ISSN: 2962-3545
Prosiding SAINTEK: Sains dan Teknologi Vol.1 No.1 Tahun 2022
Call for papers dan Seminar Nasional Sains dan Teknologi Ke-1 2022
Fakultas Teknik, Universitas Pelita Bangsa, Juli 2022

Komparasi Algoritma *Naïve Bayes* Dan *K-Nearest Neighbor* Dalam Melihat Analisis Sentimen Terhadap Vaksinasi Covid-19

Comparison of Nave Bayes and K-Nearest Neighbor Algorithm in Viewing Analysis of Sentiment on Covid-19 Vaccination

A.Yudi Permana[1], Hendri Noviyani[2]
[1,2]Teknik Informatika, Fakultas Teknik, Universitas Pelita Bangsa
[1]yudi@pelitabangsa.ac.id, [2]hendrinoviyani@gmail.com*

Abstract

Twitter is often used to deliver messages in the form of public opinion or opinion about the topic that is being reported. The government's policy to vaccinate has received various comments, ranging from praise, criticism, suggestions, and even hate speech. With so many twitter users who express their opinion, it can be used to find information. However, its use requires proper analysis, so that the resulting information can help many parties in making decisions or choices. Therefore, in this study, we tried to analyze sentiment on Covid-19 vaccination using the Naïve Bayes and K-Nearest Neighbor algorithms using the Cross Validation technique. The purpose of this study is to find out whether the Naïve Bayes and K-Nearest Neighbor algorithms in classifying produce optimal accuracy, to determine the sentiments of twitter users towards the Covid-19 vaccination and how much influence preprocessing has to measure accuracy on the classification. Based on the research that has been carried out, it can be concluded that the application of preprocessing for sentiment analysis on Covid-19 vaccination using the Naïve Bayes and K-Nearest Neighbor algorithms accompanied by the use of the Cross Validation technique got quite good results. The Naïve Bayes algorithm produces an accuracy of 77.62% and the K-Nearest Neighbor algorithm produces an accuracy of 76.43. Then for the positive response rate of the community to the Covid-19 vaccination, it was 55.63%.

Keywords: *Comparison, Naïve Bayes, K-Nearest Neighbor, Sentiment Analysis,Vaccination, RapidMiner*

Abstrak

Twitter seringkali digunakan untuk penyampaian pesan berupa pendapat atau opini masyarakat tentang bahasan yang sedang ramai diberitakan. Kebijakan pemerintah untuk melakukan vaksinasi menuai berbagai macam komentar, mulai dari pujian, kritik, saran, bahkan ujaran kebencian. Dengan banyaknya pengguna twitter yang menyampaikan pendapat tersebut dapat dimanfaatkan untuk mencari sebuah informasi. Namun dalam pemanfaatannya membutuhkan analisis yang tepat, sehingga informasi yang dihasilkan dapat membantu banyak pihak dalam menentukan keputusan atau pilihan. Oleh sebab itu, pada penelitian ini mencoba melakukan analisis sentimen terhadap vaksinasi Covid-19 menggunakan algoritma Naïve Bayes dan K-Nearest Neighbor dengan penggunaan teknik Cross Validation. Tujuan penelitian ini adalah Untuk mengetahui apakah algoritma Naïve Bayes dan K-Nearest Neighbor dalam mengklasifikasikan menghasilkan akurasi yang optimal, untuk mengetahui sentimen pengguna twitter terhadap vaksinasi Covid- 19 dan seberapa besar pengaruh preprocessing untuk mengukur akurasi terhadap klasifikasinya. Berdasarkan penelitian yang telah dilakukan maka dapat ditarik kesimpulan bahwa penerapan preprocessing untuk analisis sentimen vaksinasi Covid-19 menggunakan algoritma Naïve Bayes dan K-Nearest Neighbor disertai dengan penggunaan teknik Cross Validation mendapatkan hasil yang cukup baik. Untuk algoritma Naïve Bayes menghasilkan akurasi sebesar 77,62% dan untuk algoritma K-Nearest Neighbor menghasilkan akurasi sebesar 76,43. Kemudian untuk tingkat respon positif masyarakat terhadap vaksinasi Covid-19 sebesar 55,63%.

Kata kunci: Komparasi, Naïve Bayes, K-Nearest Neighbor, Analisis Sentimen, Vaksinasi, RapidMiner

Pendahuluan

Pada bulan Desember 2019 dunia dihebohkan dengan munculnya Covid-19 di Wuhan, Ibu Kota Provinsi Hubei Tiongkok, China. Kemudian masuk Indonesia pada bulan Maret 2020 di Depok yaitu dua WNI yang merupakan seorang ibu (64 tahun) dan putrinya (31 tahun)[1]. Berdasarkan data yang diambil dari World Health Organization (WHO) per tanggal 5 Juni 2021 tercatat sebanyak 173.357.945 kasus, 156.121.170 sembuh dan 3.728.668 meninggal. Sedangkan di Indonesia sebanyak 1.850.206 kasus, 1.701.784 sembuh dan 51.449 meninggal. Dengan banyaknya kasus yang terkena covid-19 ini tentu sangat mempengaruhi sektor kehidupan, dari ekonomi, pendidikan, politik, dan kehidupan social[2].

Untuk menangani penyebaran virus Covid-19 ini pemerintah melakukan tindakan vaksinasi, supaya dampak negatif yang ditimbulkan dapat dikendalikan. Informasi vaksinasi serta tata cara pencegahan virus ini telah tersebar dimedia sosial. Media sosial merupakan salah satu sumber yang sangat umum digunakan untuk berkomunikasi, berbagai dokumen serta data dengan jumlah komunitas yang sangat besar. Salah satu media sosial yang sering digunakan masyarakat dalam berkomunikasi atau sekedar mengutarakan opininya adalah twitter[3].

Twitter seringkali digunakan untuk penyampaian pesan berupa pendapat atau opini masyarakat tentang bahasan yang sedang ramai diberitakan[4]. Kebijakan pemerintah untuk melakukan vaksinasi menuai berbagai macam komentar, mulai dari pujian, kritik, saran, bahkan ujaran kebencian. Dengan banyaknya pengguna twitter yang menyampaikan pendapat tersebut dapat dimanfaatkan untuk mencari sebuah informasi[5]. Namun dalam pemanfaatannya membutuhkan analisis yang tepat, sehingga informasi yang dihasilkan dapat membantu digunakan untuk menganalisis pendapat adalah analisis sentimen.

Dalam penelitian ini analisis sentimen dilakukan untuk mengetahui kecenderungan opini terhadap masalah mengandung sentimen positif atau negatif dengan menggunakan algoritma Naïve Bayes dan K-Nearest Neighbor (KNN). Tahapan penelitian ini terdiri dari pengambilan data mentah (crawling), prapemrosesan data (preprocessing), klasifikasi data menggunakan algoritma Naïve Bayes dan KNN, dan evaluasi menggunakan Confusion Matrix. Untuk menunjang keberhasilan penelitian ini, maka peneliti menggunakan tools RapidMiner sebagai alat bantu dan diharapkan hasil akhir dari penelitian ini mendapat nilai akurasi yang tinggi serta dapat menjadi pertimbangan pemerintah dalam mengambil kebijakan terhadap vaksinasi Covid-19.

Berdasarkan uraian diatas, penulis ingin melakukan penelitian dengan judul "Komparasi Algoritma Naïve Bayes Dan K-Nearest Neighbor Dalam Melihat Analisis Sentimen Terhadap Vaksinasi Covid-19".

Penelitian dilakukan oleh Muhammad Syarifuddin pada tahun 2020 berfokus pada perbandingan hasil klasifikasi metode Naïve Bayes dan KNN, serta mengetahui kecenderungan masyarakat di twitter. Tahapan dalam penelitian ini adalah pengumpulan data, pengolahan data awal, dan metode yang digunakan[6].

Text Mining

Text mining adalah proses mengeksplorasi dan menganalisis sejumlah besar data teks tidak terstruktur yang dibantu oleh perangkat lunak yang dapat mengidentifikasi konsep, pola, topik, kata kunci, dan meskipun beberapa orang menarik perbedaan antara dua istilah. Dalam pandangan itu, analitik teks adalah aplikasi yang diaktifkan oleh penggunaan teknik text mining untuk memilah set data [7].

Sentimen Analisis

Analisis sentimen adalah cabang text mining yang bertujuan untuk memperjelas ulasan kedalam kelas tertentu. Review bisa diklasifikasikan menjadi kelas positif atau negatif. [8].

Preprocessing

Tahap Preprocessing diperlukan untuk membersihkan data dari teks yang tidak diperlukan, dimana data teks yang tidak terstruktur akan diubah menjadi data teks yang terstruktur atau semi terstruktur. Tahap dari preprocessing untuk mengolah data yaitu cleansing, case folding, tokenizing, stopword removal dan stemming[9].

TF IDF

Matrik ini sering digunakan oleh pencarian online mesin untuk mengambil dokumen yang paling relevan sesuai permintaan pengguna[10][2].

Cross Validation

Cross Validation merupakan teknik validasi dari pengembangan Split Validation dimana validasinya mengukur training error dari data uji[9].

Tinjauan Studi Naïve Bayes

Naïve Bayes merupakan metode klasifikasi dalam penambangan teks yang digunakan dalam analisis sentimen. Metode ini berpotensi baik dalam klasifikasi presisi dan komputasi data[6]. Naïve Bayes adalah suatu algoritma yang dapat mengklasifikasikan suatu variabel tertentu dengan menggunakan metode probabilitas dan statistik[11]. Keuntungan penggunaan Naïve Bayes adalah bahwa metode ini hanya membutuhkan jumlah data pelatihan yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian dan dapat bekerja jauh lebih baik dalam situasi dunia nyata yang kompleks[12].

Tahapan analisis sentimen menggunakan algoritma Naïve Bayes dapat dilakukan dengan mencari probabilitas P(H|X), P(c) dan P(wi|c) dapat ditulis dengan rumus sebagai berikut[13][14][15]:

1. Perhitungan Posterior Probability

$$P(H|X) = \frac{P(X|H.\ P(H)}{P(X)}$$

Keterangan :

X = Data dengan kelas yang belum diketahui
H = Hipotesis data merupakan suatu kelas spesifikasi
P(H|X) = Probabilitas hipotesis H berdasarkan kondisi X
P(X|H) = Probilitas X berdasarkan kondisi pada hipotesis H (*likelihood*)
P(H) = Probabilitas hipotesis H (*prior*)
P(X) = Probabilitas X

2. Perhitungan prior pada masing-masing kelas

$$P(c) = NcNdoc$$

Keterangan :

P(c) = *Prior Probability*
Nc = Jumlah data *training* dalam sebuah dokumen dengan kelas
Ndoc = Jumlah total dokumen dalam data *training*

3. Perhitungan *likelihood* pada masing-masing kata dalam kelas

$$P(wi|c) = count\ (Wi,c)+1\Sigma weVcount(w,c)+|V|$$

Keterangan :

Wi. = kata pada dokumen pada setiap kategori atau kelas
c = Kelas dokumen
V| = gabungan disemua jenis kelas. K-Nearest Neighbor

K-Nearest Neighbor merupakan sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Data pembelajaran diproyeksikan ke ruang berdimensi banyak, dimana masing-masing dimensi merepresentasikan fitur dari data. Ruang ini dibagi menjadi bagian-bagian berdasarkan klasifikasi data pembelajaran. Nilai k yang terbaik untuk

algoritma ini tergantung pada data, secara umum nilai k yang tinggi akan mengurangi efek noise pada klasifikasi, tetapi membuat batasan antara setiap klasifikasi menjadi lebih kabur[16]. Penerapan teknik K-Nearest Neighbor pernah diuji dan mampu mengatasi beberapa masalah yang berkaitan dengan algoritma lain[14].

Salah satu jarak yang paling umum digunakan untuk mengukur kemiripan k terdekat tetangga adalah Euclideant Distance, biasanya dikenal sebagai distence. Ruang Euclideant berdimensi -n adalah ruang yang titik-titiknya vektor dari n bilangan real. Pada Euclideant kita kuadratkan jarak disetiap dimensi, jumlahkan kuadratnya, dan ambil akar kuadrat positif yang didefinisikan sebagai berikut[17]:

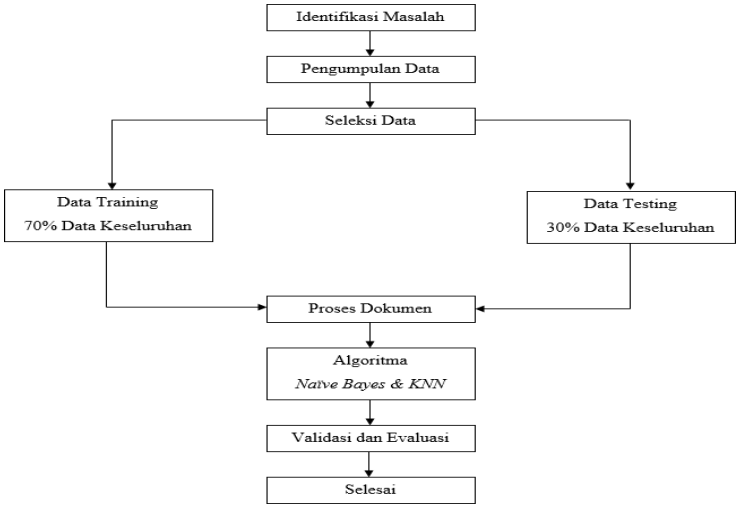
distance (x,y) $= \sqrt{\Sigma(xi - yi)2}$

Keterangan :

distance (x,y) = jarak antara data testing dan data training
xi = data testing ke-i
yi = data training ke-i

Metode Penelitian

Dalam hal ini penulis melakukan penelitian tentang analisis sentimen terhadap vaksinasi dimasa pandemi Covid-19 menggunakan algoritma Naïve Bayes dan K- Nearest Neighbor. Pemilihan objek ini didasari karena sekarang ini sedang ramai masyarakat diminta melakukan vaksinasi agar terhindar dari virus Covid-19, namun dalam pelaksanaanya tidak sedikit masyarakat yang menolak untuk di vaksin, sehingga menuai berbagai macam opini baik itu opini positif maupun opini negatif. Berikut adalah tahapan dalam penelitian ini :



Gambar 1 Kerangka Berfikir

Berikut penjelasannya:

1. Tahap pertama yaitu identifikasi masalah, pada tahap ini proses yang dilakukan adalah menggali permasalahan yang ditemukan pada objek yang akan diteliti.
2. Tahap kedua yaitu pengumpulan data, pada tahap ini pengumpulan data dilakukan dengan menggunakan crawling data twitter. Data yang dikumpulkan merupakan tweet yang berbahasa Indonesia dengan query vaksinasi.
3. Tahap ketiga yaitu seleksi data, pada tahap ini seleksi data dilakukan agar data yang digunakan lebih mudah untuk dipahami.
4. Tahap keempat yaitu pembagian data, untuk melakukan proses uji diperlukan pembagian data yang dibagi menjadi data training dan data testing. Data yang digunakan dalam penelitian ini sebanyak 600 data dan nantinya akan dibagi menjadi 70% untuk data training dan 30% untuk data testing. Untuk data masing-masing sebanyak 420 data training dan 180 data testing.

5. Tahap kelima yaitu proses dokumen, pada tahapan ini proses dokumen dilakukan dengan cara preprocessing yaitu dengan cara case folding,, tokenizing, stopword removal, dan stemming.
6. Tahapan keenam yaitu penerapan algoritma, pada penelitian ini algoritma yang digunakan adalah algoritma Naïve Bayes dan K-Nearest Neighbor.
7. Tahap ketujuh yaitu validasi dan evaluasi, pada penelitian ini proses validasi menggunakan teknik Cross Validation dan evaluasi menggunakan Confusion Matrix.

Hasil dan Pembahasan

Pengujian perhitungan manual dilakukan terhadap 5 data dimana dua data berlabel positif, dua data berlabel negatif dan satu data yang belum diketahui labelnya. Berikut adalah kelima data tersebut :

Tabel 1 Data Pengujian Manual

Kode	Sebelum Preprocessing	Setelah Preprocessing	Label
D1	Vaksin Halal Danaman Untuk Masyarakat	Vaksin Halal Aman Rakyat	Positive
D2	Jaga Kesehatandengan Vaksinasi	Jaga Sehatvaksin	Positive
D3	Vaksin Mengandung Babi Dan Chip	Vaksin Kandung Babi Chip	Negative
D4	Kenapa Orang Seperti Itu tidak diVaksin Mati Saja	Orang Vaksin Mati	Negative
D5	Vaksin Aman Masyarakat Sehat	Vaksin Aman Rakyat Sehat	?

Berdasarkan tabel 1, D1 sampai D4 merupakan data training yang sudah diketahui labelnya dan D5 merupakan data testing yang belum diketahui labelnya. Untuk mengetahui label pada D5 dilakukan pembobotan TF-IDF terlebih dahulu dengan proses sebagai berikut:

Tabel 2 Hasil Perhitungan IDF

Kata	TF					DF	IDF
	D1	D2	D3	D4	D5		
Vaksin	1	1	1	1	1	5	0
Halal	1					1	0,699
Aman	1				1	2	0,398
rakyat	1				1	2	0,398
Jaga		1				1	0,699
Sehat		1			1	2	0,398
kandung			1			1	0,699
Babi			1			1	0,699
Chip			1			1	0,699
Orang				1		1	0,699
Mati				1		1	0,699

Tabel 3 Hasil Perhitungan TF-IDF

Kata	TF*IDF				
	D1	D2	D3	D4	D5
vaksin	0	0	0	0	0
halal	0,699	0	0	0	0
aman	0,398	0	0	0	0,398
rakyat	0,398	0	0	0	0,398
jaga	0	0,699	0	0	0
Sehat	0	0,398	0	0	0,398
kandung	0	0	0,699	0	0
Babi	0	0	0,699	0	0
Chip	0	0	0,699	0	0
orang	0	0	0	0,699	0
Mati	0	0	0	0,699	0
Jumlah	1,495	1,097	2,097	1,398	1,194

Perbandingan akurasi Naïve Bayes dan K-Nearest Neighbor

Tabel 4 Tabel Perbandingan Akurasi

Nilai *K-Folds*	*Naïve Bayes* Accuracy	*K-Nearest Neighbor* Accuracy
2	77,62%	75,71%
3	75,00%	75,24%
4	74,52%	75,00%
5	74,52%	75,24%
6	75,48%	76,19%
7	72,38%	76,43%
8	74,02%	75,70%
9	72,60%	76,18%
10	75,71%	75,48%

Dari table di atas akurasi terbaik perbandingan 2 algoritma yiatu naïve bays dan KNN, terbaik berada di percobaan K-Fold 10 untuk niave bayes dengan akurasi 75,71%, dan percobaan K-fold 7 untuk KNN dengan akurasi 76,43%.

Gambar 2 Grafik Perbandingan Akurasi

Kesimpulan

Berdasarkan penelitian yang telah dilakukan maka dapat ditarik kesimpulan bahwa penerapan preprocessing untuk analisis sentimen vaksinasi Covid-19 menggunakan algoritma Naïve Bayes dan K-Nearest Neighbor disertai dengan penggunaan teknik Cross Validation mendapatkan hasil yang cukup baik. Untuk algoritma Naïve Bayes menghasilkan akurasi terbaik sebesar 77,62% dan untuk algoritma K-Nearest Neighbor menghasilkan akurasi terbaik sebesar 76,43. Kemudian untuk tingkat respon positif masyarakat terhadap vaksinasi Covid-19 sebesar 55,63%.

Daftar Rujukan

[1] Jaiz, M., Framanik, N. A., Winangsih, R., Widyastuti, N. W., & Kurniawati, R. N. K., (2021), Model Advokasi Untirta Dalam Menangani Virus Covid-19 Di Untirta Wilayah Kampus 1 (Pakupatan), Kampus 2 (Cilegon), Kampus 3 (Ciwaru), Kampus 4 (Kepandean), Dan Kampus 5 (Sindangsari).

[2] Joyosemito, I. S., & Nasir, N. M., (2021), Gelombang kedua pandemi menuju endemi covid-19: Analisis kebijakan vaksinasi dan pembatasan kegiatan masyarakat di Indonesia. *Jurnal Sains Teknologi dalam Pemberdayaan Masyarakat*, 2(1).

[3] Wahidah, I., Athallah, R., Hartono, N. F. S., Rafqie, M. C. A., & Septiadi, M. A., (2020), Pandemik COVID-19: Analisis perencanaan pemerintah dan masyarakat dalam berbagai upaya pencegahan. *Jurnal Manajemen Dan Organisasi,* 11(3), 179-188.

[4] Mentari, N. D., Fauzi, M. A., & Muflikhah, L., (2018), Analisis Sentimen Kurikulum 2013 Pada Sosial Media Twitter Menggunakan Metode K-Nearest Neighbor dan Feature Selection Query Expansion Ranking. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN*, 2548, 964X.

[5] Husna, A. N., & Faizah, R., (2021), Memberdayakan Masyarakat Digital.

[6] M. Syarifuddin, (2020), "Analisis Sentimen Opini Publik Mengenai Covid-19 Pada Twitter Menggunakan Metode Naïve Bayes Dan Knn," *Inti Nusa Mandiri*, vol. 15, no. 1.

[7] M. M. Mala Olhang, S. Achmadi, and F. . A. Wibisono, (2020), "Analisis Sentimen Pengguna Twitter Terhadap Covid-19 Di Indonesia Menggunakan Metode Naive Bayes Classifier (Nbc)" *JATI (Jurnal Mhs. Tek. Inform.*, vol. 4, no. 2, doi: 10.36040/jati.v4i2.2695.

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

[9] T. A. Mutiara, A. Yuris, A. M. Nissa, and G. Windu, (2020), "Analisis Sentimen Opini Publik Mengenai Larangan Mudik pada Twitter Menggunakan Naive Bayes," *Jurnal CoreIT*, vol. 6, no. 2.

[10] V. Amrizal, (2018), "Penerapan Metode Term Frequency Inverse Document Frequency (Tf-Idf) Dan Cosine Similarity Pada Sistem Temu Kembali Informasi Untuk Mengetahui Syarah Hadits Berbasis Web (Studi Kasus: Hadits Shahih Bukhari-Muslim)" *J. Tek. Inform.*, vol. 11, no. 2, doi: 10.15408/jti.v11i2.8623.

[11] S. Samsir, A. Ambiyar, U. Verawardina, F. Edi, and R. Watrianthos, (2021), "Analisis Sentimen Pembelajaran Daring Pada Twitter di Masa Pandemi COVID-19 Menggunakan Metode Naïve Bayes" *J. MEDIA Inform. BUDIDARMA*, vol. 5, no. 1, doi: 10.30865/mib.v5i1.2580.

[12] Y. I. Kurniawan, (2018), "Perbandingan Algoritma Naive Bayes dan C.45 dalam Klasifikasi Data Mining,"*J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 4, doi: 10.25126/jtiik.201854803.

[13] Wiyanto, W., Ngudi, T., & Saefulloh, A. (2020). ANALISA TINGKAT KEPUASAN PELANGGAN TERHADAP PELAYANAN PERUSAHAAN OTOBUS XYZ MENGGUNAKAN METODE NAÃ VE BAYES. *Pelita Teknologi*, *15*(1), 56-67.

[14] Romli, I., Prameswari, S., & Kamalia, A. Z. (2021). Sentiment Analysis about Large-Scale Social Restrictions in Social Media Twitter Using Algoritm K-Nearest Neighbor. *Jurnal Online Informatika*, *6*(1), 96-102.

[15] Nugroho, A., & Religia, Y. (2021). Analisis Optimasi Algoritma Klasifikasi Naive Bayes menggunakan Genetic Algorithm dan Bagging. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *5*(3), 504-510.

[16] T. Imandasari, E. Irawan, A. P. Windarto, and A. Wanto, (2019), "Algoritma Naive Bayes Dalam Klasifikasi Lokasi Pembangunan Sumber Air," *Pros. Semin. Nas. Ris. Inf. Sci.*, vol. 1, doi: 10.30645/senaris.v1i0.81.

[17] N. Tri Romadloni, I. Santoso, and S. Budilaksono, (2019), "Perbandingan Metode Naive Bayes, KNN, dan Decision Tree Terhadap Analisis Sentimen Transportasi KRL Commuter Line," *J. IKRA-ITH Inform.*, vol. 3, no. 2.

[18] F. Sodik and I. Kharisudin, (2021), "Analisis Sentimen dengan SVM , NAIVE BAYES dan KNN untuk Studi Tanggapan Masyarakat Indonesia Terhadap Pandemi Covid-19 pada Media Sosial Twitter," *Prisma*, vol. 4.