Jutisi: Jurnal Ilmiah Teknik Informatika dan Sistem Informasi
Jl. Ahmad Yani, K.M. 33,5 - Kampus STMIK Banjarbaru
Loktabat – Banjarbaru (Tlp. 0511 4782881), e-mail: puslit.stmikbjb@gmail.com
e-ISSN: 2685-0893
p-ISSN: 2089-3787

Prediksi Sentimen dan Pemodelan Topik dari Ulasan Aplikasi Identitas Kependudukan Digital

Primandika Hakiki[1*], Amalia Anjani Arifiyanti[2], Dhian Satria Yudha Kartika[3]
Sistem Informasi, UPN "Veteran" Jawa Timur, Surabaya, Indonesia
*e-mail *Corresponding Author:* primandikakiki@gmail.com

Abstract

In this digital era, the Indonesian government is digitizing services, for example, the Digital Population Identity (IKD) in 2023 as a continuation of the Identity card (KTP). The IKD application has received mixed reviews on the Google PlayStore and Apple AppStore. From these various reviews, sentiment analysis can be carried out to determine the level of user satisfaction, as well as topic modeling to determine topics that are frequently discussed by users. This study aims to predict sentiment using the Naive Bayes and PNN methods, and topic modeling using the LDA method. The results of sentiment prediction show that the fourth scenario has the highest accuracy with 97.12%, and positive and negative f1-scores of 97% each. The results of topic modeling show the best coherence for positive reviews on topic 2 (0.4076) and for negative reviews on topic 10 (0.5564). Interpretation of the results shows that positive reviews include identity verification, application development suggestions, ease of use, KTP digitization, and user experience, while negative reviews are related to technical constraints, access and installation, network constraints, features and data security, and application registration.

Keywords*: Digital population identity; Sentiment predict; Topic modelling; Naïve Bayes; Probabilistic Neural Network*

Abstrak

Di era digital ini, pemerintah Indonesia melakukan digitalisasi layanan, contohnya yaitu Identitas Kependudukan Digital (IKD) pada tahun 2023 sebagai kelanjutan dari Kartu Tanda Penduduk (KTP). Aplikasi IKD mendapat ulasan beragam di *Google PlayStore* dan *Apple AppStore*. Dari beragam ulasan tersebut, maka dapat dilakukan analisis sentimen untuk mengetahui tingkat kepuasan pengguna, serta pemodelan topik untuk mengetahui topik yang sering dibahas oleh pengguna. Penelitian ini bertujuan untuk memprediksi sentimen menggunakan metode *Naive Bayes* dan PNN, serta pemodelan topik menggunakan metode LDA. Hasil prediksi sentimen menunjukkan skenario keempat memiliki akurasi tertinggi dengan 97,12%, serta *f1-score* positif dan negatif masing-masing 97%. Hasil pemodelan topik menunjukkan koherensi terbaik untuk ulasan positif pada topik 2 (0,4076) dan untuk ulasan negatif pada topik 10 (0,5564). Interpretasi hasil menunjukkan bahwa ulasan positif mencakup verifikasi identitas, saran pengembangan aplikasi, kemudahan penggunaan, digitalisasi KTP, dan pengalaman penggunaan, sedangkan ulasan negatif terkait kendala teknis, akses dan instalasi, kendala jaringan, fitur dan keamanan data, dan pendaftaran aplikasi.

Kata kunci: *Identitas Kependudukan Digital; Prediksi sentimen; Pemodelan topik; Naïve Bayes; Probabilistic Neural Network*

1. Pendahuluan

Di era yang serba digital seperti saat ini, pemerintah dituntut untuk melakukan digitalisasi dalam penyediaan layanan kepada masyarakat. Teknologi Informasi (TI) tidak hanya berfungsi sebagai alat bantu bagi petugas dalam melaksanakan tugas sehari-hari, akan tetapi juga berperan penting dalam meningkatkan kepuasan masyarakat melalui pelayanan yang lebih optimal [1]. Salah satu contoh digitalisasi yang dilakukan oleh pemerintah Indonesia adalah pada Kartu Tanda Penduduk (KTP). Pada tahun 2023, pemerintah meluncurkan Identitas Kependudukan Digital (IKD) sebagai kelanjutan langkah digitalisasi pada KTP. Analisis sentimen merupakan metode yang digunakan untuk mengidentifikasi dan memahami ekspresi

atau respons pengguna terhadap suatu aplikasi berdasarkan komentar yang mereka berikan [2]. Tujuan dari analisis sentimen ini adalah untuk mengklasifikasikan teks atau kalimat ke dalam kategori sentimen, seperti positif maupun negatif, serta mengukur kepuasan pelanggan, memantau merek, menganalisis opini, serta mengevaluasi respons di media sosial [3].

Berdasarkan pantauan di *Google PlayStore* dan *Apple AppStore*, aplikasi IKD mendapat beragam ulasan. Beberapa pengguna memuji proses aktivasi dan akses aplikasi yang cepat, kemudahan mengakses data seperti KTP, KK, KIA, dan Akta Kelahiran, serta manfaatnya bagi masyarakat. Namun, ada juga ulasan negatif terkait aplikasi yang hanya bisa dibuka dengan jaringan *Wi-Fi*, proses aktivasi yang masih harus dilakukan di kecamatan atau Dispendukcapil, serta antarmuka yang kurang optimal. Berdasarkan ulasan tersebut, dapat dilakukan analisis sentimen untuk menilai tingkat kepuasan pengguna dan pemodelan topik untuk mengidentifikasi topik yang sering dibahas.

Dalam melakukan analisis sentimen, teknik klasifikasi dokumen yang akan digunakan pada penelitian ini adalah *Naïve Bayes Classifier* (NBC) dan *Probabilistic Neural Network* (PNN). Pemilihan metode *Naïve Bayes Classifier* didasarkan pada kemampuannya yang baik untuk melakukan klasifikasi dalam hal komputasi data dan presisi [4]. Sedangkan metode *Probabilistic Neural Network* (PNN) dipilih karena kerap digunakan dalam penelitian yang melibatkan dataset berupa teks maupun gambar, serta mampu memberikan akurasi yang baik dalam kasus analisis klasifikasi [5]. Kemudian untuk pemodelan topik, metode yang akan digunakan adalah *Latent Dirichlet Allocation* (LDA). Pemilihan metode *Latent Dirichlet Allocation (*LDA) didasarkan pada kemampuannya dalam mengidentifikasi topik secara akurat pada dataset yang cukup besar [6].

Hasil dari penelitian ini adalah hasil analisis sentimen dan pemodelan topik, perbandingan antara metode *Naïve Bayes* dan *Probabilistic Neural Network*, dan *website* yang berisi visualisasi berupa grafik dari analisis sentimen dan pemodelan topik yang sudah dilakukan. Dari penelitian ini, diharapkan dapat mengetahui sentimen positif dan negatif dari ulasan pengguna aplikasi IKD, mengetahui topik apa saja yang sering dibahas oleh pengguna aplikasi IKD, dan juga dapat menjadi bahan pertimbangan bagi Kemenkominfo, Kemendagri, maupun pihak-pihak terkait lainnya dalam pengembangan aplikasi IKD kedepannya.

2. Tinjauan Pustaka

Penelitian yang dilakukan oleh Ahmad Komarudin dan Atiqah Meutia Hilda membahas tentang Analisis Sentimen Ulasan Aplikasi Identitas Kependudukan Digital Pada *Play Store* Menggunakan Metode *Naïve Bayes*. Penelitian tersebut menggunakan algoritma *Naïve Bayes* dengan data uji 90% dan data latih 10% per-iterasinya. Berdasarkan hasil analisis sentimen, ditemukan bahwa jumlah sentimen negatif lebih dominan dibandingkan sentimen positif, yang mengindikasikan bahwa tanggapan pengguna terhadap aplikasi Identitas Kependudukan Digital di *Play Store* cenderung kurang memuaskan. Hal ini menunjukkan bahwa pemerintah belum sepenuhnya siap dalam menyediakan layanan identitas digital secara optimal. Selain itu, penerapan *SMOTE Upsampling* dalam proses klasifikasi terbukti meningkatkan kinerja model *Naïve Bayes*, dengan capaian akurasi sebesar 85,06%, *precision* sebesar 80,31%, dan *recall* sebesar 92,89%. Dengan hasil ini, dapat disimpulkan bahwa model klasifikasi yang digunakan mampu mengelompokkan ulasan pengguna secara efektif terhadap aplikasi tersebut [7].

Penelitian yang dilakukan Ni Luh Putu Merawati, Ahmad Zuli Amrullah, dan Ismarmiaty membahas tentang Analisis Sentimen dan Pemodelan Topik Pariwisata Lombok Menggunakan Algoritma *Naive Bayes* dan *Latent Dirichlet Allocation*. Penelitian tersebut menunjukkan bahwa algoritma *Naive Bayes* menghasilkan nilai akurasi, presisi, *recall* dan spesifisitas masing-masing sebesar 92%, 100%, 83,84% dan 100%. Hasil pemodelan topik dengan metode *LDA* pada masing-masing kelas positif dan negatif ditunjukkan dari nilai koherensi yaitu semakin tinggi nilai koherensi dari suatu topik, maka semakin mudah topik tersebut diinterpretasikan oleh manusia. Nilai koherensi tertinggi untuk kelas positif diperoleh pada topik ke-8 dengan nilai sebesar 0,613 dan untuk kelas negatif pada topik ke-12 dengan nilai sebesar 0,528. Dari hasil interpretasi, topik yang sering dibahas oleh masyarakat pada kelas positif yaitu tentang keindahan pantai di Pulau Lombok, terutama Gili Trawangan dan Pantai Senggigi. Sementara pada kelas negatif, topik yang sering dibahas adalah mengenai sampah [8].

Penelitian yang dilakukan Fikri Fahru Roji, Nava Gia Ginasta, Yayan Cahyan, Dinar Rahayu, dan Dendi Ramdani membahas tentang *Review Analysis of SatuSehat Application Using Support Vector Machine and Latent Dirichlet Allocation Modeling*. Hasil penelitian

tersebut menunjukkan bahwa pada analisis sentimen, diperoleh 171.428 ulasan positif, 131.246 ulasan negatif dan 20.819 ulasan netral, yang mengindikasikan bahwa lebih banyak pengguna memberikan tanggapan positif daripada tanggapan negatif. Pemodelan topik dengan algoritma *LDA* menghasilkan 4 topik ulasan positif dan 4 topik ulasan negatif. Topik positif yang dihasilkan berkaitan dengan perhatian terhadap kesehatan, manfaat aplikasi, peran aplikasi dalam vaksinasi, dan ungkapan apresiasi. Sementara itu, topik negatif menyoroti tantangan dalam vaksinasi dan pendaftaran aplikasi, kekhawatiran tentang data pribadi, masalah akses ke aplikasi, dan masalah fungsionalitas [9].

Penelitian yang dilakukan Ritham Tuntun, Kusrini, dan Kusnawi membahas tentang Analisis Perbandingan Kinerja Algoritma Klasifikasi dengan Menggunakan Metode *K-Fold Cross Validation*. Hasil penelitian tersebut menunjukkan bahwa algoritma *C4.5* memiliki kinerja lebih baik dengan akurasi sebesar 96,00% dibandingkan dengan algoritma *K-Nearest Neighbor* dengan akurasi sebesar 94.44%. Sedangkan untuk tingkat presisi algoritma *C4.5* sebesar 93,52% dan *K-Nearest Neighbor* sebesar 93,70% [10].

Walaupun berbagai metode telah digunakan, penelitian sebelumnya belum banyak mengeksplorasi algoritma *Probabilistic Neural Network*, yang dikenal memberikan akurasi yang baik dalam kasus analisis klasifikasi. Selain itu, penelitian terdahulu tentang analisis sentiment ulasan pada aplikasi IKD masih terbatas pada penggunaan metode *Naïve Bayes*, tanpa membandingkan kinerjanya dengan metode lain seperti *PNN*, yang memiliki potensi untuk memberikan hasil yang kompetitif. Pada penelitian ini memiliki fokus dalam penerapan *Naïve Bayes* dan *Probabilistic Neural Network* untuk klasifikasi sentimen dan *Latent Dirichlet Allocation* untuk pemodelan topik dari ulasan pengguna Identitas Kependudukan Digital, yang sebelumnya belum banyak dikaji. Selain itu, penelitian ini akan mengoptimalkan model dengan teknik *preprocessing* teks yang lebih lengkap, seperti *case folding*, *normalisasi kata*, *cleansing*, *tokenizing*, *stopword removal*, dan *stemming*, dan pembobotan *TF-IDF* serta *TF-RF*. Evaluasi akan dilakukan berdasarkan akurasi, presisi, *recall*, dan *F1-score* untuk pemodelan sentimen, serta nilai koherensi untuk pemodelan topik.

3. Metodologi

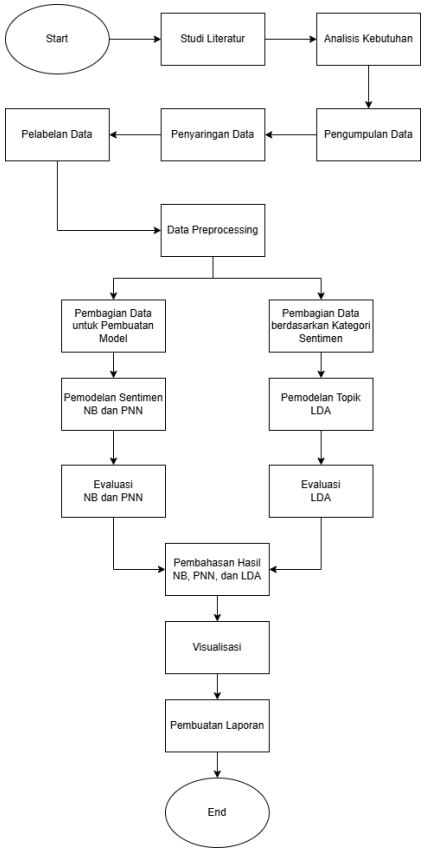
Gambar 1 merupakan kerangka metode penelitian yang akan digunakan dalam melakukan prediksi sentimen dan pemodelan topik. Setiap tahapan yang akan dilakukan memiliki pengertian dan tujuan masing-masing.

Studi literatur adalah kegiatan yang dilakukan untuk mengumpulkan informasi yang relevan. Studi literatur dilakukan dengan mengumpulkan informasi melalui pencarian pada berbagai buku, arsip, artikel, jurnal, dokumen, maupun sumber lain yang membahas analisis sentimen, penggunaan metode *Naïve Bayes Classifier* (NBC) dan *Probabilistic Neural Network* (PNN), *Latent Dirichlet Allocation* (LDA), dan referensi lain yang terkait dengan penelitian ini.

Analisis kebutuhan bertujuan untuk menjelaskan kebutuhan data dan kebutuhan *software* serta *hardware* yang diperlukan agar pelaksanaan penelitian dapat berjalan dengan optimal. Data yang dibutuhkan dalam penelitian ini adalah data ulasan aplikasi Identitas Kependudukan Digital (IKD) berbahasa Indonesia yang diambil dari *Google PlayStore* dan *Apple AppStore* region Indonesia, dengan rentang waktu pengambilan dari bulan Juni 2022 hingga bulan Mei 2024. Sedangkan untuk kebutuhan *software* dan *hardware* yang digunakan dalam penelitian ini adalah *browser* untuk menjalankan program, bahasa pemrograman *Python* untuk pemodelan data, serta laptop Asus Vivobook S14 S430UN dengan RAM sebesar 12GB.

Proses pengumpulan data dilakukan pada *Google PlayStore* dan *Apple AppStore* dengan cara memasukkan nama *package* aplikasi Identitas Kependudukan Digital (IKD) dan menetapkan bahasa “id” agar hanya memperoleh ulasan dalam bahasa Indonesia saja. Pengumpulan data ini menggunakan *library “google-play-scraper”* dan “*app_store_scraper”* karena keduanya dapat mengambil ulasan dari aplikasi yang tersedia pada *Google PlayStore* maupun *Apple AppStore*. Jumlah data yang berhasil terkumpul yaitu 1393 ulasan dari *Google PlayStore* dan 399 ulasan dari *Apple AppStore*, sehingga total keseluruhan data yang diperoleh berjumlah 1792 ulasan.

Setelah data terkumpul, data kemudian akan disaring atau difilter. Data akan disaring agar data yang terkumpul sesuai dengan keperluan dalam skripsi ini. Tahap penyaringan diperlukan karena masih terdapat kemungkinan ulasan yang kurang berhubungan dengan aplikasi IKD.



Gambar 1. Tahapan Metode Penelitian

Tahap selanjutnya yaitu pelabelan data. Skripsi ini menerapkan metode *supervised learning* yang berarti ada data yang sudah terkumpul dan ada variabel yang sudah ditentukan. Tahap pelabelan data diperlukan untuk memberikan nilai/variabel pada setiap ulasan. Pelabelan data dalam skripsi ini akan dilakukan secara manual berdasarkan skala rating, yaitu *rating* 1, 2, dan 3 diberi label sentimen negatif, serta *rating* 4 dan 5 diberi label positif [11].

Data yang sudah diberi label kemudian akan melalui proses yang dinamakan text preprocessing. Tahap text preprocessing ini berfungsi untuk mengubah bentuk data tekstual dari yang awalnya tidak terstruktur menjadi terstruktur [12] yang prosesnya terdiri dari *case folding*, *normalisasi kata*, *cleansing*, *tokenizing*, *stopword removal*, dan *stemming*.

Setelah data selesai diproses, tahap selanjutnya adalah pembagian data. Dalam pembagian data untuk pembuatan model analisis sentimen, metode yang akan digunakan ada dua, yaitu Metode *Hold Out* dan *K-Fold Cross Validation*. Metode *Hold Out* akan membagi *dataset* menjadi dua yaitu data latih (*training*) sebesar 80% dan data uji (*testing*) sebesar 20 %. Sedangkan metode *k-fold cross validation* data akan dibagi menjadi ‘k’ bagian *set* data dengan ukuran yang sama besarnya, yaitu 5 bagian.

Data yang sudah diuji selanjutnya akan diberi bobot dengan metode *TF-IDF* dan *TF-RF*. *TF-IDF* menggabungkan antara frekuensi kemunculan suatu kata dalam sebuah dokumen (TF) dan frekuensi dokumen yang mengandung suatu kata tersebut (IDF)[13]. Sedangkan pada *TF-RF* mempertimbangkan relevansi dokumen berdasarkan frekuensi kemunculan *term* pada kategori yang berkaitan [14].

Dalam pembagian data berdasarkan kategori sentimen untuk pemodelan topik, data yang sudah didapatkan akan diberi label secara manual berdasarkan skala rating. Data yang memiliki *rating* 1, 2, dan 3 diberi label sentimen negatif, serta *rating* 4 dan 5 diberi label sentimen positif.

Teknik klasifikasi yang akan digunakan dalam penelitian ini adalah *Naïve Bayes* dengan kelas *multinomial* dan *Probabilistic Neural Network* (PNN). Berikut adalah persamaan *Naïve Bayes* [15]:

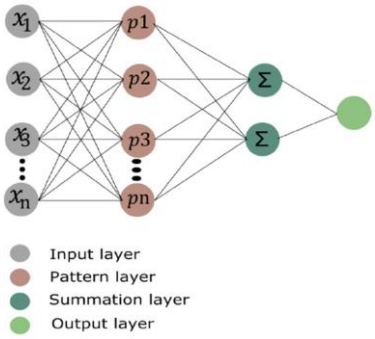
$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad \text{..... (1)}$$

Keterangan:

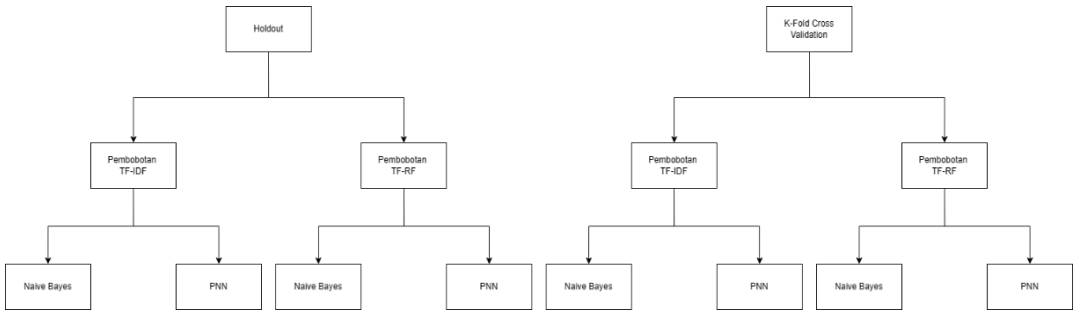
X	: Data yang *class*-nya yang belum diketahui
H	: Hipotesis data *X* merupakan suatu *class* spesifik
P(H|X)	: Probabilitas hipotesis *H* berdasar kondisi *X* (*posterior probability*)
P(H)	: Probabilitas hipotesis *H* (*prior probability*)
P(X|H)	: Probabilitas *X* berdasarkan kondisi pada hipotesis *H*
P(X)	: Probabilitas *X*

PNN merupakan penyelesaian dari algoritma statistik, yang disebut analisis diskriminasi kernel yang prosesnya disusun menjadi jaringan feed-forward berlapis-lapis yang terdiri dari empat lapisan [16], yaitu:

1) ***Input Layer***
   Layer ini menerima vektor input dari data yang akan diklasifikasi dan kemudian ditransmisikan ke *layer* berikutnya.
2) ***Pattern Layer***
   Di *layer* ini, setiap neuron mewakili data latih dan menghitung nilai kernel untuk mengukur persamaan antara vektor input dan data latih.
3) ***Summation Layer***
   Layer ini mengumpulkan hasil dari neuron pada *Pattern Layer* dan menghitung total dari tiap kelas yang ada dalam data latih.
4) ***Output Layer***
   Layer ini adalah bagian terakhir dari jaringan dan menghasilkan *output* klasifikasi/estimasi probabilitas.



Dalam proses klasifikasi, terdapat delapan skenario yang akan digunakan seperti yang dapat dilihat pada gambar 2 berikut. Untuk pemodelan topik, metode yang akan digunakan adalah *Latent Dirichlet Allocation* (LDA).



Gambar 3. Skenario Prediksi Sentimen

Evaluasi model sentimen akan dilakukan dengan menggunakan metode *confussion matrix* yang merupakan sebuah tabel yang mengklasifikasikan jumlah data uji yang benar dan jumlah data uji yang salah [17]. Sedangkan untuk pemodelan topik, nilai koherensi dapat digunakan dalam menentukan jumlah topik yang tepat dari suatu dokumen. Semakin tinggi nilai koherensi yang dihasilkan, maka pemodelan topik tersebut akan semakin bagus hasil interpretasinya.

Setelah melalui proses analisis sentimen dan pemodelan topik, maka akan diketahui bagaimana hasil sentimen pengguna dan topik apa saja yang sering dibahas pengguna dari ulasan pada aplikasi Identitas Kependudukan Digital (IKD).

Dalam tahap visualisasi, hasil dari klasifikasi akan dibuat tampilannya secara visual. Tahap visualisasi ini bertujuan untuk memudahkan pembaca dalam memahami hasil klasifikasi dan pemodelan topik. Hasil klasifikasi dan pemodelan topik akan divisualisasikan dalam bentuk grafik. Grafik tersebut akan diimplementasikan kedalam *website* yang dibuat dengan menggunakan pemrograman PHP *native*. Selain itu, terdapat juga fitur untuk memprediksi label sentimen dan topik dari ulasan yang dimasukkan ke dalam inputan.

4. Hasil dan Pembahasan

4.1 Implementasi dan Pengujian Algoritma

Dataset yang digunakan berasal dari proses pengumpulan data menggunakan *library google_play_scraper* dan *app_store_scraper*. Data yang didapatkan dari *Apple AppStore* berjumlah 399 data, sedangkan yang didapatkan dari *Google PlayStore* berjumlah 1393 data. Tabel 1 dan Tabel 2 memperlihatkan contoh data yang berhasil terambil sebelum melalui tahap *preprocessing*.

Tabel 1. Hasil *Scraping AppStore*

date	review	rating	isEdited	userName
25-05-23 03:31	Mantap	5	FALSE	tans199666383882
25-05-23 04:06	Memudahkan sekali	5	FALSE	Humaray

Tabel 2. Hasil *Scraping PlayStore*

at	content	score	userName
01-05-24 08:41	Apasih gunanya E-KTP tapi gak bisa cetak KTP-nya sesuai aplikasinya, masuk ke menu utama selalu kesalahan koneksi padahal jaringan 4G, dan tidak diberikan tutorial diawal sehingga membuat bingung usernya. KALAU TIDAK NIAT BUAT APLIKASI, GAUSAH DIPAKSAKAN LAH! Katanya ingin mempermudah masyarakat Indonesia, tapi malah 'PEMBODOHAN BERKEDOK DOWNLOAD APLIKASI'	1	Khairil Aryan TV
13-05-24 13:41	Aku kasih bintang 3 karna mau masuk aja harus pake wifi . Pakai data seluler gak mau nyambung koneksi bermasalah . Untuk lain nya sudah bagus untuk barcode nya meski ribet ke kantor desa oke gak masalah karna untuk alasan menjamin keamanan data . Mohon perbaikan untuk aplikasi agar bisa konek dengan sambungan data selular saja.	3	Abdul Muttolib

Setelah pengambilan data, tahap selanjutnya adalah penyaringan data yang berfungsi untuk menyortir data-data yang memang diperlukan dalam penelitian ini, karena tidak semua data yang sudah dikumpulkan membahas tentang aplikasi IKD. Jumlah data setelah disaring adalah menjadi 1783 data.

Data yang sudah disaring kemudian akan dilabeli. Pelabelan data dilakukan secara manual dengan membaca setiap ulasan dan memberi label sentimen positif atau negatif. Tiga anotator, termasuk peneliti, memberi label, lalu dipilih berdasarkan *majority voting*. Jika ketiganya berbeda, peneliti menentukan label akhir, yang berpotensi menimbulkan bias karena dipengaruhi pemahaman peneliti. Tabel 3 memperlihatkan contoh data yang sudah diberi label.

Tabel 3. Pelabelan Data

Ulasan	Score	Label
Seriusan apanya yang online kalo tetep disuruh dateng ke dukcapil?, sia sia gw bayar pajak tiap tahun cuma buat ngempanin orang2 tolol yg kerja di dukcapil 🤣🤣🤣	1	Neg
Katanya digital tapi masih diharuskan datang ke dukcapil untuk scan QR, seharusnya pada saat registrasi itu QRnya otomatis di kirim ke email. Mau digital elit mau QR sulit 🤨	1	Neg

Setelah data diberi label, proses selanjutnya yaitu *preprocessing*. Proses ini terdiri dari enam tahap, yaitu *case folding*, normalisasi kata, *cleansing*, *tokenizing*, *stopword removal*, dan *stemming*. Tabel 4 menunjukkan hasil data yang sudah melewati tahap *preprocessing*

Tabel 4. Hasil *Preprocessing*

Proses	Hasil
Case Folding	seriusan apanya yang online kalo tetep disuruh dateng ke dukcapil?, sia sia gw bayar pajak tiap tahun cuma buat ngempanin orang2 tolol yg kerja di dukcapil 🤣🤣🤣
Normalisasi Kata	seriusan apanya yang online kalo tetep disuruh dateng ke dukcapil?, sia sia saya bayar pajak tiap tahun cuma buat ngempanin orang2 tolol yg kerja di dukcapil 🤣🤣🤣
Cleansing	seriusan apanya yang online kalo tetep disuruh dateng ke dukcapil sia sia gw bayar pajak tiap tahun cuma buat ngempanin tolol yg kerja di dukcapil
Tokenizing	'seriusan', 'apanya', 'yang', 'online', 'kalo', 'tetep', 'disuruh', 'dateng', 'ke', 'dukcapil', 'sia', 'sia', 'gw', 'bayar', 'pajak', 'tiap', 'tahun', 'cuma', 'buat', 'ngempanin', 'tolol', 'yg', 'kerja', 'di', 'dukcapil'
Stopword Removal	seriusan online tetep disuruh dateng dukcapil sia sia gw bayar pajak ngempanin tolol dukcapil
Stemming	serius online tetep suruh dateng dukcapil sia sia gw bayar pajak ngempanin tolol dukcapil

Setelah tahap *preprocessing*, data dibagi untuk pembuatan model sentimen. Namun, terdapat *imbalanced class* dari hasil pelabelan sebelumnya yaitu 1650 data negatif dan 133 data positif. Untuk mengatasinya, digunakan metode *oversampling* data dengan menggunakan teknik *Random Over Sampling* (ROS). Total data setelah di-*oversampling* adalah 1650 data negatif dan 1650 data positif.

Untuk tahap pembagian data menggunakan metode *holdout* dimana data dibagi menjadi 80% untuk *training* dan 20% untuk *testing*, serta menggunakan metode *k-fold* dengan jumlah *fold* sebesar 5.

```python
import pandas as pd
from sklearn.model_selection import train_test_split

# 1. Baca file CSV
data = pd.read_csv('df_ulasan_balanced.csv')

# 2. Bagi data: 80% training, 20% testing
train_data, test_data = train_test_split(data, test_size=0.2, random_state=42)

# 3. Simpan ke file CSV baru
train_data.to_csv('data_train.csv', index=False)
test_data.to_csv('data_test.csv', index=False)

print('Data berhasil dibagi:')
print(f'- Data training: {train_data.shape[0]} data')
print(f'- Data testing: {test_data.shape[0]} data')
```

```python
import pandas as pd
from sklearn.model_selection import KFold

# 1. Baca data
data = pd.read_csv('df_ulasan_balanced.csv')

# 2. Inisialisasi KFold (misal 5-fold)
kf = KFold(n_splits=5, shuffle=True, random_state=42)

# 3. Loop tiap fold
fold = 1
for train_index, test_index in kf.split(data):
    train_data = data.iloc[train_index]
    test_data = data.iloc[test_index]

    # Simpan tiap fold ke file (opsional)
    train_data.to_csv(f'data_train_fold{fold}.csv', index=False)
    test_data.to_csv(f'data_test_fold{fold}.csv', index=False)

    print(f'Fold {fold}: Train = {len(train_data)}, Test = {len(test_data)}')
    fold += 1
```

Gambar 4. Source Code Pembagian Data

Pada Gambar 4 di atas menunjukkan bahwa pada metode *holdout* data dibagi menjadi 80% untuk *training* dan 20% untuk *testing*. Kemudian untuk pada metode *k-fold* menunjukkan bahwa *library* yang digunakan adalah *sklearn.model* selection fungsi *KFold* dengan jumlah *fold* sebesar 5. Setelah itu data diberi pembobotan yang menggunakan metode *TF-IDF* dan *TF-RF*.

Data yang sudah diboboti kemudian diproses menggunakan metode *Naïve Bayes* dan *PNN*. Kemudian evaluasi model dilakukan dengan menghitung nilai *accuracy*, *precision*, *recall*, *f1-score*, dan *support*. Tabel 5 berikut ini merupakan hasil dari pemrosesan sentimen beserta evaluasinya.

Tabel 5. Hasil Klasifikasi Sentimen

Skenario			Akurasi	Waktu Proses	Precision	Recall	F1-Score	Support	Nomor Skenario
Holdout	TF-IDF, Naïve Bayes	Negatif	96,06%	0,115	100%	92%	96%	343	1
		Positif			92%	100%	96%	317	
	TF-RF, Naïve Bayes	Negatif	96,36%	0,006	98%	94%	96%	343	2
		Positif			94%	98%	96%	317	
	TF-IDF, PNN	Negatif	60,60%	36,18	100%	24%	39%	343	5
		Positif			55%	100%	71%	317	
	TF-RF, PNN	Negatif	72,57%	37,02	100%	47%	64%	343	6
		Positif			64%	100%	78%	317	
K-Fold Cross Validation	TF-IDF, Naïve Bayes	Negatif	96,81%	0,052	100%	94%	97%	332	3
		Positif			94%	100%	97%	328	
	TF-RF, Naïve Bayes	Negatif	97,12%	0,014	99%	95%	97%	332	4
		Positif			96%	99%	97%	328	
	TF-IDF, PNN	Negatif	70,60%	32,32	100%	42%	59%	332	7
		Positif			63%	100%	77%	328	
	TF-RF, PNN	Negatif	78,33%	33,97	100%	57%	73%	332	8
		Positif			70%	100%	82%	328	

Dari tabel 5, dapat diketahui bahwa skenario ke-4 yaitu *Naïve Bayes* yang pembagian datanya menggunakan *k-fold cross validation* dan diberi pembobotan *TF-RF* memiliki akurasi paling tinggi yaitu 97,12% dengan waktu proses 0,014 detik. Model tersebut juga memiliki nilai *f1-score* yang tinggi dengan nilai positif 97% dan nilai negatif 97%.

Kemudian untuk pemodelan topik, data yang digunakan adalah data yang telah di-*oversampling* yaitu 1650 data negatif dan 1650 data positif. Pembentukan model *LDA* menggunakan *library gensim*. Hasil pemodelan topik ditunjukkan pada tabel 6.

Tabel 6. Hasil Pemodelan Topik

Sentimen	Topik	Hasil Nilai-Nilai Koherensi	Perkiraan Tema
Positif	1	0.071*"aplikasi" + 0.034*"ktp" + 0.032*"dukcapil" + 0.027*"bagus" + 0.024*"bantu" + 0.023*"digital" + 0.022*"scan" + 0.022*"data" + 0.015*"barcode" + 0.014*"identitas"	Verifikasi Identitas
Positif	2	0.043*"dukcapil" + 0.032*"aplikasi" + 0.030*"ktp" + 0.029*"layan" + 0.028*"moga" + 0.019*"digital" +	Saran Pengembangan Aplikasi

Sentimen	Topik	Hasil Nilai-Nilai Koherensi	Perkiraan Tema
		0.018*"bermanfat" + 0.018*"mudah" + 0.018*"bawa" + 0.016*"qr"	
Positif	3	0.037*"aplikasi" + 0.030*"layan" + 0.029*"bantu" + 0.024*"dokumen" + 0.024*"ktp" + 0.021*"mudah" + 0.019*"bagus" + 0.016*"data" + 0.014*"cepat" + 0.014*"dg"	Kemudahan Penggunaan
Positif	4	0.040*"aplikasi" + 0.038*"digital" + 0.033*"data" + 0.029*"ktp" + 0.028*"mudah" + 0.022*"mantap" + 0.017*"dukcapil" + 0.017*"bawa" + 0.016*"fitur" + 0.016*"daftar"	Digitalisasi KTP
Positif	5	0.045*"aplikasi" + 0.036*"mudah" + 0.021*"bagus" + 0.021*"ktp" + 0.018*"login" + 0.016*"data" + 0.016*"kembang" + 0.015*"bawa" + 0.015*"dokumen" + 0.014*"verifikasi"	Pengalaman Penggunaan (UX)
Negatif	6	0.031*"aplikasi" + 0.025*"ktp" + 0.019*"digital" + 0.010*"apk" + 0.010*"foto" + 0.010*"tolong" + 0.010*"data" + 0.009*"hp" + 0.009*"mudah" + 0.008*"baik"	Kendala Teknis Aplikasi
Negatif	7	0.051*"aplikasi" + 0.027*"hp" + 0.019*"data" + 0.014*"daftar" + 0.013*"buka" + 0.012*"tolong" + 0.011*"pake" + 0.011*"android" + 0.010*"ktp" + 0.010*"download"	Akses & Instalasi Aplikasi
Negatif	8	0.079*"aplikasi" + 0.049*"buka" + 0.021*"koneksi" + 0.020*"jaring" + 0.016*"baik" + 0.015*"bagus" + 0.015*"salah" + 0.014*"tolong" + 0.014*"pakai" + 0.013*"wifi"	Kendala Jaringan
Negatif	9	0.020*"aplikasi" + 0.012*"data" + 0.011*"fitur" + 0.009*"layan" + 0.007*"capil" + 0.006*"dokumen" + 0.006*"foto" + 0.006*"salah" + 0.005*"aman" + 0.005*"akses"	Fitur dan Keamanan Data
Negatif	10	0.048*"dukcapil" + 0.034*"ktp" + 0.034*"scan" + 0.032*"digital" + 0.026*"barcode" + 0.022*"qr" + 0.020*"online" + 0.019*"aplikasi" + 0.019*"mudah" + 0.017*"daftar"	Pendaftaran Aplikasi

Dari tabel 6, Pada label positif, beberapa kategori seperti verifikasi identitas, kemudahan penggunaan, dan pengalaman penggunaan (UX) menunjukkan bahwa sebagian pengguna merasa aplikasi ini sudah memberikan kemudahan dalam mengakses layanan kependudukan secara digital. Kategori seperti saran pengembangan aplikasi juga menunjukkan adanya keterlibatan aktif dari pengguna untuk memberikan masukan konstruktif, yang menandakan bahwa mereka memiliki harapan terhadap perbaikan dan keberlanjutan layanan ini. Adapun digitalisasi KTP mencerminkan dukungan terhadap inovasi pemerintah dalam memodernisasi sistem administrasi kependudukan. Sementara itu, pada label negatif, temuan seperti kendala teknis aplikasi, akses & instalasi, dan kendala jaringan menunjukkan adanya hambatan signifikan dalam proses penggunaan aplikasi, baik dari sisi perangkat lunak maupun

infrastruktur. Masalah terkait fitur dan keamanan data menandakan kekhawatiran pengguna terhadap perlindungan privasi dan keandalan sistem, yang merupakan aspek krusial dalam layanan digital. Selain itu, kesulitan dalam pendaftaran aplikasi menunjukkan bahwa proses awal penggunaan belum cukup ramah pengguna, yang bisa menjadi penghalang adopsi lebih luas.

Karena pemodelan topik termasuk ke dalam *unsupervised technique*, maka evaluasi model dilakukan dengan menggunakan nilai koherensi topik. Semakin tinggi nilai koherensi suatu topik, maka semakin mudah topik tersebut dipahami berdasarkan kumpulan kata yang membentuknya. Dengan kata lain, jika kata-kata dalam topik tersebut sering muncul secara bersamaan, maka nilai koherensinya juga semakin tinggi. Hasil evaluasi dapat dilihat pada tabel 7.

Tabel 7. Hasil Evaluasi Pemodelan Topik

Topik Positif	Nilai Koherensi	Topik Negatif	Nilai Koherensi
1	0.4032	6	0.3411
2	0.4076	7	0.4361
3	0.3760	8	0.4732
4	0.3457	9	0.3832
5	0.3734	10	0.5564

Dari Tabel 7, dapat diketahui bahwa nilai koherensi terbaik untuk kelas positif diperoleh topik ke-2 dengan nilai 0,4076. Sedangkan nilai koherensi terbaik untuk kelas negatif diperoleh topik ke-10 dengan nilai 0,5564. Ini berarti topik tersebut paling konsisten dan mudah dipahami dalam kelompoknya. Topik ke-2 mewakili tema positif yang paling relevan, meskipun ulasan positif cenderung lebih beragam. Sementara topik ke-10 mencerminkan keluhan utama pengguna secara lebih terfokus.

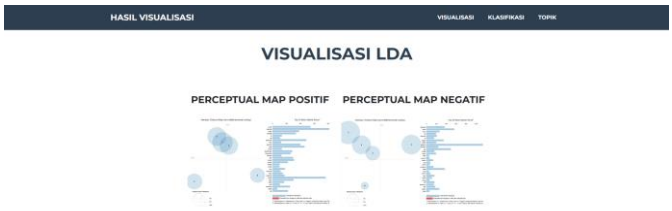
Langkah yang terakhir adalah melakukan visualisasi untuk menampilkan data yang sudah dianalisis dan membuat fitur prediksi label dan topik dari ulasan atau kata yang dimasukkan. Pada proses visualisasi ini, skenario yang digunakan dalam memprediksi sentimen adalah skenario ke-4 yaitu data yang dibagi dengan metode *k-fold cross validation*, kemudian diberi pembobotan dengan metode *TF-RF*, dan diklasifikasikan menggunakan metode *Naïve Bayes* karena mempunyai hasil akurasi paling tinggi dan hasil evaluasi yang baik. Kemudian dalam memprediksi topik, kamus dan korpus yang sebelumnya telah dibuat dijadikan dasar untuk menentukan topik apa saja yang terdapat dalam data ulasan. Visualisasi dibuat dengan menggunakan *PHP* dan *Python* yang nantinya akan menghasilkan aplikasi berbasis web. Alur dari *website* ini yaitu inputan dari *HTML* dikirimkan ke file *python* menggunakan *CMD*, setelah itu hasil inputan akan diolah dan nantinya akan menghasilkan penjumlahan nilai dari setiap kata. Hasil dari olahan tersebut kemudian akan diubah ke label positif dan negatif, serta termasuk ke dalam topik apa. Berikut ini adalah hasil visualisasi dari source code yang sudah dibuat.

Website visualisasi terdiri dari grafik, *worldcloud*, *perceptual map*, textbox untuk inputan, tombol *submit*, dan *alert box* untuk menampilkan hasil klasifikasi serta prediksi topik. Grafik batang dibuat untuk menunjukkan data awal yang awalnya tidak *balance*. Kemudian setelah melalui proses *oversampling*, data sudah menjadi *balance*. *Wordcloud* dibuat untuk mengetahui kata apa saja yang sering muncul dari sentimen positif dan sentimen negatif. Visualisasi *LDA* berupa *pyLDAvis* dibuat untuk menunjukkan seberapa berbeda setiap topik yang dihasilkan dan 30 istilah utama yang menonjol. Kemudian hasil visualisasi prediksi sentimen ulasan dan prediksi topik terdiri dari *textbox* untuk inputan, tombol *submit*, dan *alert box* untuk menampilkan hasil klasifikasi sentimen ulasan dan prediksi topik. Hasil klasifikasi tersebut berasal dari *source code python* yang berhasil mengolah kata inputan dan menampilkan hasil.

(a)

(b)

(c)



(d)

Gambar 5. Hasil Visualisasi

4.2 Pembahasan

Dalam klasifikasi sentimen, skenario ke-4 yaitu *Naïve Bayes* yang pembagian datanya menggunakan *k-fold cross validation* dan diberi pembobotan *TF-RF* memiliki akurasi paling tinggi yaitu 97,12% dengan waktu proses 0,014 detik. Model tersebut juga memiliki nilai *f1-score* yang tinggi dengan nilai positif 97% dan nilai negatif 97%. Sedangkan skenario ke-8 yaitu *PNN* yang pembagian datanya menggunakan *k-fold cross validation* dan diberi pembobotan *TF-RF* hanya memiliki akurasi sebesar 78,33% dengan waktu proses 33,97 detik. Model tersebut juga hanya memiliki nilai *f1-score* untuk positif sebesar 82% dan negatif sebesar 73%. Berdasarkan hasil tersebut, dapat diketahui bahwa model *Naive Bayes* lebih efektif jika dibandingkan PNN dalam mengklasifikasikan sentimen. Selain itu, hasil ini menunjukkan bahwa algoritma Naïve Bayes mampu mengklasifikasikan sentimen positif dan negatif dengan tingkat akurasi yang tinggi. Akurasi yang tinggi tersebut juga menunjukkan peningkatan jika dibandingkan dengan penelitian terdahulu yang menerapkan metode lain, seperti *K-Nearest* [18] *Neighbor dan C4.5* [19], yang memiliki hasil akurasi lebih rendah. Dengan demikian, penelitian ini memberikan kontribusi terhadap pemahaman penerapan *Naïve Bayes* dalam analisis sentimen dan mendukung pemanfaatannya dalam sistem evaluasi layanan aplikasi berbasis ulasan pengguna. Secara keseluruhan, penelitian ini menunjukkan bahwa algoritma

Naïve Bayes efektif dalam melakukan klasifikasi sentimen dan memperkuat bukti bahwa metode ini merupakan salah satu pilihan terbaik untuk analisis sentimen, khususnya pada aplikasi yang menuntut efisiensi dan kinerja tinggi, seperti klasifikasi ulasan pengguna. Selain itu, penelitian ini turut berkontribusi dengan menerapkan teknik *preprocessing* yang lebih lengkap, yang masih jarang dibahas dalam studi sebelumnya.

5. Simpulan

Untuk topik ulasan pengguna aplikasi IKD, penelitian ini berhasil melakukan prediksi sentimen menggunakan metode *Naïve Bayes* dan *PNN*, serta pemodelan topik menggunakan metode *LDA*. Berdasarkan pembahasan dari bagian sebelumnya, dapat ditarik kesimpulan bahwa dari hasil pemodelan sentimen yang telah dilakukan, didapatkan bahwa skenario yang memiliki akurasi terbaik adalah skenario keempat dengan data yang dibagi menggunakan metode *k-fold cross validation*, diberi pembobotan *TF-RF*, dan diklasifikasikan menggunakan *Naïve Bayes* yang memiliki akurasi 97,12% dengan hasil evaluasi *f1-score* kelas negatif 97% dan kelas positif 97%.

Kemudian dari hasil pemodelan topik yang telah dilakukan, didapatkan bahwa nilai koherensi terbaik untuk kelas positif diperoleh topik 2 dengan nilai 0,4076. Sedangkan nilai koherensi terbaik untuk kelas negatif diperoleh topik 10 dengan nilai 0,5564. Berdasarkan kata kunci, topik pada label positif dikategorikan menjadi verifikasi identitas, saran pengembangan aplikasi, kemudahan penggunaan, digitalisasi KTP, dan pengalaman penggunaan (UX). Sedangkan untuk label negatif dikategorikan menjadi kendala teknis aplikasi, akses & instalasi aplikasi, kendala jaringan, fitur dan keamanan data, serta pendaftaran aplikasi

Saran yang dapat diberikan untuk pengembangan penelitian ini lebih lanjut adalah menggunakan sumber data yang lebih seimbang antara sentimen positif dan negatifnya, kemudian bisa menggunakan metode pembagian data lain ataupun metode klasifikasi lain untuk mendapatkan hasil akurasi yang lebih baik, serta menambahkan fitur prediksi topik yang bisa memprediksi lebih dari satu topik.

Daftar Referensi

[1] R. Dimas Sulistiyo and M. R. Shihab, “Transformasi Digital dalam Pelayanan Surat Izin Mengemudi (SIM): Studi Kasus Korlantas Polri,” *Technomedia J.*, vol. 8, no. 2SP, pp. 189–204, 2023, doi: 10.33050/tmj.v8i2sp.2064.

[2] A. Kaur and N. Gumber, “Sentimental Analysis on Application Reviews on Educational Apps,” *Acad. edu, Accessed Jun*, no. 11, pp. 16–19, 2014, [Online]. Available: https://www.academia.edu/download/35954453/IJETR022706.pdf

[3] R. F. Triana, A. I. P. Sari, A. Bahtiar, and E. Wahyudin, “Implementasi Algoritma Naïve Bayes Untuk Klasifikasi Sentimen Ulasan Pengguna KAI Access,” *Jutisi J. Ilm. Tek. Inform. dan Sist. Inf.*, vol. 14, no. 1, pp. 12–21, 2025, doi: 10.35957/mdp-sc.v2i1.4326.

[4] M. Joshi and H. Vala, “Opinion Mining For Sentiment Data Classification,” *Int. J. Res. Inf. Technol.*, vol. 3, no. 1, pp. 8–14, 2014.

[5] I. Najiyah and I. Hariyanti, “Sentimen Analisis Covid-19 Dengan Metode Probabilistic Neural Network Dan Tf-Idf,” *J. Responsif Ris. Sains dan Inform.*, vol. 3, no. 1, pp. 100–111, 2021, doi: 10.51977/jti.v3i1.488.

[6] A. Alamsyah, W. Rizkika, D. D. A. Nugroho, F. Renaldi, and S. Saadah, “Dynamic large scale data on Twitter using sentiment analysis and topic modeling case study: Uber,” *2018 6th Int. Conf. Inf. Commun. Technol. ICoICT 2018*, vol. 0, no. c, pp. 254–258, 2018, doi: 10.1109/ICoICT.2018.8528776.

[7] A. Komarudin and A. M. Hilda, “Analisis Sentimen Ulasan Aplikasi Identitas Kependudukan Digital Pada Play Store Menggunakan Metode Naïve Bayes,” *Comput. Sci.*, vol. 4, no. 1, pp. 28–36, 2024, doi: 10.31294/coscience.v4i1.2955.

[8] N. L. P. M. Putu, Ahmad Zuli Amrullah, and Ismarmiaty, “Analisis Sentimen dan Pemodelan Topik Pariwisata Lombok Menggunakan Algoritma Naive Bayes dan Latent Dirichlet Allocation,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 1, pp. 123–131, 2021, doi: 10.29207/resti.v5i1.2587.

[9] F. F. Roji, N. G. Ginasta, Y. Cahyan, D. Rahayu, and D. Ramdani, “Review Analysis of SatuSehat Application Using Support Vector Machine and Latent Dirichlet Allocation Modeling,” *J. RISTEC Res. Inf. Syst. Technol.*, vol. 4, no. 1, pp. 76–88, 2023.

[10] R. Tuntun, K. Kusrini, and K. Kusnawi, “Analisis Perbandingan Kinerja Algoritma

Klasifikasi dengan Menggunakan Metode K-Fold Cross Validation," *J. Media Inform. Budidarma*, vol. 6, no. 4, p. 2111, 2022, doi: 10.30865/mib.v6i4.4681.

[11] R. Maulana, A. Voutama, and T. Ridwan, "Analisis Sentimen Ulasan Aplikasi MyPertamina pada Google Play Store menggunakan Algoritma NBC," *J. Teknol. Terpadu*, vol. 9, no. 1, pp. 42–48, 2023, doi: 10.54914/jtt.v9i1.609.

[12] A. Sasmoyo, R. Saptono, and Wiranto, "Penggunaan Jumlah Frekuensi Kata Terbanyak Sebagai Feature Set Pada Naive Bayes Classifier Untuk Mengklasifikasikan Dokumen Berbahasa Indonesia Dan Inggris," *Snik*, no. October, pp. 293–296, 2015.

[13] R. T. Wahyuni, D. Prastiyanto, and E. Supraptono, "Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen Skripsi," *J. Tek. Elektro Univ. Negeri Semarang*, vol. 9, no. 1, pp. 18–23, 2017, [Online]. Available: https://journal.unnes.ac.id/nju/index.php/jte/article/download/10955/6659

[14] M. Lan, "A New Term Weighting Method for Text Categorizationy," *PhD Theses, School of Computing, National University of Singapore,* 2007.

[15] Bustami, "Penerapan Algoritma Naive Bayes," *J. Inform.*, vol. 8, no. 1, pp. 884–898, 2014.

[16] S. Alam and N. Yao, "Probabilistic neural network and word embedding for sentiment analysis," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 7, pp. 48–53, 2018, doi: 10.14569/IJACSA.2018.090708.

[17] D. Normawati and S. A. Prayogi, "Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter," *J. Sains Komput. Inform. (J-SAKTI*, vol. 5, no. 2, pp. 697–711, 2021, [Online]. Available: http://ejurnal.tunasbangsa.ac.id/index.php/jsakti/article/view/369

[18] I.H. Kusuma, & N. Cahyono. "Analisis Sentimen Masyarakat Terhadap Penggunaan E-Commerce Menggunakan Algoritma K-Nearest Neighbor". *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 8, no. 3, pp. 302-307, 2023.

[19] Y. Akbar, & T. Sugiharto. "Analisis Sentimen Pengguna Twitter di Indonesia Terhadap ChatGPT Menggunakan Algoritma C4. 5 dan Naïve Bayes". *Jurnal Sains dan Teknologi*, vol. 5, no. 1, pp. 115-122, 2023.