

## **Analisis Sentimen Gojek Indonesia Pada Twitter Menggunakan Algoritme Naïve Bayes Dan Support Vector Machine**

**Yusuf Khoiruddin<sup>1\*</sup>, Ahmad Fauzi<sup>2</sup>, Amril Mutoi Siregar<sup>3</sup>**

Program Studi Teknik Informatika, Universitas Buana Perjuangan Karawang, Karawang, Indonesia

\*e-mail Corresponding Author: if19.yusufkhoiruddin@mhs.ubpkarawang.ac.id

### **Abstract**

*Transportation is an important element in everyday life and serves as a means of connecting between places. Online motorcycle taxi services such as Gojek have penetrated all regions, enabling users to order transportation services such as motorcycle taxis, taxis and cars online through applications. With the large number of content tweeted by Twitter users related to the use of the Gojek application, therefore sentiment analysis is needed to determine user perceptions of a topic or event. This study aims to analyze people's responses to online transportation through data collected from tweets. The data is then classified into two sentiment classes, namely positive and negative. The classification results using the Naive Bayes algorithm show an accuracy of 91%, while the use of the SVM (Support Vector Machine) algorithm produces a higher accuracy, which is equal to 99%. This indicates that the SVM algorithm is more effective in producing higher classification accuracy than the Naive Bayes algorithm.*

**Keywords:** *Sentiment analysis; Online transportation; Naive bayes; Support Vector Machine*

### **Abstrak**

Transportasi merupakan elemen penting dalam kehidupan sehari-hari dan berfungsi sebagai sarana untuk menghubungkan antar tempat. Layanan ojek *online* seperti Gojek telah merambah di seluruh wilayah, memungkinkan pengguna untuk memesan layanan transportasi seperti ojek, taksi, dan mobil secara online melalui aplikasi. Dengan banyaknya isi *tweet* pengguna *twitter* terkait dengan penggunaan aplikasi Gojek, oleh karena itu diperlukan analisis sentimen untuk mengetahui persepsi pengguna terhadap suatu topik atau peristiwa. Penelitian ini bertujuan untuk melakukan analisis respons masyarakat terhadap transportasi *online* melalui data yang terkumpul dari *tweet*. Data tersebut kemudian diklasifikasikan ke dalam dua kelas sentimen yaitu positif dan negatif. Hasil klasifikasi menggunakan algoritme *Naive Bayes* menunjukkan akurasi sebesar 91%, sedangkan penggunaan algoritme SVM (*Support Vector Machine*) menghasilkan akurasi yang lebih tinggi, yaitu sebesar 99%. Hal ini mengindikasikan bahwa algoritme SVM lebih efektif dalam menghasilkan akurasi klasifikasi yang lebih tinggi dibandingkan dengan algoritme *Naive Bayes*.

**Kata kunci:** *Analisis sentimen; Transportasi online; Naive bayes; Support Vector Machine*

### **1. Pendahuluan**

Kendaraan menjadi bagian vital dalam aktivitas harian kita. Kendaraan umum seperti bus, taksi, dan ojek pangkalan masih menjadi pilihan sebagian besar orang untuk mencapai tujuan mereka. Namun, kemajuan teknologi juga telah membawa perkembangan dalam transportasi, dengan munculnya layanan transportasi *online* yang menawarkan tarif lebih terjangkau daripada transportasi konvensional [1]. Transportasi memiliki peran yang sangat signifikan dalam mendukung pertumbuhan ekonomi suatu negara. Jika sistem transportasi tidak efektif, maka akan terjadi hambatan dalam pertukaran barang dan jasa, yang akan mengganggu aktivitas ekonomi negara tersebut. Di Indonesia, perkembangan transportasi *online*, seperti layanan ojek *online*, terutama melalui Gojek, telah menyebar ke seluruh wilayah. Fenomena ini telah menjadi suatu hal yang tidak dapat diabaikan dalam konteks studi tentang

transportasi dan ekonomi di Indonesia. Layanan ini memungkinkan pengguna untuk memesan layanan transportasi seperti ojek motor, taksi, dan mobil secara *online* melalui aplikasi Gojek. Selain itu, Gojek juga menawarkan beragam layanan lain seperti pengiriman barang, layanan kebersihan rumah, dan layanan belanja. Gojek telah berkembang menjadi salah satu perusahaan teknologi terkemuka di wilayah Asia Tenggara dan memberikan dampak besar pada perekonomian lokal [2].

Dengan semakin populernya transportasi *online*, banyak masyarakat yang mulai menggunakan layanan tersebut sebagai pilihan utama dalam beraktivitas. Fenomena tersebut mendorong banyak pihak untuk memanfaatkan data dan informasi yang bervariasi untuk memperoleh informasi mengenai tren yang sedang berkembang di masyarakat, terutama dalam diskusi di media sosial [3]. Beberapa masyarakat memberikan pendapat positif tentang kemudahan dan efisiensi yang ditawarkan oleh transportasi *online*, sementara yang lain memberikan pendapat negatif tentang masalah keamanan dan kualitas layanan. Karena itu, diperlukan analisis sentimen yang memanfaatkan proses *text mining* dan pengelompokan berdasarkan kategori sentimen yang telah ditetapkan, untuk menentukan tingkat akurasi informasi yang diperoleh [4].

*Twitter* merupakan salah satu *platform* media sosial yang populer di kalangan masyarakat global. Media sosial ini memungkinkan para pengguna untuk berbagi opini dan informasi dalam format teks yang pendek namun efektif. Dengan jumlah pengguna yang cukup besar, melalui media *Twitter* pengguna dapat menyampaikan pendapat secara bebas dan terbuka. Selain itu, *platform* ini juga menawarkan fitur trending topik, yang memungkinkan para pengguna untuk mengetahui berita terkini yang sedang populer dan banyak dibicarakan [5]. Hal ini memungkinkan untuk mengetahui opini masyarakat secara keseluruhan dan membuat keputusan yang lebih baik [6].

Penelitian ini bertujuan untuk menganalisis Sentimen Gojek Indonesia Pada Pengguna *Twitter* Menggunakan Algoritme *Naïve Bayes* Dan *Support Vector Machine*.

## 2. Tinjauan Pustaka

Pada penelitian sebelumnya tentang "Analisis Sentimen Transportasi *Online* pada *Twitter* Menggunakan Metode Klasifikasi *Naïve Bayes* dan *Support Vector Machine*" didapatkan hasil analisis bahwa penggunaan metode SVM pada GrabId menghasilkan akurasi yang cukup tinggi, dengan *class precision* positif dan negatif sebesar 86.47% dan 46.67%, serta *class recall* positif dan negatif sebesar 96.21% dan 18.06%. Akurasi secara keseluruhan mencapai 84.08%. Sedangkan untuk GojekId, metode SVM juga memberikan hasil yang cukup baik, dengan *class precision* positif dan negatif sebesar 73.90% dan 35.65%, serta *class recall* positif dan negatif sebesar 89.84% dan 15.07%. Namun, akurasi secara keseluruhan sedikit lebih rendah, yaitu sebesar 69.50%. Dari hasil tersebut, dapat disimpulkan bahwa metode SVM menunjukkan kinerja yang lebih baik dalam melakukan analisis sentimen pada data dari kedua perusahaan [7].

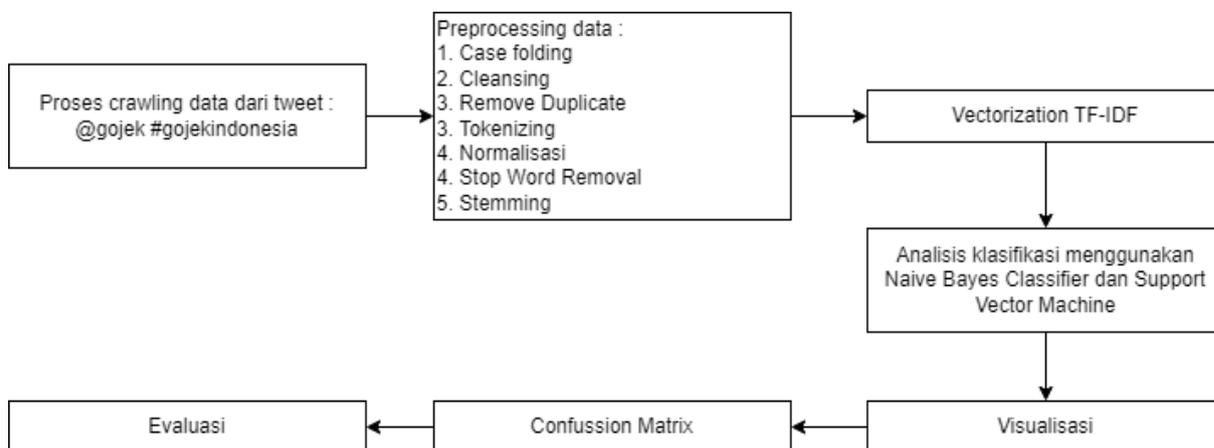
Penelitian lainnya yang berjudul "Analisis Sentimen pada *review* Aplikasi Grab di Google Play Store Menggunakan *Support Vector Machine*" hasil analisis sentimen pada 900 *review data testing* dengan menggunakan metode *Support Vector Machine* dan nilai pengujian  $fold=5$  menghasilkan akurasi sebesar 85,54%. Prediksi pada *data testing* menunjukkan bahwa terdapat 59 *review* yang dikategorikan sebagai *review* positif, dan 675 *review* lainnya dikategorikan sebagai *review* negatif [8].

Pada penelitian lain yang berjudul "Perbandingan Metode *Support Vector Machine* dan *Decision Tree* untuk Analisis Sentimen *Review* Komentar pada Aplikasi Transportasi *Online*", ditemukan bahwa metode *Support Vector Machine* memiliki akurasi sebesar 90,20%, sedangkan metode *Decision Tree* memiliki akurasi sebesar 89,80%. [9].

Penelitian kami melakukan analisis sentimen masyarakat pada *platform* media sosial *Twitter* dengan mengelompokkan data *tweet* menjadi kelas sentimen negatif dan positif. Data *tweet* kemudian melalui proses pra-pemrosesan sebelum dilakukan klasifikasi menggunakan metode algoritme *Naive Bayes* dan *Support Vector Machine*. Selanjutnya, hasil klasifikasi dari kedua algoritme dibandingkan untuk mendapatkan nilai akurasi yang terbaik. Penelitian ini diharapkan dapat memberikan kontribusi pada pengembangan metode analisis sentimen pada data sosial media. Hasil penelitian dapat dijadikan acuan untuk pengambilan keputusan dalam berbagai konteks seperti pengembangan layanan jasa transportasi online gojek yang lebih baik serta dapat meningkatkan kepuasan konsumen di Indonesia.

### 3. Metodologi

Penelitian ini bertujuan untuk menganalisis sentimen dari data komentar pengguna Gojek yang terdapat di *platform* media sosial *Twitter*. Penelitian ini akan membandingkan kinerja algoritme *Naive Bayes* dan *SVM* dalam melakukan klasifikasi sentimen pada data tersebut, dengan tujuan untuk mendapatkan nilai akurasi yang paling tinggi dari kedua algoritme tersebut. Tahapan pada proses penelitian ditampilkan pada Gambar 1.



Gambar 1. Alur Model Analisis Sentimen Twitter

#### 3.1 Pengumpulan Data

Data untuk penelitian ini dikumpulkan melalui proses *crawling* data *Twitter* dengan menggunakan *API Key Twitter*, yang diimplementasikan dalam bahasa pemrograman *Python*. Proses *crawling* data dilakukan dengan membuat program yang mengambil *tweet* berdasarkan kata kunci dan jumlah *tweet* yang ditentukan, misalnya, @gojek dan #gojekindonesia. Program akan mengambil setiap *tweet* yang mention ke akun Gojek di *Twitter*. Dataset yang didapatkan berjumlah 1500 data *tweet* yang akan digunakan dalam penelitian ini.

#### 3.2 Pre-processing data

*Preprocessing* data merupakan tahapan penting dalam persiapan data untuk analisis. Proses ini dilakukan untuk mengubah data mentah menjadi data yang siap digunakan dalam analisis, sehingga dapat memperoleh hasil analisis yang akurat dan relevan diantaranya:

1. *Case folding*, adalah proses mengubah *text tweet* menjadi *text* yang sama untuk melakukan pencarian atau analisis seperti mengkonversi *text* menjadi huruf kecil.
2. *Cleaning*, adalah proses pembersihan teks dan menghilangkan komponen yang tidak diperlukan seperti *url*, *mention*, *emoticon*, dan *hashtag*.
3. *Remove duplicate*, adalah proses menyaring suatu data untuk memisahkan data yang memiliki duplikat.
4. *Tokenizing*, adalah proses pemecahan *text* menjadi unit yang lebih kecil dengan memisahkan tanda koma atau spasi.
5. *Normalization*, adalah proses mengubah data ke bentuk standar untuk analisis dan pemrosesan data yang lebih mudah.
6. *Stop word removal*, adalah proses menghapus kata-kata umum yang dianggap memiliki makna yang sedikit dari sebuah teks.
7. *Stemming*, adalah proses menyederhanakan kata-kata dalam suatu teks menjadi kata dasar.

#### 3.3 Pembobotan data TF-IDF

Setelah tahap *preprocessing* data selesai, langkah berikutnya adalah melakukan pembobotan data menggunakan metode *Term Frequency-Inverse Document Frequency* (TF-IDF). Pada tahap ini, kata-kata dalam indeks diidentifikasi dan diberikan bobot pada setiap kata terhadap dokumen yang bersangkutan. Metode TF-IDF digunakan untuk menentukan bobot

setiap kata, berdasarkan frekuensi kemunculan kata dalam dokumen, serta frekuensi kemunculan kata tersebut dalam seluruh dokumen yang ada. Bobot ini menunjukkan seberapa pentingnya suatu kata dalam dokumen tersebut dibandingkan dengan kata lainnya di dalam dokumen yang sama. Semakin tinggi bobot suatu kata, semakin penting pula kata tersebut dalam dokumen tersebut [10]. Pembobotan TF-IDF dihitung menggunakan rumus berikut:

$$W_{t,d} = t_{ft,d} \times id_{ft} = t_{ft,d} \times \log N/d_{ft} \quad (1)$$

Keterangan:

$W_{t,d}$  : Bobot (*term*) TF-IDF  
 $t_{ft,d}$  : Jumlah frekuensi kemunculan kata  
 $id_{ft}$  : Jumlah inverse frekuensi dokumen tiap kata  
 $d_{ft}$  : Jumlah frekuensi dokumen tiap kata  
 $N$  : Jumlah total dokumen yang digunakan dalam penelitian

### 3.4 Analisis klasifikasi *Naive Bayes* dan SVM

*Naive Bayes* merupakan sebuah algoritme klasifikasi yang didasarkan pada *teorema Bayes* yang digunakan untuk menentukan kelas dari suatu input berdasarkan probabilitas yang diperoleh dari data latih. Algoritme ini menggunakan asumsi bahwa setiap fitur input adalah independen satu sama lain dan setiap fitur memberikan kontribusi yang sama terhadap kelas yang ditentukan. Algoritme ini dianggap "*naive*" karena asumsi dasar yang digunakan, yaitu bahwa semua fitur input sama-sama independen satu sama lain. Klasifikasi *Naive Bayes* adalah sebuah metode klasifikasi yang termasuk dalam *supervised learning*, yang artinya membutuhkan seorang *supervisor* sebagai guru dalam proses pembelajaran [11]. Metode ini bertujuan untuk melakukan prediksi dengan memanfaatkan informasi masa lalu dan mengestimasi kemungkinan kejadian di masa depan. Salah satu keunggulan utama dari metode ini adalah kemampuannya dalam memperlakukan setiap kondisi secara independen yang dapat memberikan pengaruh yang signifikan terhadap hasil prediksi. Dalam proses klasifikasi *Naive Bayes*, opini atau sentimen yang diberikan oleh pengguna akan dianggap independen dari sentimen lainnya [12].

$$P(H|X) = \frac{P(H|X)}{P(X)} \cdot P(H) \quad (2)$$

Keterangan:

$x$  = Data class yang belum diketahui  
 $H$  = Hipotesis data dengan class spesifik  
 $P(H|X)$  = Probabilitas hipotesis  $H$  dengan kondisi  $X$  (*posteriori probabilitas*)  
 $P(H)$  = Probabilitas hipotesis  $H$  (*prior probabilitas*)  
 $P(X|H)$  = Probabilitas  $X$  berdasarkan kondisi pada hipotesis  $H$   
 $P(X)$  = Probabilitas  $X$

*Support Vector Machine* (SVM) adalah algoritme yang digunakan untuk memisahkan dua kelas data dengan membuat sebuah garis atau *hypersurface*. SVM juga dapat digunakan untuk masalah regresi dengan mencari *hypersurface* yang meminimalkan jumlah *error* dalam memprediksi nilai target. SVM mencari garis atau *hypersurface* yang paling baik yang memisahkan dua kelas dengan mencari titik yang paling jauh dari garis tersebut, yang disebut sebagai vektor dukungan. Dalam metode SVM, digunakan konsep *maximum marginal hyperplane* yang merupakan *hyperplane* terbaik dengan margin terbesar untuk memisahkan data untuk setiap kelas dengan tingkat dan akurasi yang optimal. Dalam penelitian ini, penggunaan algoritme SVM untuk melakukan klasifikasi dilakukan melalui pustaka atau *library* yang disebut *sklearn* [13]. Rumus perhitungan SVM dapat dilihat sebagai berikut:

$$(w \cdot x_i) + b = 0 \quad (3)$$

Dalam persamaan 2, data  $x_i$  tergolong ke dalam kelas -1  
 $(w \cdot x_i + b) \leq 1, y_i = -1 \quad (4)$

Persamaan 3 digunakan untuk menentukan bahwa data xi termasuk dalam kelas +1  
 $(w \cdot x_i + b) \geq 1, y_i = -1$  (5)

### 3.5 Visualisasi

Setelah menerapkan kedua algoritme, visualisasi data dilakukan melalui penggunaan wordcloud dan grafik untuk masing-masing kelas sentimen positif dan negatif berdasarkan data latih yang telah diproses. Tujuan dari visualisasi ini adalah untuk memperoleh informasi mengenai kata-kata yang paling sering muncul dalam cuitan [14].

### 3.6 Confussion Matrix

Dengan menghasilkan *confussion matrix* sebagai hasil akhir akurasi algoritme *Naive Bayes* dan *SVM* dalam mengklasifikasikan sentimen. Evaluasi model diukur dengan menggunakan persamaan akurasi, *recall*, dan presisi [15]. Bentuk *confussion matrix* sebagai berikut:

Tabel 1. *Confussion Matrix*

| Confussion Matrix |       | Nilai sebenarnya                               |  |
|-------------------|-------|--|--|
|                   |       | True   | False  |
| Nilai Prediksi    | True  | TP ( <i>True positive</i> )<br>Correct result  | FP ( <i>False positive</i> )<br>Unexpected result        |
|                   | False | TN ( <i>False Negative</i> )<br>Missing result | TN ( <i>True Negative</i> )<br>Correct absence of result |

Keterangan :

TP : TP (*True Positive*) menunjukkan jumlah prediksi benar untuk data yang sebenarnya positif.

TN : TN (*True Negative*) menunjukkan jumlah prediksi benar untuk data yang sebenarnya negatif.

FP : FP (*False Positive*) menunjukkan jumlah prediksi salah untuk data yang sebenarnya negatif

FN : FN (*False Negative*) menunjukkan jumlah prediksi salah untuk data yang sebenarnya positif.

### 3.7 Akurasi

Akurasi adalah ukuran yang menunjukkan seberapa baik model klasifikasi dapat memprediksi kelas yang tepat dari data yang diberikan. Dalam penelitian ini, rumus akurasi dihitung dengan menghitung persentase data yang diprediksi dengan benar oleh model yang dibangun, di mana P adalah jumlah data yang benar diklasifikasikan sebagai kelas positif dan N adalah jumlah data yang benar diklasifikasikan sebagai kelas negatif. Persamaan 4 menunjukkan rumus untuk menghitung akurasi.

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \tag{6}$$

### 3.8 Recall

*Recall* adalah ukuran seberapa banyak data yang diklasifikasikan dengan benar dalam kategori tertentu dibandingkan dengan jumlah total data yang sebenarnya termasuk dalam kategori tersebut. Dalam beberapa kasus, istilah "*recall*" juga dikenal sebagai "sensitivitas". Hasil *recall* dapat dihitung menggunakan rumus seperti yang diberikan pada Persamaan 5.

$$Recall = \frac{TP}{TP+FN} \times 100\% \tag{7}$$

### 3.9 Presisi

Presisi merujuk pada tingkat ketepatan dan keakuratan respons sistem dalam memberikan informasi yang relevan dengan permintaan pengguna. Untuk menghitung nilai presisi, digunakan sebuah rumus yang dinyatakan dalam Persamaan 6.

$$Precision = \frac{TP}{TP+FP} \times 100\% \tag{8}$$

#### 4. Hasil dan Pembahasan

##### 4.1 Pengumpulan Data

Pengumpulan data pada penelitian ini dilakukan melalui *crawling* data *Twitter* menggunakan *API Key Twitter* dalam bahasa pemrograman *python* di google colab, kemudian dataset yang didapatkan disimpan ke dalam file dengan format *.csv*. Proses *crawling data* ditunjukkan pada Gambar 2.

|      | Datetime                  | Tweet Id            | Text  |
|------|---------------------------|---------------------|---|
| 0    | 2023-01-30 23:59:34+00:00 | 1620210143643451393 | @gojekindonesia Tarif nya naik lagi y?? Dlu dr... |
| 1    | 2023-01-30 23:56:45+00:00 | 1620209431136059392 | @tanyakanri gausah menyamaratakan, kita gatau ... |
| 2    | 2023-01-30 23:56:02+00:00 | 1620209253981241346 | pio, caki, barka, aya, ustaz bagas, duta warna... |
| 3    | 2023-01-30 23:53:07+00:00 | 1620208518656410625 | @tanyakanri Gapapa da ada uang banyak buat go...  |
| 4    | 2023-01-30 23:52:49+00:00 | 1620208442086803776 | @bbycopss 1. buka gojek -&gt; pilih lainnya...    |
| ...  | ...                       | ...                 | ...   |
| 1496 | 2023-01-29 16:26:52+00:00 | 1619733826275794944 | @muhamadhilal_ @tdrmedia coba cek gojek.g...      |
| 1497 | 2023-01-29 16:26:31+00:00 | 1619733741076905986 | Nonton drakor, jd pengen nyobain ice americano... |
| 1498 | 2023-01-29 16:23:30+00:00 | 1619732981773631490 | @ansoryusufs @worksfess Aamin. Lapo gojek? K...   |
| 1499 | 2023-01-29 16:22:39+00:00 | 1619732767872544772 | @tanyakanri @gojekindonesia                       |
| 1500 | 2023-01-29 16:22:06+00:00 | 1619732629489872898 | @zlotml_ @preatzel big aja/n"mama jinn kuk...     |

1501 rows × 5 columns

Gambar 2. Isi Dataset

##### 4.2 Preprocessing

Pada tahapan *preprocessing* data dilakukan dengan beberapa tahapan, yaitu *casefolding*, *cleaning*, *tokenizing*, *normalization*, *stopword removal*, dan *stemming*.

###### 4.2.1 Casefolding

*Casefolding* merupakan proses mengubah kata menjadi huruf kecil (*lowercase*) yang bertujuan untuk memudahkan pada proses selanjutnya, seperti yang terlihat pada Tabel 2.

Tabel 2. Hasil *Casefolding*

| Sebelum                          | Sesudah                        |
|----------------------------------|--------------------------------|
| HAHAHAHA GOJEK LU KENAPA SIE???? | hahaha gojek lu kenapa sie???? |

###### 4.2.2 Cleaning

*Cleaning* yaitu menghilangkan tanda baca, tautan, tagar, *emoji*, dan dan simbol yang tidak digunakan, seperti pada Tabel 3.

Tabel 3. Hasil *Cleaning*

| Sebelum                        | Sesudah                    |
|--------------------------------|----------------------------|
| hahaha gojek lu kenapa sie???? | hahaha gojek lu kenapa sie |

###### 4.2.3 Tokenizing

*Tokenizing* adalah proses penguraian teks menjadi bagian-bagian kecil atau token berdasarkan struktur kalimat, paragraf, atau dokumen.

Tabel 4. Hasil *Tokenizing*

| Sebelum                    | Sesudah                                    |
|----------------------------|--|
| hahaha gojek lu kenapa sie | ['hahaha', 'gojek', 'lu', 'kenapa', 'sie'] |

#### 4.2.4 Normalization

Normalisasi dilakukan untuk mengubah kata-kata tidak standar atau tidak resmi menjadi bentuk standar yang lebih umum dikenal atau digunakan. Hal ini bertujuan untuk menghindari variasi kata yang sama dan memudahkan proses analisis teks. Sebagai contoh, normalisasi dapat mengubah kata-kata singkatan menjadi bentuk lengkap, seperti "tidak" dari "gak" atau "tidak ada" dari "ga ada". seperti pada Tabel 5.

Tabel 5. Hasil *Normalization*

| Sebelum                                      | Sesudah                                       |
|--|---|
| ['hahahaha', 'gojek', 'lu', 'kenapa', 'sie'] | ['hahahaha', 'gojek', 'anda', 'kenapa', 'si'] |

#### 4.2.5 Stopword Removal

*Stopword removal* digunakan untuk penyaringan kata yang tidak memiliki arti seperti kata sambung, keterangan, dan yang lainnya menggunakan library NLTK berbahasa indonesia.

Tabel 6. Hasil *Stopword Removal*

| Sebelum                                       | Sesudah                     |
|---|-----------------------------|
| ['hahahaha', 'gojek', 'anda', 'kenapa', 'si'] | ['gojek', 'anda', 'kenapa'] |

#### 4.2.6 Stemming

Proses *stemming* adalah salah satu teknik dalam pemrosesan bahasa alami yang bertujuan untuk mengubah kata-kata menjadi kata dasar menggunakan *library* sastrawi.

Tabel 7. Hasil *Stemming*

| Sebelum  | Sesudah   |
|--|---|
| ['tech', 'company', 'posisi', 'junior', 'gojek', 'jutaan', 'umr', 'pengalaman', 'startup', 'risiko', 'cocok', 'belajar', 'fastpaced', 'environment'] | ['tech', 'company', 'posisi', 'junior', 'gojek', 'juta', 'umr', 'alam', 'startup', 'risiko', 'cocok', 'ajar', 'fastpaced', 'environment'] |

### 4.3 Pembobotan Data TF-IDF

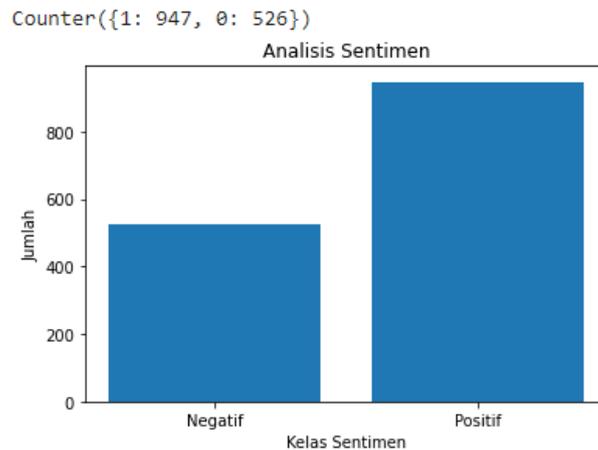
Setelah dilakukan tahap *pre-processing*, selanjutnya menentukan bobot pada setiap kata yang disebut *term weighting* menggunakan metode TF-IDF.

| term         | TF                  | TF-IDF               |
|--------------|---------------------|----------------------|
| download     | 0.03571428571428571 | 0.2116997318897704   |
| gojekgrab    | 0.03571428571428571 | 0.16218921819263146  |
| wajah        | 0.03571428571428571 | 0.17897596288426488  |
| udah         | 0.03571428571428571 | 0.1972188343716217   |
| hidup        | 0.03571428571428571 | 0.182737913765347293 |
| jakarta      | 0.07142857142857142 | 0.32004810625551755  |
| money        | 0.03571428571428571 | 0.18694447864120093  |
| kosong       | 0.03571428571428571 | 0.22197409882019215  |
| pakai        | 0.07142857142857142 | 0.18798254317923158  |
| transport    | 0.03571428571428571 | 0.1724635779238572   |
| publicparkir | 0.03571428571428571 | 0.23645408753833985  |
| banyak       | 0.07142857142857142 | 0.3944376087432434   |
| sistem       | 0.03571428571428571 | 0.16960498979185447  |
| bayar        | 0.03571428571428571 | 0.1352687066791893   |
| cashless     | 0.03571428571428571 | 0.22197409882019215  |
| last         | 0.03571428571428571 | 0.23645408753833985  |
| but          | 0.03571428571428571 | 0.18694447864120093  |
| not          | 0.03571428571428571 | 0.1972188343716217   |
| luar         | 0.03571428571428571 | 0.23645408753833985  |
| gintar       | 0.03571428571428571 | 0.2116997318897704   |
| pilih        | 0.03571428571428571 | 0.23645408753833985  |
| teman        | 0.03571428571428571 | 0.14350178448675474  |
| iya          | 0.07142857142857142 | 0.11718311174738353  |
| nder         | 0.03571428571428571 | 0.12688160842642498  |

Gambar 3. Hasil Perhitungan TF-IDF

#### 4.4. Klasifikasi

Dalam proses klasifikasi data, sebuah dataset dipilih dan dibagi menjadi dua kelas yaitu sentimen positif dan sentimen negatif. Selanjutnya, dataset tersebut dibagi menjadi data *training* dan data *testing* yang terdiri dari 1500 data dengan rasio pembagian 0.3, di mana 70% dari data digunakan untuk *training* dan 30% digunakan untuk *testing*.



Gambar 4. Hasil Klasifikasi Sentimen

#### 4.5 Visualisasi

Data hasil *pre-processing* direpresentasikan dengan menggunakan visualisasi *wordcloud*. Visualisasi *wordcloud* dilakukan pada setiap kelas berdasarkan data latih yang sudah melalui proses pengolahan data.



Gambar 5. Wordcloud Sentimen Positif



Gambar 6. Wordcloud Sentimen Negatif

#### 4.6 Evaluasi

Pada pengujian digunakan algoritme *Naïve Bayes* dan *SVM* untuk mengklasifikasikan data pada dataset yang sudah dipisahkan menjadi data *training* 1031 data dan data *testing* 442 data. Setelah dilakukan klasifikasi, *confusion matrix* digunakan untuk mengevaluasi akurasi, presisi, dan *recall* dari kedua algoritme tersebut.

Tabel 8. Hasil *Confusion Matrix*

| Algoritme          | Akurasi | Presisi | Recall |
|--------------------|---------|---------|--------|
| <b>Naïve Bayes</b> |         |         |        |
| Training           | 91%     | 93%     | 94%    |
| Testing            | 64%     | 70%     | 80%    |
| <b>SVM</b>         |         |         |        |
| Training           | 99%     | 99%     | 99%    |
| Testing            | 62%     | 71%     | 72%    |

Dalam tabel di atas, terlihat hasil evaluasi kinerja pada pengujian menggunakan data *training* dan *testing*, dapat disimpulkan bahwa algoritme SVM memiliki kinerja yang lebih baik dibandingkan dengan *Naïve Bayes*. Algoritme SVM mencapai akurasi, presisi, dan *recall* sebesar 99%, sementara *Naïve Bayes* hanya mencapai akurasi 91%, presisi 93%, dan *recall* 94%. Hasil evaluasi menunjukkan bahwa SVM memiliki kemampuan yang lebih baik dalam mengenali pola pada data yang kompleks dan memiliki keakuratan yang lebih tinggi dalam memprediksi hasil. Namun ke-efektifan algoritme tergantung pada karakteristik data yang digunakan dan ketersediaan sumber daya yang dimiliki. Oleh karena itu, pemilihan algoritme yang tepat harus didasarkan pada karakteristik data yang akan diproses dan konteks penggunaannya

## 5. Simpulan

Berdasarkan hasil analisis sentimen menggunakan algoritme klasifikasi *Naïve Bayes* dan SVM pada dataset *twitter* dengan kata kunci gojek. Kinerja kedua algoritme dilakukan dengan menggunakan data *training* dan *testing*. Hasil evaluasi menunjukkan bahwa algoritme SVM memiliki kemampuan yang lebih baik dalam mengenali pola pada data yang kompleks dan memiliki keakuratan yang lebih tinggi dalam memprediksi hasil dibandingkan dengan algoritme *Naïve Bayes*. Algoritme SVM mampu mencapai akurasi, presisi, dan *recall* sebesar 99%, sedangkan algoritme *Naïve Bayes* hanya mampu mencapai akurasi sebesar 91%, presisi sebesar 93%, dan *recall* sebesar 94%. Terdapat kecenderungan sentimen positif yang lebih banyak dari sentimen negatif terhadap layanan gojek pada dataset *twitter* yang digunakan. Hal ini menunjukkan bahwa mayoritas pengguna merasa puas dengan layanan yang disediakan oleh gojek. Namun, penelitian ini juga menekankan pentingnya pemilihan algoritme yang tepat dalam melakukan analisis sentimen terhadap data yang kompleks seperti data *twitter*.

## Daftar Referensi

- [1] B. M. Pintoko and K. M. L., "Analisis Sentimen Jasa Transportasi Online pada Twitter Menggunakan Metode Naive Bayes Classifier," *e-Proceeding Eng.*, vol. 5, no. 3, pp. 8121–8130, 2018.
- [2] M. Dwijayanti, F. Noor Hasan, and R. Zein Adam, "Analisis Sentimen Pada Ulasan Pelanggan Menggunakan Metode Naive Bayes Classifier (Studi Kasus: Grab Indonesia)," *Pros. Semin. Nas. Teknoka*, vol. 6, no. January, pp. 93–99, 2022, doi: 10.22236/teknoka.v6i1.441.
- [3] A. H. T. Siregar, Amril Mutoi, "Aplikasi Linier Regresi Dengan Algoritme Jaringan Syaraf Tiruan Untuk Sentimen Analisis," *Techno Xplore: Jurnal Ilmu Komputer dan Teknologi Informasi (2015)*, vol. 5, no. 3, pp. 248–253, 2020.
- [4] A. Agustian, "Penerapan Analisis Sentimen Dan Naive Bayes," *Jurnal Tekno Kompak* vol. 7, no. 3, pp. 243–249, 2022.
- [5] L. Luthfanida, "Analisis Sentimen Data Twitter Menggunakan Metode Naive Bayes Dan Support Vector Machine (SVM) Tentang Presiden Jokowi 3 Periode," *Djtechno J. Teknol. Inf.*, vol. 3, no. 1, pp. 5–11, 2022, doi: 10.46576/djtechno.v3i1.2143.
- [6] A. Turmudi Zy, A. Nugroho, A. Rivaldi, and I. Afriantoro, "Analisis Sentimen Terhadap Pembobolan Data pada Twitter dengan Algoritme Naive Bayes," *J. Teknol. Inform. dan Komput.*, vol. 8, no. 2, pp. 202–213, 2022, doi: 10.37012/jtik.v8i2.1240.
- [7] E. Dwianto and M. Sadikin, "Analisis Sentimen Transportasi Online pada Twitter Menggunakan Metode Klasifikasi Naive Bayes dan Support Vector Machine," *Format J. Ilm. Tek. Inform.*, vol. 10, no. 1, p. 94, 2021, doi: 10.22441/format.2021.v10.i1.009.
- [8] R. Wahyudi and G. Kusumawardana, "Analisis Sentimen pada Aplikasi Grab di Google Play Store Menggunakan Support Vector Machine," *J. Inform.*, vol. 8, no. 2, pp. 200–207, 2021, doi: 10.31294/ji.v8i2.9681.
- [9] K. A. Rokhman, B. Berlilana, and P. Arsi, "Perbandingan Metode Support Vector Machine Dan Decision Tree Untuk Analisis Sentimen Review Komentar Pada Aplikasi Transportasi Online," *J. Inf. Syst. Manag.*, vol. 3, no. 1, pp. 1–7, 2021, doi: 10.24076/joism.2021v3i1.341.
- [10] A. M. Siregar, "Perbandingan Pembobotan Kata Dalam Sistem Temu Balik Informasi," *Techno Xplore J. Ilmu Komput. dan Teknol. Inf.*, vol. 2, no. 2, pp. 1–9, 2018, doi: 10.36805/technoxplore.v2i2.302.
- [11] A. S. Rahayu and A. Fauzi, "Komparasi Algoritme Naive Bayes Dan Support Vector

- Machine ( SVM ) Pada Analisis Sentimen Spotify,” *JSON*, vol. 4, pp. 349–354, 2022, doi: 10.30865/json.v4i2.5398.
- [12] M. I. Petiwi, A. Triayudi, and I. D. Sholihati, “Analisis Sentimen Gofood Berdasarkan Twitter Menggunakan Metode Naïve Bayes dan Support Vector Machine,” *J. Media Inform. Budidarma*, vol. 6, no. 1, p. 542, 2022, doi: 10.30865/mib.v6i1.3530.
- [13] V. W. D. Thomas and F. Rumaisa, “Analisis Sentimen Ulasan Hotel Bahasa Indonesia Menggunakan Support Vector Machine dan TF-IDF,” *J. Media Inform. Budidarma*, vol. 6, no. 3, p. 1767, 2022, doi: 10.30865/mib.v6i3.4218.
- [14] A. Perdana, A. Hermawan, and D. Avianto, “Analisis Sentimen Terhadap Isu Penundaan Pemilu di Twitter Menggunakan Naive Bayes Clasifier,” *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 11, no. 2, pp. 195–200, 2022, doi: 10.32736/sisfokom.v11i2.1412.
- [15] M. Rani, D. Prawira, and N. Mutiah, “Analisis Sentimen Terhadap Vaksin COVID-19 Menggunakan Naive Bayes Classifier, Support Vector Machine dan K-Nearest Neighbors,” *CESS*, vol. 8, no. January, pp. 1–11, 2023.