

# **Pemodelan Analisis Sentimen Masyarakat terhadap Adaptasi Kebiasaan Baru (AKB) menggunakan Algoritma Naïve Bayes**

**SITI YULIYANTI, SITI SHOLIAH**

Program Studi Sistem Informasi, STMIK Bandung  
Email: [sitiyuliyanti@stmik-bandung.ac.id](mailto:sitiyuliyanti@stmik-bandung.ac.id)

Received 6 Juli 2021 | Revised 21 Juli 2021 | Accepted 18 November 2021

## **Abstrak**

*Pandemi Covid-19 hampir masuk tahun ke dua di Indonesia, pemerintah terus berupaya menekan laju peningkatan penularan Covid-19 melalui berbagai media. Sosialisasi dan informasi melalui media sosial yang merupakan wadah paling cepat untuk tersampaikan kepada masyarakat. Berbagai istilah digunakan seperti adaptasi kebiasaan baru, social distancing, PSBB sampai PPKM sehingga memicu masyarakat untuk beropini di media sosial. Penelitian ini menganalisis sentiment masyarakat terkait opini peningkatan Covid-19 dari twitter. Klasifikasi tweet menggunakan Naive Bayes dengan penambahan seleksi fitur. Penggunaan confusion matriks untuk mengetahui performance algoritma Naive Bayes. Berdasarkan pengujian, penelitian ini menghasilkan 76% dengan accuracy positif sebesar 72,727%, accuracy negatif sebesar 75% dan accuracy netral sebesar 78,947%. Sehingga disimpulkan penggunaan model klasifikasi Naive Bayes dengan fitur seleksi dapat meningkatkan akurasi.*

**Kata kunci:** analisis sentimen, seleksi fitur, twitter crawling, naïve bayes, klasifikasi, emosi

## **Abstract**

*Covid-19 pandemic is almost in its second year in Indonesia, the government continues to try to suppress the rate of increase in the transmission of Covid-19 through various media. Socialization and information through social media which is the fastest medium to be conveyed to the public. Various terms are used, such as adapting new habits, social distancing, PSBB to PPKM, thus triggering the public to share opinions on social media. This study analyzes public sentiment regarding the increasing opinion of Covid-19 from twitter. Tweet classification based on positive, negative and neutral classes using Naive Bayes with feature selection. The use of confusion matrix to determine the performance of the Naive Bayes algorithm. Based This Research, the results from the sentiment analysis system using the nave Bayes classifier of 76% with positive accuracy of 72.727%, negative accuracy of 75% and neutral accuracy of 78.947%. So it can be concluded that the use of the Naive Bayes classification model with the selection feature can increase accuracy.*

**Keywords:** sentiment analysis, fitur selection twitter crawling, naïve bayes, clasification, emotion

## 1. PENDAHULUAN

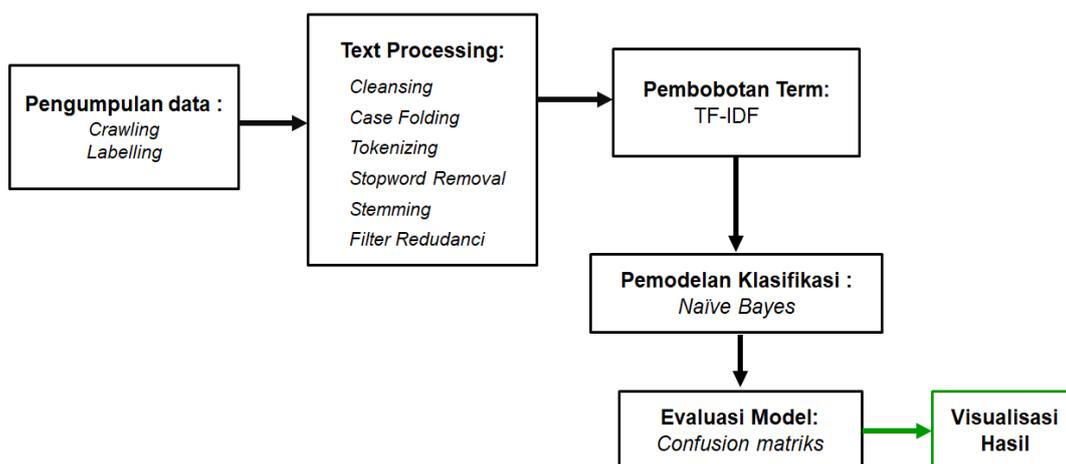
Pandemi yang disebabkan virus Covid-19 melanda berbagai Negara tak terkecuali Indonesia, sehingga penerapan berbagai kebijakan untuk mengurangi menyebarnya covid-19. Berbagai istilah muncul seperti *Social Distancing*, adaptasi kebiasaan baru atau *new normal*, termasuk PSBB (Pembatasan Sosial Berskala Besar) atau saat ini dikenal dengan PPKM. Namun ada beberapa sector yang tidak bias lepas dari rutinitas sehingga harus memulai beradaptasi dengan pemakaian masker, penggunaan handsanitizer atau penyemprotan berkala dengan disinfektan. Hal tersebut membuat banyak opini atau sentiment dari berbagai kalangan yang dituangkan diberbagai media social. Penelitian ini mengangkat objek tentang evaluasi algoritma yang digunakan untuk mengklasifikasikan sentiment masyarakat terhadap adaptasi kebiasaan baru atau new normal yang ramai diperbincangkan di mesia sosial.

Analisis sentiment dapat dimaknai langkah pemberian kelas dalam tingkatan emosi yang dapat tertuang pada cuitan di media sosial (**Mubaroka, Adiwijayab, & Aldhi, 2017**). Beragam opini engguna media social terkait kasus Covid-19. Sosial media menjadi wadah untuk melampiaskan perasaan di kalangan masyarakat, salah satunya pengguna media social. Sebuah sarana didunia masya yaitu cuitan pada Twitter men-trigger beredarnya informasi mengenai hal apapun termasuk pandemi yang disebabkan covid-19. Terkadang cuitan yang menjadi viral dapat menyebabkan pengungkapan makna yang kompleks (**Hermanto & Noviriandini, 2021**). Implementasi analisis sentiment berupa cuitan dalam twitter dapat menimbulkan beragam persepsi atau opini terhadap topik yang di tulis (**Yuliyanti, Djatna, & Sukoco, 2017**). Cuitan (*tweet*) yang berasal dari twitter tidak memiliki interferensi dan berisi banyak kesalahan (*noise*), sehingga diperlukan aplikasi atau sistem untuk memperoleh makna atau nilai yang terkandung dalam sebuah kalimat, penelitian tentang analisis dokumen tekstual paling tinggi akurasi menggunakan metode Naïve Bayes dimana performance algoritma bergantung dari jumlah data. Klasifikasi yang baik dimana hasil akurasi menunjukkan tingkat yang semakin tinggi (**Ratino, Hafidz, Anggraeni, & Gata, 2020**).

Naive Bayes Classifier (NBC) adalah yang paling sering digunakan karena algoritma ini sangat cocok untuk short dan text. Algoritma Naïve Bayes Classifier (NBC) menghitung probabilitas eksplisit untuk hipotesis dan kuat untuk kebisingan (*noise*) dalam input data. Secara statistik pengklasifikasi Bayes meminimalkan kemungkinan kesalahan klasifikasi (**Larose, 2014**). Naive Bayes classifier menggunakan prior probability (yaitu nilai probabilitas yang diyakini benar sebelum melakukan eksperimen) dari setiap label yang merupakan frekuensi masing-masing label pada training set dan kontribusi dari masing-masing fitur (**Ratnawati, 2018**). Performa algoritma klasifikasi dapat ditingkatkan dengan proses *feature selection* dengan tujuan mereduksi atribut- atribut yang tidak sesuai dengan kategorinya sehingga akurasi menjadi lebih baik. *Feature Selection* yang digunakan dalam penelitian ini adalah *Term Frequency-Inverse Document Frequency* (TF-IDF), algoritma ini diimplementasikan untuk menentukan seberapa penting sebuah kata untuk dokumen relatif terhadap *corpus*. Penelitian ini akan menguraikan evaluasi pemodelan klasifikasi sentimen pengguna twitter terhadap adaptasi kebiasaan baru atau lebih dikenal dengan istilah *new normal* dengan algoritma Naive Bayes sehingga diperoleh seberapa akurat algoritma Naïve Bayes yang digunakan sebagai teknik pengklasifikasian pada data yang berasal dari twitter dengan jumlah yang besar. Pengujian untuk mengevaluasi model klasifikasi menggunakan confusion matriks dengan parameter yang dihitung yaitu *accuration, precision, recall* untuk setiap kelas yang dihasilkan oleh model.

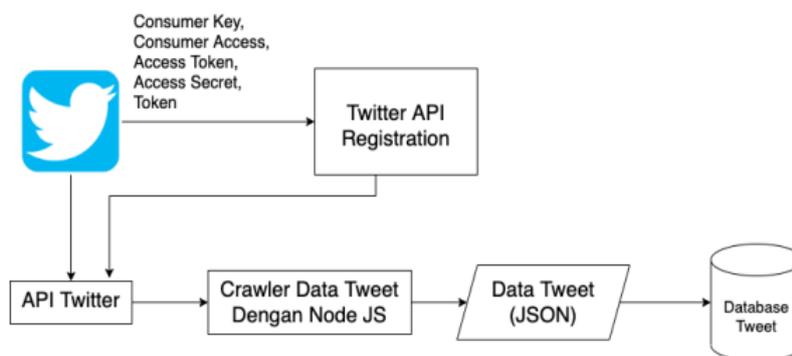
## 2. METODE PENELITIAN

Tahapan penelitian meliputi pengumpulan data cuitan (*crawling*) dan *labelling*, *text processing*, pembobotan *term*, pemodelan dan evaluasi model sebagaimana kerangka penelitian diilustrasikan Gambar 1.



Gambar 1. Kerangka Penelitian

Proses mengambil data dari media sosial *Twitter* untuk pengumpulan *dataset* lebih dikenal dengan nama *crawling*. Pengumpulan data penelitian ini mengangkat objek tweet dari Twitter yang mengandung pembahasan tentang "COVID-19", "PSBB", "Adaptasi Kebiasaan Baru", "New Normal" yang dicrawling sebanyak 10.000 tweet. Penelitian ini menggunakan API *Standard V1.1* dengan tingkat akses hingga 400 kata kunci, dengan batasan 15 permintaan per detik untuk setiap pengguna, semua jendela permintaan dibatasi per 15 menit. Penelitian ini mengambil data set berdasarkan *tweet* dari region Indonesia dan berbahasa Indonesia yang kemudian diberi label positif, negatif dan netral. Tahapan pengumpulan data diilustrasikan pada Gambar 2.



Gambar 2. Flowchart crawling data tweet

Dataset tweet dari hasil *crawling* dipecah pada data *training* dan data *testing*. Kemudian diberi label secara manual terlebih dahulu dengan memberikan label positif, negatif atau netral pada setiap *tweet*. Selain mengumpulkan *data training secara* masal, *dapat pula*

dilakukan proses input data training berupa single tweet pada form training data dalam sistem. Data training disimpan dalam dataset json agar dapat dipanggil kembali (recall) oleh system. Data testing dari hasil crawling disimpan untuk diolah ke dalam sistem sehingga output secara otomatis tampil berupa kelas *sentiment*.

Tahapan *text processing* meliputi *case folding*, *tokenizing*, *stopword removal*, *stemming*. Pembobotan term, penelitian ini menggunakan TF-IDF, pemodelan dilakukan dengan mengimplementasikan Naive Bayes Classifier sedangkan evaluasi model menggunakan confusion matriks dan divisualisasikan dengan grafik. Tahapan *text processing* dimulai dengan cleansing untuk mengurangi noise dilakukan proses pembersihan. Kata yang dihilangkan dapat berupa karakter dari *script* HTML, kata kunci, *emoticon*, *hashtag* (#), *username* dan URL. *Noise* dapat mengganggu proses penilaian analisa sentiment. Selanjutnya dilakukan *case folding*, yang merupakan perubahan semua huruf yang terdapat dalam dokumen menjadi huruf kecil atau *lowercase* dan menghilangkan karakter selain huruf. Selanjutnya dilakukan proses *tokenizing*, dimana pada proses ini dilakukan pemotongan atau pemisahan setiap kata yang terdapat dalam kalimat berdasarkan pemisah kata seperti spasi, koma (,) dan titik (.) (Yuliyanti & Rizky, 2020). Bagian yang hanya memiliki satu karakter non alfabet dan angka dibuang, sehingga menghasilkan kata-kata dalam *array*. Kemudian tahapan *stopword removal* dilakukan, dimana semua kosakata yang tidak memiliki makna dihilangkan dari dalam dokumen sehingga menyisakan kata yang bermakna saja di dalam dokumen. Sebelum proses *stopword removal* harus dibuat daftar *stopword list*. Yang termasuk ke dalam daftar ini misalnya "oleh", "pada", "di", "sebuah", "karena" dan lain-lain. Jika ada kata yang termasuk dalam *stopword list* maka dilakukan penghapusan dari deskripsi dan dianggap sebagai kata-kata yang mencirikan isi dari suatu dokumen atau kata kunci. Pada penelitian ini menggunakan kamus *stopword* yang disediakan oleh *library* dari Node JS yaitu *stopwords-iso/stopwords-id* yang mendukung kamus *stopword* berbahasa Indonesia.

Selanjutnya, penelitian ini melakukan proses *stemming* dengan menggunakan kamus kata dasar dari kata dasar untuk membandingkan dasar dan terkait. Data kata kunci diambil dari kamus *online* Bahasa Indonesia berjumlah 29.932. *Stemming* adalah proses mengelompokkan berbagai variasi morfologi suatu kata atau kalimat menjadi satu bentuk dasar (Siswandi, Permana, & Emarilis, 2021). Tahap terakhir yaitu filter *redudansi* untuk meningkatkan penghitungan frekuensi kata selama proses pembobotan, diperlukan penyaringan *redudansi* untuk mencari kata yang sama. Dalam arti tertentu, kata-kata dalam dokumen dibandingkan dengan kamus dalam *database*. Jika kata ini ditemukan, kata tersebut merupakan sinonim (Sukmana & Salsabilla, 2018).

Setelah proses *text preprocessing*, penelitian ini melakukan ekstraksi fitur untuk meminimalisir noise, sehingga akurasi tinggi. Data yang sudah siap diproses itu kemudian dihitung seberapa banyak kemunculan atau frekuensi kemunculan setiap katanya di dalam dokumen, yang dikenal dengan tahap pembobotan *term* menggunakan metode TF-IDF. Pembobotan term adalah pembobot untuk setiap kata dengan penggunaan TFIDF (*Term Frequency – Inversed Document Frequency*) (Gunawan, Pratiwi, & Pratama, 2018). Proses TF-IDF dengan menghitung bobot dengan mengintegrasikan *term frequency* (tf) dan *inverse document frequency* (idf). TF merupakan cara menghitung bobot tiap term dalam *tweet*. Jumlah dokumen yang mengandung suatu term tertentu dinamakan DF. Dokumen Frekuensi merupakan metode feature selection yang paling sederhana dengan waktu komputasi yang rendah. *Inverse Document Frequency* (IDF) TF fokus pada term di cuitan, maka IDF focus dengan munculnya term di keseluruhan lokasi *tweet* dan term yang jarang muncul pada keseluruhan koleksi term dinilai berharga. *Term Frequency Inverse Document*

*Frequency* mencantumkan bobot setelah menghasilkan term dalam tweet kemudian menghitungnya dengan mengintegrasikan  $tf$  dengan  $idf$ . Sebagaimana ditunjukkan pada Persamaan 1.

$$TF - IDF = TF * IDF = TF * \log(n/df) \quad (1)$$

TF : Frekuensi Teks  
 $df$  : Frekuensi Dokumen  
 $n$  : Jumlah Dokumen

Tahapan pembangunan model mengimplemntasikan algoritma Naïve Bayes, dimana tahap ini untuk mengetahui pola sentiment pada setiap tweet berdasarkan kelasnya sebagaimana diuraikan pada alur algoritma *Naïve Bayes*:

1. Pembacaan data latih
2. Penjumlahan dan probabilitas, dicari nilai *mean* untuk data numeric dan standar deviasi setiap parameter seperti ditunjukkan pada Persamaan 2.

$$\mu = \frac{\sum_{i=1}^n x_i}{n} \quad (2)$$

$\mu$  : rata-rata hitung (mean)  
 $x_i$  : nilai sampel ke- $i$   
 $n$  : jumlah sampe

3. Menemukan nilai dalam tabel *mean*, deviasi standar, dan probabilitas
4. Hasil klasifikasi

Nilai probabilitas dibandingkan dengan suatu sampel di kelasnya dan yang lain untuk menemukan kelas. Kemudian dilakukan penentuan kelas yang sesuai sampel dengan perbandingan nilai posterior tiap kelas dan memilih kelas dengan nilai Posterior tertinggi. Probabilitas prior dan teorema bayes sebagaimana ditunjukkan pada Persamaan 3 dan Persamaan 4. Pengklasifikasian dengan Naïve Bayes dibagi kedalam 2 proses, yaitu fase *learning* dan fase *testing*. Fase *learning* untuk menghasilkan model yang dijadikan refresnsi dalam mengklasifikasikan sentiment pada fase *testing* (**Rahayu & Zharfan, 2015**).

$$P(H) = \frac{N_j}{N} \quad (3)$$

P (H) : Peluang prior  
 $N_j$  : Jumlah data pada suatu kelas  
 $N$  : Jumlah total data

$$P(H|X) = \frac{P(H/X).P(H)}{P(X)} \quad (4)$$

X : Data yang kelasnya belum diketahui  
 $H$  : Hipotesis data merupakan suatu class spesifik  
 $P(H|X)$  : Probabilitas Hipotesis H berdasarkan kondisi X (posterior probabilitas)  
 $P(H)$  : Probabilitas hipotesis H (prior probabilitas)  
 $P(X|H)$  : probabilitas X berdasarkan kondisi pada hipotesis H  
 $P(X)$  : Probabilitas X

Penelitian ini melakukan evaluasi atau menguji validasi model bertujuan untuk menilai kinerja algoritma dengan menampilkan akurasi, *precision* dan *recall menggunakan confusion matrix*. Penggunaan Confusion Matrix untuk menghitung tingkat akurasi, dengan menghitung jumlah prediksi benar dan salah berdasarkan perbandingan data real atau prediksi target (**Arini, Wardhani, & Octaviano, 2020**). Pengujian data yang kemudian diprediksi kelas dibandingkan dengan kelas sesungguhnya dari data uji yang disimpan terlebih dahulu. Tahpan selanjutnya dapat direpresentasikan hasil nilai *accuration*, *precision*, *recall*, dan *f1-*

score (**Imron, 2018**). Terdapat beberapa parameter pengujian berdasarkan perbandingan nilai prediksi dan nilai actual yaitu:

1. *Accuracy* dapat diartikan pengujian dari kedekatan nilai prediksi dengan nilai actual dalam kumpulan data.

$$Accuracy = \frac{N_{benar}}{N} \times 100\% \quad (5)$$

2. *Precision* adalah sebuah pengujian melalui perbandingan jumlah informasi yang sesuai diperoleh sistem dengan jumlah seluruh informasi yang terambil oleh system, walaupun tidak sesuai.

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

3. *Recall* dapat diperoleh dengan membandingkan jumlah informasi relevan yang diperoleh sistem dengan jumlah seluruh informasi relevan yang ada dalam label informasi.

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

### 3. HASIL DAN PEMBAHASAN

#### 3.1. Seleksi Fitur

Proses pembobotan penelitian ini menggunakan TF-IDF sebagaimana ditunjukkan pada Persamaan 1, dan *flowchart* yang diilustrasikan pada Gambar 3. Langkah pertama yaitu Menghitung jumlah kemunculan kata, kemudian menghitung jumlah dokumen dan *query* berisi kata (DF) lalu menghitung IDF menggunakan Persamaan 6 dan menghitung TF-IDF dengan Persamaan 7 dan hasil ditunjukkan pada Tabel 1.



**Gambar 3. Flowchart Proses Pembobotan Term**

**Tabel 1. Hasil Perhitungan TF, DF, IDF dan TF-IDF**

Kata	TF										TF-IDF									
	D1	D2	D3	D4	D5	D6	D7	D8	D9	DF	IDF	D1	D2	D3	D4	D5	D6	D7	D8	D9
selamat	2	0	0	0	0	0	0	0	0	2	0.6532 1251	1.306 42503	0	0	0	0	0	0	0	0
sehat	1	1	0	1	0	0	0	0	0	3	0.4771 2125	0.477 12125	0.4771 2125	0.4771 2125	1	0	0	0	0	0
rutin	1	0	0	0	0	0	0	0	0	1	0.9542 4251	0.954 24251	0	0	0	0	0	0	0	0
pulih	0	1	0	0	0	0	0	0	0	1	0.9542 4251	0	0.9542 4251	0	0	0	0	0	0	0
protokol	1	0	0	1	0	0	0	0	0	2	0.6532 1251	0.653 21251	0	0.6532 1251	1	0	0	0	0	0
pandemi	0	1	0	0	0	0	0	0	0	1	0.9542 4251	0	0.9542 4251	0	0	0	0	0	0	0
lindungi	1	0	0	0	0	0	0	0	0	1	0.9542 4251	0.954 24251	0	0	0	0	0	0	0	0
lelah	0	2	0	0	0	0	0	0	0	2	0.6532 1251	0	1.3064 2503	0	0	0	0	0	0	0
kondisi	0	1	0	0	0	0	0	0	0	1	0.9542 4251	0	0.9542 4251	0	0	0	0	0	0	0
jaga	1	1	0	0	0	0	0	0	0	2	0.6532 1251	0.653 21251	0.6532 1251	0	0	0	0	0	0	0
covid	1	0	1	0	0	0	0	0	0	2	0.6532 1251	0.653 21251	0	0.6532 1251	0	0	0	0	0	0
ancam	1	0	0	0	0	0	0	0	0	1	0.9542 4251	0.954 24251	0	0	0	0	0	0	0	0
aktivitas	1	0	0	0	0	0	0	0	0	1	0.9542 4251	0.954 24251	0	0	0	0	0	0	0	0
adaptasi	1	0	0	0	1	1	0	0	0	3	0.4771 2125	0.477 12125	0	0	0.4 771 212 6	0.4 771 212 7	0	0	0	0

Berdasarkan Tabel 1, diilustrasikan sampel 9 dokumen yang berasal dari 9 tweet dari ribuan data yang sudah dicrawling untuk dihitung TF, DF, IDF dan TF-IDF. Tahapan pembobotan tern ini merupakan transformasi data dari data teks ke dalam sebuah matriks yang berisi numerik yang kemudian data dapat diolah dan diterapkan ke dalam model klasifikasi.

### 3.2. Pemodelan *Naïve Bayes Classification*

Mengasumsikan bahwa efek fitur tertentu dalam kelas tidak bergantung pada fitur lainnya merupakan *Naïve Bayes Classification*. Asumsi ini menyederhanakan komputasi yang disebut kebebasan bersyarat kelas. Langkah pertama adalah memahami dan mengidentifikasi fitur dan label potensial. Klasifikasi ini memiliki dua tahap, tahap pembelajaran dan tahap evaluasi. Ketepatan analisis algoritma benar jika menghasilkan luaran yang benar jika menerima masukan sesuai dengan definisinya, dan dapat diterminasi (berakhir) **(Sim, 2013)**.

Proses training pada penelitian ini seperti diilustrasikan dalam Tabel 2, dengan pengklasifikasi melatih modelnya pada *dataset* tertentu sebagaimana dihasilkan probabilitas yang merupakan hasil perhitungan setiap term yang ditambahkan dan dicari nilai paling tinggi dan dijadikan kelas untuk *term* seperti pada Tabel 3. Tahapan ini data telah berbobot seperti contoh pada Tabel 3 sebagai data latih sebagai referensi pembentukan model klasifikasi, kemudian pencarian nilai probabilitas kategori dan peluang setiap kata dari data latih. Untuk mengakumulasi nilai keluar kategori dan kata dengan mengimplementasikan Persamaan 2 pada setiap *term* untuk setiap kelas dari data latih. Jumlah seluruh *term* dalam data latih yang dipakai pada perhitungan sebanyak 9 *term* dimana 3 *term* adalah positif, 4

*term* adalah negatif dan 2 *term* adalah netral. Data hasil pembobotan term pada Tabel 3 dilakukan perhitungan probabilitas setiap term. Sebagai contoh kata "selamat" pada Tabel 3 berdasarkan perhitungan probabilitas menggunakan Persamaan 3 dan 4.

**Tabel 2. Contoh Data Training**

Kode	Sebelum Pra-Proses	Setelah Pra-Proses	Kelas
D1	@PT_Transjakarta: Selamat Pagi, selamat beraktivitas. Adaptasi Kebiasaan Baru menjadi rutinitas keseharian dengan menjaga protokol kesehatan dimulai dari diri sendiri untuk saling melindungi dari ancaman Covid-19.	selamat selamat aktivitas adaptasi rutin jaga protokol sehat lindung ancam covid	Positif
D2	@tia_alzahira: Ketika ditanya lelah nggak sih sama pandemi yang nggak terasa udah setahun berlalu?	lelah pandemi lelah kondisi pulih wajib disiplin protokol jaga sayang	Negatif
D3	@Cahyailahi6: Pemerintah Indonesia terima 83,9 juta dosis vaksin covid-19	dosis vaksin covid	Netral
D4	@Dien_Drew1: Iya lagi naik"lagi nih kak .gmn ya dilarang mudik mngkn taun lalu gak mudik .sdh lelah mereka kak dgn aturan"yg gak jelas gini	larang mudik mudik lelah aturan	?

**Probabilitas kata "selamat" pada Data Training (Positif) berdasarkan Tabel 3**

$$p("selamat"|"pos") = \frac{"selamat"|"pos" + 1}{("pos") + |kosakata|} = \left( \frac{1.306 + 1}{33 + 46} \right) = 0,0292$$

$$p("selamat"|"neg") = \frac{"selamat"|"neg" + 1}{("neg") + |kosakata|} = \left( \frac{0 + 1}{19 + 46} \right) = 0,0154$$

$$p("selamat"|"net") = \frac{"selamat"|"net" + 1}{("net") + |kosakata|} = \left( \frac{0 + 1}{23 + 46} \right) = 0,0145$$

Maka kata selamat mendapatkan nilai positif yang lebih besar dibandingkan dengan term lainnya. Hasil probabilitas kata lainnya terlihat pada Tabel 3 sebagai luarannya, probabilitas diperoleh dari nilai yang paling tinggi dan dijadikan kelas untuk term tersebut.

**Tabel 3. Perhitungan Probabilitas Kata**

Pemodelan Analisis Sentimen Masyarakat terhadap Adaptasi Kebiasaan Baru (AKB) menggunakan Algoritma Naïve Bayes

Kata	Jumlah frekuensi ( $n_{k, pos/neg/net}$ )			Probabilitas Kata $p(w_k pos/neg/net)$		
	pos	neg	net	pos	neg	net
selamat	1,3064250	3	0	0,029195	0,0153846	0,014
aktivitas	0,9542425	1	0	0,024737	0,0153846	0,014
adaptasi	1,4313637	6	0	0,030776	0,0153846	0,014
rutin	0,9542425	1	0	0,024737	0,0153846	0,014
jaga	0,6532125	1	0	0,020926	0,0153846	0,023
protokol	1,3064250	3	0	0,029195	0,0153846	0,014
sehat	0,9542425	1	0	0,024737	0,0153846	0,021
lindungi	0,9542425	1	0	0,024737	0,0153846	0,014
ancam	0,9542425	1	0	0,024737	0,0153846	0,014
covid	0,6532125	1	0	0,020926	0,0153846	0,023
lelah	0,9542425	0	0	0,024737	0,0153846	0,014
pandemi	0,6532125	0	0	0,020926	0,0153846	0,023
kondisi	0,9542425	0	0	0,024737	0,0153846	0,014
pulih	0,9542425	0	0	0,024737	0,0153846	0,014
wajib	0,9542425	0	0	0,024737	0,0153846	0,014

Langkah-langkah pada proses *testing* menyerupai dengan alur proses *training*, yaitu:

Proses 1: Pengujian memberikan data uji pada model dalam pelatihan. Contoh perhitungan pada proses 2 menggunakan D4.

Proses 2: Setelah proses seleksi data *testing* dihitung nilai probabilitas tiap kata dalam data *testing* dengan hasil probabilitas term dari data *training*.

Perhitungan probabilitas dilakukan dengan Persamaan 1. Probabilitas Data Testing (Positif) berdasarkan Tabel 1, menghasilkan 0.012 untuk peluang kata "larang", "mudik", "lelah" pada kelas positif. Sedangkan pada kelas negative menghasilkan nilai berturut-turut yaitu 0.0147, 0.035, 0.0154.

$$p(\text{"larang"}|\text{"pos"}) = \left(\frac{0 + 1}{33 + 46}\right) = 0,012$$

$$p(\text{"mudik"}|\text{"pos"}) = \left(\frac{0 + 1}{33 + 46}\right) = 0,012$$

$$p(\text{"lelah"}|\text{"pos"}) = \left(\frac{0 + 1}{33 + 46}\right) = 0,012$$

Probabilitas Data Testing (Negatif) berdasarkan Tabel 1

$$p(\text{"larang"}|\text{"neg"}) = \left(\frac{-0,4575749 + 1}{19 + 46}\right) = 0,0147$$

$$p(\text{"mudik"}|\text{"neg"}) = \left(\frac{-0,4575749 + 1}{19 + 46}\right) = 0,035$$

$$p(\text{"lelah"}|\text{"neg"}) = \left(\frac{0 + 1}{19 + 46}\right) = 0,0154$$

Probabilitas Data Testing (Netral) berdasarkan Tabel 1

$$p(\text{"larang"}|\text{"net"}) = \left(\frac{-0,4575749 + 1}{23 + 46}\right) = 0,0145$$

$$p(\text{"mudik"}|\text{"net"}) = \left(\frac{-0,4575749 + 1}{23 + 46}\right) = 0,0145$$

$$p(\text{"lelah"}|\text{"net"}) = \left(\frac{0 + 1}{23 + 46}\right) = 0,0334$$

### Proses 3: Menghitung probabilitas kategori dengan Persamaan 2

$$P(\text{testing}|\text{pos}) = P(\text{pos}) \times P(\text{larang}|\text{pos}) \times P(\text{mudik}|\text{pos}) \times P(\text{lelah}|\text{pos})$$

$$P(\text{testing}|\text{pos}) = 0,33 \times 0,0147 \times 0,034 \times 0,0154 = 2,53998\text{E-}06$$

$$P(\text{testing}|\text{neg}) = P(\text{neg}) \times P(\text{larang}|\text{neg}) \times P(\text{mudik}|\text{neg}) \times P(\text{lelah}|\text{neg})$$

$$P(\text{testing}|\text{neg}) = 0,44 \times 0,0147 \times 0,035 \times 0,0154 = 3,48625\text{E-}06$$

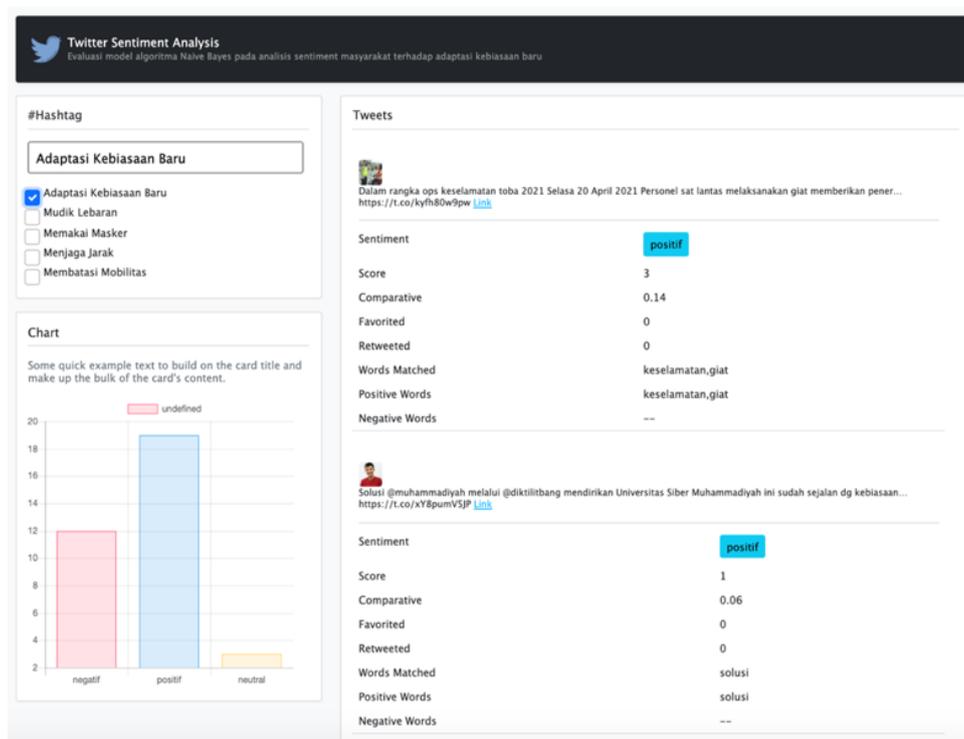
$$P(\text{testing}|\text{net}) = P(\text{net}) \times P(\text{larang}|\text{net}) \times P(\text{mudik}|\text{net}) \times P(\text{lelah}|\text{net})$$

$$P(\text{testing}|\text{net}) = 0,22 \times 0,0145 \times 0,0145 \times 0,0334 = 1,54492\text{E-}06$$

Nilai probabilitas tertinggi kategori negatif sebesar 3,48625E-06, sehingga D4 dapat dikategorikan kedalam kelas Negatif. Sebagai contoh kata "selamat" pada Tabel 2 menunjukkan perhitungan nilai positif yang lebih besar dibandingkan dengan term lainnya. Hasil probabilitas kata lainnya terlihat pada Tabel 3 di bawah yang merupakan luaran perhitungan setiap term yang sudah ditambahkan, kemudian dicari nilai yang terbesar dan digunakan sebagai kelas untuk term tersebut.

### 3.3 Antarmuka sistem

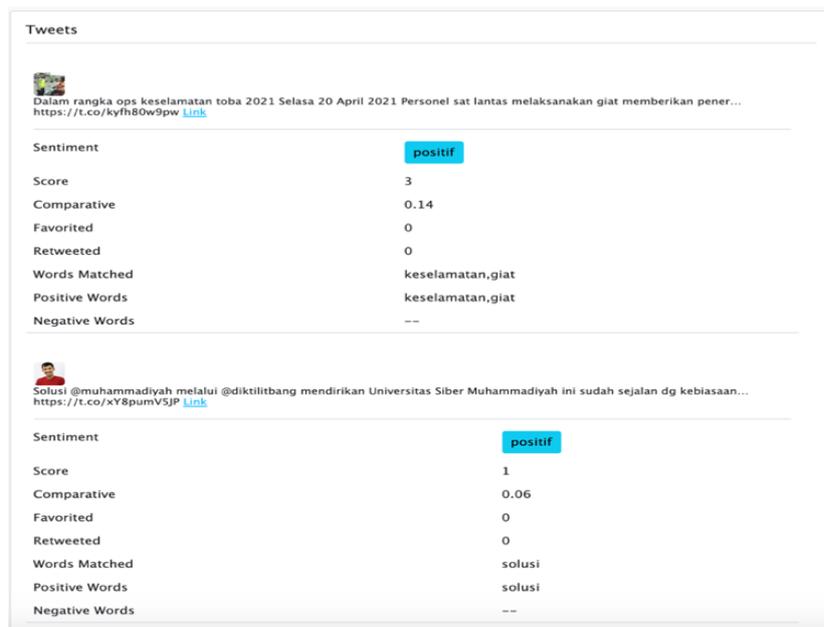
Penelitian ini menghasilkan antarmuka berbasis Web yang dinamis, dengan menyajikan informasi *tweet*, *score*, klasifikasi sentimen yang dilengkapi *summary* berupa grafik. Tampilan antarmuka Web Twitter Sentiment Analysis ditunjukkan pada Gambar 4.



Gambar 4. Tampilan Antarmuka

Pengguna dapat melakukan proses pemilihan #hashtag atau topik yang ingin dilakukan analisa *tweet*. *Tweet* berubah setelah topik di pilih. Setelah pengguna memilih topik atau hashtag yang ingin dianalisa, Data *tweet* dari twitter berada pada tab *tweets*.

Pemodelan Analisis Sentimen Masyarakat terhadap Adaptasi Kebiasaan Baru (AKB) menggunakan Algoritma Naïve Bayes



Gambar 5. Tab Tweets

**Evaluasi Model menggunakan *Confusion Matrix***

Pada tahap evaluasi, pengklasifikasi menguji kinerja pengklasifikasi. Kinerja dievaluasi berdasarkan parameter seperti *acuration*, *precision* dan *recall*.

**4.2. Pengujian *Confusion Matrix***

Pengujian dilakukan dengan menggunakan 70 data berlabel. Parameter yang dihitung dalam pengujian ini yaitu *accuration*, *precision*, *recall*. Ketiga parameter tersebut dihitung dengan menggunakan metode *Confusion Matrix*. Metode ini umumnya digunakan untuk melakukan perhitungan keakurasian pada data mining. Hasil pengujian akurasi klasifikasi tweets dengan menggunakan 20 data training dan 50 data uji.

Tabel 4. Pengujian dengan *Confusion Matrix*

<i>Predicted Class</i>		<i>Actual Class</i>		
		Positif	Negatif	Netral
	Positif	8	2	1
	Negatif	1	15	4
	Netral	0	4	15
<b>Total Kelas</b>		9	21	20

Penelitian ini menghasilkan akurasi yang lebih tinggi meskipun tanpa seleksi fitur yaitu sebagaimana terlampir pada Tabel 5 melalui Persamaan 5, Persamaan 6 dan Persamaan 7 untuk menghasilkan performance algoritma.

Tabel 5. Hasil Pengujian dengan *confussion matrix* tanpa seleksi fitur

<i>Class</i>	<i>N(truth)</i>	<i>N(classified)</i>	<i>Precision</i>	<i>Recall</i>
<b>Positif</b>	9	11	70,127%	79,84%
<b>Negatif</b>	21	20	71.7%	70,12%
<b>Netral</b>	20	19	75,247%	72.58%

Tabel 6. Hasil Pengujian dengan *confussion matrix* dengan seleksi fitur

<b>Class</b>	<b>N(truth)</b>	<b>N(classified)</b>	<b>Precision</b>	<b>Recall</b>
<b>Positif</b>	9	11	72,727%	88,88%
<b>Negatif</b>	21	20	75%	71,4%
<b>Netral</b>	20	19	78,947%	75%

Hasil pengujian menunjukkan tingkat *accuracy* sebesar 63.21% tentang analisis sentimen opini publik mengenai covid-19 pada twitter menggunakan metode naïve bayes tanpa menggunakan seleksi fitur (**Syarifuddin, 2020**). Pada penelitian ini menghasilkan akurasi yang lebih tinggi meskipun tanpa seleksi fitur yaitu sebagaimana terlampir pada Tabel 5.

Berdasarkan pengujian akurasi menggunakan Persamaan 5, Persamaan 6 dan Persamaan 7 dimana hasil akurasi klasifikasi tweets dari sistem analisis sentimen dengan menggunakan Naïve Bayes classifier yang ditambah dengan seleksi fitur sebesar 76% dengan *accuracy* positif sebesar 72,727%, *accuracy* negatif sebesar 75% dan *accuracy* netral sebesar 78,947%. Penelitian sebelumnya sebagaimana ditunjukkan pada Tabel 5, menghasilkan akurasi yang lebih rendah dibandingkan dengan Tabel 6 yang sudah menggunakan seleksi fitur sebelum pemodelan dilakukan.

## 5. KESIMPULAN

Simpulan penelitian menunjukkan bahwa dengan penambahan seleksi fitur dapat meningkatkan nilai akurasi sebanyak 5% yaitu 78,947% dibandingkan tanpa menggunakan seleksi fitur. Pengaruh akurasi dalam metode *Naïve Bayes* bergantung kepada dataset, praproses dan optimasi metode sebelum penggunaan algoritma atau metode untuk klaifikasi. Pada dataset yang digunakan untuk *training* masih belum terlalu bervariasi, namun sudah menunjukkan hasil pembelajaran yang signifikan antara klasifikasi tanpa dan dengan menggunakan fitur seleksi. Dimana akurasi tertinggi pada klasifikasi Naïve Bayes tanpa seleksi fitur sebesar 75,247% sebagaimana ditunjukkan pada Tabel 5 sedangkan klasifikasi dengan penambahan seleksi fitur ditunjukkan pada Tabel 6 mencapai nilai sebesar 78,947% pada kelas netral. Penelitian ini lebih baik hasil performance model klasifikasinya jika dibandingkan dengan penelitian sebelumnya yang menghasilkan *accuracy* sebesar 63.21% tentang analisis sentimen mengenai covid-19 pada twitter dengan mengimplementasikan naïve bayes (**Syarifuddin, 2020**). Peningkatan sebesar 15% antara klasifikasi dengan penambahan fitur seleksi sebagaimana terlampir pada Persamaan 2.

## UCAPAN TERIMA KASIH

Penulis sampaikan ucapan terimakasih kepada seluruh pihak, khususnya civitas akademik STMIK Bandung atas kerjasama dan bantuan baik secara langsung maupun tidak langsung sehingga penelitian ini dapat dirampungkan, meskipun masih harus dikembangkan demi peningkatan kualitas. Penulis juga sampaikan ucapan terimakasih yang mendalam kepada keluarga dan rekan yang selalu memberi semangat untuk penyelesaian penelitian ini, dengan harapan dapat menjadi referensi untuk para peneliti yang berminat mengembangkan penelitian ini.

## DAFTAR RUJUKAN

- Arini, Wardhani, L. K., & Octaviano, D. (2020). Perbandingan Seleksi Fitur Term Frequency & Tri-Gram Character menggunakan Algoritma Naïve Bayes Classifier pada Tweet Hashtag #2019gantipresiden. *KILAT*, 103 - 114.
- Gunawan, B., Pratiwi, H. S., & Pratama, E. E. (2018). Sistem Analisis Sentimen pada Ulasan Produk menggunakan Naive Bayes. *Jurnal Edukasi dan Penelitian Informatika*, 113-118.
- Hermanto, & Noviriandini, A. (2021). ANALISA SENTIMEN TERHADAP BELAJAR ONLINE PADA MASA COVID-19. *Jurnal Informatika Kaputama (JIK)*, 129-13.
- Imron, A. (2018). *Analisis Sentimen terhadap Tempat Wisata di Kabupaten Rembang menggunakan Metode Naive Bayes Classifier*. Yogyakarta: Universitas Islam Indonesia.
- Larose, D. T. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining*. John Wiley & Sons, Inc.
- Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Chicago: Morgan & Claypool Publisher Retrieved from. Retrieved from Chicago:.
- Mubaroka, M. S., Adiwijayab, & Aldhi, M. D. (2017). Aspect-based sentiment analysis to review products using Naïve Bayes. *International Conference on Mathematics: Pure, Applied and Computation* (pp. 1-8). Cambridge: AIP Publishing.
- Rahayu, M. I., & Zharfan, F. T. (2015). Analisis Sentimen Laporan Perkembangan Anak Didik Taman Kanak. *Jurnal Teknologi Informasi dan Komunikasi*, 30-36.
- Ratino, Hafidz, N., Anggraeni, S., & Gata, W. (2020). Sentimen Analisis Informasi Covid-19 menggunakan Support Vector Machine dan Naïve Bayes. *Jurnal JUPITER*, 01-11.
- Ratnawati, F. (2018). Implementasi Algoritma Naive Bayes Terhadap Analisis Sentimen Opini Film Pada Twitter. *Jurnal Inovtek Polbeng - Seri Informatika*, 51-59.
- Saleh, A. (2015, Desember 23). *Klasifikasi Gejala Depresi Pada Manusia dengan Metode Naïve Bayes menggunakan Java*. Retrieved from Digital Library STMIK AKAKOM: <http://eprints.akakom.ac.id/id/eprint/452>
- Sim, A. X. (2013, July 01). *Pengenalan Desain dan Analisis Algoritma*. Retrieved from [www.dev.bertzzie.com](http://www.dev.bertzzie.com): <http://dev.bertzzie.com/knowledge/analisis-algoritma/PengenalanDesaindanAnalisisAlgoritma.html#algoritma-yang-baik>
- Siswandi, A., Permana, A. Y., & Emarilis, A. (2021). Stemming Analysis Indonesian Language News Text with Porter. *Journal of Physics: Conference Series*, 01-07.
- Sukmana, R. N., & Salsabilla, Z. S. (2018). Filter Bahasa Kasar menggunakan Algoritma Naïve Bayes. *Jurnal Teknologi Informasi dan Komunikasi*, 1-5.

- Syarifuddin, M. (2020). Analisis Sentimen Opini Publik Mengenai Covid-19 Pada Twitter Menggunakan Metode Naïve Bayes Dan Knn. *Inti Nusa Mandiri*, 23-28.
- Yuliyanti, S., & Rizky. (2020). Implementasi Algoritma Rabin Karp untuk Mendeteksi Kemiripan Dokumen STMIK Bandung. *Bangkit Indonesia*, 1 -6.
- Yuliyanti, S., Djatna, T., & Sukoco, H. (2017). Sentiment Mining of Community Development Program Evaluation Based on Social Media. *TELKOMNIKA*, 1858 - 1864.