**Jurnal Komputer Antartika**
Volume 1 Nomor 1 2023
ISSN: 2988-5272

# Implementasi *Machine Learning* Untuk Klasifikasi Narasi *Informative* dan *Non-Informative* Pada Media Sosial Twitter TMC Polda Metro Jaya Menggunakan Naïve Bayes Classifier

**Nendi Riska Fadhila[1*], Tri Sutarti[2], Melly Patende[3], Galistia Yudha[4], Dedi Dwi Saputra[5]**

[1,2,3,4,5] Program Studi Sistem Informasi
Fakultas Teknologi dan Informasi
Universitas Nusa Mandiri
Jalan Kramat Raya No. 18, Jakarta Pusat 10420, Indonesia

e-mail korespondensi: 1220292@nusamandiri.ac.id

**Abstrak** - Twitter adalah layanan jejaring sosial yang memungkinkan penggunanya untuk mengirim dan membaca pesan berbasis teks hingga 140 karakter. Sejak diluncurkan, twitter telah menjadi salah satu dari sepuluh situs yang paling sering dikunjungi di internet. Adapun pengklasifikasian narasi yang bersifat *informative* atau *non-informative* pada media sosial twitter TMC Polda Metro Jaya terhadap pandangan masyarakat, menggunakan metode klasifikasi alogoritma *Naive Bayes Classifier*. Penelitian ini dilakukan dengan mengambil 1.000 sampel data dengan *keyword* @TMCPoldaMetro untuk dijadikan dataset. Dengan menggunakan *preprocessing* data, *annotation removal, remove hastag, regexp, transformation remove URL, Indonesian stemming* dan *Indonesia stop removal*. Hasil analisis penelitian yaitu nilai *Accuracy* sebesar 81,7%, nilai *Pression* sebesar 81,30%, nilai *Recall* sebesar 83,03% dan *AUC* sebesar 0,847.

Kata Kunci: Twitter, *Naive Bayes Classifier*, TMC Polda Metro Jaya

***Abstracts*** *- Twitter is a social networking service that enables its users to send and read text-based messages of up to 140 characters. Since its launch, Twitter has become one of the ten most visited websites on the internet. As for the classification of narratives that are informative or non-informative on TMC Polda Metro Jaya's social media twitter to the views of the public, using the Naive Bayes Classifier algorithm classification method. This research was conducted by taking 1,000 data samples with the keyword @TMCPoldaMetro to be used as a dataset. By using data preprocessing, annotation removal, remove hashtags, regexp, transform remove URLs, Indonesian stemming, and Indonesian stop removal. The results of the research analysis are Accuracy of 81.7%, Pressure of 81.30%, Recall of 83.03%, and AUC of 0.847.*

*Keywords: Twitter, Naive Bayes Classifier, TMC Polda Metro Jaya*

## 1. Pendahuluan

Dunia teknologi informasi dan komunikasi sudah semakin berkembang dengan pesat seiring berjalannya waktu banyaknya penyedia layanan web yang menyediakan infromasi sangat banyak. Informasi seperti misalnya data teks yang bisa dijadikan sumber informasi, dengan contohnya adalah data teks yang diambil dari Twitter.

Twitter adalah layanan jejaring sosial dan microblog daring yang memungkinkan penggunanya untuk mengirim dan membaca pesan berbasis teks hingga 140 karakter, yang dikenal dengan sebutan kicauan (*tweet*). Twitter didirikan pada bulan Maret 2006 oleh Jack Dorsey, dan situs jejaring sosialnya diluncurkan pada bulan Juli. Sejak diluncurkan, twitter telah menjadi salah satu dari sepuluh situs yang paling sering dikunjungi di internet, dan dijuluki dengan pesan singkat dari internet. Di twitter, penguna tak terdaftar hanya bisa membaca kicauan, sedangkan pengguna terdaftar bisa menulis kicauan melalui antarmuka situs web, pesan singkat (SMS) atau melalui berbagai aplikasi untuk perangkat seluler [1].

Dalam [2], didalam sebuah *Machine Learning* terdapat beberapa parameter yang mempunyai karakter objek dengan metode klasifikasi ini. Ada dua tahap dalam pembuatan model klasifikasi ini yaitu yang pertama adalah melatih data dan membedakannya dengan kelas yang berbeda, dan yang kedua adalah klasifikasi mengevaluasi objek yang tidak dikenal.

Menurut IBM, metode *machine learning* merupakan cabang dari kecerdasan buatan (*Artificial Intelligence*) dan ilmu komputer yang berfokus pada penggunaan data dan algoritma untuk meniru cara manusia



This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License. Published under licence by Antartika Media Indonesia.

belajar dan secara bertahap dapat meningkatkan akurasinya. Semakin bagus algoritma *machine learning* yang digunakan maka akan semakin baik pula keputusan yang keluar [3].

Adapun maksud dan tujuan penulis terkait impelementasi klasifikasi narasi bersifat *informative* atau *non-imformative* pada media sosial twitter TMC Polda Metro Jaya terhadap pandangan masyarakat, yang diambil dari *tweet* dengan *keyword* @TMCPoldaMetro di aplikasi Twitter. Penulis menggunakan metode berupa klasifikasi alogoritma *Naive Bayes*.

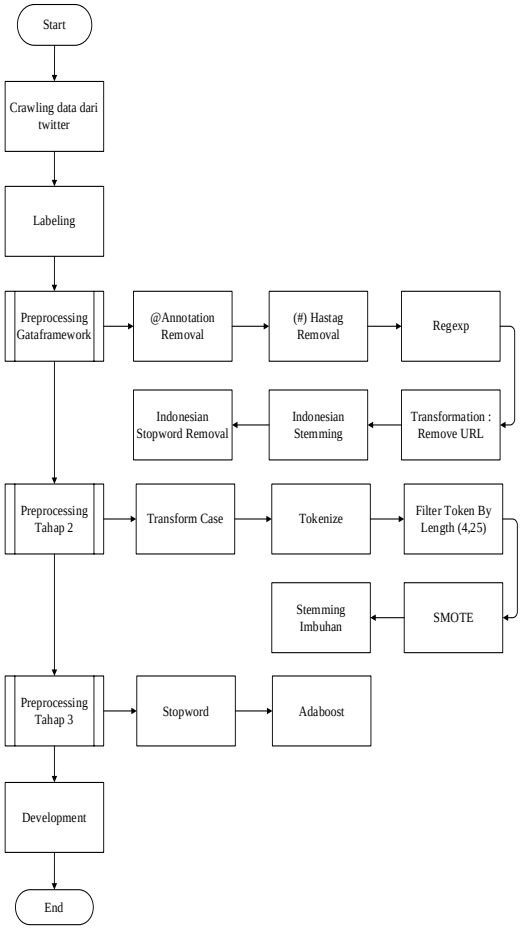
## 2. Metode Penelitian

*Naïve Bayes* atau *multinomial naïve bayes* merupakan metode yang digunakan untuk mengklasifikasikan sekumpulan dokumen. Algoritma ini memanfaatkan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes. Metode NB menempuh dua tahap dalam proses klasifikasi teks, yaitu tahap pelatihan dan tahap pengujian (klasifikasi). Pada tahap pelatihan dilakukan proses analisis terhadap sampel dokumen berupa pemilihan *vocabulary*, yaitu kata yang mungkin muncul dalam koleksi dokumen sampel yang sedapat mungkin dapat menjadi representasi dokumen.

Selanjutnya adalah penentuan probabilitas prior bagi tiap kategori berdasarkan sampel dokumen. Pada tahap klasifikasi ditentukan nilai kategori dari suatu dokumen berdasarkan term yang muncul dalam dokumen yang diklasifikasi [4].

Berikut adalah contoh dari penelitian [5] Proses klasifikasi pada penelitian ini menggunakan metode klasifikasi *Naïve bayes Classifier* (NBC) dan *Support Vector Machine* (SVM) dengan *preprocessing* data menggunakan *tokenisasi*, *cleansing* dan *filtering*. Data yang digunakan adalah tweet dalam bahasa Indonesia dengan tagar *HateSpeech* (*#HateSpeech*), dengan jumlah dataset sebanyak 522 tweet yang didistribusikan secara merata menjadi dua sentimen *HateSpeech* dan *GoodSpeech*. Hasil akurasi tertinggi didapatkan saat menggunakan metode klasifikasi *Support Vector Machine* (SVM) dengan *tokenisasi unigram*, *stopword list* Bahasa Indonesia dan *emoticons*, 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%.

Kerangka dari penelitian ditampilkan pada Gambar 1:




Gambar 1. Kerangka Penelitian

Dari gambar 1 dapat dijelaskan sebagai berikut :

1. *Crawling* data dari Twitter

Tahap pertama yaitu melakukan pengumpulan data dengan aplikasi RapidMiner Studio, *create connection* untuk menghubungkan aplikasi RapidMiner ke twitter agar dapat terhubung maka akan didapatkan nomor token untuk bisa mengakses data melalui akun twitter.

2. *Labelling*

Tahap kedua yaitu melakukan klasifikasi label pada data yang sudah terkumpul dari twitter.

3. *Preprocessing Gataframework*

*Preprocessing* menjadi tahap awal dalam klasifikasi teks untuk mempersiapkan data teks sebelum digunakan pada proses lainnya. Pada tahap ini akan mengubah data teks menjadi bentuk yang lebih baik sehingga menghasilkan informasi teks dengan kualitas yang baik dan siap digunakan pada proses selanjutnya [6].

4. *Preprocessing* Tahap 2

Tahap ini akan mengimport data dari excel yang sudah melewati *Preprocessing* tahap 1 data yang digunakan yaitu tabel *no*, *class* dan *Indonesia stopword* lalu selanjutnya diberi atribut dengan mengedit *transform case*, *tokenize*, *filter token by length (4,25*), SMOTE dan *stemming* imbuhan.

5. *Preprocessing* Tahap 3

Stopword removal, merupakan tahap penghapusan kata-kata yang terkandung pada daftar kata stopword. Kata-kata stopwords dapat menambah dimensi data pada proses klasifikasi
Stopword removal juga digunakan untuk menghilangkan kata yang tidak relevan atau yang tidak bermakna [7].

6. *Deployment*

Tahap terakhir pembuatan program menggunakan *tools MySQL* merupakan salah satu database yang bersifat *open source* yang populer di dunia, dimana saat ini digunakan lebih dari 100 juta penguna diseluruh dunia *MYSQL* turunan dari SQL (*Structured Query Languaqe*) yang sering digunakan oleh para *programmer* dalam pembuatan program. Sistem database *MYSQL* selain mudah memahami sintaks dan pengaksesan juga di dukung beberapa fitur seperti *Multithreaded*, *multi-user* dan SQL database managemen sistem [8].

Dan *tools XAMPP* adalah sebuah *software* web server *Apache* yang didalamnya sudah tersedia database server *MYSQL* dan mendukung PHP *programming*. *XAMPP* merupakan singkatan dari X (untuk empat sistem operasi), *Apache*, *MySQL*, PHP, Perl. *XAMPP* menyediakan berbagai macam layanan salah satunya adalah *Apache* untuk web server [9].

7. *Table Matrix*

Dalam penelitian ini hasil dari pengujian akan di evaluasi dalam tabel *Accuracy*, *Pression*, *Recall* dan *AUC*.
Nilai *Accuracy*, adalah perhitungan jumlah proporsi prediksi yang benar. *Accuracy* dirumuskan dalam persaman[10].

Nilai *Pression*, yakni tingkat ketepatan prediksi dari suatu sistem, dengan menghitung prediksi positif dari total data yang diprediksi sistem, termasuk prediksi salah [11].

Nilai *Recall*, yaitu, nilai tingkat keberhasilan dalam mengenali suatu kelas yang harus dikenal [11].

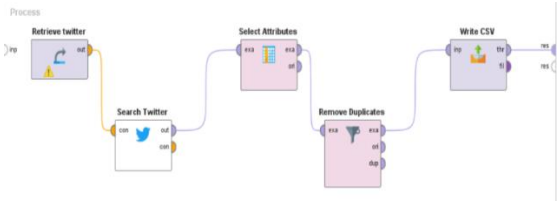
Nilai *AUC*, digunakan untuk mengukur kinerja diskriminatif dengan memperkirakan probabilitas *output* yang sudah di dapatkan hasilnya dari sampel yang dipilih secara acak dari populasi positif atau negatif, semakin besar nilai *AUC*, semakin kuat klasifikasi yang dihasilkan. Karena *AUC* merupakan bagian dari daerah unit persegi, nilainya yang dihasilkan akan selalu sama yang dihasilkannya, antara 0,0 dan 1,0 [12].

## 3. Hasil dan Pembahasan

Berdasarkan metode dan tahapan penelitian yan sudah dijabarkan, maka langkah selajutnya penulis akan menjelaskan hasil dari pembahasan terhadap penelitian yang penulis kerjakan sebagai berikut :

3.1 *Crawling* data dari Twitter.

Pengumpulan data dengan aplikasi RapidMiner Studio, *create connection* untuk menghubungkan aplikasi RapidMiner ke twitter, gunakan operasi *search* twitter menggunakan *keyword* @TMCPoldaMetro dengan jumlah data 1001. Berikut gambar dari proses *Crawling* data dari Twitter :




Gambar 2. *Crawling* data dari Twitter

Dari proses data yang dihasilkan lalu selanjutnya disimpan dalam bentuk *file excel*.

3.2 *Labelling*

Dari 1000 jumlah data, di tahap ini akan dilakukan labelling, penulis menggunakan klasifikasi narasi bersifat *Informative* dan *Non-Informative.*

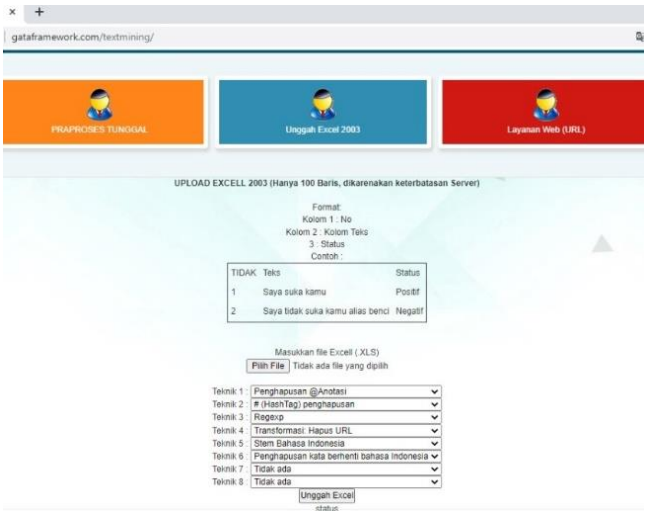
| Nc | Text | Class |
|---|---|---|
| 1 | 23.02: RT @TMCPoldaMetro: Genangan air di Jl. Kemang Raya, Kec. Mampang Prapatan Jaksel hingga sekarang masih belum bisa dilalui kendaraan. https://t.co/qlm1rbYac3 | Informative |
| 2 | RT @dangiangkuning1: cc.. @DivHumas_Polri @CCICPolri @humaspoldajbr @humas_polda_ntt @polda_bali @humaspoldagtlo @TMCPoldaMetro @polda_malu... | Non-Informative |
| 3 | #Polisi #Jakarta #DKIJakarta @TMCPoldaMetro: 'LOKASI PELAYANAN #SIMKELILING Rabu, 12 Oktober 2022 Pukul 08.00 s/d 14.00 WIB : | Non-Informative |
| 4 | RT @arya_saloka: Tolong sekitar jalan raya pasar minggu sampai pancoran diperhatikan. Ini tidak bergerak @TMCPoldaMetro | Informative |
| 5 | RT @TMCPoldaMetro: 13.12 Kegiatan Penyampaian Pendapat di depan Baharkam Polri Jaksel masih berlangsung, Situasi arus lalu lintas terpantau... | Informative |
| 6 | RT @TMCPoldaMetro: LOKASI PELAYANAN #SIMKELILING Rabu, 12 Oktober 2022 Pukul 08.00 s/d 14.00 WIB : | Non-Informative |
| 7 | Ingat Ada 25 Lokasi Ganjil Genap Jakarta Rabu 12 Oktober, Hanya untuk Kendaraan Pelat Genap | Informative |
| 8 | RT @hendratno64: Terjadi kecelakaan lalu lintas di Tol Jagorawi arah Jakarta, di ruas kanan, sebelum exit tol Rambutan. | Informative |
| 9 | Terjadi kecelakaan lalu lintas di Tol Jagorawi arah Jakarta, di ruas kanan, sebelum exit tol Rambutan | Informative |

Gambar 3. Hasil *Labelling*

Berikut terpilih 580 data dengan klasifikasi *class Informative* dan 420 untuk *class Non-Informative*.

3.3 *Preprocessing Gataframework*

Dari hasil *labelling*, maka selanjutnya akan masuk di tahan *Preprocessing* tahap 1, dengan membaginya ke dalam beberapa file yang memuat 50 data dikarenakan keterbatasan *web server* pada *Gataframework*. Di tahap ini dilakukan cleansing dengan mengunakan beberapa teknik pada website *Gataframework* seperti *@annotation Removal*, *Remove Hastag*, *Regexp*, *Transformation Remove URL*, *Indonesian Stemming* dan *Indonesian Stop Removal*.



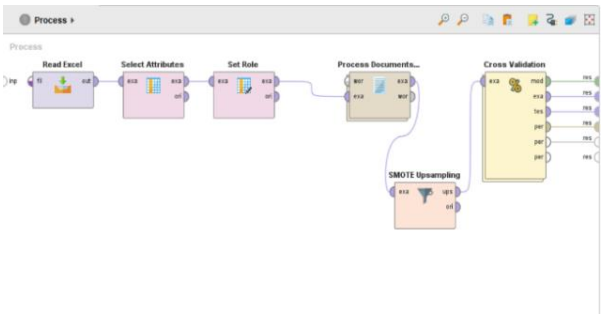

Gambar 4. *Preprocessing* Tahap 1

1.*@annotation Removal*, tujuannya untuk mengahapus *annotation* didalam teks tweet. *Annotation* sendiri salah satu noise yang tidak memiliki arti.
2. *Remove Hastag*, tujuannya untuk mengahpus tagar pada teks tweet.
3. *Regexp*, tujuanya untuk menghapus simbol-simbol pada teks tweet.
*4.Transformation Remove URL*, tujuannya untuk mengapus adanya *URL/link* yang terdapat pada teks tweet.

5. *Indonesian Stemming*, merupakan proses menemukan kata dasar dari kata berimbuhan (*affixed word*) dengan cara menghilangkan semua imbuhan *(affix*) yang terdiri dari awalan (*prefix*), sisipan (*infix*), akhiran (*suffix*) dan kombinasi awalan dan akhiran (*confix*).
6. *Indonesian Stop Removal*, tujuannya untuk menghapus kata penghubung dalam teks tweet.

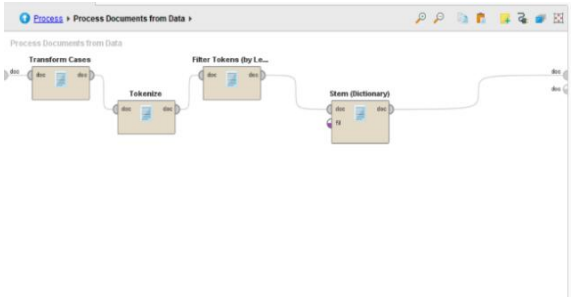
3.4 *Preprocessing* Tahap 2

Pada tahapan ini akan dilakukan pengujian melalui aplikasi RapidMiner, di *Preprocessing* Tahap 2 harus melewati beberapa proses uji coba pada data sehingga menghasilkan data yang lebih akurat.




Gambar 5. *Preprocessing* Tahap 2 *SMOTE*

Tahap ini akan mengimport data dari excel yang sudah melewati *Preprocessing* tahap 1 data yang digunakan yaitu tabel *no*, *class* dan *Indonesia stopword*, lalu kemudia melakukan perubahan pada parameter *role* menjadi status.

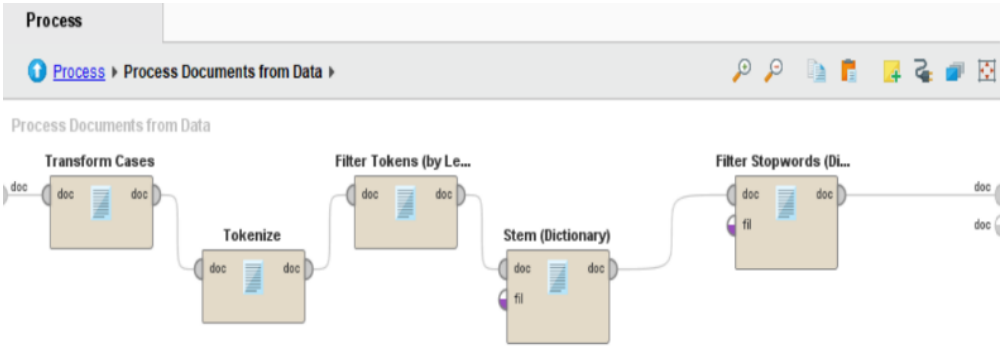



Gambar 6. *Preprocessing* Tahap 2

Pada *Preprocessing* Tahap 2 ini, ada beberapa proses yaitu :
1. *Transform Cases*, ini bertujuan untuk membuat teks tweet dari huruf kapital ke huruf kecil biasa.
2. *Tokenize*, ini bertujuan untuk memisahkan kata per kata dari suatu teks kalimat.
3. *Filter Token By Length* (4,25), ini bertujuan untuk memfilter kata-kata yang panjangnya 4 sampai 25 karakter.
4. Stemming imbuhan, ini bertujuan untuk menghilangkan imbuhan pada data seperti “di”, “nya”,”an” dan imbuhan lainnya.

3.5 *Preprocessing* Tahap 3

Pada tahap ini dilakukan pembobotan nilai pada data yang dihasilkan dari proses sebelumnya, pembobotan ini nantinya untuk mengetahui teks tersebut relevan dengan objek yang penulis teliti.




Gambar 8. *Preprocessing* Tahap 3

Dalam *Preprocessing* Tahap 3 adanya penambahan operator filter stopword. Data ini diimpor dari data hasil nilai pembobotan masing-masing kata.

| Row Labels | Sum of Bobot |
|---|---|
| cijantung | 0,2233 |
| mgkn | 0,2233 |
| amerika | 0,2376 |
| elektronik | 0,2376 |
| dollar | 0,2376 |
| kkxy | 0,2376 |
| noxk | 0,2376 |
| pelintasan | 0,2376 |
| terap | 0,2376 |
| ambulan | 0,2439 |
| bottle | 0,2439 |
| goblok | 0,2439 |
| kondisil | 0,2439 |
| kritis | 0,2439 |

Gambar 6. Hasil *Stopword*

Gambar ini menjelaskan bahwa semakin kecil nilai pembobotan maka kata tersebut tidak ada relevannya dengan objek penelitian.

**3.6 Evaluasi**

Tahap terakhir hasil dari proses metode yang sudah didapatkan, maka di tahap ini akan dilakukan untuk menguji hasil dari klasifikasi dengan menggunakan tabel *matrix* dari sejumlah data yang sudah diuji, untuk mencari nilai *Accuracy*, *Pression*, *Recall* dan *AUC*

| Teknik/Feature | Algorithm | Accuracy | Presisi | Recall | AUC |
|---|---|---|---|---|---|
| Preprocessing Gataframewor | NB | 81,74% | 81,30% | 83,03% | 0,847 |
| Transform Case | | | | | |
| Tokenize | | | | | |
| Filter Token By Length (4,25) | | | | | |
| SMOTE | | | | | |
| Stemming | | | | | |
| Stopword | | | | | |
| Adaboost | | | | | |

Gambar 7. Tabel *Matrix*

Gambar 8. Grafik ROC Algoritma *Naive Bayes*

Dengan menggunakan algoritma Naive Bayes, maka diperoleh nilai *Accuracy* sebesar 81,7%, nilai *Pression* sebesar 81,30%, nilai *Recall* sebesar 83,03% dan *AUC* sebesar 0,847.

**4. Kesimpulan**

Pengujian terhadap hasil *crawling dataset* di media sosial twitter dengan *keyword* @TMCPoldaMetro dengan menggunakan algortima *Naive Bayes* telah berhasil diteliti, metode yang digunakan *text mining* dan *naive bayes* terbukti berhasil mengklasifikasikan narasi yang bersifat *Informative* dan *Non-Imformative* dengan adanya hasil dari nilai-nilai yang diperoleh penulis yakni diperoleh nilai *Accuracy* sebesar 81,7%, nilai *Pression* sebesar 81,30%, nilai *Recall* sebesar 83,03% dan *AUC* sebesar 0,847 yang dijabarkan melalui tabel *Matrix*.

**Referensi**

[1] F. Ratnawati, "Implementasi Algoritma Naive Bayes Terhadap Analisis Sentimen Opini Film Pada Twitter," *INOVTEK Polbeng - Seri Informatika*, vol. 3, no. 1, p. 50, 2018, doi: 10.35314/isi.v3i1.335.

[2] Fahrizal, O. Fais, and H. Noer, "IMPLEMENTASI MACHINE LEARNING PADA SISTEM PETS IDENTIFICATION MENGGUNAKAN PYTHON BERBASIS UBUNTU JISICOM ( Journal of Information System , Informatics and Computing ) p-ISSN : 2579-5201 ( Printed ) JISICOM ( Journal of Information System , Informatics and," vol. 4, no. 1, pp. 86–91, 2020.

[3] I. M. Faiza and W. Andriani, "Tinjauan Pustaka Sistematis : Penerapan Metode Machine Learning untuk Deteksi Bencana Banjir," vol. 11, no. September, pp. 59–63, 2022.

[4] A. D. Wibisono, S. Dadi Rizkiono, and A. Wantoro, "Filtering Spam Email Menggunakan Metode Naive Bayes," *TELEFORTECH : Journal of Telematics and Information Technology*, vol. 1, no. 1, pp. 9–17, 2020, doi: 10.33365/tft.v1i1.685.

[5] A. Karim, "ANALISIS SENTIMEN PADA KOMENTAR SOSIAL MEDIA INSTAGRAM LAYANAN KESEHATAN BPJS MENGGUNAKAN NAÏVE BAYES CLASSIFIER," 2021.

[6] S. Khairunnisa and S. Al Faraby, "Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter ( Studi Kasus Pandemi," vol. 5, no. April, pp. 406–414, 2021, doi: 10.30865/mib.v5i2.2835.

[7] D. Muhidin and A. Wibowo, "PERBANDINGAN KINERJA ALGORITMA SUPPORT VECTOR MACHINE DAN K-NEAREST NEIGHBOR TERHADAP ANALISIS," vol. 5, no. 2, pp. 153–159, 2020.

[8] Huzaeni, Akmalulfata, Mahdi, and Ismaniar, "Pelatihan Web Design Bagi Siswa SMK Negeri 5 Kecamatan Blang Mangat Kota Lhokseumawe," vol. 3, no. 1, pp. 139–142, 2019.

[9] A. D. Rachmatsyah, B. Isnanto, S. H. Saputro, and E. Helmud, "Pelatihan Pembuatan Web Dengan PHP Dan WordPress Pada SMA Negeri 4 Pangkalpinang," 2021.

[10] D. Setiawan, R. N. Putri, and R. Suryanita, "khazanah informatika Perbandingan Algoritma Genetika dan Backpropagation pada Aplikasi Prediksi Penyakit Autoimun," pp. 21–27, 2019.

[11] M. Syarifuddin, "Inti nusa mandiri," vol. 15, no. 1, pp. 23–28, 2020.

[12] S. Dwiasnati and Y. Devianto, "Optimasi Prediksi Keputusan Calon Nasabah Potensial Berbasis Particle Swarm Optimization," vol. 6, no. 2, pp. 286–292, 2019.