*. Berdasarkan formula (2.1) jumlah minimum sampel dengan jumlah populasi diketahui dan taraf kesalahan pengambilan sampel sebesar 5% adalah sebanyak 400 sampel. Maka dari itu peneliti menentukan jumlah sampel awal yang berjumlah 5000 data *tweet*, sehingga telah memenuhi batas minimum sampel. Selanjutnya, dari sampel yang terpilih akan dilakukan *cleansing* data. *cleansing* data bertujuan untuk membuang *tweet* yang tidak mengandung sentimen terhadap penerapan kebijakan PPKM. Total sampel yang telah melalui proses *cleansing* adalah berjumlah 4354 *tweet.*

Data yang telah melalui proses cleansing kemudian akan diberikan label sentimen secara manual. Pemberian label pada *tweet* dilakukan secara objektif dengan batasan umum pemberian label mengacu pada Samsir et al (2021). Pengklasifikasikan *tweet* dilakukan kedalam tiga kategori, yaitu sentimen positif, sentimen negatif dan sentimen netral.

4.3 *Pre-Processing* Data *Tweet*

Sebelum masuk kedalam tahap analisis, data twitter harus melalui *Pre-Processing Data*, tujuanya adalah untuk menyeleksi data teks agar menjadi lebih terstruktur dan dapat lebih bermakna. Adapaun proses yang dilakukan adalah sebagai berikut.

4.3.1 *Text Cleansing*

Text Cleansing adalah proses membersihkan dan menyeragamkan setiap teks agar memiliki struktur yang sama. Adapun hal yang dilakukan pada proses *Text Cleansing* adalah menghapus tab, baris baru, *emoticon*, bahasa China, *mention*, *link*, angka, tanda baca, spasi yang berlebihan pada awal dan akhir kalimat, serta penggunaan huruf yang berlebihan pada sebuah kata. Pada saat melakukan *Text Cleansing* peneliti menggunakan bantuan *package Re* pada *python* yang digunakan untuk mencari sebuah *string* bedasarkan spesifikasi tertentu. Berikut merupakan hasil *Text Cleansing* pada tiga kalimat *tweet* tentang PPKM

Tabel 4.2 Hasil Proses *Text Cleansing*

Sebelum	Sesudah
RT @fsskroeppreal: PPKM - Pernah percaya kemudian menyesal.~ **https://t.co/GT6k6wPMFb**	RT PPKM Pernah percaya kemudian menyesal
RT @PolsekPupuan:Cegah Penyebaran Covid-19, Kapolsek Selbar bagikan masker gratis kepada masyarakat di Pasar Surabrata. #masker #pakaimasker #prokes #ppkm #BersatuLawanCorona #polrestabanan **https://t.co/d0wDwX2VTi**	RT Cegah Penyebaran Covid-19 Kapolsek Selbar bagikan masker gratis kepada masyarakat di Pasar Surabrata. masker pakaimasker prokes pkm BersatuLawanCorona polrestabanan
@Humas_SekMaos Pemerintah merevisi aturan Pemberlakuan Pembatasan Kegiatan Masyarakat (PPKM) Darurat yang diberlakukan sejak 3 Juli 2021.\n#DisiplinSuksesPPKM\nLebih Baik Dirumah	SekMaos Pemerintah merevisi aturan Pemberlakuan Pembatasan Kegiatan Masyarakat PKM Darurat yang diberlakukan sejak Juli DisiplinSuksesPKM Lebih Baik Dirumah

4.3.2 *Case Folding*

Case Folding merupakan proses membuat semua huruf pada kalimat menjadi huruf kecil. Tujuanya adalah menyeragamkan setiap kata yang sama, namun memiliki penulisan hurup kapital yang berbeda. Proses *Case Folding* menggunakan fungsi *lower()* yang merupakan salah satu fungsi bawaan pada bahasa pemrograman *python*. Berikut merupakan hasil *Case Folding* pada tiga kalimat awal *tweet* tentang PPKM

Tabel 4.3 Hasil Proses *Case Folding*

Sebelum	Sesudah
RT PPKM Pernah percaya kemudian menyesal	rt ppkm pernah percaya kemudian menyesal
RT Cegah Penyebaran Covid-19 Kapolsek Selbar bagikan masker gratis kepada masyarakat di Pasar Surabrata. masker pakaimasker prokes pkm BersatuLawanCorona polrestabanan	rt cegah penyebaran covid-19 kapolsek selbar bagikan masker gratis kepada masyarakat di pasar surabrata. masker pakaimasker prokes pkm bersatulawancorona polrestabanan
SekMaos Pemerintah merevisi aturan Pemberlakuan Pembatasan Kegiatan Masyarakat PKM Darurat yang diberlakukan sejak Juli DisiplinSuksesPKM Lebih Baik Dirumah	sekmaos pemerintah merevisi aturan pemberlakuan pembatasan kegiatan masyarakat pkm darurat yang diberlakukan sejak juli disiplinsuksespkm lebih baik dirumah

4.3.3 *Tokenizing*

Tokenizing merupakan proses pengambilan sebuah kata pada suatau kalimat. tahap *Tokenizing* pengambilan kata didasari oleh *spasi* dan *enter* sebagai pemisah dari setiap kata. Dalam melakukan proses pengambilan kata peneliti menggunakan *package NLTK* dengan fungsi *word_tokenize*, package *NLTK* digunakan pada *python* untuk membantu memahami bahasa linguistik. Berikut merupakan hasil *Tokenizing* pada tiga kalimat awal *tweet* tentang PPKM.

Tabel 4.4 Hasil Proses *Tokeneizing*

Sebelum	Sesudah
rt ppkm pernah percaya kemudian menyesal	'rt', 'ppkm', 'pernah', 'percaya', 'kemudian', 'menyesal'
rt cegah penyebaran covid-19	'rt', 'cegah', 'penyebaran', 'covid',

Sebelum	Sesudah
kapolsek selbar bagikan masker gratis kepada masyarakat di pasar surabrata. masker pakaimasker prokes pkm bersatulawancorona polrestabanan	'kapolsek', 'selbar', 'bagikan', 'masker', 'gratis', 'kepada', 'masyarakat', 'di', 'pasar', 'surabrata', 'masker' 'pakaimasker' 'prokes' 'pkm 'bersatulawancorona' 'polrestabanan'
sekmaos pemerintah merevisi aturan pemberlakuan pembatasan kegiatan masyarakat pkm darurat yang diberlakukan sejak juli disiplinsuksespkm lebih baik dirumah	'sekmaos' ,'pemerintah', 'merevisi' 'aturan' 'pemberlakuan', 'pembatasan', 'kegiatan', 'masyarakat', 'pkm, 'darurat, 'yang' 'diberlakukan', 'sejak', 'juli', 'disiplinsuksespkm', 'lebih', 'baik' 'dirumah'

4.3.4 *Normalization*

Normalization adalah proses mengganti kata gaul (*slang*) menjadi kata formal tanpa mengubah makna dari kata tersebut. misalkan kata ‘banget’ yang disingkat menjadi ‘bgt’ akan dikembalikan maknanya menjadi ‘banget’. Pada proses *Normalization* peneliti mendapatkan list kata gaul tersebut dari publikasi Salsabila, Ali, Yosef, and Ade tentang Colloquial Indonesian Lexicon dengan jumlah kosakata sebanyak 15456 kosakata. Berikut merupakan hasil *Normalization* pada tiga kalimat awal *tweet* tentang PPKM.

Tabel 4.5 Hasil Proses *Normalization*

Sebelum	Sesudah
'rt', 'ppkm', 'pernah', 'percaya', 'kemudian', 'menyesal'	'rt', 'ppkm', 'pernah', 'percaya', 'kemudian', 'menyesal'
'rt', 'cegah', 'penyebaran', 'covid', 'kapolsek', 'selbar', 'bagikan', 'masker', 'gratis', 'kepada', 'masyarakat', 'di', 'pasar', 'surabrata', 'masker' 'pakaimasker' 'prokes' 'pkm 'bersatulawancorona' 'polrestabanan'	'rt', 'cegah', 'penyebaran', 'covid', 'kapolsek', 'selbar', 'bagikan', 'masker', 'gratis', 'kepada', 'masyarakat', 'di', 'pasar', 'surabrata', 'masker', 'pakaimasker', 'prokes', 'pkm', 'bersatulawancorona', 'polrestabanan'

Sebelum	Sesudah
'sekmaos' ,'pemerintah', 'merevisi' 'aturan' 'pemberlakuan', 'pembatasan', 'kegiatan', 'masyarakat', 'pkm, 'darurat, 'yang' 'diberlakukan', 'sejak', 'juli', 'disiplinsuksespkm', 'lebih', 'baik' 'dirumah'	'sekmaos', 'pemerintah', 'merevisi', 'aturan', 'pemberlakuan', 'pembatasan', 'kegiatan', 'masyarakat', 'pkm', 'darurat', 'yang', 'diberlakukan', 'sejak', 'juli', 'disiplinsuksespkm', 'lebih', 'baik', 'dirumah'

4.3.5 *Stemming*

Stemming merupakan sebuah proses pengambilan kata dasar dari setiap kata. Dalam melakukan proses *stemming* kata peneliti menggunakan *package Sastrawi* pada *python.* Berikut merupakan hasil *Stemming* pada tiga kalimat awal *tweet* tentang PPKM.

Tabel 4.6 Hasil Proses *Stemming*

Sebelum	Sesudah
'rt', 'pkm', 'pernah', 'percaya', 'kemudian', 'menyesal'	'rt', 'ppkm', 'pernah', 'percaya', 'kemudian', 'sesal'
'rt', 'cegah', 'penyebaran', 'covid', 'kapolsek', 'selbar', 'bagikan', 'masker', 'gratis', 'kepada', 'masyarakat', 'di', 'pasar', 'surabrata', 'masker', 'pakaimasker', 'prokes', 'pkm', 'bersatulawancorona', 'polrestabanan'	'rt', 'cegah', 'sebar', 'covid', 'kapolsek', 'selbar', 'bagi', 'masker', 'gratis', 'kepada', 'masyarakat', 'di', 'pasar', 'surabrata', 'masker', 'pakaimasker', 'prokes', 'pkm', 'bersatulawancorona', 'polrestabanan'
'sekmaos', 'pemerintah', 'merevisi', 'aturan', 'pemberlakuan', 'pembatasan', 'kegiatan', 'masyarakat', 'pkm', 'darurat', 'yang', 'diberlakukan', 'sejak', 'juli', 'disiplinsuksespkm', 'lebih', 'baik', 'dirumah'	'sekmaos', 'perintah', 'revisi', 'atur', 'laku', 'batas', 'giat', 'masyarakat', 'pkm', 'darurat', 'yang', 'laku', 'sejak', 'juli', 'disiplinsuksespkm', 'lebih', 'baik', 'rumah'

4.4 Pembobotan *Tweet* menggunakan *TF-IDF*

Term frequency-inverse document frequency (TF-IDF) merupakan ukuran statistik yang mengevaluasi seberapa relevan sebuah kata pada sebuah dokumen dan dalam kumpulan dokumen. *TF-IDF* sering digunakan pada algoritma *machine learning* untuk *Natural Language Processing (NLP). TF-IDF* meningkatkan

secara proporsional dengan berapa kali sebuah kata muncul dalam dokumen tetapi diimbangi dengan jumlah dokumen yang mengandung kata tersebut. Sehingga, kata-kata yang umum di setiap dokumen, akan memiliki bobot yang rendah meskipun mungkin muncul berkali-kali, karena mereka tidak terlalu berarti bagi dokumen itu secara khusus. Sementara itu, jika sebuah muncul berkali-kali dalam sebuah dokumen, sementara tidak muncul berkali-kali di dokumen lain, itu akan membuat bobot kata tersebut tinggi. Untuk melakukan perhitungan *TF-IDF* peneliti menggunakan *package sklearn.feature_extraction.text* dengan fungsi yang digunakan adalah *TfidfVectorizer*. Berikut merupakan hasil perhitungan *TF-ID.*

Tabel 4.7 Hasil Perhitungan *TF-IDF*

Dokumen	**Kata**	***TF-IDF***
0	6550	0,551
0	3315	0,486
0	5369	0,477
0	5397	0,448
0	5499	0,095
0	6158	0,146
1	5617	0,290
…	…	…
4353	7861	0,079
4353	1370	0,085
4353	613	0,136
4353	5499	0,037

Pada tabel hasil perhitungan *TF-IDF* diatas, kolom dokumen merupakan indeks urutan dari *tweet* pada data yang berjumlah 4354 dokumen. Sementara itu, kolom kata merupakan indeks urutan kata yang terdapat dalam data, jumlah kata yang terdapat dalam data berjumlah 7933 kata. Setiap kata yang yang terdaftar akan memiliki nilai bobot tersendiri pada setiap dokumen yang terdapat pada data. Kolom *TF-IDF* merupakan hasil nilai perhitungan *TF-IDF* yang telah distandarisasikan menggunakan *L2 normalization.*

Selanjutnya, hasil perhitungan *TF-IDF* akan dibentuk matriks pembobotan kata berukuran 7933 x 4354 yang berisi nilai *TF-IDF* setiap kata pada setiap dokumen. Matriks yang dihasilkan akan digunakan untuk dalam melakukan analisis sentimen.

4.5 Data *Split*

Selanjutnya, setelah didapatkan matriks pembobotan kata berdasarkan perhitungan *TF-IDF,* data akan dibagi menjadi dua bagian, yaitu data latih dan data uji. Data latih akan digunakan untuk membangun model, sedangkan data uji akan digunakan untuk mengevaluasi model. Pembagian data menggunakan aturan proporsi 80% untuk data latih dan 20% untuk data uji. Sehingga, jumlah data yang terdapat pada data latih berjumlah 3483 data sedangkan pada data latih berjumlah 871 data. Pada saat proses pembagian data peneliti menggunakan *package sklearn.model_selection* dengan fungsi *train_test_split.*

Setelah data dibagi menjadi data latih dan data uji. Pada data latih, akan dilakukan proses *cross validation* menggunakan metode *K-fold cross validation*. *K-fold cross validation* merupakan suatu metode tambahan dari teknik *data mining* yang bertujuan untuk memperoleh hasil akurasi yang maksimal, Langkah pertama adalah menentukan jumlah *fold* sebanyak 5, Selanjutnya pada setiap *fold* akan ditentukan kelompok data latih berjumlah 4 yang digunakan untuk membangun model dan satu kelompok sisanya digunakan untuk melakukan validasi, setiap *fold* akan memiliki data latih dan data validasi yang berbeda. Hasil evaluasi dari setiap kelompok nantinya akan dilihat apakah model tersebut telah stabil dalam *dataset* yang berbeda. Pada proses ini peneliti menggunakan bantuan *package sklearn.model_selection* dengan fungsi *cross_val_score.* Performa model terbaik yang didapatkan pada proses *cross validation* akan digunakan untuk melakukan prediksi pada data latih.

4.6 *Modelling*

4.6.1 Akurasi Pada Cross Validation

Hal yang dilakukan pertama proses modelling adalah menentukan model terbaik yang akan digunakan dalam melakukan prediksi pada data uji. Pada 5 *fold,*

akan dibangun masing-masing sebuah model pada data latih, yang kemudian akan di evaluasi dengan pasangan data valid dari setiap *fold*. Metode yang digunakan dalam membangun model adalah *Multinominal Naive Bayes Classifier*. Untuk memudahkan proses peneliti menggunakan *package sklearn.naive_bayes* dengan fungsi *MultinomialNB.*

Tabel 4.8 Hasil Akurasi Pada *Cross Validation*

Fold	**Akurasi**
Ke-1	0,7144
Ke-2	0,7001
Ke-3	0,6857
Ke-4	0,6925
Ke-5	0,6939

Pada tabel 4.9 dapat dilihat hasil evaluasi model terbaik pada setiap *fold* dalam proses c*ross validation.* Perhitungan evaluasi model yang digunakan adalah akurasi dengan menggunakan *package sklearn.metrics* dengan fungsi *accuracy_score*. Hasilnya diketahui bahwa nilai akurasi tertinggi sebesar 0,714. Dan dari hasil tersebut, dapat di simpulkan bahwa model sudah cukup stabil sehingga dapat digunakan melakukan prediksi pada data latih.

4.6.2 Probabilitas Sentimen *(Prior)*

Setalah didapatkan model terbaik. Selanjutnya, akan dilihat probabilitas setiap sentimen pada model tersebut atau *prior information. prior information* merupakan merupakan sebuah nilai probabilitas yang bersifat subjektif.

Tabel 4.9 Nilai *Logaritma Prior Information*

Sentimen	**Jumlah**	**Nilai *Logaritma Prior***
Positif	1100	-1.1525
Negatif	1285	-0.9971
Netral	1098	-1.1544

$$log\hat{P}(V_{positif}) = \frac{docs_{positif}}{data\ latih} = \frac{1100}{3483} = -1.152$$

$$log\hat{P}(V_{negatif}) = \frac{docs_{negatif}}{data\ latih} = \frac{1285}{3483} = -0{,}997$$

$$log\hat{P}(V_{netral}) = \frac{docs_{netral}}{data\ latih} = \frac{1100}{3483} = -1.154$$

Dari tabel 4.10 dapat dilihat hasil nilai *Prior Information* dari setiap sentimen bedasarkan formula (2.15). Hasil perhitungan nilai *Prior Information* pada tabel di atas, merupakan nilai *logaritma natural* dari formula yang digunakan.

4.6.3 *Sample Information (Likelihood Function)*

Nilai *sample Information* merupakan nilai yang didapatkan dari perhitungan probabilitas setiap sampel kata pada setiap sentimen. *sample information* akan dihitung dengan fungsi *likelihood* dengan asumsi bahwa sampel diestimasikan berdistribusi *multinomial.*

Tabel 4.10 Nilai Bobot Kata Pada Setiap Sentimen

Kata	Negatif	Positif	Netral
aa	0	0	0,3046
abadi	0	1.171	0
abai	1,346	0	2,0645
abang	0,3165	0	0,3498
…	…	…	…
zonasi	0	0	0,2575
zone	0	0,3105	0

Tabel 4.11 merupakan hasil perhitungan nilai bobot setiap kata pada setiap sentimen, nilai ini nantinya akan di gunakan dalam perhitungan formula (2.17). Selain itu, diketahui jumlah total bobot kata untuk setiap sentimen adalah 3523,21 untuk sentiment positif, 4794,04 untuk sentimen negatif dan 4292,50 untuk sentimen netral. Sementara itu, jumlah bobot dari kata unik (tidak dikalikan oleh frekuensi) adalah 3173,19. Berikut merupakan contoh perhitungan kata pada setiap sentimen dengan menggunakan formula (2.17)

$$P(x_{aa}|V_{positif}) = \frac{W_i + 1}{\sum W_j + B'} = \frac{0 + 0.4}{3523{,}21 + 3173{,}19} = -9{,}725$$

$$P(x_{aa}|V_{netral}) = \frac{W_i + 1}{\sum W_j + B'} = \frac{0{,}3046 + 0.4}{4292{,}50 + 3173{,}19} = -9.265$$

$$P(x_{aa}|V_{negatif}) = \frac{W_i + 1}{\sum W_j + B'} = \frac{0 + 0.4}{4794{,}04 + 3173{,}19} = -9.899$$

Tabel 4.11 Nilai *Logaritma Sample Information*

Kata	Negatif	Positif	Netral
aa	-9.899	-9.725	-9.268
abadi	-9.899	-8,356	-9.834
abai	-8.425	-9.725	-8.016
abang	-9.289	-9.725	-8.872
…	…	…	…
zonasi	-9.899	-9.725	-9.337
zone	-9.899	-9.151	-9.834

Dari tabel 4.11 dapat dilihat hasil perhitungan nilai *Sample Information* pada setiap kata dalam setiap sentiment. Hasil yang ditampilkan merupakan nilai *Logaritma Natural* dari *Sample Information*

4.6.4 Akurasi Pada Data Uji

Setelah didapatkan model terbaik bedasarkan proses pada *K-fold cross validation*. Selanjutnya model tersebut akan digunakan untuk melakukan prediksi pada data uji. Hasilnya, didapatkan nilai evaluasi model dalam melakukan prediksi pada data latih sebagai berikut.

Tabel 4.12 Tabel Evaluasi Model Pada Data Uji

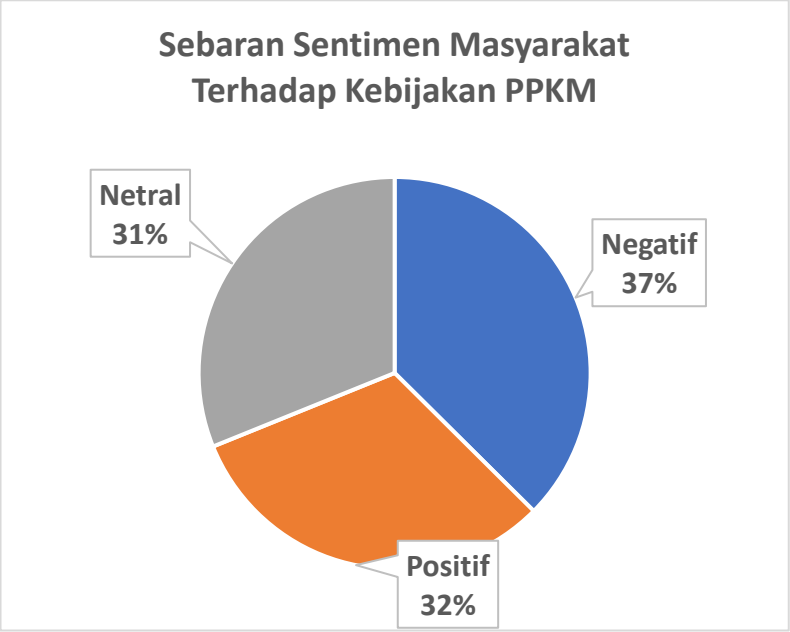
Aktual	Prediksi			Total
	Negatif	Netral	Positif	
Negatif	262	49	34	345
Netral	61	157	39	257
Positif	42	24	203	269
Total	365	230	276	871

$$acc = \frac{\sum TP}{\sum Total\ Kelas} = \frac{262 + 157 + 203}{871} = \frac{623}{871} = 0{,}714$$

Dari nilai evaluasi model pada data uji yang terdapat pada tabel 4.13 Dapat dihitung nilai akurasi pada dengan menggunakan formula 2.18. Nilai akurasi merupakan nilai pengukuran yang digunakan untuk melihat seberapa baik model bekerja. Pada nilai akurasi akan dilihat jumlah prediksi benar dari model dalam memprediksi *tweet* yang terdapat di dalam data uji. Hasilnya, penerapan model *Multinominal Naive Bayes Classifier* terbaik yang didapatkan pada proses c*ross validation* ketika pada data latih menghasilkan nilai akurasi sebesar 0,714

4.7 Hasil dan Pemaparan Sentimen

Diketahui bahwa jumlah sampel yang digunakan pada penelitian ini adalah 4354 *tweet* tentang kebijakan PPKM.

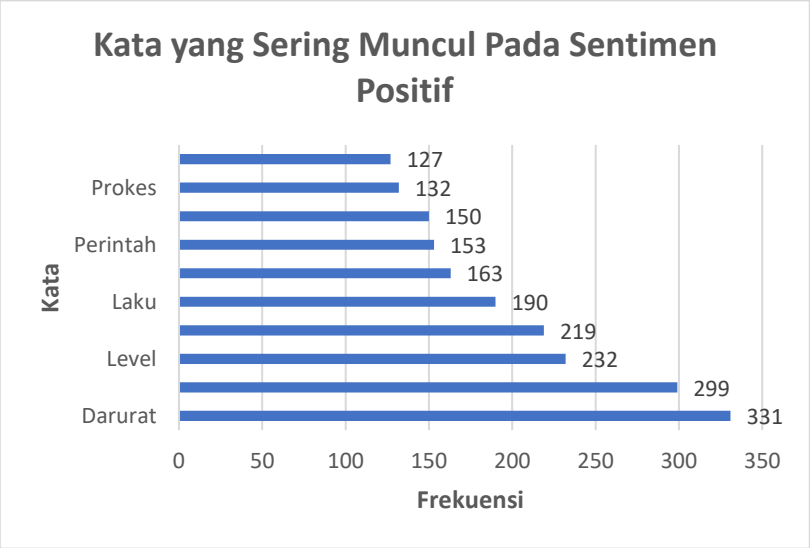


Gambar 4.3 Sebaran Sentimen Masyarakat

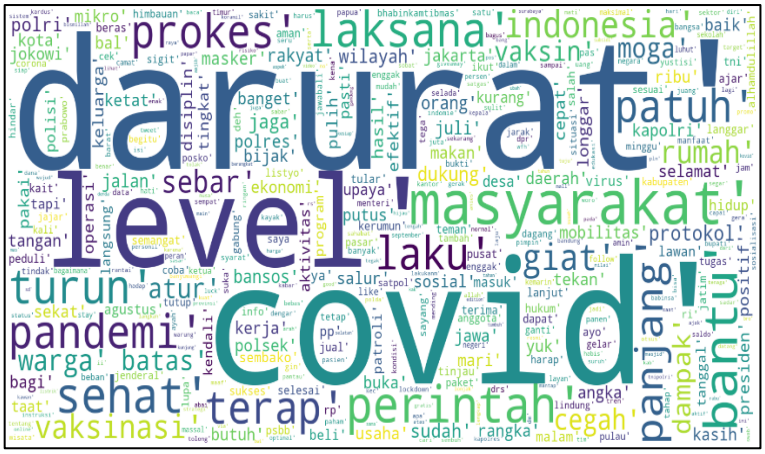
Bedasarkan gambar 4.3 dapat dilihat bahwa sebanyak 1630 (37%) *tweet* bersentimen negatif atau merasa penerapan kebijakan PPKM oleh pemerintah Indonesia berdampak buruk untuk masyarakat. Sementara itu, 1369 (32%) *tweet* bersentimen positif atau merasa penerapan penerapan kebijakan PPKM ini berdampak baik bagi serta diperlukan bagi masyarakat Indonesia. Dan 1355 (31%) sisanya bersentimen netral yang berisikan informasi terkait pelaksanaan

PPKM serta tidak ketidak berpihakan terhadap penerapan PPKM. Dari data ini juga dapat disimpulkan bahwa perbedaan jumlah sentimen dalam data tidak terlalu berbeda signifikan.

Pada penilitian kali ini, juga akan dilihat topik apakah yang menjadi pembahasan masyarakat ketika membicarakan tentang penerapan kebijakan PPKM. Adapaun untuk mempermudah pekerjaan ini peneliti menggunakan *package wordcloud* .



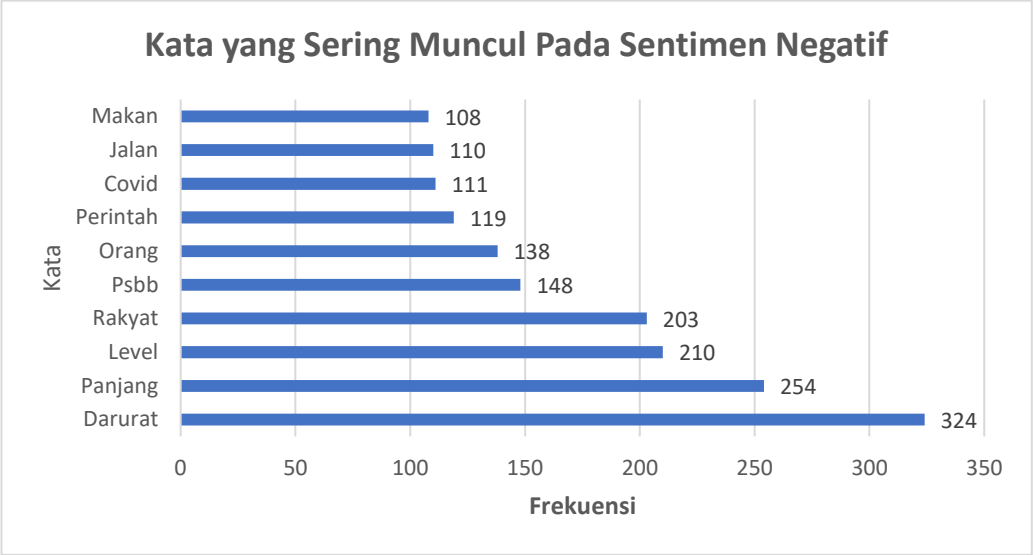
Gambar 4.4 Kata yang Sering Muncul Pada Sentimen Positif



Gambar 4.5 Gambaran Sentimen Positif

Pada gambar 4.4 dapat dilihat 10 kata dengan frekuensi muncul tertinggi pada sentimen positif, hal ini diperkuat dengan hasil pada gambar 4.5 dengan metode penyajian *wordcloud*, yang menggambarkan respon positif yang muncul pada masyarakat di media sosial Twitter terhadap penerapan kebijakan PPKM.

Adapun respon positif masyarakat antara lain, masyarakat mengganggap dengan penerapan PPKM ini mampu membuat warga Indonesia lebih patuh untuk menerapakan dan melaksanakan protokol kesehatan dalam kehidupan sehari hari. Selain itu penerapan kebijakan PPKM ini juga dianggap efektif dalam menurunkan kasus konfirmasi covid-19 atau kasus meninggal dunia yang disebabkan covid-19. Dengan adanya PPKM ini juga membuat masyarakat lebih gencar untuk meklakukan vaksinasi yang menjadi persyaratan administrasi dan beberapa kegiatan. PPKM juga secara tidak langsung meningkatkan rasa kepedulian sesama antara masyarakat, hal ini dibuktikan dengan munculnya banyak aksi serta kegiatan sosial yang bertujuan untuk saling membantu dan menolong masyarakat yang terdapak pandemi



Gambar 4.6 Kata yang Sering Muncul Pada Sentimen Negatif

Gambar 4.7 Gambaran Sentimen Negatif

Pada gambar 4.6 dapat dilihat 10 kata dengan frekuensi muncul tertinggi pada sentimen negatif. Sementara itu, gambar 4.7 menampilkan respon negatif masyarakat di media sosial Twitter terhadap penerapan kebijakan PPKM oleh pemerintah Indonesia. Banyak masyarakat yang menganggap penerapan PPKM yang telah berlangsung lama dan terus di perpanjang merugikan masyarakat dari segi ekonomi dan kegitan. Selain itu, pembatasan kegiatan masyarakat dalam penerapan PPKM seperti pemberlakuan waktu makan, penutupan jalan, hingga pelaksaan kegiatan perkantoran dan pemebelajaran dari rumah, dianggap masih kurang efektif dan menyusahkan masyarakat. Masyarakat juga mengkritik penegak hukum yang dianggap kurang tegas, dan terkesan pilih kasih dalam menegakan aturan PPKM, hal ini didadasari oleh banyaknya aparat yang melakukan kekerasan ketika melakukan penertiban, dan penerapan denda yang tidak seimbang antar kelompok orang yang melanggar PPKM.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Multinominal Naïve Bayes merupakan salah satu jenis algoritma Naïve Bayes dengan asumsi bahwa data berdistribusi *Multinominal*. *Multinominal Naïve bayes* dapat digunakan untuk melakukan analisis sentimen masyarakat terhadap penerapan kebijakan PPKM. Adapun tahapan yang dilakukan dalam analsisis sentimen dengan menggunakan *Multinominal Naïve Bayes Classifier* adalah menghitung nilai *prior* pada data latih, menghitung nilai *sampel information* pada data latih, menyesuaikan nilai pada data uji bedasarkan nilai *prior* dan *sampel information* dan menentukan sentimen bedasarkan nilai tersebut. Dari hasil analisis sentimen masyarakat terhadap penerapan kebijakan PPKM pada media sosial Twitter didapatkan hasil, algoritma *Multinominal Naïve Bayes Classifier* mendapatkan nilai akurasi sebesar 0,68 hingga 0,71 dalam proses *cross validation.* Sementara itu, penggunaan model terbaik pada yang didapatkan pada proses *cross validation* mendapatkan tingkat akurasi sebesar 0,714 pada data uji. Dari nilai akurasi dalam proses *cross validation* dan data uji dapat disimpulkan bahwa model sudah bekerja dengan cukup baik, selain itu penerapan *cross validation* juga terbukti membuat model lebih stabil, dikarenakan tidak terdapat perbedaan yang besar antara akurasi pada *cross validation* dan data uji.

Terdapat beberapa topik yang menjadi pembahasan utama dalam setiap sentimen. Pada sentimen negatif masyarakat menyoroti topik Perpanjangan PPKM, keritik terhadap penamaan PPKM berlevel, penutupan jalan, pembatasan waktu makan. Serta, penerapan sistem kerja dari rumah. Sementara itu, pada sentimen positif topik yang sering dibahas antara lain, penerapan protokol kesehatan yang semakin baik, vaksinasi, penurunan level PPKM serta kasus konfirmasi Covid-19.

5.2 Saran

Berdasarkan hasil analisis yang telah dilakukan, saran yang dapat diberikan adalah sebagai berikut:

1. Penelitian ini dapat dijadikan refrensi untuk penelitian selanjutnya tentang penerapan algoritma *Multinominal Naive Bayes* pada analisis sentimen. Hasilnya, performa dari model ini juga cukup baik untuk sebuah model tradisional dengan menghasilkan akurasi 0,714 pada data uji.
2. Pada penelitian selanjutnya diharapkan dapat menggunakan atau mendapatkan *package* yang lebih *update* untuk melakukan *pre-processing data tweet* dalam bahasa Indonesia.
3. Penlitian selanjutnya juga diharapkan dapat menemukan metode sesuai untuk proses *cleansing tweet* dan pemberian label pada data, agar menghilangkan proses secara manual, membuat proses analisis lebih cepat dan dapat menggunakan data yang lebih banyak.

DAFTAR PUSTAKA

Agustin, R., Santi, V. M., & Sumargo, B. (2019). 2 Program Studi Statistika, Fakultas Matematika Dan Ilmu Pengatahuan Alam. *Universitas Negeri Jakarta Jl. Rawamangun Muka*, *3*(1).

Bain, L. & Engelhardt. (1992). *Introduction To Probability And Mathematical Statistics Second Edition, California: Duxbury Thomson Learning.* (2nd Ed.).

Dan Jurafsky. (2021). *Multinomial Naive Bayes A Worked Example* . Artificial Intelligence - All In One.

Fahrur Rozi, I., Hadi Pramono, S., & Achmad Dahlan, E. (2012). *Implementasi Opinion Mining (Analisis Sentimen) Untuk Ekstraksi Data Opini Publik Pada Perguruan Tinggi*. *6*.

Hootsuite. (2020). *Hootsuite (We Are Social): Indonesian Digital Report 2020*.

Israel, G. D. (1992). *Determining Sample Size 1 The Level Of Precision*.

Kurniawan, B., Fauzi, M. A., & Widodo, A. W. (2017). *Klasifikasi Berita Twitter Menggunakan Metode Improved Naïve Bayes* (Vol. 1, Issue 10). Http://J-Ptiik.Ub.Ac.Id

Kurniawan, I., & Susanto, A. (2019). Implementasi Metode K-Means Dan Naïve Bayes Classifier Untuk Analisis Sentimen Pemilihan Presiden (Pilpres) 2019. *Eksplora Informatika*, *9*(1), 1–10. Https://Doi.Org/10.30864/Eksplora.V9i1.237

Murphy, K. P. (2022). *Probabilistic Machine Learning* (T. Dietterich, Ed.).

Muzaki, A., & Witanti, A. (2021). Sentiment Analysis Of The Community In The Twitter To The 2020 Election In Pandemic Covid-19 By Method Naive Bayes Classifier. *Jurnal Teknik Informatika (Jutif)*, *2*(2), 101–107. Https://Doi.Org/10.20884/1.Jutif.2021.2.2.51

Myatt, G. J. (2007). *Making Sense Of Data : A Practical Guide To Exploratory Data Analysis And Data Mining*. Wiley-Interscience.

Rahman, A., & Doewes, A. (2017a). *Online News Classification Using Multinomial Naive Bayes*. Www.Kompas.Com

Rizal, M., Afrianti, R., Abdurahman, I., Bisnis, D. A., & Bisnis, P. A. (2021). *Dampak Kebijakan Pemberlakuan Pembatasan Kegiatan Masyarakat (Ppkm) Bagi Pelaku Bisnis Coffe Shop Pada Masa Pandemi Terdampak Covid-19 Di Kabupaten Purwakarta The Impact Of The Policy For Implementing Community Activity Restrictions For Coffee Shop*

Businesses During The Covid-19 Pandemic Era Affected In Purwakarta Regency. Https://Doi.Org/10.35880/Inspirasi.V11i1.198

Sabrani, A., Gede Putu Wirarama Wedashwara, I. W., & Bimantoro, F. (2020). *Metode Multinomial Naïve Bayes Untuk Klasifikasi Artikel Online Tentang Gempa Di Indonesia (Multinomial Naïve Bayes Method For Classification Of Online Article About Earthquake In Indonesia)*. Http://Jtika.If.Unram.Ac.Id/Index.Php/Jtika/

Samsir, Ambiyar, Unung, V., & Firman, E. (2021). Analisis Sentimen Pembelajaran Daring Pada Twitter Di Masa Pandemi Covid-19menggunakan Metode Naïve Bayes. *Jurnal Media Informatika Budidarma*, *5*(1), 149. Https://Doi.Org/10.30865/Mib.V5i1.2604

Satgas Covid-19. (2021). *Peta Sebaran Covid-19*. Https://Covid19.Go.Id/Peta-Sebaran

Scheaffer, R. L., & Mendenhall. (2012). *Elementary Survey Sampling* (M. Julet, Ed.; 7th Ed.). Brooks/Cole.

Scikit-Learn Developers. (2007). *Multinomial Naive Bayes*. Https://Scikit-Learn.Org/Stable/Modules/Naive_Bayes.Html

Singh Carter, & Atenstaedt Rob. (2012). Word Cloud Analysis Of The Bjgp. *British Journal Of General Practice*.

Smrc. (2021). *Satu Tahun Covid-19: Sikap Dan Perilaku Warga Terhadap Vaksin*.

Sutoyo, E., & Almaarif, A. (2020). Twitter Sentiment Analysis Of The Relocation Of Indonesia's Capital City. *Bulletin Of Electrical Engineering And Informatics*, *9*(4), 1620–1630. Https://Doi.Org/10.11591/Eei.V9i4.2352

William M. Bolstad. (2007). *Introduction To Bayesian Statistics* (Second Edition). A John Wiley & Sons, Inc., Publication .

Yulima, S., Rembulan, N., Widayatno, A., Adina, E., Ziofani, H., Saputra, Y., Ardiansah, F., & Kegiatan, A. (2021). Di Desa Limbung. *Jabb*, *02*(01). Https://Doi.Org/10.46306/Jabb.V2i1

N. Aliyah Salsabila, Y. Ardhito Winatmoko, A. Akbar Septiandri And A. Jamal, "Colloquial Indonesian Lexicon," *2018 International Conference On Asian Language Processing (Ialp)*, 2018, Pp. 226-229, Doi: 10.1109/Ialp.2018.8629151.

LAMPIRAN

Lampiran 1 Surat Permohonan Data

KEMENTERIAN PENDIDIKAN, KEBUDAYAAN,
RISET DAN TEKNOLOGI
UNIVERSITAS NEGERI JAKARTA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

Kampus A, Gedung Hasjim Asj'arie Rawamangun, Jakarta Timur 13220
Telp/Fax : (021) 4894909, 08111937664, 08111511664, E-mail : dekanfmipa@unj.ac.id ; www.fmipa.unj.ac.id

Jakarta, 2 Juni 2022

No. : 1583/UN.39.6/FMIPA/PT.01.05/2022
Hal : Permohonan Izin Permintaan Data
Lamp. : -

Yth. PT. Ivonesia Solusi Data (Ivosights)
Jl. Tebet Barat I No. 2 Rt. 001/Rw. 02, Tebet Barat, Kec. Tebet, Jakarta Selatan

Dengan Hormat,
Dengan surat ini kami memohon kepada Bapak/Ibu untuk memberikan kesempatan kepada mahasiswa Program Studi Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Negeri Jakarta, atas nama :

No.	Nama Mahasiswa	No. Registrasi	Prodi
1	Naufal Zhafran Albaqi	1314618035	Statistika

Untuk melaksanakan pengambilan data, antara lain :
"Data Twitter Tentang Penerapan Kebijakan PPKM Dari 7 Januari 2021 Hingga 31 Oktober 2021".

Demikian permohonan ini kami sampaikan atas perhatian dan kerjasamanya yang baik diucapkan terima kasih.

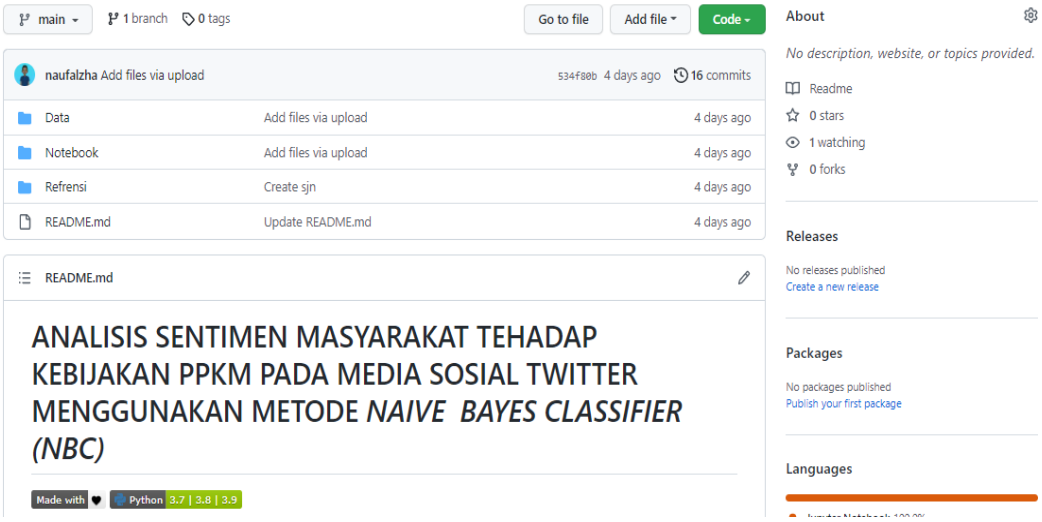
Wakil Dekan Bidang Akademik,

Dr. Esmar Budi, S.Si, MT
NIP. 197207281999031002

Lampiran 2 Jumlah Interaksi Harian Tentang PPKM pada media sosial Twitter

date	message	date	message	date	message	date	message
07/01/	9248	08/09/	27696	09/17/	3615	10/26/	2489
07/02/	50849	08/10/	18077	09/18/	5615	10/27/	2300
07/03/	58511	08/11/	13772	09/19/	2961	10/28/	1558
07/04/	61731	08/12/	12310	09/20/	4889	10/29/	1195
07/05/	49696	08/13/	10770	09/21/	4966	10/30/	1750
07/06/	28043	08/14/	10754	09/22/	4505	10/31/	2160
07/07/	27697	08/15/	11475	09/23/	4685		
07/08/	25353	08/16/	48327	09/24/	2847		
07/09/	26771	08/17/	3988	09/25/	2248		
07/10/	15735	08/18/	9634	09/26/	2347		
07/11/	43244	08/19/	9926	09/27/	3715		
07/12/	25311	08/20/	8654	09/28/	2459		
07/13/	28338	08/21/	10039	09/29/	2369		
07/14/	33015	08/22/	8056	09/30/	2318		
07/15/	65699	08/23/	14942	10/01/	1666		
07/16/	29319	08/24/	12125	10/02/	1840		
07/17/	30639	08/25/	9953	10/03/	1966		
07/18/	29241	08/26/	8859	10/04/	3006		
07/19/	27233	08/27/	7704	10/05/	3223		
07/20/	30042	08/28/	8896	10/06/	2822		
07/21/	55164	08/29/	7971	10/07/	2828		
07/22/	42117	08/30/	10371	10/08/	2267		
07/23/	22691	08/31/	14103	10/09/	2357		
07/24/	22537	09/01/	9022	10/10/	1941		
07/25/	46641	09/02/	11695	10/11/	1623		
07/26/	33864	09/03/	6074	10/12/	2912		
07/27/	25121	09/04/	4921	10/13/	2784		
07/28/	20699	09/05/	5310	10/14/	2252		
07/29/	22865	09/06/	12309	10/15/	1954		
07/30/	25588	09/07/	10382	10/16/	1735		
07/31/	20449	09/08/	9506	10/17/	1728		
08/01/	18066	09/09/	8556	10/18/	2643		
08/02/	41651	09/10/	5077	10/19/	4906		
08/03/	25582	09/11/	4633	10/20/	3535		
08/04/	19885	09/12/	4542	10/21/	3524		
08/05/	20686	09/13/	6783	10/22/	2884		
08/06/	16667	09/14/	7480	10/23/	1821		
08/07/	7030	09/15/	4942	10/24/	1831		
08/08/	5822	09/16/	3663	10/25/	2451		

Lampiran 3 Code, Refrensi, Data



Sumber: *https://github.com/naufalzha/Skripsi*

DAFTAR RIWAYAT HIDUP

Naufal Zhafran Albaqi dilahirkan pada tanggal 5 Februari 2000 di Depok, Jawa barat dari pasangan Bapak Samsir Alam dan Ibu Heni herniawati. Penulis merupakan anak pertama dari empat bersaudara. Saat ini penulis bertempat tinggal di Jl. Swadaya 2 Palsigunung-Poncol, No.105 Rt.06 Rw.01, Kecamatan Cimanggis, Kota Depok, Jawa Barat 1645.

Pada tahun 2012 penulis lulus dari SD Negeri 9 Sunghin. Setelah itu, penulis lulus dari SMP Negeri 2 Sungailiat pada tahun 2015. Lalu, penulis lulus dari SMA Negeri 1 Sungailiat pada tahun 2018. Pada tahun yang sama penulis lulus seleksi masuk Universitas Negeri Jakarta melalui jalur Penerimaan Mahasiswa Baru Seleksi (PENMABA) di Universitas Negeri Jakarta. Kemudian, penulis melanjutkan pendidikannya di Program Studi Statistika Universitas Negeri Jakarta.

Selama hidupnya, penulis pernah beberapa kali aktif di berbagai organisasi. Ketika SMA, ia pernah menjadi anggota ekstrakulikuler sepak bola sejak SMP hingga SMA. Selain itu, penulis juga aktif di kegiatan karang taruna tempat penulis tinggal, penulis juga pernah mengikuti beberapa pelatihan kepemimpinan sejak SMA. Ketika Melaksanakan pendidikan di perguruan tinggi, penulis juga aktif sebagai anggota divisi Kaderisasi Program Studi Statistika Universitas Negeri Jakarta pada tahun (2018-2020)

Penulis juga memiliki pengalaman kerja dari beberapa kesempatan yang ia dapat. Ia pernah mengikuti program Praktik Kerja Lapangan di PT. Ivonesia Solusi Data (Ivosight) pada bagian data dan operation bulan Juli-Oktober 2021. Kemudian, ia juga berkesempatan untuk daftar dan lolos seleksi Program Studi Independen Bersertifikat di Bangkit Academy, dengan topik pembelajaran yang diambil machin learning. Penulis melaksanakan program ini selama enam bulan, terhitung dari Februari- Agustus 2022. Dan saat ini penulis juga memegang sertifikasi sebagai TensorFlow Developer

.