Abstract

Di era digital, platform media sosial telah muncul sebagai saluran penting bagi individu untuk mengekspresikan pendapat dan berbagi pengalaman, berfungsi sebagai sumber data yang kaya untuk analisis sentimen. Studi ini berfokus pada penerapan Klasifikasi Naive Bayes untuk menganalisis ulasan pengguna di platform media sosial Red Note. Dengan memanfaatkan web scraping, dataset yang terdiri dari 17.727 ulasan dikumpulkan, yang kemudian menjalani proses pra-pemrosesan yang ketat, termasuk pembersihan teks, penghapusan kata-kata umum, dan tokenisasi. Ekstraksi fitur dilakukan menggunakan metode TF-IDF dan CountVectorizer, dengan kinerja model dievaluasi berdasarkan presisi, recall, F1-score, dan akurasi. Hasil menunjukkan bahwa metode TF-IDF mencapai presisi sebesar 0,88, sementara CountVectorizer unggul dalam recall (0,84) dan F1-score (0,85), menunjukkan efektivitasnya dalam mengidentifikasi kasus positif yang relevan. Kedua metode menunjukkan akurasi yang memuaskan, dengan CountVectorizer sedikit mengungguli TF-IDF dengan nilai 0,90. Temuan ini menekankan efektivitas Klasifikasi Naive Bayes dalam analisis sentimen dan memberikan wawasan berharga tentang persepsi pengguna di berbagai topik di platform tersebut. Penelitian ini menyoroti potensi untuk mengembangkan metode analisis sentimen yang lebih efisien di masa depan, berkontribusi pada bidang pengambilan keputusan berbasis data yang lebih luas dalam konteks media sosial.

1. Pendahuluan

Dalam era digital yang semakin maju, media sosial telah menjadi salah satu platform utama bagi individu untuk mengekspresikan pendapat dan berbagi pengalaman. Platform-platform ini tidak hanya berfungsi sebagai sarana komunikasi, tetapi juga sebagai sumber data yang kaya untuk analisis sentimen. Analisis sentimen, yang merupakan proses mengidentifikasi dan mengkategorikan opini yang dinyatakan dalam teks, telah menjadi alat penting dalam memahami persepsi publik terhadap berbagai isu, produk, dan layanan

Secara global, analisis sentimen telah digunakan oleh perusahaan dan organisasi untuk mendapatkan wawasan berharga tentang preferensi konsumen dan tren pasar. Misalnya, perusahaan dapat memanfaatkan analisis sentimen untuk mengukur kepuasan pelanggan dan mengidentifikasi area yang memerlukan perbaikan. Selain itu, dalam konteks politik, analisis sentimen dapat digunakan untuk memantau opini publik terhadap kebijakan pemerintah atau kandidat politik.

Di tingkat lokal, platform media sosial seperti Red Note telah menjadi tempat yang populer bagi pengguna untuk memberikan ulasan dan berbagi pengalaman mereka. Ulasan ini dapat mencakup berbagai topik, mulai dari produk dan layanan hingga isu-isu sosial dan politik. Namun, dengan volume data yang sangat besar, analisis manual terhadap ulasan-ulasan ini menjadi tidak praktis. Oleh karena itu, diperlukan pendekatan otomatis yang efektif untuk menganalisis sentimen dari ulasan-ulasan tersebut.

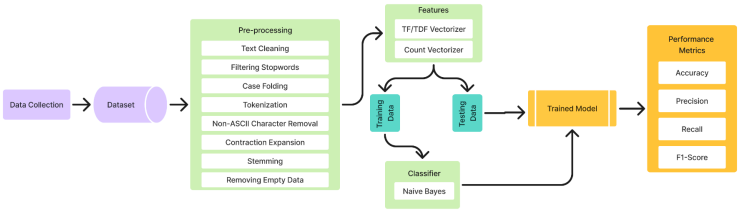
Algoritma Naive Bayes Classifier merupakan salah satu metode yang banyak digunakan dalam analisis sentimen karena kesederhanaan dan efisiensinya. Algoritma ini bekerja dengan mengklasifikasikan teks berdasarkan probabilitas kemunculan kata-kata tertentu dalam kategori sentimen yang telah ditentukan sebelumnya. Dalam konteks

ulasan platform media sosial Red Note, penggunaan Naive Bayes Classifier dapat membantu dalam mengidentifikasi sentimen positif, negatif, atau netral dari ulasan pengguna secara otomatis.

Penelitian ini bertujuan untuk menerapkan algoritma Naive Bayes Classifier dalam analisis sentimen ulasan pada platform media sosial Red Note. Dengan demikian, diharapkan dapat memberikan wawasan yang lebih mendalam tentang persepsi pengguna terhadap berbagai topik yang dibahas di platform tersebut, serta memberikan kontribusi dalam pengembangan metode analisis sentimen yang lebih efektif dan efisien.

2. Metode

Desain penelitian ini menggambarkan proses pengolahan data dan pelatihan model klasifikasi dengan metode Naive Bayes. Dimulai dari pengumpulan data untuk membentuk dataset, diikuti oleh preprocessing yang mencakup pembersihan teks, penghapusan stopwords, tokenization, dan lainnya. Setelah itu, fitur dihitung menggunakan TF/TDF dan dibagi menjadi data pelatihan dan pengujian. Model dilatih dengan data pelatihan dan dievaluasi menggunakan metrik seperti akurasi, presisi, recall, dan F1-score untuk menilai kinerjanya. Gambar dapat dilihat pada Gambar 1.



Gambar 1. Desain Penelitian

a. Pengumpulan Data

Pengambilan data dalam penelitian ini dilakukan menggunakan metode web scraping melalui platform Play Store. Metode ini memungkinkan peneliti untuk secara otomatis mengumpulkan informasi dari ulasan pengguna yang tersedia di aplikasi. Dari proses pengambilan data tersebut, berhasil dikumpulkan sebanyak 17.727 ulasan, yang semuanya ditulis dalam bahasa Inggris. Data yang diperoleh ini akan digunakan untuk analisis sentimen, dengan fokus pada pemahaman persepsi pengguna terhadap aplikasi yang diteliti.

b. Pra-preprocessing

Tahapan pertama dalam pre-processing adalah pembersihan teks, yang bertujuan untuk menghapus elemen yang tidak diperlukan, seperti karakter

khusus, angka, dan tanda baca, sehingga hanya informasi relevan yang tersisa. Setelah pembersihan, langkah berikutnya adalah penghapusan stopwords dan tokenization. Penghapusan stopwords melibatkan penghilangan kata-kata umum yang tidak memberikan makna signifikan, seperti "dan" dan "atau". Tokenization memecah teks menjadi unit-unit kecil, seperti kata atau frasa, yang memudahkan analisis dan membantu dalam mengidentifikasi kata kunci penting dalam dataset.

Tahapan terakhir mencakup normalisasi teks, yang meliputi case folding, penghapusan karakter non-ASCII, ekspansi kontraksi, stemming, dan penghapusan data kosong. Proses ini memastikan konsistensi dan relevansi data, dengan case folding mengubah huruf menjadi kecil dan penghapusan karakter non-ASCII menghilangkan karakter yang tidak standar. Setelah melalui semua tahapan pre-processing, jumlah data yang tersisa adalah 15.314 entri. Dari jumlah tersebut, analisis lebih lanjut menunjukkan bahwa terdapat 11.743 entri yang dikategorikan sebagai positif dan 3.571 entri sebagai negatif. Pembagian ini memberikan wawasan penting mengenai persepsi pengguna, yang akan meningkatkan akurasi dan efektivitas model klasifikasi yang akan diterapkan.

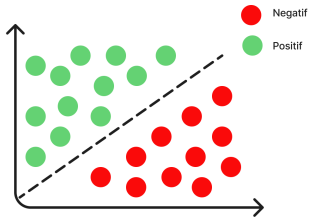
c. Feature Extraction

Tahapan ekstraksi fitur dalam proses ini bertujuan untuk mengubah data teks yang telah diproses menjadi representasi numerik yang dapat digunakan oleh model klasifikasi. Dua metode utama yang digunakan adalah TF/TDF Vectorizer dan Count Vectorizer. TF/TDF Vectorizer menghitung frekuensi kata dengan mempertimbangkan seberapa umum atau jarangnya kata tersebut di seluruh dataset, sedangkan Count Vectorizer menghitung jumlah kemunculan setiap kata dalam dokumen. Fitur yang dihasilkan dari proses ini kemudian digunakan untuk melatih model klasifikasi, seperti Naive Bayes, dengan membagi dataset menjadi data pelatihan dan pengujian untuk evaluasi kinerja model.

d. Classification

Dalam penelitian ini, digunakan algoritma Naive Bayes, yang merupakan metode klasifikasi berbasis probabilistik berdasarkan teorema Bayes dengan asumsi independensi antar fitur. Algoritma ini terkenal karena kesederhanaan dan efisiensinya dalam memproses data, khususnya dalam analisis teks dan sentimen.

Keunggulan Naive Bayes terletak pada struktur yang sederhana, yang memudahkan pemahaman dan implementasi. Algoritma ini juga sangat efisien dalam waktu komputasi, baik saat pelatihan maupun klasifikasi, menjadikannya pilihan ideal untuk dataset besar. Kinerjanya yang baik pada data teks, termasuk analisis sentimen, memungkinkan algoritma ini untuk menangkap pola meskipun asumsi independensi tidak selalu valid. Selain itu, Naive Bayes dapat berfungsi dengan baik meskipun terdapat ketidakseimbangan dalam jumlah data antar kelas. Arsitektur dari Naive Bayes dapat dilihat pada gambar 2.



Gambar 2. Arsitektur Naive Bayes

e. Evaluation

Tahapan evaluasi bertujuan untuk mengukur kinerja model klasifikasi, seperti Naive Bayes, setelah dilatih dengan data pelatihan. Model diuji menggunakan data pengujian untuk menilai kemampuannya dalam mengklasifikasikan data baru. Kinerja dievaluasi melalui metrik seperti akurasi (proporsi prediksi yang benar), presisi (ketepatan prediksi positif), recall (kemampuan menemukan semua contoh positif), dan F1-score (keseimbangan antara presisi dan recall). Metrik-metrik ini membantu menilai efektivitas model dan mengidentifikasi area perbaikan. Rumusnya dapat dilihat pada persamaan

Rumus

3. Hasil Dan Diskusi

Hasil evaluasi divisualisasikan dalam bentuk tabel yang ditampilkan pada Tabel 1.

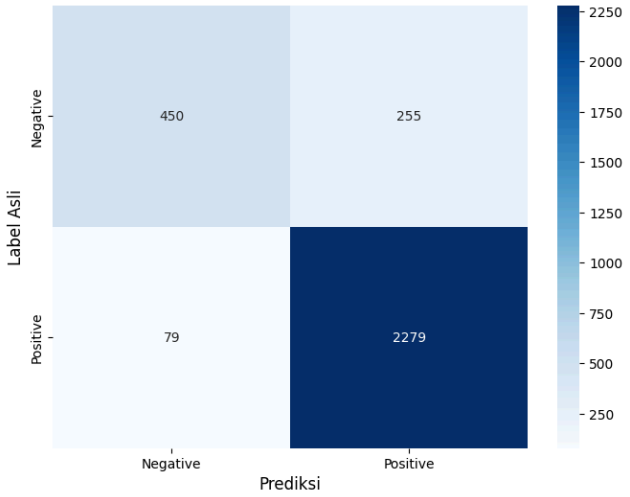
Tabel 1. Hasil Evaluasi

Metrics	TF-IDF	CountVectorizer
Precision	0.88	0.86
Recall	0.80	0.84
F1-Score	0.83	0.85
Accuracy	0.89	0.90

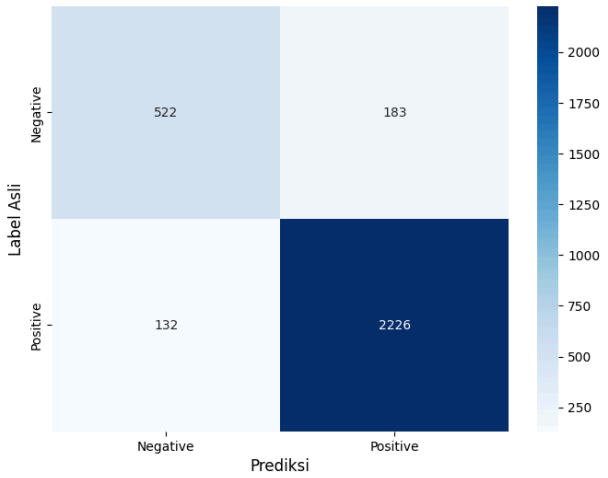
Tabel 1 menyajikan hasil evaluasi kinerja model klasifikasi yang menggunakan dua metode ekstraksi fitur, yaitu TF-IDF dan CountVectorizer. Metrik yang digunakan untuk menilai kinerja model meliputi precision, recall, F1-Score, dan accuracy.

Dari hasil yang ditampilkan, model yang menggunakan TF-IDF menunjukkan nilai precision sebesar 0.88, yang mengindikasikan bahwa 88% dari prediksi positif yang dihasilkan adalah benar. Sebaliknya, CountVectorizer mencatat nilai precision sebesar 0.86, yang menunjukkan sedikit penurunan dalam ketepatan prediksi positif. Meskipun demikian, dalam hal recall, CountVectorizer unggul dengan nilai 0.84 dibandingkan dengan TF-IDF yang hanya mencapai 0.80. Hal ini menunjukkan bahwa CountVectorizer lebih efektif dalam mengidentifikasi semua contoh positif dalam dataset, sehingga lebih mampu menangkap kasus-kasus yang relevan.

Metrik F1-Score, yang merupakan ukuran keseimbangan antara precision dan recall, menunjukkan nilai 0.83 untuk TF-IDF dan 0.85 untuk CountVectorizer. Ini menegaskan bahwa CountVectorizer memberikan keseimbangan yang lebih baik antara kedua metrik tersebut. Terakhir, dalam hal accuracy, kedua metode menunjukkan hasil yang baik, dengan TF-IDF mencapai 0.89 dan CountVectorizer 0.90. Ini menunjukkan bahwa kedua model secara keseluruhan mampu mengklasifikasikan data dengan baik, meskipun CountVectorizer sedikit lebih unggul dalam akurasi. Hasil ini memberikan wawasan penting mengenai efektivitas masing-masing metode dalam konteks klasifikasi teks.



Gambar 3. matriks konfusi dengan menggunakan TF-IDF



Gambar 4. matriks konfusi dengan menggunakan CountVectorizer

4. Kesimpulan

Hasil evaluasi menunjukkan bahwa model yang menggunakan TF-IDF memiliki precision yang lebih tinggi (0.88) dibandingkan CountVectorizer (0.86), namun CountVectorizer unggul dalam recall (0.84) dan F1-Score (0.85). Kedua metode menunjukkan akurasi yang baik, dengan CountVectorizer mencapai 0.90. Temuan ini menegaskan efektivitas Naive Bayes dalam analisis sentimen dan memberikan wawasan berharga tentang persepsi pengguna terhadap berbagai topik di platform tersebut. Penelitian ini juga menyarankan potensi pengembangan metode analisis sentimen yang lebih efisien di masa depan.