

PENERAPAN ALGORITMA *K-NEAREST NEIGHBOR* PADA ANALISIS SENTIMEN TERHADAP ULASAN APLIKASI *MYBCA* DI *GOOGLE PLAY STORE*

Muhammad Rifka Noor Ikhsan¹⁾, Pitrasacha Adytia²⁾, Wahyuni³⁾

Teknik Informatika, STMIK Widya Cipta Dharma
Jl. M. Yamin No. 25, Samarinda, Kalimantan Timur, 75123
Email : rifkanoorikhsan@gmail.com¹⁾, pitra@wicida.ac.id²⁾, wahyuni@wicida.ac.id³⁾

ABSTRAK

Banyak dari pengguna *mobile banking* ini mengeluhkan tentang kinerja aplikasi terkait fitur teknis dan fungsional dari aplikasi *myBCA*, seperti gagal proses saat *top-up*, proses transaksi, login ke aplikasi *mobile banking* dan aplikasi berjalan lambat meskipun koneksi jaringan stabil. Untuk meningkatkan kualitas dan reputasi dari aplikasi *mobile banking myBCA*, serta tetap menjadi layanan *banking* yang dipercaya oleh nasabah atas keandalan, efektivitas, dan efisiensi pada aplikasi *mobile banking* nya, diperlukan suatu analisis sentimen sebagai bahan evaluasi pihak PT. *Bank Central Asia*, Tbk. terutama pada bagian layanan perbankan *mobile banking*.

Penelitian ini menggunakan metode dari Keputusan Menteri Ketenagakerjaan Republik Indonesia Nomor 299 Tahun 2020 tentang Penetapan Standar Kompetensi Kerja Nasional Indonesia (SKKNI). Pengambilan data (*scraping*) dilakukan menggunakan *library google-play-scraper*, sedangkan tahap *NLP* dan *preprocessing text* menggunakan *library NLP-ID*. Proses *labeling* dilakukan menggunakan kamus sentimen dari *InSet* dan *SentiWord*. Teknik pembobotan yang digunakan adalah *TF-IDF*, dengan data yang sudah dibobot dilakukan *balancing* menggunakan *SMOTE* untuk meningkatkan performa model *K-Nearest Neighbors(KNN)*. Pengklasifikasian data ulasan dilakukan menggunakan model *K-Nearest Neighbors (KNN)*, dengan penentuan jumlah tetangga optimal melalui *cross-validation*, mencari 10 skor terbaik, dan *metode elbow* untuk menentukan rentang terbaik dari 1 hingga 11. Model *K-Nearest Neighbors(KNN)* yang dihasilkan memiliki akurasi 81%, dan dievaluasi menggunakan *Confusion Matrix*.

Hasil dari analisis dan pembangunan model *K-Nearest Neighbors* diimplementasi dalam bentuk *website* menggunakan *web service* dari *Streamlit* yang dipadukan dengan *web hosting* dari *GitHub Codespaces*, untuk menganalisis sentimen terhadap Layanan yang ada pada Aplikasi *myBCA* di *Google Play Store* berdasarkan ulasan-ulasan dari pengguna aplikasi, serta diuji dengan menggunakan *White Box Testing* dan *Black Box Testing*.

Kata Kunci : *K-Nearest Neighbor, Analisis Sentimen, SKKNI, Layanan*

1. PENDAHULUAN

PT. *Bank Central Asia*, Tbk. (*BCA*) adalah bank swasta terbesar di Indonesia yang melayani berbagai segmen nasabah, baik individu maupun bisnis, melalui jaringan perbankan elektronik dan kantor cabang yang tersebar di hampir seluruh kota besar di Indonesia. Salah satu layanan unggulan *BCA* adalah *mobile banking*, dengan aplikasi andalannya, *myBCA*. Aplikasi ini memungkinkan nasabah untuk melakukan transaksi perbankan dan memperoleh informasi produk serta layanan *BCA*, termasuk dari mitra kerja *BCA*. *MyBCA* dapat diakses melalui *browser* atau diunduh dari *platform* resmi seperti *Google Play Store* [1].

Di *Google Play Store*, pengguna dapat memberikan rating berupa jumlah bintang dan ulasan untuk aplikasi, termasuk *myBCA*. Selain rating bintang 5 dan 4, ada juga rating bintang 3, 2, dan 1. Penting untuk memperhatikan ulasan pengguna, mengingat aplikasi *mobile banking myBCA* ini dirilis sejak Mei 2021 dan terakhir diperbarui pada Juni 2024, serta telah digunakan oleh banyak nasabah. Banyak pengguna mengeluhkan kinerja aplikasi terkait fitur teknis dan fungsional, seperti kegagalan proses saat *top-up*, transaksi, login, dan lambatnya aplikasi meskipun koneksi jaringan stabil. Oleh karena itu, penting bagi pengembang *myBCA* untuk mengetahui respon nasabah guna memperbaiki dan meningkatkan kinerja serta fitur aplikasi,

sehingga pengguna merasa nyaman. Untuk meningkatkan kualitas dan reputasi aplikasi, serta tetap menjadi layanan perbankan yang dipercaya oleh nasabah atas keandalan, efektivitas, dan efisiensi, diperlukan analisis sentimen sebagai bahan evaluasi bagi PT *Bank Central Asia*, terutama pada layanan *mobile banking* [2]. Proses klasifikasi ini menggunakan data latih sebagai referensi untuk mengenali pola dan menghasilkan prediksi yang akurat [3].

Dari latar belakang yang sudah dibahas, maka Peneliti akan melakukan penelitian mengenai bagaimana menerapkan pemodelan *Machine Learning* dengan Algoritma *K-Nearest Neighbor* Pada Analisis Sentimen Terhadap Aplikasi *myBCA* di *Google Play Store*.

2. BATASAN MASALAH

Adapun batasan-batasan masalah dalam penelitian ini adalah :

1. Analisis Sentimen yang dilakukan didasarkan dari permasalahan yang ada di layanan perbankan aplikasi *myBCA* pada *Bank BCA*.
2. Data pada penelitian ini diambil dari ulasan nasabah terkait aplikasi *myBCA* di *Google Play Store* *Bank BCA*.
3. *Tools* atau *library* yang digunakan pada penelitian ini dalam mengambil data atau proses *scraping* adalah

google-play-scrapper untuk mengambil data dari aplikasi *Google Play Store*.

4. *Tools* atau *library* yang digunakan pada penelitian ini untuk proses visualisasi data dan algoritma *K-Nearest Neighbor* adalah *matplotlib* dari *python*.
5. *Tools* atau *library* yang digunakan pada penelitian ini untuk proses data *frame* atau menyimpan dan mengorganisasi data teks yang akan dianalisis adalah *Pandas*.
6. *Tools* atau *library* yang digunakan pada penelitian ini untuk awal pemrosesan sebelum dilakukan analisis data adalah *NLP-ID*.
7. Kamus sentimen yang digunakan dalam penelitian ini adalah Indonesia *Sentiment Lexicon (InSet)* dan *SentiWord*.
8. *Tools* atau *library* yang digunakan pada penelitian ini untuk model *machine learning*, algoritma *K-Nearest Neighbor* dan proses pembobotan kata pada Teknik TF-IDF adalah *Scikit-Learn*.
9. Metode Penelitian yang digunakan adalah Metode SKKNI bidang keahlian *Artificial Intelligence* Subbidang *Data Science* yang berdasarkan Keputusan Menteri Ketenagakerjaan RI No. 299 Tahun 2020.
10. Pemodelan yang digunakan pada penelitian ini adalah *Machine Learning* dengan Algoritma *K-Nearest Neighbor*.
11. Evaluasi terhadap model dan algoritma yang digunakan adalah dengan menggunakan metode *Confussion Matrix*.
12. Implementasi atau *deployment* hasil analisis sentimen berupa *website* menggunakan *web service* dari *python* yaitu *Streamlit* dan *codespaces* dari *GitHub*.
13. Bahasa pemrograman yang digunakan pada penelitian ini adalah *Python*.
14. Evaluasi terhadap *website* yang digunakan untuk menampilkan hasil analisis sentimen menggunakan metode *Black Box Testing & White Box Testing*.

3. DASAR TEORI

Berikut beberapa teori yang digunakan untuk membahas suatu permasalahan yang dihadapi dalam penelitian ini.

3.1 Algoritma K-Nearest Neighbor (KNN)

Metode sederhana yang mudah diimplementasikan yang menggunakan data berlabel untuk memudahkan pengelompokkan ke kelas yang sesuai dan unggul dalam mengklasifikasikan data dengan adanya data latih dan data uji [4], salah satu tugas utama dalam algoritma *KNN* adalah mengukur kesamaan atau jarak [5]. Algoritma *k-nearest neighbor* adalah metode *supervised learning* yang mengklasifikasikan objek berdasarkan data latih terdekat dengan data uji. *KNN* menggunakan tetangga terdekat untuk memprediksi nilai sampel uji baru. Keunggulannya adalah menghasilkan data akurat dan efektif jika data latihnya cukup besar. Langkah-langkah klasifikasinya adalah:

1. Tentukan nilai *K*, contohnya $k=23$ yang mempunyai arti bahwa 23 dokumen yang memiliki jarak terdekat dengan dokumen training yang akan diambil.
2. Melakukan perhitungan jarak pada data terbaru di masing-masing label data dengan jarak seluruh model data *training*. Kemudian untuk melakukan perhitungan tingkatan kesamaan pada dokumen tersebut menggunakan *euclidean distance* seperti pada rumus :

$$D(X, Y) = \sqrt{\sum_{k=1}^n (X_k - Y_k)^2} \quad (1)$$

Keterangan :

D = Jarak antara dua titik *x* dan *y*

X = Data *Training*

Y = Data *Testing*

n = Dimensi Data

k = Variabel Data

3. Selanjutnya urutkan hasil euclidean distance sesuai jarak *K* yang telah ditentukan. nilai jarak yang dipakai adalah nilai jarak yang paling terkecil.
4. Kemudian gunakan model mayoritas atribut kelas terhadap tetangga terdekat yang telah dipilih untuk melakukan penentuan prediksi kelas terhadap data terbaru [2].

3.2 Web Service : Streamlit

Streamlit adalah *framework opensource* dari *Python* yang memungkinkan untuk membuat aplikasi *web* menggunakan bahasa *Python* dalam mengaplikasi model dari *machine learning* atau *data science* [6] dengan cepat dan minimalis tetapi juga memiliki tampilan yang cukup baik serta ramah pengguna. Setiap kali pengguna berinteraksi dengan aplikasi *Streamlit*, sintak *Python* dijalankan kembali dari atas ke bawah.

3.3 Standar Kompetensi Kerja Nasional Indonesia

Standar Kompetensi Kerja Nasional Indonesia (SKKNI) adalah rumusan kemampuan kerja yang mencakup pengetahuan, keterampilan, dan sikap kerja relevan dengan tugas dan syarat jabatan. SKKNI menetapkan unit kompetensi sebagai alat ukur kemampuan kerja sesuai kebutuhan industri [7].

Peraturan Menteri tenaga Kerja Republik Indonesia nomor 05 tahun 2012 menjelaskan bahwa SKKNI adalah rumusan kemampuan kerja yang mencakup aspek pengetahuan, keterampilan dan/atau keahlian, serta sikap kerja relevan dengan pelaksanaan tugas dan syarat jabatan yang ditetapkan sesuai dengan ketentuan peraturan perundang-undangan. SKKNI disusun untuk menetapkan unit-unit kompetensi yang akan dijadikan sebagai alat ukur kemampuan kerja seseorang yang meliputi aspek pengetahuan, keterampilan dan sikap kerja sebagaimana yang disyaratkan oleh pengguna jasa sumber daya manusia, dalam hal ini industri.

Adapun proses tahapan pengembangan dari metodologi berdasarkan keputusan menteri ketenagakerjaan RI nomor 299 tahun 2020 tentang penetapan SKKNI kategori informasi dan komunikasi golongan pokok aktivitas pemrograman, konsultasi computer dan kegiatan YBDI bidang keahlian artificial intelligence subbidang *data science*.

Tabel 1. Pemetaan SKKNI Nomor 299 Tahun 2020

Tujuan Utama	Fungsi Kunci	Fungsi Utama	Fungsi Dasar
Menemukan pengetahuan insight atau pola yang bermanfaat dari data untuk	Meng-analisis kebutuhan organisasi	<i>Business understanding</i>	Menentukan objektif bisnis
			Menentukan tujuan teknis <i>data science</i>
			Membuat rencana proyek <i>data science</i>
			Mengumpulkan data

Tujuan Utama	Fungsi Kunci	Fungsi Utama	Fungsi Dasar
berbagai keperluan		<i>Data understanding</i>	Menelaah data
			Memvalidasi data

Tabel 1. Pemetaan SKKNI Nomor 299 Tahun 2020 (Lanjutan)

Tujuan Utama	Fungsi Kunci	Fungsi Utama	Fungsi Dasar
	Mengembangkan model	<i>Data preparation</i>	Memilah data
			Membersihkan data
			Mengkonstruksi data
			Menentukan label
			Mengintegrasikan data
		<i>Modeling</i>	Membangun skenario pengujian
			Membangun model
		<i>Model evaluation</i>	Mengevaluasi hasil pemodelan
			Melakukan review proses pemodelan
		Menggunakan model yang dihasilkan	<i>Deployment</i>
Melakukan deployment model			
Membuat rencana pemeliharaan			
Melakukan pemeliharaan model			
<i>Evaluation</i>	Melakukan review proyek data science		
	Membuat laporan akhir proyek data science		

4. METODOLOGI PENELITIAN

4.1 Waktu dan Tempat Penelitian

Penelitian ini dilaksanakan sejak tanggal 4 Maret 2024 sampai dengan 7 Agustus 2024 dan dilaksanakan PT. Bank Central Asia yang beralamat di Jl. Wahid Hasyim I, Sempaja Sel., Kec. Samarinda Utara, Kota Samarinda, Kalimantan Timur 75119.

4.2 Teknik Pengumpulan Data

Teknik pengumpulan data yang digunakan dalam penelitian :

a. Studi Pustaka

Teknik studi pustaka merupakan kegiatan yang dilakukan oleh peneliti untuk menghimpun informasi yang relevan dengan topik atau masalah yang akan atau sedang diteliti.

b. Observasi

Pada teknik observasi ini peneliti melakukan pengamatan secara daring atau online sekaligus mengambil data untuk diteliti pada ulasan dari pengguna aplikasi myBCA di Google Play Store.

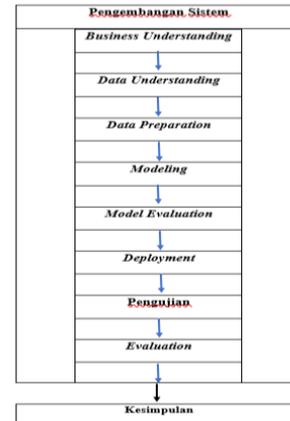
c. Wawancara

Dalam teknik wawancara ini, peneliti melakukan tanya jawab dengan pihak cabang Bank BCA KCP Wahid Hasyim di Kota Samarinda mengenai pendapat nasabah terkait peng-

gunaan aplikasi mobile banking, khususnya myBCA, dan perlunya analisis sentimen pengguna aplikasi tersebut.

4.3 Metode Pengembangan Sistem

Pada penelitian ini menggunakan metode dari Keputusan Menteri Ketenagakerjaan Republik Indonesia Nomor 299 Tahun 2020 Tentang Penetapan Standar Kompetensi Kerja Nasional Indonesia (SKKNI) dan Penambahan tahap pengujian sebelum evaluasi untuk memastikan bahwa analisis sentimen yang dilakukan berfungsi dengan benar. Berikut kerangka kerja pada penelitian ini :



Gambar 1. Kerangka kerja pengembangan sistem

a. Business Understanding

Pada tahap ini, peneliti mengikuti ketentuan yang ada, yaitu menentukan objektif bisnis, yang dalam penelitian ini adalah aplikasi mobile banking myBCA. Selanjutnya, tujuan teknis penelitian ini adalah mengetahui apakah ulasan pengguna aplikasi myBCA cenderung positif atau negatif, memahami alasan di balik rating bintang 1, 2, dan 3, serta mengidentifikasi ulasan lain terkait aplikasi tersebut. Hasil penelitian ini akan digunakan untuk merencanakan proyek ke depan.

b. Data Understanding

Tahap Data Understanding meliputi pengumpulan data menggunakan teknik crawling di Google Play Store dengan library google-play-scrapper. Selanjutnya, data yang dikumpulkan dieksplorasi dan dianalisis melalui deskripsi statistik, analisis univariat dan multivariat untuk mendukung analisis sentimen, serta memvalidasi data tersebut.

c. Data Preparation

Pada tahap Data Preparation, peneliti menggunakan library NLP-ID untuk beberapa langkah. Pertama, pengumpulan data dengan memfilter data yang akan dianalisis. Selanjutnya, pembersihan data dengan mengubah seluruh kalimat menjadi huruf kecil, mengganti emoji dengan teks, menghapus URL dan simbol, serta mengubah angka menjadi huruf untuk memudahkan analisis. Terakhir, preprocessing menggunakan metode Natural Language Processing dengan tools atau library dari NLP-ID untuk tugas seperti:

1. *Normalization*, Membuat kamus berdasarkan ulasan yang diambil dan memperbaiki kata yang kurang benar secara manual jika diperlukan. Data diimpor menjadi file .csv, lalu dilakukan perbaikan manual untuk kata atau kalimat yang salah, seperti slang, huruf berulang, atau

hasil normalisasi yang kurang maksimal. Selanjutnya, memperpanjang kata yang disingkat dan menghapus kata henti (*Stopword Removal*).

2. *Tokenizing*, mengubah suatu kalimat menjadi kata-kata yang terpisah agar pada data tersebut setiap kata nya memiliki nilai.
3. *Part of Speech*, Menandai setiap kata dalam teks dengan label kelas kata seperti *noun*, *verb*, *adjective*, dll.
4. *Stemming*, Menghilangkan imbuhan di awal dan akhir kata dalam data

Setelah *preprocessing*, tahap selanjutnya dalam *Data Preparation* adalah *Labeling* menggunakan kamus sentimen *InSet*, *SentiWord*, dan kamus sentimen yang dibuat berdasarkan ulasan. Ulasan tersebut diubah menjadi bahasa baku dan diberikan skor sentimennya. Selanjutnya, dilakukan pembobotan kata menggunakan metode *Term Frequency-Inverse Document Frequency (TF-IDF)* dengan *library* dari *Scikit-Learn*. Data kemudian dibagi menjadi data latih dan uji. Setelah pembobotan kata, tentunya data tersebut tidakimbang maka dari itu diseimbangkan dengan metode *SMOTE* untuk meningkatkan performa model.

d. *Modelling*

Pada tahap modeling, perhitungan dilakukan dengan algoritma *K-Nearest Neighbor*. Pertama, nilai K atau N ditentukan menggunakan Metode *Cross Validation*, yang kemudian dioptimalkan dengan *Elbow Method*. Data yang telah dilabeli dilatih menggunakan model *Machine Learning* dengan algoritma *K-Nearest Neighbor*. Selanjutnya, data uji diuji menggunakan model dan algoritma yang sama.

e. *Model Evaluation*

Pada tahap ini peneliti melakukan evaluasi dan mengukur kinerja dari model yang sudah diuji di tahap *modeling* menggunakan metode *confusion matrix*.

f. *Deployment*

Pada tahap *deployment*, peneliti mengimplementasikan model *machine learning* dengan algoritma *K-nearest neighbor* dan hasil analisis sentimen menggunakan *web service Streamlit* dari *Python*. Skrip program yang dikerjakan di *Visual Studio Code* disalin ke *Codespaces* di *GitHub* untuk terhubung dengan *Streamlit*. Hal ini memungkinkan analisis sentimen ditampilkan secara visual di *website* yang *user-friendly*, sehingga mudah dipahami dan digunakan oleh pengguna.

g. *Pengujian*

Pada tahap pengujian, peneliti memeriksa kode program analisis sentimen menggunakan *white box testing*. Selain itu, karena penelitian ini sampai tahap *deployment*, peneliti juga menguji hasil *web service* yang diimplementasikan menggunakan *black box testing*.

h. *Evaluation*

Setelah dilakukan *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Model Evaluation*, *Deployment*, dan Pengujian. Maka pada tahap ini peneliti membuat laporan akhir guna mencatat mengenai analisis sentimen yang sudah dibangun, dalam kasus ini membuat laporan akhir dalam bentuk Skripsi.

5. IMPLEMENTASI

Berikut adalah penjabaran mengenai tahapan – tahapan dalam membangun sistem analisis sentimen menggunakan Metode SKKNI dan algoritma *K-nearest neighbor* berbasis *website*.

a. *Business Understanding*

Pada tahap ini peneliti melakukan pemahaman bisnis guna memahami kebutuhan dan tujuan dari penelitian yang diteliti oleh peneliti yaitu analisis sentimen terhadap ulasan aplikasi *mobile banking myBCA*. Aplikasi *myBCA* sendiri merupakan aplikasi *mobile banking* yang dimiliki oleh PT *Bank Central Asia Tbk. (BCA)*, sebagai sarana penting bagi nasabah untuk melakukan berbagai transaksi perbankan.

1. *Objektif Bisnis*

Objektif bisnis dari penelitian ini adalah untuk membantu *BCA* dalam meningkatkan kepuasan dan loyalitas nasabah dengan cara memahami dan memberikan hasil analisis sentimen pengguna terhadap aplikasi *myBCA* berdasarkan ulasan di *Google Play Store*.

2. *Tujuan Teknis*

Tujuan teknis dari penelitian ini seperti Mengumpulkan data ulasan dari *Google Play Store* menggunakan teknik *web scraping*, Mengkategorikan ulasan dari pengguna menjadi sentiment (positif, negatif, atau netral) menggunakan *machine learning* dengan algoritma *K-Nearest Neighbor (KNN)*, dan Menyusun hasil analisis terhadap ulasan atau komentar dari pengguna aplikasi *myBCA* dalam format yang jelas serta informatif agar dapat digunakan oleh manajemen *BCA* untuk pengambilan keputusan.

b. *Data Understanding*

1. *Pengambilan Data (Scraping)*

Pada tahap ini peneliti melakukan pengambilan data komentar pengguna aplikasi *myBCA* di *Google Play Store* dengan menggunakan *library google-play-scraper*. Data diambil dengan App ID "com.bca.mybca.omni.android", dengan pengaturan bahasa Indonesia dan negara id, serta diurutkan berdasarkan ulasan terbaru. Semua rating dari bintang 1 hingga 5 ditampilkan tanpa filter.

2. *Eksplorasi Data (EDA)*

Dalam eksplorasi data penelitian ini, peneliti menganalisis data yang telah diambil dari aplikasi *myBCA* di *Google Play Store*, meliputi:

- *Analisis Univariat*

Analisis statistik yang hanya melibatkan satu variabel. Tujuan dari analisis univariat ini adalah untuk memahami dan mendeskripsikan karakteristik dasar dari variabel tersebut. Diantaranya Total rating yang diberikan pada *myBCA* di *Google Play Store* dan Total ulasan yang diberikan para Pengguna dari tahun ke tahun.

- *Analisis Multivariat*

Analisis statistik yang melibatkan dua atau lebih variabel. Berguna untuk memahami hubungan dan interaksi antara satu variabel dengan variabel yang lain. Diantaranya banyaknya rating yang diberikan berdasarkan update versi aplikasi dan trend ulasan pengguna dari waktu ke waktu di *google play store*.

c. *Data Preparation*

1. *Pengumpulan Data (Data Collection)*

Menampilkan keseluruhan dari fitur atau kolom dari data yang sudah diambil menggunakan teknik crawling di *Google*

Play Store. Lalu menghapus kolom yang tidak relevan, untuk memfokuskan pada kolom yang penting seperti `userName`, `score`, dan `content`,

2. Pembersihan Data (*Data Cleaning*)

Pada tahap ini peneliti melakukan perubahan kalimat menjadi huruf kecil dan emoji menjadi teks. Menghapus kode URL, simbol, dan sejenisnya. Dan mengubah angka menjadi teks.

8	Pengguna Google	1	SUSAH LOG IN JELEK KKKKK	susah log in jelekkkkk
9	Pengguna Google	1	Apasi jelek banget otp gak masuk? malah disuru...	apasi jelek banget otp gak masuk malah disuru...
10	Pengguna Google	3	Saat biometrik login digunakan, jika jari dita...	saat biometrik login digunakan jika jari dita...
11	Pengguna Google	1	Perbaiki lagi lah masa saya login ulang dari k...	perbaiki lagi lah masa saya login ulang dari k...
12	Pengguna Google	5	Mantap	mantap setuju
13	Pengguna Google	1	aku kan lupa email yang aku gunakan itu apa. t...	aku kan lupa email yang aku gunakan itu apa te...
14	Pengguna Google	5	Oke	oke

Gambar 2. Hasil data cleaning

3. *Preprocessing Text*

Pada tahap ini peneliti melakukan *Normalization* dengan menggantikan kata-kata tidak baku, *slangword*, Bahasa Daerah, Bahasa Inggris, angka, dan lainnya dalam teks ulasan dengan kata-kata baku berdasarkan kamus yang dibuat secara manual dengan fungsi `replace_taboo_words` untuk menggantikan kata-kata tidak baku dengan kata-kata baku yang sesuai.

Selanjutnya melakukan *Tokenizing* untuk memecah kalimat menjadi per-kata dan memberikan token pada setiap kata dalam teks ulasan yang telah dinormalisasi sebelumnya untuk dapat dilakukan analisis lebih lanjut dalam tahap *Preprocessing Text*.

Kemudian, pada tahap *part-of-speech* yaitu memproses pemberian label kata-kata dalam teks dengan bagian-bagian kalimat seperti kata benda, kata kerja, dan kata sifat.

Selanjutnya melakukan *stemming* dan *lemmatization* pada teks ulasan serta menghitung jumlah kata yang diubah.

term	count	rank
bagus	374	191124
aplikasi	373	322993
oke	367	252123
pulsa	276	588915
bantu	246	975664
mantap	229	873449
verifikasi	215	294154
sms	213	493792
mudah	211	248143
masuk	204	6482143

Gambar 3. Hasil stemming

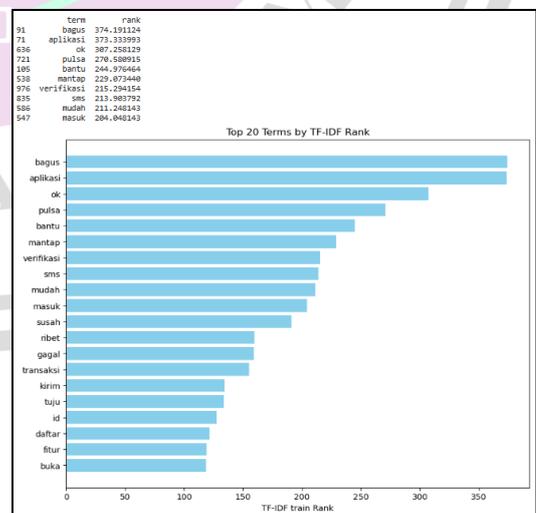
Setelah melakukan *Stemming* yaitu membuat *WordCloud* menggunakan `NLP_id` untuk memvisualisasikan frekuensi kemunculan kata. Ini memudahkan untuk melihat kata-kata yang paling sering muncul dan memberikan wawasan cepat tentang data teks.



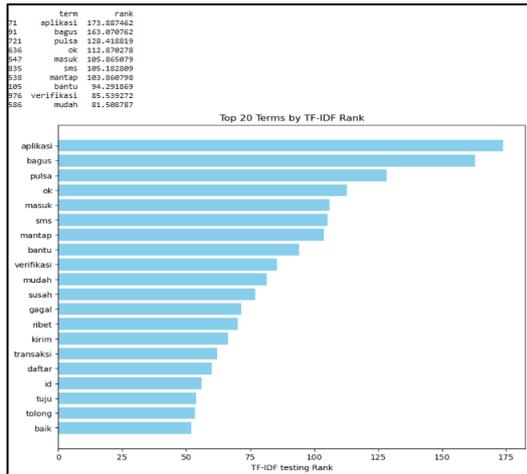
Gambar 4. Wordcloud hasil stemming

Selanjutnya peneliti mengunduh lexicon sentimen Bahasa Indonesia dari URL eksternal dan membersihkan lexicon default VADER, yang awalnya berbahasa Inggris. Peneliti menghapus isi ketiga kamus VADER dan mem-perbaruinya dengan isi dari kamus `InSet` yang diunduh dari GitHub. Ini dilakukan dengan membuat tiga instance `SentimentIntensityAnalyzer` dan mengunduh data lexicon negatif dan positif dari `InSet` serta lexicon `SentiWords`. Lexicon ini diubah menjadi dictionary dan digunakan untuk memperbarui lexicon VADER agar sesuai dengan Bahasa Indonesia. Sebagian dari lexicon yang diperbarui ditampilkan untuk verifikasi, memungkinkan analisis sentimen yang lebih akurat untuk teks dalam Bahasa Indonesia.

Pada tahap pembobotan kata dengan `TF-IDF`, peneliti membagi data ulasan yang telah diproses dan label sentimen numerik atau label menjadi data latih dan data uji dengan perbandingan 70% untuk data latih dan 30% untuk data uji. Pembobotan data ulasan menggunakan `TfidfVectorizer` untuk mengubah teks ulasan yang telah diproses menjadi *matriks TF-IDF* dengan maksimum 1000 fitur. Matriks `TF-IDF` memberikan informasi tentang seberapa penting suatu kata dalam dokumen relatif terhadap seluruh corpus, membantu model untuk memahami dan memprediksi sentimen dengan lebih akurat. Selanjutnya peneliti memvisualisasikan 20 istilah teratas dalam data latih berdasarkan peringkat `TF-IDF` yang dapat memberikan wawasan tentang kata-kata yang memiliki pengaruh terbesar dalam kumpulan data ulasan yang telah diproses. Peringkat `TF-IDF` dihitung dengan menjumlahkan frekuensi `TF-IDF` setiap istilah di seluruh dokumen.

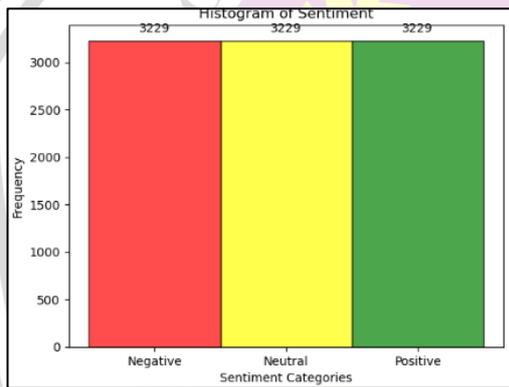


Gambar 5. Hasil ranking kata dari data latih



Gambar 6. Tampilan ranking kata dari data uji

Selanjutnya, Peneliti mengimbangkan Sentimen menggunakan teknik SMOTE untuk menangani masalah ketidakseimbangan data dalam data latih yang telah diubah menjadi matriks TF-IDF. Teknik menggunakan SMOTE ini cukup penting dikarenakan cukup berpengaruh terhadap keakuratan model yang dibangun.

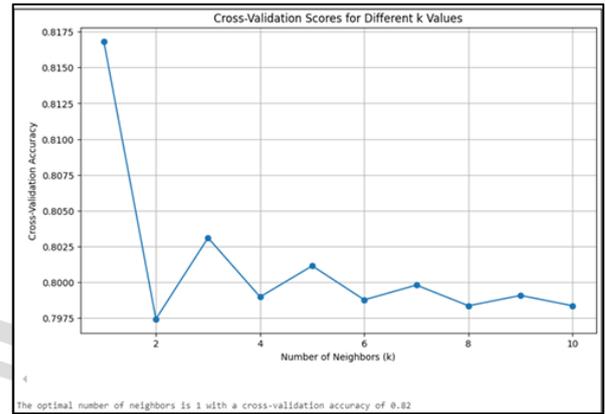


Gambar 7. Hasil resempling menggunakan SMOTE

d. Modelling

Pada tahap Modelling ini, sebelum peneliti melakukan penerapan algoritma dari K-Nearest Neighbor, yaitu mencari nilai K optimal yang akan diterapkan ke dalam model K-Nearest Neighbor nantinya. Untuk menentukan atau mencari nilai K dalam penelitian ini, peneliti menggunakan 2 tahap, yaitu Cross validation dan Elbow Method. Dalam penelitian ini, Cross Validation di masukan untuk pencarian nilai terbaik sebanyak 10, yang kemudian dari 10 nilai tersebut akan diseleksi kembali menggunakan Elbow Method untuk mencari yang terbaik diantara 10 besar dari Cross Validation untuk nilai K yang lebih optimalisasi.

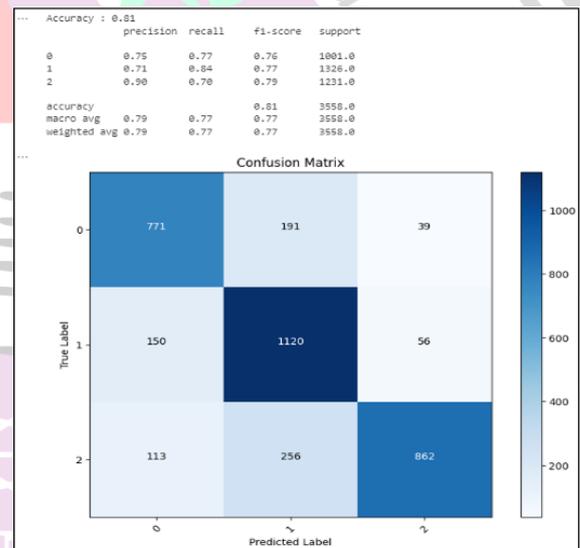
Pada penelitian ini nilai K dimasukan secara manual setelah mendapatkan hasil dari kedua metode tersebut. Ada pilihan lainnya yaitu nilai K dapat juga dimasukan secara otomatis atau mengambil secara otomatis hasil dari kedua metode tersebut setelah dijalankan, hanya saja terkadang cara tersebut masih kurang efektif dalam menentukan atau mencari nilai K yang optimal. Dalam Penelitian ini mendapatkan nilai K = 5, yang sudah dicoba pada model dengan menghasilkan akurasi senilai 81%.



Gambar 8. Hasil pencarian nilai K Optimal dengan Cross Validation dan Elbow Method

e. Model Evaluation

Model Evaluasi yang digunakan dalam penelitian ini adalah Confusion Matrix. Pada Confusion Matrix menyediakan raport klasifikasi seperti precision, recall, dan F1-Score yang berguna untuk melihat, memahami kekuatan serta kelemahan dari model yang sudah kita terapkan sehingga membantu dalam analisa yang kita lakukan lebih lanjut.



Gambar 8. Hasil Evaluasi dari Model K-Nearest Neighbor

Dari hasil confusion matrix yang ditampilkan dalam grafik heatmap, semakin gelap warnanya, semakin banyak data dalam kategori tersebut. Sebaliknya, semakin muda atau terang warnanya, berarti lebih sedikit data dalam kategori tersebut. Dari gambar 8, dapat peneliti simpulkan bahwa dengan hasil akurasi 81%, menunjukkan bahwa model cukup baik dalam mengklasifikasikan data meskipun tetapi masih ada ruang untuk perbaikan.

Dari confusion matrix yang sudah dibangun, mengungkapkan bahwa ada beberapa kesalahan klasifikasi antara kelas-kelas, terutama untuk Negatif memiliki Precision 0.75, Recall 0.76, dan F1-Score 0.76. Netral dengan Precision 0.71, Recall 0.84, dan F1 Score 0.77. Dan, (Positif) memiliki Precision 0.90, Recall 0.70, dan F1-Score 0.79 yang menghasilkan Akurasi senilai 81%. Hal ini merupakan hasil dari melakukan percobaan beberapa kali di tahap pencarian Nilai K di Cross Validation dan Elbow Method.

f. Deployment

Setelah model dibangun, tahap yang dilakukan selanjutnya adalah, mengimplementasikan ke dalam bentuk website agar pengguna dapat melihat hasil dari analisis sentimen yang sudah dibangun dengan tampilan yang mudah dipahami. Tahap deployment pada penelitian ini menggunakan web services streamlit dipadukan dengan github codespaces sebagai web hostingnya. Berikut adalah tahapan implementasi machine learning dan model yang telah diterapkan:

1. Mendaftar akun GitHub.
2. Membuat repository di GitHub untuk server website Streamlit.
3. Membuat file utama website (app.py), fungsi analisis sentimen dengan K-Nearest Neighbor (function.py), dan data kata slang serta singkatan (global_data).
4. Menyalin sintaks dari Visual Studio Code ke file Python di GitHub Codespaces.
5. Menambahkan library Streamlit pada sintaks.
6. Menguji secara lokal dengan menjalankan streamlit run app.py di terminal. Membuka URL yang muncul di browser dan membuat website publik dengan opsi "Make Public" atau melalui tab "Ports".

Berikut hasil atau tampilan dari website yang sedang dibangun dengan streamlit.

1. Halaman Dashboard
 Pada halaman ini berisikan hasil dari eksplorasi data seperti, menampilkan total rating yang diberikan dari pengguna, banyaknya ulasan tiap tahunnya, banyaknya ulasan berdasarkan versi update aplikasi, dan trend ulasan dari waktu ke waktu



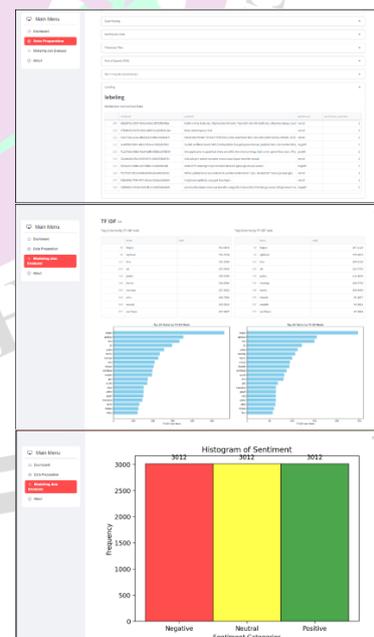
Gambar 9. Halaman Dashboard

2. Halaman Data Preparation

Pada halaman ini, menampilkan seperti hasil dari pengambilan data(drop data), hasil proses case folding, hasil dari normalisasi kata dengan NLP-ID dan dikombinasi dengan tambahan kamus slang word manual dari file.csv, hasil dari proses tokenizing dengan NLP-ID, part-of-speech dengan NLP-ID, stemming dengan NLP-ID, labeling dengan menggunakan kamus InSet dan SentiWord, imbalance-balance data dengan SMOTE, dan pembobotan dengan TF-IDF



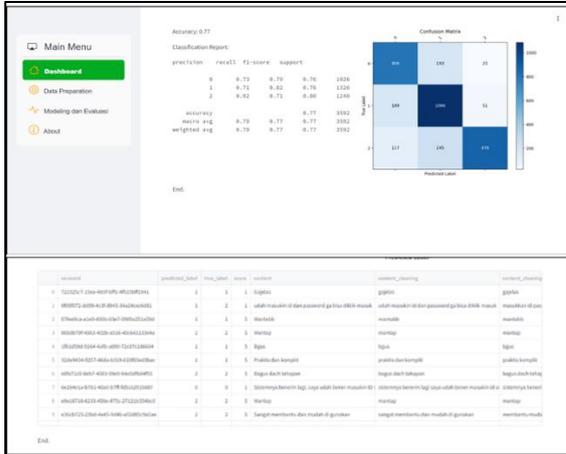
Gambar 10. Halaman Data Preparation(1)



Gambar 11. Halaman Data Preparation(2)

3. Halaman Modeling dan Evaluasi
 Pada halaman ini menampilkan hasil dari model yang sudah dibangun, dalam penelitian ini menggunakan model dari machine learning yaitu

algoritma K-Nearest Neighbor, disertai dengan hasil evaluasi dari model K-Nearest Neighbor yang sudah dibangun dengan confusion matrix. Dan, menampilkan proses keseluruhan dari proses dan hasil dari proses di tahap Data Preparation.



Gambar 11. Hasil Proses Modeling dan Evaluasi

- Halaman About Untuk halaman ini hanya berisikan informasi terkait aplikasi mobile banking myBCA dari Bank BCA.

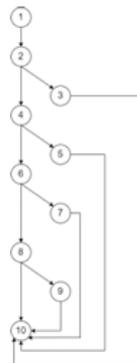


Gambar 12. Halaman About

g. Pengujian

Pada penelitian ini, pengujian menggunakan 2 metode, yaitu White box Testing dan Black box Testing. Untuk White box testing, peneliti menggunakan metode basis path yang terdiri dari :

- Testing dengan Skrip Program pada proses deployment
- Kemudian, testing dengan menggunakan flowgraph



Gambar 13. Flowgraph white box testing

- Menggunakan cyclomatix complexity untuk melihat kompleksitas yang dimiliki di dalam skrip program selama proses deployment pada web service streamlit

$$V(G) = E - N + 2$$

$$V(G) = 13 - 10 + 2$$

$$V(G) = 5$$

Yang dimana berarti terdapat 5 edges yang dimiliki untuk sampai ke nodes terakhir(10).

- Dan, menggunakan test case, yang dimana urutannya di peroleh dari tahap dlowgraph yang sudah dihitung kompleksitas nya dengan cyclomatix complexity yang dimana tahap nya diurutkan dari nodes 1 hingga nodes akhir, apakah memiliki output atau hasil yang berhasil atau tidak.

Kemudian, melakukan tahap black box testing untuk menguji tiap tombol atau fitur yang mendukung aplikasi yang sudah dibangun.

Tabel 1. Hasil Black box Testing

No	Kegiatan Pengujian	Hasil yang Diharapkan	Hasil Pengujian
1	Tombol <i>Dashboard</i>	Klik untuk berpindah ke Halaman Dashboard dan menampilkan semua isinya	Berhasil
2	Tombol <i>Data Preparation</i>	Klik untuk berpindah ke Halaman Data Preparation dan menampilkan semua isinya	Berhasil
3	Tombol <i>Modeling dan Evaluasi</i>	Klik untuk berpindah ke Halaman Modeling dan Evaluasi serta menampilkan semua isinya	Berhasil
4	Tombol <i>About</i>	Klik untuk berpindah ke Halaman About dan menampilkan semua isinya	Berhasil
5	Tombol <i>Sliding</i> pada Halaman <i>Data Preparation</i>	Klik untuk mengeksekusi dan menampilkan hasil proses	Berhasil

h. Evaluation

Pada tahapan ini, peneliti meninjau seluruh proses pengembangan sistem dan menyusun laporan atau skripsi sebagai referensi untuk peneliti lain.

6. KESIMPULAN

Berdasarkan hasil pembahasan dari Penerapan Algoritma K-Nearest Neighbor Pada Analisis Sentimen Terhadap Ulasan Aplikasi myBCA Di Google Play Store, Maka dapat ditarik kesimpulan :

- Sistem penerapan algoritma K-Nearest Neighbor untuk analisis sentimen ulasan aplikasi myBCA di Google Play Store dibuat sebagai website menggunakan library Streamlit dan GitHub Codespaces.
- Penerapan model dan algoritma K-Nearest Neighbors pada sistem mengikuti metode SKKNI dengan tujuh tahapan Business Understanding, Data Understanding, Data Preparation, Modelling, Model Evaluation, Deployment, Testing, dan Evaluation. Proses ini terbukti efisien dalam waktu penerapan model untuk analisis sentimen hingga tahap deployment, Dengan tingkat akurasi yang cukup bagus.
- Jumlah ulasan yang di scraping dari aplikasi myBCA di Google Play Store dengan google-play-scraper yaitu 8.250 ulasan.
- Pada proses labeling data, digunakan kamus InSet dan SentiWords untuk menentukan sentimen setiap ulasan yang telah di-scraping. Selain itu, kamus Vader diadaptasi dengan kata-kata yang telah diperbaiki secara

manual untuk mengatasi kata atau ulasan yang tidak terdeteksi oleh kamus sentimen.

5. Pada proses Data Preparation, sebelum Peneliti menggunakan library secara keseluruhan dengan NLP-ID. Peneliti mencoba menggunakan beberapa library, seperti sastrawi dan GPT 4o versi Plus. Pada Penelitian ini, untuk library sastrawi dengan jumlah ulasan 8.250, pada tahap data preparation terutama stemming memiliki waktu pemrosesan yang normal tidak lama dan tidak juga cepat tetapi proses menghilangkan kata imbuhan pada stemming masih belum semua dapat dihilangkan. Sedangkan untuk GPT 4o versi Plus dengan jumlah ulasan 8.250, memiliki waktu pemrosesan yang agak sedikit lebih lama dibandingkan sastrawi, tetapi memiliki pemrosesan yang cukup bagus dalam menghilangkan kata imbuhan pada ulasan, hanya saja untuk library GPT 4o versi Plus ini terkadang jika sudah mencapai titik tertentu dalam memproses data bisa force stop process dikarenakan terlalu banyak membaca data melebihi kapasitasnya. Maka dari itu Peneliti menggunakan library NLP-ID pada tahap Data Preparation ini dikarenakan memiliki keunggulan di waktu memproses data dan sedikit lebih baik dalam menghilangkan imbuhan pada kata dengan jumlah ulasan 8.250.
6. Penggunaan SMOTE untuk *oversampling* data latih guna menyamakan jumlah setiap kategori sentimen dan meningkatkan performa model.
7. Pada penentuan nilai optimal K, Peneliti menggunakan 2 metode yaitu Cross-Validation guna mencari nilai skor yang tinggi dan Elbow Method untuk menentukan skor yang sudah dirangkingkan di tahap Cross Validation untuk mencari nilai skor yang memiliki nilai error yang rendah dan akurasi yang tinggi. Pada Penelitian ini, penggunaan 2 metode penentuan nilai K itu sangat berpengaruh terhadap penelitian yang peneliti lakukan. Dikarenakan sebelum tingkat akurasi KNN mencapai 81%, sangat susah dalam mencapai nilai akurasi tersebut setelah dilakukan percobaan dengan merubah nilai perulangan cross validation dan range dari elbow method nya akurasi hanya berkisar di 69-72%.
8. Pada pembuatan model K-Nearest Neighbor, dibangun menggunakan perbandingan 70% untuk data latih dan 30% untuk data uji dari ±8.250 data ulasan pada aplikasi myBCA di Google Play Store.
9. Pada penelitian ini guna meningkat akurasi dari model, peneliti juga mencoba melakukan tahap manual labeling dengan 50% dari total ulasan dan menghasilkan akurasi pada model KNN yang masih belum mencukupi yaitu 60-64%. Maka dari itu Peneliti tetap menggunakan labeling dengan kamus sentimen InSet dan SentiWords.

DAFTAR PUSTAKA

- [1] Nurian, A., & Sari, B. N. (2023). Analisis Sentimen Ulasan Pengguna Aplikasi Google Play Menggunakan Naïve Bayes. *JITET (Jurnal Informatika dan Teknik Elektro Terapan)*, 829-835.
- [2] Habibah, N., Budianta, E., Fikry, M., & Iskandar, I. (Februari, 2023). Analisis Sentimen Mengenai Penggunaan E-Wallet Pada Google Play Menggunakan Lexicon Based dan K-Nearest Neighbor. *JURIKOM (Jurnal Riset Komputer)*, Vol. 10 No. 1, 192-200.

- [3] Alamsyah, A. A., & Mulyati, S. (Agustus 2023). Implementasi Algoritme K-Nearest Neighbour Dan Lexicon Based Untuk Analisis Sentimen Kepuasan Pengguna Aplikasi Gamedia Digital Pada Media Sosial Twitter. *Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI)*, Volume 2, Nomor 2, September 2023, 521-529.
- [4] Furqan, M., Sriani, & Sari, S. M. (Februari 2022). Analisis Sentimen Menggunakan K-Nearest Neighbor Terhadap New Normal Masa Covid-19 Di Indonesia. *Techno.COM*, 21, 52-61.
- [5] Alrajak, M. S., Ernawati, I., & Nurlaili, I. (Agustus 2020). Analisis Sentimen Terhadap Pelayanan Pt Pln Di Jakarta Pada Twitter Dengan Algoritma K-Nearest Neighbor (K-NN). *SENAMIKA*, 110-122.
- [6] Ferdyanadi, M., Setiawan, N. Y., & Bachtiar, F. A. (Februari 2022). Prediksi Potensi Penjualan Makanan Beku berdasarkan Ulasan Pengguna Shopee menggunakan Metode Decision Tree Algoritma C4.5 dan Random Forest (Studi Kasus Dapur Lilis). *J-PTIHK*, 588-596.
- [7] Falakh, F. (2022). Identifikasi Standar Kompetensi Kerja Nasional Indonesia (SKKNI) pada Pengembangan Kurikulum Program Studi Teknik Lingkungan. *Journal on Education*, 1796-1802.