Analisis Sentimen Tingkat Kepuasan Pengguna Penyedia Layanan Telekomunikasi Seluler Indonesia Pada Platform Twitter Dengan Algoritma Gaussian Naïve Bayes dan Support Vector Machine

Latifatuzikra Suhairi
ASIMO

Latar Belakang

- Tingkat kebutuhan manusia untuk dapat berkomunikasi menggunakan teknologi telepon selular semakin lama semakin tinggi.
- Hal ini mengakibatkan berbagai penyedia layanan telekomunikasi selular harus dapat menunjang kebutuhan tersebut dengan berlomba-lomba meningkatkan layanan mereka.
- Banyaknya perusahaan penyedia layanan telekomunikasi selular di Indonesia menjadikan tiap provider memiliki kelebihan dan kelemahan tersendiri di mata pengguna.
- Untuk dapat mengetahui kelebihan dan kekurangan dari provider, perusahaan penyedia dapat melihatnya dari tingkat kepuasan pengguna yang biasanya tersampaikan lewat cuitan atau perkataan di sosial media, salah satunya Twitter.
- Oleh karena itu, diperlukan analisis sentimen pada Twitter pengguna menyangkut penyedia layanan telekomunikasi selular tersebut.
- Analisis sentimen dapat mengelompokkan polaritas dari teks apakah termasuk opini positif atau tidak.
- Dalam tugas ini, proses analisis sentimen pada Twitter pengguna menyangkut penyedia layanan telekomunikasi selular dapat dimodelkan dengan algoritma Naive Bayes dan Support Vector Machine.

Rumusan Masalah

- Bagaimana langkah mengembangkan model analisis sentiment kepuasan pengguna penyedia layanan telekomunikasi seluler Indonesia pada platform Twitter menggunakan metode GaussianNaiveBayes dan SupportVectorMachine?
- Model manakah yang terbaik untuk dapat mengembangkan analisis sentiment kepuasan pengguna penyedia layanan telekomunikasi seluler Indonesia ?

Urgensi

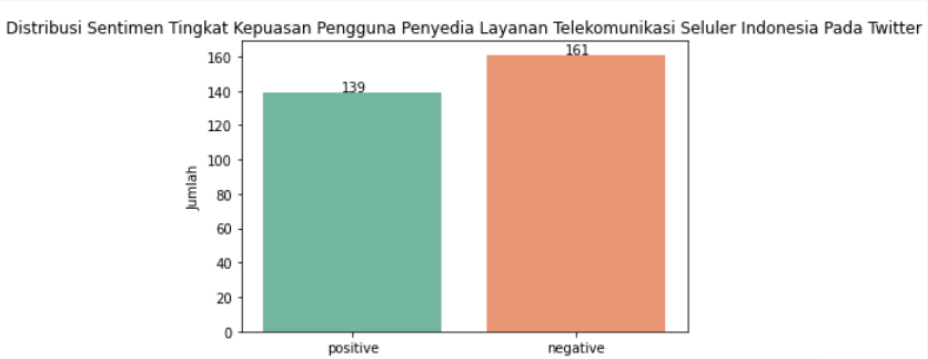
- Dengan dikembangkannya model ini, diharapkan perusahaan penyedia layanan telekmonukasi selular mampu melihat bagaimana tingkat kepuasan pengguna terhadap produk layanan mereka. Sehingga, mereka terpacu untuk terus memperbaiki layanan dan mutu produk.

Dataset

Dataset:

dataset_tweet_sentiment_cellular_service_provider.csv

- Berasal dari Github. Link: https://github.com/rizalespe/Dataset-Sentimen-Analisis-Bahasa-Indonesia.
- Berisikan 300 data tentang Sentimen Tingkat Kepuasan Pengguna Penyedia Layanan Telekomunikasi Seluler Indonesia Pada Platform Twitter
- Terdiri atas 3 kolom: Id, Sentiment, dan Text Tweet.
- Jenis sentiment : positive dan negative



Dataset

index	Id	Sentiment	Text Tweet
0	1	positive	<USER_MENTION> #BOIKOT_<PROVIDER_NAME> Gunakan Produk Bangsa Sendiri <PROVIDER_NAME>
1	2	positive	Saktinya balik lagi, alhamdulillah :v <PROVIDER_NAME>
2	3	negative	Selamat pagi <PROVIDER_NAME> bisa bantu kenapa di dalam kamar sinyal 4G hilang yang 1 lagi panggilan darurat saja <URL>
3	4	negative	Dear <PROVIDER_NAME> akhir2 ini jaringan data lemot banget padahal H+ !!!!
4	5	negative	Selamat malam PENDUSTA <PROVIDER_NAME>

Preprocessing Data

1. Text Normalize
   - Dilakukan untuk mengubah kata slang menjadi bentuk baku yang sesuai dengan kaidah penulisan Bahasa Indonesia.
   - Dilakukan agar komputer dapat memahami makna tweet.
   - Proses ini menggunakan bantuan dari data csv github https://raw.githubusercontent.com/ksnugroho/klasifikasi-spam-sms/master/data/key_norm.csv yang didalamnya terdapat penulisan bahasa slang dan konversinya ke bahasa Indonesia yang baik dan benar.

Preprocessing Data

2. Case Folding

- Dilakukan untuk pembersihan data teks.
- Pada proses ini, dilakukan pengubahan teks menjadi lower case, menghapus angka menggunakan regex yang sudah ditetapkan, menghapus karakter tanda baca menggunakan regex yang sudah ditetapkan, dan menghapus whitespace.

Preprocessing Data

3. Filtering (Remove Stopwords)

- Filtering: pemilihan kata-kata penting atau kata-kata apa saja yang di gunakan untuk mewakili dokumen.
- Dilakukan penghapusan stopwords berdasarkan corpus Indonesia milik library nltk. Kemudian, juga ditambahkan beberapa kata: 'url', 'provider_name', 'user_mention', 'product_name', 'boikot_provider_name', 'boikotprovider_name' pada corpus stopword karena kata tersebut dianggap tidak penting dan belum ada pada corpus stopword ntlk.

index	Id	Sentiment	Text Tweet
0	1	positive	<USER_MENTION> #BOIKOT_<PROVIDER_NAME> Gunakan Produk Bangsa Sendiri <PROVIDER_NAME>
1	2	positive	Saktinya balik lagi, alhamdulillah :v <PROVIDER_NAME>
2	3	negative	Selamat pagi <PROVIDER_NAME> bisa bantu kenapa di dalam kamar sinyal 4G hilang yang 1 lagi panggilan darurat saja <URL>
3	4	negative	Dear <PROVIDER_NAME> akhir2 ini jaringan data lemot banget padahal H+ !!!!
4	5	negative	Selamat malam PENDUSTA <PROVIDER_NAME>

Preprocessing Data

4. Stemming

- Stemming: melibatkan proses pemetaan dan penguraian bentuk dari suatu kata menjadi bentuk kata dasarnya.
- Dalam implementasinya, karena dataset berupa text Bahasa Indonesia, maka digunakan bantuan library Sastrawi. Sastrawi merupakan library sederhana yang dapat mengubah kata berimbuhan bahasa Indonesia menjadi bentuk dasarnya.

Preprocessing Data

Gunakan Pipeline untuk melakukan PreProcessing pada data dengan tahapan berikut:

1. Text Normalize
2. Casefolding
3. Filtering (remove Stopwords)
4. Stemming

Lama PreProcessing

```
CPU times: user 44.5 s, sys: 182 ms, total: 44.7 s
Wall time: 50.4 s
```

Hasil PreProcessing

	Id	Sentiment	Text Tweet	clean_teks
0	1	positive	<USER_MENTION> #BOIKOT_<PROVIDER_NAME> Gunakan...	produk bangsa
1	2	positive	Saktinya balik lagi, alhamdulillah :v <PROVIDE...	sakti alhamdulillah v
2	3	negative	Selamat pagi <PROVIDER_NAME> bisa bantu kenap...	selamat pagi bantu kamar sinyal g hilang pangg...
3	4	negative	Dear <PROVIDER_NAME> akhir2 ini jaringan data ...	dear jaring data lot banget h
4	5	negative	Selamat malam PENDUSTA <PROVIDER_NAME>	selamat malam dusta
5	6	negative	Untuk penembakan paket dari <PRODUCT_NAME> mas...	tembak paket ganggu ya
6	7	positive	<PROVIDER_NAME> aku pakai <PROVIDER_NAME>, pa...	pakai paket nya off ganti paket gratis youtube...
7	8	negative	RT <USER_MENTION>: <PROVIDER_NAME> tak ada lag...	rt kamus perhapean
8	9	negative	keluhan gak ditanggapi. bikin emosi aja. pulsa...	keluh tanggap bikin emosi aja pulsa curi soak ...
9	10	negative	#Bilboard iklan <PROVIDER_NAME> kok ada pesan ...	bilboard iklan pesan sembunyi ramadhan

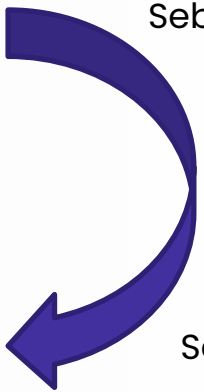
Preprocessing Data

5. Pelabelan Data Label

- Dilakukan karena kolom Sentimen masih berbentuk String → numerical
- Hasilnya: nilai 0 mewakili data negative dan nilai 1 mewakili data positive.

	Id	Sentiment	Text Tweet	clean_teks
0	1	positive	<USER_MENTION> #BOIKOT_<PROVIDER_NAME> Gunakan...	produk bangsa
1	2	positive	Saktinya balik lagi, alhamdulillah :v <PROVIDE...	sakti alhamdulillah v
2	3	negative	Selamat pagi <PROVIDER_NAME> bisa bantu kenap...	selamat pagi bantu kamar sinyal g hilang pangg...
3	4	negative	Dear <PROVIDER_NAME> akhir2 ini jaringan data ...	dear jaring data lot banget h
4	5	negative	Selamat malam PENDUSTA <PROVIDER_NAME>	selamat malam dusta

Sebelum pelabelan



Setelah pelabelan

	Id	Sentiment	Text Tweet	clean_teks
0	1	1	<USER_MENTION> #BOIKOT_<PROVIDER_NAME> Gunakan...	produk bangsa
1	2	1	Saktinya balik lagi, alhamdulillah :v <PROVIDE...	sakti alhamdulillah v
2	3	0	Selamat pagi <PROVIDER_NAME> bisa bantu kenap...	selamat pagi bantu kamar sinyal g hilang pangg...
3	4	0	Dear <PROVIDER_NAME> akhir2 ini jaringan data ...	dear jaring data lot banget h
4	5	0	Selamat malam PENDUSTA <PROVIDER_NAME>	selamat malam dusta

Ekstraksi Feature

- Mengubah data text menjadi vektor agar mudah dipahami oleh komputer.
- Menggunakan TF-IDF. TF-IDF biasa digunakan ketika kita ingin mengubah data teks menjadi vektor namun dengan memperhatikan apakah sebuah kata tersebut cukup informatif atau tidak. Mudahnya, TF-IDF membuat kata yang sering muncul memiliki nilai yang cenderung kecil, sedangkan untuk kata yang semakin jarang muncul akan memiliki nilai yang cenderung besar.
- Menggunakan TF-IDFVectorizer dengan n_gram = (1,3)

Ekstraksi Feature

- Hasilnya, terdapat 2959 fitur.
- Berikut Matriks Jumlah Token dengan TFIDF

	acara	acara live	acara live streaming	aceh	aceh singkil	aceh singkil stabil	adhan	adhan styles	adhan styles mekah	aja	...	youtube top deh	youtubenya	youtubenya pakai	youtubenya pakai jam	youtubeyondergenflix	yuk	yuk pakai	yuk pakai rp	zalim	zalim ya
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
295	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
296	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
297	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
298	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
299	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

300 rows × 2959 columns

- Selanjutnya, dilakukan seleksi fitur. Karena jumlah fitur yang terlalu banyak.

Ekstraksi Feature

- Seleksi fitur dilakukan dengan menggunakan chisquare.
- Dari 2959 fitur, diambil 1200 fitur terbaik. Seleksi fitur melalui proses mask dengan Mask yang bernilai false menandakan bahwa fitur tidak akan dipilih dan True menandakan fitur tersebut akan dipilih (digunakan).
- Berikut Matriks Jumlah Token dengan TFIDF Setelah di seleksi fitur

	ajaib min	ajaib min kartu	ajar pakai	ajar pakai produk	aksi	aksi pakai	aksi pakai bagus	alhamdulilah	alhamdulilah kunjung	alhamdulilah kunjung vendor	...	youtube lancar banget	youtube ramadhan	youtube ramadhan faedah	youtube tahun	youtube tahun kuota	yuk	yuk pakai	yuk pakai rp	zalim	zalim ya
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
295	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
296	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
297	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
298	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
299	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

300 rows × 1200 columns

Modelling

01 Model 1 : Gaussian Naïve Bayes

Algoritma *Naive Bayes* mempelajari probabilitas suatu objek berdasarkan ciri-ciri tertentu yang termasuk dalam kelompok atau kelas tertentu. (pengklasifikasian probabilistic). *Gaussian* NB adalah tipe *Naïve Bayes* yang mengikuti distribusi normal *Gaussian* dan mendukung data kontinu.

Pemodelan Menggunakan GridSearchCV untuk Tuning Hyperparameter

Parameter sebelum Tuning

```
naive_bayes = GaussianNB()
model_nb = naive_bayes.fit(X_train, y_train)
```

Parameter Setelah Tuning

```
print("Best param: ", model_nb_grid.best_params_)
print("Best score: ", model_nb_grid.best_score_)
```

```
Best param:  {'var_smoothing': 0.0533669923120631}
```

Modelling

02 Model 2 : Support Vector Machine Classifier

Algoritma supervised learning untuk klasifikasi dan regresi yang bekerja menggunakan konsep Structural Risk Minimization. dirancang untuk mengolah data menjadi Hyperplane yang mengklasifikasikan ruang input menjadi dua kelas.

Pemodelan Menggunakan GridSearchCV untuk Tuning Hyperparameter

Parameter sebelum Tuning

```
svc = SVC(C=0.0001, kernel='poly', degree=2, max_iter=50)
model_svc = svc.fit(X_train, y_train)
```

Parameter Setelah Tuning

```
Best param:  {'C': 1.0, 'gamma': 0.1, 'kernel': 'linear', 'max_iter': 50}
Best score:  0.8380952380952382
```

Performa Model Yang Dihasilkan

Evaluasi model menggunakan Confussion Matrix dan Classification Report

01 Model 1 : Gaussian Naïve Bayes

Sebelum Tuning

```
Confusion Matrix Model 1:
 [[55  2]
 [ 8 25]]
```

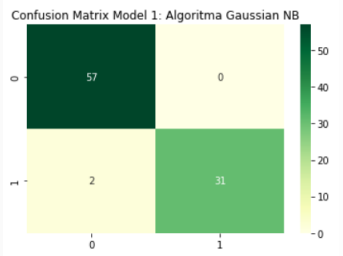
Prediksi Benar: 80
Prediksi Salah : 10
Akurasi : 89%

```
Classification report Model 1:
              precision    recall  f1-score   support

           0       0.87      0.96      0.92        57
           1       0.93      0.76      0.83        33

    accuracy                           0.89        90
   macro avg       0.90      0.86      0.88        90
weighted avg       0.89      0.89      0.89        90
```

Setelah Tuning



Prediksi Benar: 88
Prediksi Salah 2
Akurasi : 98%

```
Classification report Model 1 Setelah Di Tuning:
              precision    recall  f1-score   support

           0       0.97      1.00      0.98        57
           1       1.00      0.94      0.97        33

    accuracy                           0.98        90
   macro avg       0.98      0.97      0.98        90
weighted avg       0.98      0.98      0.98        90
```

Performa Model Yang Dihasilkan

02 Model 2 : Support Vector Machine

Sebelum Tuning

```
Confusion Matrix Model 2:
 [[14 43]
 [ 0 33]]
```

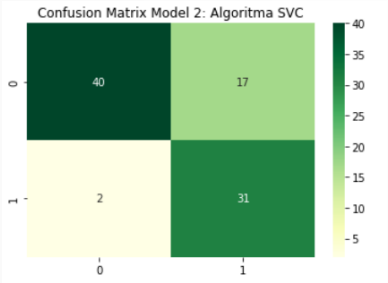
Prediksi Benar: 47
Prediksi Salah : 43
Akurasi : 52%

```
Classification report Model 2:
              precision    recall  f1-score   support

           0       1.00      0.25      0.39        57
           1       0.43      1.00      0.61        33

    accuracy                           0.52        90
   macro avg       0.72      0.62      0.50        90
weighted avg       0.79      0.52      0.47        90
```

Setelah Tuning



Prediksi Benar: 71
Prediksi Salah 17
Akurasi : 79%

```
Classification report Model 2 Setelah Di Tuning:
              precision    recall  f1-score   support

           0       0.95      0.70      0.81        57
           1       0.65      0.94      0.77        33

    accuracy                           0.79        90
   macro avg       0.80      0.82      0.79        90
weighted avg       0.84      0.79      0.79        90
```

Kesimpulan

01

Dari tugas analisis, berdasarkan data **dataset_tweet_sentiment_cellular_service_provider.csv** dapat dibuatkan model analisis sentimen dengan langkah preprocessing data (text normalize – case folding – filtering – stemming) dan label encoder, kemudian feature extraction, dan modelling (naïve bayes dan svc) – evaluasi model – tuning hyperparameter – deployment.

02

Model analisis sentiment Tingkat Kepuasan Pengguna Penyedia Layanan Telekomunikasi Seluler Indonesia Pada Platform Twitter yang terbaik adalah model Gaussian Naive Bayes dengan nilai akurasi model 98% dibandingkan dengan model Support Vector Machine Classifier yang hanya memiliki nilai akurasi 79%.