# BAB II

# TINJAUAN PUSTAKA DAN DASAR TEORI

## 2.1 Tinjauan Pustaka

Ada beberapa penelitian terdahulu yang sejenis sehingga digunakan sebagai acuan studi pustaka dalam penelitian ini. Adapun tinjauan pustaka pada penelitian terdahulu diantaranya adalah Febriyani & Februariyanti, (2021), Verena et al., (2021), Duei Putri et al., (2022), Wibowo et al., (2022) , Setya Ananto & Hasan, (2023).

Penelitian Febriyani & Februariyanti, (2021) melakukan analisis sentimen terhadap program kampus merdeka menggunakan algoritma *naive bayes classifier* di *twitter*. Pada penelitian ini pengumpulan data dilakukan menggunakan *website* vicinitas.io, analisis terhadap data *tweet* hanya dilakukan dengan mengklasifikasikan teks dalam bentuk negatif dan positif menggunakan algoritma *naive bayes classifier*. Dari hasil analisis dan pengujian yang telah dilakukan dapat disimpulkan bahwa klasifikasi menggunakan algoritma *naive bayes classifier* yang digunakan pada penelitian ini mendapatkan hasil kurang memuaskan dikarenakan jumlah data yang sedikit yaitu sejumlah 501 data, namun evaluasi pemodelan menggunakan *k-fold cross validation* pada penelitian ini cukup efektif digunakan untuk mendapatkan hasil akurasi lebih maksimal.

Kemudian, adanya penelitian dari Verena et al., (2021) yang menganalisis sentimen twitter menggunakan metode *naive bayes* dengan *relevance frequency feature selection*. (studi kasus: opini masyarakat mengenai kebijakan *new normal*). Klasifikasi menggunakan metode *naive bayes* dengan menggunakan dataset

sejumlah 300 data opini masyarakat, dengan pembagian data menggunakan *k-fold cross validation* dengan k=5. Hasil dari pengujian sebanyak 5 pengujian menggunakan klasifikasi *naive bayes*, berdasarkan hasil dari pengujian yang telah dilakukan didapatkan beberapa simpulan pada penelitian ini yaitu penggunaan metode *naive bayes* pada analisis sentimen mengenai opini masyarakat pada twitter tentang kebijakan new normal mendapatkan akurasi tertinggi sebesar 73,3% yang terjadi pada fold ke-4 dalam 5 kali pengujian dengan pembagian dataset menggunakan 5-*fold cross validation* dan dihasilkan rata-rata akurasi sebesar 62,6%. Selain itu penggunaan metode seleksi fitur *relevance frequency feature selection* berpengaruh dalam meningkatkan hasil akurasi. Dari 5 kali pengujan yang dilakukan, terjadi peningkatan hasil akurasi di setiap *fold*-nya. Terjadi peningkatan nilai rata-rata akurasi sebesar 2,7%. Dengan hasil rata-rata akurasi yang didapatkan sebesar 65,3%.

Penelitian mengenai analisis sentimen kinerja Dewan Perwakilan Rakyat (DPR) pada twitter menggunakan metode *naive bayes classifier* yang dilakukan oleh Duei Putri et al., (2022) mengemukakan bahwa sistem *sentiment analyzer* tersebut telah berhasil dibuat untuk mendapatkan tweet dari twitter mengenai kinerja DPR. Data hasil analisis berupa jumlah klasifikasi, *accuracy score* berhasil disajikan ke dalam *website*, berdasarkan penelitian menggunakan algoritma *naive bayes* didapatkan *accuracy score* sebesar 0.8 atau 80% hal ini berarti sistem mampu memprediksi 80% secara akurat dari total data *testing* sebesar 20%. Hasil analisis dapat digunakan untuk memprediksi *dataset* baru tanpa harus dilakukan *labeling* terlebih dahulu, lalu hasil analisis sistem mendapatkan klasifikasi *tweet* dari *twitter*

mengenai DPR sebanyak 95 positif, 693 netral dan 758 negatif dari data hasil *crawling* sebanyak 1546.

Penelitian oleh Wibowo et al., (2022) bertujuan menganalisis *opini* publik terhadap pembelajaran *daring* pada masa pandemi COVID-19 di Indonesia pada awal november 2020. Penelitian dilakukan dengan penambangan teks berbasis dokumen pada twitter yang dianalisis menggunakan algoritma *naive bayes*. Temuan menunjukkan bahwa pembelajaran daring memiliki 30% sentimen positif, 69% sentimen negatif, dan 1% netral pada periode tersebut. Tingginya sentimen negatif dihasilkan karena ketidakpuasaan masyarakat terhadap pembelajaran daring. Beberapa *tweet* menunjukkan kekecewaan dengan kata ‘stress’ dan ‘malas’ merupakan kata yang memiliki frekuensi tinggi dalam percakapan.

Penelitian yang membahas mengenai implementasi algoritma *naive bayes* terhadap analisis sentimen ulasan media sosial MyPertamina pada *google play store* yang dilakukan oleh (Setya Ananto & Hasan, 2023), pengambilan data pada ulasan media sosial MyPertamina diproses menggunakan teknik *web scrapping* menggunakan *website google collab*. Data yang telah didapat kemudian diberi pelabelan antara positif atau negatif. Klasifikasi ini bertujuan untuk mencari nilai *accuracy*, *precision*, dan *recall* dari data ulasan media sosial *mypertamina* yang sudah didapat. Setelah melakukan klasifikasi, data kemudian di evaluasi dan didapatkan hasil nilai *accuracy* sebesar 77.42%, nilai *precision* sebesar 49.98%, dan nilai *recall* sebanyak 76.87%.

**Tabel 2. 1 Perbandingan Dengan Penelitian Sebelumnya**

| No | Penulis | Tahun | Judul | Objek | Hasil |
|---|---|---|---|---|---|
| 1. | Febriyani & Februariyanti, (2021) | 2021 | Analisis sentimen terhadap program kampus merdeka Menggunakan algoritma *naive bayes classifier* di twitter | Twiter | Penelitian dan pengklasifikasian yang dilakukan oleh sistem mendapatkan hasil klasifikasi sentimen positif sebanyak 272 opini dan sentimen negatif sebanyak 229 opini dengan rata-rata akurasi 60%, presisi 64%, recall 58% dan f1-score 58%. |
| 2. | Verena et al., (2021) | 2021 | Analisis sentimen twitter menggunakan metode *naive bayes* dengan *relevance frequency feature selection* (studi kasus: opini masyarakat mengenai kebijakan new normal) | Twiter | Dari pengujian sebanyak 5 pengujian menggunakan klasifikasi *naive bayes*, diperoleh rata-rata akurasi sebesar 62,6%, sementara hasil pengujian akurasi klasifikasi dengan penambahan rffs diperoleh rata-rata akurasi sebesar 65,3%. |
| 3. | Duei Putri et al., (2022) | 2022 | Analisis sentimen kinerja Dewan Perwakilan Rakyat (DPR) pada twitter menggunakan metode *naive bayes classifier* | Twiter | Penelitian ini menggunakan sebanyak 1546 data tweet. Hasil dari penelitian ini didapatkan bahwa DPR mendapatkan 95 tweet positif dengan polaritas 0.75 atau 75% sentimen positif, 693 tweet netral dengan polaritas 0.79 atau 79% sentimen netral dan 758 tweet negatif dengan polaritas 0.82 atau 82% sentimen negatif dengan *accuracy score* 0.8 atau 80% berdasarkan data testing sebanyak 20%. |
| 4. | Wibowo et al., (2022) | 2021 | Analisis sentimen pembelajaran *daring* pada twitter di masa pandemi | Twiter | Penelitian pada periode tersebut menunjukkan 30% sentimen positif, 69% sentimen negatif, dan 1% netral. |

| | | | | | |
|---|---|---|---|---|---|
| | | | covid-19 menggunakan metode *naive bayes* | | |
| 5. | Setya Ananto & Hasan, (2023) | 2023 | Implementasi algoritma *naïve bayes* terhadap analisis sentimen ulasan media sosial *mypertamina* pada *google play store* | Aplikasi MyPertamina | Data bersih 1289 dengan jumlah data positif 285 dan data negatif 1004. Pada tahap klasifikasi menggunakan teknik cross validation dengan metode k-fold cross dan juga menggunakan implementasi algoritma naïve bayes yang menghasilkan nilai accuracy 77.42%, precision 49.98%, dan recall 76.87%. |
| 6. | Salsabila Nur Isara, (2024) | 2024 | Penerapan metode *naive bayes classifier* guna menganalisis sentimen pengguna media sosial X pada program pencegahan *stunting* | Twiter | Dari data sampel 5.630 didapatkan jumlah sentimen positif sebanyak 4.610 data dan sentimen negatif sebanyak 1.020 data. Klasifikasi *naive bayes* mndapatkan akurasi dengan persentase sebesar 85,26%, model evaluasi dengan *confusion matrix* diperoleh hasil presisi 83.80%, recall 85.26%, dan f1-score 83.42%. |

## 2.2 Dasar Teori

Pada bagian ini berisi definisi, penjelasan dan uraian yang diperoleh dari berbagai referensi yang dipublikasikan pada media sosial berkaitan dengan topik penelitian.

### 2.2.1 Analisis Sentimen

Analisis sentimen dikenal sebagai *opinion mining*, yang merupakan proses untuk mengidentifikasi dan mengekstraksi opini atau emosi dari teks, seperti komentar, ulasan, atau tweet. Analisis sentimen dapat digunakan untuk berbagai tujuan, seperti memahami sikap pelanggan, mengukur kepuasan karyawan, atau

memantau sentimen publik terhadap isu sosial. Analisis sentimen dilakukan dengan mengumpulkan, mengelompokkan, dan menganalisis data berupa teks dari berbagai sumber, termasuk media sosial, survei, dan ulasan pelanggan. Analisis sentimen dapat digunakan untuk menentukan apakah sentimen yang diekspresikan dalam teks bernada positif, negatif, atau netral (Camelia, 2023).

### 2.2.2 *Crawling* Data

Crawling adalah proses pengambilan data dari media sosial yang kemudian dikumpulkan menjadi satu untuk dievakuasi dan dibentuk agar menjadi sebuah penelitian. Proses crawling data terbagi menjadi dua cara, yaitu dengan menggunakan *API* dan tanpa *API* (Tineges, 2021).

### 2.2.3 *Tweet Harvest*

*Tweet Harvest* merupakan program perintah yang menggunakan *Playwright* untuk mendapatkan *tweets* dari hasil pencarian X berdasarkan kata kunci dan rentang tanggal yang ditentukan. Tweets yang berhasil didapatkan kemudian akan disimpan dalam file berbentuk CSV. Dalam penggunaan *tweet harvest* dibutuhkan *authorization token* yang didapatkan dengan login ke akun X di *browser* anda kemudian mengekstrak *authorization token* (Vincent, 2023).

### 2.2.4 *Machine Learning*

Menurut Nils John Nilsson dalam buku *Learning Machines* yang di publikasikan oleh (Hill-McGraw, 1965), *Machine Learning* atau Pembelajaran Mesin adalah “suatu bidang yang mempelajari bagaimana membuat program komputer yang dapat meningkatkan kinerjannya berdasarkan pengalaman”. Dalam pendefinisiannya yang lain, *Machine Learning* atau Pembelajaran Mesin adalah

“suatu bidang yang mempelajari bagaimana membuat mesin yang dapat belajar secara mandiri tanpa arahan langsung dari pengguna”.

Dari definisi di atas dapat disimpulkan bahwa *Machine Learning* atau bisa disebut dengan Pembelajaran Mesin merupakan cabang dari AI yang fokus belajar dari data (*learn from data*), yaitu fokus pada pengembangan sistem yang mampu belajar secara “mandiri” tanpa harus berulang kali diprogram manusia.

### 2.2.5 Python

Menurut Ljubomir Perkovic pada bukunya, *python* merupakan bahasa pemrograman interpretatif multiguna dengan filosofi perancangan yang berfkus pada tingkat keterbacaan kode. Tujuan dikembangkannya pemrograman *python* ini untuk membuat *source code* mudah dibaca. *Python* juga memiliki *library* yang lengkap sehingga memungkinkan *programmer* untuk membuat media sosial yang mutakhir dengan menggunakan *source code* yang terlihat sederhana (Wiley, 2011).

### 2.2.6 X (Twitter)

X atau yang sebelumnya bernama twitter adalah sebuah platform media sosial yang memungkinkan pengguna untuk berbagi pesan singkat, gambar, video, dan konten lainnya dengan orang-orang di seluruh dunia. Sehubungan dengan perubahannya, X juga memiliki fitur-fitur yang baru seperti X *spaces*, X *fleets*, X *trends*, dan X *moments* yang dapat digunakan untuk berinteraksi dengan pengguna lain, mengikuti topik yang sedang populer, dan mengetahui berita terkini (Columbres Duane, 2023).

X didirikan pada tahun 2006 oleh Jack Dorsey, Noah Glass, Biz Stone, dan Evan Williams dengan nama twitter. Pada tahun 2022, twitter diakuisisi oleh Elon

Musk, seorang pengusaha dan visioner yang juga memiliki perusahaan seperti tesla, *spacex*, *neuralink*, dan *the boring company*. Musk kemudian mengubah nama twitter menjadi X pada tahun 2023, dengan alasan bahwa nama twitter tidak sesuai lagi dengan fitur-fitur baru yang akan ditambahkan (Isnanto Ardi, 2023).

### 2.2.7 *Stunting*

Menurut *World Health Organization* (WHO) (2015), *stunting* adalah gangguan pertumbuhan dan perkembangan anak akibat kekurangan gizi kronis dan infeksi berulang, yang ditandai dengan panjang atau tinggi badannya berada di bawah standar.

Selanjutnya menurut *World Health Organization* (WHO) (2020) *stunting* adalah pendek atau sangat pendek berdasarkan panjang atau tinggi badan menurut usia yang kurang dari -2 standar deviasi (sd) pada kurva pertumbuhan *World Health Organization* (WHO).

### 2.2.8 *Text Processing*

*Text processing* merupakan proses pengolahan data teks yang bertujuan untuk mengambil informasi tentang polaritas atau emosi dari teks tersebut, misalnya positif, negatif, atau netral. *Text processing* meliputi beberapa langkah, antara lain:

1. *Case Folding*

*Case folding* merupakan proses penyamaan format huruf dalam sebuah dokumen. Dilakukannya tahap ini sebab tidak semua dokumen teks konsisten dalam penggunaan huruf kapital . oleh karena itu diperlukan case folding dalam mengubah

keseluruhan teks dalam suatu dokumen dari huruf kapital menjadi huruf kecil (Izzah & Girsang 2021).

2. *Cleansing*

Cleansing merupakan tahap proses pembersihan kata dan karakter pada data yang tidak digunakan untuk mengurangi noise pada proses klasifikasi. Kata yang perlu dibersihkan seperti *username*, *mention*, *link url*, angka, *hashtag* dan sebagainya. Untuk karakternya sendiri berupa (@#$%^&()_+":{}<.,?!/[]).

3. *Tokenizing*

Dikutip dari Leravio, (2021), proses *tokenizing* adalah proses membagi teks menjadi unit-unit yang lebih kecil, seperti kata, frasa, atau simbol. Proses ini berguna untuk memudahkan analisis sentimen, karena dapat mengidentifikasi kata-kata kunci yang mengungkapkan perasaan, seperti "senang", "marah", atau "kecewa".

4. *Normalization*

Tahap normalisasi dilakukan untuk merubah kata yang tidak baku atau terdapat kesalahan ejaannya menjadi kata baku dengan menggunakan kamus normalisasi yang dibuat secara manual berdasarkan pengecekan data secara keseluruhan. Contoh : "tdk" merupakan singktan dari kata penghubung "tidak".

5. *Stopword Removal*

Proses ini dilakuakan untuk penghapusan kata yang dirasa tidak memiliki makna yang jelas. Seperti kata sambung, agar data yang didapatkan nantinya hanya berupa kalimat kalimat pokok (Wibowo et al., 2022b).

6. *Stemming*

*Stemming* merupakan proses pemetaan dan penguraian yang berbentuk dari suatu kata menjadi bentuk kata dasarnya. Algortima stemming dikembangkan berdasarkan aturan mafologi bahasa Indonesia, yang mengelompokkan imbuhan menjadi awalan (*prefix*), sisipan (*infix*), akhiran (*sufflix*) dan gabungan awalan akhiran (*confixes*) (Syarifuddin, 2020).

### 2.2.9 *TextBlob*

*TextBlob* merupakan salah satu *library* pada bahasa pemrograman *python* berbasis *lexicon* yang digunakan untuk pemrosesan data dalam teks. *Library textblob* ini menyediakan API yang sederhana untuk melaksanakan tugas Natural *Language Processing* (NLP) (Loria, 2018). Berikut adalah fitur-fitur yang ada pada *textlob* yaitu :

1. *noun phrase extraction* : Mengidentifikasi frasa benda dari teks, yaitu kumpulan kata yang berfungsi sebagai benda dalam kalimat.
2. *part-of-speech tagging* : Menentukan kategori sintaksis dari setiap kata dalam teks, seperti kata benda, kata kerja, kata sifat, dll.
3. *sentiment analysis* : Menganalisis emosi atau sikap dari teks, seperti positif, negatif, atau netral.
4. *classification* : Mengklasifikasikan teks ke dalam kategori tertentu, seperti topik, genre, atau label.
5. *tokenization* : Membagi teks menjadi unit yang lebih kecil, seperti kata atau kalimat.
6. *word and phrase frequencies* : Menghitung frekuensi kemunculan kata atau frasa dalam teks.

7. *parsing* : Menganalisis struktur gramatikal dari teks, seperti subjek, predikat, objek, dll.
8. *n-gram* : Menghasilkan kumpulan n kata yang berurutan dalam teks.
9. *word inflection* : Mengubah bentuk kata, seperti menjadi jamak atau tunggal, atau menjadi kata kerja berimbuhan.
10. *and lemmatization* : Mengubah kata ke bentuk dasarnya, seperti mengubah “menulis” menjadi “tulis” atau “mice” menjadi "mouse".
11. *spelling correction* : Memperbaiki ejaan kata yang salah dalam teks.
12. *add new models or language through extension* : Menambahkan model atau bahasa baru yang tidak tersedia secara default pada TextBlob melalui ekstensi.
13. *wordnet integration* : Mengintegrasikan WordNet, yaitu basis data leksikal yang berisi sinonim, antonim, hipernim, hiponim, dan relasi semantik lainnya antara kata.

Dalam konteks pelabelan dengan *textblob*, terdapat ekspresi lambda yang biasanya digunakan untuk menerapkan fungsi pda setiap elemen dalam sebuah kolom data frame. Berikut merupakan contoh penggunaan ekspresi lambda dengan *textblob*.

```
Df[‘sentimen’] = df[‘text’].apply(lamda x: TextBlob(str(x)),sentiment)
```

Keterangan :

Lambda x : Mendefinisikan fungsi anonim yang mengambil satu argumen.

X : Mewakili teks dari setiap baris dalam kolom data frame.

TextBlob(str(x)),sentiment) : Menghitung sentimen dari teks yang mengembalikan namedtuple dengan polaritas dan subjektifitas.

Subjektivitas adalah ukuran seberapa subjektif atau objektif sebuah teks, sedangkan polaritas adalah ukuran seberapa positif atau negatif sebuah teks.

### 2.2.10 *Naive Bayes Classifiers*

Menurut Raschka Sebastian, (2014) dalam bukunya "*Python Machine Learning*", *Naive Bayes Classifiers* adalah "sekelompok algoritma klasifikasi probabilistik yang didasarkan pada teorema Bayes dengan asumsi bahwa semua fitur yang terkait dengan kelas target adalah independen satu sama lain".

Sedangkan menurut M. Mitchell. Tom, (2020) dalam bukunya "*Machine Learning*", *Naive Bayes Classifiers* adalah "metode klasifikasi probabilistik yang didasarkan pada teorema Bayes dengan asumsi bahwa setiap fitur yang terkait dengan kelas target adalah independen satu sama lain".

Secara umum *Naive Bayes Classifiers* adalah salah satu metode klasifikasi yang digunakan dalam *machine learning*. Metode ini didasarkan pada teorema Bayes yang menyatakan bahwa probabilitas suatu hipotesis dapat dihitung berdasarkan probabilitas kondisional dari setiap fitur yang terkait dengan hipotesis tersebut. *Naive Bayes Classifiers* sering digunakan dalam klasifikasi teks, seperti klasifikasi spam email atau klasifikasi dokumen berita.

Persamaan dari teorema *bayes* adalah sebagai berikut :

$$P(H \mid X) = \frac{P(X|X).P(X)}{P(X)} \qquad (2.1)$$

Keterangan :

X : data dengan class yang belum diketahui

H : hipotesis data X merupakan suatu *class* spesifik

P(H/X) : probabilitas hipotesis H berdasar kondisi X (*posteriori probability*)

P(H) : probabilitas hipotesis H (*prior probability*)

P(X/H) : probabilitas X berdasarkan kepada kondisi pada hipotesis H

P(X) : probabilitas X

Rumus yang digunakan untuk klasifikasi Naive Bayes adalah sebagai berikut :

$$P(C_i \mid X) = \frac{P(X|C_i)P(C_i)}{P(X)} \quad (2.2)$$

Keterangan :

$P(X \mid C_i)$ :probabilitas posterior, yaitu probabilitas kelas $C_i$ jika diberikan data $X$.

$P(X \mid C_i)$ : probabilitas likelihood, yaitu probabilitas data $X$ jika kelasnya adalah $C_i$

$P(C_i)$ : probabilitas prior, yaitu probabilitas kelas $C_i$ sebelum diberikan data $X$.

$P(X)$ : probabilitas evidence, yaitu probabilitas data $X$ tanpa memperhatikan kelasnya.

### 2.2.11 *Confusion Matrix*

*Confusion Matrix* adalah sebuah tabel yang menampilkan hasil prediksi dan nilai aktual dari sebuah model klasifikasi machine learning (Afifah Lutfia, 2023). Confusion Matrix terdiri dari empat elemen, yaitu dapat dilihat pada tabel 2.2 berikut :

**Tabel 2.2 Ilustrasi tabel confusion matrix**

| Nilai Aktual | Nilai Prediksi | |
|---|---|---|
| | Positif | Negatif |
| Positif | TP (True) | FN (False Negatif) |
| Negatif | FP (False) | TN |

Keterangan :

TP (*True Positif*), merupakan data kelas positif yang terdeteksi benar.

FP (*False Positif*), merupakan data kelas negatif yang terdeteksi sebagai data positif.

FN (*False Negatif*), merupakan data kelas positif yang salah diklasifikasikan sehingga tergolong ke dalam data negatif.

TN (*True Negatif*), merupakan data kelas negatif yang terdeteksi benar

Berdasarkan nilai TP (*True Positif*), FP (*False Positif*), FN (*False Negatif*), TN (*True Negatif*), maka dapat diperoleh nilai Akurasi, Presisi, Recall, F1-Score sebagai berikut :

1. Akurasi : rasio prediksi benar (positif dan negatif) dengan keseluruhan data. Akurasi menjawab pertanyaan "berapa persen tweet yang benar diprediksi sentimennya?”

$$\text{Akurasi} = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}} \qquad (2.3)$$

2. Presisi : rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif. Presisi menjawab pertanyaan "berapa persen tweet yang diprediksi positif benar-benar positif?”

$$\text{Presisi} = \frac{\text{TP}}{\text{TP+FP}} \qquad (2.4)$$

3. Recall : rasio prediksi benar positif dibandingkan dengan keseluruhan data yang sebenarnya positif. Recall menjawab pertanyaan "berapa persen tweet yang sebenarnya positif berhasil diprediksi positif?"

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (2.5)$$

4. F1-score : rata-rata harmonik dari presisi dan recall yang memperhitungkan kedua metrik tersebut. F1-score menjawab pertanyaan "bagaimana keseimbangan antara presisi dan recall?"

$$\text{F1-score} = \frac{2\times\text{Presisi}\times\text{Recall}}{\text{Presisi+Recall}} \quad (2.6)$$