The 2st Seminar Nasional dan Prosiding Scitech 2023
Fakultas Sains dan Teknologi
Jepara, 22 Juni 2023
Avalaible Online at: https://conference.unisnu.ac.id/scitech/semprotek23
ISSN 2830-0319 (e)

ANALISIS SENTIMEN TERHADAP KULIAH DARING MENGGUNAKAN *MULTINOMIAL NAIVE BAYES* DENGAN SELEKSI FITUR *INFORMATION GAIN*

SENTIMENT ANALYSIS OF ONLINE LECTURES USING MULTINOMIAL NAIVE BAYES WITH INFORMATION GAIN FEATURE SELECTION

Bayu Restu Adji[1*], J.B. Budi Darmawan[2]

[1,2]Universitas Sanata Dharma Yogyakarta
Email : [1*]bayurestua@gmail.com, [2]b.darmawan@usd.ac.id

Abstrak - Pandemi virus corona yang terjadi pada tahun 2020 menyebabkan kegiatan perkuliahan dilakukan secara daring untuk mencegah penyebaran virus corona. Perkuliahan yang dilakukan secara daring mendapat banyak opini dari masyarakat maupun mahasiswa. Banyaknya opini terkait perkuliahan daring dapat dilakukan analisis sentimen untuk mengetahui opini apa yang banyak disampaikan masyarakat dan mahasiswa. Proses analisis sentimen dilakukan menggunakan metode *Multinomial Naïve Bayes*. Data yang digunakan untuk proses analisis sentimen sebanyak 4.014 dengan kata kunci "kuliah daring" dan "belajar online". Data akan dilakukan pembersihan data terlebih dahulu melalui proses *preprocessing* kemudian diberi label menggunakan *textblob* yang akan dikategorikan ke dalam kelas positif, negatif, dan netral. Dalam penelitian ini juga akan dilakukan perbandingan hasil akurasi menggunakan seleksi fitur *Information Gain* dengan harapan dapat meningkatkan hasil akurasi. Berdasarkan hasil pengujian yang telah dilakukan seleksi fitur *Information Gain* terbukti dapat meningkatkan akurasi dalam proses analisis sentimen menggunakan *Multinomial Naïve Bayes.* Hasil akurasi tertinggi menggunakan seleksi fitur *Information Gain* sebesar 79.54%. Sedangkan hasil akurasi tertinggi tanpa menggunakan seleksi fitur *Information Gain* sebesar 78.43%.

Kata kunci: Analisis Sentimen; *Information Gain*; *Multinomial Naive Bayes*;

Abstract *- The corona virus pandemic that occurred in 2020 caused lecture activities to be carried out online to prevent the spread of the corona virus. Lectures that are held online receive a lot of opinion from the public and students. The large number of opinions regarding online lectures can be carried out by sentiment analysis to find out what opinions are expressed by the public and students. The process of sentiment analysis is carried out using the Multinomial Naïve Bayes method. The data used for the sentiment analysis process is 4,014 with the keywords "online lectures and online learning". The data will be cleaned first through a preprocessing process and then labeled using a text blob and categorized into positive, negative, and neutral classes. In this study, a comparison of accuracy results will also be carried out using the Information Gain feature selection in the hope of increasing accuracy results. Based on the test results that have been carried out, the Information Gain feature selection is proven to increase accuracy in the sentiment analysis process using Multinomial Naïve Bayes. The highest accuracy results using the Information Gain feature selection of 79.54%. While the highest accuracy results without using the Information Gain feature selection are 78.43%.*

Keywords*: Sentiment Analysis; Information Gain; Multinomial Naive Bayes;*

1. PENDAHULUAN

Coronavirus Disease-19 atau lebih dikenal dengan Covid-19 merupakan jenis penyakit yang menyerang saluran pernapasan dengan gejala seperti demam, batuk, dan sesak napas. Virus ini pertama kali terdeteksi di Wuhan pada Desember 2019 dan dengan cepat menyebar ke berbagai negara termasuk Indonesia. Pemerintah Republik Indonesia berupaya memutus rantai penyebaran covid-19 dengan mengeluarkan beberapa kebijakan. Salah satunya kebijakan yang dikeluarkan Menteri Pendidikan dan Kebudayaan Republik Indonesia yaitu Nadiem Makarim tentang pelaksanaan Pendidikan selama masa darurat penyebaran covid-19 akan dilakukan secara daring [1].

Pemberlakuan kebijakan Menteri Pendidikan dan Kebudayaan membuat masyarakat banyak memberikan komentar dan opini di media sosial, salah satunya pada media sosial Twitter. Komentar maupun opini masyarakat ini dapat dijadikan instrumen dalam mengukur kepuasan masyarakat terkait kuliah daring dengan mengelompokkan ke dalam tiga kelas yaitu positif, negatif, dan netral. Pada penelitian ini pengukuran tingkat kepuasan masyarakat terkait kuliah daring diukur menggunakan analisis sentimen dengan algoritma *Multinomial Naïve Bayes*. Algoritma *Multinomial Naïve Bayes* dipilih untuk proses analisis sentimen karena memiliki beberapa kelebihan antara lain, sederhana, menghasilkan tingkat akurasi yang tinggi, dan waktu komputasi yang rendah[2][3].

Beberapa penelitian terkait analisis sentimen menggunakan algoritma *Multinomial Naïve Bayes* telah di lakukan oleh peneliti sebelumnya seperti penelitian analisis sentimen terhadap pengguna transportasi Grab yang mendapatkan hasil akurasi sebesar 86.57% [3]. Penelitian lain terkait analisis sentimen menggunakan *Multinomial Naïve Bayes* juga pernah dilakukan terhadap ulasan produk *Colearn* pada *Google Play Store* yang mendapatkan hasil akurasi sebesar sebesar 88.89% [4].

Penelitian yang telah dilakukan peneliti terdahulu terkait algoritma *Multinomial Naïve Bayes* pada proses analisis sentimen mendapatkan hasil yang baik. Akan tetapi pada proses klasifikasi teks sering terjadi permasalahan dimensi dan fitur yang berlebihan sehingga ruang pencarian semakin tinggi yang menyebabkan pemprosesan menjadi lama dan dapat menurunkan nilai evaluasi. Permasalahan tersebut dapat diatasi dengan melakukan pemangkasan dimensi yang tidak penting menggunakan seleksi fitur, salah satu seleksi fitur yang dapat di gunakan adalah *Information Gain* [5] [6]. Penelitian menggunakan seleksi fitur *Information Gain* pernah dilakukan peneliti sebelumnya untuk kasus analisis sentimen *Cyberbullying* di Twitter menggunakan algoritma *Support Vector Machine* mendapatkan peningkatan akurasi 6% yang semula sebesar 80% menjadi 86% [7].

Berdasarkan pemaparan diatas, peneliti tertarik untuk mengoptimalkan hasil akurasi menggunakan seleksi fitur *Information Gain* untuk proses analisis sentimen dengan kata kunci "kuliah daring" dan "belajar online" yang terjadi selama pandemi covid-19 menggunakan algoritma *Multinomial Naïve Bayes.*

2. METODE PENELITIAN

Pada penelitian ini menerapkan algoritma *Multinomial Naïve Bayes* dan seleksi fitur *Information Gain* untuk melakukan proses analisis sentimen terkait perkuliahan daring selama masa pandemi covid-19. Gambar alur penelitian secara menyeluruh dapat dilihat pada Gambar 1. Tahap-tahap analisis sentimen pada penelitian ini adalah sebagai berikut :

1) Pengumpulan data

Proses pengumpulan data atau *crawling* data dilakukan melalui *Application Programming Integration* (API) Twitter dengan menggunakan kata kunci "kuliah daring" dan "belajar online". Proses *crawling* data akan diimplementasikan menggunakan bahasa pemrograman *python*.

2) Preprocessing

Pada proses analisis sentimen *preprocessing* adalah tahapan yang digunakan untuk menyiapkan data teks sebelum digunakan untuk proses klasifikasi teks. Pada langkah ini data teks akan diubah menjadi bentuk yang lebih relevan agar menghasilkan informasi teks dengan kualitas yang lebih baik. *Preprocessing* yang dilakukan terdiri dari beberapa tahapan yaitu *case folding, clean text, tokenizing, normalization, stopword removal, dan stemming.*

3) Pelabelan data

Proses pelabelan pada data dilakukan dengan menggunakan *tools Textblob* dari bahasa pemrograman *python*. *Textblob* akan menghitung nilai polaritas dan subjektivitas pada setiap *tweet*. Setelah nilai polaritas diperoleh kemudian akan dikategorikan ke dalam 3 kelas yaitu positif, negatif, dan netral [8]. Saat ini pelabelan data menggunakan *textblob* hanya dapat digunakan untuk teks berbahasa inggris, sehingga sebelum masuk ke tahap pelabelan data sudah dilakukan *translate* terlebih dahulu ke dalam bahasa Inggris.

4) Pembobotan Kata

Pembobotan kata dilakukan untuk mengubah kata pada dataset menjadi bentuk angka. Proses pembobotan akan dilakukan dengan menghitung bobot nilai pada setiap kata berdasarkan frekuensi kemunculan kata atau *term* dalam dokumen menggunakan *TF-IDF* [8] . Untuk menghitung *TF-IDF* dapat menggunakan persamaan (1).

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N+1}{df_{i+1}}\right) + 1 \qquad (1)$$

Keterangan :

$tf_{i,j}$: Jumlah kata-*i* pada dokumen ke-j

N : Jumlah keseluruhan dokumen

df_i : Jumlah dokumen yang mengandung kata ke-*i*

5) Information Gain

Metode *Information Gain* adalah salah satu metode seleksi yang dapat mengurangi kata atau fitur yang tidak relevan untuk proses klasifikasi [5]. Proses untuk mendapatkan *nilai Information Gain* adalah dengan menghitung nilai *entropy* terlebih dahulu agar menghasilkan suatu atribut yang akan menjadi parameter pengukur atas beberapa kumpulan data. Untuk mencari nilai *entropy* dapat menggunakan persamaan (2) [9].

$$Info(D) = -\Sigma_{i=1}^{c} P(i) \log_2 P(i) \quad (2)$$

Keterangan :
C : Banyaknya nilai pada atribut target (Banyaknya kelas untuk melakukan proses klasifikasi)
P(i) : Banyaknya sample pada kelas I dengan sampe keseluruhan

Setelah mendapatkan nilai *entropy* total langkah selanjutnya mencari nilai *entropy* setelah di berikan bobot nilai untuk masing-masing fitur menggunakan persamaan (3).

$$InfoA^{(D)} = -\sum_{J=1}^{v} \frac{|Dj|}{|D|} InfoDj \quad (3)$$

Keterangan :
A : atribut
|D| : Banyaknya keseluruhan sample data
$|D_j|$: Banyaknya sample untuk nilai j
V : Nilai kemungkinan yang akan digunakan atribut A
Info D_j : Nilai *entropy* untuk setiap partisi j

Langkah selanjutnya akan mencari nilai *information gain* menggunakan persamaan (4).

$$InfoGain(A) = -Info(D) - |Info\ A^{(D)} \quad (4)$$

6) Multinomial Naïve Bayes

Proses klasifikasi akan dilakukan menggunakan algoritma *Multinomial Naïve Bayes*. Algoritma *Naïve Bayes* banyak digunakan untuk pemprosesan *data mining* karena penggunaannya yang mudah dan waktu pemprosesan yang cepat. Algortima *Multinomial Naïve Bayes* hanya akan menghitung probabilitas dari setiap kata yang muncul dalam suatu dokumen. Untuk menghitung *posterior probability* dapat menggunakan persamaan (5) [10].

$$P(c) = P(c|d) \propto \prod_{1 \le k \le nd} P(t_k|c) \quad (5)$$

Keterangan :
$P(c|d)$: Probabilitas suatu dokumen yang termasuk dalam kelas c
$P(c)$: Probabilitas prior pada kelas c
$P(t_k|c)$: Probabilitas kata t_k dengan di ketahui kelas c

Nilai probabilitas kelas c dapat dituliskan dengan persamaan (6).

$$P(C) = \frac{N_C}{N} \quad (6)$$

Keterangan :
N_c : Jumlah kelas c pada keseluruhan dokumen
N : Jumlah total dokumen

Untuk menghitung nilai probabilitas dalam kelas c (*likehood probability*) dapat digunakan persmaan (7).

$$P(t_k|c) = \frac{T_{tc}}{\sum_{t' \in v} T_{ct''}} \quad (7)$$

T_{tc} :Jumlah kemunculan kata t dalam dokumen yang masuk kedalam kelas c
$\sum_{t' \in v} T_{ct'}$:Jumlah keseluruhan untuk kemunculan semua kata yang masuk kedalam kelas c

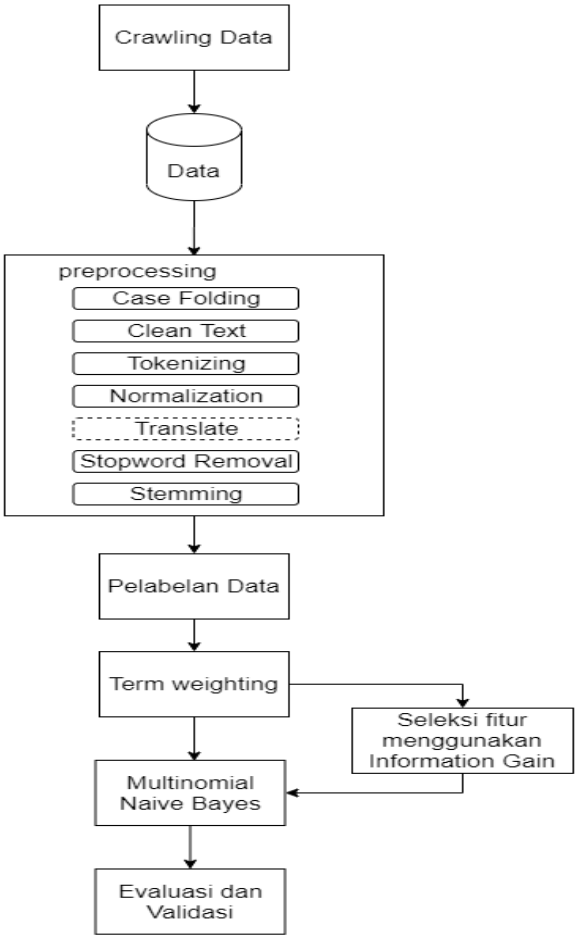
Untuk menghitung nilai probabilitas kata ke-n dapat menggunakan teknik *Laplace Smoothing*. *Laplace smoothing* digunakan untuk mengantisipasi jika probabilitas kata dalam kelas *likehood probability* bernilai 0 dan menyebabkan *posterior probability* bernilai 0. Sehingga akan dilakukan penambahan angka 1 pada *numerator* dan jumlah kosakata *denominator*. *Laplace smoothing* dapat di tuliskan pada persamaan (8).

$$P(t|c) = \frac{T_{ct}+1}{\left(\sum_{t' \in v} T_{ct'}\right)+B'} \quad (8)$$

Keterangan :
B' : Total kata pada keseluruhan kelas

7) Evaluasi dan hasil

Pada tahap ini akan dilakukan evaluasi hasil percobaan, membandingkan dan menganalisa terhadap kinerja seleksi fitur *Information Gain*. Pada penelitian ini evaluasi akan dilakukan dengan *confusion matrix*. *Confusion matrix* akan menghitung nilai *accuracy*, *precison*, *recall*, dan *f1-score*.



Gambar 1 Tahap-tahap analisis sentimen

Pada penelitian ini skenario penelitian akan dilakukan proses *preprocessing* terlebih dahulu. Setelah data melalui proses *preprocessing* kemudian data akan dilakukan pelabelan menggunakan *tools textblob* dari bahasa pemrograman *python* lalu dilakukan proses pembobotan kata menggunakan *TF-IDF*. Selanjutnya dilakukan proses seleksi fitur menggunakan *Information Gain* untuk mengurangi kata atau fitur yang kurang relevan pada proses analisis sentimen dengan harapan dapat meningkatkan akurasi dan efektivitas dalam proses klasifikasi. Untuk mengetahui seberapa besar pengaruh penggunaan seleksi fitur *Information Gain* dalam proses analisis sentimen, maka dilakukan perbandingan nilai akurasi dengan seleksi fitur *Information Gain* dan tanpa menggunakan seleksi fitur *Information Gain*. Dalam penggunaan seleksi fitur *Information Gain* kata atau fitur yang dihapus akan dibatasi dengan *treshold*. Pada penelitian ini nilai *treshold* yang digunakan adalah 0.00001, 0.00005, 0.0001, 0.0005, 0.001, 0.005, 0.01. Sebelum masuk ke proses klasifikasi data akan dibagi menjadi 2 yaitu data *training* dan data *testing* menggunakan *k-fold cross validation*. Nilai *k-fold* yang digunakan adalah 3,5,7,9,10. Setelah pembagian data dilakukan tahap selanjutnya melakukan proses klasifikasi menggunakan *Multinomial Naïve Bayes* dengan nilai parameter *alpha* sebesar 10.

3. HASIL DAN PEMBAHASAN

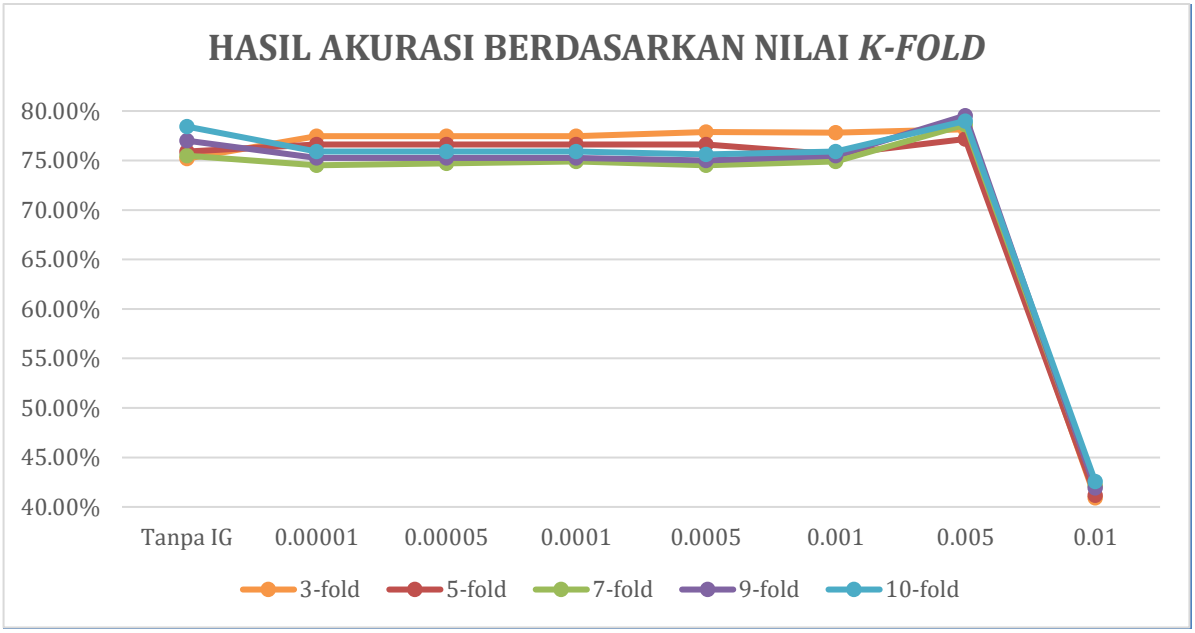
Pada penelitian ini pengumpulan data dilakukan dengan menggunakan kata kunci "belajar online" menghasilkan data sebanyak 1.328 data sedangkan kata kunci "kuliah daring menghasil data sebanyak 2.686 data. Data tersebut kemudian akan di gabungkan sehingga total data yang digunakan sebanyak 4.014 data. Selanjutnya data dilakukan proses *preprocessing* dan penghapusan duplikat sebanyak 444 data sehingga data yang digunakan tersisa sebanyak 3.570 data. Tahap selanjutnya setelah dilakukan *preprocessing* adalah melakukan pelabelan data menggunakan *textblob* dan akan dikategorikan ke dalam kelas positif, negatif, dan netral. Pada penelitian ini hasil pelabelan menggunakan *textblob* terdiri dari label positif berjumlah 828, label negatif berjumlah 401, dan label netral berjumlah 2.341.

Setelah tahap pelabelan data dilakukan, tahap selanjutnya adalah pemberian bobot menggunakan *TF-IDF*. Data yang telah diberikan bobot menggunakan *TF-IDF* kemudian dilakukan proses seleksi fitur *Information Gain* menggunakan 8 variasi *treshold*. Seleksi fitur dilakukan untuk mengurangi kata yang dianggap tidak terlalu berpengaruh dalam proses analisis sentimen dengan harapan dapat meningkatkan hasil akurasi.

Tahap selanjutnya dilakukan proses klasifikasi menggunakan *Multinomial Naïve Bayes*. Data yang akan digunakan untuk proses klasifikasi menggunakan *Multinomial Naïve Bayes* akan dilakukan pembagian data menjadi data *training* dan data *testing* menggunakan *k-fold cross validation*. Proses pembagian data menggunakan *k-fold cross validation* dilakukan untuk mencari hasil akurasi yang paling optimal. Setelah dilakukan akan dilakukan proses klasifikasi kemudian dilakukan evaluasi pengujian menggunakan *confusion matrix*. Hasil pengujian menggunakan *confusion matrix* pada algoritma *Multinomial Naïve Bayes* dengan seleksi fitur *Information Gain* dan tanpa seleksi fitur *Information Gain* pada proses analisis sentimen terkait kuliah daring dapat di lihat pada Tabel 1 dan Gambar 2 .

Tabel 1. Tabel Hasil Akurasi Berdasarkan Nilai *K-Fold*

Treshold	Tanpa IG	0.00001	0.00005	0.0001	0.0005	0.001	0.005	0.01
3-fold	75.21%	77.47%	77.47%	77.47%	77.89%	77.81%	78.15%	40.92%
5-fold	75.91%	76.61%	76.61%	76.61%	76.61%	75.63%	77.17%	41.17%
7-fold	75.49%	74.50%	74.70%	74.90%	74.50%	74.90%	78.62%	42.15%
9-fold	77.02%	75.25%	75.25%	75.25%	75%	75.50%	79.54%	41.91%
10-fold	78.43%	75.91%	75.91%	75.91%	75.63%	75.91%	78.99%	42.57%



Gambar 2 Hasil akurasi berdasarkan nilai *k-fold*

Berdasarkan Tabel 1 dan Gambar 2 dapat dilihat bahwa seleksi fitur *Information Gain* terbukti dapat meningkatkan hasil akurasi. Hasil akurasi tertinggi ketika tidak menggunakan seleksi fitur *Information Gain* diperoleh menggunakan nilai variasi *k-fold* 10 dengan hasil akurasi yang diperoleh sebesar 78.43%, nilai *precission* sebesar 80.43%, *recall* sebesar 78.43%, *f1-score* sebesar 78.92%. Sedangkan hasil akurasi tertinggi ketika menggunakan seleksi fitur *Information Gain* diperoleh menggunakan nilai variasi *k-fold 9* dan nilai variasi *treshold* 0.005 dengan hasil akurasi yang diperoleh sebesar 79.54%, nilai *precission* sebesar 81.09%, *recall* sebesar 79.54%, *f1-score* sebesar 80.08%.

4. KESIMPULAN

Penelitian analisis sentimen dilakukan menggunakan dataset dari Twitter dengan kata kunci "kuliah daring" dan "belajar online" yang terjadi selama pandemi covid-19 dengan jumlah data sebanyak 4.014 data. Penelitian analisis sentimen ini menggunakan algoritma *Multinomial Naïve Bayes* dengan seleksi fitur *Information Gain*. Berdasarkan hasil pengujian analisis sentimen diperoleh hasil bahwa penggunaan seleksi fitur *Information Gain* terbukti dapat meningkatkan akurasi pada klasifikasi teks terkait perkuliahan daring selama pandemi covid-19 menggunakan algoritma *Multinomial Naïve Bayes* dengan akurasi sebesar 79.54%, sedangkan hasil akurasi tertinggi yang diperoleh ketika tidak menggunakan seleksi fitur *Information Gain* sebesar 78.43%.

DAFTAR PUSTAKA

[1] W. A. F. Dewi, "Dampak COVID-19 terhadap Implementasi Pembelajaran Daring di Sekolah Dasar," *Edukatif J. Ilmu Pendidik.*, vol. 2, no. 1, pp. 55–61, 2020, doi: 10.31004/edukatif.v2i1.89.

[2] C. S. Sriyano and E. B. Setiawan, "Pendeteksian Berita Hoax Menggunakan Naive Bayes Multinomial Pada Twitter dengan Fitur Pembobotan TF-IDF," *e-Proceeding Eng. Vol.8, No.2*, vol. 8, no. 2, pp. 3396–3405, 2021.

[3] S. Mandasari, B. H. Hayadi, and R. Gunawan, "Analisis Sentimen Pengguna Transportasi Online Terhadap Layanan Grab Indonesia Menggunakan Multinomial Naive Bayes Classifier," *J-SISKO TECH (Jurnal Teknol. Sist. Inf. dan Sist. Komput. TGD)*, vol. 5, no. 2, p. 118, 2022, doi: 10.53513/jsk.v5i2.5635.

[4] F. Hadaina and U. Budiyanto, "Implementasi Metode Multinomial Naïve Bayes Untuk Sentiment Analysis Terhadap Data Ulasan Produk Colearn Pada Google Play Store Implementation Of Multinomial Naive Bayes Method For Sentiment Analysis Of Colearn Product Review Data On Google Play Store," no. September, pp. 256–263, 2022.

[5] R. I. Pristiyanti, M. A. Fauzi, and L. Muflikhah, "Sentiment Analysis Peringkasan Review Film Menggunakan Metode Information Gain dan K-Nearest Neighbor," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 3, pp. 1179–1186, 2018, [Online]. Available: http://j-ptiik.ub.ac.id.

[6] R. Prakosa, "Analisis Sentimen Terhadap Penggunaan Marketplace Di Indonesia Menggunakan Metode Support Vector Machine," pp. 314–323, 2022.

[7] C. Destitus, W. Wella, and S. Suryasari, "Support Vector Machine VS Information Gain: Analisis Sentimen Cyberbullying di Twitter Indonesia," *Ultim. InfoSys J. Ilmu Sist. Inf.*, vol. 11, no. 2, pp. 107–111, 2020, doi: 10.31937/si.v11i2.1740.

[8] A. Baita, Y. Pristyanto, and N. Cahyono, "Analisis Sentimen Mengenai Vaksin Sinovac Menggunakan Algoritma Support Vector Machine (SVM) dan K-Nearest Neighbor (KNN)," *Inf. Syst. J.*, vol. 4, no. 2, pp. 42–46, 2021, [Online]. Available: https://jurnal.amikom.ac.id/index.php/infos/article/view/687.

[9] M. R. Hasibuan and Marji, "Pemilihan Fitur dengan Information Gain untuk Klasifikasi Penyakit Gagal Ginjal menggunakan Metode Modified K-Nearest Neighbor (MKNN)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 11, pp. 10435–10443, 2019.

[10] A. A. Farisi, Y. Sibaroni, and S. Al Faraby, "Sentiment analysis on hotel reviews using Multinomial Naïve Bayes classifier," *J. Phys. Conf. Ser.*, vol. 1192, no. 1, 2019, doi: 10.1088/1742-6596/1192/1/012024.