Analisis Sentimen Masyarakat Terhadap Pandemi Covid-19 Pada Sosial Media Menggunakan Naïve Bayes Clasifier

Ade Fitriadin*, Agus Sidiq Purnomo**
*,**Program Studi Informatika, Fakultas Teknologi Informasi, Universitas Mercu Buana Yogyakarta, Jl. Wates Km. 10 Yogyakarta 55753, Indonesia
Email: *fitriadin98@gmail.com, **sidiq@mercubuana-yogya.ac.id

ABSTRACT

Social Media is an application internet based of communication. Twitter is one tool that Internet users frequently utilize. Twitter is one tool that Internet users frequently utilize., Twitter is a website that offers microblogging services so that users can communicate ideas, opinions, or just their daily lives through brief writings called Tweets. Social media is a communication tool that Internet users are currently using in large numbers. Tweets from Twitter users cover a wide range of topics, and from these tweets, data can be extracted for sentiment analysis, which can provide information to a variety of parties. The purpose of this project is to develop a system for sentiment analysis that can produce data and information in the form of positive sentiment, negative sentiment, and neutral sentiment. The Nave Bayes Classifier technique is used to categorize feelings. This system receives input in the form of public tweets on the Covid-19 Pandemic. Data visualizations of positive, negative, and neutral sentiment are produced by this system as its output. The Public Sentiment Analysis System with the Nave Bayes Classifier Algorithm can automatically do sentiment analysis with an accuracy of 70% when classifying a tweet using the Nave Bayes Classifier approach. The analysis's findings are presented in tabular form and visually using word clouds and diagrams.

Keyword: Sentiment Analysis; Twitter; Pandemic Covid-19; Naïve Bayes Classifier

1. Pendahuluan

Perkembangan teknologi di Indonesia telah menghadirkan berbagai layanan yang memudahkan manusia, salah satunya media sosial. Media sosial merupakan sebuah alat komunikasi yang menggunakan internet yang dapat digunakan untuk mengungkapkan, dan menyampaikan argumen tentang berbagai isu dan masalah. Pengguna layanan ini dapat menulis tentang apa saja, termasuk kehidupan mereka sendiri, berbagi pendapat tentang berbagai topik, dan mendiskusikan masalah terkini.. Salah satu bahasan yang kerap muncul di twitter yaitu *Pandemic Covid-19*. Berbagai macam komentar dan tanggapan terkait *pandemic covid-19* muncul semenjak diumumkannya kasus pertama *covid-19* pada awal bulan maret 2019, komentar yang muncul antara lain sanjungan dan pujian pada penanganan masalah, kritik dan masukan kepada pemerintah dalam upaya menangani sebuah masalah, bahkan sampai kalimat bernada ujaran kebencian. Besarnya jumlah pengguna Twitter di Indonesia yang menyampaikan komentar dapat memberi peluang untuk memanfaatkan data banyaknya komentar dan mengumpulkan informasi yang dapat membantu berbagai pihak dalam mendukung sebuah keputusan. Tetapi untuk melakukan hal tersebut diperlukan sebuah Analisis yang tepat, salah satu metode yang bisa dimanfaatkan untuk menganalisa berbagai tanggapan di Twitter yaitu Analisis Sentimen. Untuk melakukan analisis sentimen pada Twitter yang umumnya berbentuk teks dapat dilakukan dengan *text mining*. *Text mining* dapat didefinisikan sebagai suatu proses untuk mengekstrak informasi yang berguna dari suatu sumber data melalui identifikasi dan eksplorasi pola tertentu [1]. Analisis sentimen atau *opinion mining* adalah studi komputational dari opini-opini orang, sentimen dan emosi melalui entitas atau atribut yang dimiliki yang diekspresikan dalam bentuk teks [2]. Analisis sentimen bisa digunakan untuk mencari nilai persentase sentimen positif, nilai presentasi sentimen negatif dan nilai presantase sentiment netral pada suatu produk, layanan, tokoh, lembaga, intitusi atau sebuah kondisi tertentu [3]

Peneltian mengenai Analisis sentiment sebelumnya pernah dilakukan yang menganalisis Sentimen Twitter terhadap Bom Bunuh Diri di Surabaya 13 Mei 2018 menggunakan Pendekatan *Support Vector Machine*. Hasil dari analisis sentimen yang dilakukan adalah sentimen negatif lebih mendominasi dengan jumlah 1921 Tweet dan sentimen positif sebanyak 121 Tweet [4].

Selanjutnya penelitian Analisis Sentimen Terhadap Tayangan Televisi Berdasarkan Opini Masyarakat pada Media Sosial Twitter menggunakan Metode K-Nearest Neighbor dan Pembobotan Jumlah Retweet. Dari hasil pengujian akurasi menggunakan pembobotan tekstual diperoleh 82,50%, menggunakan pembobotan non-tekstual 60%, dan menggunakan penggabungan keduanya 83,33% [5].

Selanjutnya penelitian Analisis Sentimen Twitter Debat Calon Presiden Indonesia Menggunakan Metode *Fined-Grained Sentiment Analysis*. Hasil yang didapat dari penelitian ini adalah sebuah kecendurangan bernilai

positif yang lebih mendominasi dari pada kecendurangan negatif maupun kecendurangan netral diantara kedua calon pasangan presiden dan wakil presiden Republik Indonesia [6].

Kemudian penelitian analisis sentimen *hatexpeech* pada twitter dengan Metode *Naïve Bayes Classifier* dan *Support Vector Machine*. Hasil tertinggi didapat dengan menggunakan metode klasifikasi *Support Vector Machine* dengan nilai rata-rata akurasi mencapai 66.6% nilai presisi 67.1%, nilai *recall* 66.7% nilai TP *rate* *66.7%* dan nilai TN *rate* 75.8% [7].

Kemudian penelitian analisis sentimen mengenai komentar media sosial terhadap UMBY. Hasil sentimen analisis dengan menggunakan tagar #umby diperoleh data tweet sebanyak 179 data. Dengan hasil klasifikasi positif sebanyak 57 data (32%), negatif sebanyak 30 data (17%), dan netral sebanyak 92 data (51%) [8].

2. Metode Penelitian

Data dalam penelitian ini diperoleh dengan memanfaatkan API Twitter yang diambil dari awal bulan Desember 2020 menggunakan kata kunci yang berkaitan dengan *covid-19* seperti *pandemi covid-19, covid-19, corona, virus corona*. Data-data ini digunakan sebagai basisdata data latih pada proses klasifikasi. Sementara untuk data yang akan dianalisis oleh sistem bisa diambil kapan saja secara *real time* melalui sistem.

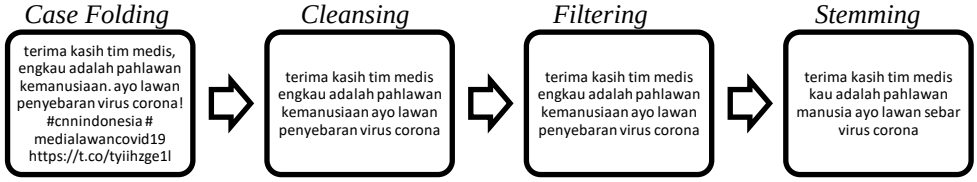
Jumlah data yang digunakan untuk data latih sebanyak 1525 data dan telah dilakukan proses *tagging* secara manual berupa jenis sentiment positif untuk opini berupa dukungan, netral untuk opini berupa saran dan negatif berupa kritik dan ujaran kebencian. Rincian data latih dan data uji dapat dilihat pada Tabel 1.

Tabel 1. Rincian jumlah data latih dan data uji

Sentiment	Jumlah Data Latih	Jumlah Data Uji
Positif	515	103
Netral	500	100
Negatif	510	102

Data sampel didapatkan melalui proses *crawling* yang dilakukan satu kali untuk setiap kata kunci dengan batasan 500 *tweet* setiap kata kunci. Data yang telah dikumpulkan selanjutnya harus melewati proses *text preprocessing* yang memliki beberapa tahap. Tahapan pertama dalam proses *text preprocessing* adalah *case folding* yang bertujuan untuk mengubah semua huruf menjadi huruf kecil (*lowercase*). Selanjutnya adalah tahapan *cleansing* yaitu menghilangkan semua atribut yang tidak mempengaruhi proses klasifikasi seperti angka, tanda baca, url. Tahapan berikutnya adalah proses *filtering* atau penyaringan yang memiliki tujuan untuk membuang *stopwords* atau kata yang diabaikan dalam pemrosesan misalnya kata penghubung seperti "dan","atau","tapi". Dan proses terakhir adalah *stemming* yaitu proses mengembalikan sebuah kata ke kata dasar atau *root word.*

Contoh tahapan *text preprocessing* bisa dilihat dari Gambar 1.



Gambar 1. Proses Text Preprocessing

Pada tahap klasifikasi dibagi menjadi dua proses yaitu proses *training* dan *testing.* Proses *training* menggunakan Data latih yang sudah diproses pada *text preprocessing* kemudian digunakan untuk bahan *learning* atau pembelajaran dalam proses *testing* yang digunakan dalam menentukan jenis sentiment pada tanggapan yang berbentuk sebuah *tweet.*

Berikut adalah contoh perhitungan manual *Naïve Bayes* yang coba penulis buat untuk menghitung *probabilitas* sebuah *tweet* dengan menggunakan contoh data latih seperti pada Tabel 2.

Tabel 2. Contoh Data Uji

ID	Text	Sentiment
*Tweet*1	alhamdulillah dapat bantuan sosial masa pandemi covid	Positif
Tweet2	Terimakasih tim medis pahlawan manusia cegah sebar covid	Positif
Tweet3	Uji coba vaksin covid	Netral
Tweet4	Update kembang covid	Netral
Tweet5	Pandemi covid lumpuh ekonomi	Negatif
Tweet6	corona hanya sebuah bisnis vaksin	Negatif

Data dari Tabel 1 selanjutnya dibuat sebuah model *probabilitas* dengan menggunakan Persamaan 1.

$$P(a_i|v_j) = \frac{n+1}{n+kosakata} \quad \text{.....(1)}$$

Dimana $P(a_i|v_j)$ adalah *probabilitas* kata ai untuk setiap kategori, n merupakan nilai jumlah kosakata yang muncul pada setiap kategori vj dan kosakata adalah nilai jumlah kata unik disemua data *training*

$P(a_{terimaksih} \mid V_{positif}) = \frac{1+1}{17+32} = \frac{2}{49}$
$P(a_{terimakasih} \mid V_{netral}) = \frac{0+1}{7+32} = \frac{1}{39}$
$P(a_{terimakasih} \mid V_{negatif}) = \frac{0+1}{9+32} = \frac{1}{41}$

Setelah mendapat nilai *Probabilitas* setiap kata, maka kita bisa menghitung data uji. Berikut penulis berikan contoh perhitungan *probabilitas* pada proses *testing*, menggunakan data pada Tabel 3.

Tabel 3. Contoh data uji

Id	Text
Tweet7	Pemerintah keluarkan bantuan dampak covid

Untuk menghitung *probabilitas* pada proses *testing* dan mencari *probabilitas* tertinggi digunakan Persamaan 2.

$$V_{MAP} = \underset{vj \in v}{argmax} P(V_j)\ x\ \prod_i P(\ a_i|v_j) \quad \text{.....(2)}$$

$P(Tweet7|VPositif) = P(a_{pemerintah}|V_{Positif})\ x\ P(a_{keluarkan}|V_{Positif})\ x\ P(a_{bantuan}|V_{Positif})\ x\ P(a_{dampak}|V_{Positif})\ x\ P(a_{covid}|V_{Positif})\ x\ P(V_{Positif})$
$= 1\ x\ 1\ x\ \frac{2}{49}\ x\ 1\ x\ 1\ x\ \frac{3}{49}\ x \frac{1}{3}$
= 0.0008329863
$P(Tweet7|VNetral) = P(a_{pemerintah}|V_{Netral})\ x\ P(a_{keluarkan}|V_{Netral})\ x\ P(a_{bantuan}|V_{Netral})\ x\ P(a_{dampak}|V_{Netral})\ x\ P(a_{covid}|V_{Netral})\ x\ P(V_{Netral})$
$= 1\ x\ 1\ x\ \frac{1}{39}\ x\ 1\ x\ 1\ x\ \frac{3}{39}\ x \frac{1}{3}$
= 0.0006574622
$P(Tweet7|VNegatif) = P(a_{pemerintah}|V_{Negatif})\ x\ P(a_{keluarkan}|V_{Negatif})x\ P(a_{bantuan}|V_{Negatif})\ xP(a_{dampak}|V_{Negatif})\ x\ P(a_{covid}|V_{Negatif})\ x\ P(V_{Negatif})$
$= 1\ x\ 1\ x\ \frac{1}{41}\ x\ 1\ x\ 1\ x\ \frac{3}{41}\ x \frac{1}{3}$
= 0.000594884

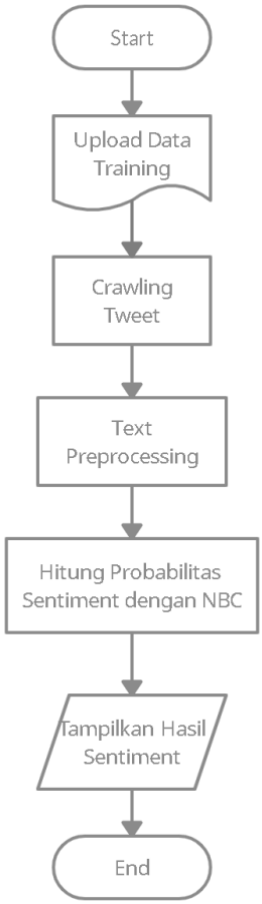
Setelah menghitung *probabilitas* data uji, diperoleh hasil seperti terlihat pada Tabel 4.

Tabel 4. Nilai probabilitas data uji

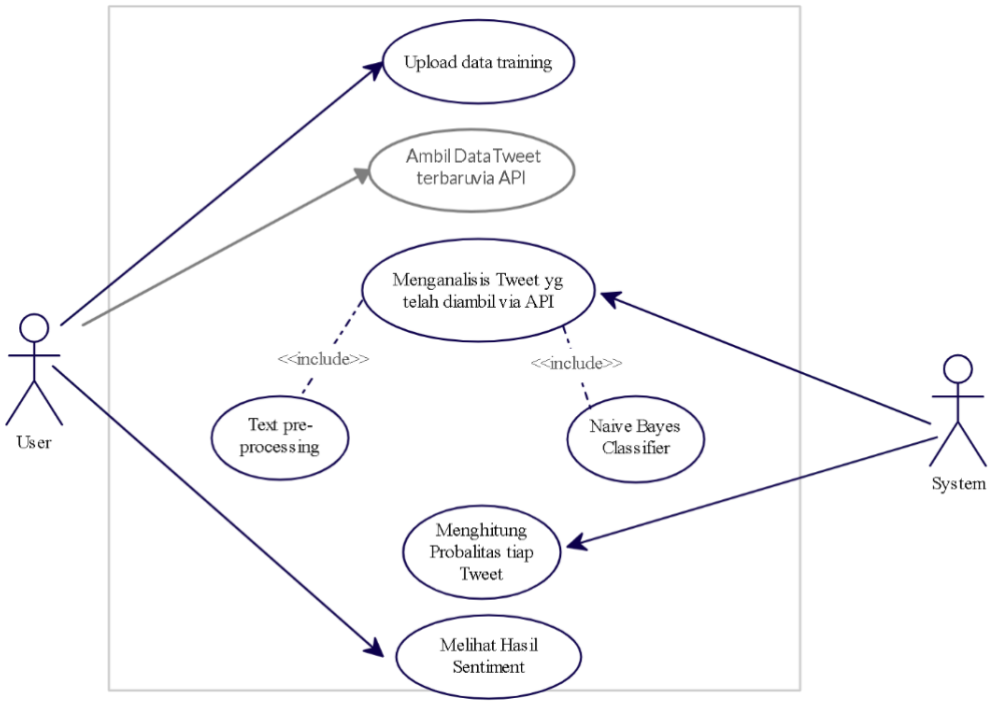
Tweet	Probabilitas		
	Positif	Netral	Negatif
Pemerintah keluarkan bantuan dampak covid	0.0008329863	0.0006574622	0.000594884

Pada Tabel 4 terlihat nilai *probabilitas* Positif mempunyai nilai tertinggi, sehingga *Tweet* bisa disimpulkan termasuk sentiment positif.

Flowchart system jalannya Analisis Sentiment Masyarakat terhadap *pandemic covid-19* menggunakan metode *Naïve Bayes Classifier* dapat dilihat pada Gambar 2. Sedangkan *usecase diagram* jalannya Analisis Sentiment Masyarakat terhadap *pandemic covid-19* menggunakan metode *Naïve Bayes Classifier* dapat dilihat pada Gambar 3.



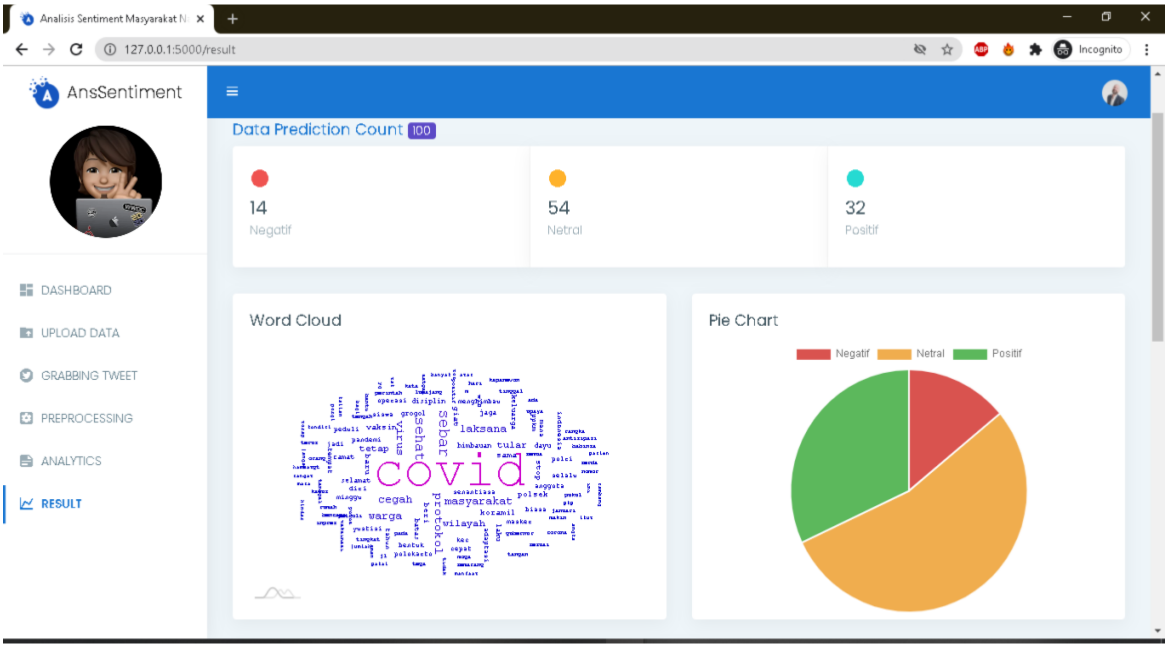
Gambar 2. Flowchart System



Gambar 3. usecase diagram

3. Hasil Dan Analisis

Hasil dari penelitian ini adalah sistem dapat memprediksi *tweet* yang diambil secara *realtime* yang memanfaatkan API Twitter seperti pada Gambar 4.



Gambar 4. Hasil Analisis Tweet secara Realtime

Analisis dilakukan deng an melakukan uji performa dari model pembelajaran dengan memanfaatkan *Confusion Matrix*. *Confussion Matrix* merupakan tabel pencatat hasil kerja klasifikasi [9]. Confusion matrix merupakan alat pengukuran yang dapat digunakan untuk menghitung kinerja atau tingkat kebenaran proses klasifikasi. Dengan confusion matrix dapat dianalisa seberapa baik *classifier* dapat mengenali *record* dari kelas-kelas yang berbeda. *Confusion Matrix* merepresentasikan prediksi dan kondisi sebenarnya dari data yang dihasilkan oleh algoritma *Machine Learning* [10]. Berdasarkan *Confusion Matrix*, maka bisa menentukan *Accuracy*, *Precission*, *Recall*. Dalam peneltian ini untuk melakukan pengujian dilakukan proses klasifikasi menggunakan sebaran data uji acak yang bervariasi. Data dari pengujian klasifikasi yang telah dilakukan dapat dilihat pada Tabel 5.

Pengujian pada tahap klasifikasi ini menggunakan data uji sebanyak 20% dari jumlah data latih yang telah dimasukan ke sistem seperti yang sudah dijelaskan pada Tabel 1 dan terlihat pada nilai *Actual* di Tabel 5.

Tabel 5. Confussion Matrix

		Prediksi		
		Negatif	Netral	Positif
Actual	Negatif	80	6	16
	Netral	21	52	27
	Positif	13	10	80

Berdasarkan *Confusion Matrix*, maka dapat menentukan nilai *Accuracy*, *Precission*, *Recall* dan *F1-Score* yang digunakan sebagai *Classification Report*.

Accuracy merupakan rasio prediksi benar dari keselurahan data. Untuk menghitung *accuracy* digunakan Persamaan 4.

$$\text{accuracy}=\frac{\text{TP(Kelas-1)+TP(kelas-2)+...+TP(kelas-n)}}{\text{Total(kelas-1)+Total(kelas-2)+...+Total(kelas-n)}} \quad \text{.......} \quad (3)$$

Berdasarkan Persamaan 4 maka nilai *accuracy* dari proses klasifikasi menggunakan *Naïve Bayes* sebagai berikut:

$$\text{accuracy}=\frac{80+52+80}{80+6+16+21+52+27+13+10+80}$$

$$accuracy = 0.7$$

 ISSN : 2503 – 250X

Precission digunakan untuk mencari ketepatan nilai yang diminta oleh pengguna dengan respon system. Untuk menghitung *precission* digunakan Persamaan

$$Precission = \frac{TP(Kelas-i)}{Prediksi(kelas-i)} \quad \text{..... (4)}$$

Berdasarkan Persamaan 5 maka nilai *Precission* untuk sentimen positif, netral dan negatif adalah sebagai berikut:

$$Precission\ Positif = \frac{80}{16+27+80} = 0{,}65$$

$$PrecissionNetral = \frac{52}{6+52+10} = 0{,}76$$

$$PrecissionNegatif = \frac{80}{80+21+13} = 0{,}7$$

Recall digunakan untuk mencari nilai ketepatan antara data yang sama dengan data sebelumnya yang dipanggil. Untuk menghitung *recall* digunakan Persamaan sebgai berikut

$$Recall = \frac{TP(Kelas-i)}{Total(kelas-i)} \quad \text{.....(5)}$$

Berdasarkan Persamaan 6 maka nilai *recall* untuk sentimen positif, netral dan negatif adalah sebagai berikut

$$Recall\ Positif = \frac{80}{13+10+80} = 0{,}78$$

$$Recall\ Netral = \frac{52}{21+52+27} = 0{,}52$$

$$Recall\ Negatif = \frac{80}{80+6+16} = 0{,}78$$

*F1-Scor*e adalah perbandingan rata-rata presisi dan recall yang dibobotkan. *F1-score* dirumuskan dengan Persamaan sebagai berikut.

$$F1-Score = 2\left(\frac{Recall\ x\ Precission}{Recall+Precission}\right) \quad \text{.....(6)}$$

Berdasarkan Persamaan 7 maka nilai *F1-score* untuk sentimen positif, netral dan negatif adalah sebagai berikut:

$$F1-Score\ positif = 2\left(\frac{0.78\ x\ 0.65}{0.78+0.65}\right) = 0.74$$

$$F1-Score\ netral = 2\left(\frac{0.76\ x\ 0.52}{0.76+0.52}\right) = 0.62$$

$$F1-Score\ negatif = 2\left(\frac{0.78\ x\ 0.7}{0.78+0.7}\right) = 0.71$$

Jika dibuat sebuah tabel maka hasil dari *Classifiacation Report* disajikan pada Tabel 6.

Tabel 6 Tabel Classification Report

Class	Precision	F1-Score	Recall
Negatif	70%	74%	78%
Netral	76%	62%	52%
Positif	65%	71%	78%
Accuracy		70%	

Berdasarkan *Classification Report* menunjukan bahwa model yang dibangun memiliki nilai akurasi prediksi sebesar 70% sehingga bisa dikatakan pemodelan cukup berhasil. Dengan nilai *precision* 70% pada *tweet* berlabel Negatif, 76% pada *tweet* berlabel Netral dan 65% pada *tweet* berlabel positif.

4. Kesimpulan

Dari hasil perancangan, implementasi serta pengujian analisis sentiment masyarakat pada twitter dapat diambil kesimpulan bahwa kicauan pengguna twitter yang berupa data teks berhasil didapatkan dan dikumpulkan dengan memanfaatkan API Twitter yang tersedia, kemudian dengan memanfaatkan metode *Naive Bayes Classifier*, tanggapan atau opini pengguna twitter berkaitan dengan *Pandemic Covid-19* dapat dianalisis sentiment secara otomatis, dengan nilai akurasi sebesar 70%.

Daftar Pustaka

[1] R. Feldman e J. Sanger, The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, Cambridge, England: Cambridge University Press, 2006.

[2] C. S. Aditya, M. Hani'ah, A. A. Fitrawan, A. Z. Arifin e D. Purwitasari, "Deteksi Bot Spammer pad twitter berbasis Sentiment Analysis," 2016.

[3] B. Pang e L. Lee, Opinion Mining and Sentiment Analysis, Boston: now Publisher, 2008.

[4] Listari, M. Ihsan, E. R. Paradistia e E. Widodo, "Analisis Sentimen Twitter terhadap Bom Bunuh Diri di Surabaya 13 Mei 2018 menggunakan Pendekatan Support Machine," *PRISMA, Prosiding Seminar Nasional Matematika,* pp. 416-426, 2019.

[5] W. E. Nurjanah, R. S. Perdana e . M. A. Fauzi, "Analisis Sentimen Terhadap Tayangan Televisi Berdasarkan Opini Masyarakat pada Media Sosial Twitter menggunakan Metode K-Nearest Neighbor dan Pembobotan Jumlah Retweet," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer,* pp. 1750-1757, 2017.

[6] S. F. Pratama, R. Andrean e A. Nugroho, "Analisis Sentimen Twitter Debat Calon Presiden Indonesia Menggunakan Metode Fined-Grained Sentiment Analysis," *JOINTECS (Journal of Information Technology and Computer Science) Vol. 4, No. 2,* pp. 39-44, 2019.

[7] G. A. Buntoro, "Analisis Sentimen Hatespeech pada twitter dengan Naive Bayes Classifier Dan Support vector Machine," *Jurnal Dinamika Informatika,* 2016.

[8] A. S. Purnomo e A. F. Rozi, "Sentiment Analysis Komentar Media Sosial terhadap UMBY," Universitas Mercu Buana Yogyakarta, Yogyakarta, 2020.

[9] E. Prasetyo, Data Mining Konsep dan Aplikasi Menggunakan MATLAB., Yogyakarta: Penerbit ANDI, 2012.

[10] R. Arthana, "Mengenal Accuracy, Precision, Recall dan Specificity serta yang diprioritaskan dalam Machine Learning," 5 April 2019. [Online]. Available: https://rey1024.medium.com/mengenal-accuracy-precission-recall-dan-specificity-serta-yang-diprioritaskan-b79ff4d77de8. [Acesso em 28 Desember 2020].