



---

## PERBANDINGAN METODE NAÏVE BAYES, SUPPORT VECTOR MACHINE DAN RECURRENT NEURAL NETWORK PADA ANALISIS SENTIMEN ULASAN PRODUK E-COMMERCE

Tjut Awaliyah Zuraiyah<sup>1)</sup>, Mulyati<sup>2)</sup>, Gilang Haikal Fiqri Harahap<sup>3)</sup>

<sup>1), 2), 3)</sup> Ilmu Komputer Universitas Pakuan Bogor

Jl. Pakuan, Tegallega. Kecamatan Bogor Tengah, Kota Bogor, Jawa Barat, Indonesia

Email : [tjut.awaliyah@unpak.ac.id](mailto:tjut.awaliyah@unpak.ac.id)

Received: May 02, 2023. Accepted: July 18, 2023

### Abstrak

Analisis sentimen digunakan sebagai alat bantu untuk mendapatkan pendapat dari konsumen atau masyarakat luas. Ulasan produk pada e-commerce memberikan pengaruh pada penjualan produk. Penelitian ini bertujuan untuk melakukan analisis sentimen terhadap ulasan produk pada platform e-commerce menggunakan algoritma Naïve Bayes, Support Vector Machine (SVM), dan Recurrent Neural Network (RNN). Penelitian juga melibatkan tahapan seleksi data, preprocessing, transformasi, data mining, dan evaluasi/interpretasi. Selain itu, penelitian ini juga bertujuan untuk mengatasi masalah imbalanced data yang terjadi antara sentimen positif dan negatif dengan menerapkan teknik oversampling menggunakan library SMOTE. Dengan melakukan penelitian ini, diharapkan dapat memberikan wawasan dan pemahaman yang lebih baik tentang analisis sentimen dan kontribusinya dalam memahami pendapat konsumen serta meningkatkan keputusan pembelian produk. Dalam penelitian ini, dilakukan analisis sentimen terhadap ulasan produk e-commerce menggunakan algoritma Naïve Bayes, SVM, dan RNN. Data opini diklasifikasikan menjadi positif, negatif, atau netral. Terdapat perbedaan jumlah data antara sentimen positif dan negatif (imbalanced data), yang diperlakukan secara berbeda dalam model. Hasil penelitian menunjukkan bahwa Naïve Bayes memiliki akurasi 86%, SVM memiliki akurasi 88%, dan RNN memiliki akurasi 96% dengan epoch 100.

Kata kunci: Analisis Sentimen, Naïve Bayes, SVM, RNN, SMOTE.

### Abstract

*Sentiment analysis serves as a valuable tool for capturing consumer opinions and broader public sentiment. Product reviews posted on e-commerce platforms significantly influence product sales. The objective of this research is to perform sentiment analysis on e-commerce product reviews utilizing Naïve Bayes, Support Vector Machine (SVM), and Recurrent Neural Network (RNN) algorithms. The study encompasses data selection, preprocessing, transformation, data mining, and evaluation/interpretation as crucial stages. Moreover, addressing the issue of imbalanced data, particularly the disparity between positive and negative sentiments, is achieved through the application of oversampling techniques utilizing the SMOTE library. This research aims to enhance the understanding of sentiment analysis, its significance in comprehending consumer opinions, and its role in improving product purchase decisions. The sentiment analysis of e-commerce product reviews was conducted using Naïve Bayes, SVM, and RNN algorithms. The opinions were classified as positive, negative, or neutral. Notably, there is a distinction in the data distribution between positive and negative sentiments (imbalanced data), which necessitates distinct treatment within the models.*

*The findings revealed an accuracy of 86% for Naïve Bayes, 88% for SVM, and 96% for RNN with an epoch of 100.*

*Keyword: Sentiment Analysis, Naïve Bayes, SVM, RNN, SMOTE.*

## PENDAHULUAN

Pasar belanja online diperkirakan terus berkembang seiring dengan meningkatnya akses internet. Faktor kemudahan yang ditawarkan oleh pelaku E-commerce juga mendorong masyarakat untuk melakukan pembelian secara online. Di negara maju, belanja online telah menjadi tren dalam lima tahun terakhir. Melalui akses internet, calon pembeli dapat memperoleh informasi tentang ketersediaan stok barang di berbagai toko dan membandingkan harga antar toko. Hal yang sama juga terjadi di Indonesia, di mana pertumbuhan belanja melalui E-commerce terus meningkat. Pembelian handphone (hp) tidak terbatas pada hp baru, melainkan juga banyak terjadi pembelian hp bekas atau setengah pakai. Hp bekas ini merupakan hp yang telah digunakan sebelumnya dan dijual kembali, atau hp yang dibeli dalam kondisi tidak dalam kemasan pabrikan. Membeli hp bekas menjadi cara untuk mendapatkan spesifikasi yang bagus dengan harga yang lebih terjangkau[1][2].

Analisis sentimen adalah teknik pemrosesan bahasa alami yang digunakan untuk mengidentifikasi dan mengukur sentimen atau pendapat yang terkandung dalam teks. Teknik ini sering digunakan dalam berbagai aplikasi, termasuk pemasaran, media sosial, dan politik, untuk mengevaluasi respons masyarakat terhadap produk, peristiwa, atau individu tertentu. Analisis sentimen melibatkan klasifikasi teks menjadi kategori positif, negatif, atau netral berdasarkan konteks dan nada teks tersebut. Tahapan-tahapan dalam analisis sentimen meliputi seleksi data, transformasi/reduksi data, data mining, dan interpretasi/evaluasi.

Berdasarkan riset Google Temasek, untuk memahami pendapat pembeli terhadap suatu produk yang dijual, diperlukan sistem analisis sentimen berdasarkan ulasan produk yang ada di platform E-commerce seperti Shopee. Penelitian ini bertujuan untuk mengklasifikasikan komentar-

komentar tersebut sebagai positif atau negatif, serta melakukan analisis sentimen menggunakan tiga metode, yaitu Naïve Bayes, Recurrent Neural Network (RNN), dan Support Vector Machine (SVM). Ketiga algoritma tersebut merupakan metode yang unggul dalam analisis sentimen, namun belum ada perbandingan yang cukup untuk menentukan algoritma mana yang paling efektif [3][4][5].

Penelitian sebelumnya yang menggunakan klasifikasi Naïve Bayes dan SVM dalam layanan komplain mahasiswa menunjukkan bahwa algoritma Support Vector Machine memiliki akurasi yang lebih tinggi dibandingkan Naïve Bayes[6]. Penelitian lain juga mengatasi masalah ketidakseimbangan kelas pada dataset penyakit diabetes Pima Indian menggunakan teknik oversampling k-means-SMOTE[7], dengan membandingkan penerapan algoritma C4.5, SVM, dan Naïve Bayes. Penelitian lainnya juga mengusulkan penggunaan RNN[8] untuk mengatasi kelemahan dalam analisis sentimen dengan menggunakan sebagian data yang tersedia. Model tersebut berhasil mengungguli kinerja metrik akurasi dan regresi. Dalam konteks transaksi pembelian hp bekas melalui E-commerce yang semakin banyak, diperlukan alat bantu untuk mendapatkan ulasan produk yang akurat serta mengetahui sentimen terhadap produk yang dijual.

## METODE PENELITIAN

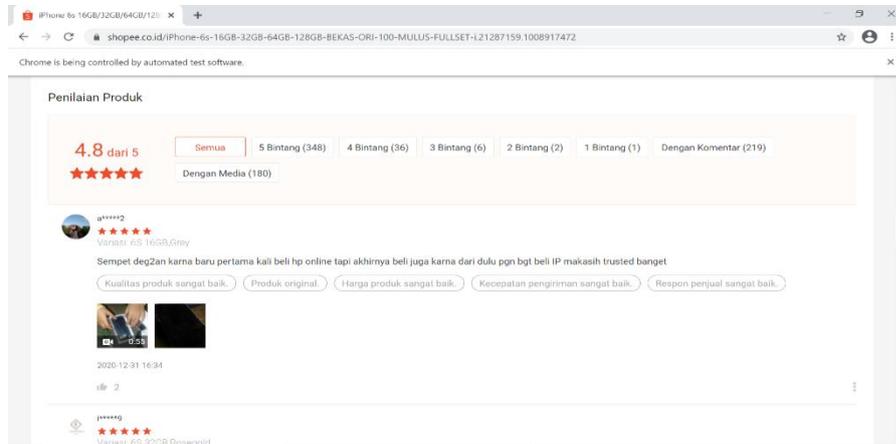
Tahapan penelitian menggunakan metode *Knowledge Discovery and Data Mining* (KDD) yang meliputi tahap: seleksi data, preprocessing, transformasi, data mining dan interpretasi /evaluasi.

### 1. Data Selection

Data yang digunakan pada penelitian ini didapatkan dari hasil scraping pada webpage salah satu E-commerce, yaitu Shopee. Pada proses ini, dipilih

beberapa webpage yang sesuai dengan topik yang akan dianalisis[9]. Pemilihan webpage berdasarkan dengan ulasan terbanyak di setiap produknya, yang kemudian ulasan dapat di scraping dari webpage

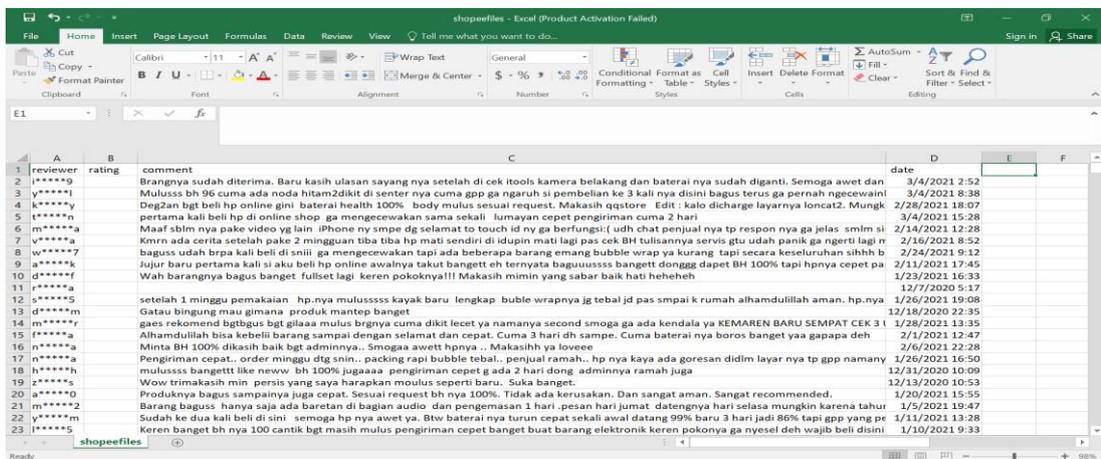
shopee tersebut dan selanjutnya disimpan ke dalam format csv. Proses scraping data dapat dilihat pada Gambar 1.



Gambar 1. Proses Scraping Pada Data Selection.

Setelah proses scraping dilakukan, selanjutnya data disimpan ke dalam format csv yang dapat dilihat pada Gambar 2. Penyeleksian data yang dilakukan pada penelitian ini yaitu dengan scraping data

ulasan produk hp bekas yang ada pada E-Commerce Shopee, yang kemudian akan disimpan dengan format csv agar mempermudah pemanggilan data pada saat proses klasifikasi.



Gambar 2. Hasil Scraping Ulasan HP Bekas di Shopee.

2. Preprocessing

Sebelum melakukan pembobotan kata, dilakukan tahap pre-processing dengan tujuan mengubah data yang tidak terstruktur menjadi data yang terstruktur. Data yang digunakan dalam penelitian ini berasal dari ulasan produk elektronik hp bekas di platform e-commerce Shopee. Tahap pre-processing terdiri dari beberapa langkah, yaitu case folding, tokenizing, normalization, stop word removal, dan stemming.

Case folding adalah proses mengubah semua karakter menjadi huruf kecil. Tujuan dari proses ini adalah agar seluruh dokumen tetap konsisten menggunakan huruf kecil. Tokenization digunakan untuk membersihkan karakter-karakter tertentu, seperti tanda baca, dan mengubah kata-kata menjadi potongan string. Proses normalisasi dilakukan untuk mengubah kata-kata yang tidak baku menjadi bentuk yang baku. Stop word removal adalah tahap di mana kata-kata yang tidak

memiliki makna dihilangkan. Sedangkan stemming merupakan proses pemetaan atau pengurangan kata-kata menjadi kata dasarnya.

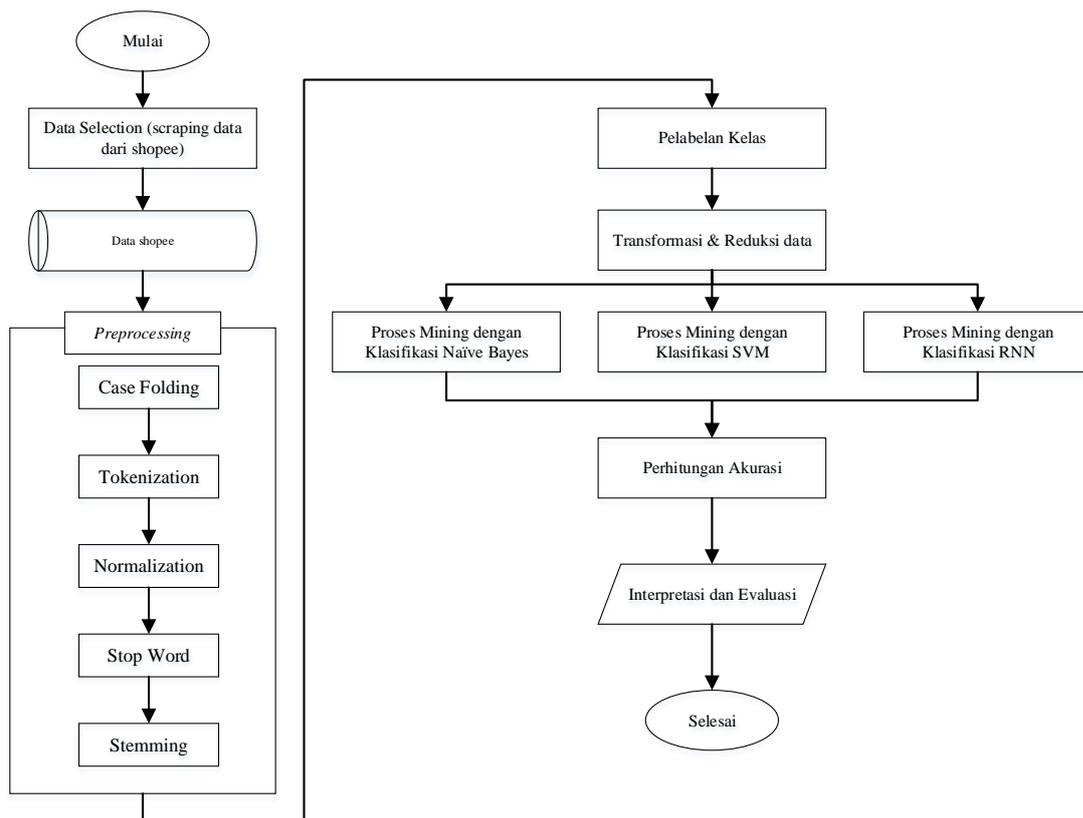
### 3. Transformasi Data & Reduksi Data

Proses transformasi dan reduksi data dilakukan dengan membersihkan data yang mengandung noise pada baris, atribut, serta menghapus data yang berulang. Transformasi dan reduksi data ini melibatkan penghapusan atribut yang tidak relevan, yaitu rating, serta menghilangkan data yang kosong dengan nilai NaN.

### 4. Data Mining

Tahap ini melibatkan implementasi algoritma klasifikasi untuk menemukan pola dan memprosesnya sehingga menghasilkan pengetahuan yang berguna. Dalam penelitian ini, proses mining data dilakukan dengan menerapkan

algoritma Naïve Bayes untuk menganalisis sentimen pada ulasan hp bekas di marketplace Shopee. Sentimen positif ditentukan oleh kalimat yang mengandung kata-kata seperti bagus, memuaskan, lancar, normal, dan cepat. Sedangkan sentimen negatif ditentukan oleh kalimat yang mengandung kata-kata seperti rusak, mati, buruk, kecewa, dan lambat. Proses pemilihan data hingga klasifikasi dapat dilihat pada Gambar 3.



Gambar 3. Flowchart Klasifikasi Naïve Bayes, Support Vector Machine, dan Recurrent Neural Network.

Data yang didapat dari proses scraping pada webpage shopee disimpan ke dalam format csv untuk memudahkan pengambilan data. Kemudian

akan dilakukan tahap preprocessing sampai dengan proses klasifikasi menggunakan Naïve Bayes,

Support Vector Machine, dan Recurrent Neural Network.

#### 5. Naïve Bayes

Algoritma Naïve Bayes adalah metode statistik yang digunakan untuk melakukan klasifikasi dan memprediksi probabilitas keanggotaan dalam suatu kelas. Dalam proses klasifikasi teks, metode Naïve Bayes melibatkan dua tahap utama, yaitu tahap pelatihan dan tahap klasifikasi. Pada tahap pelatihan, dilakukan analisis terhadap sampel dokumen dengan memilih kosakata yang mungkin muncul dalam kumpulan dokumen sampel sehingga dapat mewakili dokumen secara keseluruhan. Kemudian, probabilitas untuk setiap kategori ditentukan berdasarkan sampel dokumen tersebut. Pada tahap klasifikasi, nilai kategori dari suatu dokumen ditentukan berdasarkan kemunculan term (kata) yang terdapat dalam dokumen yang sedang diklasifikasikan[6].

#### 6. Support Vector Machine

SVM adalah metode klasifikasi yang digunakan untuk data linear maupun nonlinier. Secara sederhana, SVM adalah algoritma yang menggunakan pemetaan nonlinier untuk mengubah data pelatihan awal menjadi dimensi yang lebih tinggi. Di dimensi baru ini, SVM mencari hyperplane yang dapat memisahkan secara linear antara tuple-tuple dari satu kelas dengan kelas yang lain. Dengan menggunakan pemetaan nonlinier yang sesuai ke dimensi yang lebih tinggi, data dari dua kelas selalu dapat dipisahkan oleh hyperplane. SVM berupaya menemukan hyperplane dengan menggunakan vektor dukungan, yaitu tuple-tuple pelatihan yang penting, serta margin yang ditentukan oleh vektor dukungan tersebut.

#### 7. Recurrent Neural Network

RNN memproses data input dari inputan satu sampai inputan ke-n satu persatu secara sekuensial. Setiap inputan akan diteruskan dari hidden layer satu ke hidden layer selanjutnya sampai pada hasil output. RNN[10] bekerja dengan proses looping. RNN tidak membuang informasi sebelumnya namun menyimpan dan memproses pada hidden

layer[11]. Pada umumnya, jaringan syaraf tiruan terdiri dari 3 layer, yaitu input layer, hidden layer, dan output layer. Pada penelitian analisis sentimen input layer berupa kumpulan dari sequence of words. Pada jaringan syaraf tiruan yang hanya memiliki 1 hidden layer, jumlah nodes pada input layer bergantung pada dimensi data. Hubungan nodes dari input layer dengan hidden layer disebut dengan "synapses"[12]. Sedangkan, hubungan antara setiap dua nodes dari input layer ke hidden layer memiliki koefisien yang disebut dengan bobot (weight), yang merupakan sinyal untuk membuat keputusan. Selanjutnya, proses pelatihan dilanjutkan dengan menyesuaikan bobotnya sampai optimal[12]. Lalu, nodes yang ada pada hidden layer yang berasal dari input layer menerapkan fungsi sigmoid atau tangen (tanh) pada jumlah bobot yang berasal dari input yang disebut dengan aktivasi. Transformasi ini akan menghasilkan nilai untuk proses pelatihan dan evaluasi dengan menggunakan fungsi Softmax. Nilai dari hasil transformasi akan terlihat pada output layer, namun nilai yang dihasilkan belum optimal sehingga proses backpropagation diterapkan untuk menargetkan nilai error yang optimal, proses ini juga disebut sebagai loss validation. Backpropagation menghubungkan hidden layer dengan output layer dengan cara mengirimkan sinyal yang menunjukkan bobot dan nilai error yang optimal dengan jumlah epochs yang telah ditentukan[12].

#### 8. Interpretasi dan Evaluasi

Tahap evaluasi dilakukan dengan menggunakan confusion matrix untuk mendapatkan informasi apakah hasil yang diperoleh sudah sesuai dengan harapan. Hal ini bertujuan untuk memudahkan proses pengambilan keputusan dan penemuan pengetahuan yang signifikan. Dalam penelitian ini, interpretasi atau visualisasi hasil evaluasi ditampilkan dalam bentuk grafik yang menunjukkan nilai akurasi dari data latih dan data uji untuk ketiga algoritma yang dibandingkan. Hubungan antara aktual data dengan hasil prediksi dapat dilihat pada Tabel 1.

Tabel 1. *Confusion Matrix*

Aktual data	Prediksi Klasifikasi	
	Positif	Negatif
Positif	TP	FN
Negatif	FP	TN

Pada Tabel 1 terdapat *True Positive* (TP) merupakan hasil yang benar pada aktual data diprediksi benar oleh sistem, *True Negative* (TN) merupakan hasil yang salah pada aktual data diprediksi salah oleh sistem, *False Positive* (FP) merupakan hasil yang benar pada aktual data diprediksi salah oleh sistem, *False Negative* (FN) merupakan hasil yang salah pada aktual data diprediksi benar oleh sistem. Dapat dihitung nilai accuracy, precision, recall dan F-1 score menggunakan rumus dibawah ini:

$$Accuracy = (TP+TN) / (TP+FP+FN+TN) \tag{1}$$

$$Precision = (TP) / (TP + FP) \tag{2}$$

$$Recall = TP / (TP + FN) \tag{3}$$

$$F-1 \text{ Score} = (2 * Recall * Precision) / (Recall + Precision) \tag{4}$$

## HASIL DAN PEMBAHASAN

Berdasarkan langkah-langkah yang telah diimplementasikan, terdapat beberapa tantangan yang dihadapi. Saat melakukan proses scraping, tidak semua komentar pada setiap halaman produk dapat diambil secara otomatis. Oleh karena itu, dalam penelitian ini dipilih 14 halaman produk dengan jumlah ulasan tertinggi untuk dilakukan proses scraping. Selama proses scraping berlangsung, dilakukan encoding simbol dan emotikon secara otomatis. Selanjutnya, pada tahap preprocessing, library Sastrawi tidak berfungsi sepenuhnya, sehingga banyak kata yang harus

dimasukkan secara manual untuk proses stop word dan normalisasi. Tahap preprocessing melibatkan case folding, tokenisasi, normalisasi, stop word, dan stemming. Proses-proses ini bertujuan untuk menghasilkan model klasifikasi yang optimal untuk data yang akan digunakan. Dalam penelitian ini, akan dilakukan klasifikasi sentimen dengan dua kelas, yaitu positif dan negatif. Selanjutnya, karena ketidakseimbangan data, di mana jumlah sampel kelas positif lebih banyak daripada kelas negatif, akan dilakukan proses oversampling. Setelah proses oversampling selesai, model akan dilatih menggunakan tiga algoritma, yaitu Naïve Bayes, Support Vector Machine, dan Recurrent Neural Network.

### 1. Data Selection

Data ulasan produk hp bekas diambil dari salah satu e-commerce yaitu Shopee. Data yang berhasil di scraping sebanyak 1934 data, yang berisi sentiment positif dan negatif. Data tersebut masih uncleaned atau kotor, sehingga belum dapat dilakukan proses mining.

### 2. Preprocessing

#### a. Case Folding

Tahap pertama yang dilakukan pada preprocessing adalah case folding, yaitu mengubah huruf capital menjadi huruf kecil yang dapat dilihat pada Gambar 4.

```
In [18]: df['comment'] = df['comment'].str.lower()

print('Case Folding Result : \n')
print(df['comment'].head(5))

Case Folding Result :

0    barangnya sudah diterima. baru kasih ulasan say...
1    mulusss bh 96 cuma ada noda hitam2dikit di sen...
2    deg2an bgt beli hp online gini baterai health...
5    pertama kali beli hp di online shop ga mengec...
6    maaf sbml nya pake video yg lain iphone ny sm...
Name: comment, dtype: object
```

Gambar 4. Tahap Case Folding.

Dilakukannya case folding karena jika terdapat 2 kata yang sama namun salah satu kalimat tersebut menggunakan huruf kapital dianggap sebagai salah penulisan oleh sistem[13].

Sebuah teks dokumen terdiri dari beberapa kata, proses tokenisasi dilakukan dengan cara memecah dokumen menjadi beberapa bagian dari kata-kata yang biasa disebut sebagai token. Proses tokenisasi dapat dilihat pada Gambar 5.

b. Tokenization

```
Tokenizing Result :

0    [barangnya, sudah, diterima, baru, kasih, ulasa...
1    [mulusss, bh, cuma, ada, noda, hitamdikit, di,...
2    [degan, bgt, beli, hp, online, gini, baterai, ...
5    [pertama, kali, beli, hp, di, online, shop, ga...
6    [maaf, sbml, nya, pake, video, yg, lain, iphon...
Name: comment_token, dtype: object
```

Gambar 5. Tahap Tokenization.

Tahap tokenisasi berfungsi untuk melakukan encode dokumen seperti penghapusan emoticon, link, tanda baca, dan nomor.

menyisipkan file excel yang berisi kata-kata yang baku, sehingga jika terdapat kata-kata yang tidak baku pada dokumen akan otomatis terganti. Proses normalisasi dapat dilihat pada Gambar 6.

c. Normalization

Normalisasi dilakukan untuk mengubah kalimat yang tidak baku menjadi baku dengan cara

```
df['comment_normalized'].head(10)

C:\Users\ASUS\Anaconda3\lib\site-packages\ipykernel_launcher.py:12: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-vs-copy
rsus-a-copy
if sys.path[0] == '':

Out[21]: 0    [barangnya, sudah, diterima, baru, kasih, ulas...
1    [mulus, battery health, cuma, ada, noda, hitam...
2    [dengan, banget, beli, hp, online, gini, bater...
5    [pertama, kali, beli, hp, di, online, shop, ti...
6    [maaf, sebelum, nya, pake, video, yang, lain, ...
7    [kemarin, ada, cerita, setelah, pake, mingguan...
8    [bagus, udah, berapa, kali, beli, di, sini, ti...
9    [jujur, baru, pertama, kali, si, aku, beli, hp...
10   [wah, barangnya, bagus, banget, fullset, lagi,...
12   [setelah, minggu, pemakaian, hpnya, mulussss,...
Name: comment_normalized, dtype: object
```

Gambar 6. Tahap Normalization.

Terdapat 239 kata baku pada file excel untuk menjadi parameter jika terdapat kata yang tidak baku pada dokumen akan otomatis terganti.

Tahap ini dilakukan untuk menghilangkan kata-kata yang tidak penting atau tidak berpengaruh kepada sentimen, contohnya yaitu imbuhan yang terdapat pada kata-kata. Proses stop word menggunakan library Sastrawi pada python, serta menginputkan

d. Stop Word

manual kata-kata yang tidak penting. Tahap stop word dapat dilihat pada Gambar 7.

```
In [22]: from Sastrawi.StopWordRemover.StopWordRemoverFactory import StopWordRemoverFactory
more_stopwords=['yang', 'nya', 'di', 'cuma', 'untuk', 'nih', 'sih',
                'gini', 'dikit', 'si', 'kalau', 'gak', 'gapapa', 'kali', 'gitu', 'kalo', 'kak', 'dsh', 'yaa', 'ya', 'ih', 'an']
factory = StopWordRemoverFactory()
stopwords = factory.get_stop_words()+more_stopwords

def stopwords_removal(words):
    return [word for word in words if word not in stopwords]

df['comment_filtered'] = df['comment_normalized'].apply(stopwords_removal)

print(df['comment_filtered'].head())

0      [barangnya, diterima, baru, kasih, ulasan, say...
1      [mulus, battery health, noda, hitam, senter, g...
2      [banget, beli, hp, online, baterai, health, bo...
3      [pertama, beli, hp, online, shop, mengecewakan...
4      [maaf, pake, video, iphone, selamat, to, touch...
5      [maaf, pake, video, iphone, selamat, to, touch...
6      [maaf, pake, video, iphone, selamat, to, touch...
Name: comment_filtered, dtype: object

C:\Users\ASUS\Anaconda3\lib\site-packages\ipykernel_launcher.py:12: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
if sys.path[0] == '':
```

Gambar 7. Tahap Stop Word.

e. Stemming

Tahap stemming dilakukan untuk mengubah kata ke bentuk dasarnya. Tahap ini merupakan proses

yang penting di dalam tahapan preprocessing, karena hasil stemming yang baik dapat mempengaruhi baik atau tidaknya text mining[13]. Proses stemming dapat dilihat pada Gambar 8.

```
print(len(term_dict))
print("-----")

for term in term_dict:
    term_dict[term] = stemmed_wrapper(term)
    print(term, ":", term_dict[term])

print(term_dict)
print("-----")

# apply stemmed term to dataframe
def get_stemmed_term(document):
    return [term_dict[term] for term in document]

df['comment_stemmed'] = df['comment_filtered'].swifter.apply(get_stemmed_term)
print(df['comment_stemmed'])

2607
-----
barangnya : barang
diterima : terima
baru : baru
kasih : kasih
ulasan : ulas
sayang : sayang
cok : cok
itools : itools
kamera : kamera
```

Gambar 8. Tahap Stemming.

Sebanyak 2607 kata telah diubah ke bentuk dasarnya. Proses stemming ini dilakukan dengan menggunakan library Sastrawi pada python. Library ini merupakan library text mining untuk proses penyederhanaan kata yang menggunakan bahasa Indonesia.

f. Pelabelan Kelas

Proses pelabelan data dilakukan dengan cara memberikan score pada setiap dokumen. Pembobotan kata dilakukan dengan menghitung frekuensi kemunculan kata pada sebuah dokumen teks. Frekuensi kemunculan kata di dalam dokumen yang diberikan menunjukkan seberapa penting kata itu di dalam dokumen tersebut.

Cara menentukan kelas sentimen adalah dengan menghitung skor jumlah kata positif dikurangi skor

jumlah kata negatif dalam setiap kalimat ulasan. Kalimat yang memiliki skor > 0 akan diklasifikasikan ke dalam kelas positif, sedangkan kalimat yang memiliki skor < 0 diklasifikasikan ke dalam kelas negatif. Contoh pelabelan corpus dapat dilihat pada Tabel 6. Contoh D1 – D8 hasil pemeriksaan corpus:

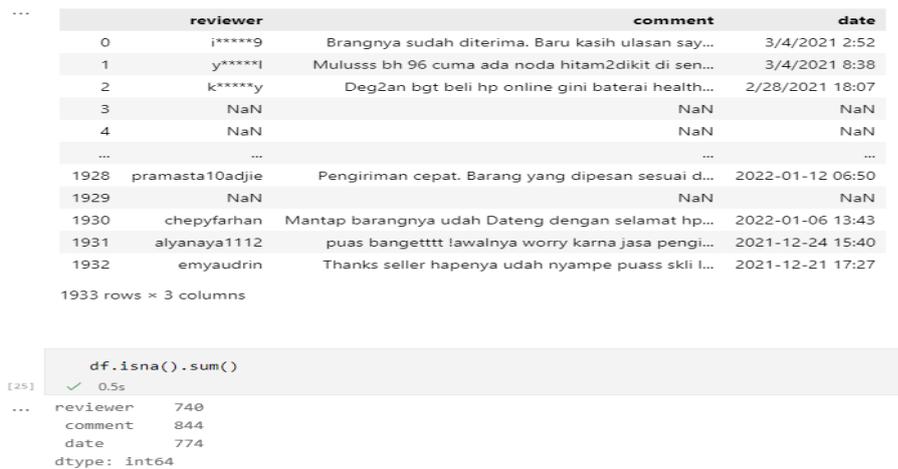
- D1. bagus sesuai cepat : (1) - (1) - (1), skor = 3
- D2. bagus baru : (1) - (1), skor = 2
- D3. rekomen bagus sesuai aman : (1) - (1) - (1) - (1), skor = 4
- D4. aman rusak tipu : (1) - (-1) - (-1), skor = -1
- D5. baik lama rusak lecet mulus boros bagus : (1) - (-1) - (-1) - (-1) - (1) - (-1) - (1), skor = -4
- D6. kecewa rusak : (-1) - (-1), skor = -2
- D7. bagus baru rekomen jelek buruk :- (1) - (1) - (-1) - (-1), skor = 1
- D8. Rusak : (-1), skor = -1

Tabel 2. Pelabelan Kelas

Kode	Label Kelas
D1	Positif
D2	Positif
D3	Positif
D4	Negatif
D5	Negatif
D6	Negatif
D7	Positif
D8	Negatif

### 3. Transformasi Data & Reduksi Data

Proses transformasi dan reduksi data dilakukan dengan menghapus cell yang kosong atau disebut dengan NaN, serta atribut yang tidak relevan. Hal ini dilakukan untuk memperkecil volume dokumen yang akan dilakukan analisis sentiment. Proses transformasi dan reduksi data dapat dilihat pada Gambar 9.



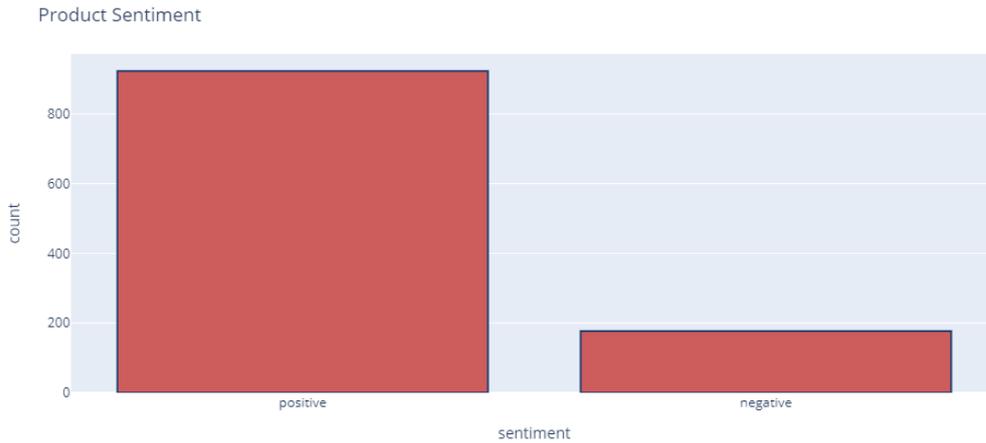
Gambar 9. Proses Transformasi dan Reduksi Data.

Berikut ditampilkan summary dari data yang kosong atau missing values. Kemudian pada kolom comment dan date hanya menggunakan data yang terisi saja, sehingga data yang siap digunakan untuk proses selanjutnya berjumlah 1100 data.

### 4. Data Mining

Pada tahap mining dilakukan perbandingan 3 algoritma, yaitu Naïve Bayes, Support Vector Machine, dan Recurrent Neural Network. Ketiga

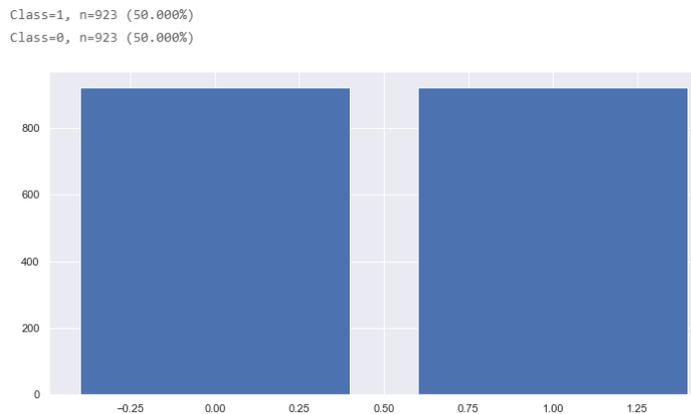
algoritma ini digunakan untuk melakukan analisis sentiment terhadap ulasan produk hp bekas di shopee. Dataset didapatkan dari proses scraping pada shopee yang sudah melalui tahap preprocessing, label sentimen yang digunakan yaitu positif dan negatif.



Gambar 10. Jumlah Kelas Sentimen.

Berdasarkan Gambar 10, jumlah sentimen positif berjumlah 923 data, dan sentimen negatif berjumlah 177 data. Jumlah tersebut menunjukkan terdapat ketidakseimbangan kelas yang akan berdampak pada model. Hal ini dapat mempengaruhi proses pelatihan dan pengujian model, namun hal ini dapat

diatasi dengan proses oversampling dengan library SMOTE[14]. Jumlah kelas mayoritas terdapat pada sentimen positif, maka kelas minoritas akan disamakan jumlahnya dengan kelas mayoritas. Proses ini disebut juga sebagai oversampling yang dapat dilihat pada Gambar 11.



Gambar 11. Proses Oversampling untuk Metode Naïve Bayes dan Support Vector Machine.

Proses oversampling untuk metode Recurrent Neural Network berbeda dengan dua metode lainnya. Hal ini karena metode Recurrent Neural Network menggunakan sequence of words untuk

proses inputnya, sedangkan dua metode lainnya menggunakan feature extraction yakni Count Vectorizer. Hasil oversampling untuk RNN dapat dilihat pada Gambar 12.

```

kelas mayoritas sebelum oversampling: (738, 2)
kelas minoritas sebelum oversampling: (142, 2)
After oversampling
1    738
0    738
Name: label, dtype: int64
    
```

Gambar 12. Proses Oversampling untuk Metode Recurrent Neural Network.

a. Pemanggilan Library

Proses pemanggilan library digunakan untuk memudahkan ketika menjalankan sebuah proses

yang dibutuhkan dalam program yang dibuat. Proses pemanggilan library dapat dilihat pada Gambar 13.

```
In [1]: from sklearn.naive_bayes import MultinomialNB
from sklearn.model_selection import train_test_split, KFold, cross_val_score
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
from sklearn.svm import SVC
import nltk
from keras.models import Sequential
from keras.layers import Dense, Embedding, Conv1D, MaxPooling1D, Dropout, Flatten
from keras.preprocessing.text import Tokenizer
from keras.preprocessing.sequence import pad_sequences
from keras.utils.np_utils import to_categorical
from sklearn.preprocessing import LabelEncoder

C:\Users\ASUS\Anaconda3\lib\site-packages\statsmodels\tools\_testing.py:19: FutureWarning: pandas.util.testing is deprecated. Use the functions in the public API at pandas.testing instead.
import pandas.util.testing as te
```

Gambar 13. Pemanggilan *Library*.

b. Proses Word Embedding dan Count Vectorizer

Proses word embedding dilakukan untuk mengkonversikan kata yang berupa alphanumeric ke dalam bentuk vector [10], word embedding

digunakan pada algoritma recurrent neural network. Pada algoritma Naïve Bayes dan Support Vector Machine digunakan Count Vectorizer yang mempunyai kegunaan yang sama dengan Word Embedding. Proses keduanya dapat dilihat pada Gambar 14

```
In [18]: from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer()
X = cv.fit_transform(X).todense()
y = y
```

Gambar 22. Proses *Count Vectorizer*.

```
In [37]: tokenizer = Tokenizer()
tokenizer.fit_on_texts(data['comment_normal'].values)

In [38]: max_kata = 100
X = tokenizer.texts_to_sequences(data['comment_normal'])
X = pad_sequences(X, maxlen = max_kata, padding = 'post')
X

Out[38]: array([[ 33, 202, 14, ..., 0, 0, 0],
 [ 4, 3, 2, ..., 0, 0, 0],
 [ 1, 9, 10, ..., 0, 0, 0],
 ...,
 [ 33, 69, 546, ..., 0, 0, 0],
 [ 40, 34, 18, ..., 0, 0, 0],
 [117, 14, 22, ..., 0, 0, 0]])
```

Gambar 14. Proses *Word Embedding*.

c. Proses Pembagian Data Latih dan Data Uji

Pembagian data latih dan data uji dilakukan dengan mengimplementasikan 20% data uji dan 80% data

latih serta 25% data uji dan 75% data latih. Pembagian ini dilakukan setelah menyeimbangkan kelas dengan library SMOTE. Proses pembagian dapat dilihat pada Gambar 15.

```
In [22]: X_train, X_test, y_train, y_test = train_test_split(X_SMOTE, y_SMOTE, test_size=0.2)
```

Gambar 15. Pembagian Data Latih dan Data Uji.

d. Pembuatan Model Naïve Bayes, Support Vector Machine, dan Recurrent Neural Network

Dalam proses ini, model Naïve Bayes dibuat dengan menggunakan jenis classifier multinomial, kemudian model Support Vector Machine dibuat

dengan menggunakan kernel linear. Selanjutnya, model Recurrent Neural Network (RNN) dibuat dengan menggunakan long short term memory (LSTM). Hasil dari proses word embedding dimasukkan ke dalam layer LSTM, dan dari layer

LSTM, keluaran diteruskan ke output layer. Proses pembuatan model ini dapat dilihat pada Gambar 16, 17, dan 18.

```
In [23]: classifierNB = MultinomialNB()
classifierNB.fit(X_train, y_train)
clf_NB = classifierNB.predict(X_train)
NB_train = accuracy_score(y_train, clf_NB)
```

Gambar 16. Pembuatan Model Naïve Bayes.

```
In [25]: from sklearn.svm import SVC

In [26]: classifierSVM = SVC(kernel='linear')
classifierSVM.fit(X_train, y_train)
clf_SVM = classifierSVM.predict(X_train)
SVM_train = accuracy_score(y_train, clf_SVM)
```

Gambar 17. Pembuatan Model Support Vector Machine.

```
In [61]: embed_dim = 128
lstm_out = 192

model = Sequential()
model.add(Embedding(max_features, embed_dim, input_length = X_train_rnn.shape[1]))
model.add(SpatialDropout1D(0.4))
model.add(LSTM(lstm_out, dropout=0.4, recurrent_dropout=0.4))
model.add(Dense(2, activation='softmax'))
model.compile(loss = 'binary_crossentropy', optimizer='adam', metrics = ['accuracy'])
print(model.summary())

Model: "sequential_1"
Layer (type) Output Shape Param #
-----
embedding_1 (Embedding) (None, 29, 128) 512000
spatial_dropout1d_1 (Spatial (None, 29, 128) 0
lstm_1 (LSTM) (None, 192) 246528
dense_1 (Dense) (None, 2) 386
-----
Total params: 758,914
Trainable params: 758,914
Non-trainable params: 0
None
```

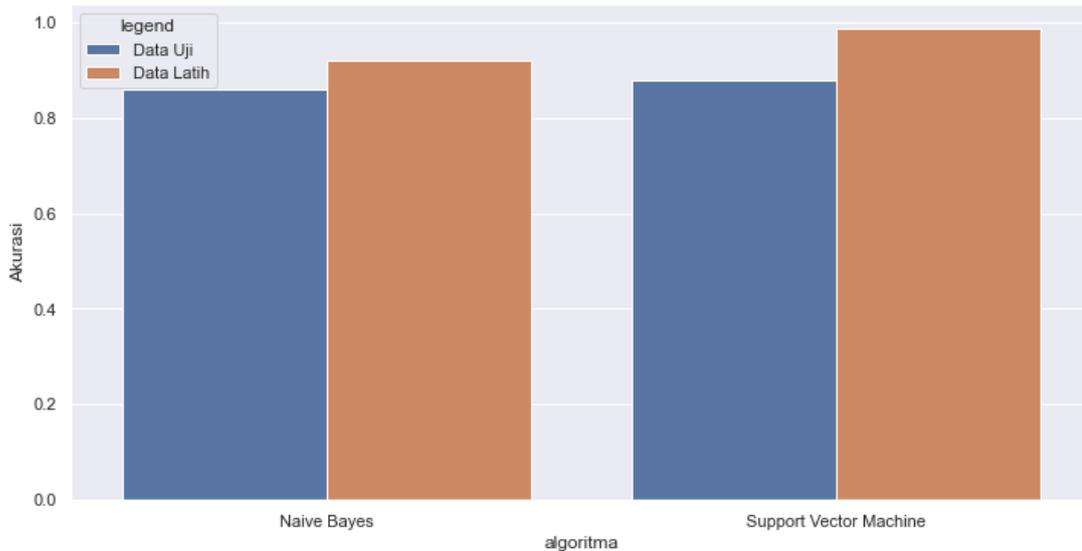
Gambar 18. Pembuatan Model Recurrent Neural Network.

Neural Network dilatih dengan menggunakan gradient stochastic yang mengharuskan untuk memilih loss function ketika merancang atau mengatur model yang akan dilatih.

### 5. Perbandingan Algoritma

Dalam proses pengujian model, setelah mencapai keseimbangan kelas dengan jumlah sentiment positif dan negatif yang sama, dataset akan dibagi menjadi data latih dan data uji untuk diterapkan pada model. Model akan dibangun menggunakan

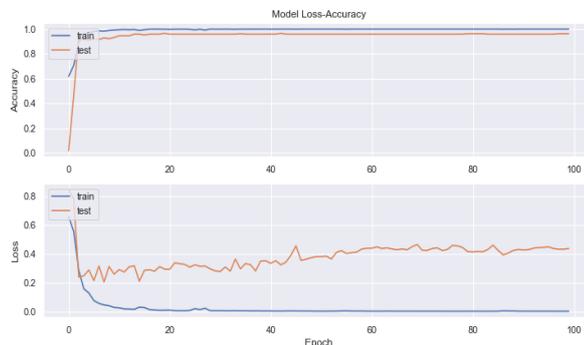
data latih yang terdiri dari 1846 data, kemudian model akan diuji menggunakan data uji yang berjumlah 370 data. Pada proses ini, akan digunakan tiga algoritma yaitu Naïve Bayes, Support Vector Machine, dan Recurrent Neural Network untuk membandingkan algoritma mana yang menghasilkan tingkat akurasi tertinggi.



Gambar 19. Nilai akurasi Algoritma Naïve Bayes dan Support Vector Machine.

Berdasarkan ilustrasi pada Gambar 19, dengan pembagian data latih sebesar 80% dan data uji sebesar 20%, Naïve Bayes mencapai tingkat akurasi sebesar 92% pada data latih dan 86% pada data uji. Sementara itu, Support Vector Machine mencapai tingkat akurasi sebesar 99% pada data latih dan

88% pada data uji. Grafik RNN untuk hasil akurasi tidak dapat digabungkan dengan Algoritma Naïve Bayes dan Support Vector Machine karena RNN melatih modelnya secara sekuen seperti yang terlihat dalam kode sumber pada Gambar 20.



Gambar 20. Nilai Akurasi dan Loss Function Recurrent Neural Network.

Berdasarkan Gambar 20, pembagian data latih 80% dan data uji 20% serta epochs yang digunakan adalah sebanyak 100 kali dan mendapatkan nilai akurasi tertinggi mencapai 99% pada data latih, dan 96% pada data uji. Kemudian terdapat loss function yang berperan sangat penting dalam neural network yang berfungsi untuk mengukur seberapa bagus performa dari neural network dalam melakukan prediksi terhadap target. Loss function bekerja

ketika model pembelajaran memberikan kesalahan yang perlu diperhatikan. Loss function yang baik adalah fungsi yang menghasilkan error yang diharapkan paling rendah[15]. Pada penelitian ini dapat dilihat data latih memiliki loss function yang rendah dibandingkan data uji.

Tabel 3. Hasil Akurasi Data Latih dan Data Uji.

Nilai Akurasi	Algoritma					
	Naïve Bayes		Vector Machine		Reccurent Neural Network	
	Data latih 75%, Data Uji 25%	Data latih 80%, Data Uji 20%	Data latih 75%, Data Uji 25%	Data latih 80%, Data Uji 20%	Data latih 75%, Data Uji 25%, Epoch 100	Data latih 80%, Data Uji 20%, Epoch 100
Data Latih	78%	92%	98%	99%	98%	99%
Data Uji	71%	86%	88%	88%	96%	96%

Berdasarkan hasil pada tabel 3, nilai akurasi terbaik pada Naïve Bayes sebesar 92% data latih dan 86% data uji dengan pembagian data latih 80% dan data uji 20%. Nilai akurasi terbaik pada Support Vector Machine sebesar 99% data latih dan 88% data uji dengan pembagian data latih 80% dan data uji 20%.

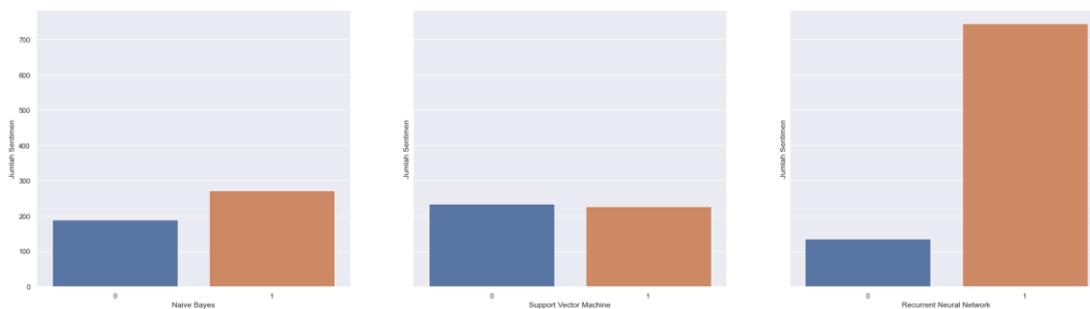
Nilai akurasi terbaik pada Reccurent Neural Network sebesar 99% data latih dan 96% data uji dengan pembagian data latih 80% dan data uji 20% dan epoch 100.

Tabel 4. Hasil Akurasi Data Latih dan Data Uji Metode Reccurent Neural Network Dengan Berbagai Epoch.

Nilai Akurasi	Reccurent Neural Network					
	Data Latih 80% Data Uji 20%			Data Latih 75% Data Uji 25%		
	Epoch 20	Epoch 50	Epoch 100	Epoch 20	Epoch 50	Epoch 100
Data Latih	99%	99%	99%	99%	99%	98%
Data Uji	96%	96%	96%	95%	96%	96%

Berdasarkan hasil pada tabel 4, nilai akurasi data latih dan data uji untuk metode Reccurent Neural Network pada penelitian ini menggunakan epoch sebesar 20, 50 dan 100. Pada pembagian data latih 80% dan data uji 20%, epoch 20, 50 dan 100

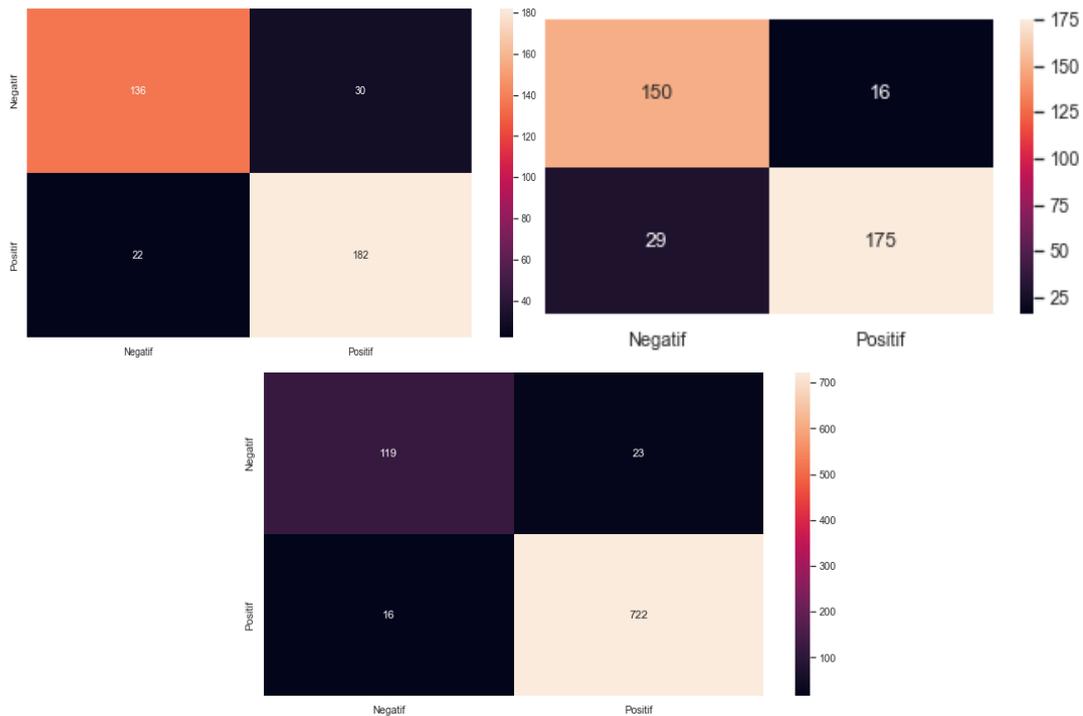
mendapatkan nilai akurasi yang sama yaitu data latih 99% dan data uji 96%. Pada pembagian data latih 75% dan data uji 25%, epoch 50 mendapatkan nilai akurasi terbaik yaitu data latih 99% dan data uji 96%.



Gambar 21. Hasil Analisis Sentimen Ketiga Algoritma.

Gambar 21 merupakan hasil prediksi sentimen dengan menggunakan data uji, dapat dilihat sentimen positif yang ditunjukkan dengan angka 1 dapat diprediksi lebih banyak pada algoritma Reccurent Neural Network dibandingkan dengan

kedua algoritma lainnya. Untuk melihat seberapa banyak yang terklasifikasi dan termisklasifikasi dapat menggunakan confusion matrix dapat dilihat pada Gambar 22.



Gambar 22. Confusion Matrix Naïve Bayes (kiri atas), Support Vector Machine (kanan atas), dan Recurrent Neural Network.

Confusion matrix merupakan pilihan yang tepat untuk melihat performa model karena dapat mengamati hubungan hasil prediksi dan data aktualnya. Model yang memiliki true positive tertinggi yaitu Recurrent Neural Network sebanyak 722 data dapat diprediksi benar oleh sistem.

Dengan menggunakan persamaan (2)(3) dan (4) dihasilkan tabel 3 perhitungan precision, recall dan F1-score untuk *Naïve Bayes*, SVM dan RNN.

Tabel 5. Hasil Precision, Recall, dan F1-score.

Sentimen	Algoritma								
	<i>Naïve Bayes</i>			<i>Support Vector Machine</i>			<i>Recurrent Neural Network</i>		
	<i>Precision</i>	<i>Recall</i>	<i>F1-score</i>	<i>Precision</i>	<i>Recall</i>	<i>F1-score</i>	<i>Precision</i>	<i>Recall</i>	<i>F1-score</i>
0	86%	82%	84%	84%	90%	87%	88%	84%	86%
1	86%	89%	88%	92%	86%	89%	97%	98%	97%

Berdasarkan hasil pada Tabel 3, pembagian data latih 80% dan data uji 20% serta epoch 100 untuk metode Recurrent Neural Network. Precision merupakan ketepatan antara data yang diminta dengan prediksi yang diberikan oleh model. Recall menggambarkan keberhasilan model dalam menemukan kembali sebuah informasi. Pada kelas negatif yang ditandai dengan angka 0, Support Vector Machine merupakan model dengan

performance paling baik untuk sentimen negative yakni mendapatkan precision sebesar 84%, recall sebesar 90%, dan f1 score sebesar 87%. Sedangkan untuk sentimen positif yang ditandai dengan angka 1, Recurrent Neural Network mendapatkan precision sebesar 97%, recall sebesar 98% dan f1 score sebesar 97% menjadikan model Recurrent Neural Network memiliki performance yang baik untuk sentimen positif.

## KESIMPULAN

Penelitian ini fokus pada membandingkan metode klasifikasi Naïve Bayes, SVM, dan RNN. Untuk meningkatkan akurasi, dilakukan balancing data pada Naïve Bayes dan SVM menggunakan library SMOTE, yang menciptakan data sintetik berdasarkan data terdekat dengan sentimen. Sementara itu, pada RNN dilakukan oversampling dengan resampling sentimen. Proses pelatihan model dengan menggunakan data latih menghasilkan tingkat akurasi yang tinggi.

Hasil analisis sentimen menunjukkan bahwa Naïve Bayes mencapai akurasi terbaik pada pembagian data latih 80% dan data uji 20% dengan nilai 86%. SVM mencapai akurasi terbaik pada pembagian yang sama dengan nilai 88%. Sementara itu, RNN mencapai akurasi terbaik pada pembagian data latih 80% dan data uji 20% dengan 100 epoch, mencapai nilai 96%. SVM menunjukkan kinerja yang baik untuk sentimen negatif, sementara RNN menunjukkan kinerja yang baik untuk sentimen positif, hal ini dapat dilihat dari nilai f1-score yang dievaluasi.

Saran untuk penelitian selanjutnya, pada tahap *preprocessing* bisa menambahkan kata-kata yang akan dijadikan parameter pada proses normalisasi sehingga bisa meningkatkan nilai akurasi dan dapat memprediksi dengan langsung melalui inputan *keyword*. Tahap interpretasi dapat ditingkatkan lagi dengan membuat *Graphic User Interface* agar terlihat lebih interaktif.

## UCAPAN TERIMA KASIH

Kami ingin mengucapkan terima kasih kepada yayasan pakuan Siliwangi Universitas Pakuan yang telah memberikan dukungan dana hibah internal untuk penelitian ini.

## DAFTAR RUJUKAN

- [1] N. A. Rakhmawati, A. E. Permana, A. M. Reyhan, and H. Rafli, "Analisa Transaksi Belanja Online Pada Masa Pandemi Covid-19," *J. Teknoinfo*, vol. 15, no. 1, p. 32, 2021, doi: 10.33365/jti.v15i1.868.
- [2] V. Jain, B. Malviya, and S. Arya, "An Overview of Electronic Commerce (e-Commerce)," *J. Contemp. Issues Bus. Gov.*, vol. 27, no. 3, 2021, doi: 10.47750/cibg.2021.27.03.090.
- [3] L. Yue, W. Chen, X. Li, W. Zuo, and M. Yin, "A survey of sentiment analysis in social media," *Knowl. Inf. Syst.*, vol. 60, no. 2, pp. 617–663, 2019, doi: 10.1007/s10115-018-1236-4.
- [4] R. Obiedat, O. Harfoushi, R. Qaddoura, L. Al-Qaisi, and A. M. Al-Zoubi, "An evolutionary-based sentiment analysis approach for enhancing government decisions during covid-19 pandemic: The case of jordan," *Appl. Sci.*, vol. 11, no. 19, 2021, doi: 10.3390/app11199080.
- [5] P. P. A., "Performance Evaluation and Comparison using Deep Learning Techniques in Sentiment Analysis," *J. Soft Comput. Paradig.*, vol. 3, no. 2, pp. 123–134, 2021, doi: 10.36548/jscp.2021.2.006.
- [6] H. Hermanto, A. Mustopa, and A. Y. Kuntoro, "Algoritma Klasifikasi Naive Bayes Dan Support Vector Machine Dalam Layanan Komplain Mahasiswa," *JITK (Jurnal Ilmu Pengetah. dan Teknol. Komputer)*, vol. 5, no. 2, pp. 211–220, 2020, doi: 10.33480/jitk.v5i2.1181.
- [7] H. Hairani, K. E. Saputro, and S. Fadli, "K-means-SMOTE for handling class imbalance in the classification of diabetes with C4.5, SVM, and naive Bayes," *J. Teknol. dan Sist. Komput.*, vol. 8, no. 2, pp. 89–93, 2020, doi: 10.14710/jtsiskom.8.2.2020.89-93.
- [8] A. Shenoy and A. Sardana, "Multilogue-Net: A Context-Aware RNN for Multi-modal Emotion Detection and Sentiment Analysis in Conversation," pp. 19–28, 2020, doi: 10.18653/v1/2020.challengehml-1.3.
- [9] S. R. Ahmad, M. Z. M. Rodzi, N. S. Shapiei, N. M. M. Yusop, and S. Ismail, "A review of feature selection and sentiment analysis technique in issues of propaganda," *Int. J.*

- Adv. Comput. Sci. Appl.*, vol. 10, no. 11, pp. 240–245, 2019, doi: 10.14569/IJACSA.2019.0101132.
- [10] R. G, “A Study to Find Facts Behind Preprocessing on Deep Learning Algorithms,” *J. Innov. Image Process.*, vol. 3, no. 1, pp. 66–74, 2021, doi: 10.36548/jiip.2021.1.006.
- [11] N. Alvi Hasanah, Nanik Suciati, and Diana Purwitasari, “Pemantauan Perhatian Publik terhadap Pandemi COVID-19 melalui Klasifikasi Teks dengan Deep Learning,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 1, pp. 193–202, 2021, doi: 10.29207/resti.v5i1.2927.
- [12] A. Moghar and M. Hamiche, “Stock Market Prediction Using LSTM Recurrent Neural Network,” *Procedia Comput. Sci.*, vol. 170, pp. 1168–1173, 2020, doi: 10.1016/j.procs.2020.03.049.
- [13] M. A. Rosid, A. S. Fitriani, I. R. I. Astutik, N. I. Mulloh, and H. A. Gozali, “Improving Text Preprocessing for Student Complaint Document Classification Using Sastrawi,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 874, no. 1, 2020, doi: 10.1088/1757-899X/874/1/012017.
- [14] X. Pei, F. Mei, and J. Gu, “The real-time state identification of the electricity-heat system based on Borderline-SMOTE and XGBoost,” *IET Cyber-Physical Syst. Theory Appl.*, no. March, pp. 1–11, 2022, doi: 10.1049/cps2.12032.
- [15] H. Utama and A. Masruro, “Analisis Sentimen pada Twitter menggunakan Word Embedding dengan Pendekatan Word2Vec,” *J. Sist. Cerdas*, vol. 05, no. 02, pp. 128–134, 2022.