Journal of Manufacturing in Industrial Engineering and Technology (MINE-TECH)
Vol.5, No.1, Juni 2026, E-ISSN: 2985-3516
DOI : 10.30651/mine-tech.v5i1.31396

Deployment Klasifikasi Sentimen Pada Ulasan Pengguna Aplikasi PLN Mobile Dengan Machine Learning

Akhmad Ghiffary Budianto[1,2*], Aqli Mursadin[2]
Program Studi Pendidikan Profesi Insinyur, Universitas Lambung Mangkurat[1]
Jurusan Teknik Mesin, Universitas Lambung Mangkurat[2]
Professional Engineer Program, Lambung Mangkurat University[1]
Department of Mechanical Engineering, Lambung Mangkurat University[2]
ghiffaryb04@gmail.com[1], a.mursadin@ulm.ac.id[2]

Informasi Artikel

Riwayat Artikel:

Disubmit April 27, 2026
Diterima Juni 03, 2026
Diterbitkan Juni 11, 2026

Kata Kunci:

Deployment
Klasifikasi
Sentimen
Logistic regression
Gradio

ABSTRAK

PT. PLN menyediakan layanan pengaduan bagi pelanggan terkait masalah kelistrikan baik melalui social media maupun aplikasi mobile PLN di android. Dalam google playstore, mobile PLN mendapatkan ulasan pengguna berupa rating dan ulasan. Sentimen terhadap ulasan pengguna bertujuan untuk memahami konteks dari ulasan yang diberikan pengguna. Penelitian ini bertujuan untuk dapat melakukan klasifikasi sentimen pada ulasan pengguna tersebut dan melakukan proses *deployment* ke dalam *user interface* agar mudah digunakan. Metode yang digunakan yaitu mengumpulkan ulasan pengguna dengan web scraping lalu mengolahnya dengan TF-IDF dan klasifikasi dengan logistic regression dan naïve bayes. *K-folds cross validation* dengan k=5 dan metrik performa seperti akurasi dan f1-score digunakan untuk mengukur kinerja model. Model klasifikasi sentimen dengan naïve bayes menghasilkan tingkat akurasi sebesar 91%±2,42% dan f1-score sebesar 90,98%±2,44%. Model Naïve Bayes sedikit lebih baik dibandingkan dengan hasil model logistic regression dengan ingkat akurasi 87,9%±2,36% dan f1-score 87,87%±2,39%. Total misklasifikasi yang terjadi dengan model naïve bayes (17 kali) juga lebih sedikit terjadi dibandingkan dengan model logistic regression (22 kali). *Deployment* berhasil dilakukan dengan menggunakan Gradio *user interface* sederhana untuk melakukan trial dari model klasifikasi sentimen.

***Penulis Korespondensi*:**

Akhmad Ghiffary Budianto
Program Studi Pendidikan Profesi Insinyur
Universitas Lambung Mangkurat
Jl. Jenderal Achmad Yani KM 35,5, Banjarbaru, Indonesia
Email: ghiffaryb04@gmail.com

1. PENDAHULUAN

Listrik merupakan kebutuhan utama dalam era teknologi dan digitalisasi. PT. PLN (Persero) sebagai Badan Usaha Milik Negara (BUMN) yang memiliki kewenangan dalam pengurusan kelistrikan dan bertugas dalam memenuhi kebutuhan listrik di seluruh wilayah di Indonesia. Dalam rangka menjaga kualitas pelayanan dari penyediaan listrik nasional, PT. PLN menyediakan layanan pengaduan bagi pelanggan terkait masalah kelistrikan. Sistem pengaduan dapat diakses oleh pelanggan melalui berbagai platform antara lain telepon PLN 123, PLN Mobile

dan media sosial PLN. Dalam hal ini, PLN memiliki banyak jenis keluhan pelanggan baik menggunakan rekaman suara maupun komentar tertulis. Dalam proses menindaklanjuti keluhan pelanggan secara harian ada ratusan bahkan sampai ribuan komentar terkait keluhan pelanggan listrik. Oleh karena itu, keluhan pelanggan berupa komentar ini perlu diolah dan diklasifikasi oleh sistem agar dapat ditindaklanjuti secepat mungkin. Keluhan pelanggan listrik umumnya meliputi pemadaman listrik, fluktuasi tegangan, tagihan listrik yang tidak sesuai, dan sebagainya.

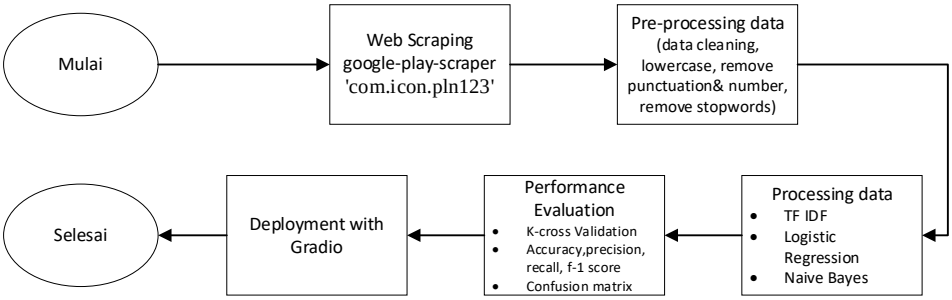
Big data analytics khususnya pada *text mining* menjadi salah satu cara dalam memahami komentar pelanggan. *Text mining* dan klasifikasi teks membantu dalam upaya memahami pengguna tentang produk/layanan dengan interaksi dalam bisnis. Umumnya hal ini dikenal sebagai data *Voice of Customer* (VOC) [1]. Sistem pakar berbasis AI memiliki potensi yang sangat besar dalam pengalaman pelanggan (CX) [2]. Dalam survei *MIT Technology Review Insights* Maret 2020 yang melibatkan lebih dari 1000 pemimpin bisnis, hasilnya menunjukkan bahwa departemen layanan pelanggan kemungkinan besar menggunakan AI (73%), diikuti oleh departemen penjualan dan pemasaran (59%). Survei tersebut juga memprediksi layanan pelanggan akan menjadi departemen terdepan pada tahun 2022 [3]. Pada literatur lain menyatakan bahwa untuk memanfaatkan analisis *big data* dan mewujudkan peningkatan kinerja, perusahaan harus mengembangkan kemampuan analisis *big data* yang kuat dan AI untuk mencapai hasil yang sangat baik di area ini [4].

Dengan adanya platform digital, pelanggan listrik dapat langsung memberikan komentar dan keluhan terkait permasalahan yang dialami. Manajemen PLN dalam mengelola keluhan harus berinovasi dengan membuat kecerdasan buatan untuk deteksi dari permasalahan kelistrikan. *Natural Language Processing* (NLP) dalam hal ini dapat memberikan solusi untuk melakukan klasifikasi dari komentar pengguna dalam jumlah besar. NLP juga digunakan untuk analisis sentimen, yaitu untuk mendeteksi informasi subjektif dari dokumen, jejaring sosial, dan sebagainya. Ini berguna untuk mengidentifikasi tren dalam opini pengguna yang dapat dimanfaatkan [5]. Salah satu cabang dari NLP yaitu analisis sentimen. Analisis sentimen umum digunakan untuk mengetahui respon sentimen pelanggan terhadap produk yang digunakan.

Pengintegrasian analisis sentimen ke dalam sistem pemantauan berbasis data menjadi instrumen yang penting dalam ranah manajemen sistem layanan digital, dimana pada ulasan yang diberikan oleh konsumen bertindak sebagai *Voice of Customer*. Dalam konteks Qualtity 4.0 (kualitas pada era industri digital), organisasi dituntut untuk mengesktraksi data teks terstruktur dan tidak terstruktur dari berbagai kanal digital [6]. Hal ini bertujuan untuk memberi wawasan secara real-time mengenai emosi serta preferensi konsumen demi meningkatkan customer engagement, loyalitas dan posisi market perusahaan [7].

Berdasarkan uraian diatas, penelitian ini bertujuan untuk membangun model klasifikasi otomatis untuk mengidentifikasi sentimen dari komentar pelanggan pengguna aplikasi PLN mobile di android. Selain itu, penelitian ini juga menganalisis kata kunci apa yang paling sering muncul pada setiap komentar pelanggan pada ulasan aplikasi PLN mobile. Model yang dibangun dengan algoritma *Logistic Regression* selanjutnya diukur performanya dengan metrik Akurasi, presisi, *recall, F-1 Score* dan *specificity*. Model yang memiliki tingkat akurasi yang bagus selanjutnya dapat dibuatkan pipeline dan user interface sederhana untuk melakukan klasifikasi sentimen secara otomatis. Harapannya model yang dibangun dapat membantu manajerial PLN dalam memahami sentimen dari pelanggan dan dapat mengambil tindakan yang responsif.

2. METODE



Gambar 1. Diagram Alir Metode Penelitian

Secara umum, gambar 1 menunjukkan alur proses metode penelitian yang digunakan. Tahap pengumpulan data dengan metode web scraping menggunakan *google colab* dan *library* google-play-scraper. Ini bertujuan untuk mendapatkan komentar dari pengguna aplikasi PLN mobile. Selanjutnya, tahap *pre-processing data*, data berupa teks yang didapatkan sebelumnya dilakukan proses *lowercase*, penghapusan tanda baca, angka dan *stop words*. Setelah itu dataset dipreoses dengan metode TF-IDF (*Term Frequency-Inverse Document Frequency*) dan algoritma logistic regression untuk tugas klasifikasi biner. Model klasifikasi sentiment selanjutnya dilakukan proses evaluasi dengan K-fold cross validation, pengukuran metrik akurasi, presisi, *recall* dan *f1-score*. Untuk memperkuat gambaran misklasifikasi yang terjadi *confusion matrix* juga dapat digunakan untuk mengilustrasikan performa model. Yang terakhir, model dapat di *deploy* dengan *user interface* sederhana dengan menggunakan Gradio.

Pengumpulan data dengan web scraping

Untuk pengumpulan data, metode web scraping dengan menggunakan library python (google-play-scraper). Pengaturan yang dilakukan adalah sebagai berikut:

```
result_all, continuation_token = reviews(
    'com.icon.pln123',
    lang='id', # defaults to 'en'
    country='id', # defaults to 'us'
    # Removed sort=Sort.NEWEST to potentially get a more diverse set of reviews
    count=100000, # Increased count to get more data for filtering
    filter_score_with=None
)
```

'com.icon.pln123' digunakan sebagai token yang didapat dari halaman situs google playstore (https://play.google.com/store/apps/details?id=com.icon.pln123&hl=id). Jumlah yang dikumpulkan sebagai initial ada sebanyak 100.000 baris data. Hal ini untuk memastikan semua rating memiliki keterwakilan yang berimbang didalam dataset. Selanjutnya melakukan proses filter data untuk mendapatkan rating skor 1-3 sebanyak 500 baris data dan skor 4-5 sebanyak 500 data dengan setting sebagai berikut:

```
# Filter for scores 1-3
df_low_score = df_pln_all[df_pln_all['score'].isin([1, 2, 3])]
# Take all available reviews for scores 1-3, as there are limited number based on initial
df_pln
reviews_low_score = df_low_score.head(500) # Take up to 500 if available, but will likely
be less

# Filter for scores 4-5
df_high_score = df_pln_all[df_pln_all['score'].isin([4, 5])]
# Take up to 500 reviews for scores 4-5
reviews_high_score = df_high_score.head(500)
```

Dengan pengaturan ini, maka data akan memiliki 500 baris dengan skor 1-3 dan 500 baris data dengan skor 4-5. Hal ini akan membantu mendapatkan input ulasan pengguna pada rating rendah dan tinggi secara berimbang.

Web scraping dengan menggunakan *library* google-play-scraper akan mengumpulkan data berupa:

1) *Username* yaitu nama akun dari gmail yang digunakan pengguna
2) *Score* yaitu nilai rating ulasan yang diberikan antara 1 sampai dengan 5
3) *At* yaitu waktu dan tanggal pemberian ulasan pada google playstore
4) *Content* yaitu berupa ulasan komentar pengguna terhadap aplikasi yang digunakan

Dari keempat kolom ini, yang digunakan yaitu hanya pada *score* dan *content*. *Score* digunakan untuk melihat rating rendah dan tinggi, sedangkan *content* bertujuan untuk mengumpulkan data teks pada rating rendah maupun tinggi. Rating rendah dengan skor 1-3 akan memiliki label negatif dan skor 4-5 akan memiliki label sentimen positif.

Preprocessing data

Preprocessing data menjadi tahapan yang penting untuk menjamin dataset yang digunakan akan memiliki pengaruh langsung terhadap model klasifikasi yang dibuat [8]. Adapun pada preprocessing ini melakukan aktivitas antara lain:

1) *Remove empty row* yaitu pembersihan data jika ada baris data yang kosong
2) *Lowercase* yaitu mengubah teks data menjadi huruf kecil tanpa adanya huruf kapital
3) *Remove punctuation & number* yaitu menghapus adanya simbol dan angka pada data
4) *Remove stopwords* yaitu penghapusan kata-kata yang frekuensinya sangat tinggi namun tidak memiliki makna signifikan seperti “di”, “kan”, dan “saya”. Biasanya terdiri atas kata penghubung, kata depan, atau ganti.

Processing data* dengan TF-IDF dan *machine learning

Data yang akan diproses harus diubah dalam bentuk numerik oleh karena itu perlu diubah dengan metode TF-IDF (*Term Frequency-Inverse Document Frequency)*. TF IDF merupakan metode yang digunakan untuk mengukur keterhubungan kata (*term*) terhadap dokumen dengan pemberian bobot pada setiap katanya. Hal ini berdasarkan kepada frekuensi kemunculan kata dalam dokumen dan inverse frekuensi dokumen yang terdapat kata tersebut [9].

Logistic Regression mampu melakukan klasifikasi binary dengan baik. Tujuan regresi logistik adalah untuk menemukan model yang paling sesuai (namun tetap masuk akal secara logis) guna menggambarkan hubungan antara variabel dependen dikotomis (variabel terikat = variabel respons atau hasil) dan serangkaian variabel independen (variabel prediktor atau variabel penjelas) [10]. Secara matematis, fungsi dari logistic regression dapat dilihat pada persamaan (1) berikut:

$$logit(S) = b_0 + b_1M_1 + b_2M_2 + b_3M_3 + \cdots + b_kM_k \quad (1)$$

Keterangan:

S = probabilitas fitur yang diamati
$M_1, M_2, ..., M_K$ = nilai prediktor
$B_0, B_1, ..., B_k$ = intercept pada model

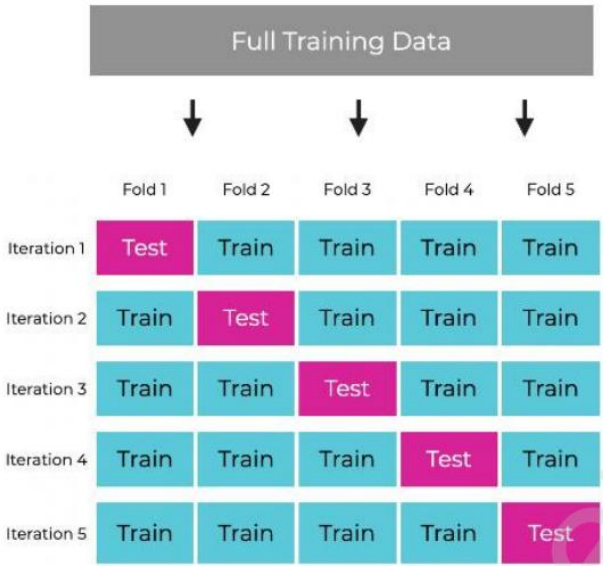
Klasifikasi Naive Bayes banyak digunakan untuk klasifikasi teks dalam machine learning didasarkan pada probabilitas bersyarat dari fitur-fitur yang dimiliki oleh suatu kategori. Metode naive bayes memiliki keunggulan dimana metode ini simpel tangkas dan memiliki kecepatan yang cukup tinggi [11]. Naive Bayes memiliki batasan penggunaan yang sederhana pada tugas klasifikasi teks. Secara matematis, persamaan naive bayes dapat dilihat sebagai berikut:

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

Keterangan:

X = data yang memiliki kelas tidak diketahui
H = asumsi data X melambangkan kelas tertentu
P(H|X) = peluang asumsi H berlandaskan pada keadaan X (*posterior probability*)
P(H) = peluang asumsi H (*prior probability*)
P(X|H) = peluang asumsi X berlandaskan pada keadaan H (*likelihood*)
P(X) = peluang dari X (*predictor prior probability*)

Setelah data diproses dengan metode TF-IDF untuk mengubah kata menjadi vector numerik dan logistic regression sebagai classifier machine learning, maka data perlu dilakukan proses split menjadi 80:20 (80% dari data untuk data training dan 20% dari data untuk data testing). Dalam menjamin output model tidak terjadi *overfitting* (output terlalu bagus saat dilatih namun saat pengujian model outputnya kurang akurat) maupun *underfitting* (output tidak terlalu bagus saat dilatih namun saat pengujian model outputnya malah lebih akurat), maka penggunaan *k- folds cross validation* digunakan. Dalam *k-folds cross validation*, adalah teknik dalam pembelajaran mesin dan pemodelan statistik, yang terutama digunakan untuk menilai sejauh mana hasil analisis statistik dapat diterapkan pada kumpulan data yang independen [12].



Gambar 2. Ilustrasi K-folds Cross Validation k=5 [13]

Gambar 2 mengilustrasikan k-folds cross validation dengan k=5. Dalam kasus validasi silang 5 folds, data dibagi menjadi lima lipatan. Dalam pendekatan ini, model dilatih dan divalidasi beberapa kali melalui berbagai iterasi. Untuk setiap iterasi, satu lipatan dipilih sebagai himpunan validasi, dan model dilatih pada empat lipatan yang tersisa dengan ketentuan 80% dari data untuk data training dan 20% dari data untuk data testing. Kinerja rata-rata model kemudian dihitung dari lima iterasi tersebut [13].

***Metric performance* pada machine learning**

Kinerja klasifikasi dievaluasi menggunakan *Confusion Matrix*, sebuah alat yang umum digunakan untuk menilai tugas klasifikasi. Metrik *Precision* mengacu pada proporsi prediksi positif benar di antara semua prediksi positif, sedangkan *recall* mengukur proporsi positif benar di antara semua positif aktual. *F1-score* menggabungkan *precision* dan *recall* menjadi satu skor tertimbang. Akurasi, yang mewakili rasio pengamatan yang diprediksi dengan benar terhadap jumlah total pengamatan, sering dianggap sebagai metrik kinerja yang paling penting, dengan nilai yang lebih tinggi menandakan kinerja model yang lebih baik [14]. Nilai rata-rata tertimbang F1 yang tinggi menunjukkan bahwa algoritma tersebut berkinerja baik di semua kelas, dan sebaliknya [15]. Persamaan untuk presisi, *recall*, akurasi, dan skor F1 dalam klasifikasi adalah sebagai berikut:

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

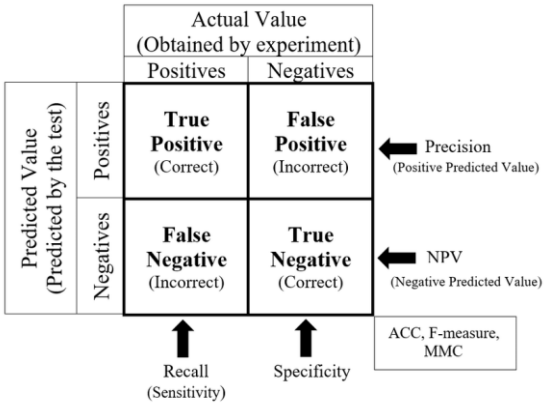
$$Recall = \frac{TP}{TP + FN} \quad (4)$$

$$F1 - score = 2\,x\,\frac{precision\ .recall}{precision + recall} \quad (5)$$

$$Accuracy = \frac{TP+TN}{TP + TN + FP + FN} \quad (6)$$

$$Weighted\ avg\ f1-score\ = \frac{\sum_{i-1}^{N} F_1 \times support_i}{\sum_{i-1}^{N} support_i} \quad (7)$$

Where:
TP : *True Positive*
TN : *True Negative*
FP : *False Positive*
FN : *False Negative*
N : Jumlah kelas
Support : Jumlah sampel

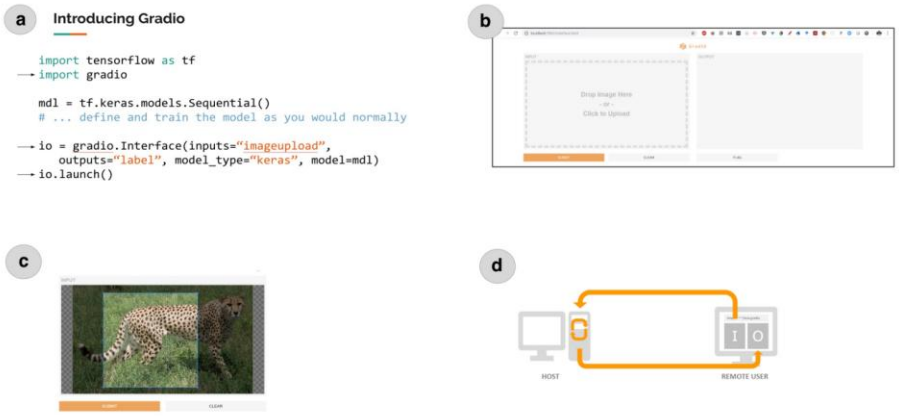


Gambar 3. Confusion Matrix pada Klasifikasi Biner [16]

Gambar 3 mengilustrasikan *confusion matrix* pada tugas klasifikasi biner. Dimana dalam pengukuran metrik sebagai bentuk evaluasi menggunakan *precision* untuk mengukur jumlah prediksi label positif, *recall* (sensitifitas) untuk mengukur *true positive* dan *false negative* serta nilai acc (akurasi) dan *f-measure* (*f1-score*) sebagai 2 indikator utama dalam mengukur performa suatu machine learning.

Deployment model dengan gradio

Gradio diimplementasikan sebagai pustaka Python dan dapat diinstal dari PyPi2. Setelah terinstal, menjalankan antarmuka Gradio hanya memerlukan sedikit perubahan pada alur kerja yang sudah ada bagi pengembang pembelajaran mesin. Setelah model dilatih, pengembang membuat objek *interface* dengan empat parameter wajib seperti pada Gambar 4. Parameter pertama dan kedua adalah input dan output, yang mengambil antarmuka input/output yang akan digunakan sebagai argumen. Pengembang dapat memilih subkelas mana pun dari `Gradio.AbstractInput` dan `Gradio.AbstractOutput`, masing-masing. Saat ini, ini mencakup pustaka antarmuka standar untuk menangani data gambar, teks, dan audio [17].



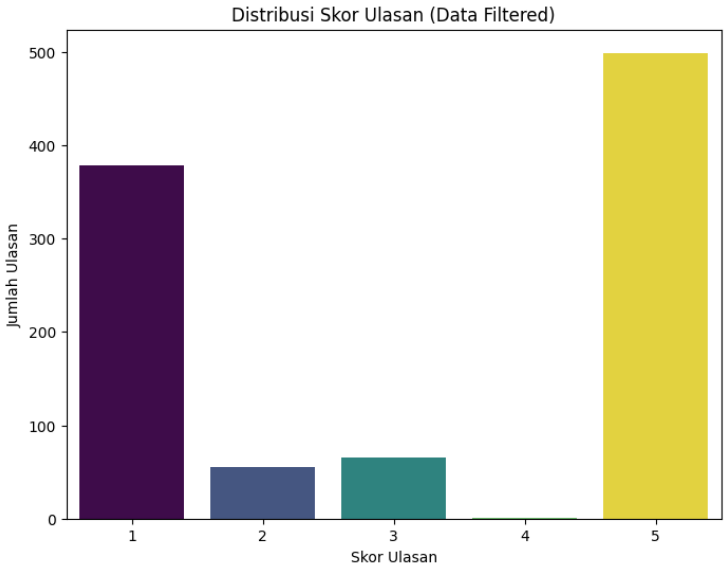
Gambar 4. Proses Desain *User Interface* pada Gradio, (a) Menentukan Input dan Output, (b) Tampilan Rancangan *User Interface*, (c) Proses Penyesuaian Input (Gambar), (d) Proses Komputasi *Machine Learning* dari *Host* ke *Remote User* dan Sebaliknya [17]

3. HASIL DAN ANALISIS

Hasil pengumpulan data

Data yang digunakan adalah data hasil web scraping pada aplikasi PLN Mobile di google play store sebanyak 1.000 baris data ulasan pengguna. Namun, untuk proses pengampilan data tidak menggunakan data terbaru, melainkan menggunakan fitur balance score review. Fitur ini akan memastikan data akan seimbang untuk mendapatkan jumlah data sebanyak 500 baris untuk rating skor 1-3 dan 500 data untuk rating skor 4-5. Pada ulasan online, seorang customer juga memberikan ulasan tekstual mengenai suatu produk atau layanan serta memberikan penilaian numerik 1-5. Hal ini untuk menunjukkan pengalaman yang dirasakan customer, rating bintang 1 untuk pengalaman terburuk dan sebaliknya rating bintang 5 untuk menunjukkan pengalaman terbaik (memuaskan). Penggunaan metode ini ditujukan sebagai baseline label yang seimbang untuk efisiensi pemrosesan data skala besar dan pelabelan awal otomatis [18]. Pada penelitian terdahulu yang dilakukan oleh Asri, dkk. [19] melakukan pengumpulan data yang tidak seimbang antara label, dimana label positif yang lebih banyak pada dataset. Hal ini berdampak pada data input label lain mengalami ketidakcukupan input data dan menghasilkan akurasi klasifikasi di 70%.

Berdasarkan pertimbangan keseimbangan dataset sebagai input, maka penggunaan dataset dengan 50% memiliki rating 1-3 dan 50% dataset memiliki rating 4-5 dijadikan sebagai *data training*. Hasil rekapitulasi rating skor ulasan pengguna dan frekuensinya pada penelitian ini dapat dilihat pada Gambar 5 berikut.



Gambar 5. Frekuensi Ulasan pada Skor Ulasan 1 – 5

Gambar 5 menunjukkan frekuensi ulasan pengguna pada tiap rating skor. Rating skor menunjukkan persebaran skor rating yang cukup merata. Hal ini akan menjamin data akan memiliki keseimbangan 50:50 untuk label data positif dan negatif. Dataset yang baik akan memberikan keseimbangan antara label untuk mendapatkan data *training* yang sama besarnya. Sehingga outputnya pun juga akan mencerminkan data *training* dan *testing* yang telah dilakukan.

Hasil *pre-processing data*

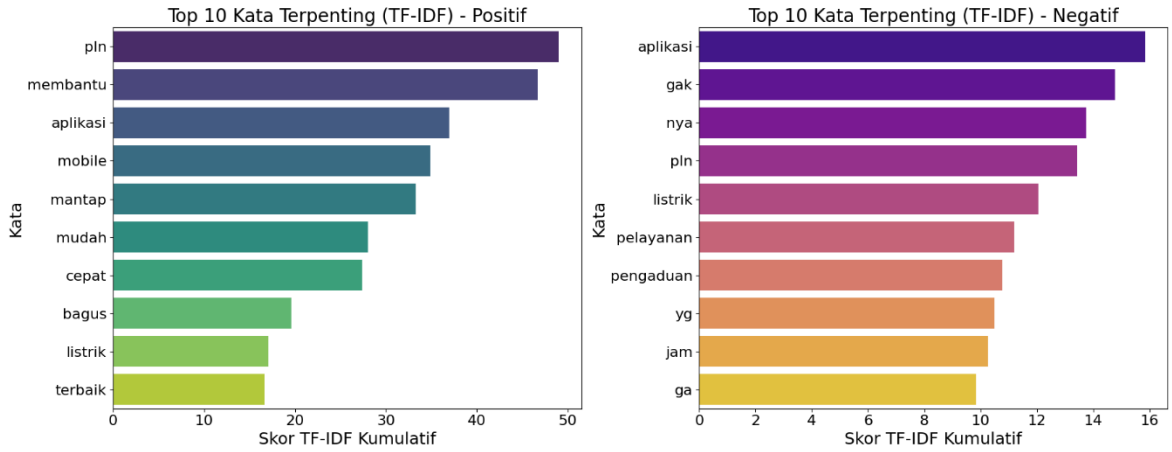
Data yang telah dilakukan proses *web scraping*, selanjutnya perlu untuk dilakukan *data cleaning*, *lowercase*, menghapus tanda baca, angka dan *stop words*, serta melakukan tokenisasi. Hasilnya dapat dilihat pada Gambar 6 berikut.

	content	processed_content
0	banyak lokasi non kantor PLN tidak ada dan sul...	lokasi non kantor pln sulit ditemukankalopun a...
1	Pengalaman sangat buruk Pengajuan perubahan da...	pengalaman buruk pengajuan perubahan daya resp...
2	sangat puas atas layanan ya	puas layanan ya
3	sebaiknya kwh meter dulu jangan diganti lebih ...	kwh meter diganti pengecekan ganti mati pas dg...
4	pengaduan dibatalkan sepihak oleh petugas, vol...	pengaduan dibatalkan petugas voltase rendah di...

Gambar 6. Contoh Sampel Dataset untuk *Pre-Processing*

Hasil *processing data* dengan TF-IDF

Dataset selanjutnya diproses dengan metode TF-IDF untuk mendapatkan skor kumulatif dari tiap label sentimennya. Hasil rekapitulasi skor TF-IDF dapat dilihat pada gambar 3 berikut.



Gambar 7. Hasil Skor TF-IDF Kumulatif dan Top 10 Kata Terpenting pada Label Positif dan Negatif

Gambar 7 menunjukkan hasil rekapitulasi skor TF-IDF secara kumulatif beserta 10 kata teratas yang muncul pada label positif dan negatif. Skor kumulatif TF-IDF yang tinggi nilainya menunjukkan bahwa kata tersebut bukan hanya sekedar sering muncul secara berulang-ilang namun juga memiliki karakteristik yang merepresentasi label tersebut. Pada label positif, 10 kata terpenting secara berurutan dari paling tinggi ke rendah yaitu pln, membantu, aplikasi, mobile, mantap, mudah, cepat, bagus, listrik, dan terbaik. Kata kunci ini menunjukkan bahwa aplikasi dinilai positif oleh pengguna saat aplikasinya sangat membantu, mudah dan cepat dalam prosesnya. Sedangkan pada label negatif, 10 kata terpenting yang secara berurutan dari paling tinggi ke rendah yaitu aplikasi, gak, nya, pln, listrik, pelayanan, pengaduan, yg, jam, dan ga. Hal ini menunjukkan bahwa pelayanan dan pengaduan masih dinilai negatif oleh pengguna.

Hasil metrik performa *Logistic Regression* dengan *K-folds Cross Validation*

Pada proses prediksi sentimen ulasan pengguna aplikasi PLN mobile, algoritma logistic regression dipilih karena kemampuannya yang baik dalam mengklasifikasikan label biner (0/1 atau positif / negatif). Pada proses data training dan testing, data dilakukan proses split dengan perbandingan 80:20. Sehingga jumlah dataset sebanyak 1.000 akan menjadi 800:200. Selanjutnya

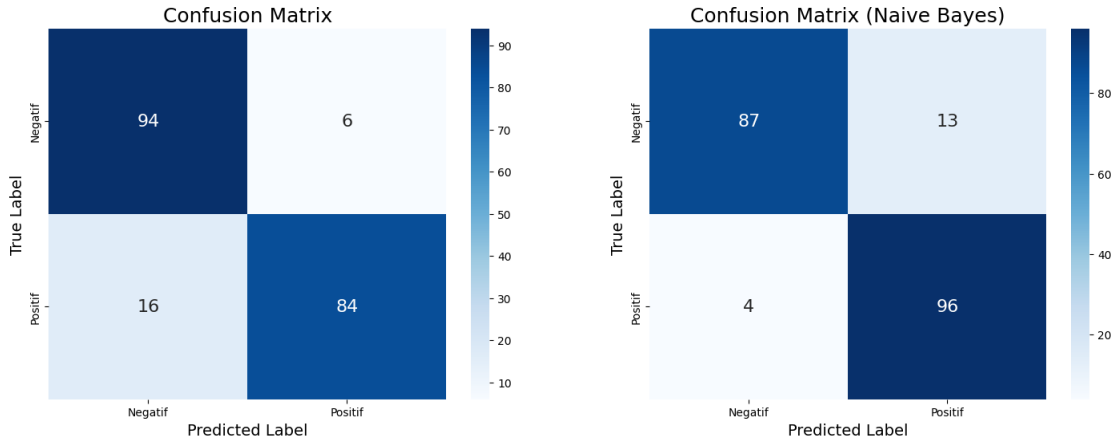
untuk memastikan output tidak terjadi *overfitting* maupun *underfitting*, validasi dengan *k-folds cross validation* dilakukan dengan k=5. Sehingga dataset dilakukan pengujian *training* dan *testing* pada 5 lipatan data yang berbeda. Performa algoritma klasifikasi diuji dengan metrik akurasi, precision, recall dan f-1 score. Hasilnya dapat dilihat pada tabel 1.

Tabel 1. Metrik Performa *Machine Learning* pada *K-folds Cross Validation*

	Logistic Regression				**Naïve Bayes**			
	Accuracy	**Precision**	**Recall**	**F1-Score**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Fold 1	0.8600	0.8703	0.8600	0.8596	0.9200	0.9240	0.9200	0.9196
Fold 2	0.9150	0.9151	0.9150	0.9148	0.9350	0.9411	0.9350	0.9352
Fold 3	0.8600	0.8661	0.8600	0.8587	0.9100	0.9131	0.9100	0.9101
Fold 4	0.8900	0.8976	0.8900	0.8906	0.8700	0.8698	0.8700	0.8695
Fold 5	0.8700	0.8723	0.8700	0.8697	0.9150	0.9247	0.9150	0.9146
Mean	0.8790	0.8843	0.8790	0.8787	0.9100	0.9145	0.9100	0.9098
Std Dev	0.0236	0.0212	0.0236	0.0239	0.0242	0.0269	0.0242	0.0244

Berdasarkan Tabel 1, model logistic regression mencapai nilai rata-rata akurasi sebesar 87,9% $\pm$ 2,36% sedangkan model naive bayes memperoleh rata-rata tingkat akurasi 91% $\pm$ 2,42%. Rata-rata akurasi yang tinggi ini mengindikasikan bahwa model TF-IDF dan Naive Bayes mampu membedakan sentimen positif dan negatif dengan sangat akurat. Selain itu, fold 2 memiliki tingkat akurasi 93,5% yang mengindikasikan nilai tertinggi pada fold ke 2 untuk data training dan testing ini tidak terlalu berbeda jauh dibandingkan dengan fold dataset yang lain. Sehingga kasus *overfitting* maupun *underfitting* tidak terjadi pada model klasifikasi ini. Nilai rata-rata precision dan recall juga memiliki nilai yang tidak berbeda jauh yaitu 91,45% dan 91%. Hal ini menunjukkan bahwa model naive bayes tidak bias pada salah satu label. Model ini tidak hanya akurat dalam menebak satu label tetapi juga mampu mengenali karakteristik ulasan pengguna dengan sentimen positif dan negatif dengan sama baiknya. Nilai *precision* dan *recall* yang tidak berbeda jauh ini juga mengakibatkan nilai f-1 score yang serupa yaitu 90,98%. Hal ini berarti model mampu merepresentasikan pola sentimen secara sangat baik tanpa adanya bias kategori sentimen tertentu.

Confusion Matrix Logistic Regression dan Naive Bayes



(a) (b)

Gambar 8. Confusion Matrix pada (a) Logistic Regression dan (b) Naïve Bayes

Gambar 8 mengilustrasikan confusion matrix dari model dengan algoritma logistic regression dan naïve bayes. Confusion matrix ini menggambarkan misklasifikasi yang terjadi pada data testing. Total data testing ada sebanya 200 data terdiri dari 100 label positif dan 100 label negatif. Terlihat pada gambar 8(a) label yang sebenarnya positif tetapi diprediksi label negatif masih terjadi misklasifikasi sebanyak 16 kali. Sedangkan label yang sebenarnya negatif namun diprediksi sebagai label positif terjadi misklasifikasi sebanyak 6 kali. Hal ini mengindikasikan

kecenderungan model dalam memprediksi lebih banyak label negatif yang menyebabkan nilai recall kelas positif lebih rendah dibandingkan kategori label negatif.

Pada Gambar 8(b) label yang sebenarnya positif tetapi diprediksi sebagai lable negatif terjadi sebanyak 4 kali. Sedangkan label yang sebenarnya negatif namun diprediksi sebagai label positif terjadi misklasifikasi sebanyak 13 kali. Hal ini berbanding terbalik dengan model logistic regression. Namun secara keseluruhan misklasifikasi prediksi terjadi lebih sedikit yaitu sebanyak 17 kali dibandingkan pada model logistic regression sebanyak 22 kali.

Deployment model dengan Gradio

Pada tahap deployment, model yang telah dilakukan proses data training, testing dan *k-folds cross validation* selanjutnya dapat diuji secara *real-time* dengan memasukkan data baru yang belum ada dalam dataset pengujian. Pada penelitian ini, desain sederhana untuk *user interface* diterapkan untuk demonstrasi model. Model naive bayes yang memiliki akurasi lebih tinggi dan total misklasifikasi yang lebih sedikit dipilih menjadi model machine learning pada deployment. Pada deployment model naive bayes menggunakan antar muka dari Gradio seperti yang ditunjukkan pada gambar 9.

Gambar 9. Desain *User Interface Model Naïve Bayes* dengan Gradio

Gambar 9 mengilustrasikan desain interface sederhana dari model klasifikasi sentimen dengan Gradio. Pada deployment dengan Gradio memiliki struktur dan fitur *user interface* antara lain *input field, processing, output field*, dan atribut identitas. Fitur *input field* bertujuan untuk memberikan input berupa teks ulasan/komentar pengguna. Pada bagian ini, teks akan secara otomatis dilakukan fungsi *pre-processing data* saat klik *submit*. Fitur *processing* bertujuan untuk menghapus (jika klik *clear*) atau memproses data (jika klik *submit*) untuk mengkalkulasi output label dengan algoritma logistic regression. Fitur *output field* bertujuan untuk menampilkan output dari pemrosesan data ke dalam salah satu label positif atau negatif. Sedangkan tampilan antar muka berupa logo memberikan identitas kepada *user interface* yang dibuat. Sebagai contoh penerapan, komentar baru berupa "pengaduan di aplikasi dibatalkan secara sepihak" ini akan memberikan output dengan sentimen negatif.

4. KESIMPULAN

Proses klasifikasi sentimen pada aplikasi mobile PLN di google playstore menggunakan data sebanyak 1.000 ulasan pengguna yang terdiri atas rating skor 1-3 sebanyak 500 data dan rating skor 4-5 sebanyak 500 data. Hal ini bertujuan untuk mendapatkan data yang seimbang dan beragam dalam kosa kata yang digunakan dalam memberikan ulasan. Dalam pengolahan data dengan TF-IDF didapatkan 10 kata terpenting pada label positif secara berurutan yaitu pln, membantu,

aplikasi, mobile, mantap, mudah, cepat, bagus, listrik, dan terbaik. Sedangkan pada label negatif 10 kata terpentingnya yaitu aplikasi, gak, nya, pln, listrik, pelayanan, pengaduan, yg, jam, dan ga. Setelah melakukan k-folds cross validation sebanyak k=5, didapatkan tingkat akurasi model logistic regression sebesar 87,9% $\pm$ 2,36% dan model naive bayes sebesar 91% $\pm$ 2,42%. Total misklasifikasi yang dihasilkan model naive bayes (17 kali) juga lebih sedikit dibandingkan model logistic regression (22 kali). Hal ini menunjukkan model mampu melakukan prediksi klasifikasi sentimen dengan sangat baik. Model selanjutnya di *deployment* dengan Gradio *User Interface* agar mempermudah proses trial pada input ulasan pengguna dan output klasifikasi.

Model klasifikasi sentiment ini memiliki batasan dimana penggunaan pelabelan dataset masih menggunakan metode *star rating* sehingga belum bisa mendeteksi ulasan sarkasme atau satire. Harapannya pada penelitian berikutnya, model dapat melakukan proses mendeteksi ulasan sarkasme dan satire pada tiap ulasan sehingga tidak bergantung pada pelabelan otomatis berdasarkan pemberian rating bintang *customer*.

UCAPAN TERIMAKASIH

Penulis mengucapkan terima kasih kepada DRTPM BIMA atas dana hibah penelitian dosen pemula. Selain itu, penulis juga mengucapkan terima kasih kepada pihak PLN dan pendidikan profesi insinyur ULM atas proyek penelitian terapan pada bidang keilmuan big data dan *data science*.

REFERENSI

[1] C. Gallagher, E. Furey, and K. Curran, "The application of sentiment analysis and text analytics to customer experience reviews to understand what customers are really saying," *Int. J. Data Warehous. Min.*, vol. 15, no. 4, pp. 21–47, 2019, doi: DOI: 10.4018/IJDWM.2019100102.

[2] Ming-Hui Huang and Roland T Rust, "Artificial Intelligence in Service," *J. Serv. Res.*, vol. 21, no. 2, pp. 155–172, Feb. 2018, doi: 10.1177/1094670517752459.

[3] D. McCauley, "The global AI agenda: Promise, reality, and a future of data sharing." MIT Technology Review Insights, 2020.

[4] P. Mikalef, M. Boura, G. Lekakos, and J. Krogstie, "Big data analytics and firm performance: Findings from a mixed-method approach," *J. Bus. Res.*, vol. 98, pp. 261–276, 2019.

[5] F. Scarcello, "Artificial Intelligence," S. Ranganathan, M. Gribskov, K. Nakai, and C. B. T.-E. of B. and C. B. Schönbach, Eds. Oxford: Academic Press, 2019, pp. 287–293.

[6] F. Barravecchia, L. Mastrogiacomo, and F. Franceschini, "Digital VoC analysis for product/service quality tracking in the era of Quality 4.0," *Int. J. Qual. Reliab. Manag.*, vol. 43, no. 2, pp. 499–515, 2025, doi: https://doi.org/10.1108/IJQRM-01-2025-0027.

[7] L. Owusu-Berko, "Harnessing big data, machine learning, and sentiment analysis to optimize customer engagement, loyalty, and market positioning," *Int. J. Comput. Appl. Technol. Res*, vol. 14, pp. 1–16, 2025.

[8] K. Hasanah, "Comparison of Sentiment Analysis Model for Shopee Comments on Google Play Store," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 13, no. 1, pp. 21–30, 2024.

[9] A. Deolika, K. Kusrini, and E. T. Luthfi, "Analisis pembobotan kata pada klasifikasi text mining," *J. Teknol. Inf.*, vol. 3, no. 2, pp. 179–184, 2019.

[10] K. Bhargava and R. Katarya, "An improved lexicon using logistic regression for sentiment analysis," in *2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN)*, 2017, pp. 332–337.

[11] N. Khoirunnisaa, K. N. N. Kesuma, S. Setiawan, and A. Y. P. Yusuf, "Klasifikasi Teks Ulasan Aplikasi Netflix Pada Google Play Store Menggunakan Algoritma Naive Bayes dan SVM," *SKANIKA Sist. Komput. dan Tek. Inform.*, vol. 7, no. 1, pp. 64–73, 2024.

[12] D. Wilimitis and C. G. Walsh, "Practical considerations and applied examples of cross-validation for model development and evaluation in health care: tutorial," *Jmir ai*, vol. 2, p. e49023, 2023.

[13] V. W. Lumumba, D. Kiprotich, M. Lemasulani Mpaine, N. Grace Makena, and M. Daniel Kavita, "Comparative analysis of cross-validation techniques: LOOCV, K-folds cross-validation, and repeated K-folds cross-validation in machine learning models," *K-folds Cross-Validation, Repeated K-folds Cross-Validation Mach. Learn. Model. (June 01, 2024)*, 2024.

[14] M. Wankhade, A. C. S. Rao, S. Dara, and B. Kaushik, "A sentiment analysis of food review using logistic regression," *Int J Sci Res Comput Sci Eng InformTechnol*, pp. 2–17, 2017.

[15] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: an overview," *arXiv Prepr. arXiv2008.05756*, 2020.

[16] S. Mokhtari, K. K. Yen, and J. Liu, "Effectiveness of artificial intelligence in stock market prediction based on machine learning," *arXiv Prepr. arXiv2107.01031*, 2021.

[17] A. Abid, A. Abdalla, A. Abid, D. Khan, A. Alfozan, and J. Zou, "Gradio: Hassle-free sharing and testing of ml models in the wild," *arXiv Prepr. arXiv1906.02569*, 2019.

[18] S. Al-Natour and O. Turetken, "A comparative assessment of sentiment analysis and star ratings for consumer reviews," *Int. J. Inf. Manage.*, vol. 54, p. 102132, 2020, doi: https://doi.org/10.1016/j.ijinfomgt.2020.102132.

[19] Y. Asri, W. N. Suliyanti, D. Kuswardani, and M. Fajri, "Pelabelan Otomatis Lexicon Vader dan Klasifikasi Naive Bayes dalam menganalisis sentimen data ulasan PLN Mobile," *PETIR J. Pengkaj. dan Penerapan Tek. Inf.*, vol. 15, no. 2, pp. 264–275, 2022.