INFORMATICS AND DIGITAL EXPERT (INDEX) - VOL. 3 NO. 2 (2021) 75-83
Terbit *online* pada laman web jurnal : https://e-journal.unper.ac.id/index.php/informatics
| ISSN (Print 2775-2208 | ISSN (Online) 2715-0453 |

ANALISIS SENTIMENT PENGGUNAAN VAKSIN *COVID-19* MENGGUNAKAN *GEO-TAGGED TWEETS* DAN ALGORITMA *NAIVE BAYES*

Azy Mushofy Anwary [1], Asep Id Hadiana, Puspita Nurul Sabrina[3]

[1,2,3] *Universitas Jenderal Achmad Yani Cimahi Jl. Terusan Sudirman, Cimahi 40513, Indonesia*
e-mail: azy.anwary9@gmail.com[*1], asep.hadiana@lecture.unjani.ac.id[2], puspita.sabrina@lecture.unjani.ac.id

INFORMASI ARTIKEL

Sejarah Artikel:
Diterima Redaksi:28 Oktober 2021
Revisi Akhir: 13 November 2021
Diterbitkan *Online*: 30 November 2021

Kata Kunci :
Twitter, Aanalisis Sentiment, Naïve Bayes, Geo-Tagged

Korespondensi :
Telepon / Hp : 022-6656190
E-mail : **azy.anwary9@gmail.com**

A B S T R A K

Sentimen analisis adalah salah satu teknik yang dapat dilakukan untuk mengolah suatu opini dari masyarakat salah satunya pada media sosial yaitu twitter. Dengan sentiment analisis data twitter tersebut dapat diketahui apakah polaritas suatu data tersebut akan mengarah pada sifat positif, negatif, atau netral. Penelitian ini menggunakan topik vaksin Covid-19 yang didapat dari Twitter. Metode yang digunakan dalam penelitian ini adalah metode naïve bayes. Metode *naïve bayes* adalah metode yang sering digunakan dalam mengkategorikan teks dan sangat cocok digunakan untuk implementasi analisis sentiment. Pada penelitian ini juga terdapat fitur tambahan yaitu fitur *Geo-Tagged*, fitur ini berguna untuk mengambil data pengguna twitter agar mengetahui lokasi dan waktu pengguna pada saat melakukan tweet. Ada beberapa proses yang dilakukan pada penelitian ini diantaranya pengumpulan data, pelabelan data, *preprocessing data*, *feature extraction*, penyeimbangan kelas label, mengklasifikasikan data menggunakan metode *naïve bayes*, melakukan visualisasi data berupa maps dan yang terakhir yaitu evaluasi hasil. Penelitian ini menghasilkan nilai akurasi (79%) dengan dibantu oleh metode *synthetic minority oversampling technique*. Data yang digunakan sebesar 1132 dataset yang diambil langsung menggunakan Teknik *crawling* dengan *liblary twint*. Wilayah yang melakukan tweet terbanyak jatuh kepada wilayah Karawang dengan sentimen positif 70 tweet, sentimen negatif 12 tweet dan sentimen netral 13 tweet..

1. PENDAHULUAN

Sentimen adalah istilah yang dapat menggambarkan topik yang objektif dan subjektif serta topik non-faktual ataupun faktual yang memiliki hasil berbeda diantaranya topik positif atau topik negatif[1]. Analisis sentimen merupakan suatu proses untuk menentukan suatu isi dari dataset yang berbentuk text (kalimat, dokumen, kata, paragram,dll)[2][3]. Analisis sentimen ini bertujuan untuk mengetahui subjektivitas opini, hasil review atau tweet. Berdasarkan analisis sentimen, opini dari seseorang dapat diklasifikasikan ke dalam berbagai kategori diantaranya positif, negatif atau netral berdasarkan data tekstual[4]. Salah satu teknik yang dapat digunakan pada analisis sentiment adalah menggunakan metode *Naïve Bayes Classifie*. Metode Naïve Bayes adalah algoritma yang sering digunakan dalam mengkategorikan teks. Naïve Bayes merupakan teknik pembelajaran mesin yang berbasis probabilistik[5].

Twitter adalah platform mikro-blog standar yang telah dianalisis oleh para peneliti karena basis pengguna yang besar lebih dari 319 juta pengguna aktif[6]. Oleh karena itu, situs semacam itu kaya akan sumber data untuk dimanfaatkan[7]. Pada twitter juga terdapat informasi tambahan seperti informasi geografis tentang pengguna pada saat melakukan tweet atau biasa disebut twitter geo-tagged[8]. Twitter geo-tagged adalah tweet yang berisi koordinat geografis (*latitude, longitude*) yang menunjukkan lokasi dimana tweet tersebut dibuat[6]. Pengguna tweet yang diberi geo-tagged hanya ada sedikit dibandingkan dengan tweet yang tidak di beri geo tagged. namun tweet yang di beri geo tagged akan memberikan nilai lebih karena dapat mengolah data tersebut menjadi lebih unik[9]. Pemanfaatan Tweet Geo-tagged dapat digunakan dalam beberapa hal salah satunya dapat divisualisasikan dalam bentuk peta.

Pada penelitian terdahulu menjelaskan bahwa Analisis sentimen menggunakan data Twitter yang diberi geotag, dapat memungkinkan analisis polaritas sentimen pada tingkat yang sangat detail[10]. Pada penelitian sebelumnya menggunakan metode Naïve Bayes mendapat nilai akurasi dengan nilai rata-rata akurasi mencapai 93%, dengan kasus "Analisis Sentimen Terhadap Opini Masyarakat Tentang Vaksin Covid-19 Menggunakan Algoritma Naïve Bayes Classifier"[11]

Topik yang akan diangkat pada penelitian ini adalah vaksin Covid-19. Dikarenakan hangatnya perbincangan di twitter terkait vaksin Covid-19 karena masih diupayakan penelusuran lebih lanjut mengenai vaksin covid-19 yang berpotensi agar aman dan efektif saat digunakan[12].

Maskud dari penelitian ini adalah untuk menganalisis klasifikasi sentiment pada data twitter mengenai vaksin Covid-19 yang diberi geo-tagged menggunakan metode naïve bayes yang diperloleh dari masyarakat yang ada di Indonesia kususnya masyarat yang ada di Pulau Jawa. Data yang akan dihasilkan berupa data yang berbentuk sentiment positif, negatif ataupun netral dan divisualisasikan berdasarkan lokasi tweet tersebut.

Penelitian ini menggunakan metode naïve bayes dengan menggunakan beberapa tahapan penelitian

seperti, pengumpulan data, pelabelan data, *preprocessing* data, *feature extraction*, penyeimbangan kelas label, klasifikasi naïve bayes, evaluasi hasil dan implementasi data berupa visualisasi yang disajikan dalam bentuk peta.

2. METODE PENELITIAN

A. Perolehan Data

Pada penelitian ini, tweet dikumpulkan dengan kata kunci vaksin Covid-19 menggunakan teknik Crawling twint dari media sosial Twitter. Teknik Crawling adalah teknik mengambil/mengekstrak data dari suatu website secara spesifik.[1]

Tweet yang berhasil di kumpulkan adalah tweet yang muncul berdasarkan lokasi pengguna di tiap kabupatennya dengan kata kata kunci atau topik vaksin Covid-19.

B. Preprocessing Data

Pada pre-processing data terdapat beberapa tahapan diantaranya tokenizing, filtering, stemming, serta stopwords.

1. *Case folding* adalah tahapan proses mengubah semua huruf dalam teks dokumen menjadi huruf kecil, serta menghilangkan karakter selain a-z
2. *Tokenization* adalah proses untuk motongan urutan karakter dari sebuah dokumen menjadi potongan-potongan kata atau karakter yang sesuai dengan kebutuhan sistem.
3. *Filtering* adalah tahapan mengambil kata kata penting dari hasil token berdasarkan stopwords. Stopwords merupakan sebagai menghilangkan karakter, tanda baca, serta kata-kata umum yang tidak memiliki makna atau informasi yang dibutuhkan.
4. *Stemming* merupakan salah satu proses dari mengubah token yang berimbuhan menjadi kata dasar, dengan menghilangkan semua imbuhan yang ada pada token tersebut.

C. Feature Extraction

Feature extraction pada penelitian ini TF-IDF adalah suatu proses untuk melakukan transformasi data dari data tekstual ke dalam data numerik untuk dilakukan pembobotan pada tiap kata atau fitur[13]. TF-IDF ini adalah sebuah ukuran statistik yang digunakan untuk mengevaluas seberapa penting sebuah kata di dalam sebuah dokumen atau dalam sekelompok kata[9].

Rumus TF :

$$\boldsymbol{TF_{ij} = tf_{ij}} \quad (1)$$

Keterangan:

TF_{ij} : Nilai *Term Frequency* pada *term i* dalam dokumen *j*

tf_{ij} : Frekuensi kemunculan *term i* dalam dokumen *j*

Rumus IDF :

$$\boldsymbol{IDF_{ij} = \log\left(\frac{N}{n_i}\right) + 1} \quad (2)$$

Keterangan :

IDF_i : Nilai *Invers Document Frequency* pada *term i* dalam dokumen *j*

N : Jumlah keseluruhan dokumen (*tweet*) dalam dataset

n_i : Jumlah dokumen yang memiliki kemunculan *term i*

Sehingga didapat untuk persamaan dari TF-IDF adalah sebagai berikut :

Rumus TF-IDF :

$$\boldsymbol{TF_{ij} - IDF_{ij} = tf_{ij}\, x \log\left(\frac{N}{n_i}\right) + 1} \quad (3)$$

D. Penyeimbangan Kelas Label

Synthetic Minority Oversampling Technique (SMOTE) adalah metode oversampling yang digunakan untuk menangani masalah ketidakseimbangan kelas, dimana data pada kelas minoritas diperbanyak dengan menggunakan data sintetik yang berasal dari replikasi data pada kelas minoritas[14]. Metode ini menambahkan jumlah data kelas minor agar setara dengan kelas mayor dengan cara membangkitkan data buatan[15].

Rumus SMOTE :

$$\boldsymbol{Xnew = Xi + (X\hat{i} - Xi) \times \delta} \quad (1)$$

Keterangan :

$X\,i$: vektor dari fitur pada kelas minoritas

$X\hat{i}$: k-nearest neighbors untuk Xi

δ : angka acak antara 0 sampai 1

E. Klarifikasi Naive Bayes

Naive Bayes Classifier merupakan salah satu metode yang populer untuk keperluan data mining karena penggunaannya yang mudah dan dalam pemroresan memiliki waktu yang cepat, mudah diimplementasikan dengan strukturnya yang cukup sederhana dan untuk tingkat efektifitasnya memiliki efektifitas yang tinggi. Klasifikasi Naive Bayes juga memperlihatkan tingginya akurasi dan cepat ketika digunakan untuk dataset dengan jumlah besa.[16].

Secara umum proses dari klasifikasi Naive Bayes dapat dilihat pada Persamaan dibawah ini

Rumus :

$$P(\boldsymbol{H\,X}) = \frac{\boldsymbol{P(X|H)\,P(H)}}{\boldsymbol{P(X)}} \quad (1)$$

Keterangan :

$P(H|X)$: Probabilitas kelas H dari dokumen (*tweet*) yang diinputkan, *posterior probability*

$P(X|H)$: Probabilitas *term X* dalam kelas H, *conditional probability*

$P(H)$: Probabilitas kemunculan kelas H dalam dataset, *class prior probability*
$P(X)$: Probabilitas kemunculan *term* X dalam dataset, *predictor prior probability*

Pada proses perhitungan klasifikasi peluang kemunculan kata sebenarnya dapat dihilangkan, hal ini dikarenakan peluang tersebut tidak berpengaruh pada perbandingan hasil klasifikasi dari setiap kategori. Sehingga proses pada klasifikasi dapat disederhanakan dengan Persamaan dibawah ini[16].

Rumus :

$$\boldsymbol{P(H|X) = P(X|H)\ P(H)} \quad (2)$$

Untuk menghitung *class pior probability* atau P(H) dapat dilakukanoleh persamaan
Rumus :

$$\mathrm{P(H_i)} = \frac{\mathbf{N_i}}{\mathbf{N}} \quad (3)$$

Keterangan :

Ni : Jumlah data dalam kelas Hi pada data latih
N : Jumlah keseluruhan data latih

Karena data memiliki banyak atribut, maka sangat sulit untuk menghitung P(X|H).Agar dapat meminimalkan perhitungan P(X|H) maka persamaannya menjadi:[17]
Rumus :

$$\boldsymbol{P(X|H_i)} = \prod_{k=1}^{n} \boldsymbol{P(x_k|H_i)} \quad (4)$$
$$= \boldsymbol{P(x_1|H_i)} \times \boldsymbol{P(x_2|H_i)} \times \ldots \times \boldsymbol{P(x_n|H_i)}$$

Keterangan :

$P(x_k|H_i)$: Probabilitas kemunculan *term* x_k dalam kelas H_i

Penggunaan persamaan Probabilitas diatas ditunjukan agar lebih mudah untuk diperkirakan. Untuk menghitung nilai dari P(xk | Hi) digunakan persamaan dibawah ini:
Rumus :

$$\boldsymbol{P(x_k|H_i)} = \frac{\boldsymbol{T_{ki}}}{\boldsymbol{T_i}} \quad (5)$$

Keterangan :

T_{ki} : Jumlah *term* x_k dalam kelas H_i pada data latih
T_i : Jumlah seluruh Tern dalam kelas H_i pada data latih

Pada dataset memiliki beberapa contoh pemilihan data latih secara acak yang mungkin akan menghasilkan data yang bernilai nol terhadap suatu kelas. Maka dari itu digunakannya metode *laplace smooting* untuk mencegah situasi probabilitas nol dan untuk memastikan bahwa setiap kata memiliki peluang kemunculan, berdasarkan setidaknya satu hitungan maka dari itu digunakan persamaan seperti dibawah ini[18]
Rumus :

$$\boldsymbol{P(x_k|H_i)} = \frac{\boldsymbol{T_{ki}+1}}{\boldsymbol{T_i+\beta}} \quad (6)$$

Keterangan :

β : Jumlah total fitur dalam data latih

Maka persamaan naïve bayes sebeagai berikut :

Rumus :

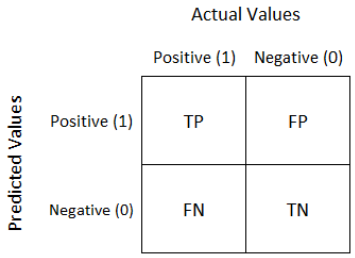
$$P(\boldsymbol{H|X}) = \boldsymbol{P(H_i)} \prod_{k=1}^{n} \boldsymbol{P(x_k|H_i)} \quad (7)$$
$$= \boldsymbol{P(H_i)} \times \boldsymbol{P(x_1|H_i)} \times \boldsymbol{P(x_2|H_i)} \times \ldots$$
$$\times \boldsymbol{P(x_n|H_i)}$$

F. Twitter Geo-Tagged

Twitter memberi penggunanya pilihan untuk 'memberi tag geografis' pada Tweet saat diposting. Penandaan geografis ini dapat didasarkan pada lokasi yang tepat, diberi Tempat Twitter atau keduanya. *Twitter Places* dapat dianggap sebagai tingkat lingkungan, yang menyediakan informasi lebih lengkap dengan koordinat lintang dan bujur yang menentukan area lokasi. Jenis metadata geografis ini, yang disebut sebagai "Lokasi Tweet" memberikan tingkat presisi tertinggi. Lokasi Tweet tidak memerlukan proses atau penguraian bahasa untuk mengakses informasi geografis[19]. Lokasi Tweet akan menghasilkan Tweet yang telah 'diberi tag geo.' Lokasi Tweet tersebut dapat ditetapkan dengan menggunakan antarmuka pengguna Twitter atau saat memposting Tweet menggunakan API[20]. Kelemahan utama jika menggunakan Lokasi Tweet adalah hanya 1-2% Tweet yang diberi tag geo. Selain itu, memberikan jangkauan yang sangan luas (misal seluruh negara bagian atau Provinsi) memerlukan penggunaan serangkaian aturan *PowerTrack* yang signifikan untuk mencakup seluruh area[19].

G. Confusion Matrix

Confusion Matrix merupakan sebuah metode untuk evaluasi yang menggunakan tabel matrix seperti pada Gambar I.



Gambar 1. Confusion Matrix

Pada gambar I dapat kita lihat bahwa jika dataset terdiri dari dua kelas, kelas yang satu dianggap sebagai positif dan yang lainnya negatif, Confusion Matrix ini biasa digunakan untuk mengukur kinerja dari *machine learning* dan dapat digunakan sebagai alat visual untuk mengevaluasi hasil klasifikasi [21].

Pada Gambar 2.1. True Positive (TP) menyatakan data positif yang diprediksi benar, True Negative (TN) menyatakan data negatif yang diprediksi benar. kemudian False Positive (FP) sebagai kesalahan tipe 1 merupakan data negatif namun diprediksi sebagai data positif, sebaliknya False Negative (FN) sebagai

kesalahan tipe 2 merupakan data postif namun diprediksi sebagai data negatif. Selain itu kita dapat menghitung nilai tersebut di antaranya Accuracy, Precision, Recall, dan F1-Score[21].

$$\boldsymbol{ACCURASY} = \frac{\boldsymbol{TN+TP}}{\boldsymbol{TP+FP+FN+TN}} \quad (1)$$

Akurasi Klasifikasi (ACC): Untuk mengukur tingakat akurasi dari pengklasifikasi data yang akan dievaluasi.

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \quad (2)$$

Presisi: Ini adalah rasio jumlah sentimen yang diprediksi secara akurat dengan jumlah total sentimen yang diprediksi.

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \quad (3)$$

Recall : Ini adalah rasio jumlah sentimen yang diprediksi secara akurat dengan jumlah total sentimen aktual

$$\boldsymbol{F-1\ Score} = \frac{\boldsymbol{2*(Recall*Precision)}}{\boldsymbol{Recall+Precision}} \quad (4)$$

F-1 Score: ini menggambarkan perbandingan rata-rata precision dan recall yang dibobotkan. Accuracy tepat kita gunakan sebagai acuan performansi algoritma jika dataset kita memiliki jumlah data False Negatif dan False Positif yang sangat mendekati (symmetric). Namun jika jumlahnya tidak mendekati, maka sebaiknya kita menggunakan F1 Score sebagai acuan.

3. PERANCANGAN SISTEM

Perancangan sistem ini membahas bagian isi dari tahapan yang dilakukan pada metode penelitian.

A. Perolehan Data

Perolehan data yang digunakan pada penelitian ini diambil dari kumpulan tweet yang berasal dari data public yang ada di twitter. Data tweet ini diperoleh dengan menggunakan proses crawling data menggunakan libalary twint dengan Bahasa pemograman phyton. Data yang diambil sebanyak 1132 data dengan kata kunci "vaksin covid-19" yang dimulai dari tanggal 1 Januari 2021 sampai 30 September 2021 berdasarkan longitude dan latitude di setiap pusat kabupatennya. dilakukannya filter pada data yang telah didapat agar tidak adanya data yang sama atau duplikat. Contoh Data tweet yang sudah diperoleh dapat dilihat pada table I :

Tabel 1 . Contoh Tweet

Id	1367353684674230000	1428978859014110000
Usernaname	imambdn11	ratugaluh73
Tweet	Vaksin Covid 19 tahap 1 https://t.co/niFCPUMAQV	Hari ke-4, vaksin Covid-19 dosis kedua ðŸ¤— https://t.co/T8XilzM00K
Geo-Lokasi	-7.010637831252968 ,107.52634034196025	-6.647046904720037 ,106.21435454457047
Nama Loaksi	Bandung	Lebak

B. Pelabelan Data

Pelabelan data merupakan tahapan yang digunakan untuk memberi kelas atau label pada data tweet yang mentah agar dapat digunakan untuk proses selanjutnya. Pelabelan ini dibedakan dalam beberapa kategori diantaranya positf negatif ataupun netral. Pelabelan dilakukan oleh seorang ahli dibidangnya secara manual atau satu persatu. Untuk contoh tweet atau dataset yang telah dilabeli terdapat pada tabel dibawah ini

Tabel 2. Perolehan Data

No	Tweet	Label
1	Barusan dapat cerita, teman yang adalah ibu menyusui, ga boleh divaksin COVID-19.vaksin https://t.co/ySvqMQqMsh	Negatif
2	Rabu, 27 Januari 2021 Giat Camat Lengkong mengikuti Kegiatan *Rapat Evaluasi Covid-19 dan Vaksin Kota Bandung* secara virtual melalui Zoom Meeting @ Kecamatan Lengkong https://t.co/YOMIJ8tR3f	Netral
3	Alhamdullilah, Vaksin bpk lancar hari ini ðŸ™• Semoga diberikan kekebalan buat semua yg sudah menerima vaksin..ðŸ’ª #Covid_19	Positif
4	Saya siap di vaksin covid-19 https://t.co/hC5q5XqEn0	Positif
5	Kalau cara vaksin nya model skrng dan dimana masyarakat masih setengah2 percaya ada Covid-19 ya wasalam	Negatif

C. Preprocessing Data

Preprocessing data merupakan teknik awal data mining untuk mengubah data mentah atau biasa dikenal dengan raw data yang dikumpulkan dari berbagai sumber menjadi informasi yang lebih bersih dan bisa digunakan untuk pengolahan selanjutnya. Pada penelitian ini terdapat 4 langkah dalam Preprocessing data diantaranya:

1. Case Folding

Case folding atau mengubah semua huruf dalam kalimat menjadi huruf kecil. Hanya huruf 'a' sampai 'z' yang diterima agar sistem dapat mengolah data lebih efisien dan efektif.

Tabel 3. Case Folding

Input	Output
Vaksin untuk keselamatan Stop Hoaks Vaksinasi	vaksin untuk keselamatan stop hoaks vaksinasi

2. Tokenazition

Tokenization merupakan proses motongan urutan karakter dari sebuah dokumen menjadi potongan-potongan kata atau karakter yang sesuai dengan kebutuhan sistem, Tokenisasi dilakukan untuk mempermudah pengamatan makna tiap kata yang berpengaruh dalam menentukan sentimen yang dimaksud termasuk kedalam sentimen positif, negatif ataupun netral.

Tabel 4. Tokenization

Input	Output
vaksin untuk keselamatan stop hoaks vaksinasi	[‘vaksin’ ‘untuk’ ‘keselamatan’ ‘stop’ ‘hoaks’ ‘vaksinasi’]

3. Filtering

Filtering adalah proses mengambil kata kata penting token berdasarkan stopwords. Stopwords merupakan sebagai menghilangkan karakter, tanda baca, serta kata-kata umum yang tidak memiliki makna atau informasi yang dibutuhkan.

Tabel 5. Filtering

Input	Output
[‘vaksin’ ‘untuk’ ‘keselamatan’ ‘stop’ ‘hoaks’ ‘vaksinasi’]	[‘vaksin’ ‘keselamatan’ ‘stop’ ‘hoaks’ ‘vaksinasi’]

4. Stemming

Pada proses *Stemming* akan mengubah token yang asalnya memiliki imbuhan menjadi kata dasar dengan menghilangkan imbuhannya..

Tabel 6. Stemming

Input	Output
[‘vaksin’ ‘keselamatan’ ‘stop’ ‘hoaks’ ‘vaksinasi’]	[‘vaksin’ ‘selamat’ ‘stop’ ‘hoaks’ ‘vaksinasi’]

D. Feature Extraction (TF-IDF)

Dalam pembobotan TF-IDF terdapat 3 langkah yang harus dilalui diantaranya mendapatkan vector dari hasil perhitungan TF-IDF, mendapatkan vektor nilai dari hasil perhitungan IDF dan yang terakhir mengalikan hasil dari perhitungan TF dan IDF. Dataset yang akan digunakan dalam proses TF-IDF adalah data yang telah melalui tahapan praproses. Untuk contoh kalimatnya terdapat pada tabel dibawah ini.

Tabel 7. Contoh Kalimat

Text tweet	Hasil Praproses
Vaksin untuk keselamatan Stop Hoaks Vaksinasi	[vaksin, selamat, stop, hoaks, vaksinasi]
Berharap pemberian vaksin COVID-19 bisa kayak posyandu.	[harap, beri, vaksin, covid19, bisa, kaya, posyandu]
Saya Tidak Percaya Covid-19	[saya, tidak, percaya, covid19,]
Alhamdulillah sudah vaksin	[alhamdulillah, sudah, vaksin]
Saya ragu untuk di vaksin covid19	[saya , ragu, vaskin, covid19]

Tahap TF: Nilai vector TF didapatkan dari banyaknya jumlah kemunculan kata yang ada didalam sebuah text. Dari tabel diatas terdapat lima buah text tweet dan didalam satu text tweet itu terdapat beberapa kata yang telah melalui tahapan praproses dan dari lima text tweet tersebut terdapat 17 kata yang berbeda. Panjang dari vektor yang dihasilkan menyesuaikan dengan banyaknya fitur yang digunakan, nilai yang dimasukan kedalam vektor adalah frekuensi kemunculan kata dalam teks.yang memiliki frekuensi kemunculan satu kali dalam satu text diberi nilai satu, jika frekuensi kemunculannya dua kali diberi nilai dua, dan selanjutnya. Tetapi jika dalam text tersebut tidak terdapat frekuensi kemunculan kata tersebut maka diberi nilai 0.

Tabel 8 TF

Fitur	TF				
	Text 1	Text 2	Text 3	Text 4	Text 5
vaksin	1	1	0	1	1
selamat	1	0	0	0	0
stop	1	0	0	0	0
hoaks	1	0	0	0	0
vaksinasi	1	0	1	0	0
harap	0	1	0	0	0
beri	0	1	0	0	0
covid19	0	1	1	0	1
bisa	0	1	0	0	0
kaya	0	1	0	0	0
prosyandu	0	1	0	0	0
saya	0	0	1	0	1
tidak	0	0	1	0	0
percaya	0	0	1	0	0
alhamdulillah	0	0	0	1	0
sudah	0	0	0	1	0
ragu	0	0	0	0	1

Tahap IDF : Panjang dari vektor IDF menyesuaikan dengan banyaknya jumlah fitur, nilai yang dimasukan kedalam vektor adalah hasil dari perhitungan $IDF_{ij} = \log\left(\frac{N}{n_i}\right) + 1$. sebagai contoh untuk menghitung vaktor IDF pada kata vaksin adalah sebagai berikut:

$$IDF_{fitur} = \log\left(\frac{N}{n_{fitur}}\right) + 1 \quad (1)$$

$$IDF_{fitur} = \log\left(\frac{5}{1}\right) + 1$$

$$IDF_{fitur} = \log\ (5) + 1$$

$$IDF_{fitur} = 2{,}60945$$

Nilai N adalah banyaknya text pada data yang akan dihitung. Sebagai contoh jumlah text yang digunakan pada tabel diatas sebanyak 5 text dan n*fitur* adalah banyaknya kemunculan kata dalam dataset. karena kata “vaksin” hanya muncul satu kali pada teks pertama maka n*fitur* bernilai 1. Sehingga IDF pada kata vaksin adalah 2,60945. Hal yang sama akan dilakukan oleh setiap kata yang lain yang dapat dilihat pada tabe.

Tabel 9. IDF

Fitur	IDF
vaksin	2,60945
selamat	2,60945
stop	2,60945
hoaks	2,60945
vaksinasi	2,60945
harap	2,60945
beri	2,60945
covid19	2,60945
bisa	2,60945
kaya	2,60945
prosyandu	2,60945
saya	2,60945
tidak	2,60945
percaya	2,60945
alhamdulillah	2,60945
sudah	2,60945
ragu	2,60945

Tahap TF-IDF: Tahap ini adalah tahap yang melakukan pengkalian dari vektor hasil TF dan vektor hasil IDF sehingga hasil dari pembobotan TF-IDF terdapat pada tabel dibawah ini.

Tabel 10. TF-IDF

Fitur	TF				
	Text 1	Text 2	Text 3	Text 4	Text 5
vaksin	2,60945	2,60945	0	2,60945	2,60945
selamat	2,60945	0	0	0	0
stop	2,60945	0	0	0	0
hoaks	2,60945	0	0	0	0
vaksinasi	2,60945	0	2,60945	0	0
harap	0	2,60945	0	0	0
beri	0	2,60945	0	0	0
covid19	0	2,60945	2,60945	0	2,60945
bisa	0	2,60945	0	0	0
kaya	0	2,60945	0	0	0
prosyandu	0	2,60945	0	0	0
saya	0	0	2,60945	0	2,60945
tidak	0	0	2,60945	0	0
percaya	0	0	2,60945	0	0
alhamdulillah	0	0	0	2,60945	0
sudah	0	0	0	2,60945	0
ragu	0	0	0	0	2,60945

E. Penyeimbangan Kelas Label (SMOTE)

Metode *synthetic minority oversampling technique* (SMOTE) digunakan jika terdapat data yang tidak seimbang. Teknik ini merupakan teknik penambahan jumlah sampel pada kelas minor dengan melakukan replikasi data pada kelas nimor secara acak sehingga menghasilkan jumlah data yang sama dengan data pada kelas mayor. Data yang direplikasi merupakan data yang berasal dari kelas minor.

Sebagai Contoh akan dijelaskan ilustrasi mengenai metode SMOTE dalam kasus data yang tidak seimbang.

Tabel 10. SMOTE

Teks	Label		Kelas
	vaksin	covid-19	
Teks 1	4	1	Positif
Teks 2	4	2	Positif
Teks 3	5	1	Positif
Teks 4	6	2	Positif
Teks 5	4	1	Positif
Teks 6	5	3	Positif
Teks 7	1	1	Negatif
Text 8	2	4	Negatif
Teks 9	1	2	Negatif
Teks 10	3	4	Netral
Teks 11	3	2	Netral

Tabel diatas adalah data yang akan digunakan untuk menghitung rumus manual SMOTE, data yang digunakan sebagai contoh hanya menggunakan 10 data text, terdiri dari atribut vaksin, covid-19, dan kelas sebagai keterangan. Kemudian dihitung menggunakan rumus SMOTE seperti di bawah ini.

Diketahui input K = 2
N = 2

Output 1. for i in range N = 2 do,
2. select (1,1)
3. random select a neigbhor (1,2)
4. diff = (1,2) – (1,1) = (0,1)
δ = 0,2
5. new sample = (1,1) + [(0,1) x 0,2]
6. synthetic = (1,1) + (0,0.2)
= (1,1.2)

1. for i in range N = 2 do,
2. select (1,2)
3. random select a neigbhor (2,4)
4. diff = (2,4) – (1,2) = (1,2)
δ = 0,2
5. new sample = (1,2) + [(1,2) x 0,2]
6. synthetic = (1,2) + (02,0.4)
= (1.2,2.4)

End for.

Rumus diatas merupakan hasil dari perhitungan manual SMOTE, dimana k tetangga terdekat yang akan dicari adalah 2 dari 11 data text tweet pada tabel 3.10 kemudian nilai N yang akan dicari 2 dari data text tweet, nilai N disini merupakan jumlah data yang akan ditemukan setelah menggunakan teknik k-nearest neighbour seperti pada rumus diatas.

F. Klasifikasi Naïve Bayes

Setelah mendapatkan hasil pembobotan TF-IDF maka tahap selanjutnya yaitu melakukan proses klasifikasi dengan menggunakan metode naïve bayes. Pada proses klasifikasi ini terdapat tiga tahapan untuk mendapatkan nilai probabilitas tertinggi berdasarakan data latih yang dimiliki. Untuk dataset latih yang digunakan menggunakan contoh dataset pada tahapan sebelumnya yang dapat dilihat pada *tabel XI.*

Tabel 11 Data Latih Klasifikasi

Teks	FItur				
	vaksin	harap	...	saya	Label
Teks 1	2,60945	0	...	0	Positif
Teks 2	2,60945	2,60945	...	0	Netral
Teks 3	2,60945	0	...	2,60945	Negatif
Teks 4	2,60945	0	...	0	Positif
Teks 5	2,60945	0	...	2,60945	Negatif

Data uji yang akan digunakan menggunakan kalimat “Saya berharap tidak vaksin” yang tellah melalui praproses terlebih dahulu. Data uji tersebut dapat dilihat pada tabel dibawah ini.

Tabel 12. Data Uji

Sebelum praproses	Sesudah praproses
Saya berharap tidak vaksin	Saya harap tidak vaksin

Tahap 1 : Untuk menghitung class *prior probability* menggunakan persamaan $P(H_i) = \frac{N_i}{N}$. Sebagai contoh dilakukannya perhitungan untuk kelas pertama seperti ini:

$$(H_{\text{Positif}}) = \frac{N_1}{N} \quad (1)$$

$$(H_{\text{Positif}}) = \frac{2}{5}$$

$P(H_{\text{Positif}}) = 0.4$

N_1 adalah banyaknya kemunculan kata pada kelas Positif didalam data latih. contohnya pada kelas positif muncul sebanyak tiga kali. Lalu N adalah banyaknya data pada data latih. Pada contoh data latih ini terdapat 5 data maka nilai dari kelas positif adalah 0,4. Begitupun dengan kelas yang lain untuk perhitungannya sama dengan kelas positif.

Tabel 13. Class Prior Probability

Class prior probability	Hasil
$P(H_{positf})$	0.4
$P(H_{negatif})$	0.2
$P(H_{netral})$	0.4

Tahap 2 : Untuk melakukan tahap selanjutnya dilakukannya perhitungan P(X|H) atau *conditional probability* dengan persamaan $P(X|H_{\text{Positif}}) = P(x_k|H_{\text{Positif}})$.

$$P(X|H_A) = \prod_{k=1}^{n} P(x_k|H_{\text{Positif}}) \quad (2)$$
$$= \prod_{k=1}^{n=3} P(x_k|H_{\text{Positif}})$$

Contoh perhitungannya menggunakan data latih "saya harap tidak vaksin" dengan kelas postif. untuk menghitung nilai dari $P(x_k|H_A)$ menggunakan persamaan sebagai berikut ini:

$$\prod_{k=1}^{n=3} P(x_k|H_{Positif}) = \left(\frac{T_{1\text{Positif}}+1}{T_{\text{Positif}}+\beta}\right) * \dots * \left(\frac{T_{4\text{Positif}}+1}{T_{\text{Positif}}+\beta}\right) \quad (3)$$

$$\text{Saya} = \left(\frac{0+1}{(2{,}60945)+\cdots+(2{,}60945)+17}\right)$$

$$\prod_{k=1}^{n=3} P(x_k|H_{Positif}) = \left(\frac{1}{20{,}8756+17}\right)\left(\frac{1}{37{,}8756}\right)$$

$$\text{Harap} = \left(\frac{0+1}{(2{,}60945)+\cdots+(2{,}60945)+17}\right)$$

$$\prod_{k=1}^{n=3} P(x_k|H_{Positif}) = \left(\frac{1}{20{,}8756+17}\right)\left(\frac{1}{37{,}8756}\right)$$

$$\text{Tidak Vaksin} = \left(\frac{1}{37{,}8756}\right)$$

$$\prod_{k=1}^{n=3} P(x_k|H_{Positif}) = \left(\frac{2{,}60945+2{,}60945+1}{(2{,}60945)+\cdots+(2{,}60945)+17}\right)\left(\frac{5{,}2189+1}{20{,}8756+17}\right)\left(\frac{6{,}2189}{37{,}8756}\right)$$

$$\text{Hasil} = \prod_{k=1}^{n=3} P(x_k|H_{Positif}) = \left(\frac{1}{37{,}8756}\right) * \left(\frac{1}{37{,}8756}\right) * \left(\frac{1}{37{,}8756}\right) * \left(\frac{6{,}2189}{37{,}8756}\right)$$

$$\mathbf{3{,}021867 * 10^{-6}}$$

Dimana $T_{1\text{Positif}}$ – $T_{4\text{Positif}}$ memiliki jumlah bobot kata masing masing dari data uji terhadap kata " saya harap tidak vaksin" dalam kelas Positif. Terdapat bobot kata yang bernilai 0 yaitu saya dan harap, maka nilai tersebut bernilai nol, , kemudian T_A adalah jumlah bobot dari keseluruhan kata dalam kelas Positif, lalu β adalah jumlah dari banyaknya kata. Sehingga hasil dari tahap ini pada kelas Positif adalah $3{,}021867 * 10^{-6}$. Begitupun untuk setiap kelas yang lain, sehingga dapat dilihat pada tabel dibawah ini.

Tabel 14. Conditional probability

Conditional probability	Hasil
$\prod_{k=1}^{n=3} P(x_k\|H_{Positif})$	$3{,}021867 * 10^{-6}$
$\prod_{k=1}^{n=3} P(x_k\|H_{Netral})$	$8{,}229375 * 10^{-6}$
$\prod_{k=1}^{n=3} P(x_k\|H_{Negatif})$	$3{,}015899 * 10^{-5}$

Tahap 3 : Tahapan terakhir yaitu mengalikan hasil perhitungan *class prior probability* dan *conditional probability* sehingga hasilnya dapat dilihat pada tabel dibawah ini.

Tabel 15. Posterior probability

Posterior probability	Class prior probability * Conditional probability	Hasil
$P(H_{Positif}\|X)$	$0.4 * 3{,}021867 * 10^{-6}$	$1{,}20874 * 10^{-6}$
$P(H_{Neral}\|X)$	$0.2 * 8{,}229375 * 10^{-6}$	$4{,}65015 * 10^{-6}$
$P(H_{Negatif}\|X)$	$0.4 * 3{,}015899 * 10^{-5}$	$1{,}20635 * 10^{-5}$

G. Visualisasi Data dengan Geo-Tageed

Tahapan ini dilakukan untuk menampilkan suatu data yang telah diklasifikasi dalam bentuk peta. Data yang telah diambil sebelumnya terdapat longitude dan latitudenya di tiap kabupaten atau kota. Longitude dan latitudenya itu digunakan dalam pembentukan peta sekaligus untuk menampilkan berapa banyak sentimen positif, sentimen netral, atau sentimen negatif ditiap kabupaten atau kotanya. Contoh longitude dan latitutude ditiap kabupatennya terdapat pada tabel dibawah ini.

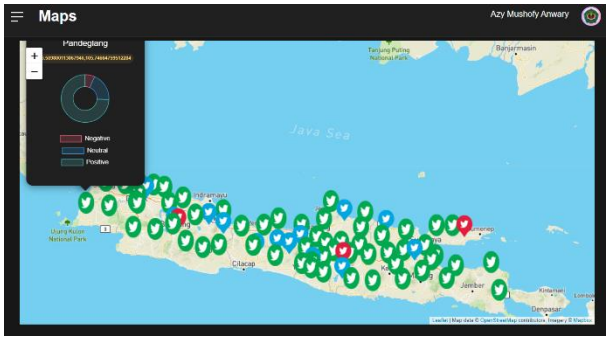
Tabel 16. Longitude Lattitude

Nama Kota	Longitude dan Latitude
Bandung	-7.010637831252968,107.52634034196025
Bekasi	-6.213739385639549,107.12379723294846
Bogor	-6.559089391925341,106.76741460079916
Jakarta Pusat	-6.181592519130218,106.8352147370192
Serang	-6.143175233103424,106.15451430954502
Brebes	-7.058330891657038,108.92961411951245
Blora	-7.074348515501978,111.3847691260547
Purwokerto	-7.43127294583079,109.2331319784154
…	…
Sidoarjo	-7.450231809240246,112.70016879442531

4. HASIL DAN PEMBAHASAN

A. Hasil

Pada tahap ini dilakukannya implementasi dari perancangan sistem yaitu menampilkan sentimen masyarakat mengenai vaksin covid-19 dalam bentuk peta di pulau jawa.



Gambar 1. Visualisasi Data

Dari hasil visualisasi diatas, dapat dilihat bahwa respon masyarakat mengenai sentimen positif di pulau jawa mengenai vaksin covid-19 lebih mendominasi dan mendukung.

B. Evaluasi Hasil / Pengujian Akurasi

Pada tahap ini dilakukan pengukuran kinerja suatu model klasifikasi naive bayes terhadap pengujian data setiap kelas yang bertujuan untuk mengetahui hasil akurasi terbaik dari setiap kelas. Pengujian akurasi merupakan persentase dari total data yang diidentifikasi dan dinilai benar menggunakan *Confusion Matrix* berbentuk tabel matriks.

Tabel 17. Pengujian Akurasi

	Precision (%)	Recall (%)	F1-Score (%)
Negatif	40%	50%	44%
Positif	88%	92%	90%
Netral	68%	49%	57%
Accuracy (%)			79%

Dari hasil pengujian akurasi diatas dapat disimpulkan bahwa tingkat akurasi menggunakan *Synthetic Minority Over-sampling Technique* menghasilkan akurasi sebesar 79%. Data uji yang digunakan sebanyak 227 data tweet dengan pembagian untuk kelas 0 (Negatif) sebanyak 20 data, untuk kelas 1 (Positif) sebanyak 160 data dan untuk kelas 2 (Neteral) sebanyak 47. Walaupun telah menggunakan *Synthetic Minority Over-sampling Technique* tingkat akurasinya hanya naik 2% dengan nilai *presisi* tertinggi pada nilai positif 88%, nilai *recall* tertinggi 92%, dan nilai *f1-score* tertinggi 90%.

5. KESIMPULAN

Berdasarkan hasil penelitian tugas akhir dapat disimpulkan bahwa sentimen masyarakat mengenai vaksin covid-19 di pulau jawa lebih banyak melakukan tweet positif. Model klarifikasi naïve bayes pada pada penelitian ini mendapatkan nilai akurasi (79%) dengan dibantu oleh metode *synthetic minority oversampling technique*. Data yang digunakan sebesar 1132 dataset yang diambil langsung menggunakan Teknik crawling dengan liblary twint. Wilayah yang melakukan tweet terbanyak jatuh kepada wilayah Karawang dengan sentimen positif 70 tweet, sentimen negatif 12 tweet dan sentimen netral 13 tweet. Untuk sentimen positif terbanyak terdapat pada wilayah Karawang, Sentimen negatif terbanyak terdapat pada wilayah Karawang Dan untuk sentimen netral terbanyak terdapat pada wilayah Karawang dan Banyuwangi. Dengan tingkat akurasi sebesar (79%) menunjukan bahwa sentimen analisis menggunakan metode naïve bayes dengan data tweet vaksin covid-19 kurang baik karena dataset memiliki jumlah sentimen positif yang lebih mendominasi oleh karena itu hasil yang didapat lebih kecil dari penelitian sebelumnya.

REFERENSI

[1] M. Wongkar and A. Angdresey, “Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter,” *Proc. 2019 4th Int. Conf. Informatics Comput. ICIC 2019*, pp. 1–5, 2019, doi: 10.1109/ICIC47613.2019.8985884.

[2] V. Chandani, F. I. Komputer, and U. D. Nuswantoro, “Komparasi Algoritma Klasifikasi Machine Learning Dan Feature Selection pada Analisis Sentimen Review Film,” *J. Intell. Syst.*, vol. 1, no. 1, pp. 56–60, 2015.

[3] A. I. H. Raflialdy Raksanagara, Yulison Herry Chrisnanto, “Analisis Sentimen Jasa Ekspedisi Barang Menggunakan Metode Naïve Bayes,” *Pros. SNST Fak. Tek.*, vol. 1, no. 1, 2016.

[4] R. K. Poluru, B. Bhushan, B. S. Muzamil, P. K.

Rayani, and P. K. Reddy, "Applications of Domain-Specific Predictive Analytics Applied to Big Data," no. January 2019, pp. 289–306, 2018, doi: 10.4018/978-1-5225-4999-4.ch016.

[5] D. G. Nugroho, Y. H. Chrisnanto, and A. Wahana, "Analisis Sentimen Pada Jasa Ojek Online ... (Nugroho dkk.)," pp. 156–161, 2015.

[6] W. L. Lim, C. C. Ho, and C.-Y. Ting, "Sentiment Analysis by Fusing Text and Location Features of Geo-Tagged Tweets," *IEEE Access*, vol. 8, no. September, pp. 181014–181027, 2020, doi: 10.1109/access.2020.3027845.

[7] Y. Sharma and V. Mangat, "Pendekatan Praktis untuk Analisis Sentimen Tweet Hindi," no. September, pp. 4–5, 2015.

[8] A. Puri, P. Arora, and N. Sardana, "Analysis and Visualisation of Geo-Referenced Tweets for Real-Time Information Diffusion," *Procedia Comput. Sci.*, vol. 132, no. Iccids, pp. 1138–1146, 2018, doi: 10.1016/j.procs.2018.05.028.

[9] V. Amrizal, "Penerapan Metode Term Frequency Inverse Document Frequency (Tf-Idf) Dan Cosine Similarity Pada Sistem Temu Kembali Informasi Untuk Mengetahui Syarah Hadits Berbasis Web (Studi Kasus: Hadits Shahih Bukhari-Muslim)," *J. Tek. Inform.*, vol. 11, no. 2, pp. 149–164, 2018, doi: 10.15408/jti.v11i2.8623.

[10] Z. Song and C. Xia, "Analisis Sentimen Spasial dan Temporal data Twitter," pp. 205–221.

[11] W. Yulita *et al.*, "Analisis Sentimen Terhadap Opini Masyarakat Tentang Vaksin Covid-19 Menggunakan Algoritma Naïve Bayes Classifier," *Jdmsi*, vol. 2, no. 2, pp. 1–9, 2021.

[12] A. Makmun and S. F. Hazhiyah, "Tinjauan Terkait Pengembangan Vaksin Covid 19," *Molucca Medica*, vol. 13, pp. 52–59, 2020, doi: 10.30598/molmed.2020.v13.i2.52.

[13] A. Deviyanto and M. D. R. Wahyudi, "Penerapan Analisis Sentimen Pada Pengguna Twitter Menggunakan Metode K-Nearest Neighbor," *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 3, no. 1, p. 1, 2018, doi: 10.14421/jiska.2018.31-01.

[14] E. Sutoyo, M. Asri Fadlurrahman, J. Telekomunikasi Jl Terusan Buah Batu, K. Dayeuhkolot, K. Bandung, and J. Barat, "JEPIN (Jurnal Edukasi dan Penelitian Informatika) Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Television Advertisement Performance Rating Menggunakan Artificial Neural Network."

[15] A. Nur Rais, W. Kurniawan, R. Ardianto, S. Informasi Akuntansi Universitas Bina Sarana Informatika, J. Kramat Raya No, and J. Pusat, "Analisa Akurasi Dan F1 Score Pada Algoritma Smote Dan Naïve Bayes Pada Dataset Bank Direct Marketing," *J. Speed-Sentra Penelit. Eng. dan Edukasi*, vol. 11, no. 4, 2019.

[16] M. A. F. Prananda Antinasari, Rizal Setya Perdana, "Analisis Sentimen Tentang Opini Film Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Naive Bayes Dengan Perbaikan Kata Tidak Baku," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 12, pp. 1733–1741, 2017, [Online]. Available: http://j-ptiik.ub.ac.id.

[17] T. E. Putri, R. T. Subagio, Kusnadi, and P. Sobiki, "Classification System of Toddler Nutrition Status using Naïve Bayes Classifier Based on Z- Score Value and Anthropometry Index," *J. Phys. Conf. Ser.*, vol. 1641, no. 1, 2020, doi: 10.1088/1742-6596/1641/1/012005.

[18] H. T. Sueno, "Converting Text to Numerical RepresentationusingModified Bayesian Vectorization Technique for Multi-Class Classification," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 4, pp. 5618–5623, 2020, doi: 10.30534/ijatcse/2020/211942020.

[19] D. Twitter, "Tweet geospatial metadata | Twitter Developer." https://developer.twitter.com/en/docs/tutorials/tweet-geo-metadata.

[20] D. Twitter, "Filtering Tweets by location | Twitter Developer." https://developer.twitter.com/en/docs/tutorials/filtering-tweets-by-location.

[21] A. Andriani, "Sistem Pendukung Keputusan Berbasis Decision Tree Dalam Pemberian Beasiswa Studi Kasus : Amik ' BSI Yogyakarta ,'" *Semin. Nas. Teknol. Inf. dan Komun. 2013 (SENTIKA 2013)*, vol. 2013, no. SENTIKA, pp. 163–168, 2013, [Online]. Available: https://repository.bsi.ac.id/index.php/unduh/item/48930/Sentika_2013Anik-Andriani.pdf.