

Analisis Sentimen *Physical Distancing* pada Twitter Menggunakan *Text Mining* dengan Algoritma *Naive Bayes Classifier*

¹Nilia Hardi, ²Yuris Alkahfi, ³Popon Handayani, ⁴Windu Gata, ⁵Muhammad Rifqi Firdaus*

^{1,2,4,5}Program Studi Magister Ilmu Komputer, STMIK Nusa Mandiri

³Program Studi Sistem Informasi, STMIK Nusa Mandiri

Jl. Kramat Raya No.18, RT.5/RW.7, Kwitang, Kec. Senen, Kota Jakarta Pusat, Daerah Khusus
Ibukota Jakarta 10450

*e-mail: 14002358@nusamandiri.ac.id

(received: 11 November 2020, revised: 21 November 2020, accepted: 6 Desember 2020)

Abstrak

Physical distancing kini sedang ramai menjadi perbincangan publik sebagai salah satu cara pemerintah dalam menekan penyebaran virus covid-19 yang sedang melanda beberapa negara di belahan dunia. Tidak tersaringnya cuitan terkait *physical distancing* bisa memunculkan berbagai macam opini, tidak hanya opini yang positif tapi juga yang negatif. Maka dari itu, Twitter di anggap lebih diminati oleh masyarakat indonesia dikarenakan twitter dirasa lebih mudah untuk mengungkapkan opininya. Metode yang digunakan yaitu *Naive Bayes Classifier* (NBC). Data terkumpul dilakukan filter dari cuitan tersebut dengan menghapus data yang double maka setelah di filter data yang terambil yaitu sebanyak 547 *tweets*. Pada perhitungan analisis sentimen terhadap *physical distancing* di tengah pandemi covid-19 menggunakan metode NBC memperoleh hasil akurasi sebesar 50,26%. Tujuan dari penelitian ini, agar dapat mengkategorikan opini negatif atau positif, dari pembahasan *physical distancing*. Nantinya informasi terkait kebijakan *Physical Distancing* bisa sampai tepat informasinya kepada masyarakat.

Kata Kunci: *naïve bayes, physical distancing, twitter*

Abstract

Physical distancing is now busy becoming a public conversation as a way for the government to spread the Covid-19 virus which is currently hitting several countries around the world. There are public tweets related to physical distance that is free from various kinds of opinions, not only positive but also negative ones. Therefore, Indonesian people consider Twitter to be more desirable because it is easier for Indonesians to express their opinion. The method used is the *Naive Bayes Classifier* (NBC). The data collected was filtered from the tweets with double data, then after filtering the data were taken as many as 547 tweets. In calculating the sentiment analysis of physical distance in the middle of the Covid-19 pandemic using the NBC method, it gets an accurate result of 50.26%. The purpose of this study, to find and categorize negative or positive opinions, from the discussion of *physical distancing*. So that the implementation of the *Physical Distance* policy can get accurate information to the public.

Keywords: *naïve bayes, physical distancing, twitter*

1 Pendahuluan

Sebagai media komunikasi masyarakat, jejaring sosial semakin populer digunakan. Salah satu jejaring sosial yang cukup populer yaitu twitter. Twitter dapat dimanfaatkan sebagai sarana promosi produk, iklan, kampanye politik maupun sebagai sarana menyampaikan pendapat terkait kritik, saran, isu-isu dan opini-opini publik[1]. Angka penyebaran COVID-19 yang begitu besar membuat beberapa masyarakat menjadi sedikit panik. Pemerintah pun mengeluarkan kebijakan *Physical Distancing* [2][3]. *Physical distancing* kini sedang ramai menjadi perbincangan publik sebagai salah satu cara pemerintah dalam menekan penyebaran virus covid-19 yang sedang melanda beberapa

<http://sistemasi.ftik.unisi.ac.id>

Negara di belahan dunia. Terdapat cuitan-cuitan masyarakat yang terkait *physical distancing* yang memunculkan berbagai macam opini, tidak hanya opini yang positif tapi juga yang negatif [4]. Maka dari itu, Twitter di anggap lebih diminati oleh masyarakat Indonesia dikarenakan twitter dirasa lebih mudah untuk mengungkapkan opininya [5].

Berdasarkan latar belakang yang dikemukakan di atas maka dapat dirumuskan masalah penelitian bahwa, Bagaimana implementasi algoritma *Naive Bayes Classifier* (NBC) dalam mengklasifikasikan data *tweet* pada media sosial twitter apakah data *tweet* tersebut termasuk ke dalam kelas bersentimen positif atau negatif? Tujuan dari penelitian ini, untuk menemukan dan mengkategorikan opini negatif atau positif, dari pembahasan *physical distancing*. Agar penerapan kebijakan *Physical Distancing* bisa sampai tepat informasinya kepada masyarakat. Selain itu agar masalah yang sedang di tinjau lebih terarah dan tercapai sasaran yang telah ditentukan, maka harus diberikan batasan masalah yakni, *tweet* yang akan digunakan yaitu kalimat *tweet* yang berada di lingkup Indonesia [6].

2 Tinjauan Literatur

Penelitian sebelumnya yang pernah dilakukan berkaitan dengan masalah yang dihadapi penulis yaitu dengan judul “Analisis Sentimen Masyarakat pada Media Sosial Twitter Terhadap Partai Politik Peserta Pemilu 2019 Menggunakan Naive Bayes Classifier” penelitian dilakukan oleh Adiaty et al. pada tahun 2019 Pada penelitian ini hasil pengujian dan analisis yang dilakukan oleh penulis dengan akurasi mencapai 78,03% [7].

Penelitian lainnya yaitu “Analisis Sentimen Opini Masyarakat Terhadap Acara Televisi pada Twitter dengan *Retweet Analysis* dan *Naive Bayes Classifier*” yang dilakukan oleh (Berlian et al. pada tahun 2019 dengan hasil pengujian dan analisis mencapai tingkat akurasi 61% [8].

Penelitian selanjutnya yaitu “*Sentiment Analysis of the Indonesian Police Mobile Brigade Corps Based on Twitter Posts Using the SVM and NB Methods*” yang dilakukan oleh Pratama et al. tahun 2019 dengan tingkat akurasi 86,96% [9] dan “*Optimization Sentiments of Analysis from Tweets in myXLCare using Naive Bayes Algorithm and Synthetic Minority Over Sampling Technique Method*” yang di lakukan oleh Saputra et al. pada tahun 2020 yang akurasinya mencapai 86,33% [10]. Ada pula penelitian lain yaitu “Analisis Sentimen Tentang Opini Pilkada DKI 2017 Pada Dokumen Twitter Berbahasa Indonesia Menggunakan *Naive Bayes* dan Pembobotan Emoji” oleh Rossi et al. tahun 2017 dengan tingkat akurasi 68,52% untuk kondisi pembobotan tekstual, 75,93% untuk pembobotan non-tekstual, dan 74,81% [11].

Selain itu ada penelitian lain pula yaitu “Implementasi Metode K-Means dan *Naive Bayes Classifier* untuk Analisis Sentimen Pemilihan Presiden (Pilpres) 2019” yang di lakukan oleh Kurniawan and Susanto di tahun 2019 dengan nilai akurasi rata-rata 93.35% dan *error rate* sebesar 6.66% [12].

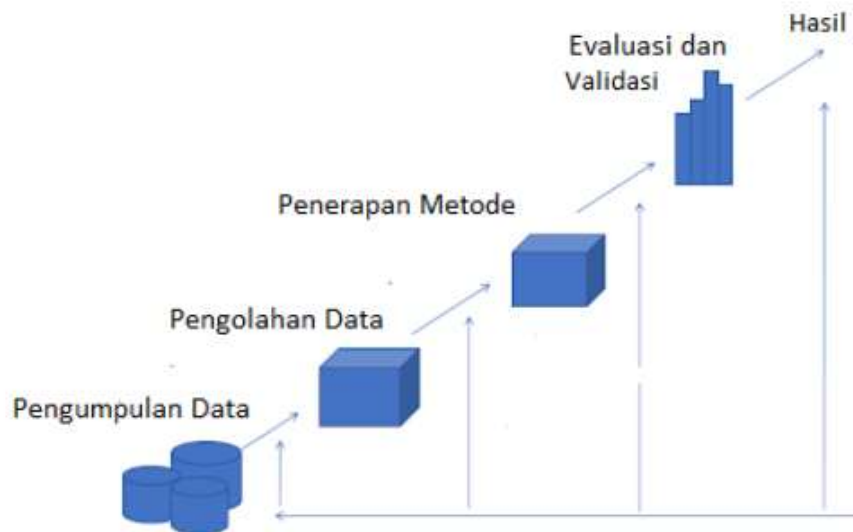
Tujuan dari penelitian ini adalah untuk mengklasifikasikan *tweet* mengenai *physical distancing* di Twitter. Proses klasifikasi dilakukan dengan memisahkan *tweet* positif atau negatif melalui analisis sentimen. Algoritma yang akan di gunakan untuk pengklasifikasian dalam penelitian ini adalah *Naive Bayes Classifier* (NBC). *Naive bayes* merupakan salah satu metode data mining yang dapat digunakan dalam penelitian *text mining* khususnya analisis sentimen, seperti yang diungkap oleh Xhemali et al., pada tahun 2009 Salah satu teknik pembelajaran dari *text mining* untuk analisis sentimen adalah *Naive Bayes Classifier*. Metode *Naive Bayes Classifier* dianggap sebagai metode yang berpotensi baik untuk melakukan klasifikasi data daripada metode klasifikasi lainnya dalam hal akurasi dan komputasi [13].

Dari beberapa penelitian sebelumnya yang dijadikan sebagai referensi, algoritma *Naive Bayes Classification* yang dipakai untuk kasifikasi dalam penelitian *text mining* efektif untuk dipakai. Rata-rata hasil penelitian menghasilkan nilai baik meski ada beberapa yang nilainya masih dibawah standar baik. Dalam penelitian ini, menghasilkan nilai akurasi yang baik dan penelitian selanjutnya bisa menggunakan fitur optimasi untuk lebih meningkatkan nilai akurasi lebih baik lagi.

3 Metode Penelitian

Metode penelitian yang dilakukan untuk menganalisis permasalahan diatas menggunakan teknik Data Mining dengan algoritma *Naive Bayes* seperti gambar dibawah ini :

<http://sistemasi.ftik.unisi.ac.id>



Gambar 1. Alur Diagram Penelitian

Tahapan penelitian yang dilakukan oleh penulis dalam melakukan penelitian ini seperti kegiatan di bawah ini :

3.1. Pengambilan Data

Dataset yang digunakan dalam penelitian ini adalah data mengenai kebijakan pemerintah mengenai *physical distancing* di dapat dari twitter dengan *query* “*physical distancing*” dengan metode *crawling* menggunakan *RapidMiner*.

3.2. Pengolahan Data

Tahap awal ini digunakan untuk mengolah data teks menjadi suatu analisa sentimen. Terdapat beberapa metode yang digunakan pada tahap ini diantaranya:

1. *Tokenization*
Tokenization dapat diartikan sebuah proses untuk memisahkan data teks menjadi beberapa token
2. *Stopwords*
Tahap ini berguna untuk membuang kata yang lebih dari 3 huruf dan termasuk dalam *stopwrds*.
3. *Transform Case*
Digunakan untuk mengubah karakter huruf kapital menjadi huruf kecil.
4. *Filter Token (by length)*
Digunakan untuk pemilihan token dengan ukuran panjang minimal 3 karakter (huruf) meski diantaranya adalah *stopwords*.
5. *Weighting*
Dilakukan berdasarkan jumlah kemunculan kata-kata dalam suatu dokumen, sehingga dokumen dapat direpresentasikan dalam vektor. Bobot fitur yang digunakan adalah *Term Frequency-Inverse Frecuency Document (TF-IDF)*.

3.3. Penerapan Metode

Penerapan metode yang dilakukan menggunakan metode *Naive Bayes Classification*, pada tahap ini data dianalisis, lalu diterapkan model-model yang sesuai dengan jenis data. Model yang diusulkan dalam penelitian ini yaitu *naive bayes*.

3.4. Evaluasi dan Validasi

Validasi merupakan pembuktian dari suatu proses atau metode yang memberikan hasil yang konsisten. Evaluasi kinerja dilakukan untuk menguji hasil klasifikasi dengan mengukur nilai kinerja yang dilakukan oleh sistem. Parameter pengujian yang digunakan untuk mengevaluasi adalah akurasi atau kebenaran tingkat proses klasifikasi yang perhitungannya diperoleh dari table matriks.

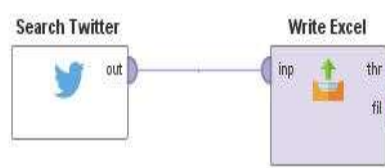
Metode Penelitian memberikan penjelasan tentang langkah-langkah, data, lokasi penelitian, metode evaluasi yang digunakan serta penjelasan terstruktur tentang algoritma atau metode dari penelitian yang dibahas.

4 HASIL DAN PEMBAHASAN

Perhitungan dilakukan dengan menerapkan langkah-langkah yang sudah dituliskan pada bab 3, diantaranya :

4.1. Pengambilan Data

Data yang digunakan dalam penelitian ini merupakan opini berbahasa indonesia mengenai kebijakan *Physical distancing* yang dikeluarkan oleh pemerintah guna menekan penyebaran virus covid-19 yang sedang melanda beberapa negara di belahan dunia. Data diambil dari media sosial twitter dengan prses *Crawling*. *Crawling* data menggunakan aplikasi *RapidMiner* dengan *query Physical distancing* dan diambil sebanyak 1000 data tetapi setelah dilakukan filter dari cuitan tersebut dengan menghapus data yang *double* maka setelah di filter data yang terambil yaitu sebanyak 547 *tweets*.

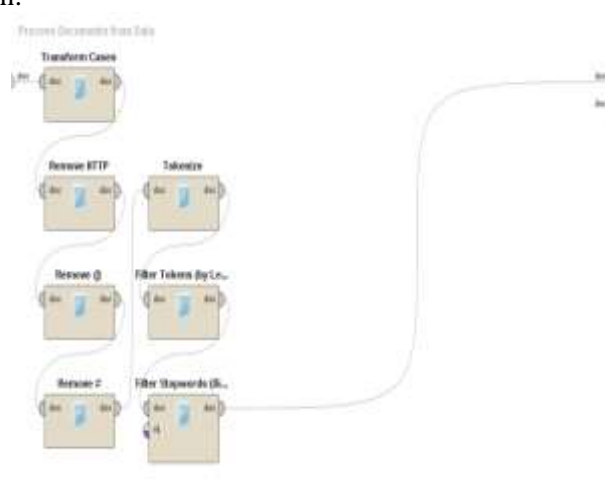


Gambar 2. Proses *Crawling*

Proses *crawling* dilakukan untuk mendapatkan data teks dari twitter secara manual dengan bahasa pemrograman perl.

4.2. Pengolahan Data

Pada tahapan ini dilakukan pengolahan data teks menjadi analisis sentimen [14]. Berikut tahapan *preprocessing* yang dilakukan:



Gambar 3. Proses *Preprocessing* Pada *Rapidminer*

Pada tahapan ini, data yang sudah dikumpulkan, kemudian dilakukan proses *cleansing* dan hanya menyisakan kata-kata saja tanpa ada tanda baca dalam data.

1. *Tokenization*

Tahap *tokenization* dilakukan guna menghilangkan tanda baca, simbol dan karakter yang bukan berupa huruf pada setiap komentar

Tabel 1. Hasil dari Tokenization

Teks sebelum dilakukan proses	Teks setelah dilakukan proses
masih belum paham apa bedanya "social distancing" dan "physical distancing"?	masih belum paham apa bedanya social distancing dan physical distancing
masyarakat harus disiplin dalam menerapkan jaga jarak sosial (physical / social distancing), dan menjaga kebersihan tubuh.	masyarakat harus disiplin dalam menerapkan jaga jarak social physical social distancing), dan menjaga kebersihan tubuh.

Dalam proses Tokenization, tanda baca yang terdapat dalam kalimat akan dihapus.

2. Stopwords

Pada tahap ini akan menyempurnakan tahap filter token. Kata-kata yang dibuang yaitu kata-kata yang tidak relevan atau tidak memiliki makna. Dalam proses stopwords, juga dilakukan proses *stemming*, yakni dengan mengubah kata hubung menjadi kata dasar dalam bahasa Indonesia.

3. Transform Case

Rapidminer akan mengubah kapitalisasi karakter (huruf) menjadi kecil untuk semua kata atau huruf pada tahap ini. Kemudian seluruh isi komentar akan menjadi non-kapital [15].

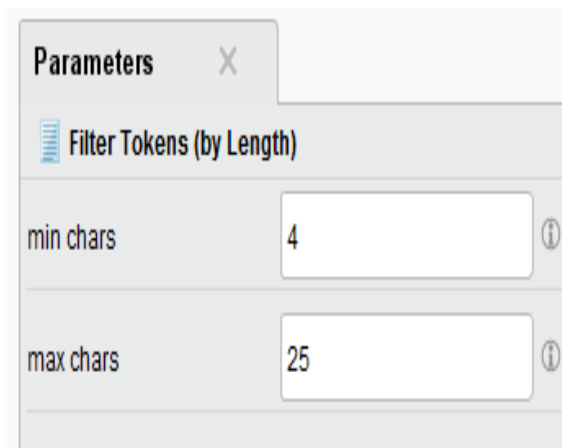
Tabel 2. Hasil dari Transform Case

Teks sebelum dilakukan proses	Teks setelah dilakukan proses
Masih belum paham apa bedanya "social distancing" dan "physical distancing"?	masih belum paham apa bedanya "social distancing" dan "physical distancing"?
Masyarakat harus disiplin dalam menerapkan jaga jarak sosial (physical / social distancing), dan menjaga kebersihan tubuh.	masyarakat harus disiplin dalam menerapkan jaga jarak sosial (physical / social distancing), dan menjaga kebersihan tubuh.

4. Filter Token (by length)

Pada tahap ini dilakukan pemilihan token dengan ukuran panjang minimal 4 karakter (huruf).

<http://sistemasi.ftik.unisi.ac.id>

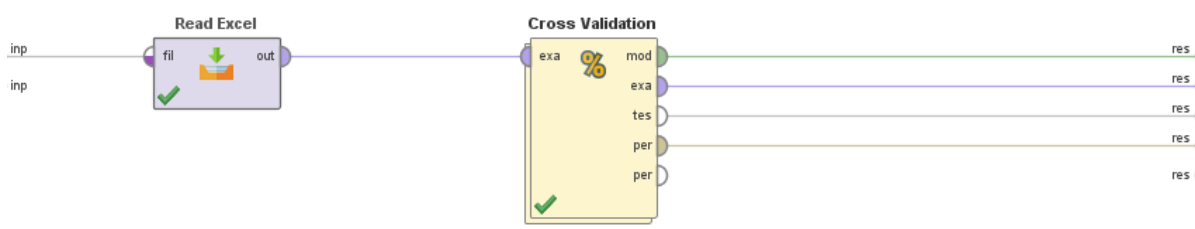


Gambar 4. Parameter dari Filter Token (by length)

Pada gambar diatas, dijelaskan bahwa jika kata yang kurang dari 4 huruf namun termasuk dalam *stopwords* akan tetap dibuang.

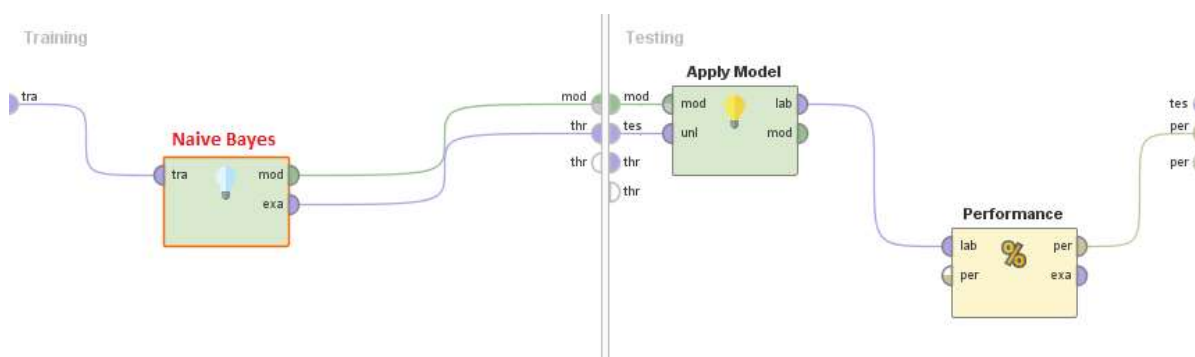
4.3. Penerapan Metode

Proses perhitungan untuk menentukan akurasi menggunakan Model *Naive Bayes* terdapat proses *Training*, *Learning* dan *Testing*. Proses ini merupakan proses dengan memberikan pola data terhadap mesin. Nantinya mesin akan mempelajari pola data tersebut sehingga hasilnya akan dijadikan sebagai rujukan pola baru yang akurat dalam pemodelan hasil. Dalam tahap ini urutan – urutan pengujian yang dilakukan sama hal nya dengan pengujian model lainnya pada Rapid Miner.



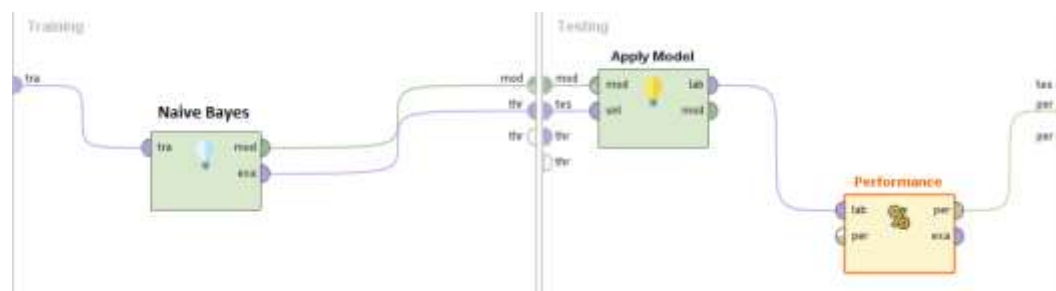
Gambar 5. Operator Cross Validation K-Fold

Pada tahapan ini, mengatur proses validasi dengan *number of folds* 10.



Gambar 6. Proses Insert Training

Proses selanjutnya yaitu *insert* model yang akan di *training*, dalam hal ini memasukkan algoritma yang dipilih yakni *Naive Bayes* dengan menggunakan dataset dengan *cycles* 200, *learning rate* 0,01 dan *momentum* 0,9.



Gambar 7. Proses Insert Apply dan Performance

Proses terakhir yaitu proses testing, dimana memasukkan *apply model*, *performance* dan AUC. Pada proses ini, data dibagi menjadi data training sebesar 80% dan data tes sebesar 20%.

4.4. Validasi dan Evaluasi

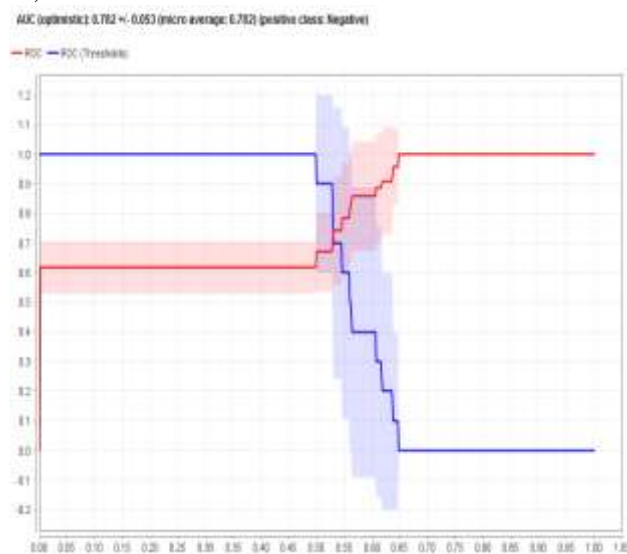
Pada tahap ini proses evaluasi menggunakan metode *confusion matrix*. Model *confusion matrix* akan membentuk matriks yang terdiri dari *accuracy*, *true positif* dan *true negatif*, *recall* serta *precision*. Berikut merupakan hasil dari *confusion matrix* pada algoritma *naïve bayes classifier*.

accuracy: 50.26% +/- 4.25% (micro average: 50.27%)

	true Positive	true Negative	class precision
pred Positive	144	81	64.00%
pred Negative	181	131	40.68%
class recall	42.99%	61.76%	

Gambar 8. Accuracy Analyst

Hasil *accuracy* pengujian *naïve bayes* untuk analisis sentimen pada media sosial instagram memperoleh hasil sebesar 50,26%.



Gambar 9. Nilai ROC Curve

Hasil dari Uji AUC sebesar 0.782 dan garis yang dibentuk semakin melebar dan tidak mengerucut. Hal tersebut menunjukkan hasil akurasi yang diperoleh masuk dalam kategori baik.

5 Kesimpulan

Berdasarkan hasil analisa yang dilakukan terkait sentimen terhadap kebijakan *physical distancing* dengan menggunakan metode *naïve bayes* dapat diambil beberapa kesimpulan bahwa Metode *naïve bayes* merupakan metode yang cukup baik dalam pengklasifikasian data mining ataupun *text mining*. hal ini dikarenakan algoritma tersebut dapat menghasilkan nilai akurasi yang cukup tinggi yakni diatas lima puluh persen.

Referensi

- [1] A. Rossi, T. Lestari, R. Setya Perdana, and M. A. Fauzi, "Analisis Sentimen Tentang Opini Pilkada Dki 2017 pada Dokumen Twitter Berbahasa Indonesia Menggunakan Naïve Bayes dan Pembobotan Emoji," *J. Pengemb. Teknol. Inf. Dan Ilmu Komput.*, Vol. 1, No. 12, pp. 1718–1724, 2017, [Online]. Available: <http://J-Ptiik.Ub.Ac.Id>.
- [2] I. Indriati, M. Marji, and S. Pakpahan, "Analisis Sentimen Tentang Opini Performa Klub Sepak Bola pada Dokumen Twitter Menggunakan Support Vector Machine, dengan perbaikan Kata Tidak Baku," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, 2019.
- [3] I. Hanafi and R. Ferdiansyah, "Sentimen Analisis Twitter Pada Hastag PSBB Di Indonesia dalam Menghadapi Pandemi Covid-19," Vol. 1, No. 3, pp. 1011–1022, 2020.
- [4] B. U. Manalu, "Analisis Sentimen Pada Twitter Menggunakan Text Mining Skripsi Boy Utomo Manalu," *Teknol. Inf. Fak. Ilk. Unsut*, 2014.
- [5] A. Syakuro, "Pada Media Sosial Menggunakan Metode Naïve Bayes Classifier (NBC) Dengan Seleksi Fitur Information Gain (IG) Halaman Judul Skripsi Oleh : Abdan Syakuro," *Anal. Sentimen Masy. Terhadap E-Commerce Pada Media Sos. Menggunakan Metod. Naive Bayes Classif. Dengan Sel. Fitur Inf. Gain*, 2017.
- [6] A. V. Sudiantoro and E. Zuliarso, "Analisis Sentimen Twitter Menggunakan Text Mining dengan Algoritma Naïve Bayes Classifier," *Pros. Sintak 2018*, pp. 398–401, 2018.
- [7] A. R. Adiati, A. Herdiani, W. Astuti, and M. Kom, "Analisis Sentimen Masyarakat pada Media Sosial Twitter terhadap Partai Politik Peserta Pemilu 2019 Menggunakan Naive Bayes Classifier," Vol. 6, No. 2, 2019.
- [8] T. F. Berlian, A. Herdiani, W. Astuti, and M. Kom, "Analisis Sentimen Opini Masyarakat terhadap Acara Televisi pada Twitter dengan Retweet Analysis dan Naive Bayes Classifier," Vol. 6, No. 2, pp. 8660–8669, 2019.
- [9] B. Pratama, D. D. Saputra, D. Novianti, and E. P. Purnamasari, "Sentiment Analysis Of The Indonesian Police Mobile Brigade Corps Based on Twitter Posts using he SVM and NB Methods Sentiment Analysis Of the Indonesian Police Mobile Brigade Corps Based On Twitter Posts Using The SVM and Nb Methods," 2019, doi: 10.1088/1742-6596/1201/1/012038.
- [10] D. D. Saputra, W. Gata, N. K. Wardhani, and K. Sakh, "Optimization Sentiments Of Analysis From Tweets In Myxlcare Using Naïve Bayes Algorithm and Synthetic Minority Over Sampling Technique Method Optimization Sentiments of Analysis From Tweets In Myxlcare Using Naïve Bayes Algorithm and Synthetic Minority Ov," 2020, doi: 10.1088/1742-6596/1471/1/012014.
- [11] Y. S. Mahardhika and E. Zuliarso, "Analisis Sentimen Terhadap Pemerintahan Joko Widodo Pada Media Sosial Twitter Menggunakan Algoritma Naives Bayes," *Pros. Sintak 2018*, 2018.
- [12] I. Kurniawan and A. Susanto, "Implementasi Metode K-Means dan Naïve Bayes Classifier Untuk Analisis Sentimen Pemilihan Presiden (Pilpres) 2019," *Eksplora Inform.*, Vol. 9, No. 1, pp. 1–10, 2019, doi: 10.30864/Eksplora.V9i1.237.
- [13] D. Xhemali, C. J. Hinde, and R. G. Stone, "Naïve Bayes vs Decision Trees vs Neural Networks in The Classification of Training Web Pages," Vol. 4, No. 1, pp. 16–23, 2009.
- [14] T. G. Laksana and P. Rizkiyah, "Analisa Sentimen Menggunakan Naïve Bayes untuk Mengetahui Presentase Komentar pada Aplikasi Go-Jek," 2018.
- [15] E. Junianto and D. Riana, "Penerapan Pso Untuk Seleksi Fitur Pada Klasifikasi Dokumen Berita Menggunakan NBC," Vol. 4, No. 1. pp. 38–45, 2017.