sudo
Jurnal Teknik Informatika
https://jurnal.ilmubersama.com/index.php/sudo

Data Mining

Analisis Sentimen Keefektifan Belajar Bahasa Asing di Aplikasi Duolingo dengan Algoritma *Naïve Bayes*

Muhammad Aqshal Anindya Tratama, Fadli Santoso Murmita, Dimas Arsya Maulana, Cindy Renata, Raras Ailsa, Fuad Nur Hasan *

Fakultas Teknik, Program Studi Teknik Informatika, Universitas Bina Sarana Informatika, Margonda, Indonesia

INFORMASI ARTIKEL

Diterima Redaksi: 01 November 2025
Revisi Akhir: 19 Januari 2026
Diterbitkan *Online*: 05 Maret 2026

KATA KUNCI

Analisis Sentimen
Duolingo
Efektivitas Pembelajaran
Naïve Bayes
Pembelajaran Bahasa Asing

KORESPONDENSI (*)

Phone: +62 818-253-851
E-mail: fuad.fnu@bsi.ac.id

ABSTRAK

Aplikasi pembelajaran bahasa Duolingo telah diunduh lebih dari 500 juta kali, namun terdapat perdebatan mengenai keefektifannya. Ulasan pengguna di Google Play Store merupakan sumber data masif untuk mengevaluasi persepsi ini, namun volumenya terlalu besar untuk dianalisis secara manual. Penelitian ini bertujuan untuk menganalisis sentimen pengguna guna memahami lebih dalam persepsi efektivitas Duolingo sebagai media pembelajaran bahasa asing. Penelitian ini menggunakan 5.000 ulasan yang dikumpulkan melalui *scraping* dari Google Play Store. Metode *Natural Language Processing* (NLP) diterapkan, meliputi *text pre-processing* serta vektorisasi TF-IDF. Sebuah model klasifikasi sentimen biner yang telah dilabeli positif dan negatif dibangun menggunakan algoritma Multinomial *Naïve Bayes*. Hasil analisis data menunjukkan sentimen keseluruhan sangat positif, dimana 85.34% ulasan diklasifikasikan sebagai positif. Analisis kualitatif mengidentifikasi pendorong sentimen positif adalah efektivitas ("mudah paham", "bantu ajar") dan aspek gamifikasi ("seru"). Sebaliknya, sentimen negatif didominasi oleh keluhan pada fitur "sistem hati" dan "iklan" yang mengganggu. Model Naïve Bayes yang telah dilatih berhasil mencapai akurasi 92.81% dalam mengklasifikasikan data uji, membuktikan efektivitasnya untuk tugas ini terutama dalam tugas klasifikasi sentimen positif. Hasil ini mengonfirmasi persepsi positif terhadap keefektifan Duolingo, sekaligus memberikan masukan kritis mengenai model bisnis *freemium*-nya dan kekurangan lain yang dimiliki aplikasi pembelajaran ini.

PENDAHULUAN

Di era globalisasi saat ini, penguasaan bahasa asing telah menjadi sebuah kebutuhan fundamental, yang mendorong pergeseran signifikan dalam metodologi pembelajaran dari ranah konvensional ke platform digital. Transformasi ini menuntut standar baru, di mana pembelajaran daring bukan sekedar memberikan materi, tugas atau soal soal melalui media internet, namun pembelajaran daring harus direncanakan, dilaksanakan, serta dievaluasi sama halnya dengan pembelajaran yang terjadi di kelas [1]. Evaluasi terhadap efektivitas platform digital menjadi krusial untuk memastikan bahwa tujuan pedagogis tercapai.

Menjawab tantangan ini, hadirlah berbagai platform *Mobile-Assisted Language Learning* (MALL). Salah satu yang paling fenomenal adalah Duolingo. Duolingo adalah aplikasi pembelajaran bahasa gratis yang diluncurkan pada tahun 2011 oleh Luis von Ahn dan Severin Hacker. Dengan diperkenalkannya Duolingo, solusi inovatif yang mengusung pendekatan modern dan interaktif dalam proses pembelajaran bahasa, tanggap terhadap kebutuhan akan metode pembelajaran yang lebih efektif dan kontekstual [2]. Cara kerja Duolingo berfokus pada gamifikasi, mengubah proses belajar menjadi sebuah

https://doi.org/10.56211/sudo.v5i1.1354

permainan yang adiktif dan menyenangkan karena tampilan visual dan fitur-fiturnya yang menarik. Pengguna menyelesaikan pelajaran-pelajaran singkat (*bite-sized lessons*), mendapatkan poin (XP), dan bersaing dalam liga mingguan.

Popularitas pendekatan ini didukung oleh berbagai fitur berguna yang dirancang untuk menjaga motivasi pengguna. Fitur-fitur tersebut di antaranya adalah *Streaks* (rentetan belajar harian), *Health* (sistem 'hati' yang membatasi kesalahan), *Duolingo Stories* (latihan pemahaman bacaan interaktif), serta pohon pembelajaran (*learning path*) yang adaptif. Keberhasilan model ini tercermin jelas dari basis penggunanya yang masif. Hingga saat ini, Duolingo telah diunduh lebih dari 500 juta kali di Google Play Store dengan akumulasi rating yang secara umum sangat memuaskan.

Volume unduhan yang masif ini berbanding lurus dengan jumlah ulasan pengguna yang ditinggalkan di Play Store. Ulasan-ulasan ini merupakan repositori data yang kaya akan opini, kritik, dan testimoni langsung mengenai pengalaman belajar. Menganalisis data dalam jumlah besar, dalam penelitian ini sebanyak 5000 ulasan secara manual adalah tugas yang mustahil. Oleh karena itu, penelitian ini akan menerapkan metode komputasi *Natural Language Processing* (NLP) melalui teknik analisis sentimen. Algoritma yang dipilih untuk mengklasifikasikan sentimen ini adalah *Naïve Bayes* Classifier. Metode ini dipilih karena efisiensinya, keandalannya dalam menangani data teks bervolume besar, dan performanya yang telah terbukti andal dalam berbagai tugas klasifikasi teks.

Berlandaskan latar belakang tersebut, tujuan utama dari penelitian ini adalah untuk menganalisis sentimen atau memahami lebih dalam lagi seberapa efektif belajar bahasa asing menggunakan Duolingo berdasarkan 5000 ulasan otentik dari pengguna yang telah memberikan ulasan di Google Play Store. Dengan menerapkan algoritma Naïve Bayes, penelitian ini akan mengklasifikasikan ulasan ke dalam kategori sentimen positif dan negatif, mengidentifikasi aspek-aspek spesifik dari aplikasi yang paling sering dikomentari pengguna. Hasil analisis ini diharapkan dapat memberikan evaluasi berbasis data mengenai keefektifan Duolingo sebagai alat pembelajaran bahasa.

TINJAUAN PUSTAKA

Duolingo

Duolingo adalah salah satu platform *Mobile-Assisted Language Learning* (MALL) paling populer di dunia. Diluncurkan pada tahun 2011, aplikasi ini merevolusi pembelajaran bahasa dengan menerapkan model *freemium* (gratis dengan layanan premium opsional) dan pendekatan utama berbasis gamifikasi. Proses belajar dirancang seperti permainan, di mana pengguna menyelesaikan pelajaran singkat, mendapatkan poin pengalaman (XP), dan naik level, yang terbukti efektif dalam menjaga motivasi. Duolingo tidak hanya berfokus pada aplikasi pembelajaran utama, tetapi juga telah memperluas ekosistemnya ke ranah sertifikasi akademik dan profesional. Selain pembelajaran bahasa sebagai platform intinya, Duolingo juga mengembangkan Duolingo English Test, pilihan sertifikasi bahasa yang nyaman, terjangkau, dan diterima oleh ribuan institusi di seluruh dunia [3]. Popularitasnya yang masif, dibuktikan dengan lebih dari 500 juta unduhan di Google Play Store, menjadikan basis ulasan penggunanya sebagai sumber data yang sangat kaya untuk menganalisis persepsi publik terhadap keefektifan metode pembelajarannya.

Analisis Sentimen

Analisis sentimen merupakan bagian dari *Natural Language Processing* (NLP) yang bertujuan untuk mengenali, mengambil, dan mengukur pandangan, perasaan, serta subjektivitas yang terkandung dalam sebuah teks. Tujuannya adalah untuk menentukan sikap atau sentimen penulis terhadap suatu topik, produk, atau layanan. Dalam konteks bisnis dan sosial, analisis sentimen merupakan alat yang sangat berharga untuk mendapatkan pemahaman dari *feedback* pelanggan dalam skala besar. Analisis sentimen sering digunakan dalam pemantauan media sosial untuk mendeteksi sentimen yang ada, termasuk opini positif dan negatif [4]. Dalam penelitian ini, analisis sentimen diterapkan pada 5000 ulasan pengguna Duolingo untuk mengekstrak dan mengklasifikasikan persepsi mereka mengenai efektivitas aplikasi tersebut sebagai alat bantu belajar.

 https://doi.org/10.56211/sudo.v5i1.1354

Text Pre-processing

Data teks yang diperoleh dari sumber seperti ulasan Play Store pada dasarnya adalah data tidak terstruktur. Data ini sering kali mengandung banyak *noise* atau elemen yang tidak relevan, seperti penggunaan huruf kapital yang tidak konsisten, tanda baca, angka, *stopwords*, dan kata-kata berimbuhan.

Pre-processing adalah tahapan fundamental dalam NLP untuk membersihkan data mentah yang digunakan supaya lebih terstruktur dan dapat dipahami oleh algoritma *machine learning*. Pre-processing dilakukan untuk membersihkan *noise* untuk mendapatkan informasi akurat sebanyak mungkin dari teks [5].

Algoritma Naïve Bayes

Naïve Bayes adalah salah satu algoritma klasifikasi probabilistik, metode atau pendekatan yang menggunakan konsep dan data peluang untuk menganalisis ketidakpastian dan membuat prediksi yang didasarkan pada Teorema Bayes. Algoritma ini bekerja dengan menghitung probabilitas sebuah data termasuk dalam kelas tertentu berdasarkan fitur-fitur atau kata-kata di dalamnya.

Disebut "*Naïve*" (lugu) karena algoritma ini membuat asumsi yang kuat namun sederhana yang berarti setiap fitur atau kata bersifat independen satu sama lain. Meskipun asumsi ini jarang terpenuhi dalam data teks di dunia nyata, *Naïve Bayes* terbukti sangat cepat, efisien, dan memiliki performa yang kuat, terutama untuk klasifikasi teks. Metode ini efektif dalam mengklasifikasikan opini ke dalam kategori positif, netral, atau negatif mengenai suatu produk atau isu [6]. Untuk data teks, varian yang paling sering digunakan adalah Multinomial *Naïve Bayes*, yang ideal untuk menghitung frekuensi kemunculan kata.

Confusion Matrix

Evaluasi model adalah tahap krusial untuk mengetahui seberapa baik performa model klasifikasi yang telah dibangun. *Confusion matrix* adalah tabel yang menyatakan klasifikasi jumlah data uji yang benar dan jumlah data uji yang salah [7]. Dalam kasus klasifikasi dua kelas, *confusion matrix* akan berukuran 2x2. Matriks ini menunjukkan:

True Positive (TP) dan True Negative (TN)

Merupakan jumlah data yang berhasil diklasifikasikan dengan benar sesuai dengan kelas aslinya. TP menunjukkan data positif yang terdeteksi benar, sedangkan TN menunjukkan data negatif yang terdeteksi benar.

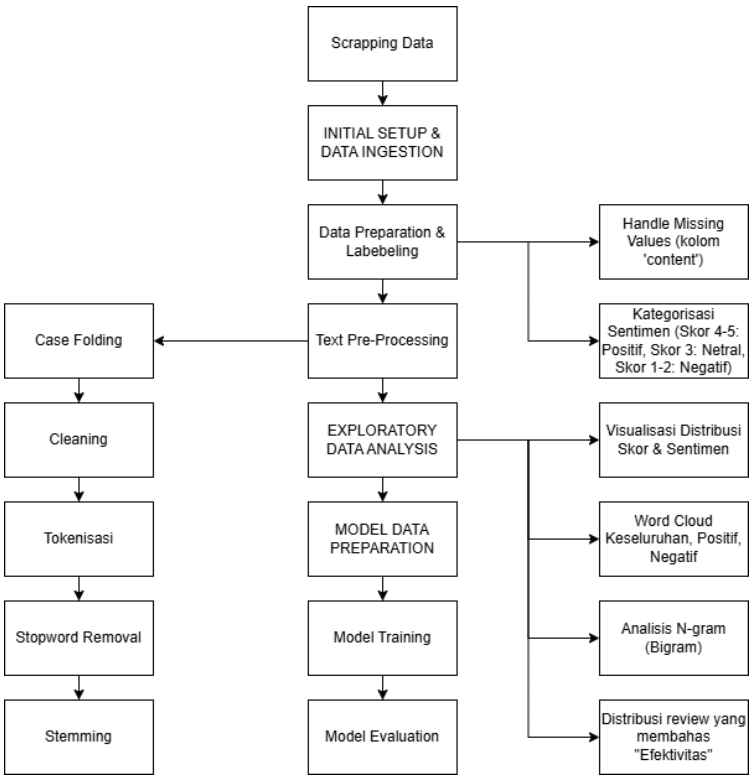
False Positive (FP) dan False Negative (FN)

Menunjukkan jumlah data yang salah diklasifikasikan. FP terjadi ketika data negatif diklasifikasikan sebagai positif, sedangkan FN terjadi ketika data positif diklasifikasikan sebagai negatif.

Dari nilai-nilai dalam *confusion matrix* inilah metrik evaluasi lain seperti akurasi, presisi, dan Perolehan dapat dihitung untuk mengukur keandalan model secara kuantitatif.

METODOLOGI

Metodologi penelitian ini menguraikan tahapan sistematis yang dilakukan untuk menganalisis sentimen pengguna Duolingo. Alur kerja penelitian ini, mulai dari pengumpulan data hingga evaluasi model, dirancang untuk memastikan bahwa data diproses secara valid dan model yang dihasilkan dapat dipertanggungjawabkan. Gambar 1 memperlihatkan setiap tahapan penelitian:



Gambar 1. Tahapan Penelitian

Data Scraping

Tahap awal penelitian adalah pengumpulan data mentah. Teknik Scraping merupakan teknik untuk mengubah data web yang tidak terstruktur menjadi data terstruktur yang dapat disimpan dan dianalisis dalam database atau spreadsheet pusat [8]. Data yang digunakan adalah ulasan pengguna aplikasi Duolingo yang bersumber dari platform Google Play Store. Proses *scraping* dilakukan untuk mengumpulkan 5.000 ulasan terbaru. Data yang diambil dari proses ini tidak hanya berisi teks ulasan, tetapi juga atribut pendukung seperti rating bintang 1 sampai 5 dan tanggal ulasan. *Scraping* dilakukan menggunakan *tools* google-play-scraper.

Initial Setup & Data Ingestion

Setelah data mentah berhasil dikumpulkan dan dikemas dalam format .CSV, data tersebut dimuat ke dalam lingkungan kerja analisis. Tahap ini melibatkan persiapan *environment* Google Colab dan mengimpor *library* yang dibutuhkan. Data mentah kemudian dibaca dan diubah menjadi struktur data tabular (*DataFrame*) menggunakan Pandas untuk mempermudah manipulasi dan analisis.

Data Preparation

Sebelum masuk ke tahap pemrosesan teks, data tabular terlebih dahulu dibersihkan agar kualitas dataset tetap terjaga. Proses pembersihan ini mencakup beberapa langkah utama, yaitu penanganan data kosong (*handling missing values*) dengan memeriksa dan menghapus baris data yang tidak memiliki isi ulasan, penghapusan duplikat (*removing duplicates*) dengan mengidentifikasi serta menghapus ulasan yang identik atau terposting lebih dari satu kali, dan seleksi fitur (*feature selection*) dengan memilih kolom yang relevan untuk penelitian, seperti kolom Ulasan dan Skor.

Text Pre-processing

Ini adalah tahap krusial dalam *Natural Language Processing* (NLP) untuk membersihkan *noise* dari data teks ulasan. Data yang dikumpulkan merupakan data yang tidak terstruktur sehingga membutuhkan *Pre-Processing* [9]. Tujuannya adalah mengubah teks tidak terstruktur menjadi format yang bersih dan konsisten, siap untuk dianalisis. Tahapan yang dilakukan meliputi:

 https://doi.org/10.56211/sudo.v5i1.1354

Case Folding

Menyeragamkan seluruh teks ulasan menjadi huruf kecil (*lowercase*) untuk menghindari duplikasi makna. Misalnya, kata “Bagus” dan “bagus” dianggap sama.

Cleaning

Menghapus karakter, angka, atau simbol yang tidak relevan dengan sentimen, seperti URL, tanda baca, *mention* (@username), *hashtag* (#), dan emoji.

Tokenizing

yaitu memecah teks ulasan menjadi unit-unit kata individual yang disebut *token*. Proses ini bertujuan mengubah kalimat panjang menjadi daftar kata agar dapat dianalisis secara statistik atau linguistik. Setiap *token* mewakili satuan makna terkecil yang nantinya digunakan untuk ekstraksi fitur atau pembentukan representasi vektor teks.

Stopword Removal

Menghapus kata-kata umum dalam Bahasa Indonesia yang frekuensi kemunculannya sering namun tidak memiliki bobot sentimen, seperti "yang", "di", "dan", "dari", "ke". Proses ini menggunakan daftar *stopwords* standar Bahasa Indonesia.

Stemming

Mengubah setiap kata berimbuhan ke bentuk kata dasarnya, misalnya "mempelajari" atau "dipelajari" diubah menjadi "ajar". Proses ini menggunakan *stemmer* Bahasa Indonesia, seperti *library* Sastrawi.

Exploratory Data Analysis (EDA)

Setelah data dibersihkan, dilakukan analisis eksploratif untuk memperoleh wawasan awal terhadap karakteristik ulasan. Analisis ini berfokus pada visualisasi dan pemahaman distribusi data. Langkah-langkah yang dilakukan meliputi visualisasi distribusi rating bintang dari 1 hingga 5 untuk melihat proporsi kepuasan pengguna, serta pembuatan *word cloud* dari kata-kata yang paling sering muncul setelah proses *pre-processing*. Melalui *word cloud* ini, dapat diidentifikasi topik utama yang sering dibicarakan oleh pengguna.

Model Data Preparation

Data teks yang bersih perlu diubah menjadi format yang dapat dipahami oleh algoritma *machine learning*. Tahap ini meliputi:

Pelabelan Data (Labeling)

Tahap ini menentukan sentimen dari setiap ulasan berdasarkan rating bintang. Dalam penelitian ini, rating 4 dan 5 dikategorikan sebagai positif, rating 3 sebagai netral, dan rating 1 serta 2 sebagai negatif

Ekstraksi Fitur (Feature Extraction)

Data teks yang telah ditokenisasi dikonversi menjadi representasi numerik atau vektor menggunakan metode TF-IDF (Term Frequency–Inverse Document Frequency). Metode ini memberikan bobot pada setiap kata berdasarkan frekuensi kemunculannya di satu ulasan serta tingkat keunikannya di seluruh korpus. Representasi vektor inilah yang kemudian menjadi masukan bagi algoritma klasifikasi.

Pembagian Data (Data Splitting)

Dataset dibagi menjadi dua bagian, yaitu data latih dan data uji. Data latih digunakan untuk membangun dan “mengajarkan” model *Naïve Bayes*, sedangkan data uji digunakan untuk mengevaluasi performa model dalam mengenali sentimen baru yang belum pernah dilihat sebelumnya.

Model Training

Pada tahap ini, model klasifikasi dibangun. Data latih dimasukkan ke dalam Algoritma *Naïve Bayes Classifier*. Secara spesifik, varian yang digunakan adalah Multinomial *Naïve Bayes*, yang sangat cocok untuk data teks hasil perhitungan frekuensi. Model akan mempelajari pola probabilitas kata-kata yang cenderung muncul, terutama pada sentimen positif dan negatif.

Model Evaluation

Setelah model selesai dilatih, kinerjanya diuji menggunakan 20% Data Uji. Model diminta untuk memprediksi sentimen dari data uji, dan hasilnya dibandingkan dengan label sentimen yang sebenarnya. Evaluasi dilakukan menggunakan:

Confusion Matrix

Sebuah tabel yang merangkum hasil prediksi, menunjukkan berapa banyak prediksi Benar dan berapa banyak prediksi Salah. Dimana ada 4 faktor dalam pengukuran performa dengan *Confusion Matrix*, diantaranya yaitu *Accuracy, Precision, Recall, dan F1-Score* [10].

Metrik Kinerja

confusion matrix, dihitung dari nilai *Accuracy, Precision, Recall, dan F1-Score* untuk mengukur seberapa andal dan akurat model yang telah dibangun.

HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil dari analisis sentimen terhadap ulasan aplikasi Duolingo menggunakan metode klasifikasi *Naive Bayes*. Pembahasan dimulai dari analisis data eksploratif untuk memahami karakteristik dataset, dilanjutkan dengan tahapan implementasi model, hingga evaluasi kinerja model berdasarkan metrik yang telah ditentukan.

Data Scraping

Penelitian ini berdasar pada data primer yang telah dikumpulkan langsung dari ulasan pengguna aplikasi Duolingo di Google Play Store. Sesuai dengan metodologi yang telah dijelaskan, proses *web scraping* dilakukan untuk mengumpulkan 5.000 data ulasan. Data mentah yang berhasil diperoleh terdiri dari lima kolom. Struktur data dari hasil *scraping* disajikan pada Tabel 1.

Tabel 1. Hasil *Scraping*

reviewId	userName	content	score	at
8d6ceeb5-d273-43e3-a75d-2b971aa510b1	Pengguna Google	bagus untuk belajar dan mudah dipahami	5	10/25/2025
00161091-9070-4e11-b7a1-88eae84e260d	Pengguna Google	Sistem baru = downgrade. Alih-alih membantu pengguna...	1	10/25/2025
06854f1e-a20e-4293-adc6-b25149a32d33	Pengguna Google	seruuuu banget dapat membantu belajar dari nol	5	10/25/2025
455cef56-4066-4a2d-a3ac-6515a6740b79	Pengguna Google	maaf sedikit kritik untuk anda pembuat aplikasi ini…	2	10/25/2025
ce233805-f0b9-4583-a330-75cb2c453cf8	Pengguna Google	baik, saya jadi tau cara bicara atau penyampaian bahasa Inggris yang benar	4	10/25/2025

Dari kelima kolom di Tabel 1 tersebut, dua kolom utama yang menjadi fokus dalam penelitian ini adalah *content* (isi ulasan) dan *score* (rating bintang). Kolom content akan menjadi data teks utama yang akan diproses dan dianalisis sentimennya. Kolom *score* akan digunakan sebagai dasar untuk pelabelan sentimen positif, negatif, dan netral pada tahap persiapan data model.

Text Pre-processing

Setelah data mentah dikumpulkan, tahapan krusial selanjutnya adalah *text pre-processing* atau pembersihan data teks. Tahapan ini bertujuan untuk mengubah ulasan yang tidak terstruktur menjadi data bersih yang konsisten dan siap untuk diubah menjadi vektor numerik. Untuk mengilustrasikan proses ini, contoh penerapan tahapan *pre-processing* secara bertahap pada salah satu ulasan pengguna dapat dilihat pada Tabel 2.

 https://doi.org/10.56211/sudo.v5i1.1354

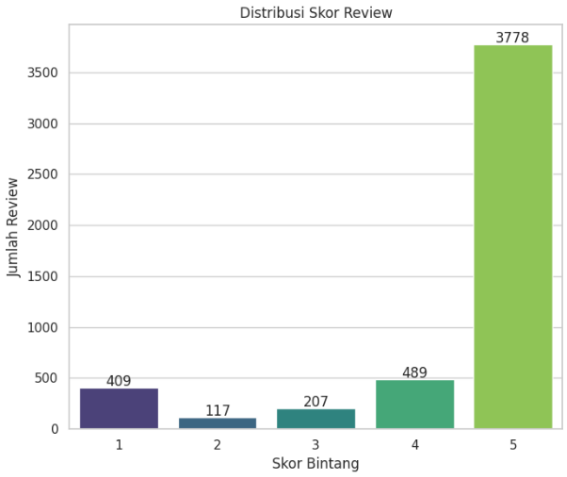
Table 2. *Pre-processing*

Proses	Hasil
Data Mentah	Karena lebih mudah untuk belajar bahasa terutama bahasa Inggris dan juga Duolingo menyenangkan,seru,asik,menambah ilmu berbahasa kepada orang di luar negeri mau pun di Asia,maupu di Eropa.
Case Folding	karena lebih mudah untuk belajar bahasa terutama bahasa inggris dan juga duolingo menyenangkan,seru,asik,menambah ilmu berbahasa kepada orang di luar negeri mau pun di asia,maupu di eropa
Cleaning	karena lebih mudah untuk belajar bahasa terutama bahasa inggris dan juga duolingo menyenangkan seru asik menambah ilmu berbahasa kepada orang di luar negeri mau pun di asia maupu di eropa
Tokenisasi	['karena', 'lebih', 'mudah', 'untuk', 'belajar', 'bahasa', 'terutama', 'bahasa', 'inggris', 'dan', 'juga', 'duolingo', 'menyenangkan', 'seru', 'asik', 'menambah', 'ilmu', 'berbahasa', 'kepada', 'orang', 'di', 'luar', 'negeri', 'mau', 'pun', 'di', 'asia', 'maupu', 'di', 'eropa']
Stopword Removal	['mudah', 'belajar', 'bahasa', 'terutama', 'bahasa', 'inggris', 'menyenangkan', 'seru', 'asik', 'menambah', 'ilmu', 'berbahasa', 'negeri', 'asia', 'maupu', 'eropa']
Stemming	['mudah', 'ajar', 'bahasa', 'utama', 'bahasa', 'inggris', 'senang', 'seru', 'asik', 'tambah', 'ilmu', 'bahasa', 'negeri', 'asia', 'maupu', 'eropa']

Distribusi Data

Distribusi Rating Pengguna

Analisis pertama adalah melihat distribusi skor asli yang diberikan oleh pengguna. Distribusi ini penting untuk mendapatkan gambaran umum tingkat kepuasan pengguna sebelum dilabeli sentimen.

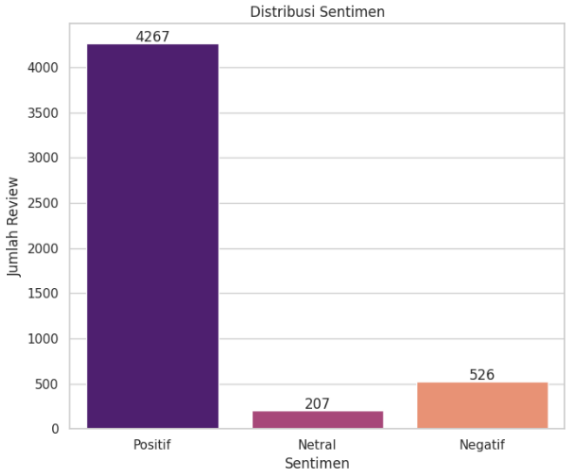


Gambar 2. Visualisasi Distribusi Skor

Data ini menunjukkan distribusi yang sangat *right-skewed* (miring ke kanan), di mana rating bintang 5 memiliki sebanyak 3.778 ulasan dan sangat mendominasi. Hal ini mengindikasikan bahwa secara umum, mayoritas pengguna memiliki impresi awal yang sangat positif terhadap aplikasi Duolingo di Play Store.

Distribusi Sentimen

Langkah selanjutnya dalam persiapan pemodelan adalah mengonversi skor rating menjadi label sentimen. Dari tiga kategori yang ada (positif, negatif, netral), hanya sentimen positif dan negatif yang akan dianalisis lebih lanjut dalam penelitian ini, mengingat sifat ambigu dari data netral. Perbandingan distribusi dapat dilihat pada Gambar 3.

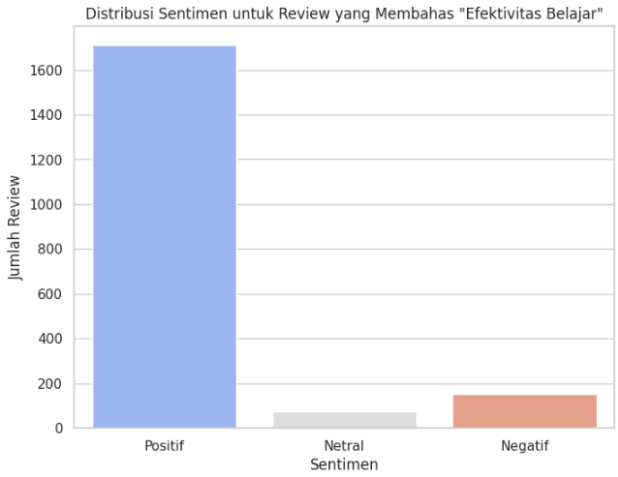


Gambar 3. Visualisasi Distribusi Sentimen

Jika dipersentasekan, 85.34% ulasan bersifat positif, 10.52% bersifat negatif, dan 4.14% bersifat netral. Ini mengonfirmasi temuan awal bahwa sentimen umum terhadap Duolingo sangatlah positif.

Distribusi Sentimen Spesifik "Efektivitas Belajar"

Analisis ini adalah inti dari tujuan penelitian, yaitu untuk memahami sentimen tidak hanya secara umum, tetapi secara spesifik terkait keefektifan belajar. Dilakukan pemfilteran terhadap 5.000 ulasan untuk menemukan data yang mengandung kata kunci yang relevan dengan proses belajar, yaitu: 'ajar', 'efektif', 'paham', 'lancar', 'ngerti', 'kosakata', 'baca', 'tulis', 'bicara', dan 'mudah'. Perbandingan distribusi dapat dilihat pada Gambar 4.



Gambar 4. Visualisasi Distribusi Keefektifan

Dari proses pemfilteran tersebut, ditemukan 1.938 ulasan yang secara spesifik membahas topik keefektifan belajar. Secara visual, terlihat bahwa bahkan di dalam sub-set ulasan yang spesifik ini, sentimen Positif tetap sangat dominan. Ulasan negatif dan netral ada, namun jumlahnya jauh lebih kecil. Hal ini memberikan indikasi awal yang kuat bahwa pengguna yang secara eksplisit mendiskusikan fungsi Duolingo sebagai alat bantu belajar cenderung memberikan sentimen positif terhadap keefektifan aplikasi tersebut.

Analisis Wordcloud

Untuk lebih mendalami tema dan kata kunci yang paling sering dibicarakan pengguna, *Word Cloud* dibuat. Visualisasi ini menampilkan frekuensi kata, di mana kata yang paling sering muncul akan ditampilkan dengan ukuran *font* terbesar. Besar kecil ukuran dalam word cloud berpengaruh dengan frekuensi kemunculan kata dalam data opini [11]. Analisis ini dilakukan pada tiga level penting: keseluruhan data, ulasan positif, dan ulasan negatif.

Word Cloud Sentimen Keseluruhan

Word Cloud pertama dibuat dari gabungan 5.000 ulasan setelah melalui *pre-processing*. Ini memberikan gambaran umum tentang topik yang paling dominan di seluruh dataset.

Gambar 5. Visualisasi *Word Cloud* Gabungan

Seperti yang terlihat pada Gambar 5, kata-kata yang paling dominan adalah "bagus", "ajar", "bahasa", "suka", "bantu", dan "seru". Dominasi kata-kata positif ini sejalan dengan temuan pada analisis distribusi, di mana 85.34% ulasan memang diklasifikasikan sebagai positif.

Word Cloud Sentimen Positif

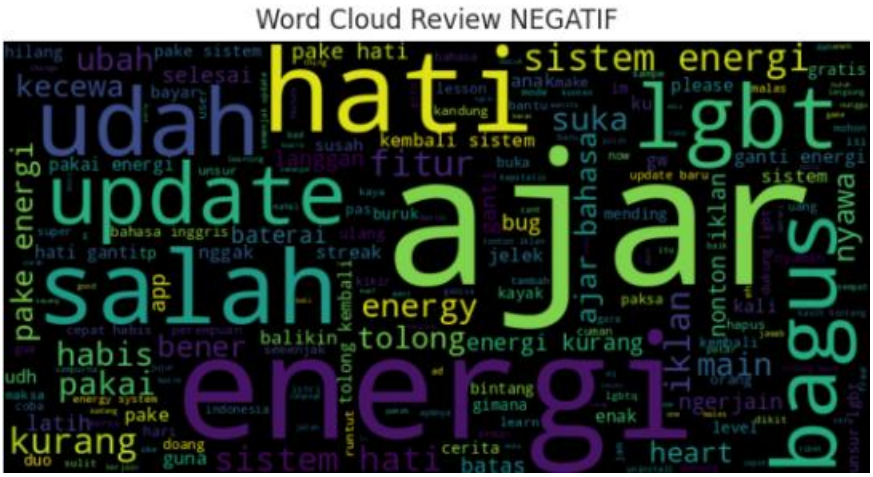
Untuk mempertajam analisis, *Word Cloud* kedua dibuat khusus dari 4.267 ulasan yang telah teridentifikasi sebagai sentimen positif.

Gambar 6. Visualisasi *Word Cloud Positif*

Pada Gambar 6, kata-kata yang lebih sering muncul pada ulasan positif menjadi lebih jelas. Kata-kata seperti "bagus", "suka", "bantu", "mudah", "paham", "senang", "seru", dan "manfaat" sangat menonjol. Hal ini mengindikasikan bahwa kepuasan pengguna sangat erat kaitannya dengan persepsi bahwa aplikasi ini efektif, membantu, dan menyenangkan.

Word Cloud Sentimen Negatif

Analisis *Word Cloud* yang paling krusial adalah pada ulasan dengan sentimen negatif sebanyak 526 ulasan. Visualisasi ini membantu mengidentifikasi titik masalah utama (*pain points*) yang dikeluhkan pengguna.



Gambar 7. Visualisasi *Word Cloud* Negatif

Berbeda drastis dengan dua *Word Cloud* sebelumnya, Gambar 7 didominasi oleh kata-kata yang berkonotasi keluhan. Kata-kata yang paling menonjol adalah "hati", "energi", "iklan", "salah", "kurang", dan "update".

Analisis "Hati" dan "Energi." Dua kata ini merujuk pada fitur gamifikasi utama Duolingo. Pada aplikasi Duolingo, jika pengguna salah menjawab soal, "energi" atau "hati" mereka akan berkurang. Ketika energi ini habis, pengguna tidak dapat melanjutkan pelajaran dan dipaksa untuk berhenti atau membayar versi premium. Kemunculan dominan kata ini di ulasan negatif menunjukkan bahwa fitur ini menimbulkan frustrasi besar bagi pengguna gratis.

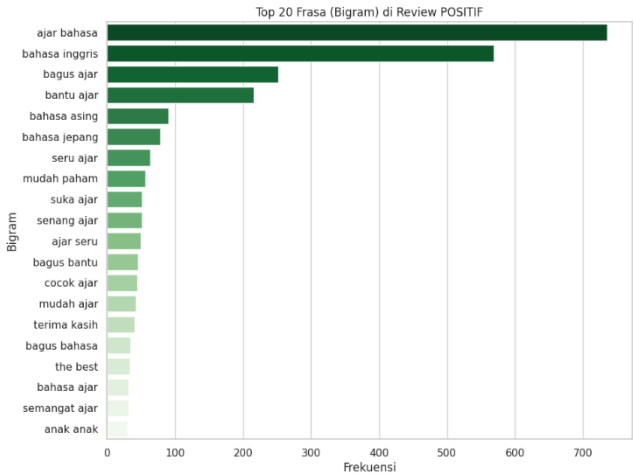
Analisis "Iklan." Kata "iklan" juga muncul dengan frekuensi tinggi. Duolingo menayangkan iklan di beberapa kesempatan. Banyaknya keluhan terkait "iklan" menunjukkan bahwa pengguna merasa kemunculannya terlalu sering atau intrusif, sehingga mengganggu pengalaman belajar mereka dan menyebabkan keresahan.

Analisis Bigram Frasa

Untuk mendapatkan pemahaman yang lebih kontekstual, analisis dilanjutkan menggunakan N-gram, secara spesifik bigram. Analisis *Word Cloud* efektif untuk melihat kata kunci, namun analisis bigram dapat mengungkap frasa atau konsep yang lebih spesifik. Misalnya, kata "hati" pada *Word Cloud* negatif baru mendapatkan makna penuh ketika dipasangkan dengan kata "sistem" menjadi "sistem hati". Frasa-frasa yang dihasilkan oleh bigram memberikan frasa yang lebih bermakna [12]

Bigram Frasa Pada Sentimen Positif

Analisis bigram pada 4.267 ulasan positif dilakukan untuk mengidentifikasi pujian yang paling sering diungkapkan pengguna dapat dilihat pada Gambar 8.



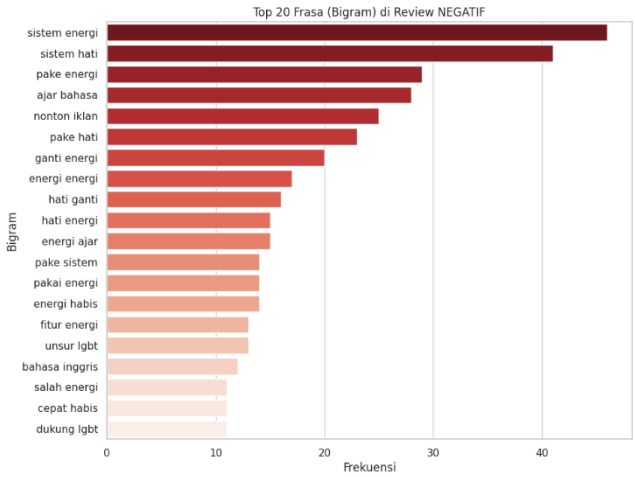
Gambar 8. Visualisasi Bigram Positif

 https://doi.org/10.56211/sudo.v5i1.1354

Temuan pada Gambar 8 sangat memperkuat hasil EDA sebelumnya. Analisis frasa menunjukkan bahwa "ajar bahasa" dan "bahasa inggris" adalah tema yang paling dominan. Hal ini mengindikasikan bahwa kepuasan pengguna sangat didorong oleh fungsi inti aplikasi dan ketersediaan subjek yang paling diminati, yaitu bahasa Inggris. Selain itu, pujian spesifik terhadap keefektifan aplikasi juga terlihat jelas melalui frasa seperti "bagus ajar", "bantu ajar", "mudah paham", dan "seru ajar". Temuan-temuan ini mengonfirmasi bahwa persepsi positif pengguna terbentuk karena aplikasi Duolingo berhasil menggabungkan tiga pilar utama, yaitu efektivitas pembelajaran yang tercermin dari frasa "mudah paham", relevansi subjek yang diwakili oleh "bahasa inggris", dan aspek pengalaman yang menyenangkan yang terlihat dari "seru ajar".

Bigram Frasa Pada Sentimen Negatif

Analisis bigram pada 526 ulasan negatif sangat krusial untuk memvalidasi hipotesis dari *Word Cloud* negatif dan menunjuk sumber masalah secara presisi.



Gambar 9. Visualisasi Bigram Negatif

Hasil pada Gambar 9 memberikan bukti yang sangat kuat mengenai sumber frustrasi pengguna. Keluhan ini diperjelas dengan frasa turunan seperti "pake energi", "pake hati", "energi habis", "fitur energi", dan "salah energi". Frasa-frasa ini melukiskan skenario yang jelas di mana pengguna frustrasi karena energi mereka habis ketika salah menjawab, dan kemudian dipaksa untuk menggunakan sistem tersebut. Frasa kedua yang paling sering dikeluhkan, "nonton iklan", secara langsung mengikat masalah "energi habis" ini dengan model monetisasi aplikasi. Pengguna merasa frustrasi karena proses belajar mereka terinterupsi oleh keharusan menonton iklan untuk memulihkan energi atau melanjutkan pelajaran.

Hasil Pemodelan dengan Naïve Bayes

Setelah data dibersihkan dan dipahami melalui EDA, langkah berikutnya adalah membangun dan melatih model klasifikasi sentimen. Seperti yang telah dijelaskan dalam metodologi, algoritma yang digunakan adalah Multinomial *Naïve Bayes*, dan proses ini difokuskan pada klasifikasi biner.

Data Splitting

Untuk menyederhanakan masalah dan fokus pada sentimen yang paling kuat dan jelas, ulasan dengan label 'Netral' (sebanyak 207 data) tidak diikutsertakan dalam proses pelatihan model. Dengan demikian, model hanya dilatih untuk membedakan antara sentimen 'Positif' dan 'Negatif', menyisakan total 4.793 data untuk model yang terdiri dari 4.267 ulasan positif dan 526 ulasan negatif.

Dataset ini kemudian dibagi menjadi dua bagian, yaitu 80% untuk bagian data latih (*training data*) dan 20% untuk bagian data uji. Pembagian ini dilakukan secara terstratifikasi untuk memastikan bahwa proporsi kelas positif dan negatif di kedua set data tetap seimbang. Hasil pembagian data dapat dilihat pada Tabel 3.

Tabel 3. *Data Spitting*

Skenario Rasio Perbandingan	Total Data Untuk Model	Jumlah Data Latih	Jumlah Data Latih
80:20	4793	3834	959

Proses Pelatihan Model

Model dilatih dengan menggunakan fungsionalitas Pipeline dari *library* Scikit-learn. Pendekatan *pipeline* ini sangat efisien karena menggabungkan dua langkah utama ke dalam satu alur kerja.

Langkah pertama dalam *pipeline* ini adalah Vektorisasi (TfidfVectorizer). Tahap ini mengubah teks bersih dari data latih menjadi matriks vektor numerik menggunakan TF-IDF. Tujuannya adalah memberi bobot tinggi pada kata yang penting (sering muncul di satu ulasan, tapi jarang muncul di ulasan lain) dan bobot rendah pada kata yang terlalu umum. Bobot TF-IDF, *W(t,d)*, untuk sebuah kata (*t*) dalam sebuah ulasan (d) dihitung dengan:

Term Frequency (TF), atau *TF(t,d)*, mengukur seberapa sering kata *t* muncul dalam ulasan *d*. Rumus untuk TF adalah:

$$TF(t,d) = \frac{Jumlah\ kemunculan\ kata\ t\ di\ ulasan\ d}{Total\ kata\ di\ ulasan\ d} \tag{1}$$

Inverse Document Frequency (IDF), atau *IDF(t)*, mengukur seberapa unik kata *t* di seluruh korpus (semua ulasan). Rumus untuk IDF adalah:

$$IDF(t) = log\left(\frac{Total\ jumlah\ ulasan\ (N)}{Jumlah\ ulasan\ yang\ mengandung\ kata\ t\ (df_t)}\right) \tag{2}$$

Langkah kedua adalah Klasifikasi (MultinomialNB). Setelah teks diubah menjadi vektor TF-IDF, Tujuannya adalah menghitung probabilitas sebuah ulasan *d* masuk ke dalam kelas *c* (misalnya, 'Positif' atau 'Negatif'). Rumus ini didasarkan pada Teorema Bayes, dan untuk klasifikasi teks (Multinomial) disederhanakan menjadi:

$$P(c|d) \propto P(c) \prod_{i=1}^{n} P(t_i|c) \tag{3}$$

Di mana:

P(c|d) : Probabilitas ulasan *d* termasuk dalam kelas *c* (ini yang ingin kita cari).
P(c) : Probabilitas *prior* dari kelas *c*
$P(t_i|c)$: Probabilitas kata (term) t_i muncul dalam ulasan yang termasuk kelas *c*.

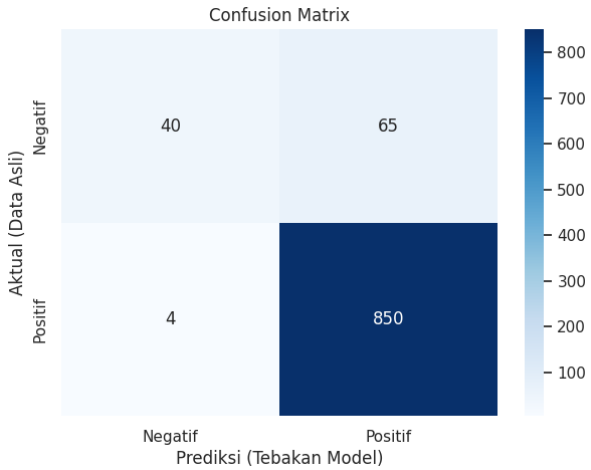
Inilah yang "dipelajari" oleh model selama pelatihan, yaitu seberapa sering kata "bagus" muncul di ulasan 'Positif' versus seberapa sering muncul di ulasan 'Negatif'. Model akan menghitung probabilitas ini untuk 'Positif' dan 'Negatif', dan menetapkan ulasan tersebut ke kelas dengan probabilitas tertinggi.

Hasil Evaluasi Model

Setelah model dilatih menggunakan 3.834 data latih, model tersebut diuji kinerjanya menggunakan 959 data uji yang belum pernah dilihat sebelumnya. Evaluasi ini sangat penting untuk mengetahui seberapa baik model *Naïve Bayes* dapat menggeneralisasi dan memprediksi sentimen pada data baru.

Confusion Matrix

Metrik evaluasi pertama adalah *Confusion Matrix*, yang memvisualisasikan performa model dengan membandingkan prediksi model (sumbu x) dengan data asli/aktual (sumbu y).



Gambar 10. *Confusion Matrix*

Dari matriks ini saja, terlihat bahwa model sangat akurat dalam memprediksi ulasan positif (hanya 4 kesalahan dari 854 data positif). Namun, model mengalami kesulitan lebih besar pada kelas negatif, di mana 65 dari 105 ulasan negatif salah diprediksi sebagai positif.

Classification Report

Analisis performa model dirangkum secara kuantitatif dalam *Classification Report* dan skor akurasi keseluruhan. Untuk mengevaluasi kinerja algoritma, kita menggunakan empat komponen dasar: *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Komponen-komponen ini, yang didapat dari *Confusion Matrix*, digunakan untuk menghitung metrik performa seperti akurasi, presisi, *recall* (perolehan), dan rata-rata harmonik (*f1-score)*, dengan rincian formula sebagai berikut:

Metrik pertama yang dievaluasi adalah Akurasi. Mengukur seberapa banyak prediksi yang benar (TP+TN) dari keseluruhan data uji. Nilai akurasi adalah matrik evaluasi yang digunakan untuk mengukur seberapa baik model klasifikasi atau prediksi dapat memberikan hasil yang benar atau sesuai dengan data yang ada [13].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{850 + 40}{850 + 40 + 65 + 4} = \frac{890}{959} = 0{,}93 \tag{4}$$

Selanjutnya adalah Presisi, yang mengukur tingkat ketepatan model saat memprediksi ‘Positif’. Dari semua yang diprediksi ‘Positif’ (TP+FP), seberapa banyak yang benar-benar ‘Positif’.

$$Precission = \frac{TN}{TN + FN} = \frac{40}{40 + 4} = \frac{40}{44} = 0{,}91 \tag{5}$$

Berbeda dengan Presisi, Recall mengukur seberapa baik model dapat menemukan kembali (me-recall) semua data ‘Positif’ yang ada. Dari semua data yang *seharusnya* ‘Positif’ (TP+FN), seberapa banyak yang berhasil ditemukan model.

$$Recall = \frac{TN}{TN + FP} = \frac{40}{105} = 0{,}38 \tag{6}$$

Terakhir, untuk menyeimbangkan Presisi dan Recall, digunakan F1-score. Metrik ini merupakan rata-rata harmonik (*harmonic mean*) dari *Precision* dan *Recall*. Metrik ini sangat berguna ketika terjadi ketidakseimbangan kelas (jumlah data positif dan negatif tidak sama), karena ia menyeimbangkan kedua metrik tersebut.

$$F1 - score = 2 \times \frac{Precission \times Recal}{Precission + Recall} = \frac{0.90909 \times 0.38095}{0.90909 + 0.38095} = \frac{0{,}6294}{1{,}29004} = 0{,}54 \tag{7}$$

Jika semua kalkulasi sudah diselesaikan, maka keseluruhannya dapat dilihat pada tabel 4.

Tabel 4. Hasil Evaluasi

	Precision	Recall	F1-score	Support
Negatif	0.91	0.38	0.54	105
Positif	0.93	1.0	0.96	854
Accuracy			0.93	959
Macro avg	0.92	0.69	0.75	959
Weighted avg	0.93	0.93	0.91	959

Akurasi keseluruhan 92.81% menunjukkan bahwa algoritma *Naïve Bayes* sangat efektif untuk tugas klasifikasi sentimen pada ulasan Duolingo. Performa model yang timpang antara kelas Positif dan Negatif adalah fenomena yang sangat umum dan dapat dijelaskan oleh ketidakseimbangan data. Data latih kita didominasi oleh 85.34% sentimen positif. Akibatnya, model menjadi "terlalu terlatih" untuk mengenali pola-pola positif dan memiliki kecenderungan untuk menebak 'Positif' ketika dihadapkan pada ulasan yang ambigu. Meskipun model ini *melewatkan* banyak ulasan negatif (low recall), model ini sangat akurat ketika *menemukan* ulasan negatif (high precision 0.91). Dalam konteks bisnis, model ini tetap sangat berguna. Model ini dapat digunakan sebagai sistem penyaring otomatis untuk mengekstrak keluhan pengguna: ulasan yang ditandai 'Negatif' oleh model dapat dipastikan adalah keluhan valid yang perlu segera ditindaklanjuti oleh tim pengembang Duolingo.

KESIMPULAN DAN SARAN

Kesimpulan

Berdasarkan analisis sentimen terhadap 5.000 ulasan pengguna Duolingo di Google Play Store dengan menggunakan algoritma Naïve Bayes, dapat ditarik beberapa kesimpulan. Pertama, sentimen pengguna secara keseluruhan terhadap aplikasi Duolingo adalah sangat positif, dengan 85.34% ulasan diklasifikasikan sebagai positif. Temuan ini diperkuat oleh analisis pada 1.938 ulasan yang secara spesifik membahas "efektivitas belajar", di mana sentimen positif juga tetap dominan. Analisis kualitatif menggunakan *Word Cloud* dan Bigram menunjukkan bahwa sentimen positif ini didorong oleh persepsi pengguna bahwa Duolingo adalah alat yang efektif, membantu, dan menyenangkan (dengan frasa kunci seperti "ajar bahasa", "mudah paham", dan "seru ajar"). Kedua, sentimen negatif, meskipun hanya mencakup 10.52% data, memiliki tema yang sangat spesifik. Keluhan utama pengguna tidak berfokus pada metode pembelajaran, melainkan pada model bisnis *freemium*, terutama pada batasan "sistem hati" atau "sistem energi" dan gangguan akibat "nonton iklan" yang dianggap menghambat proses belajar. Ketiga, model klasifikasi Multinomial Naïve Bayes yang dibangun untuk membedakan sentimen positif dan negatif terbukti sangat berhasil dengan mencapai tingkat akurasi keseluruhan 92.81%. Model ini menunjukkan kinerja luar biasa dalam mengenali ulasan positif dan memiliki presisi yang sangat tinggi saat mengidentifikasi ulasan negatif, meskipun memiliki keterbatasan dalam menjaring *semua* ulasan negatif akibat ketidakseimbangan data.

Saran

Menyadari bahwa penulis masih jauh dari kata sempurna, kedepannya penulis akan lebih fokus dan details dalam menjelaskan tentang jurnal di atas dengan sumber-sumber yang lebih banyak yang tentunya dapat di pertanggung jawabkan, Selain itu, mengingat aplikasi Duolingo digunakan secara global, pengembangan analisis sentimen dengan mempertimbangkan berbagai bahasa juga dapat menjadi arah penelitian berikutnya agar pemahaman terhadap persepsi pengguna di berbagai negara menjadi lebih menyeluruh.

 https://doi.org/10.56211/sudo.v5i1.1354

DAFTAR PUSTAKA

[1] N. C. Dewi, "Pembelajaran Daring Berbasis Konten pada Program Kursus Bahasa Asing," *IJALR*, vol. 3, no. 1, Mar. 2022, doi: 10.21009/ijalr.31.02.

[2] R. M. Simanjuntak, A. Sitorus, F. B. Manurung, and M. Kaban, "Sosialisasi Pengenalan Software Duolingo di SMAN 1 Pantai Cermin".

[3] M. R. P. Hardiyanto, G. Pahlevi, and M. F. Nugroho, "Pengaruh Fitur-Fitur Aplikasi Duolingo Terhadap Popularitasnya," *SNATI*, vol. 3, no. 1, Dec. 2023, doi: 10.20885/snati.v3i1.28.

[4] M. Apriliyani, M. I. Musyaffaq, S. Nur'Aini, M. R. Handayani, and K. Umam, "Implementasi analisis sentimen pada ulasan aplikasi Duolingo di Google Playstore menggunakan algoritma Naïve Bayes," *AITI*, vol. 21, no. 2, pp. 298–311, Sept. 2024, doi: 10.24246/aiti.v21i2.298-311.

[5] Ernianti Hasibuan and Elmo Allistair Heriyanto, "ANALISIS SENTIMEN PADA ULASAN APLIKASI AMAZON SHOPPING DI GOOGLE PLAY STORE MENGGUNAKAN NAIVE BAYES CLASSIFIER," *JTS*, vol. 1, no. 3, pp. 13–24, Oct. 2022, doi: 10.56127/jts.v1i3.434.

[6] A. Saninah, W. Prihartono, and C. L. Rohmat, "ANALISIS SENTIMEN ULASAN PENGGUNA TERHADAP APLIKASI PEMBELAJARAN BERBAHASA DUOLINGO MENGGUNAKAN ALGORITMA NAÏVE BAIYES CLASSIFIER," *JITET*, vol. 13, no. 1, Jan. 2025, doi: 10.23960/jitet.v13i1.5691.

[7] D. Normawati and S. A. Prayogi, "Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter," vol. 5, 2021.

[8] A. A. Munandar, F. Farikhin, and C. E. Widodo, "Sentimen Analisis Aplikasi Belajar Online Menggunakan Klasifikasi SVM," *JOINTECS*, vol. 8, no. 2, p. 77, July 2023, doi: 10.31328/jointecs.v8i2.4747.

[9] H. Tuhuteru, "Analisis Sentimen Masyarakat Terhadap Pembatasan Sosial Berksala Besar Menggunakan Algoritma Support Vector Machine," *INFORMATION SYSTEM DEVELOPMENT*, vol. 4.

[10] T. Safitri, Y. Umaidah, and I. Maulana, "Analisis Sentimen Pengguna Twitter Terhadap BTS Menggunakan Algoritma Support Vector Machine," vol. 7, no. 1.

[11] E. Febriyani and H. Februariyanti, "Analisis Sentimen Terhadap Program Kampus Merdeka Menggunakan Algoritma Naive Bayes Classifier Di Twitter," *JTK*, vol. 17, no. 1, p. 25, Feb. 2023, doi: 10.33365/jtk.v17i1.2061.

[12] A. F. Nurhaliza, B. D. Setiawan, and R. S. Perdana, "Penerapan Pemodelan Topik menggunakan Metode Latent Dirichlet Allocation terhadap Pembehasan Pemilu Indonesia tahun 2024 di Twitter".

[13] S. Rabbani, D. Safitri, N. Rahmadhani, A. A. F. Sani, and M. K. Anam, "Perbandingan Evaluasi Kernel SVM untuk Klasifikasi Sentimen dalam Analisis Kenaikan Harga BBM: Comparative Evaluation of SVM Kernels for Sentiment Classification in Fuel Price Increase Analysis," *MALCOM*, vol. 3, no. 2, pp. 153–160, Oct. 2023, doi: 10.57152/malcom.v3i2.897.

NOMENKLATUR

Simbol TF-IDF

W(t,d) : Bobot (nilai) TF-IDF dari kata t pada ulasan d
TF(t,d) : Term Frequency, frekuensi kemunculan kata t pada ulasan d
IDF(t) : Inverse Document Frequency, tingkat keunikan kata t dalam seluruh ulasan
N : Jumlah total ulasan (dokumen) dalam dataset
df_t : Document Frequency, jumlah ulasan yang mengandung kata t
t : Term atau kata
d : Document atau ulasan

Simbol Naïve Bayes

P(c|d) : Probabilitas Posterior (peluang ulasan d termasuk ke kelas c)
P(c) : Probabilitas Prior (peluang awal dari kelas c)
$P(t_i|c)$: Likelihood, probabilitas kata t_i muncul dalam kelas c
C : Class (kelas sentimen; Positif atau Negatif)
D : Dokumen atau ulasan yang dianalisis

Metrik Evaluasi & Akronim

TP : True Positive — data positif yang diprediksi positif
TN : True Negative — data negatif yang diprediksi negatif
FP : False Positive — data negatif yang diprediksi positif
FN : False Negative — data positif yang diprediksi negatif
NLP : Natural Language Processing (Pemrosesan Bahasa Alami)
EDA : Exploratory Data Analysis (Analisis Eksploratif Data)
NB : Naïve Bayes, algoritma klasifikasi probabilistik

 https://doi.org/10.56211/sudo.v5i1.1354