e-ISSN 2962-6129
Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)
Jakarta-Indonesia, 20 Agustus 2022

Analisis Sentimen Terhadap Ppkm Darurat Pada Media Sosial *Twitter* Menggunakan Metode *Naïve Bayes* Dengan Seleksi Fitur *Information Gain*

Albet Dwi Pangestu [1], Iin Ernawati S.Kom., M.Si. [2], Nurul Chamidah S.Kom, M.Kom. [3],
S1 Informatika / Fakultas Ilmu Komputer
Program Studi Informatika, Universitas Pembangunan Nasional Veteran Jakarta Jl. RS. Fatmawati Raya, Pd. Labu, Kec. Cilandak, Kota Depok, Jawa Barat 12450
email: albetdp@upnvj.co.id [1], iinernawati@upnvj.ac.id [2], nurul.chamidah@upnvj.ac.id [3]

Abstrak. *Twitter* merupakan media sosial yang digunakan masyarakat sebagai media berkomunikasi dan pengutaraan pendapat. Semenjak pandemik COVID-19 melanda Indonesia, pemerintah banyak mengeluarkan kebijakan-kebijakan untuk menekan penyebaran COVID-19, salah satunya adalah PPKM Darurat. Banyak opini masyarakat yang mengkritik ataupun mendukung kebijakan tersebut di media sosial, khususnya *twitter*. Penelitian ini bertujuan untuk membangun model analisis sentimen terhadap PPKM Darurat di media sosial *twitter* dengan tagar #ppkmdarurat. Pada penelitian ini akan menggunakan metode *Naïve Bayes* serta metode *Information Gain* sebagai seleksi fitur. Pengumpulan data akan dilakukan dengan cara *crawling.* Setelah dilakukan *filtering* data menjadi 770 dengan label 335 positif dan 335 negatif. Hasil dari pengujian model klasifikasi *Naïve Bayes* terjadi peningkatan performa apabila menggunakan seleksi fitur *Information Gain* dengan nilai pengambilan *top ranking* '>0.0001' yaitu akurasi 0.81, *recall* 0.82, *precision* 0.84, *f1 score* 0.83 dan *specificity* 0.79 dibandingkan sebelumnya yaitu akurasi 0.79, *recall* 0.81, *precision* 0.81, *f1 score* 0.81 dan *specificity* 0.76.

Kata kunci: Analisis sentimen, PPKM Darurat, *Twitter*, *Naïve Bayes*, *Information Gain*

1. Pendahuluan

Berkat majunya teknologi di masa sekarang, masyarakat dimudahkan untuk bersosialisasi antar individu menggunakan media sosial tanpa dibatasi oleh ruang dan waktu. Selain digunakan untuk bersosialisasi, media sosial dapat dimanfaatkan sebagai sarana masyarakat dalam mengutarakan pendapat. Media sosial yang sering digunakan masyarakat untuk menyampaikan pendapat atau pandangannya adalah *twitter*. *Twitter* merupakan media sosial yang sangat terkenal dikalangan masyarakat global, khususnya masyarakat Indonesia. Menurut data Statista pada bulan Juli 2021, negara Indonesia menempati posisi ke 6 pengguna *twitter* terbanyak dengan 15,7 juta pengguna [1].

Semenjak pandemik COVID-19 melanda Indonesia, pemerintah telah lakukan segala cara untuk mencegah penyebaran COVID-19 di masyarakat. Usaha pemerintah untuk menekan penyebaran COVID-19 adalah dengan mengeluarkan kebijakan-kebijakan atau peraturan, salah satunya ialah PPKM Darurat. Kebijakan PPKM Darurat telah mulai diberlakukan pada tanggal 3 – 20 Juli 2021 dengan menyasar daerah Jawa dan Bali. Kebijakan PPKM Darurat menimbulkan banyak opini masyarakat di media sosial, khususnya media sosial *twitter*. Banyak yang beropini mendukung kebijakan PPKM Darurat atau malah sebaliknya. Pihak yang sangat terdampak dengan adanya kebijakan PPKM Darurat ini adalah UMKM (Usaha Mikro, Kecil dan Menengah). Pelaku UMKM mengalami penurunan omset yang diakibatkan sepinya pengunjung.

Metode *Naïve Bayes* adalah salah satu metode yang cukup baik dalam analisis sentimen. Terlihat pada penelitian yang dilakukan oleh Julianto *et al.* menggunakan metode *Naïve Bayes* dalam analisis sentimen mengenai pemerintahan Joko Widodo. Hasil pada penelitian ini, didapatkan nilai akurasi sebesar 79 % [2]. Penelitian dilakukan oleh Syarifuddin dengan membandingkan performa antara metode *Naïve Bayes* dan *K-Nearest Neighbor*, didapatkan hasil yaitu nilai akurasi *Naïve Bayes* 63.21% dan *K-Nearest Neighbor* 58.10% [3]. Metode *Naïve Bayes* mampu mengolah data *tweet* dengan jumlah besar. Metode *Naïve Bayes* memiliki kelemahan yaitu, sensitif terhadap fitur yang terlalu banyak. Oleh karena itu, pemilihan fitur perlu dilakukan dengan menggunakan metode *Information Gain.* Seleksi fitur menggunakan *Information Gain* dapat meningkatkan nilai akurasi dari metode *Naïve Bayes.* Penelitian yang dilakukan oleh Bijaksana *et al*., analisis sentimen menggunakan metode *Naïve Bayes* dan dikombinasikan dengan seleksi fitur *Information Gain* pada ulasan customer maskapai penerbangan. Hasil dari penelitian menunjukan bahwa perbandingan nilai akurasi metode *Naïve Bayes* menggunakan seleksi fitur *Information Gain* 86,5% sedangkan tanpa menggunakan seleksi fitur 81%. Terdapat kenaikan dari nilai akurasi apabila penggunaan seleksi fitur *Information Gain* sebesar 5,5%[4].

Berdasarkan latar belakang dan studi literatur yang telah dilangsungkan, oleh karena itu perlu adanya suatu model analisis sentimen terhadap opini publik mengenai kebijakan PPKM Darurat di media sosial *twitter*. Dalam melakukan analisis sentimen, penelitian ini akan memakai metode *Naïve Bayes* dan seleksi fitur menggunakan metode *Information Gain*.

2. Landasan Teori

2.1. Analisis Sentimen

Analisis sentimen adalah menganalisis dan pemodelan dari sentimen, emosi, serta opini dalam bentuk penyampaian secara tekstual [5].

2.2. *Twitter*

Twitter merupakan suatu media sosial yang memberikan layanan *microblogging* sehingga pengguna dapat menulis pesan atau *tweet* secara *real time* [6].

2.3. PPKM Darurat

PPKM Darurat merupakan suatu peraturan yang dikeluarkan oleh pemerintah Indonesia, dengan tujuan menekan penyebaran COVID-19 di kalangan masyarakat. Peraturan ini diberlakukan selama dua pekan, terhitung pada tanggal 3 - 20 Juli 2021 yang menyasar di Jawa dan Bali [7].

2.4. *Preprocessing*

Preprocessing merupakan tahapan pertama dalam *text mining*. *Preprocessing* bertujuan untuk penyiapan data sehingga data dapat digunakan pada tahapan selanjutnya [8].
Tahapan atau proses yang dilakukan pada *preprocessing* ini adalah sebagai berikut:

a. *Cleaning*, merupakan suatu proses dengan tujuan menghilangkan kata tidak dibutuhkan seperti kata kunci, HTML, *hashtags* (#), *username*, email dan simbol atau tanda baca [9].
b. *Case Folding*, adalah langkah yang dilakukan untuk mengubah semua kalimat dalam data menjadi huruf kecil (lowercase) [8].
c. Normalisasi Bahasa, menurut Adiyasa (2013) dalam Buntoro, Normalisasi bahasa adalah tahapan untuk merubah bahasa yang tidak baku atau bahasa gaul menjadi bahasa baku yang sesuai dengan ketentuan KBBI (Kamus Besar Bahasa Indonesia) [10].
d. *Stopword Removal*, merupakan langkah yang bertujuan untuk menghilangkan kata-kata yang sering muncul pada data, tetapi kata tersebut tidak berpengaruh secara signifikan dan semantik pada data [11].
e. *Stemming*, merupakan sebuah proses menghilangkan imbuhan pada suatu kata dengan mencari akar dari kata tersebut, sehingga menjadi kata dasar [12].
f. Tokenisasi, merupakan suatu tahapan yang bertujuan untuk pemotongan teks berdasarkan tiap kata dan disebut sebagai token[5].

2.5. Pembobotan *Trem* (TF-IDF)

Term Frequency – Inverse Document Frequency (TF-IDF) merupakan pemodelan yang dimanfaatkan untuk menemukan hubungan antara *term* terhadap data dengan cara memberikan bobot pada setiap kata atau *term* [13].
Rumus mencari TF, yaitu sebagai berikut :

$$TF(k) = \sum Fk, d \quad \textbf{(1)}$$

Keterangan :
F = Frekuensi kemunculan
k = Kata
d = Dokumen
Rumus mencari IDF, yaitu sebagai berikut :

$$IDF(k) = Log \frac{D}{Df(k)} \quad \textbf{(2)}$$

Keterangan :
D = Total jumlah dokumen
DF(k) = Total dokumen yang mengandung frekuensi kemunculan kata k
Rumus mencari TF-IDF, yaitu sebagai berikut :

$$TF - IDF\ (k) = TF(k) \times IDF\ (k) \quad \textbf{(3)}$$

2.6. *Information Gain*

Information Gain adalah pemodelan pemilihan fitur dengan melakukan *rangking* pada setiap fitur yang ada. Metode *Information Gain* banyak digunakan pada klasifikasi teks, analisis data citra, dan analisis data *microarray*. Penggunaan metode *Information Gain* dalam seleksi fitur juga dapat mengurangi jumlah *noise* pada data disebabkan oleh label yang tidak tepat [14].
Rumus hitung nilai *entropy*, yaitu sebagai berikut:

$$Entropy\ (S) = \sum_{i=1}^{c} - p_i \cdot log_2\ p_i \quad \textbf{(4)}$$

Rumus hitung nilai *entropy* pada fitur, yaitu sebagai berikut:

$$Entropy_F(S) = \sum_{j=1}^{n} - \frac{|D_j|}{D} \cdot Entropy(S_j) \quad \textbf{(5)}$$

Rumus hitung nilai *Information Gain*, yaitu sebagai berikut:

$$IG(S, F) = Entropy(D) - |Entropy_F(S)| \quad \textbf{(6)}$$

Keterangan:
S = Sampel
c = Jumlah fitur target
F = Fitur
n = Jumlah nilai pada kelas
pi = Proporsi jumlah sampel pada kelas i dengan sampel data
|Dj| = Jumlah sampel data untuk nilai partisi j
D = Jumlah sampel data
Entropy (Sj) = Nilai *Entropy* pada sampel j

2.7. Klasifikasi *Naïve Bayes*

Naïve Bayes adalah pengklasifikasi berdasarkan algoritma bayes yang mengasumsikan semua variabel bersifat individu, sehingga tidak memiliki kaitan dengan variabel lain [2].

$$P(a|b) = \frac{P(a) \times P(b|a)}{P(b)} \quad \textbf{(7)}$$

Keterangan:
a = Hipotesis data yang merupakan suatu kelas spesifik.
b = Data dengan kelas yang belum diketahui.
P(a|b) = Probabilitas hipotesis a jika evidence b terjadi (posteriori probability)
P(b|a) = Probabilitas munculnya evidence b berdasar kondisi hipotesis a.
P(a) = Probabilitas hipotesis a tanpa memandang evidence apapun.
P(b) = Probabilitas evidence b tanpa memandang apapun.
Rumus yang menghitung nilai probabilitas setiap kelas, yaitu sebagai berikut:

$$P(a) = \frac{|dok\ a|}{|dokumen|} \quad \textbf{(8)}$$

Keterangan:
P(a) = Probabilitas kemunculan suatu data yang memiliki kelas a.
|dok a| = Total dari dokumen untuk tipe kelas a.
|dokumen| = Total dari setiap kelas.
Rumus untuk mengukur peluang pada setiap kata dalam dokumen berdasarkan kategori adalah sebagai berikut:

$$P(W_i|a) = \frac{Count(W_i,a)+1}{|a|+|V|} \quad \textbf{(9)}$$

Keterangan:
Wi = Kata
a = Kelas
P(Wi|a) = Peluang kata Wi pada kelas a.
Count(Wi,a) = Total kemunculan kata Wi pada kelas a.

|a| = Total keseluruhan kata pada kelas a.
|V| = Total keseluruhan kata (trem)
Untuk proses klasifikasi data uji dapat menggunakan persamaan, yaitu sebagai berikut:

$$a_{MAP} = \underset{a \in V}{arg\,max}\ P(a) \prod_i P(W_i|a) \quad \textbf{(10)}$$

Keterangan:
a = Kelas
P = Peluang
Wi = Kata
P(a) = Peluang kemunculan suatu dokumen yang memiliki kelas a.
P(Wi|a) = Peluang kata Wi pada kelas a.

2.8. Evaluasi

Evaluasi merupakan tahapan dengan bertujuan untuk mengetahui kebenaran hasil metode yang dipakai, dengan cara menghitung nilai yang diperoleh. Untuk melakukan tahapan evaluasi ini menggunakan perhitungan *confusion matrix* [9].
Berikut merupakan tabel perhitungan dari *confusion matrix*:

Tabel 1. *Confusion matrix*

		Nilai Aktual	
		Positif	**Negatif**
Nilai Prediksi	**Positif**	*True Positive (TP)*	*False Positif (FP)*
	Negatif	*False Negatif (FN)*	*True Negatif (TN)*

Keterangan:
True Positive (TP) = Data dengan aktual positif dan diprediksi positif
False Positive (FP) = Data dengan aktual negatif dan diprediksi positif
False Negative (FN) = Data dengan aktual positif dan diprediksi negatif
True Negative (TN) = Data dengan aktual negatif dan diprediksi negatif
Rumus perhitungan nilai akurasi, yaitu sebagai berikut:

$$Akurasi = \frac{TP+TN}{TP+FN+FP+TN} \quad \textbf{(11)}$$

Rumus perhitungan nilai *recall*, yaitu sebagai berikut:

$$Recall = \frac{TP}{TP+FN} \quad \textbf{(12)}$$

Rumus perhitungan nilai *precision*, yaitu sebagai berikut:

$$Precision = \frac{TP}{TP+FP} \quad \textbf{(13)}$$

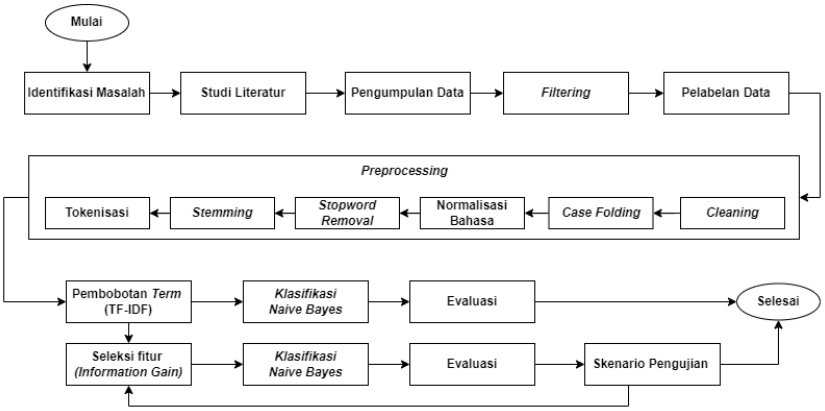
Rumus perhitungan nilai *f1 score*, yaitu sebagai berikut :

$$F1\ Score = 2\ \frac{Precision\ x\ Recall}{Precision + Recall} \quad \textbf{(14)}$$

Rumus Perhitungan nilai *specificity*, yaitu sebagai berikut :

$$Specificity = \frac{TN}{TN+FP} \quad \textbf{(15)}$$

3. Metodologi Penelitian



Gambar. 1. adalah proses alur dalam penelitian yang akan dilakukan. Berikut ini merupakan penjelasan setiap alur prosesnya.

3.1. Identifikasi Masalah

Identifikasi masalah merupakan sebuah proses atau tahapan untuk memperjelas masalah yang akan dibahas dalam penelitian. Permasalahan yang akan dikupas dari penelitian ini adalah, mengklasifikasikan data *tweet* opini masyarakat mengenai PPKM Darurat menggunakan metode *Naïve Bayes*, serta seberapa besar pengaruh seleksi fitur *Information Gain* dalam meningkatkan nilai akurasi dari metode *Naïve Bayes*.

3.2. Studi Literatur

Studi literatur merupakan suatu proses pengumpulan teori-teori mendasar dari berbagai sumber berkaitan dengan penelitian dan akan digunakan sebagai acuan serta pengetahuan dalam penelitian ini. Sumber pustaka dalam penelitian ini berasal dari *e-book*, jurnal, website, artikel berita, dan lain-lainnya yang sejalan dengan penelitian.

3.3. Pengumpulan Data

Pengumpulan data dilakukan bertujuan untuk mendapatkan data yang akan digunakan pada penelitian. Data yang digunakan pada penelitian ini adalah *tweet* opini masyarakat mengenai PPKM Darurat pada sosial media *twitter*.

3.4. *Filtering*

Data yang telah berhasil di *crawling* masih diperlukan proses *filtering*. Proses *filtering* ini bertujuan untuk menghapus *tweet double* serta *tweet* yang tidak memiliki sentimen seperti *tweet* netral, iklan, berita dan lain-lain. Proses *filtering* ini dilakukan secara manual oleh peneliti.

3.5. Pelabelan Data

Proses pelabelan data dilakukan bertujuan untuk memberikan label positif apabila *tweet* mendukung kebijakan PPKM Darurat dan label negatif diberikan apabila *tweet* tidak pendukung kebijakan PPKM Darurat.

3.6. *Preprocessing*

Berikut ini merupakan tahapan *preprocessing*:

a. *Cleaning*, proses cleaning bertujuan untuk membersihkan data *tweet* dengan cara menghapus karakter yang tidak digunakan pada penelitian atau tidak memiliki *value* seperti URL, tanda baca, angka, *hashtag*, *username*, dan *emoji*.
b. *Case Folding*, pada tahapan *case folding* setiap kata data *tweet* akan diubah menjadi huruf kecil.
c. Normalisasi Bahasa, merupakan tahapan untuk merubah kata-kata yang tidak baku seperti kata singkatan atau kata gaul (*slang word*) menjadi kata baku sesuai dengan aturan KBBI (Kamus Besar Bahasa Indonesia).

d. *Stopword Removal*, pada tahapan *stopword removal* digunakan sebagai proses untuk menghapus kata-kata yang tidak terlalu penting dan tidak memiliki makna terhadap analisis sentimen.
e. *Stemming*, pada proses *stemming* akan merubah setiap kata menjadi kata dasar dengan cara menghapus kata imbuhan.
f. Tokenisasi, pada tahapan tokenisasi *tweet* akan dipisah menjadi tiap-tiap kata atau sering disebut dengan token.

3.7. Pembobotan *Trem* (TF-IDF)

Pada proses ini dilakukan setelah data *tweet* telah melewati tahap *preprocessing*. Tahapan pembobotan *term* ini bertujuan untuk memberikan nilai kepada setiap kata agar dapat dilakukan klasifikasi metode *Naïve Bayes*. Perhitungan pada pembobotan *term* ini menggunakan TF-IDF (*Term Frequency – Inverse Document Frequency*). Perhitungan TF-IDF dilakukan dengan rumus persamaan (1), (2), dan (3).

3.8. Seleksi Fitur *Information Gain*

Tahapan seleksi fitur merupakan suatu proses penyeleksian fitur dengan nilai terbaik untuk analisis sentimen. Metode yang digunakan adalah *Information Gain*. Berikut proses *Information Gain* untuk seleksi fitur:

a. Pemisahan Fitur Sesuai Label, tahapan ini merupakan tahapan pertama yang akan berisi variasi *record* pada setiap fitur, label positif dan negatif untuk setiap kemunculan variasi *record*, dan total semua variasi *record* pada setiap fitur beserta label.
b. Hitung Nilai *Entropy*, pada tahapan ini merupakan tahapan untuk menghitung nilai *entropy* pada setiap *record* dan total semua variasi *record* pada setiap fitur beserta label. Nilai dari hasil perhitungan *entropy* ini akan digunakan pada tahapan selanjutnya yaitu perhitungan nilai *Information Gain*. Perhitungan nilai *entropy* dilakukan dengan rumus persamaan (4).
c. Hitung Nilai *Entropy* Pada Fitur, tahapan ini merupakan tahapan untuk menghitung nilai *entropy* pada setiap fitur. Nilai dari hasil perhitungan *entropy* pada setiap fitur ini akan digunakan pada tahapan selanjutnya yaitu perhitungan nilai *Information Gain* dengan mengacu pada rumus persamaan (5).
d. Hitung Nilai *Information Gain*, pada tahapan ini merupakan tahapan perhitungan untuk nilai *Information Gain*. Nilai hasil dari perhitungan pada tahapan ini akan menentukan setiap posisi suatu fitur. Nilai hasil dari perhitungan pada tahapan ini akan menentukan setiap posisi suatu fitur. Perhitungan nilai *Information Gain* dilakukan dengan rumus persamaan (6).
e. Pengurutan Serta Pengambilan *Top ranking*, pada tahapan ini merupakan tahapan pengurutan nilai hasil dari perhitungan *Information Gain*. Pengurutan ini akan dilakukan untuk mengetahui fitur mana yang memiliki nilai yang baik dan yang tidak, dengan pengurutan dari kecil ke besar. Pengambilan fitur akan memilih fitur dengan nilai terbaik atau *top ranking* sesuai dengan nilai yang ditentukan.

3.9. Klasifikasi *Naïve Bayes*

Tahapan klasifikasi akan dilakukan sebanyak 2 kali yaitu, data *tweet* tanpa menggunakan seleksi fitur *Information Gain* dan data *tweet* yang menggunakan seleksi fitur *Information Gain*. Sebelum masuk pada tahapan klasifikasi, data *tweet* akan dibagi menjadi data *training* dan data *testing*. Data *training* akan digunakan sebagai data latih model klasifikasi *Naïve Bayes* sebanyak 90% sedangkan data *testing* akan digunakan sebagai evaluasi model klasifikasi *Naïve Bayes* sebanyak 10%. Perhitungan metode *Naïve Bayes* dilakukan dengan rumus persamaan (7), (8), (9) dan (10).

3.10. Evaluasi

Model pengkasifikasian menggunakan metode *Naïve Bayes* akan melakukan tahapan evaluasi, agar mengetahui seberapa besar performa yang dihasilkan. Pengukuran performa metode *Naïve Bayes* akan menggunakan *confusion matrix* dengan perhitungan nilai tingkat akurasi, *recall*, *precision*, *f1 score* dan *specificity*. Perhitungan nilai akurasi, *recall*, *precision*, *f1 score* dan *specificity*. Perhitungan nilai akurasi, *recall*, *precision*, *f1 score* dan *specificity* dilakukan dengan rumus persamaan (11), (12), (13), (14), dan (15).

3.11. Skenario Pengujian

Tahapan skenario pengujian dilakukan dengan tujuan mendapatkan pola terbaik dalam menentukan nilai pengambilan *top ranking* untuk seleksi fitur *Information Gain* dan membandingkan tanpa penggunaan seleksi fitur *Information Gain* dengan menggunakan seleksi fitur *Information Gain*.

4. Hasil dan Pembahasan

Data yang digunakan berupa *teks tweet* dari *twitter* dengan kata kunci #ppkmdarurat yang didapatkan dengan cara *crawling* menggunakan bahasa pemrograman R, dan terhubung dengan *Application Programming Interface* (API) yang telah difasilitasi oleh *twitter* di website *developer.twitter.com*. Proses *crawling* data *tweet* dilakukan sebanyak 5 kali pada tanggal 3, 10, 17, 24, 31 Juli 2021 dan didapatkan sebanyak 5000 *tweet* dan akan disimpan dengan format .csv. Setelah data berhasil di *crawling*, selanjutnya akan melalui proses *filtering* dengan tujuan menghapus *tweet* yang tidak digunakan dalam analisis sentimen seperti *tweet* netral, berita, iklan, *tweet double*, dan lain-lain. Setelah dilakukan *filtering*, dari 5000 *tweet* hasil *crawling* menjadi 770 *tweet* dan akan digunakan pada proses selanjutnya yaitu pelabelan. Setelah dilakukan pelabelan oleh tiga orang penilai, dari 770 *tweet* terdapat 385 *tweet* berlabel positif dan 385 *tweet* berlabel negatif. Setelah melalui tahapan pelabelan, selanjutnya proses *preprocessing* dengan 6 tahapan yaitu *Cleaning*, *Case Folding*, Normalisasi Bahasa, *Stopword Removal*, *Stemming* dan Tokenisasi. Berikut ini merupakan tabel salah satu contoh *tweet* dalam tahapan *preprocessing*.

Tabel 2. *Preprocessing*

Sebelum	Sesudah
mencegah kita dari tertular covid dukung penerapan ppkm darurat	['cegah', 'tular', 'covid', 'dukung', 'terap', 'ppkm', 'darurat']

Tabel 2 merupakan hasil dari tahapan *preprocessing* dengan kolom “sebelum” merupakan *tweet* belum melalui proses *preprocessing* dan kolom “sesudah” merupakan *tweet* yang sudah melalui proses *preprocessing*.

Setelah data dibersihkan, untuk melihat kata-kata yang memiliki volume kemunculan terbesar atau sering muncul dilakukan visualisasi menggunakan *word cloud*. Berikut ini gambar visualisasi *word cloud* data *tweet* berlabel positif dan negatif.

(a) (b)

Gambar. 2. visualisasi *wordc cloud* (a) *tweet* berlabel positif dan (b) *tweet* berlabel negatif. Pada gambar 2 diatas menunjukan tiap-tiap kata yang memiliki volume kemunculan yang lebih jelas. Seperti pada gambar (a), kata yang memiliki volume kemunculan yang besar adalah ppkm, darurat, dukung, perintah, covid, masyarakat dan lain-lain. Sedangkan pada gambar (b), kata yang memiliki volume kemunculan yang besar adalah ppkm, darurat, panjang, perintah, rakyat, tutup, covid dan lain-lain.

Setelah data *tweet* sudah melalui proses *preprocessing*, tahapan selanjutnya yaitu pembobotan *term* menggunakan metode *Term Frequency - Inverse Document Frequency* (TF-IDF). Terdapat 1899 kata dari 770 *tweet* yang akan melalui tahapan pembobotan *term*. Setelah melalui tahapan (TF-IDF) akan masuk pada proses seleksi fitur menggunakan metode *Information Gain.* Terdapat 5 tahapan yang akan dilakukan metode *Information Gain* dalam menyeleksi fitur yaitu memisahkan fitur sesuai label, hitung nilai *entropy*, hitung nilai *entropy* pada fitur, hitung nilai *Information Gain*, dan pengurutan serta pengambilan *top ranking*. Pada proses pengambilan *top ranking* akan menggunakan pengujian dengan nilai ‘>0.01’, ‘>0.001’, dan ‘>0.0001’. Proses selanjutnya adalah klasifikasi menggunakan metode *Naïve Bayes*. Proses klasifikasi akan dilakukan dua kali dengan input dari TF-IDF dan seleksi fitur *Information Gain*. Tujuan dilakukannya dua kali input dalam model yaitu untuk mengetahui seberapa besar pengaruh seleksi fitur *Information Gain* dalam meningkatkan performa model yang telah dibuat. Sebelum data diproses model klasifikasi, data akan dibagi menjadi data *training* dan data *testing* dengan perbandingan 90 :10. Berikut ini merupakan tabel pembagian data *training* dan *testing*.

Tabel 3. Pembagian Data *Training* dan *Testing*

	Label Positif	**Label Negatif**	**Total**
Data *Training*	346	346	692
Data *Testing*	39	39	78
Total	385	385	770

Tabel 3 di atas, pembagian data *training* dan *testing*, *tweet* data *training* berlabel positif dan negatif berjumlah 692 *tweet*, sedangkan data *testing* berlabel positif dan negatif berjumlah 78 *tweet*.

Setelah data dibagi, proses klasifikasi ini akan menggunakan algoritma *Multinomial Naïve Bayes* dan akan dievaluasi menggunakan metode *confusion matrix*. Berikut ini merupakan tabel evaluasi tanpa menggunakan seleksi fitur *Information Gain*.

Tabel 4. *Confusion matrix* Tanpa *Information Gain*

		Nilai Aktual	
		Positif	Negatif
Nilai Prediksi	Positif	35	8
	Negatif	8	26

Tabel 4 merupakan *confusion matrix* tanpa menggunakan *Information Gain*. Dari Tabel 4.21 diketahui nilai TP = 35, FP = 8, FN = 8, dan TN = 26. Berikut ini perhitungan untuk mencari nilai akurasi, *recall*, *precision*, *f1 score* dan *specificity*.

$$Akurasi = \frac{TP+TN}{TP+FN+FP+TN} = \frac{35+26}{35+8+8+26} = \frac{61}{77} = 0.79$$

$$Recall = \frac{TP}{TP+FN} = \frac{35}{35+8} = \frac{35}{43} = 0.81$$

$$Precision = \frac{TP}{TP+FP} = \frac{35}{35+8} = \frac{35}{43} = 0.81$$

$$F1\ Score = 2\ \frac{Precision\ x\ Recall}{Precision\ +\ Recall} = 2\ \frac{0.81\ x\ 0.81}{0.81\ +\ 0.81} = \frac{1.321}{1.620} = 0.81$$

$$Specificity = \frac{TN}{TN+FP} = \frac{26}{26+8} = \frac{26}{34} = 0.76$$

Dari perhitungan diatas, didapatkan hasil nilai akurasi 0.79, *recall* 0.81, *precision* 0.81, *f1 score* 0.81 dan *specificity* 0.76. Pada tahapan terakhir yaitu skenario pengujian yang bertujuan untuk untuk mengetahui seberapa besar pengaruh penggunaan seleksi fitur *Information Gain* dalam meningkatkan performa metode *Naïve Bayes*. Skenario pengujian dilakukan tanpa menggunakan seleksi fitur dan menggunakan seleksi fitur *Information Gain* dengan pengambilan *top ranking* yaitu '>0.01', '>0.001' dan '>0.0001'. Berikut ini tabel hasil skenario pengujian.

Tabel 5. Hasil Skenario Pengujian

Confusion Matrix	**Tanpa Menggunakan *Information Gain***	**Menggunakan *Information Gain***		
		>0.01	>0.001	>0.0001
Akurasi	0.79	0.73	0.77	0.81
Recall	0.81	0.78	0.80	0.82
Precision	0.81	0.70	0.77	0.84
F1 Score	0.81	0.74	0.79	0.83
Specificity	0.76	0.66	0.72	0.79

Dari Tabel 5 di atas menunjukan adanya peningkatan performa apabila menggunakan seleksi fitur *Information Gain* dengan nilai pengambilan *top ranking* '>0.0001' yaitu akurasi 0.81, *recall* 0.82, *precision* 0.84, *f1 score* 0.83 dan *specificity* 0.79 dibandingkan sebelumnya yaitu akurasi 0.79, *recall* 0.81, *precision* 0.81, *f1 score* 0.81 dan *specificity* 0.76. Berbeda halnya apabila menggunakan seleksi fitur dengan pengambilan *top ranking* '>0.01' dan '>0.001' terjadi penurunan performa. Penurunan performa ini disebabkan fitur yang terseleksi terlalu banyak. Berikut ini Tabel 6 ringkasan jumlah fitur.

Tabel 6. Ringkasan Jumlah Fitur

	Jumlah Fitur
Tanpa Menggunakan IG	1899
>0.01	94
>0.001	894
>0.0001	1878

Dari Tabel 6 diatas, menunjukan terjadinya perbedaan jumlah fitur apabila tidak menggunakan seleksi fitur *Information Gain* dan menggunakan seleksi fitur *Information Gain* dengan pengambilan nilai *top ranking* '>0.01', '>0.001', dan '>0.0001'. Jumlah fitur tanpa menggunakan *Information Gain* = 1,873, '>0.01' = 94, '>0.001' = 894, dan '>0.0001' = 1,878.

5. Kesimpulan dan Saran

5.1. Kesimpulan

1. Data yang digunakan pada penelitian ini berupa teks *tweet* dari *twitter* dan didapatkan dengan *crawling* menggunakan pemrograman R. *crawling* dilakukan sebanyak 5 kali pada tanggal 3, 10, 17, 24, 31 Juli 2021 menggunakan kata kunci #ppkmdarurat dan didapatkan sebanyak 5.000 *tweet*. Setelah data melalui tahapan filtering dan pelabelan, data menjadi 770 dengan 50% *tweet* berlabel 'positif' dan 50% *tweet* berlabel 'negatif'. Data akan melalui tahapan *preprocessing* yang terdiri dari 6 tahapan yaitu *cleaning, case folding*, normalisasi bahasa, *stopword removal, stemming* dan tokenisasi. Data akan melalui tahapan pembobotan *term* (TF-IDF) bertujuan untuk memberikan bobot pada setiap kata. Selanjutnya data akan melalui tahapan seleksi fitur *Information Gain* dengan tujuan untuk meningkatkan performa metode *Naïve Bayes*. Setelah melalui klasifikasi *Naïve Bayes* dan dievaluasi menggunakan metode *confusion matrix*.
2. Dari hasil penelitian menunjukkan adanya peningkatan model *Naïve Bayes* apabila menggunakan seleksi fitur *Information Gain* dengan nilai pengambilan *top ranking* '>0.0001' yaitu akurasi 0.81, *recall* 0.82, *precision* 0.84, *f1 score* 0.83 dan *specificity* 0.79 dibandingkan sebelumnya yaitu akurasi 0.79, *recall* 0.81, *precision* 0.81, *f1 score* 0.81 dan *specificity* 0.76.

5.2. Saran

1. Diharapkan untuk penelitian selanjutnya dapat lebih mengoptimalkan proses *preprocessing*, seperti menambah corpus normalisasi bahasa.
2. Diharapkan untuk penelitian selanjutnya dapat menggunakan metode seleksi fitur lainnya untuk meningkatkan performa algoritma *Naïve Bayes*.
3. Diharapkan untuk penelitian selanjutnya dapat mencoba algoritma analisis sentimen yang lainnya seperti *Support Vector Machine* (SVM) atau *K-Nearest Neighbor* (K-NN).

Referensi

[1] D. A. Ramadhanty, "Indonesia Peringkat 6 Negara dengan Pengguna Twitter Terbanyak di Dunia 2021," 2021. https://www.goodnewsfromindonesia.id/2021/11/19/indonesia-peringkat-6-negara-dengan-pengguna-twitter-terbanyak-di-dunia-2021

[2] R. Julianto, E. D. Bintari, and Indrianti, "Analisis Sentimen Layanan Provider Telepon Seluler pada Twitter menggunakan Metode Naïve Bayesian Classification," *J. Big Data Anal. Artif. Intell.*, vol. 3, no. 1, 2017.

[3] M. Syarifuddin, "Analisis Sentimen Opini Publik Mengenai Covid-19 Pada Twitter Menggunakan Metode Naïve Bayes Dan Knn," *Inti Nusa Mandiri*, vol. 15, no. 1, pp. 23–28, 2020.

[4] A. Bijaksana, P. Negara, H. Muhardi, and I. M. Putri, "Analisis Sentimen Maskapai Penerbangan Menggunakan Metode Naive Bayes Dan Seleksi Fitur Information Gain Sentiment Analysis on Airlines Using Naïve Bayes

Method and Feature Selection Information Gain," *Core.Ac.Uk*, vol. 7, no. 3, pp. 599–606, 2020, doi: 10.25126/jtiik.202071947.
[5] J. Ipmawati, Kusrini, and E. Taufiq Luthfi, "Komparasi Teknik Klasifikasi Teks Mining Pada Analisis Sentimen," *Indones. J. Netw. Secur.*, vol. 6, no. 1, pp. 28–36, 2017.
[6] M. S. Hadna, P. I. Santosa, and W. W. Winarno, "Studi Literatur Tentang Perbandingan Metode Untuk Proses Analisis Sentimen Di Twitter," *Semin. Nas. Teknol. Inf. dan Komun.*, vol. 2016, no. Sentika, pp. 57–64, 2016, [Online]. Available: https://fti.uajy.ac.id/sentika/publikasi/makalah/2016/95.pdf
[7] S. Sukendar, A. P. A. Santoso, R. A. Rifai, and R. D. Hermawan, "Kebebasan Berdagang Di Tengah PPKM Darurat Ditinjau Dari Sudut Pandang Sosiological Jurisprudence Dan Konsep Keadilan," *JISIP (Jurnal Ilmu Sos. dan Pendidikan)*, vol. 5, no. 3, pp. 593–602, 2021, doi: 10.36312/jisip.v5i3.2226.
[8] D. Rustiana and N. Rahayu, "Analisis Sentimen Pasar Otomotif Mobil: Tweet Twitter Menggunakan Naïve Bayes," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 8, no. 1, pp. 113–120, 2017, doi: 10.24176/simet.v8i1.841.
[9] M. Yasid, "Analisis Sentimen Maskapai Citilink Pada Twitter Dengan Metode Naãve Bayes," *J. Ilm. Inform.*, vol. 7, no. 02, p. 82, 2019, doi: 10.33884/jif.v7i02.1329.
[10] G. A. Buntoro, "Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter," *INTEGER J. Inf. Technol.*, vol. 1, no. 1, pp. 32–41, 2017, [Online]. Available: https://www.researchgate.net/profile/Ghulam_Buntoro/publication/316617194_Analisis_Sentimen_Calon_Gubernur_DKI_Jakarta_2017_Di_Twitter/links/5907eee44585152d2e9ff992/Analisis-Sentimen-Calon-Gubernur-DKI-Jakarta-2017-Di-Twitter.pdf
[11] N. M. A. J. Astari, Dewa Gede Hendra Divayana, and Gede Indrawan, "Analisis Sentimen Dokumen Twitter Mengenai Dampak Virus Corona Menggunakan Metode Naive Bayes Classifier," *J. Sist. dan Inform.*, vol. 15, no. 1, pp. 27–29, 2020, doi: 10.30864/jsi.v15i1.332.
[12] N. Ruhyana, "Analisis Sentimen Terhadap Penerapan Sistem Plat Nomor Ganjil / Genap Pada Twitter Dengan Metode Klasifikasi Naive Bayes," *J. IKRA-ITH Inform.*, vol. 3, no. 1, pp. 94–99, 2019.
[13] A. Deolika, K. Kusrini, and E. T. Luthfi, "Analisis Pembobotan Kata Pada Klasifikasi Text Mining," *J. Teknol. Inf.*, vol. 3, no. 2, p. 179, 2019, doi: 10.36294/jurti.v3i2.1077.
[14] R. Khalida and S. Setiawati, "Analisis Sentimen Sistem E-Tilang Menggunakan Algoritma Naive Bayes Dengan Optimalisasi Information Gain," *J. Inform. Inf. Secur.*, vol. 1, no. 1, pp. 19–26, 2020, doi: 10.31599/jiforty.v1i1.137.