Vol.8 No.1 Tahun 2022, pp.32-38
Jurnal Teknologi dan Manajemen Informatika
http:// http://jurnal.unmer.ac.id/index.php/jtmi
P-ISSN: 1693-6604 E-ISSN: 2580-8044

Analisis Sentimen dan Klasifikasi *Tweet* Terkait Mutasi COVID-19 Menggunakan Metode *Naïve Bayes Classifier*

Aryo Dewandaru [1], Jati Sasongko Wibowo [2]

Universitas Stikubank Semarang, Indonesia

Article Info

Article History
Received: 24-11-2021
Revised : 13-12-2021
Accepted: 16-12-2021

Keywords
COVID-19
Sentiment Analysis
Nave Bayes Classifier
Text Mining.

✉ Corresponding Author
Aryo Dewandaru,
Universitas Stikubank Semarang,
Tel. +62 82327704799
aryodewan162@gmail.com

ABSTRACT

Towards the end of 2019 in Wuhan City, China, a new type of Corona Virus was discovered which has the scientific name COVID-19 and is a type of virus that causes acute disorders in the human respiratory system. The spread of this virus is very fast and causes mutations of this virus to a more lethal stage than before. Thus, sentiment analysis is expected to be able to determine the trend of public assessment of the COVID-19 mutation. Naïve Bayes Classifier is a method used in research. This method can classify data or opinions into two sentiments, namely positive and negative. The research data comes from Twitter which is taken using the Twitter API with the keyword "covid mutation", for data processing several processes are carried out, namely sentiment classification, data cleaning, and preprocessing so that the final result is obtained. The test results from this study show that the Naïve Bayes Classifier method has an accuracy of 86.67% with an f1-score of 82.00% on positive sentiment and 89.00% on negative sentiment. Based on the results of the study, it can be concluded that the Naïve Bayes Classifier method can be used to analyze sentiment data from tweets about the COVID-19 mutation with an accuracy of 86.67%.

PENDAHULUAN

Virus Corona atau SARS-CoV-2 merupakan virus yang menjadi penyebab dari penyakit COVID-19. COVID-19 sendiri ditemukan pada bulan Desember tahun 2019 di Wuhan, China. Virus ini mempunyai sifat mudah menular dan menginfeksi sistem pernafasan manusia sehingga menyebar hampir keseluruh belahan dunia[1]. Seiring dengan perkembangan waktu virus yang menyebabkan COVID-19 mengalami mutasi gen. Mutasi gen ini menyebabkan semakin banyaknya varian virus, dengan demikian penyebarannya lebih cepat dan mematikan dari virus penyebab COVID-19 biasa, oleh karena itu hal ini menjadi masalah baru bagi seluruh negara di dunia termasuk Indonesia[2]. Berbagai cara telah dilakukan pemerintah Indonesia demi memutus mata rantai penyebaran infeksi COVID-19.

Pemerintah melalui Kementerian Kesehatan RI sebagai ujung tombak dalam menangani penyebaran COVID-19 di Indonesia selalu memberikan berita atau informasi terbaru tentang pandemi COVID-19 setiap hari melalui media sosial salah satunya Twitter. Namun dalam pelaksanaannya tidak luput dari komentar yang muncul dari kalangan masyarakat pengguna Twitter tentang mutasi COVID-19. Komentar yang diberikan tidak hanya bersifat positif, tetapi juga bersifat negatif, oleh karena itu komentar dari masyarakat Twitter merupakan data yang dapat diolah dan dijadikan bahan dari penelitian ini.

Dengan adanya permasalahan di atas cenderung dijadikan sebagai sumber perspektif untuk menentukan akurasi opini atau anggapan masyarakat luas terhadap mutasi COVID-19.

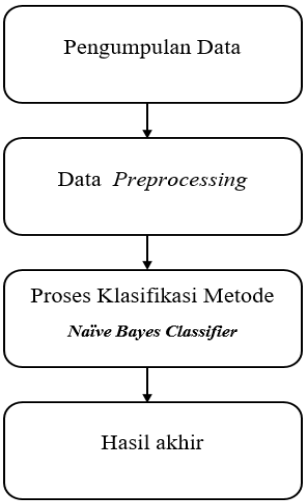
Penelitian ini memanfaatkan informasi yang diambil dari Twitter dan diolah terlebih dahulu dengan *text mining*. *Text mining* merupakan siklus untuk mendapatkan informasi yang berguna dari suatu catatan yang memiliki format tidak terstruktur. Proses ini digunakan untuk menangani permasalahan proses ekstraksi informasi dan dapat dimanfaatkan untuk mencari informasi yang belum ditemukan dan diketahui, secara umum *text mining* terdapat proses dalam pembersihan data, yaitu *preprocessing* data[3][4][5]. Sedangkan analisis sentimen adalah metode yang melibatkan pemahaman, penggalian, dan pengolahan data sastra secara otomatis untuk mengambil data sentimen yang berada dalam kalimat opini. Proses ini merupakan turunan dari *text mining* yang berguna untuk mengklasifikasikan ulasan ke dalam kelas positif dan negatif [6][7]. Salah satu metode yang sering digunakan untuk analisis sentimen adalah *Naïve Bayes Classifier.*

Penelitian [8] melakukan analisis sentimen untuk mengetahui opini dari film yang diperoleh melalui Twitter dengan memanfaatkan metode *Naïve Bayes Classifier*. Hasil penelitian menunjukan tingkat akurasi *Naïve Bayes Classifier* yang didapatkan sebanyak 90% dengan ulasan masing-masing diperoleh nilai dari *precission* 92%, *recall* 90% dan *f-measure* 90%. Penelitian [9] memakai perhitungan *Naïve Bayes Classifier* yang digunakan untuk mendapati pendapat dari masyarakat terhadap informasi harian COVID-19 yang terdapat pada akun Twitter Kementerian Kesehatan Republik Indonesia dengan presentase kelas positif 11%, kelas negatif 85%, dan 4% kelas natural

Naïve Bayes Classifier umumnya digunakan dalam proses pengelompokan statistik yang dapat digunakan untuk meramalkan kemungkinan keanggotaan *class,* dan memiliki kecepatan dan presisi yang sangat besar saat diterapkan pada data yang besar misalnya teks[10]. Kelebihan dari *Naïve Bayes Classifier* adalah bahwa teknik ini hanya memerlukan data latih yang relatif sedikit dalam menentukan batas-batas nilai yang diperlukan untuk satu siklus klasifikasi[11].

METODE

Dalam penelitian ini terdapat beberapa tahapan alur penelitian yang akan dilakukan. Alur disusun terstruktur dan divisualisasikan dalam bentuk gambar dengan tahapan-tahapan yang berurutan tentang apa saja yang akan dilakukan. Berikut merupakan alur untuk penelitian:

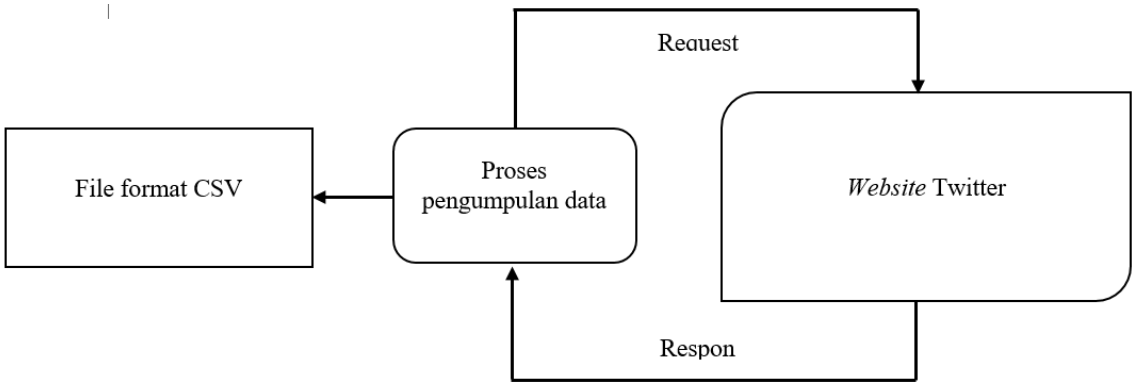


Gambar 1. Tahapan dari Penelitian

Pengumpulan Data

Pengumpulan data dikerjakan dengan mencari data berupa *tweet* yang terdapat pada situs www.twitter.com. Twitter dipilih dikarenakan memiliki API yang kompleks sehingga dapat mempermudah dalam proses pengumpulan data dengan mencari kata kunci tertentu sesuai data yang ingin dicari dan data yang disajikan merupakan opini masyarakat luas dalam

menyikapi hal atau kejadian tertentu misalnya mutasi COVID-19. Untuk penelitian kali ini dipergunakan data hasil dari kata kunci "mutasi covid-19" yang disimpan dalam bentuk *file* berformat CSV. Berikut gambaran dari proses pengumpulan data.



Gambar 2. Proses Pengumpulan Data

Data yang diambil dengan jumlah 300 *tweet* yang akan dibagi menjadi dua yaitu data beropini positif sebanyak 102 dan data beropini negatif sebanyak 198.

Data *Preprocessing*

Data *tweet* yang sudah didapatkan akan diproses pada tahap *preprocessing* yang terdiri dari:

1. *Case Folding*
   Case Folding adalah rangkaian prosedur untuk penyeragaman ukuran huruf pada data *tweet* menjadi huruf kecil. Karakter selain huruf akan dihilangkan atau dianggap sebagai pembatas. Proses penyeragaman dijalankan dengan cara memeriksa seluruh data dari awal sampai selesai. Apabila ditemukan huruf kapital, maka huruf yang tercantum akan diproses sebagai huruf kecil.
2. *Tokenizing*
   Tokenizing adalah tahapan pemotongan sebuah data *tweet* menjadi beberapa bagian kata. Pada saat yang sama *tokenizing* juga berfungsi untuk membuang karakter tertentu.
3. *Stopword Removal*
   Stopword Removal adalah proses penghapusan kata yang tidak penting. Setiap kata ganti, kata depan, kata sambung atau kata yang tidak penting akan dihilangkan. Proses pembuangan kata-kata yang tidak perlu dilakukan dengan membandingkan kata hasil sebelumnya dengan daftar kata yang terdapat pada *database*, lantas dilaksanakan pemeriksaan pada kata apakah ada yang serupa pada daftar atau tidak, apabila didapatkan kata yang serupa terletak dalam *database*, maka kata yang tertera bakal dihapus atau dihilangkan.
4. *Stemming*
   Stemming adalah tahapan untuk mengubah kata tambahan menjadi kata dasar yang sesuai standar Bahasa Indonesia. *Stemming* merupakan teknik pemrosesan teks dasar yang kerap digunakan untuk mengembangkan kemampuan lebih lanjut dalam pengklasifikasian teks dan pemulihan teks.

Proses Klasifikasi

Setelah data telah selesai pada tahap *preprocessing*, langkah berikutnya yaitu proses klasifikasi. Dalam penelitian yang dijalankan memanfaatkan metode *Naïve Bayes Classifier* pada saat proses klasifikasi.

HASIL DAN PEMBAHASAN

Pada penelitian ini dipergunakan data yang berasal dari Twitter memakai *library tweepy* dengan kata kunci "mutasi covid" menggunakan program berikut: data yang berjumlah 300 akan dipisah menjadi dua bagian masing-masing dengan 102 data positif dan 198 data negatif disimpan pada *file* berformat CSV. Selanjutnya, data tersebut melalui tahap *preprocessing*, *testing* dan *training* sehingga didapatkan nilai *accuracy*, *recall*, dan *f1-score*. Pada Tabel 1 berisi sampel data yang sudah melalui *preprocessing*.

Tabel 1. Sampel Data

Data	Label
Disiplin kesehatan untuk mencegah munculnya mutasi covid	Positif
Gak kelar mutasi covid	Negatif
Gk perlu gue pingin cepat vaksin mutasi covid khawatir	Negatif
Waspada mutasi covid masuk ke tahap yang lebih berbahaya	Positif
Alhamdulillah semoga wabah segera hilang	Positif
Virus banyak sekali mutasi makin ngeri	Negatif

Data yang sudah melalui tahap data *preprocessing* dan diberi label selanjutnya diuji akurasinya. Pengetesan akurasi pada proses klasifikasi dijalankan agar dapat diketahui tingkat akurasi dalam pengklasifikasian data berupa *tweet* memanfaatkan metode pengklasifikasian *Naïve Bayes Classifier*. Teknik yang dipergunakan dalam pengetesan adalah *confusion matrix* merupakan matriks yang diperlukan dalam pembandingan antara prediksi dengan data asli yang sebelumnya sudah di-*input*. Tabel 2 memperlihatkan representatif kelas data dan kelas prediksi yang tersaji pada matrix 2×2.

Tabel 2. Representatif *Confusion Matrix*

Data Aktual	Data Prediksi	
	Positif	**Negatif**
Positif	TP	FP
Negatif	FN	TN

Pada Tabel 2 dijelaskan bahwa *confusion matrix* merupakan hasil dari prediksi menggunakan metode *Naïve Bayes Classifier* dengan mengukur performa tiap kelas dengan menghitung *recall, precission,* dan *F1-score*. Selanjutnya dilakukan *input* data sesuai yang ada pada Tabel 3.

Tabel 3. Hasil *confusion matrix*

Data Aktual	Data Prediksi	
	Positif	**Negatif**
Positif	92	10
Negatif	30	168

Berdasarkan tabel didapatkan jumlah data yang diprediksi oleh sistem, yaitu sebanyak 92 data diprediksi sebagai *true positive*, 10 data diprediksi sebagai *false positive*, 30 data diprediksi sebagai *false negative*, dan 168 data diprediksi sebagai *true negative*.

Selanjutnya data yang sudah diprediksi dilakukan pengujian akurasi menggunakan library *Naïve Bayes Classifier* dengan bantuan *software jupyter notebook*. Hasil pengujian diperlihatkan pada Gambar 3.

	precision	recall	f1-score	support
0	0.85	0.94	0.89	18
1	0.90	0.75	0.82	12
accuracy			0.87	30
macro avg	0.88	0.85	0.86	30
weighted avg	0.87	0.87	0.86	30

Accuracy Score = 86.66666666666667

Gambar 3. Hasil pengujian dengan Jupyter Notebook

Berdasarkan gambar di atas pengujian akurasi menggunakan bantuan *software jupyter notebook* menghasilkan nilai akurasi sebesar 86.67%. Angka 0 menunjukan kelas negatif dengan nilai *precision* sebesar 85%, *recall* sebesar 94%, dan *f1-score* sebesar 89%, sedangkan angka 1 menunjukan kelas positif dengan nilai *precision* sebesar 90%, *recall* sebesar 75%, dan nilai *f1-score* sebesar 82%. Hasil perhitungan manual dan pengetesan untuk penelitian dapat dilihat pada Tabel 4.

$$\text{Akurasi} = \frac{\text{jumlah data benar}}{\text{jumlah data keseluruhan}} = \frac{92+168}{300} = 0.8667 \times 100\%$$

$$= 86.67\%$$

$$\textit{Recall positive} = \frac{TP}{TP+FN} = \frac{92}{92+30} = 0.7540 \times 100\%$$

$$= 75.40\%$$

$$\textit{Recall negative} = \frac{TN}{TN+FP} = \frac{168}{168+10} = 0.943 \times 100\%$$

$$= 94.38\%$$

$$\textit{Precision positive} = \frac{TP}{TP+FP} = \frac{92}{92+10} = 0.901 \times 100\%$$

$$= 90.19\%$$

$$\textit{Precision negative} = \frac{TN}{TN+FN} = \frac{168}{168+30} = 0.848 \times 100\%$$

$$= 84.84\%$$

$$\textit{F1-Score positive} = 2 * \frac{recall*precission}{recall+precission} = 2 * \frac{0.75*0.90}{0.75+0.90} = 0.818 \times 100\%$$

$$= 81.81\%$$

$$\textit{F1-Score negative} = 2 * \frac{recall*precission}{recall+precission} = 2 * \frac{0.85*0.94}{0.85+0.94} = 0.892 \times 100\%$$

$$= 89.27\%$$

Tabel 4. Hasil Pengujian

	Nilai
Akurasi	86.67%
Recall positive	75.40%
Recall negative	94.38%
Precision positive	90.19%
Precision negative	84.84%
F1-score positive	81.81%
F1score negative	89.27%

Hasil dari Tabel 4 di atas dapat disederhanakan seperti Tabel 5 dibawah ini.

Tabel 5. Hasil Pengujian Keseluruhan

	Nilai
Akurasi	86.67%
Recall	84.89%
Precision	87.51%
F1-score	85.54%

Berdasarkan hasil Tabel 5 di atas, didapatkan format hasil seperti yang ada pada penelitian [8] dan penelitian [9]. Untuk penelitian ini memiliki nilai akurasi dari klasifikasi data *tweet* memakai metode *Naïve Bayes Classifier* menghasilkan nilai sebesar 86.67%, *recall* dengan nilai sebesar 84.89%, *precision* dengan nilai sebesar 87.51%, dan *f1-score* dengan nilai sebesar 85.54%.

SIMPULAN DAN SARAN

Berdasar hasil yang terdapat pada penelitian yang telah dilakukan di atas bisa disimpulkan bahwa metode *Naïve Bayes Classifier* bisa dipergunakan dalam proses menganalisis data sentimen masyarakat luas terhadap mutasi COVID-19 dari media sosial Twitter. Untuk hasil analisis disimpulkan bahwa sentimen masyarakat terhadap mutasi COVID-19 berkencenderungan sentimen negatif seperti takut, sedih, marah dan sebagainya. Pada penelitian selanjutnya penulis menyarankan untuk menggunakan jumlah data lebih besar, semakin besar data yang dipergunakan maka dapat menghasilkan nilai akurasi yang lebih baik dari penelitian ini dan diharapkan dapat menggunakan data dari media sosial lain sehingga dapat dibandingkan data media sosial mana yang lebih akurat.

DAFTAR PUSTAKA

[1] M. K. Dr. Mashuri Masri, S.Si., M. K. Delima Engga Maretha, S.Pd., and M. S. Rusny, S.Pt., *Everything About Corona*. 2020.

[2] E. Parwanto, "Virus Corona (SARS-CoV-2) penyebab COVID-19 kini telah bermutasi," *J. Biomedika dan Kesehat.*, vol. 4, no. 2, 2021, doi: 10.18051/jbiomedkes.2021.v4.47-49.

[3] S. Samsir, A. Ambiyar, U. Verawardina, F. Edi, and R. Watrianthos, "Analisis Sentimen Pembelajaran Daring Pada Twitter di Masa Pandemi COVID-19 Menggunakan Metode Naïve Bayes," *J. Media Inform. Budidarma*, vol. 5, no. 1, pp. 157–163, 2021.

[4] D. Ariyanti and K. Iswardani, "Teks Mining untuk Klasifikasi Keluhan Masyarakat Pada Pemkot Probolinggo Menggunakan Algoritma Naïve Bayes," *J. IKRA-ITH Inform.*, vol. 4, no. 3, 2020.

[5] R. Rasenda, H. Lubis, and R. Ridwan, "Implementasi K-NN Dalam Analisa Sentimen Riba Pada Bunga Bank Berdasarkan Data Twitter," *J. MEDIA Inform. BUDIDARMA*,

vol. 4, no. 2, 2020, doi: 10.30865/mib.v4i2.2051.

[6] A. Harun and D. P. Ananda, "Analisa Sentimen Opini Publik Tentang Vaksinasi Covid-19 di Indonesia Menggunakan Naïve Bayes dan Decission Tree," *Indones. J. Mach. Learn. Comput. Sci.*, vol. 1, no. April, 2021.

[7] F. A. Maulana, I. Ernawati, P. Labu, and J. Selatan, "Analisa sentimen cyberbullying di jejaring sosial twitter dengan algoritma naïve bayes," *Semin. Nas. Mhs. Ilmu Komput. dan Apl. (SENAMIKA*, 2020.

[8] F. Ratnawati, "Implementasi Algoritma Naive Bayes Terhadap Analisis Sentimen Opini Film Pada Twitter," *INOVTEK Polbeng - Seri Inform.*, vol. 3, no. 1, 2018, doi: 10.35314/isi.v3i1.335.

[9] E. T. Handayani and A. Sulistiyawati, "Analisis Setimen Respon Masyarakat Terhadap Kabar Harian Covid-19 Pada Twitter Kementerian Kesehatan Dengan Metode Klasifikasi Naive Bayes," *J. Teknol. dan Sist. Inf.*, vol. 2, no. 3, pp. 32–37, 2021.

[10] J. Song, K. T. Kim, B. Lee, S. Kim, and H. Y. Youn, "A novel classification approach based on Naïve Bayes for Twitter sentiment analysis," *KSII Trans. Internet Inf. Syst.*, vol. 11, no. 6, pp. 2996–3011, 2017.

[11] 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, 2017, doi: 10.24176/simet.v8i1.841.