# BAB II

# TINJAUAN PUSTAKA

## 2.1 Studi Sebelumnya

Penelitian mengenai analisis sentimen pada media sosial telah banyak dilakukan dengan menggunakan berbagai metode klasifikasi, terutama pada *platform* X sebagai sumber data opini publik. Berbagai penelitian terdahulu menunjukkan bahwa analisis sentimen mampu memberikan gambaran mengenai persepsi masyarakat terhadap suatu isu yang sedang berkembang. Pada tahun 2025, Windarti [16] melakukan penelitian yang bertujuan untuk menganalisis sentimen publik terhadap sistem pajak Coretax di media sosial X dengan membandingkan metode Naïve Bayes dan Support Vector Machine (SVM). Penelitian ini menggunakan data sebanyak 7.632 *tweet* berbahasa Indonesia yang dikumpulkan melalui teknik *scraping*. Tahapan yang dilakukan meliputi *preprocessing* data, pelabelan sentimen menggunakan *Indonesia Sentiment Lexicon*, serta proses klasifikasi menggunakan kedua algoritma. Hasil penelitian menunjukkan bahwa mayoritas sentimen yang muncul bersifat negatif. Selain itu, metode SVM memperoleh akurasi sebesar 90%, lebih tinggi dibandingkan Naïve Bayes yang memperoleh akurasi sebesar 87%, sehingga menunjukkan bahwa SVM memiliki performa yang lebih baik dalam klasifikasi sentimen pada kasus tersebut.

Penelitian lain yang dilakukan oleh Nathania [17] pada tahun 2025 bertujuan untuk menganalisis sentimen publik terhadap kebijakan efisiensi anggaran pendidikan di media sosial X menggunakan algoritma *Support Vector Machine* (SVM). Data dikumpulkan dari unggahan pengguna di *platform* X dan diklasifikasikan ke dalam tiga kategori sentimen, yaitu positif, negatif, dan netral. Proses penelitian meliputi *preprocessing* data, pelatihan model, serta evaluasi menggunakan metrik akurasi, *precision*, *recall*, dan F1-*score*. Hasil penelitian menunjukkan bahwa sentimen negatif mendominasi persepsi masyarakat terhadap kebijakan tersebut. Model SVM dengan kernel RBF menghasilkan performa terbaik dengan akurasi sebesar 89,41%, sehingga dinilai efektif dalam mengklasifikasikan data teks tidak terstruktur.

Selanjutnya, penelitian yang dilakukan oleh Virginia [18] pada tahun 2025 bertujuan untuk menganalisis sentimen pengguna terhadap layanan *Buy Now Pay Later* (BNPL) di Indonesia menggunakan metode Naïve Bayes Classifier dan Latent Dirichlet Allocation (LDA). Penelitian ini tidak hanya melakukan klasifikasi sentimen, tetapi juga mengidentifikasi topik utama yang dibahas oleh pengguna. Data dikumpulkan dari media sosial X dan melalui tahapan *preprocessing*, klasifikasi, serta pemodelan topik. Berdasarkan hasil klasifikasi dengan algoritma *Naïve Bayes Classifier*, diketahui bahwa sebagian besar pengguna X memiliki pandangan negatif terhadap layanan BNPL, dengan persentase mencapai 86,94%. Sementara itu, sentimen positif dan netral hanya sebesar 9,39% dan 3,69%.

Penelitian lain yang dilakukan oleh Subrata [19] pada tahun 2026 bertujuan untuk menganalisis sentimen masyarakat terhadap profesionalisme Generasi Z di dunia kerja melalui *platform* X menggunakan metode *Support Vector Machine* (SVM). Data yang digunakan dalam penelitian ini berjumlah 2.095 *tweet* yang diperoleh melalui proses crawling, dengan pelabelan manual yang terdiri dari 1.092 sentimen negatif, 855 sentimen positif, dan 145 sentimen netral. Tahapan penelitian meliputi *preprocessing* data, ekstraksi fitur menggunakan TF-IDF, serta klasifikasi menggunakan algoritma SVM dengan kernel linear dan tiga skenario pembagian data, yaitu 70:30, 80:20, dan 90:10. Hasil penelitian menunjukkan bahwa performa terbaik diperoleh pada rasio 90:10 dengan nilai akurasi sebesar 69,52%, *precision* 66%, dan *recall* 70%. Temuan ini menunjukkan bahwa metode SVM cukup efektif dalam mengklasifikasikan data teks tidak terstruktur.

Sementara itu, Hutahaean [20] pada tahun 2026 melakukan penelitian yang bertujuan untuk menganalisis sentimen terhadap isu #BoikotProdukIsrael di media sosial X dengan membandingkan metode Naïve Bayes dan Random Forest. Data yang digunakan dalam penelitian ini berjumlah 5.896 *tweet* yang diperoleh melalui teknik *scraping* dan melalui tahapan *preprocessing* serta klasifikasi sentimen. Hasil penelitian mengindikasikan bahwa tingkat akurasi *Random Forest* sebesar 73,98% lebih unggul dibandingkan *Naïve Bayes* yang memperoleh akurasi sebesar 66,36%.

.

Berdasarkan penelitian sebelumnya, analisis sentimen di media sosial X telah banyak dilakukan pada isu kebijakan, layanan digital, dan fenomena sosial, seperti sistem pajak Coretax, efisiensi anggaran pendidikan, layanan BNPL, profesionalisme Generasi Z, hingga gerakan #BoikotProdukIsrael. Namun, penelitian yang secara khusus membahas sentimen publik terhadap isu kelangkaan BBM akibat konflik geopolitik internasional masih sangat terbatas. Isu ini memiliki karakteristik yang berbeda karena berkaitan langsung dengan kebutuhan pokok yang memicu kepanikan dan kekhawatiran ekonomi secara nyata di masyarakat.

Dari segi metodologi, meskipun algoritma *Naïve Bayes Classifier* sering digunakan dalam studi sentimen, kinerjanya lebih banyak dijadikan sebagai pembanding dan sering dianggap kurang unggul jika dibandingkan dengan *Support Vector Machine* (SVM) atau *Random Forest*. Oleh karena itu, evaluasi kinerja algoritma *Naïve Bayes Classifier* secara spesifik pada data teks yang berisi keluhan dan rasa panik masyarakat terkait kelangkaan energi masih belum banyak dieksplorasi, sehingga penelitian ini hadir untuk mengisi celah tersebut.

| No | Peneliti | Tahun | Metode | Hasil |
|---|---|---|---|---|
| 1 | Windarti | 2025 | Naïve Bayes dan Support Vector Machine | Hasil penelitian menunjukkan bahwa penggunaan dua algoritma klasifikasi menghasilkan perbedaan kinerja yang cukup mencolok. Algoritma *Naïve Bayes* mengelompokkan seluruh data ke dalam sentimen negatif dengan tingkat akurasi sebesar 87%. Di sisi lain, *Support Vector Machine (SVM)* juga mendominasi klasifikasi pada sentimen negatif, namun menunjukkan performa yang lebih baik dengan akurasi mencapai 90%. Secara keseluruhan, *SVM* terbukti lebih unggul dibandingkan *Naïve Bayes* dalam menganalisis sentimen terhadap sistem Coretax di *platform* X, baik dari aspek akurasi maupun distribusi hasil klasifikasi. Temuan ini juga mengindikasikan bahwa mayoritas opini pengguna terhadap Coretax cenderung bersifat negatif. |
| 2 | Nathania | 2025 | Support Vector Machine | Hasil penelitian menunjukkan bahwa opini masyarakat dapat dikaji secara efektif melalui pendekatan analisis sentimen. Pemetaan sentimen publik terhadap kebijakan efisiensi anggaran pendidikan berhasil dilakukan dengan menggunakan algoritma *Support Vector Machine (SVM)*. Model yang dibangun mampu mengidentifikasi opini masyarakat di media sosial dengan tingkat akurasi terbaik sebesar 89,41% menggunakan kernel RBF. Hasil klasifikasi menunjukkan bahwa sebagian besar sentimen publik terhadap kebijakan tersebut cenderung negatif, terutama berkaitan dengan dampaknya terhadap mahasiswa penerima KIP Kuliah serta isu kenaikan uang kuliah tunggal. Oleh karena itu, penelitian ini memberikan kontribusi dalam memahami persepsi masyarakat secara sistematis dan |

| No | Peneliti | Tahun | Metode | Hasil |
|---|---|---|---|---|
| | | | | berbasis data, yang dapat dimanfaatkan sebagai bahan pertimbangan bagi pemerintah dalam menyusun kebijakan yang lebih responsif terhadap kebutuhan publik. |
| 3 | Virginia | 2025 | Naïve Bayes Classifier Dan Latent Dirichlet Allocation | Hasil penelitian menunjukkan bahwa klasifikasi sentimen pengguna X terhadap layanan BNPL menggunakan algoritma *Naïve Bayes Classifier* menunjukkan bahwa mayoritas opini cenderung negatif dengan persentase sebesar 86,94%. Sementara itu, sentimen positif dan netral memiliki proporsi yang jauh lebih kecil, masing-masing sebesar 9,39% dan 3,69%. Selain itu, hasil pemodelan topik menunjukkan bahwa pada sentimen positif, pembahasan didominasi oleh isu seperti promo dan kemudahan akses, kecepatan layanan pinjaman online, serta kepercayaan terhadap layanan kredit. Pada sentimen netral, topik yang sering muncul meliputi diskusi umum terkait BNPL, tingkat aksesibilitas di Indonesia, serta penggunaan layanan tersebut. Adapun pada sentimen negatif, topik yang paling banyak dibicarakan mencakup keluhan dalam pembayaran belanja online, beban cicilan atau utang, serta risiko gagal bayar. |
| 4 | Subrata | 2026 | Support Vector Machine | Hasil penelitian menunjukkan bahwa analisis sentimen terhadap profesionalisme Generasi Z di dunia kerja yang bersumber dari data *platform* X periode 1 Januari hingga 30 November 2025 menunjukkan bahwa persepsi masyarakat Indonesia cenderung didominasi oleh sentimen negatif. Dari total 2.095 *tweet* yang dianalisis, sebanyak 1.092 termasuk sentimen negatif, 855 sentimen positif, dan 145 sentimen netral. Penerapan |

| No | Peneliti | Tahun | Metode | Hasil |
|---|---|---|---|---|
| | | | | metode *Support Vector Machine (SVM)* dalam klasifikasi sentimen menghasilkan tingkat akurasi yang bervariasi tergantung pada rasio pembagian data. Pada rasio 70:30 diperoleh akurasi sebesar 48,81%, kemudian meningkat menjadi 64,92% pada rasio 80:20, dan mencapai nilai tertinggi sebesar 69,52% pada rasio 90:10. Hal ini menunjukkan bahwa semakin besar proporsi data latih yang digunakan, semakin baik kemampuan model dalam mengenali pola sentimen. |
| 5 | Hutahaean | 2026 | Naïve Bayes dan Random Forest | Hasil penelitian menunjukkan bahwa algoritma *Naïve Bayes* dan *Random Forest* dapat digunakan untuk mengklasifikasikan sentimen publik terhadap gerakan #BoikotProdukIsrael dengan data 5.896 *tweet* berbahasa Indonesia menggunakan fitur TF-IDF. Proses *pre-processing* seperti *Cleaning*, *case folding*, *tokenizing*, *stopword removal*, dan *stemming* terbukti membantu meningkatkan kualitas data.<br>Dari sisi kinerja, *Random Forest* memperoleh akurasi lebih tinggi (73,98%) dibandingkan *Naïve Bayes* (66,36%). Meskipun *Naïve Bayes* lebih cepat, *Random Forest* memberikan hasil klasifikasi yang lebih stabil dan seimbang antar kelas. |

*Tabel 2.1 Studi Sebelumnya*

## 2.2 Dasar Teori

### 2.1.1 Analisis Sentimen

Analisis sentimen merupakan salah satu cabang dari *Natural Language Processing* (NLP) yang berfokus pada identifikasi, ekstraksi, dan klasifikasi opini atau emosi yang terkandung dalam data teks. Analisis ini bertujuan untuk mengetahui kecenderungan sentimen, seperti positif, negatif, atau netral, dari suatu pernyataan atau dokumen teks secara otomatis[21].

Untuk mengolah dan memahami opini publik yang tersebar di media sosial, analisis sentimen menjadi metode yang penting dalam mengetahui persepsi masyarakat terhadap suatu produk, layanan, maupun kebijakan tertentu. Informasi yang diperoleh dari analisis sentimen dapat digunakan sebagai dasar dalam pengambilan keputusan, baik dalam bidang bisnis, pemerintahan, maupun penelitian[22].

Analisis sentimen juga dikenal sebagai *opinion mining*, yaitu proses untuk mengekstraksi informasi subjektif dari data teks[23]. Teknik ini memungkinkan sistem komputer untuk mengenali emosi, sikap, serta pandangan yang disampaikan oleh pengguna dalam bentuk tulisan. Dengan demikian, analisis sentimen tidak hanya berfungsi untuk mengelompokkan opini, tetapi juga memberikan wawasan terhadap perilaku dan preferensi pengguna [21].

### 2.1.2 Text Mining

*Text mining* merupakan salah satu teknik dalam bidang data mining yang digunakan untuk mengekstraksi informasi dan pengetahuan dari data teks yang tidak terstruktur[24]. Seiring dengan perkembangan teknologi informasi, jumlah data teks terus mengalami peningkatan secara pesat. Kondisi tersebut menyebabkan proses pengolahan data secara manual menjadi semakin sulit, sehingga diperlukan metode yang mampu mengolah data tersebut menjadi informasi yang bernilai[25].

*Text mining* memungkinkan proses eksplorasi data teks secara otomatis untuk menemukan pola, konsep, serta hubungan yang tersembunyi di dalamnya[26]

### 2.1.3 Media Sosial

Media sosial merupakan *platform* komunikasi berbasis internet yang memungkinkan pengguna untuk berinteraksi, berbagi informasi, serta membangun hubungan sosial secara daring[27]. Media sosial dapat digunakan untuk menyampaikan berbagai bentuk konten seperti teks, gambar, video, dan audio secara *real-time* tanpa dibatasi oleh ruang dan waktu[28]. *Platform* seperti X, WhatsApp, YouTube, Facebook, dan Instagram menjadi media yang paling sering digunakan untuk berbagai keperluan, termasuk komunikasi jarak jauh dan penyebaran informasi[29].

### 2.1.4 X

*Platform* X (sebelumnya dikenal sebagai Twitter) merupakan salah satu bentuk media sosial berbasis microblogging yang memungkinkan pengguna untuk berbagi informasi secara singkat, cepat, dan interaktif. Hal ini menjadikan *Platform* X sebagai media yang efektif dalam menyampaikan informasi secara *real-time* kepada pengguna di berbagai wilayah tanpa batas ruang dan waktu [30].

Selain sebagai media komunikasi, *Platform* X juga berfungsi sebagai sarana distribusi informasi yang sangat cepat. Informasi yang dipublikasikan dapat dengan mudah menyebar melalui interaksi pengguna seperti *like*, *retweet*, dan *reply*, sehingga membentuk jaringan penyebaran informasi yang luas[31]. Dengan adanya fitur *trending topic*, pengguna dapat mengetahui isu - isu yang sedang berkembang dan menjadi perhatian publik, sehingga *platform* ini juga berperan dalam membentuk opini dan wacana sosial di masyarakat[32].

### 2.1.5 TF-IDF

*Term Frequency–Inverse Document Frequency* (TF-IDF) merupakan salah satu metode ekstraksi fitur dan pembobotan kata yang digunakan dalam pengolahan teks dan *information retrieval*. Metode ini memberikan bobot pada setiap kata berdasarkan frekuensi kemunculannya dalam suatu dokumen dan tingkat kepentingannya dalam keseluruhan kumpulan dokumen[33].

Secara umum, TF-IDF terdiri dari dua komponen utama, yaitu *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF). TF merupakan ukuran frekuensi kemunculan suatu kata dalam sebuah dokumen, yang menunjukkan seberapa penting kata tersebut dalam dokumen tersebut. Namun, nilai TF saja belum cukup karena tidak mempertimbangkan distribusi kata pada seluruh dokumen. Oleh karena itu, digunakan IDF yang berfungsi memberikan bobot pada kata – kata yang memiliki frekuensi kemunculan rendah dalam keseluruhan dokumen[34].

Nilai TF dapat dihitung dengan menggunakan rumus[35]:

$$\mathrm{TF} = \frac{\text{(jumlah kemunculan kata dalam dokumen)}}{\text{(jumlah kata dalam dokumen)}}$$

*Inverse Document Frequency* (IDF) berfungsi untuk menurunkan bobot kata yang terlalu umum. Secara matematis[36]:

$$IDF(t) = \log \frac{N}{df_t}$$

Langkah terakhir dalam proses ini adalah menghitung nilai TF-IDF, yaitu dengan mengalikan nilai *Term Frequency* (TF) dengan nilai *Inverse Document Frequency* (IDF) [37].

$$W_{t,d} = tf_t \times idf_t$$

Perkalian antara TF dan IDF menghasilkan bobot akhir yang mencerminkan kepentingan relatif suatu kata dalam dokumen tertentu. Representasi ini memungkinkan dokumen diubah menjadi vektor numerik yang dapat digunakan dalam perhitungan probabilitas pada algoritma Naïve Bayes[37].

### 2.1.6 Naïve Bayes Classifier (NBC)

Naive Bayes merupakan salah satu algoritma klasifikasi dalam bidang *machine learning* yang berbasis pada teori probabilitas, khususnya Teorema Bayes[17]. Algoritma ini digunakan untuk menentukan probabilitas suatu data termasuk ke dalam kelas tertentu berdasarkan fitur - fitur yang dimilikinya[38]. Dalam penerapannya, Naive Bayes mengasumsikan bahwa setiap fitur bersifat independen atau tidak saling bergantung ketika kelas data telah diketahui. Proses klasifikasi dilakukan dengan menghitung probabilitas setiap kelas, kemudian data dimasukkan ke dalam kelas yang memiliki nilai probabilitas paling tinggi. Algoritma ini dikenal sederhana, memiliki proses komputasi yang cepat, serta mampu menangani data dengan jumlah fitur yang besar. Oleh karena itu, Naive Bayes banyak digunakan dalam klasifikasi teks, analisis sentimen, penyaringan spam, sistem rekomendasi, diagnosis medis, dan deteksi kecurangan[39]

Secara matematis, dasar dari Naive Bayes adalah Teorema Bayes, yang dirumuskan sebagai berikut:

$$(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

Keterangan

(C|X) : Probabilitas terjadinya suatu hipotesis C ketika bukti X diketahui telah terjadi.

P(X|C) : Probabilitas munculnya data X dengan asumsi bahwa hipotesis atau kelas C benar.

P(C) : Probabilitas awal dari hipotesis C sebelum mempertimbangkan data X.

P(X) : Probabilitas kemunculan data X secara keseluruhan.

Selain dari cara kerjanya yang sederhana secara probabilitas, pemilihan Naïve Bayes Classifier sebagai metode utama dalam penelitian ini juga didasari oleh tingkat efisiensinya jika dibandingkan dengan algoritma lain. Dalam bidang analisis sentimen, terdapat berbagai metode *machine learning* dan deep learning yang dapat digunakan, seperti Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest, hingga Decision Tree. Sebagai landasan justifikasi penggunaan metode, Tabel berikut menyajikan perbandingan kelebihan dan kekurangan dari masing - masing algoritma tersebut berdasarkan literatur terdahulu:

| Algoritma | Kelebihan | Kekurangan |
|---|---|---|
| **Naïve Bayes Classifier (NBC)** | Proses komputasinya sangat cepat dan efisien[40]. | Asumsi independensi fiturnya kurang realistis, namun tetap menghasilkan akurasi tinggi[16]. |
| **Support Vector Machine (SVM)** | Mampu menghasilkan tingkat akurasi yang sangat tinggi[17]. | Waktu pelatihan model sangat lama dan memakan memori besar[40]. |
| **K-Nearest Neighbor (KNN)** | Sangat mudah diimplementasikan tanpa proses pelatihan khusus[41]. | Kinerjanya menjadi sangat lambat saat memproses dataset berskala besar[41]. |
| **Random Forest** | Tahan terhadap *overfitting* dan baik untuk data tidak seimbang[42]. | Waktu prediksi lebih lama dan logika modelnya sulit diinterpretasikan[42]. |
| **Decision Tree** | Kesederhanaan dan kecepatan pelatihan[42]. | Rentan mengalami *overfitting[42]*. |

*Tabel 2.2 Tabel Perbandingan Kelebihan dan Kekurangan Beberapa Algoritma*

Berdasarkan perbandingan pada Tabel 2.2, dapat diketahui bahwa meskipun metode seperti SVM dan Random Forest menawarkan tingkat akurasi yang lebih tinggi, kedua model tersebut memiliki kelemahan signifikan pada lambatnya waktu komputasi dan besarnya sumber daya memori yang dibutuhkan.

Di sisi lain, metode seperti KNN dan *Decision Tree* memiliki kerentanan terhadap penurunan performa ketika dihadapkan pada *dataset* teks yang berdimensi tinggi dan rentan mengalami *overfitting*.

Oleh karena itu, *Naïve Bayes Classifier* dipilih sebagai metode utama dalam penelitian ini karena algoritma tersebut paling optimal dalam menyeimbangkan tingkat akurasi yang baik dengan kecepatan komputasi yang sangat tinggi, sehingga sangat ideal untuk memproses dan mengklasifikasikan data opini publik dari media sosial X yang berjumlah masif.

### 2.1.7 Confusion Matrix

*Confusion matrix* merupakan tabel yang digunakan untuk mengevaluasi kinerja model klasifikasi dengan membandingkan hasil prediksi model dengan kelas atau label yang sebenarnya. Tabel ini menunjukkan jumlah data yang diprediksi dengan benar maupun salah oleh model[17]. Pada klasifikasi sentimen yang terdiri dari kelas negatif, netral, dan positif, nilai pada bagian diagonal menunjukkan data yang berhasil diklasifikasikan dengan benar, sedangkan nilai di luar diagonal menunjukkan kesalahan prediksi antar kelas[16].

Komponen utama dalam confusion matrix pada klasifikasi biner terdiri dari[17]:

a. *True Positive* (TP), yaitu data positif yang diprediksi dengan benar sebagai positif.
b. *True Neutral* (TNt), yaitu data netral yang diprediksi dengan benar sebagai netral.
c. *True Negative* (TNg), yaitu data negatif yang diprediksi dengan benar sebagai negatif.
d. *False Positive* (FP), yaitu data positif yang salah diprediksi sebagai positif.
e. *False Neutral* (FNt), yaitu data netral yang salah diprediksi sebagai netral.
f. *False Negative* (FNg), yaitu data positif yang salah diprediksi sebagai negatif.