

Implementasi *n-Gram* dalam Analisis Sentimen Masyarakat DIY Terhadap PSBB Jawa-Bali Jilid II Menggunakan *Naive Bayes Classifier*

SYIFA AYU ANJANI¹, ACHMAD FAUZAN²

^{1,2}Program Studi Statistika Fakultas MIPA Universitas Islam Indonesia, Indonesia
e-mail: ¹17611102@students.uii.ac.id, ²176110102@uui.ac.id

ABSTRAK

Coronavirus Disease (COVID-19) telah menyebar dengan sangat cepat ke seluruh dunia, termasuk Indonesia. Dalam upaya pencegahan penularan COVID-19, pemerintah menerapkan Pembatasan Sosial Berskala Besar (PSBB) di seluruh provinsi di Pulau Jawa dan Bali dari tanggal 26 Januari 2021 s.d. 8 Februari 2021. Pemberlakuan PSBB Jawa-Bali menimbulkan tanggapan positif maupun negatif dari masyarakat dalam media sosial *twitter*, khususnya masyarakat yang berlokasi di Daerah Istimewa Yogyakarta. Sehingga penting untuk dilakukan analisis terhadap tanggapan tersebut agar dapat mengetahui secara umum hal apa yang sering dikeluhkan masyarakat DIY terkait PSBB Jawa-Bali jilid II dan menjadi bahan evaluasi bagi pemerintah DIY dalam menjalankan kebijakan PSBB yang telah dibuat. Penelitian ini menggunakan teknik *crawling* untuk memperoleh data *tweet* (tanggapan) masyarakat dalam *twitter*. Kemudian, dilakukan *preprocessing* pada data agar siap untuk dianalisis. *Tokenization* dilakukan menggunakan metode *n-gram*, yang terdiri dari *unigram*, *bigram*, dan *trigram*. Selanjutnya, data diklasifikasikan berdasarkan kategori sentimen positif dan sentimen negatif menggunakan *Naive Bayes Classifier* (NBC). Setelah itu akan dilakukan evaluasi pada hasil klasifikasi menggunakan *confusion matrix* yang akan menghasilkan nilai akurasi. Berdasarkan hasilnya, klasifikasi dengan NBC memiliki tingkat akurasi sebesar 82,14%. Klasifikasi dengan NBC dan *unigram* menghasilkan akurasi sebesar 83,04%, dengan *bigram* menghasilkan akurasi sebesar 80,36%, dan dengan *trigram* menghasilkan akurasi sebesar 82,14%.

Kata Kunci: Analisis Sentimen, *Naive Bayes Classifier*, PSBB Jawa-Bali, *Twitter Crawling*.

ABSTRACT

Coronavirus Disease (COVID-19) has spread very quickly throughout the world, including Indonesia. In an effort to prevent the spread of COVID-19, the government implemented Large-Scale Social Restrictions (PSBB) in all provinces on the islands of Java and Bali from January 26, 2021 to February 8, 2021. The implementation of the Java-Bali PSBB is a positive or negative response from the community on Twitter social media, especially people located in the Special Region of Yogyakarta. So it is important to analyze these responses in order to find out in general what the DIY community often complains about the Java-Bali PSBB volume II and become an evaluation material for the DIY government in implementing the PSBB policies that have been made. This study uses a crawling technique to obtain data on people's tweets (responses) on Twitter. Then, the data is preprocessed so that it is ready for analysis. Tokenization is carried out using the *n-gram* method, which consists of unigram, bigram, and trigram. Furthermore, the data is classified according to the categories of positive sentiment and negative sentiment using the *Naive Bayes Classifier* (NBC). After that, an evaluation will be carried out on the results of the classification using a confusion matrix which will produce an accuracy value. Based on the results, the classification with NBC has an accuracy rate of 82.14%. Classification with NBC and unigram produces an accuracy of 83.04%, with bigram produces an accuracy of 80.36%, and with trigram produces an accuracy of 82.14%.

Keywords: Sentiment Analysis, *Naive Bayes Classifier*, Java-Bali PSBB, *Twitter Crawling*.

1. PENDAHULUAN

Coronavirus Disease atau biasa disebut dengan COVID-19 mulai menghantui masyarakat dunia pada akhir tahun 2019. Penyebaran virus *corona* terjadi dengan sangat cepat ke seluruh dunia, termasuk Indonesia. Salah satu strategi pemerintah untuk menekan peningkatan jumlah kasus tersebut adalah menerapkan kebijakan Pembatasan Sosial Berskala Besar (PSBB). Kementerian Koordinator Bidang Pembangunan Manusia dan Kebudayaan Republik Indonesia (KEMENKO PMK) menjelaskan bahwa PSBB adalah pembatasan kegiatan tertentu penduduk dalam suatu wilayah yang diduga terinfeksi COVID-19. Pada tanggal 26 Januari 2021 pemerintah melakukan perpanjangan kebijakan PSBB Jawa-Bali atau PPKM Jawa-Bali hingga tanggal 8 Februari 2021, sehingga dinamakan dengan PSBB Jawa-Bali Jilid II.

Daerah Istimewa Yogyakarta (DIY) merupakan salah satu provinsi yang menjalankan kebijakan dari pemerintah pusat terkait PSBB Jawa Bali. Hal tersebut dikarenakan dari empat parameter yang ditetapkan, DIY memenuhi salah satunya yaitu tingkat keterisian rumah sakit. Di provinsi DIY keterisian rumah sakit lebih dari 70%, angka kasus aktifnya di atas rata-rata nasional dengan kesembuhan di bawah angka nasional. Selain itu, Yogyakarta merupakan kota pariwisata yang mengandalkan pendapatannya dari industri wisata. Pergerakan yang telah dibatasi untuk orang dari luar daerah keluar masuk DIY, berdampak kepada pendapatan asli daerahnya yang mengalami penurunan selama masa pandemi ini.

Wisatawan yang berlibur ke kota Yogyakarta dan menginap di hotel biasanya akan dikenakan pajak hotel. Berdasarkan data dalam *open data* Pemerintah Kota Yogyakarta, penurunan pajak hotel menjadi faktor yang cukup penting terhadap penurunan pendapatan daerah, begitu juga dengan pajak restoran. Hal tersebut dikarenakan berkurangnya kuantitas wisatawan yang belibur ke Yogyakarta selama masa COVID-19. Pemberlakuan PSBB Jawa-Bali menimbulkan tanggapan positif maupun negatif dari masyarakat dalam media sosial *twitter*, khususnya masyarakat yang berlokasi di Daerah Istimewa Yogyakarta. Sehingga penting untuk dilakukan analisis terhadap tanggapan tersebut agar dapat mengetahui secara umum hal apa yang sering dikeluhkan masyarakat DIY terkait PSBB Jawa-Bali jilid II.

Dalam penelitian ini data diambil dari media sosial *twitter* menggunakan teknik *crawling*. Selanjutnya dilakukan *preprocessing* pada data, yang merupakan tahap penyiapan data sebelum dilakukan pengolahan atau pembelajaran oleh komputer. Beberapa tahapan yang akan dilakukan dalam *preprocessing* data seperti mengonversi teks menjadi *lower case* (*case folding*), melakukan pembersihan data (*cleaning data*), dan *tokenizing* (Nugraha, Harani, & Habibi, 2020). *Tokenizing* adalah proses pemecahan kalimat menjadi kata. Pada penelitian ini akan dilakukan *tokenizing* menggunakan fitur *n-gram* dengan *unigram*, *bigram*, dan *trigram*. *n-gram* dapat berfungsi dalam pengambilan *n* karakter dalam suatu kalimat tertentu. Pada proses analisisnya, data *tweet* diklasifikasi terlebih dahulu.

Proses klasifikasi dilakukan untuk menentukan mana ulasan yang berbentuk positif, netral, dan negatif. Kemudian data dibagi menjadi data *training* dan data *testing* dengan perbandingan 80:20. *Machine learning* akan menggunakan data *training* sebagai alat untuk mempelajari pola data dalam pembentukan model *classifier*. Sedangkan, data *testing* digunakan untuk mengukur sebaik apa *classifier* mampu mengklasifikasikan data dengan benar. Dalam penelitian ini, *classifier* yang digunakan adalah *Naïve Bayes*. Setelah melakukan klasifikasi, data akan diekstrak dan dieksplorasi untuk memperoleh informasi yang dianggap penting. Proses ekstraksi dan eksplorasi yang digunakan adalah statistika deskriptif untuk memperoleh gambaran umum dari *tweet* seputar PSBB masyarakat DIY, *wordcloud* untuk mengetahui kata yang paling sering digunakan dalam *tweet*, dan asosiasi antarkata untuk menemukan topik yang sering dibahas oleh masyarakat DIY di *twitter*. Analisis tersebut berguna untuk melihat hal apa yang sering dikeluhkan masyarakat terkait PSBB Jawa-Bali jilid II sehingga bisa dijadikan bahan evaluasi bagi pemerintah dalam menjalankan kebijakan PSBB yang telah dibuat.

2. METODE PENELITIAN

Data yang digunakan dalam penelitian ini adalah data *tweet* masyarakat Daerah Istimewa Yogyakarta dengan kata kunci "PSBB" dari tanggal 26 Januari hingga 8 Februari 2021 yang diambil dari media sosial *twitter* menggunakan teknik *crawling*. Adapun variabel yang digunakan dalam penelitian ini terdapat tujuh macam, yaitu:

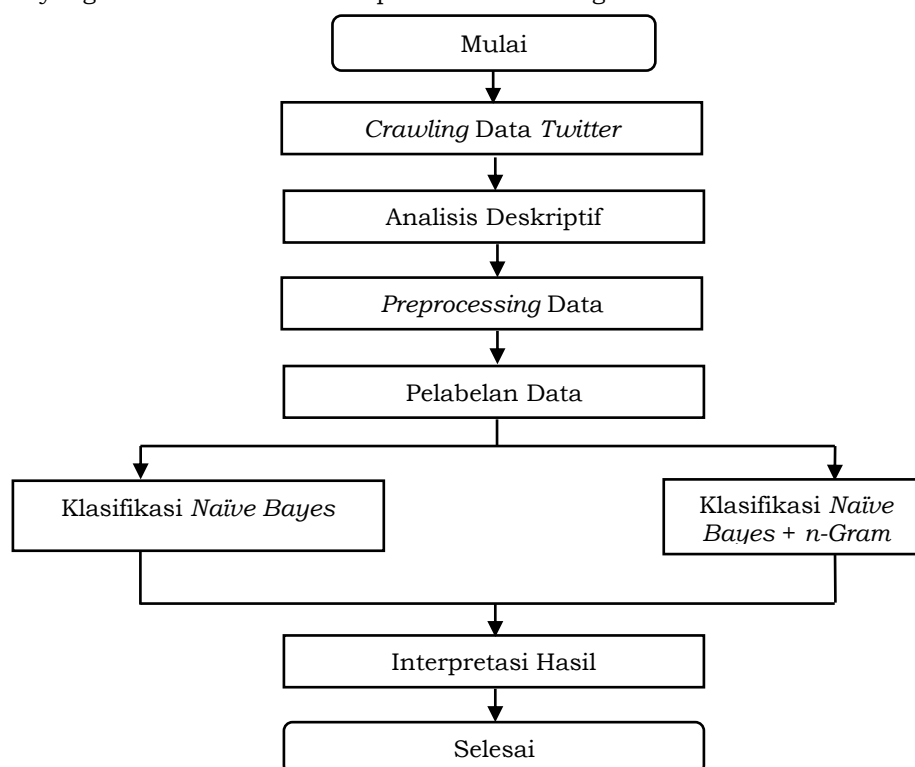
Tabel 1 Definisi Variabel

Variabel	Definisi Variabel	Contoh
<i>User</i>	Nama pengguna <i>twitter</i>	@romaragain
<i>Location</i>	Lokasi pengguna <i>twitter</i> saat mengirim <i>tweet</i>	Yogyakarta
<i>Text</i>	<i>Tweet</i> atau cuitan dari pengguna <i>twitter</i>	Psbbs membuatku bangun lebih pagi dan berangkat lebih awal
<i>Date</i>	Tanggal dibuatnya <i>tweet</i> oleh pengguna <i>twitter</i>	26/01/2021
<i>Time</i>	Waktu dibuatnya <i>tweet</i> oleh pengguna <i>twitter</i>	02.12
<i>Retweet Count</i>	Frekuensi suatu <i>tweet</i> di- <i>retweet</i>	0
<i>Favorite Count</i>	Frekuensi suatu <i>tweet</i> difavoritkan	0

Proses analisis dalam penelitian ini menggunakan beberapa bantuan software, yaitu *Python* dan *RStudio*. Metode yang digunakan dalam penelitian ini, antara lain sebagai berikut.

1. *Web crawling*, digunakan untuk mengumpulkan data *tweet* masyarakat yang dengan kata kunci PSBB selama masa PSBB Jawa-Bali jilid II, yaitu dari tanggal 26 Januari 2021 hingga 8 Februari 2021.
2. Analisis deskriptif, digunakan untuk melihat gambaran umum mengenai sentimen masyarakat Daerah Istimewa Yogyakarta terhadap PSBB Jawa-Bali Jilid II.
3. Analisis sentimen, digunakan untuk proses pengklasifikasian *tweet* pengguna yang bersifat sentimen positif dan negatif. Algoritma klasifikasi yang digunakan dalam penelitian ini adalah *naïve bayes classifier* dan tokenisasi *n-gram*.

Tahapan yang akan dilakukan dalam penelitian ini sebagai berikut.



Gambar 1. Diagram Alir Penelitian

3. HASIL DAN PEMBAHASAN

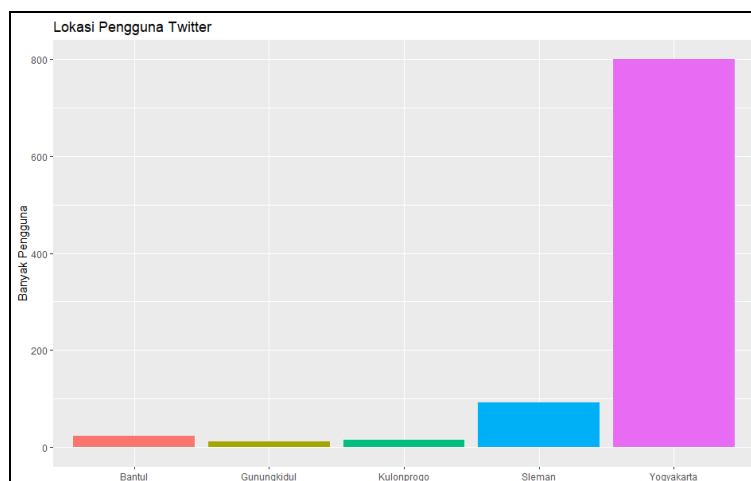
Selanjutnya akan dijelaskan lebih lanjut mengenai penyelesaian rumusan masalah yang telah dijabarkan dalam bab sebelumnya.

Pengumpulan Data

Dalam penelitian ini, data yang digunakan adalah data *tweet* yang diambil dari *twitter* melalui *twitter* API. *Twitter* API adalah sebuah aplikasi yang diciptakan oleh pihak *twitter* agar mempermudah pihak *developer* lain untuk mengakses informasi *web twitter* tersebut. Pengambilan data dari *twitter* biasanya disebut dengan *crawling*. Data dari *twitter* dapat diambil dengan memperoleh kode akses melalui akun *developer twitter* yang dapat diakses di <https://developer.twitter.com>. Data yang diperlukan dalam penelitian ini adalah data *tweet* dari pengguna *twitter* yang melakukan cuitan tentang PSBB dan sedang berada di Daerah Istimewa Yogyakarta. Topik PSBB tersebut difokuskan kepada PSBB Jawa-Bali Jilid II yang dilaksanakan pada tanggal 26 Januari 2021 hingga 8 Februari 2021 sehingga pengambilan data *twitter* dilakukan dengan menggunakan kata kunci “PSBB” dalam periode 26 Januari 2021 hingga 8 Februari 2021.

Analisis Deskriptif

Analisis deskriptif bertujuan untuk melihat gambaran data secara umum. Pada Gambar 1 dapat dilihat bahwa pengguna *twitter* yang paling banyak berkomentar tentang PSBB adalah pengguna yang berlokasi di kota Yogyakarta dengan frekuensi sebanyak 799 pengguna. Setelah kota Yogyakarta, terdapat 93 pengguna yang berlokasi di kabupaten Sleman, 24 pengguna yang berlokasi di kabupaten Bantul, 15 pengguna yang berlokasi di Kulonprogo, dan 12 pengguna yang berlokasi di Gunungkidul.



Gambar 2 Grafik Masyarakat Berdasarkan Lokasi

Text Mining

Preprocessing

Data komentar pengguna *twitter* seputar PSBB adalah data mentah yang belum dapat digunakan. Hal ini dikarenakan data tersebut merupakan data yang tidak terstruktur. Maka dari itu, perlu dilakukan proses pembersihan data agar data layak untuk diolah pada tahap selanjutnya

1. Cleansing

Pengguna *twitter* mengirim *tweet*-nya ada yang menggunakan emoji, simbol, tanda baca, angka, link dari situs web, mention, dan hastag. Dalam tahap ini, karakter-karakter di dalam *tweet* yang bukan merupakan alfabet akan dihapus karena karakter tersebut tidak memiliki arti dalam analisis sentimen.

2. Translate Tweet Bahasa Jawa

Dalam data *tweet* tentang PSBB ini, masih terdapat *tweet* yang menggunakan bahasa Jawa. Hal tersebut dikarenakan data *tweet* yang diambil adalah data *tweet* dengan pengguna yang

berlokasi di Daerah Istimewa Yogyakarta, dimana provinsi tersebut mayoritas penduduknya berbahasa daerah Jawa.

3. *Spelling Normalization*

Tweet yang dikirim oleh para pengguna *twitter*, mayoritas menggunakan kata-kata dalam bahasa informal, disingkat, atau bahkan diperpanjang tiap katanya. Untuk memperbaiki kata-kata tersebut menjadi kata baku dalam bahasa Indonesia, maka perlu dilakukan tahap *spelling normalization* ini.

4. *Case Folding*

Tahap *case folding* adalah tahap untuk menyeragamkan huruf. Huruf kapital akan diseragamkan ke dalam huruf kecil atau *lowercase*.

5. *Tokenizing*

Tahap *case folding* adalah tahap untuk menyeragamkan huruf. Huruf kapital akan diseragamkan ke dalam huruf kecil atau *lowercase*.

6. Ekstraksi *n-Gram*

Selanjutnya akan dilakukan proses tokenisasi menggunakan jenis token *n-gram* pada data *tweet*. Pada penelitian ini akan digunakan tiga jenis pemecahan kata, yaitu unigram, bigram, dan trigram. Unigram adalah pemisahan kata pada data *tweet* dengan $n=1$ atau term tunggal, bigram adalah pemisahan *n-kata* pada kalimat dengan $n=2$, sedangkan trigram adalah pemisahan *n-kata* pada kalimat dengan $n=3$. Berikut adalah ilustrasi penerapan *n-gram* pada salah satu *tweet*.

Contoh kalimat:

"Psbb tanpa solusi adalah omong kosong!"

Tabel 2 Ilustrasi Penerapan *Character n-Gram*

<i>Unigram</i>	'Psbb', 'tanpa', 'solusi', 'adalah', 'omong', 'kosong'
<i>Bigram</i>	'Psbb tanpa', 'tanpa solusi', 'solusi adalah', 'adalah omong', 'omong kosong'
<i>Trigram</i>	'Psbb tanpa solusi', 'tanpa solusi adalah', 'solusi adalah omong', 'adalah omong kosong'

7. *Filtering*

Dalam *tweets* terdapat beberapa kata yang tidak bermakna atau *stopwords*. Tahap *filtering* ini adalah tahap untuk membuang kata-kata yang tidak mempunyai makna, contohnya "yang", "di", "dari", dan masih banyak lagi.

8. *Stemming*

Tahap *stemming* adalah tahap untuk mencari kata dasar dengan menghilangkan kata imbuhan. Fitur yang digunakan dalam proses ini adalah *katadasaR* dalam software R. Fitur tersebut dapat menghapus imbuhan terhadap suatu kata, bukan frase atau kalimat. Selain menggunakan fitur *katadasaR*, proses *stemming* juga dilakukan secara manual. Hal tersebut dikarenakan masih terdapat kata yang belum sempurna dalam penghilangan kata imbuhan.

Sentiment Scoring dan Labelling

Sentiment scoring adalah teknik yang digunakan untuk memberikan label suatu pernyataan yang diklasifikasikan menjadi sentimen positif, netral, dan negatif. *Labelling* ke dalam kategori positif dan negatif ditentukan menggunakan kumpulan kata dengan bahasa Indonesia yang terdiri dari kumpulan kata-kata positif dan negatif. Perhitungan skor dilakukan dengan mengurangi jumlah kata positif dan jumlah kata negatif. Jika perhitungan skor suatu *tweet* menghasilkan skor lebih dari 0, maka akan diklasifikasikan ke dalam kelas positif, jika suatu *tweet* menghasilkan skor sama dengan 0, maka akan diklasifikasikan ke dalam kelas netral, sedangkan jika suatu *tweet* menghasilkan skor kurang dari 0, maka akan diklasifikasikan kelas negatif.

$$\text{Skor} = (\text{jumlah kata positif}) - (\text{jumlah kata negatif}) \quad (1)$$

Simulasi perhitungan skor sentimen disajikan pada Tabel 3.

Tabel 3 Perhitungan Skor Sentimen

<i>Tweet</i>	Kata Positif	Kata Negatif	Kelas
<i>stres</i> melanda karena psbb	- 0	Stres 1	Negatif
psbb rumah saja luar <i>banyak</i> godaan ghibah	Banyak 1	- 0	Positif
psbb cuma malam kok	- 0	- 0	Netral
kamu ini kayak psbb, merusak ke <i>bahagia</i> anku	Rusak 1	Bahagia 1	Netral
psbb saja terus sampai <i>mampus efektif</i> juga <i>tidak</i>	Efektif 1	Mampus Tidak 2	Negatif

Kemudian banyaknya *tweet* berdasarkan kelasnya disajikan pada Tabel 4.

Tabel 4 Perhitungan Skor Sentimen

Kelas	Jumlah <i>Tweet</i>
Positif	111
Netral	399
Negatif	450

Suatu *tweet* dikategorikan sebagai netral ketika *tweet* tersebut tidak mengandung kata positif ataupun kata negatif. Selain itu, *tweet* juga dikategorikan sebagai netral jika memiliki jumlah kata positif dan kata negatif yang seimbang. Misalnya dalam suatu *tweet* terdapat 2 kata positif dan 2 kata negatif. Maka perhitungan skor sentimen untuk *tweet* tersebut sebagai berikut.

$$\text{Skor} = 2 - 2 = 0$$

Berdasarkan hasil perhitungan di atas, diperoleh skor sentimen sebesar 0. Sehingga *tweet* tersebut dikategorikan sebagai *tweet* yang netral.

Pembuatan Data Training dan Data Testing

Dalam algoritma klasifikasi data akan dibagi menjadi data *training* dan data *testing*. Data *training* digunakan untuk membentuk sebuah model *classifier* yang akan menjadi representasi pengetahuan untuk prediksi kelas data baru yang belum pernah ada. Data *testing* digunakan untuk mengukur sebaik apa *classifier* berhasil melakukan klasifikasi dengan benar. Jumlah data *training* dan data *testing* yang akan digunakan memiliki perbandingan 80:20. Pembagian data *training* dan data *testing* disajikan pada Tabel 5.

Tabel 5 Pembagian Data Training dan Data Testing

	Data Training	Data Testing
Rasio	80%	20%
Jumlah	449	112

Klasifikasi Sentimen

Algoritma klasifikasi yang akan digunakan dalam penelitian ini adalah *naïve bayes classifier*. Namun, untuk meningkatkan nilai akurasi, algoritma *naïve bayes* akan dilakukan menggunakan fitur *n-gram* pada tahap *preprocessing*.

Klasifikasi dengan Algoritma Naïve Bayes tanpa Fitur *n-gram*

Algoritma *naïve bayes* akan mempelajari pola data untuk mendeteksi karakteristik data pada masing-masing kelas yang telah dibentuk pada data *training*. Karakteristik dari data *training* akan digunakan dalam memprediksi klasifikasi data baru. Berikut adalah perbandingan data aktual dengan data hasil prediksi.

Tabel 6 Perbandingan Data Aktual dan Data Prediksi

No.	Aktual	Prediksi
1.	Negatif	Negatif
2.	Negatif	Negatif
3.	Negatif	Negatif
4.	Negatif	Negatif
5.	Positif	Negatif
6.	Negatif	Negatif
7.	Negatif	Negatif
8.	Positif	Positif
9.	Positif	Negatif
10.	Negatif	Negatif
⋮	⋮	⋮
112.	Negatif	Negatif

Dalam tabel 6 dapat dilihat terdapat *miss classification* pada baris kelima, dan kesembilan karena sentimen positif terprediksi menjadi sentimen negatif atau disebut juga dengan *false negative* (FN). Pada baris pertama dapat dilihat bahwa sentimen negatif terprediksi menjadi sentimen negatif juga atau bisa juga disebut dengan *true negative* (TN). Evaluasi menggunakan metode *confusion matrix* untuk mengetahui nilai akurasi klasifikasi disajikan pada Tabel 7.

Tabel 7 Confusion Matrix Hasil Prediksi Naïve Bayes Classifier

Prediksi	Aktual	
	Negatif	Positif
Negatif	84	16
Positif	4	8

Tabel 7 merupakan *confusion matrix* hasil prediksi dengan data aktual. Dari tabel tersebut dapat dilihat bahwa terdapat 84 sentimen negatif yang terprediksi negatif yang bisa disebut juga dengan *true negative* (TN). Lalu, terdapat 4 sentimen negatif yang terprediksi positif yang bisa disebut juga dengan *false positive* (FP). Kemudian, terdapat 8 sentimen positif yang terprediksi positif atau bisa juga disebut dengan *true positive* (TP) dan yang terakhir, terdapat 16 sentimen positif yang terprediksi negatif atau bisa juga disebut dengan *false negative* (FN). Berikut adalah perhitungan evaluasi dengan nilai *recall*, *precision*, *accuracy*, dan *specificity*.

1. *Recall* digunakan untuk mengukur proporsi positif yang benar diidentifikasi.

$$Recall = \frac{TP}{TP + FN} \times 100 = \frac{8}{8 + 16} \times 100 = 33,33\%$$

Pada perhitungan di atas, diperoleh nilai *recall* sebesar 33,33%. Artinya identifikasi sentimen positif yang dilakukan kurang baik.

2. *Precision* adalah ukuran ketepatan dalam klasifikasi.

$$Precision = \frac{TP}{TP + FP} \times 100 = \frac{8}{8 + 4} \times 100 = 66,67\%$$

Pada perhitungan di atas, diperoleh nilai *precision* sebesar 66,67%. Artinya tingkat ketepatan dalam klasifikasi cukup baik.

3. *Accuracy* adalah ukuran untuk mengetahui seberapa bagus model bisa mengklasifikasikan data dengan benar.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100 = \frac{8 + 84}{8 + 4 + 84 + 16} \times 100 = 82,14\%$$

Pada perhitungan di atas, diperoleh nilai *accuracy* sebesar 82,14% sehingga dapat dikatakan bahwa model dapat mengklasifikasikan data dengan benar.

4. *Specificity* digunakan untuk mengukur proporsi sentimen negatif yang benar diidentifikasi.

$$Specificity = \frac{TN}{TN + FP} \times 100 = \frac{84}{84 + 4} \times 100 = 95,46\%$$

Pada perhitungan di atas, diperoleh nilai *specificity* sebesar 95,46% sehingga dapat disimpulkan bahwa identifikasi sentimen negatif yang dilakukan sangat baik.

Klasifikasi dengan Algoritma Naïve Bayes Menggunakan Fitur *n-gram*

Selanjutnya, akan dilakukan klasifikasi menggunakan algoritma *naïve bayes* dengan fitur *n-gram*. Dengan penerapan fitur *n-gram* diharapkan dapat meningkatkan nilai akurasi. Fitur *n-gram* yang akan diterapkan dalam penelitian ini adalah *unigram*, *bigram*, dan *trigram*. Berikut adalah perbandingan antara data aktual dan data prediksi dari hasil klasifikasi data baru menggunakan *naïve bayes* dan *unigram*.

Tabel 8 Perbandingan Data Aktual dan Data Prediksi *Naïve Bayes* dan *n-gram*

No.	Aktual	Hasil Prediksi		
		Unigram	Bigram	Trigram
1.	Positif	Positif	Negatif	Negatif
2.	Negatif	Negatif	Negatif	Negatif
3.	Negatif	Negatif	Negatif	Negatif
4.	Positif	Negatif	Positif	Positif
5.	Negatif	Negatif	Negatif	Negatif
6.	Negatif	Negatif	Negatif	Negatif
7.	Negatif	Negatif	Positif	Negatif
8.	Positif	Negatif	Positif	Negatif
9.	Negatif	Negatif	Negatif	Negatif
10.	Negatif	Negatif	Negatif	Negatif
:	:	:	:	:
112.	Negatif	Negatif	Negatif	Negatif

Tabel 8 menunjukkan perbandingan data aktual dengan hasil prediksi dari *naïve bayes classifier* dengan tambahan fitur *unigram*, *bigram*, dan *trigram*. Pada klasifikasi dengan tambahan fitur *unigram* dapat dilihat terdapat *miss classification* pada baris keempat karena sentimen positif terprediksi menjadi sentimen negatif atau dapat disebut dengan *false negative* (FN). Selanjutnya, pada klasifikasi dengan tambahan fitur *bigram* dapat dilihat terdapat *miss classification* pada baris pertama karena sentimen positif terprediksi menjadi sentimen negatif atau dapat disebut dengan *false negative* (FN). Selain itu, juga terdapat *miss classification* pada baris ketujuh karena sentimen negatif terprediksi menjadi sentimen positif atau dapat disebut dengan *false positive* (FP). Kemudian, pada klasifikasi dengan tambahan fitur *trigram* dapat dilihat terdapat *miss classification* pada baris kelima pertama dan kedelapan karena sentimen positif terprediksi menjadi sentimen negatif atau dapat disebut dengan *false negative* (FN). Selanjutnya, akan dilakukan evaluasi untuk hasil prediksi menggunakan *confusion matrix*.

Confusion matrix hasil prediksi dari klasifikasi menggunakan *naïve bayes* dan *unigram* disajikan pada Tabel 9.

Tabel 9 Confusion Matrix Hasil Prediksi Naïve Bayes Classifier & Unigram

Prediksi	Aktual	
	Negatif	Positif
Negatif	87	16
Positif	3	6

Berdasarkan Tabel 9 dapat dilihat bahwa terdapat 87 sentimen negatif yang terprediksi negatif yang bisa disebut juga dengan *true negative* (TN). Lalu, terdapat 3 sentimen negatif yang terprediksi positif yang bisa disebut juga dengan *false positive* (FP). Kemudian, terdapat 6 sentimen positif yang terprediksi positif atau bisa juga disebut dengan *true positive* (TP) dan yang terakhir, terdapat 16 sentimen positif yang terprediksi negatif atau bisa juga disebut dengan *false negative* (FN). Untuk hasil klasifikasi menggunakan *naïve bayes* dan *unigram* dari beberapa *tweet* dapat dilihat pada Tabel 10. Berikut adalah *confusion matrix* hasil prediksi dari klasifikasi menggunakan *naïve bayes* dan *bigram*.

Tabel 10 Confusion Matrix Hasil Prediksi Naïve Bayes Classifier & Bigram

Prediksi	Aktual	
	Negatif	Positif
Negatif	86	18
Positif	4	4

Berdasarkan tabel 10 dapat dilihat bahwa terdapat 86 sentimen negatif yang terprediksi negatif yang bisa disebut juga dengan *true negative* (TN). Lalu, terdapat 4 sentimen negatif yang terprediksi positif yang bisa disebut juga dengan *false positive* (FP). Kemudian, terdapat 4 sentimen positif yang terprediksi positif atau bisa juga disebut dengan *true positive* (TP) dan yang terakhir, terdapat 18 sentimen positif yang terprediksi negatif atau bisa juga disebut dengan *false negative* (FN). *Confusion matrix* hasil prediksi dari klasifikasi menggunakan *naïve bayes* dan *trigram* disajikan pada Tabel 11.

Tabel 11 Confusion Matrix Hasil Prediksi Naïve Bayes Classifier & Trigram

Prediksi	Aktual	
	Negatif	Positif
Negatif	89	19
Positif	1	3

Berdasarkan Tabel 11 dapat dilihat bahwa terdapat 89 sentimen negatif yang terprediksi negatif yang bisa disebut juga dengan *true negative* (TN). Lalu, terdapat 1 sentimen negatif yang terprediksi positif yang bisa disebut juga dengan *false positive* (FP). Kemudian, terdapat 3 sentimen positif yang terprediksi positif atau bisa juga disebut dengan *true positive* (TP) dan yang terakhir, terdapat 19 sentimen positif yang terprediksi negatif atau bisa juga disebut dengan *false negative* (FN). Setelah dilakukan perhitungan, klasifikasi untuk kalimat baru menggunakan algoritma *naïve bayes* ditambah dengan fitur *n-gram* disajikan pada Tabel 12.

Tabel 12 Hasil Klasifikasi pada Kalimat Baru

Kalimat	NB	NB + Unigram	NB + Bigram	NB + Trigram
PSBB tidak jelek di mataku	Negatif	Negatif	Negatif	Positif
Sebagian masyarakat ada yang menganggap PSBB tidak buruk	Negatif	Negatif	Positif	Positif

Evaluasi Model

Setelah melalui tahap klasifikasi, tahap selanjutnya adalah evaluasi. Dalam tahap ini akan dilakukan perbandingan untuk seluruh hasil prediksi yang telah dilakukan sebelumnya. Hasil evaluasi dari masing-masing model disajikan pada Tabel 13.

Tabel 13 Hasil Akurasi Klasifikasi dengan *Naïve Bayes* dan *n-gram*

	NB	NB + Unigram	NB + Bigram	NB + Trigram
Accuracy	82,14%	83,04%	80,36%	82,14%
Recall	33,33%	27,27%	18,18%	13,64%
Specificity	95,46%	96,67%	95,56%	98,89%
Precision	66,67%	84,47%	82,68%	82,41%

Berdasarkan evaluasi menggunakan *confusion matrix* dapat dibuktikan bahwa penerapan *character n-gram* pada proses *preprocessing* dapat meningkatkan nilai akurasi klasifikasi algoritma *naïve bayes*. Hasil klasifikasi menggunakan *naïve bayes* dengan tambahan *unigram* menghasilkan nilai akurasi yang lebih tinggi dibandingkan dengan klasifikasi tanpa tambahan *n-gram*. Nilai akurasi untuk model algoritma *naïve bayes* sebesar 82,14%, sedangkan nilai akurasi untuk model algoritma *naïve bayes* dengan penerapan *n-gram* pada tahap *preprocessing* memiliki akurasi sebesar 83,04% dengan selisih 0,9%.

4. SIMPULAN DAN SARAN**Simpulan**

Berdasarkan hasil analisis terhadap data tweet masyarakat Daerah Istimewa Yogyakarta seputar PSBB pada masa PSBB Jawa-Bali Jilid II, kesimpulan yang dapat diambil adalah sebagai berikut.

1. Gambaran umum mengenai tweet masyarakat Daerah Istimewa Yogyakarta berdasarkan kabupaten atau kota adalah pengguna yang paling banyak berkomentar tentang PSBB adalah pengguna yang berlokasi di kota Yogyakarta dengan frekuensi sebanyak 799 pengguna.
2. Penerapan *character n-gram* pada tahap *preprocessing* algoritma *naïve bayes* terbukti dapat meningkatkan nilai akurasi hingga 0,9%. Sebelum menggunakan *n-gram*, nilai akurasi dari algoritma *naïve bayes* sebesar 82,14%. Namun, setelah menggunakan *n-gram*, nilai akurasi menjadi 83,04%.

Saran

Dalam penelitian ini terdapat data *tweet* masyarakat yang menggunakan bahasa Jawa. Untuk penelitian selanjutnya dapat melakukan penanganan yang lebih efisien terhadap hal tersebut sehingga tidak akan memakan waktu yang lama pada tahap *preprocessing*. Selain itu, untuk penelitian selanjutnya dapat melakukan pengambilan data dari media sosial lain ataupun pengambilan data secara langsung agar diperoleh data yang lebih banyak dan dengan keberagaman yang lebih tinggi sehingga bisa lebih representatif untuk masyarakat DIY secara keseluruhan. Bagi peneliti selanjutnya, dapat menggunakan *character n-gram* dengan $n=4$ atau *quadgram* sebagai pembanding untuk penerapan *unigram*, *bigram*, dan *trigram*. Selain itu, peneliti selanjutnya juga dapat menggunakan *machine learning* lainnya, seperti *Support Vector Machine* dan *Decision Tree*.

DAFTAR PUSTAKA

- Darmawidjadja, M. I., Alamsah, D., & Rahman, M. F. (2017). KLASIFIKASI UNTUK DIAGNOSA DIABETES MENGGUNAKAN METODE BAYESIAN REGULARIZATION NEURAL NETWORK (RBNN). *Jurnal Informatika*, 41.
- Fitriyani, & Arifin, T. (2020). PENERAPAN WORD N-GRAM UNTUK SENTIMENT ANALYSIS REVIEW MENGGUNAKAN METODE SUPPORT VECTOR MACHINE. *Jurnal Sistem Informasi*, 610-621.
- Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques*. USA: Elsevier Inc.
- Larose, D. T. (2005). *Discovering Knowledge in Data*. USA: John Wiley & Sons, Inc.
- Larose, D. T. (2006). *Data Mining Methods and Models*. USA: John Wiley & Sons Inc.
- Liu, B. (2015). *Sentiment Analysis*. USA: Cambridge University.
- Luthfi, E. T., & Kusriani. (2009). *Algoritma Data Mining*. Yogyakarta: Andi offset.
- Melita, R., Amrizal, V., Suseno, H. B., & Dirjam, T. (2018). PENERAPAN METODE TERM FREQUENCY INVERSE DOCUMENT FREQUENCY (TF-IDF) DAN COSINE SIMILARITY PADA SISTEM TEMU KEMBALI INFORMASI UNTUK MENGETAHUI SYARAH HADITS BERBASIS WEB. *Jurnal Teknik Informatika*, 149.
- Nugraha, F. A., Harani, N. H., & Habibi, R. (2020). *Analisis Sentimen Terhadap Pembatasan Sosial Menggunakan Deep Learning*. Bandung: Kreatif Industri Nusantara.
- Nugroho, A. (2018). Analisis Sentimen Pada Media Sosial Twitter Menggunakan Naive Bayes Classifier Dengan Ekstraksi Fitur N-Gram. *Jurnal Sains Komputer & Informatika*, 200-209.
- Nugroho, A., & Santoso, E. B. (2019). Analisis Sentimen Calon Presiden Indonesia 2019 Berdasarkan Komentar Publik di Facebook. *Jurnal Eksplorasi Informatika*, 63.
- Pintoko, B. M., & L., K. M. (2018). Analisis Sentimen Jasa Transportasi Online pada Twitter Menggunakan Metode Naive Bayes Classifier. *e-Proceeding of Engineering*, 8122
- Prasetyo, E. (2012). *Data Mining Konsep dan Aplikasi Menggunakan Matlab*. Yogyakarta: Andi.
- Saleh, A. (2015). Implementasi Metode Klasifikasi Naive Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga. *Citec Journal*, 209-210.
- Septiani, L., & Sibaroni, Y. (2019). Sentiment Analysis Terhadap Tweet Bernada Sarkasme Berbahasa Indonesia. *Jurnal Linguistik Komputasional*, 62.
- Setiawan, D. (2009). *Panduan Praktis Mengoptimalkan Twitter*. Jakarta: PT TransMedia.
- Tan, P. N., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining*. USA: Addison-Wesley.
- Tuhuteru, H. (2020). Analisis Sentimen Masyarakat Terhadap Pembatasan Sosial Berskala Besar Menggunakan Algoritma Support Vector Machine. *Information System Development*, 3.
- Ulwan, M. N. (2016). *Pattern Recognition Pada Unstructured Data Teks Menggunakan Support Vector Machine Dan Association*. Yogyakarta: Program Studi Statistika FMIPA UII.
- Zaman, B., Hariyanti, E., & Purwanti, E. (2015). Sistem Deteksi Bahasa pada Dokumen Menggunakan N-Gram. *Jurnal Multinetics*, 22.