# Analisis Dan Klasifikasi Sentimen Terhadap *Brand* Infinix Tecno dan Itel Menggunakan Kombinasi Metode *Naive Bayes* Dan *Cosine Similarity*

Rakka Yusuf Hidayat[1]
[1]Teknik Informatika, [2]Fakultas Sains dan Teknologi, [3]Universitas Muhammadiyah Sukabumi
Email : rakkayusufhidayat@gmail.com[1]

**Abstrak**

Di era digital yang serba cepat ini, smartphone telah menjadi bagian tak terpisahkan dari kehidupan sehari-hari. Pasar smartphone yang kompetitif mendorong berbagai brand, termasuk Infinix, Tecno, dan Itel, untuk menawarkan produk dengan harga terjangkau namun berkualitas tinggi. Untuk memahami persepsi dan pengalaman konsumen terhadap brand ini, penelitian ini melakukan analisis sentimen menggunakan metode Naive Bayes pada komentar di media sosial YouTube. Data komentar dikumpulkan dari video YouTube yang berkaitan dengan Infinix, Tecno, dan Itel, lalu dipreproses dan dianalisis.Hasil penelitian menunjukkan bahwa model Naive Bayes memiliki performa yang baik, dengan akurasi rata-rata sebesar 78% untuk Infinix, 72% untuk Tecno, dan 77% untuk Itel. Evaluasi dengan metrik precision dan recall memberikan hasil yang beragam. Precision tertinggi dicapai pada label negatif untuk dataset Infinix, yaitu 100%, sementara recall tertinggi pada label positif untuk dataset yang sama mencapai 100%. Proses evaluasi menggunakan k-fold cross-validation menunjukkan konsistensi hasil model pada setiap iterasi. Secara keseluruhan, mayoritas sentimen konsumen terhadap Infinix adalah positif, sementara Tecno dan Itel memiliki distribusi sentimen yang lebih beragam.

**Kata kunci**: Analisis Sentimen, *Naive Bayes*, Media Sosial, YouTube, *Smartphone*

## PENDAHULUAN

Di pasar *smartphone* yang kompetitif, berbagai *brand* berlomba-lomba menawarkan produk dengan fitur dan spesifikasi unggulan untuk menarik minat konsumen. Di antara banyak *brand* yang beredar, Infinix, Tecno, dan Itel merupakan beberapa nama yang cukup populer, terutama di pasar negara berkembang. *Brand* ini dikenal karena menawarkan *smartphone* dengan harga terjangkau namun memiliki kualitas yang tidak kalah dengan *brand* ternama lainnya. Infinix, Tecno, dan Itel telah berhasil menarik perhatian konsumen dengan strategi pemasaran yang efektif dan produk yang inovatif. Dengan meningkatnya jumlah pengguna *smartphone* dari *brand* ini, penting untuk memahami bagaimana persepsi dan pengalaman konsumen terhadap produk-produk tersebut.

Di era digital ini, media sosial telah menjadi salah satu platform utama bagi konsumen untuk berbagi pengalaman dan pendapat mereka mengenai berbagai produk, termasuk *smartphone*. Salah satu platform media sosial yang sangat populer untuk tujuan ini adalah YouTube. Dengan jutaan pengguna aktif setiap harinya, YouTube menyediakan wadah bagi konsumen untuk menonton review, tutorial, *unboxing*, dan berbagai konten lainnya yang berkaitan dengan *smartphone*. Konsumen seringkali mencari informasi di YouTube sebelum memutuskan untuk membeli sebuah produk, menjadikannya sumber informasi yang sangat berpengaruh dalam proses pengambilan keputusan. Oleh karena itu, analisis terhadap konten YouTube dapat memberikan wawasan berharga mengenai persepsi konsumen terhadap *brand smartphone* tertentu.

Memahami persepsi konsumen adalah aspek penting dalam strategi bisnis, karena persepsi tersebut dapat mempengaruhi

keputusan pembelian dan loyalitas *brand*. Salah satu cara untuk mengukur dan memahami persepsi konsumen adalah melalui analisis sentimen. Analisis sentimen adalah proses untuk mengidentifikasi dan mengategorikan opini yang terkandung dalam teks, seperti ulasan atau komentar, untuk menentukan sikap penulis terhadap topik tertentu (Fitrana et al., 2024) . Dengan analisis sentimen, perusahaan dapat memahami apakah konsumen memiliki pandangan positif, negatif, atau netral terhadap produk mereka. Informasi ini sangat berharga untuk mengidentifikasi kekuatan dan kelemahan produk, serta untuk merumuskan strategi pemasaran yang lebih efektif.

Untuk melakukan analisis sentimen, terdapat berbagai metode yang dapat digunakan, salah satunya adalah metode Naive Bayes. Naive Bayes adalah algoritma klasifikasi yang didasarkan pada Teorema Bayes dengan asumsi independensi antara fitur-fitur (Ananda & Suryono, 2024). Meskipun asumsi independensi ini seringkali tidak sepenuhnya benar dalam konteks nyata, metode Naive Bayes terbukti efektif dan efisien dalam berbagai aplikasi, termasuk analisis sentimen. Metode ini relatif mudah diimplementasikan dan memiliki kinerja yang baik meskipun dengan data pelatihan yang terbatas. Selain itu, Naive Bayes mampu menangani data teks dengan baik, menjadikannya pilihan yang tepat untuk analisis sentimen terhadap ulasan-ulasan konsumen di media sosial seperti YouTube.

Metode Naive Bayes bekerja dengan menghitung probabilitas bahwa sebuah teks termasuk dalam kategori tertentu (misalnya, positif, negatif, atau netral) berdasarkan distribusi kata-kata dalam teks tersebut (Hasri & Alita, 2022). Dalam konteks analisis sentimen, kata-kata yang sering muncul dalam ulasan positif akan memiliki probabilitas tinggi untuk diklasifikasikan sebagai ulasan positif, dan sebaliknya untuk ulasan negatif. Proses ini melibatkan beberapa langkah, mulai dari preprocessing teks (seperti tokenisasi dan penghilangan kata-kata umum), perhitungan probabilitas kata, hingga klasifikasi akhir berdasarkan probabilitas gabungan. Keandalan dan kesederhanaan metode ini membuatnya sangat berguna dalam analisis sentimen skala besar.

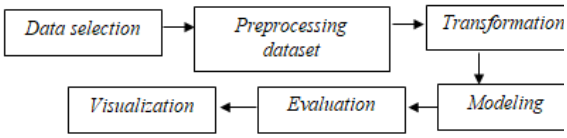
Berikut adalah beberapa penelitian terdahulu mengenai analisis sentimen komentar di *youtube* tentang islamofobia menggunakan algoritma *Random Forest* dengan nilai akurasi 79% (Afdhal et al., 2022). Penelitian lainnya yaitu analisis sentimen ulasan aplikasi "MyPertamina" pada *google play* store yang menghasilkan nilai akurasi 87% (Maulana et al., 2023). Dan penelitian lainnya menggunakan metode K-Nearest Neighbor dengan judul "analisis sentimen pada teks opini penilaian kinerja dosen menggunakan menghasilkan hasil nilai akurasi 80% (Permana & Makmun, 2020). Adapun penelitian terkait sentimen *brand smartphone* dengan judul "Klasifikasi Penjualan Produk Menggunakan Algoritma Naive Bayes pada Konter HP Bayu" yang menghasilkan nilai 0.005211 pada salah satu *brand smartphone* (Ayu et al., 2024).

## METODE PENELITIAN

Dalam metode penelitian ini, metode *Knowledge Discovery in Database* (KDD) digunakan peneliti untuk mengaplikasikan algoritma *Naïve Bayes* untuk mengolah dan menganalisis data komentar. *Knowledge Discovery in Database* adalah Metode di mana semua pengetahuan data terhubung ke organisasi atau lembaga yang memiliki sejumlah besar data dan berfokus pada metodologi yang luas digabungkan. (Buani, 2024)

Metode ini digunakan dengan tujuan agar peneliti dapat mengekstraksi informasi yang berharga disembunyikan dalam kumpulan data ulasan sehingga dapat

ditemukan pengetahuan yang sebelumnya tidak diketahui, berguna dan potensial untuk pengambilan keputusan di masa depan. Tahapan metode penelitian ini adalah sebagai berikut.




**Gambar 1** Tahapan Analisis Sentimen pada penelitian

**1. *Data Selection***

*Data Selection* merupakan tahap awal penelitian ini yang akan menyeleksi dan mengambil ulasan dari komentar pengguna media sosial youtube terhadap *brand smartphone* Infinix, Tecno, dan Itel. Proses pengumpulannya menggunakan teknik *crawling* data melalui Youtube API menggunakan Google Colaboratory. Data yang terkumpul terhadap *brand smartphone* Infinix sebanyak 1000 komentar, Tecno 1000 komentar dan Itel 1000 komentar sejak bulan Maret 2024 sampai dengan September 2024 yang selanjutnya akan diproses kembali sesuai tahap analisis sentimen.

**2. *Pre-processing Dataset***

Tahap *pre-processing dataset* merupakan tahap pengolahan *dataset* yang semula berupa data mentah atau data kotor yang akan diolah menjadi data sesuai kebutuhan untuk analisis selanjutnya karena dalam *preprocessing* ini banyak tahapan yang akan dilakukan untuk mengolah data tersebut, yaitu *cleaning, casefolding, tokenizing, stopword removal dan stemming.*

**3. *Transformation***

Tahap *transformation* (transformasi) merupakan bagian dari proses KDD yang bertujuan untuk mengubah atau memodifikasi data ulasan agar lebih mudah diolah dan dianalisis. Pada Tahap ini biasanya dilakukan setelah tahapan *preprocessing* dataset dan dapat mencakup berbagai macam proses, seperti pembobotan kata menggunakan *Term Frequency Invers Document Frequency* (TF-IDF). Dengan melakukan transformasi, data komentar akan lebih mudah diolah dan dianalisis sehingga hasil analisis lebih akurat dan berguna.

**4. *Modeling***

Tahap pemodelan (*modeling*) merupakan tahapan dalam proses *Knowledge Discovery in Database* (KDD) yang bertujuan untuk membangun model yang mampu menerapkan pola atau hubungan meninjau data. Pada langkah ini, data *training* yang telah diproses dan transformasi sebelumnya akan digunakan untuk membangun model. Pada penelitian ini algoritma *naïve bayes* akan digunakan untuk melakukan proses klasifikasi pada dataset menjadi data latih (*training*) dan data pengujian (*testing*).Dalam prosesnya, dataset yang diambil dan dikumpulkan akan dibagi menjadi dua, yaitu data latih dan data uji. Data latih tersebut akan digunakan dalam pelatihan algoritma klasifikasi, dan sebelumnya dataset harus sudah diberi label dan harus melalui proses *text preprocessing* dan pembobotan kata menggunakan TF-IDF. Data pengujian akan digunakan untuk menguji algoritma dan model yang telah dilatih sehingga akan memberikan hasil yang baik dalam proses klasifikasi untuk data yang mencakup data positif maupun data negatif.

Kombinasi metode *Naive Bayes* dan *Cosine Similarity* dalam analisis sentimen melibatkan beberapa tahapan, yaitu :

1) *Preprocessing data*

   Tahap ini dilakukan dengan membersihkan komentar agar data komentar siap sesuai dengan kebutuhan analisis, lalu komentar diubah menjadi vektor fitur menggunakan TF-IDF.

2) *Naive Bayes Classifier*

   Metode ini digunakan untuk mengklasifikasikan komentar menjadi sentimen positif atau negatif.

3) *Cosine Similarity*

Setelah sentimen telah ditentukan, selanjutnya metode *cosine similarity* digunakan untuk mengelompokkan komentar ke dalam kategori tertentu seperti kinerja, pengalaman pengguna, atau perbandingan merek berdasarkan tingkat kesamaannya dengan masing-masing kategori.

Kombinasi kedua metode ini memberikan keunggulan berupa klasifikasi sentimen yang akurat melalui pendekatan probabilistik *Naive Bayes* dan penambahan kemampuan kategorisasi berbasis konteks dengan *Cosine Similarity*, serta proses yang dapat diotomatisasi untuk menghasilkan analisis sentimen yang lebih mendalam, seperti yang diterapkan dalam penelitian ini pada komentar YouTube untuk *brand smartphone* Infinix, Tecno, dan Itel.

**5. *Evaluation***

Tahapan evaluasi dalam *Knowledge Discovery in Database* (KDD) adalah tahap penilaian hasil model yang telah dibangun pada tahap pemodelan. Pada langkah ini model telah dibangun dan akan diterapkan pada data pengujian yang telah disiapkan sebelumnya. Hasil penerapan model akan dianalisis dan dibandingkan dengan hasil yang di harapkan, sehingga dapat diketahui tingkat keakuratan model. Ada beberapa metrik yang dapat digunakan dalam evaluasi, seperti *accuracy, precision, recall,* dan *f1-score*. Selain itu, *confusion matrix* juga dapat digunakan untuk mengetahui seberapa baik model dapat memprediksi hasil sebenarnya. Tahap evaluasi ini penting untuk dilakukan agar mengetahui sejauh mana model yang dibangun dapat menangkap pola atau hubungan yang terdapat pada data komentar.

**6. *Visualization***

Tahapan visualisasi pada *Knowledge Discovery in Database* (KDD) merupakan tahap akhir dari proses KDD yang bertujuan untuk menampilkan hasil model yang telah dibangun pada tahap sebelumnya dalam bentuk visual yang mudah dipahami orang lain.

## HASIL DAN PEMBAHASAN

**1. *Data Collection***

Data yang dipakai yaitu berupa komentar berbahasa Indonesia mengenai ulasan terhadap *brand smartphone* Infinix, Tecno, dan Itel yang diperoleh dari proses *crawling* menggunakan Bahasa pemrograman *python*. Sumber data diambil pada komentar untuk setiap *brand* pada video youtube yang membahas terkair 3 *brand* tersebut dengan rentang periode waktu Maret – September 2023 Peneliti memilih *brand smartphone* Infinix, Tecno dan Itel karena *brand* tersebut sedang naik di pasar *smartphone* di Indonesia.

Dari proses *crawling* yang diambil dalam kurun 2 bulan yang terhitung dari bulan Maret sampai dengan September 2024, menghasilkan 3 dataset yang berbeda. Jumlah keseluruhan dari ketiga dataset tersebut adalah sebanyak 3000 dengan masing-masing dataset terdiri dari 1000 data. Dengan rinciannya dapat dilihat pada tabel berikut.

**Tabel 1** Jumlah Data yang Digunakan

| Dataset | Jumlah Data |
|---|---|
| Infinix | 1000 |
| Tecno | 1000 |
| Itel | 1000 |
| **Jumlah keseluruhan data yang digunakan** | **3000** |

**2. *Data Labelling***

Hasil data *crawling* yang disimpan dalam bentuk csv, kemudian dibagi menjadi 2 bagian yaitu *data training* dan *data testing*. Keseluruhan data yang digunakan adalah *data training* 80% dan untuk *data testing* sebanyak 20%. *Data training* untuk masing-masing dataset akan diberi 2 kategori label yaitu positif dan negatif. Berikut adalah tabel untuk setiap data yang sudah diberi label

positif dan negatif untuk masing-masing dataset.

**Tabel 2** Pelabelan Data

| Dataset | Jumlah Data | |
|---|---|---|
| | Positif | Negatif |
| Infinix | 685 | 315 |
| Tecno | 668 | 332 |
| Itel | 570 | 430 |
| **Jumlah keseluruhan data yang digunakan** | **1923** | **1077** |

**3. *Pre-processing***

**1) *Data Cleaning***

Tahapan awal dari *pre-processing* ini diawali dengan melakukan peninjuan terhadap data untuk melihat apakah terdapat data duplikat, *missing values*, data tidak valid atau terdapat *noise* dalam data. Karena tidak semua atribut dalam dataset akan digunakan, sehingga atribut yang tidak yang tidak mempengaruhi hasil sentimen akan dihilangkan.

**Tabel 3** Tahap *Cleaning*

| Sebelum *Cleaning* | Sesudah *Cleaning* |
|---|---|
| infinix sih speknya gg bgt, mantap pokoknya! | infinix sih speknya gg bgt mantap pokoknya |
| kalo itel mah biasa aja, ga worth it bro. | kalo itel mah biasa aja ga worth it bro |
| tecno sih lumayan lah buat budget low, tp oke kok! | tecno sih lumayan lah buat budget low tp oke kok |
| hape infinix gw gahar parah, puas bgt make nya | hape infinix gw gahar parah puas bgt make nya |
| serius tecno murah tp spek nya ngalahin hape lain! | serius tecno murah tp spek nya ngalahin hape lain |
| susah banget service infinix, ga direkomendasikan. | susah banget service infinix ga direkomendasikan |

**2) *Case Folding***

Tahap ini bertujuan untuk merubah huruf kapital pada teks komentar menjadi huruf kecil.

**Tabel 4** Tahap *Case Folding*

| Sebelum *Case Folding* | Sesudah *Case Folding* |
|---|---|
| Infinix sih | infinix sih |
| speknya GG bgt | speknya gg bgt |
| mantap pokoknya | mantap pokoknya |
| Kalo itel mah | kalo itel mah |
| biasa aja ga | biasa aja ga worth |
| WORTH IT | it bro |

**3) *Tokenizing***

*Tokenizing* bertujuan untuk membagi teks dengan cara memotong kata yang dipisahkan spasi menjadi potongan tunggal. Proses ini menggunakan *library Natural Language Toolkit (*NLTK) pada bahasa pemrograman *python*.

**Tabel 5** Tahap *Tokenizing*

| Sebelum *Tokenizing* | Sesudah *Tokenizing* |
|---|---|
| infinix sih speknya gg bgt mantap pokoknya | ['infinix', 'sih', 'speknya', 'gg', 'bgt', 'mantap', 'pokoknya'] |
| kalo itel mah biasa aja ga worth it bro | ['kalo', 'itel', 'mah', 'biasa', 'aja', 'ga', 'worth', 'it', 'bro'] |
| tecno sih lumayan lah buat budget low tp oke kok | ['tecno', 'sih', 'lumayan', 'lah', 'buat', 'budget', 'low', 'tp', 'oke', 'kok'] |
| hape infinix gw gahar parah puas bgt make nya | ['hape', 'infinix', 'gw', 'gahar', 'parah', 'puas', 'bgt', 'make', 'nya'] |
| serius tecno murah tp spek nya ngalahin hape lain | ['serius', 'tecno', 'murah', 'tp', 'spek', 'nya', 'ngalahin', 'hape', 'lain'] |
| susah banget service infinix ga direkomendasikan | ['susah', 'banget', 'service', 'infinix', 'ga', 'direkomendasikan'] |

**4) *Stopword Removal***

Tahap *Stopword* bertujuan untuk menghilangkan kata yang tidak bermakna yang tidak memiliki pengaruh terhadap akurasi dalam proses klasifikasi. Berikut adalah tabel hasil *stopword removal*.

**Tabel 6** Tahap *Stopword Removal*

| **Sebelum *Stopword*** | **Sesudah *Stopword*** |
|---|---|
| ['infinix', 'sih', 'speknya', 'gg', 'bgt', 'mantap', 'pokoknya'] | ['infinix', 'speknya', 'gg', 'bgt', 'mantap', 'pokoknya'] |
| ['kalo', 'itel', 'mah', 'biasa', 'aja', 'ga', 'worth', 'it', 'bro'] | ['itel', 'biasa', 'ga', 'worth', 'bro'] |
| ['tecno', 'sih', 'lumayan', 'lah', 'buat', 'budget', 'low', 'tp', 'oke', 'kok'] | ['tecno', 'lumayan', 'budget', 'low', 'oke'] |
| ['hape', 'infinix', 'gw', 'gahar', 'parah', 'puas', 'bgt', 'make', 'nya'] | ['hape', 'infinix', 'gahar', 'parah', 'puas', 'bgt', 'make'] |
| ['serius', 'tecno', 'murah', 'tp', 'spek', 'nya', 'ngalahin', 'hape', 'lain'] | ['serius', 'tecno', 'murah', 'spek', 'ngalahin', 'hape', 'lain'] |
| ['susah', 'banget', 'service', 'infinix', 'ga', 'direkomendasikan'] | ['susah', 'banget', 'service', 'infinix', 'ga', 'direkomendasikan'] |

**5) *Slangword***

Proses ini bertujuan untuk mengganti kata *slang* atau kata yang dugunakan dalam kehidupan sehari-hari menjadi bentuk yang sebenarnya. Berikut adalah tabel hasil proses *slangword*.

**Tabel 7** Contoh *Slangword*

| **Sebelum *Slangword*** | **Setelah *Slangword*** |
|---|---|
| ['infinix', 'speknya', 'gg', 'bgt', 'mantap', 'pokoknya'] | ['infinix', 'speknya', 'gahar', 'banget', 'mantap', 'pokoknya'] |
| ['itel', 'biasa', 'ga', 'worth', 'bro'] | ['itel', 'biasa', 'tidak', 'worth', 'teman'] |
| ['tecno', 'lumayan', 'budget', 'low', 'oke'] | ['tecno', 'lumayan', 'budget', 'rendah', 'oke'] |
| ['hape', 'infinix', 'gahar', 'parah', 'puas', 'bgt', 'make'] | ['hape', 'infinix', 'gahar', 'luar biasa', 'puas', 'banget', 'make'] |
| ['serius', 'tecno', 'murah', 'spek', 'ngalahin', 'hape', 'lain'] | ['serius', 'tecno', 'murah', 'spek', 'mengalahkan', 'hape', 'lain'] |
| ['susah', 'banget', 'service', 'infinix', 'ga', 'direkomendasikan'] | ['susah', 'banget', 'service', 'infinix', 'tidak', 'direkomendasikan'] |

**6) *Steeming***

*Steeming* bertujuan untuk menghilangkan infleksi kata ke bentuk dasarnya. Proses *steeming* Bahasa Indonesia dilakukan dengan cara menghilangkan sufiks (imbuhan akhir) dan prefix (imbuhan awal) menggunakan *library* pada bahasa pemrograman *phyton*.

Berikut adalah tabel yang menampilkan data yang telah melalui tahap *stemming*.

**Tabel 8** Tahap *Steeming*

| **Sebelum *Steeming*** | **Sesudah *Steeming*** |
|---|---|
| ['infinix', 'speknya', 'gahar', 'banget', 'mantap', 'pokoknya'] | ['infinix', 'spek', 'gahar', 'banget', 'mantap', 'pokok'] |
| ['itel', 'biasa', 'tidak', 'worth', 'teman'] | ['itel', 'biasa', 'tidak', 'worth', 'teman'] |
| ['tecno', 'lumayan', 'budget', 'rendah', 'oke'] | ['tecno', 'lumayan', 'budget', 'rendah', 'oke'] |
| ['hape', 'infinix', 'gahar', 'luar biasa', 'puas', 'banget', 'make'] | ['hape', 'infinix', 'gahar', 'luar', 'puas', 'banget', 'make'] |
| ['serius', 'tecno', 'murah', 'spek', | ['serius', 'tecno', 'murah', 'spek', |

| | |
|---|---|
| 'mengalahkan', 'hape', 'lain'] | 'kalah', 'hape', 'lain'] |
| ['susah', 'banget', 'service', 'infinix', 'tidak', 'direkomendasikan'] | ['susah', 'banget', 'service', 'infinix', 'tidak', 'rekomendasi'] |

**4. TF-IDF**

Setelah tahap *pre-processing* makan dilanjutkan dengan menghitung bobot suatu kata dalam kalimat. Peneliti menggunakan algortima TF-IDF pada tahap ini. Menurut (Cahyono & Anggista Oktavia Praneswara, 2023) pendekatan TF-IDF menghitung bobot setiap istilah yang paling umum digunakan dalam pengambilan informasi. Berikut adalah kata yang telah melalui pra-pemrosesan data dan akan menjalani pembobotan kata menggunakan metode TF-IDF. Berikut adalah tabel hasil proses TF-IDF.

**Tabel 9** Hasil Proses Perhitungan TF-IDF

| Token | TF-IDF | | | | | |
|---|---|---|---|---|---|---|
| | D1 | D2 | D3 | D4 | D5 | D6 |
| itel | 0,78 | 0 | 0 | 0 | 0 | 0 |
| biasa | 0,78 | 0 | 0 | 0 | 0 | 0 |
| tidak | 0,48 | 0 | 0 | 0 | 0,48 | 0 |
| worth | 0,78 | 0 | 0 | 0 | 0 | 0 |
| teman | 0,78 | 0 | 0 | 0 | 0 | 0 |
| tecno | 0 | 0,48 | 0 | 0,48 | 0 | 0 |
| lumayan | 0 | 0,78 | 0 | 0 | 0 | 0 |
| budget | 0 | 0,78 | 0 | 0 | 0 | 0 |
| rendah | 0 | 0,78 | 0 | 0 | 0 | 0 |
| oke | 0 | 0,78 | 0 | 0 | 0 | 0 |
| hape | 0 | 0 | 0,48 | 0,48 | 0 | 0 |
| infinix | 0 | 0 | 0,30 | 0 | 0,30 | 0,30 |
| gahar | 0 | 0 | 0,78 | 0 | 0 | 0 |
| luar | 0 | 0 | 0,78 | 0 | 0 | 0 |
| puas | 0 | 0 | 0,78 | 0 | 0 | 0 |
| banget | 0 | 0 | 0,48 | 0 | 0,48 | 0 |
| make | 0 | 0 | 0,78 | 0 | 0 | 0 |
| serius | 0 | 0 | 0 | 0,78 | 0 | 0 |
| murah | 0 | 0 | 0 | 0,78 | 0 | 0 |
| spek | 0 | 0 | 0 | 0,78 | 0 | 0 |
| kalah | 0 | 0 | 0 | 0,78 | 0 | 0 |
| lain | 0 | 0 | 0 | 0,78 | 0 | 0 |
| susah | 0 | 0 | 0 | 0 | 0,78 | 0 |
| service | 0 | 0 | 0 | 0 | 0,78 | 0 |
| rekomendasi | 0 | 0 | 0 | 0 | 0,78 | 0 |
| terbaik | 0 | 0 | 0 | 0 | 0 | 0,78 |
| kualitas | 0 | 0 | 0 | 0 | 0 | 0,78 |
| mantap | 0 | 0 | 0 | 0 | 0 | 0,78 |
| harga | 0 | 0 | 0 | 0 | 0 | 0,78 |
| ramah | 0 | 0 | 0 | 0 | 0 | 0,78 |
| kantong | 0 | 0 | 0 | 0 | 0 | 0,78 |

**5. *Cosine Similarity***

Selanjutnya dilakukan proses perhitungan *cosine similarity*, yang dilakukan pada saat perhitungan TF-IDF. *Cosine similarity* adalah algoritma yang digunakan untuk mengukur tingkat

kemiripan antar dokumen dengan menghitung ukuran kemiripan ruang vektor menggunakan kata kunci dari dokumen tertentu sebagai dasar perhitungan. (No et al., 2024)

Tahap ini bertujuan untuk mengklasifikasikan komentar ke dalam 5 jenis kategori yaitu performa, riwayat pemakaian, harga dan keputusan pembelian, perbandingan merek, kinerja fisik/komponen, dan apabila terdapat data yang tidak dapat dihitung akan masuk kedalam kategori "*unclassified*". Tahapan ini dilakukan dengan cara membandingkan similaritas antar data komentar dengan kata kunci kategori yang sudah ditentukan.. Lalu jika nilainya mendekati 1 maka dokumen dinyatakan mirip dan jika hasilnya 0 maka dokumen dinyatakan sebaliknya.berikut adalah tabel hasil *cosine similarity*.

Berikut ini adalah tabel yang menampilkan tabel hasil dari penghitungan menggunakan persamaan kosinus yang membagi 5 kategori pada setiap dataset.

**Tabel 10** Tabel Hasil Pembagian Kategori *Cosine Similarity* Infinix

| *Brand* | Kategori |
|---|---|
| Infinix | 1. Riwayat pemakaian<br>2. Performa<br>3. Harga dan keputusan pembelian<br>4. *Unclassified*<br>5. Perbandingan *brand*<br>6. Kinerja komponen |
| Tecno | 1. Riwayat pemakaian<br>2. Harga dan keputusan pembelian<br>3. *Unclassified*<br>4. Performa<br>5. Perbandingan *brand*<br>6. Kinerja komponen |
| Itel | 1. Harga dan keputusan pembelian<br>2. Riwayat pemakaian<br>3. *Unclassified*<br>4. Perbandingan *brand*<br>5. Performa<br>6. Kinerja komponen |

**6. Modelling Klasifikasi *Naive Bayes***

Setelah data yang sudah sesuai dengan bentuk uang dibutuhkan untuk pemodelan melalui tahapan sebelumnya. Makan dilanjutkan untuk proses klasifikasi menggunakan pemodelan *naive bayes classifier*. Naïve Bayes merupakan salah satu algoritma klasifikasi dengan memprediksi peluang masa depan berdasarkan data atau pengalaman masa lalu. Dalam proses klasifikasi Naïve Bayes diasumsikan ada atau tidaknya fitur tertentu pada suatu kelas tidak berhubungan dengan fitur pada kelas lainnya (Pebdika et al., 2023).

Pada proses pemodelan data akan dibagi menjadi 2 yaitu 80% untuk *data training* dan 20% untuk *data testing*. Data yang digunakan telah selesai melalui tahap pembobotan kata menggunakan TF-IDF, selanjutnya adalah melakukan klasifikasi menggunakan pemodelan *naive bayes classifier*. Setelah tahap klasifikasi selesai dilakukan, langkah berikutnya adalah menghitung hasil prediksi pada data tersebut untuk mengetahui peluang antar kelas dan mengukur performa dari algoritma *naive bayes classifier*.

**7. Evaluation**

**1) *Cofusion Matrix***

*Confusion matrix* adalah tabel yang digunakan untuk menilai kinerja model

klasifikasi. Hal ini memungkinkan melihat seberapa banyak data yang berhasil dan salah diklasifikasi oleh model, serta jenis kesalahan yang terjadi. *Confusion matrix* paling sering digunakan dalam masalah klasifikasi yang melibatkan banyak kelas *(multiclass)* atau dua kelas *(binary).* Hasil dari *confusion matrix* digunakan untuk menghitung nilai *precision, accuracy,f-score dan recall* (Raihan et al., 2021)

Hasil analisis sentimen menggunakan *naive bayes classifier* akan dievaluasi dengan menghitung presisi, *recall* dan *F1-Score*. Hail ini dilakukan untuk mengetahui akurasi performa dari algoritma pada setiap dataset. Hasilnya adalah akurasi dan tabel *confusion matrix* untuk setiap dataset. Berikut adalah gambar dari hasil evaluasi menggunaan *confusion matrix*.

Berikut adalah gambar yang menampilkan nilai akurasi dan tabel *confusion matrix* untuk setiap dataset.

```
Average Accuracy  : 0.78

Confusion Matrix:
[[ 19  31]
 [  0 150]]
```

**Gambar 2** Nilai Akurasi dan Tabel *Confusion Matrix* Infinix

```
Average Accuracy  : 0.728

Confusion Matrix:
[[  9  53]
 [  1 137]]
```

**Gambar 3** Nilai Akurasi dan Tabel *Confusion Matrix* Tecno

```
Average Accuracy  : 0.775

Confusion Matrix:
[[ 42  44]
 [  5 109]]
```

**Gambar 4** Nilai Akurasi dan Tabel *Confusion Matrix* Itel

Selanjutnya adalah mendapatkan nilai kinerja dari algoritma klasifikasi yang digunakan pada setiap dataset, diketahui bahwa nilai *precision* sebesar 75%, nilai *recall* sebesar 99%, dan nilai *f1-score* sebesar 85% untuk dataset Infinix. Lalu nilai *precision* sebesar 71%, nilai *recall* sebesar 98%, dan nilai *f1-score* sebesar 82% untuk dataset Tecno. Kemudian nilai *precision* sebesar 73%, nilai *recall* sebesar 95%, dan nilai *f1-score* sebesar 82% untuk dataset Itel.

Berikut adalah tabel yang menampilkan nilai kinerja setiap label pada dataset Infinix, Tecno, dan Itel.

**Tabel 11** Nilai Kinerja Setiap Label

| **Dataset** | **Precision** | | **Recall** | | **F1-Score** | |
|---|---|---|---|---|---|---|
| | **(+)** | **(-)** | **(+)** | **(-)** | **(+)** | **(-)** |
| Infinix | 0.83 | 1.00 | 1.00 | 0.38 | 0.91 | 0.55 |
| Tecno | 0.72 | 0.90 | 0.99 | 0.15 | 0.84 | 0.25 |
| Itel | 0.71 | 0.89 | 0.96 | 0.49 | 0.82 | 0.63 |

Tabel diatas menunjukkan bahwa akurasi antara informasi yang tersedia pada sistem yang diperoleh adalah presentase 83% untuk label positif dan 100% untuk label negatif pada dataset Infinix, sedangkan presentase 72% untuk label positif dan 90% untuk label negatif pada dataset Tecno, lalu presentase 71% untuk label positif dan 89% untuk label negatif pada dataset Itel. Selanjutnya, sistem juga berhasil memulihkan informasi dengan tingkat

keberhasilan sebesar 100% untuk label positif dan 38% untuk tabel negatif pada dataset Infinix, kemudian 99% untuk label positif dan 15% untuk label negatif pada dataset Tecno, sedangkan pada dataset Itel label positif menncapai 96% untuk label positif dan 49% untuk label negatif.

**2) *K-fold Cross Validation***

*K-Fold Cross validation* merupakan salah satu metode data mining yang tujuannya untuk memperoleh hasil akurasi yang maksimal ketika data dibagi menjadi dua subset dan salah satu jenis pengujiannya adalah menilai kinerja dalam proses metode algoritmik dengan cara membagi sampel data secara acak kemudian mengelompokkan data pada nilai K di *K-Fold* (Prasetya & Ferdiansyah, 2022).

Selanjutnya dilakukan pengujian ulang dengan K-fold Cross Validation yang bertujuan agar hasil uji dan evaluasi kinerja algoritma memperoleh hasil yang maksimal. Hasil pengujian menggunakan *k-fold cross validation* menunjukkan bahwa sistem mempunyai rata-rata tingkat akurasi untuk setiap dataset yang dapat dilihat pada tabel berikut.

**Tabel 12** Tingkat Akurasi

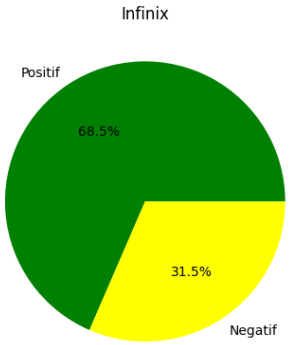
| Dataset | Akurasi |
|---|---|
| Infinix | 78% |
| Tecno | 72% |
| Itel | 77% |

**Tabel 13** *Recall*

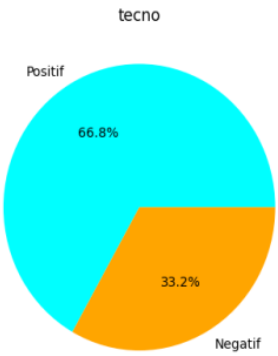
| Dataset | Recall |
|---|---|
| Infinix | 99% |
| Tecno | 95% |
| Itel | 98% |

**8. Visualisasi**

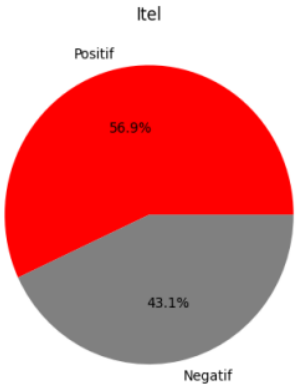
Pada tahap visualisasi ini hasil proses klasifikasi ke dalam bentuk chart agar lebih mudah menyajikan informasi yang ingin disampaikan. Tahap ini memanfaatkan *library matplotlib* dan *seaborn* untuk memvisualisasi data. Berikut adalah gambar grafik *pie chart* untuk melihat presentase sintimen positif dan negatif untuk setiap dataset




**Gambar 5** Diagram *Pie Chart* Dataset Infinix




**Gambar 6** Diagram Pie Chart Dataset Tecno




**Gambar 7** Diagram *Pie Chart* Dataset Itel

## KESIMPULAN

Berdasarkan hasil analisis yang telah dilaksanakan terkait sentimen konsumen terhadap merek smartphone Infinix, Tecno, dan Itel menggunakan metode *Naive Bayes* dan *Cosine Similarity* pada komentar di media sosial YouTube, dapat disimpulkan beberapa hal sebagai berikut :

1. Analisis Sentimen Konsumen

   Sentimen positif cenderung mendominasi komentar pada merek smartphone Infinix dan Tecno, menunjukkan bahwa kedua merek tersebut berhasil memenuhi ekspektasi konsumen dalam hal performa dan pengalaman pengguna.

   Sebaliknya, sentimen negatif lebih banyak ditemukan pada merek Itel, yang dapat menjadi indikasi bahwa konsumen kurang puas terhadap kualitas atau fitur yang ditawarkan oleh merek ini.
2. Kinerja Model *Naive Bayes*

   Model *Naive Bayes* yang digunakan menunjukkan performa yang baik dalam mengklasifikasikan komentar dengan akurasi rata-rata sebesar 78% untuk Infinix, 72% untuk Tecno, dan 77% untuk Itel.

   Metrik evaluasi lainnya, seperti *precision* dan *recall*, menunjukkan hasil yang beragam pada setiap dataset. *Precision* untuk dataset Infinix adalah 83% untuk label positif dan 100% untuk label negatif. Pada dataset Tecno, precision mencapai 72% untuk label positif dan 90% untuk label negatif, sedangkan pada dataset Itel precision tercatat sebesar 71% untuk label positif dan 89% untuk label negatif.

   Recall juga menunjukkan tingkat keberhasilan yang bervariasi. Untuk dataset Infinix, recall mencapai 100% untuk label positif dan 38% untuk label negatif. Pada dataset Tecno, recall mencapai 99% untuk label positif dan 15% untuk label negatif. Sementara itu, pada dataset Itel, recall tercatat sebesar 96% untuk label positif dan 49% untuk label negatif.Proses evaluasi menggunakan k-fold cross-validation memberikan gambaran konsistensi hasil model pada setiap iterasi, memastikan bahwa model dapat diandalkan.
3. Penerapan *Cosine Similarity*

   *Cosine Similarity* efektif dalam mengkategorikan komentar berdasarkan tiga kategori utama yaitu “performa”, “pengalaman pengguna”, dan “perbandingan merek”. Hal ini memberikan wawasan yang lebih mendalam tentang aspek yang menjadi perhatian utama konsumen terhadap masing-masing merek.
4. Analisis Tren

   Dari analisis tren sentimen selama periode Maret hingga Desember, ditemukan bahwa terdapat peningkatan sentimen positif pada merek Infinix dan Tecno, yang kemungkinan disebabkan oleh peluncuran produk baru atau promosi yang berhasil. Sentimen negatif terhadap merek Itel relatif stabil sepanjang periode tersebut, menunjukkan perlunya perbaikan strategis yang signifikan.

## DAFTAR PUSTAKA

Afdhal, I., Kurniawan, R., Iskandar, I., Salambue, R., Budianita, E., & Syafria, F. (2022). Penerapan Algoritma Random Forest Untuk Analisis Sentimen Komentar Di YouTube Tentang Islamofobia. *Jurnal Nasional Komputasi Dan Teknologi Informasi*, *5*(1), 122–130. http://ojs.serambimekkah.ac.id/jnkti/article/view/4004/pdf

Ananda, D., & Suryono, R. R. (2024). Analisis Sentimen Publik Terhadap

Pengungsi Rohingya di Indonesia dengan Metode Support Vector Machine dan Naïve Bayes. *Jurnal Media Informatika Budidarma*, *8*(April), 748–757. https://doi.org/10.30865/mib.v8i2.7517

Ayu, P., Purnama, W., & Putra, T. A. (2024). Klasifikasi Penjualan Produk Menggunakan Algoritma Naive Bayes pada Konter HP Bayu Cell. *Remik: Riset Dan E-Jurnal Manajemen Informatika Komputer*, *8*(1), 286–292. http://doi.org/10.33395/remik.v8i1.13207

Buani, D. C. P. (2024). Deteksi Dini Penyakit Diabetes dengan Menggunakan Algoritma Random Forest. *EVOLUSI : Jurnal Sains Dan Manajemen*, *12*(1), 1–8. https://doi.org/10.31294/evolusi.v12i1.21005

Cahyono, N., & Anggista Oktavia Praneswara. (2023). Analisis Sentimen Ulasan Aplikasi TikTok Shop Seller Center di Google Playstore Menggunakan Algoritma Naive Bayes. *Indonesian Journal of Computer Science*, *12*(6), 3925–3940. https://doi.org/10.33022/ijcs.v12i6.3473

Fitrana, L. A., Linawati, S., Herlinawati, N., Seimahuria, S., Informasi, S., Bina, U., Informatika, S., Data, S., Mandiri, U. N., Raya, J. K., Pusat, J., & Mining, T. (2024). *ANALISIS SENTIMEN PENGGUNA TWITTER TERHADAP BRAND INDOSAT*. *8*(3), 4291–4297.

Hasri, C. F., & Alita, D. (2022). Penerapan Metode Naã• Ve Bayes Classifier Dan Support Vector Machine Pada Analisis Sentimen Terhadap Dampak Virus Corona Di Twitter. *Jurnal Informatika Dan Rekayasa Perangkat Lunak*, *3*(2), 145–160. https://doi.org/10.33365/jatika.v3i2.2026

Maulana, R., Voutama, A., & Ridwan, T. (2023). Analisis Sentimen Ulasan Aplikasi MyPertamina pada Google Play Store menggunakan Algoritma NBC. *Jurnal Teknologi Terpadu*, *9*(1), 42–48. https://doi.org/10.54914/jtt.v9i1.609

No, V., Hal, J., Azizah, N., & Fauzan, A. (2024). *Sistem Rekomendasi Produk Somethinc Menggunakan Metode Content-based Filtering*. *6*(3), 461–468.

Pebdika, A., Herdiana, R., & Solihudin, D. (2023). Klasifikasi Menggunakan Metode Naive Bayes Untuk Menentukan Calon Penerima Pip. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(1), 452–458. https://doi.org/10.36040/jati.v7i1.6303

Permana, A. Y., & Makmun, M. (2020). *Analisis Sentimen pada Teks Opini Penilaian Kinerja Dosen dengan Pendekatan Algoritma KNN*. *19*, 39–50.

Prasetya, F., & Ferdiansyah, F. (2022). Analisis Data Mining Klasifikasi Berita Hoax COVID 19 Menggunakan Algoritma Naive Bayes. *Jurnal Sistem Komputer Dan Informatika (JSON)*, *4*(1), 132. https://doi.org/10.30865/json.v4i1.4852

Raihan, M., Allaam, R., & Wibowo, A. T. (2021). Klasifikasi Genus Tanaman Anggrek Menggunakan Metode Convolutional Neural Network (CNN). *E-Proceeding of Engineering*, *8*(2), 1153.