# BAB II

# LANDASAN TEORI

## 2.1. *Scrapping* Data

*Scrapping* data merupakan teknik yang mengacu pada operasi teknis ekstraksi informasi yang berkaitan dengan pemrosesan data secara otomatis. Teknik *scrapping* digunakan untuk pengumpulan data secara online terkait bentuk penelitian social digital saat ini (Juniarsih, Faja and Esyudha, 2020). Umumnya data *scrapping* digunakan untuk beberapa pekerjaan yang berkaitan dengan data seperti *research* untuk konten website, keperluan bisnis dalam komparasi harga, atau melakukan riset penelitian pada sumber data public.

### 2.1.1. Dataset

Dataset adalah istilah informal yang merujuk pada kumpulan data dan digunakan untuk klasifikasi dengan metode data mining. Beberapa definisi tentang dataset menyatakan sebagai berikut:

a. Kumpulan data yang berasal dari informasi-informasi pada masa lalu dan siap untuk di kelola menjadi informasi baru (Nurul Hidayati,2021)
b. Representasi data yang disimpan di memori dalam kondisi tidak terhubung (kumparan, 2021)
c. Representasi di memori dari satu tabel atau lebih dan digunakan untuk menyimpan baris yang didapatkan saat permintaan dikirim ke basis data. Dataset dapat ditambahkan, dihapus, atau memperbarui baris dalam memori (Jubile Enterprise, 2021)

## 2.2. Peduli Lindungi

Peduli Lindungi merupakan pengaplikasian teknologi digital dalam mendukung penerapan protocol kesehatan dan 3T *(Testing, Tracing, Treatment)* Dalam upaya krusial untuk pengendalian pandemic Covid-19 (Srlll, 2021)

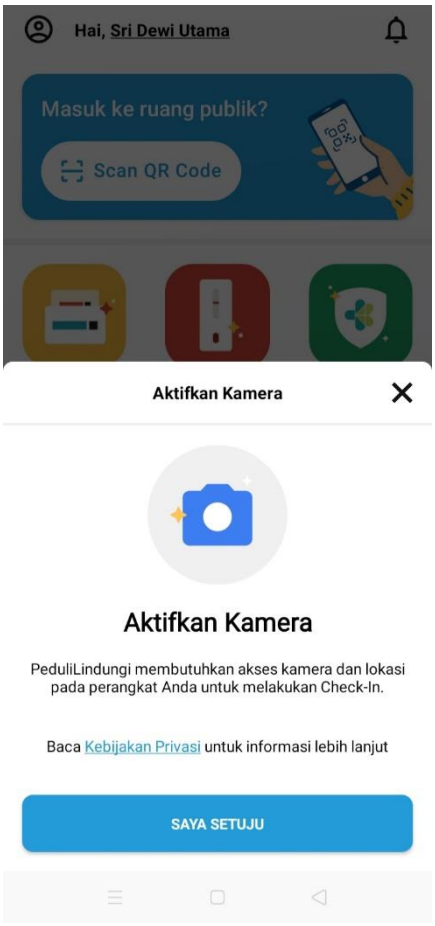
Fungsi utama aplikasi Peduli Lindungi dapat membantu setiap warga melakukan surveilans kesehatan berupa penelusuran *(tracing)*, pelacakan *(tracking)* dan pengurungan *(fencing)* terhadap anggota masyarakat yang terpapar Covid-19.


Gambar 2.1 Tampilan Awal Aplikasi Peduli Lindungi

Gambar 2.1 merupakan tampilan awal aplikasi Peduli Lindungi berguna untuk pendaftaran vaksinasi, melihat status vaksinasi, melihat dan mengunduh sertifikat vaksin, dan melihat status keamanan bepergian tiap individu. Pemanfaatan aplikasi tersebut sangat penting terhadap penerapan perpanjangan PPKM, di mana pemerintah melaksanakan beberapa uji coba penyesuaian aktivitas masyarakat. Salah satunya, sebagai fungsi skrining untuk memasuki suatu tempat atau área. Melalui aplikasi peduli lindungi, seseorang dapat diperiksa status vaksinasinya, hasil tes COVID-19 atau apakah ada kontak erat dengan pasien COVID-19.

Terdapat 6 sektor yang menjadi fokus pemanfaatan aplikasi Peduli Lindungi dalam hal skrining, yaitu (COVID-19, 2022):

a. Perdagangan (pusat perbelanjaan, pasar modern, dan pasar tradisional)
b. Transportasi (darat, laut, udara)
c. Pariwisata (hotel, restorant, event/pertunjukkan)
d. Kantor/pabrik (pemerintah, swasta, bank, pabrik besar, UMKM/IRT)
e. Keagamaan (masjid, gereja, wihara, pura, kegiatan keagamaan)
f. Pendidikan (PAUD, SD, SMP, /SMA, Perguruan Tinggi)




Gambar 2.2 *Scan QR* Aplikasi Peduli Lindungi

Gambar 2.2 merupakan salah satu contoh penerapan Penggunaan aplikasi peduli lindungi dalam penerapan protokol kesehatan di pusat perbelanjaan/mall. Pengunjung wajib *scan* barcode untuk *check in* dengan aplikasi tersebut, kemudian pengecekan suhu badan, serta mendapatkan barcode sesuai riwayat vaksinasi dan test COVID-19. Saat ini terdapat 28.627.905 pengguna yang telah mengunduh aplikasi PeduliLindungi lewat *Apps Store* dan *Google Play Store*. Pemerintah melalui Kemenkominfo akan terus meningkatkan performa dari peduli lindungi

agar masyarakat tidak menemukan kendala penggunaan, dapat mempermudah serta memberikan rasa aman bagi masyarakat ketika beraktivitas.

### 2.3. *Google Play Store*

*Google play store* merupakan layanan distribusi digital yang dioperasikan dan dikembangkan oleh google berfungsi sebagai tool aplikasi resmi untuk sistem informasi android, yang memungkinkan pengguna untuk menelusuri dan mengunduh aplikasi yang dikembangkan dengan *software development kit* (SDK) dan diterbikan oleh google.

Menurut (jatimtech, 2021) ada beberapa jenis layanan *google play store* diantaranya sebagai berikut:

a. Google play books
b. Google play film & TV
c. Musik
d. Google play apps and games

### 2.4. Data Mining

Data mining adalah suatu proses menemukan hubungan, pola dan kecenderungan dengan memeriksa data dalam ukuran besar menggunakan suatu teknik seperti teknik *statistic* atau matematika (Thi Bi Dan, Widya Sihwi and Anggrainingsih, 2016). Data mining adalah inti dari Knowledge *Discovery in Database* (KDD) yang melibatkan penyimpulan algoritma dalam ekstrasi data, mengembangkan model, dan menemukan pola yang belum diketahui sebelumnya (Chailes, Hermawan and Kurnaedi, 2020). Data

mining adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak di ketahui secara manual (Faisal, 2019) Berdasarkan definisi tersebut maka dapat disimpulkan bahwa data mining merupakan studi untuk menganalisis data dalam ukuran besar secara otomatis yang bertujuan untuk menemukan korelasi, pola dan tren baru yang memiliki makna atau arti menggunakan teknologi pengenalan pola serta teknik *statistic* atau matematika.

Dalam data mining, pengelompokan data juga dilakukan. Tujuannya adalah agar penulis dapat mengetahui pola dan tindak lanjut yang diambil. Semua hal tersebut bertujuan untuk mendukung kegiatan evaluasi agar sesuai dengan yang diharapkan. Menurut (Thi Bi Dan, Widya Sihwi and Anggrainingsih, 2016) pengelompokkan data mining terbagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu sebagai berikut:

a. Deskripsi

Hasil dari model data mining yang berguna untuk menggambarkan pola dan kecenderungan yang terdapat dalam data.

b. Estimasi

Model menggunakan *record* lengkap yang menyediakan nilai dari *variable* target sebagai nilai prediksi. Estimasi diperlukan contohnya untuk menentukan berapa lama proyek terselesaikan dan berapa biayanya.

c. Prediksi

Metode yang digunakan untuk memprediksi nilai yang akan dicapai pada satu periode, noise data dan nilai pada periode sebelumnya dijadikan dasar bahan prediksi.

d. Klasifikasi

Proses menemukan kesamaan karakteristik dalam suatu kelompok atau kelas *(class)*. Bertujuan untuk memperkirakan kelas dari suatu objek yang labelnya belum diketahui.

e. Pengklusteran

Metode *segmentation* atau Pengelompokkan suatu class kedalam beberapa segmen berdasarkan atribut yang ditentukan.

f. Asosiasi

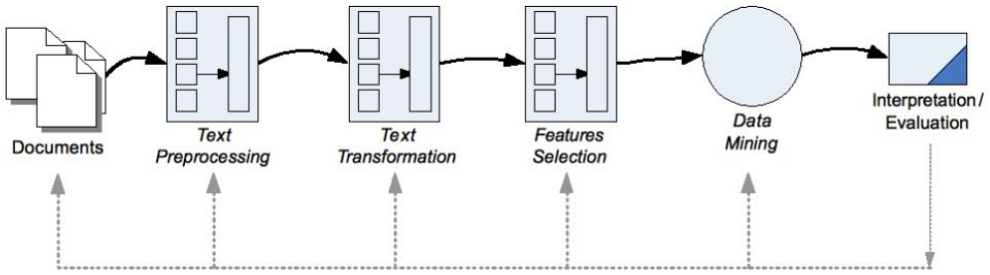
Asosiasi atau *Market basket analysis* (analisa keranjang pasar) berhubungan dengan pemasaran, metode ini bertujuan untuk mengidentifikasi produk yang sering dibeli dalam satu waktu atau bersamaan oleh pelanggan.

## 2.5. *Text Mining*

*Text mining* adalah proses menambang teks untuk menemukan suatu informasi yang berguna dalam koleksi dokumen teks sehingga diperoleh pola, tren, atau keterhubungan antar teks. Hasil *text mining* berupa teks, yang pada umumnya terdapat noise yang tinggi maka perlu dilakukan *text preprocessing* (Sodik and Kharisudin, 2021). Permasalahan yang dihadapi *pada text mining* umumnya sama dengan permasalahan yang terdapat pada data mining, yaitu jumlah data yang besar, dimensi yang tinggi, struktur data yang dapat berubah (tidak konsisten), dan data *noise*. hal tersebut menjadi suatu tantangan untuk ditangani melalui *text mining*. Diharapkan melalui proses *text mining*, informasi yang ada dapat dikeluarkan secara jelas di dalam

teks tersebut dan dapat dipergunakan dalam proses analisis menggunakan alat bantu computer.

Proses *text mining* dikategorikan kedalam *framework* "*Knowledge Discovery in Database (KDD)* yaitu proses mengidentifikasi *pattern* di dalam data yang benar, unik, berguna dan dimengerti. Tahapan KDD (Pajri, Umaidah and Padilah, 2020) sebagai berikut:




Gambar 2.3 Tahapan *KDD*

a. *Documents*

Document dapat berupa text merupakan fragmen yang dianggap sebagai unit. Documents dapat berupa buku, paragraph, abstrak, maupun judul.

b. *Text Prepocessing*

Tahap ini bertujuan untuk mengurangi atribut yang kurang berpengaruh terhadap proses klasifikasi. Secara umum *text preprocessing* terdiri dari *case folding*, *cleaning*, *convert emoticon*, *tokenizing*, *stopword removal*, dan *stemming*.

c. *Text Transformation*

Sebuah dokumen diwakili oleh fitur (kata-kata) yang dikandungnya. Dokumen harus ditransformasikan dari versi teks lengkap menjadi bentuk vektor dokumen.

d. *Feature Selection*

Sebuah dokumen diwakili oleh fitur (kata-kata) yang dikandungnya. Dokumen harus ditransformasikan dari versi teks lengkap menjadi bentuk vektor dokumen.

e. *Data Mining*

Tahap yang digunakan pada text mining sama seperti pada data mining. Misalnya klasifikasi, asosiasi dan lain-lain.

f. *Interpretation / evaluation*

Evaluasi dilakukan terhadap hasil dari *pattern discovery*, umumnya menggunakan suatu nilai performansi.

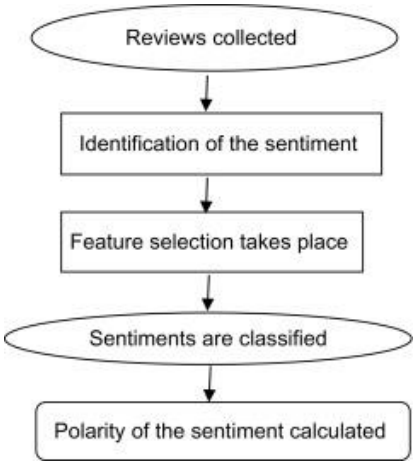
## 2.6. *Sentiment Analysis*

*Sentiment analysis* umumnya merupakan proses klasifikasi teks yang berfokus pada ulasan yang mengungkapkan apakah ulasan tersebut negatif atau positif yang berperan penting dalam memahami minat pengguna. Beberapa definisi mengenai *sentiment analysis* menyatakan sebagai berikut:

a. Riset komputasional dari opini, sentiment, dan emosi yang diekspresikan secara tekstual (Liu, 2011).
b. Bidang ilmu dengan pendekatan Penambangan Teks *(Text Mining)* yang dapat mengekstraksi teks untuk mendapatkan sentimen bahkan emosi seseorang (Irianandа *et al.*, 2021)

c. Proses komputasi dengan memahami, mengekstrak dan mengolah data tekstual secara otomatis untuk mendapatkan sebuah informasi sentimen yang terdapat dalam suatu kalimat opini atau pendapat, bertujuan untuk menyediakan informasi berharga yang terkandung dari sebuah dataset yang tidak terstruktur (Herlinawati *et al.*, 2020)
d. Proses eksplorasi ekstensif data yang disimpan di Web untuk mengidentifikasi dan mengkategorikan pandangan yang diungkapkan dalam bagian teks (Aqlan, Manjula and Lakshman Naik, 2019)

Berikut ini gambaran proses pada *Sentiment Analysis*:




Gambar 2.4 *The Process Of Sentiment Analysis* (Chakraborty *et al.*, 2018)

Gambar 2.4 merupakan gambaran proses Analisis Sentiment, dimulai dengan pengumpulan review data kemudian pendeteksian sentimen, langkah selanjutnya yaitu *feature selection* atau pemilihan fitur ekstrasi, kemudian pengklasifikasian sentimen dan hasil akhir yaitu didapatkan perhitungan polaritas sentimen. Pengklasifikasian *Sentiment Analysis* dapat diselesaikan dengan beberapa metode machine learning,

seperti Naive Bayes Classifier (NBC), K-nearest neighbor (KNN), Support vector machine (SVM), Random forest, dan Maximum Entropy.

## 2.7. *Text Preprocessing*

*Text preprocessing* adalah suatu proses pengubahan bentuk data yang belum terstruktur menjadi data yang terstruktur sesuai dengan kebutuhan, untuk proses text mining lebih lanjut (*analysis sentiment*, peringkasan, clustering dokumen. *Text preprocessing* diperlukan untuk menyelesaikan beberapa jenis masalah seperti *missing value, data redudant, outlier* ataupun format data yang tidak sesuai dengan sistem. Berikut ini tahapan dalam proses *preprocessing* adalah sebagai berikut (Tanggu Mara, Sediyono and Purnomo, 2021):

1. *Cleaning*

Data *cleaning* atau tahapan membersihkan data, tahap tersebut meliputi pengisian data yang hilang agar data nya konsisten, juga memperhalus *noisy* data, bertujuan agar data dalam keadaan bersih. Hal tersebut berimbas pada tingkat akurasi mining yang tinggi. Banyak hal yang bisa menyebabkan kurangnya akurasi data baik itu kesalahan computer atau manusia. Data yang tidak konsisten, data ganda, data *noisy* merupakan salah satu penyebab kurangnya akurasi mining dikarenakan tahap *cleaning* yang tidak baik.

2. *Case Folding*

*Case folding* merupakan proses untuk mengkonversi teks ke dalam format kecil *(lowercase)*. Hal ini bertujuan untuk memberikan bentuk standar pada teks

Contoh:

**ulasan** = ‘Data sudah masuk semua tp tdk ditemukan sertifikat. Sudah coba lewat web sama aja. Jd bagaimana nih sistem nya kok ga bagus’

**Output** = ‘data sudah masuk semua tp tdk ditemukan sertifikat. sudah coba lewat web sama aja. jd bagaimana nih sistem nya kok ga bagus’.

3. Tokenisasi

Tokenisasi atau disebut juga tahap *Lexical Analysis* adalah proses pemotongan teks menjadi bagian-bagian yang lebih kecil, yang disebut *token*. Pada proses ini juga dilakukan penghilangan angka, tanda baca dan karakter lain yang dianggap tidak memiliki pengaruh terhadap pemrosesan teks.

4. *Stopword Removal*

Penghapusan kata-kata yang terlalu umum dan tidak memiliki (atau sedikit) arti, ciri-ciri pada kata ini adalah frekuensi kemunculannya yang jumlahnya cukup banyak dibandingkan dengan kata yang lainnya, contoh kata: aku, kamu, dengan, yang dan lain-lain.

5. *Stemming*

*Stemming* merupakan proses pengubahan bentuk kata menjadi kata dasar atau tahap mencari root dari tiap kata.

6. *Convert negation*

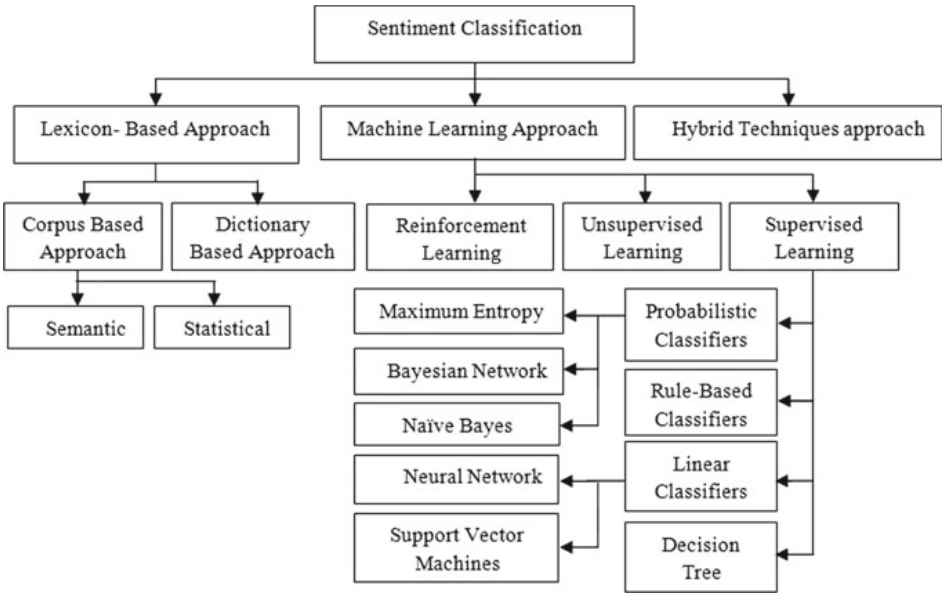
Pada tahap ini kata yang mengandung kata negatif dapat mengubah orientasinya opini seperti “tidak baik” sama dengan “buruk”

Tidak ada tahapan baku mengenai *text preprocessing* sendiri, proses proses tersebut memiliki manfaat untuk memperlancar proses *machine learning* sehingga data lebih

mudah dimengerti mesin, beban representasi dalam data dapat berkurang, dan mempermudah analisis datanya.

## 2.8. Klasifikasi dengan *Machine Learning*

Metode analisis sentimen untuk klasifikasi data terbagi menjadi beberapa teknik utama seperti dilihat pada gambar 2.5 di bawah ini: (Aqlan, Manjula and Lakshman Naik, 2019)




Gambar 2.5 Sentiment Classification Techniques (Aqlan, Manjula and Lakshman Naik, 2019)

Gambar 2.5 merupakan teknik-teknik yang yang digunakan untuk klasifikasi sentimen dimana terbagi kedalam 3 teknik utama yaitu pendekatan *lexicon-based approach,* pendekatan *machine learning,* dan pendekatan *hybrid approach.* Manfaat skema klasifikasi memungkinkan untuk mengidentifikasi jenis dan komentar untuk

dianalisa. Analisa ini dapat bersifat kualitatif untuk menyelidiki opini pengguna ataupun kuantitatif untuk analisis sentimen.

Pendekaan *lexicon-based* menggunakan berbagai kata yang dinilai dengan skor polaritas untuk mengetahui tanggapan/ opini seseorang mengenai suatu topik. Pada umumnya menggunakan kamus *(Dictionary Based Approach)* untuk menggambarkan popularitas (positif, negatif, netral). Pendekatan dengan corpus dan kamus termasuk kedalam metode *Lexicon-based* (Thomas, Yuliana and Noviyanti. P, 2021). Pendekatan *machine learning* Digunakan untuk memecahkan masalah yang berkaitan dengan klasifikasi teks yang mengandung fitur sintaksis atau *linguistic*. Teknik pembelajaran mesin ini terbagi kedalam 3 jenis yaitu:

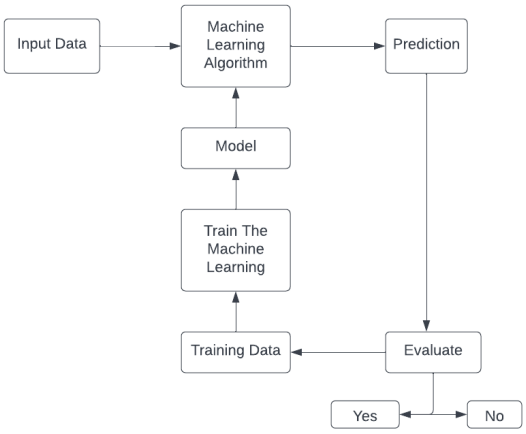
*a.* *Supervised Learning,* pembelajaran dengan menggunakan algoritma yang bertipe klasifikasi. Data data yang digunakan dalam algoritma ini merupakan data yang berlabel artinya algoritma mengidentifikasi fitur secara eksplisit dan melakukan prediksi atau klasifikasi yang sesuai. *Supervised learning* sendiri memiliki metode atau algoritma klasifikasi yaitu meliputi Naïve Bayes, Regresi Linier Sederhana, Decision Tree, K-Nearest Neighbor, Linier Discriminant Analysis (LDA) Dan SVM.

*b.* *Unsupervised Learning,* pembelajaran dengan menggunakan algoritma yang bertipe *clustering.* dalam metode ini Berbeda dengan *supervised Learning* data yang digunakan tidak memerlukan label atau tujuan, nantinya metode ini memungkinkan model untuk belajar sendiri menggunakan data yang telah diberikan dan algoritma akan membantu model dalam membentuk klaster dari jenis

data yang serupa. Model yang dimaksud akan membentuk klaster berdasarkan kesamaan fitur. Algoritma K-Means, EM Klastering dan Spektral Klastering merupakan salah satu metode pembelajaran *Unsupervised Learning.*

c. *Reinforcement Learning,* proses membuat model untuk belajar bagaimana membuat keputusan, model ini biasa digunakan dengan variasi dari teknik learning lainnya. Sebagai contoh dibuat sebuah robot dengan reinforcement learning untuk mengambil barang dari satu tempat ketempat lain, robot ini akan dilatih untuk menghafalkan objek dan melakukan pekerjaan dengan kecepatan dan presisi yang akurat.

Pendekatan *machine learning* merupakan inferensi terhadap data dengan pendekatan matematis (matematika dan statistika). Inti dari ML ini adalah untuk membuat model sistematis yang merefleksikan pola-pola data (Thomas, Yuliana and Noviyanti. P, 2021).




Gambar 2.6 Algoritma *Machine Learning*

Pada proses diatas menggunakan algoritma *machine learning* sebagai penerapan teknik statistika, dimulai dengan mempersiapkan data kemudian melatih sebuah model dengan algoritma tertentu. Untuk memastikan model yang terbentuk data akan dibagi menjadi data pembelajaran *(data training)* dan data pengujian *(data testing),* kemudian dilakukan evaluasi model dan terakhir yaitu meningkatkan kinerja *machine*.

Metode terakhir yaitu *Hybrid Approach* Merupakan gabungan 2 metode untuk mendapatkan klasifikasi sentimen yang lebih. Pendekatan ini menggabungkan *knowledge-based approach* dan *machine learning approach*. Beberapa penelitian berhasil mengaplikasikan keduanya secara bersamaan.

## 2.9. *Term Frequency-Inverse Document Frequency* (TF-IDF)

Setelah tahap *preprocessing,* selanjutnya dilakukan proses pembobotan data berupa kata menjadi numerik menggunakan Metode *Term Frequency - Inverse Document Frequency* (TF-IDF). Proses pembobotan dilakukan berdasarkan jumlah kemunculan suatu kata *(term)* dalam sebuah dokumen dan dalam keseluruhan dokumen.

Berikut ini merupakan beberapa metode pembobotan menggunakan TF-IDF (Fazar *et al.,* 2020)

1. *Term Frequency (TF)* merupakan frekuensi kemunculan kata pada suatu dokumen teks.semakin besar nilai TF nya , maka semakin besar pula nilai bobot nya. berikut ini merupakan persamaan TF:

$$\mathbf{TF_{t,d} = f_{(t,d)}} \qquad \mathbf{(2.1)}$$

**Keterangan:**

TF: Frekuensi kemunculan kata pada suatu dokumen teks.

*f:* jumlah kata pada suatu dokumen.

2. *Invers Document Frequency (IDF)* menunjukkan hubungan ketersediaan sebuah term dalam seluruh dokumen. Semakin sedikit TF nya, maka nilai IDF semakin besar. Berikut ini merupakan persamaan IDF:

$$\boldsymbol{IDF = Log\ \frac{N}{DFt}} \qquad \mathbf{(2.2)}$$

**Keterangan:**

N : jumlah dokumen

DFt : nilai TF

3. *Term Frequency – Inverse Document Frequency* (TF-IDF) merupakan penggabungan dari formula perhitungan raw TF dan formula IDF dengan cara mengalikan nilai *Term Frequency (TF)* dengan nilai *Inverse Document Frequency (IDF)* digunakan untuk menganalisa hubungan antara sebuah dokumen dan metode untuk menghitung bobot setiap kata yang paling umum digunakan. Berikut ini merupakan persamaan TF-DF:

$$\mathbf{WT = TF\ x\ IDF} \qquad \mathbf{(2.3)}$$

**Keterangan:**

Wt : TF-IDF

TF: Nilai TF

IDF: Nilai IDF

## 2.10. *Synthetic Minority Oversampling Technique* (SMOTE)

Ketidakseimbangan data terjadi jika jumlah objek suatu kelas data lebih banyak dibandingkan kelas lain, kelas data yang objeknya lebih banyak disebut kelas mayor sedangkan lainnya disebut kelas minor (Azmatul Barro, Sulvianti and Afendi, 2013). Ketidakseimbangan tersebut berpengaruh terhadap pembuatan model yang diperoleh, sehingga pengolahan algoritma yang tidak menghiraukan ketidakseimbangan data akan cenderung diliputi oleh kelas mayor dan mengacuhkan kelas minor. Teknik SMOTE berguna untuk menangani masalah ketidakseimbangan tersebut dikarenakan pengklasifikasian dari dataset tidak di presentasikan secara merata, ia bekerja dengan cara menambah jumlah data kelas minor agar setara dengan kelas mayor dengan teknik membangkitkan data buatan.

Beberapa algoritma yang dipadukan dengan metode SMOTE telah dibuktikan dapat menghasilkan kinerja pengklasifikasian yang lebih baik (Chohan *et al.*, 2020)

## 2.11. Algoritma Naïve Bayes Classifier

Naïve Bayes Classifier merupakan salah satu metode yang banyak digunakan untuk melakukan *sentiment analysis* dengan menggunakn teknik prediksi yang berbasis probabilitas berdasarkan kategori yang ada pada data latih. Klasifikasi ini

menggabungkan pengetahuan sebelumnya dengan pengetahuan baru (Gunawan, Fauzi and Adikara, 2017)

Proses klasifikasi Naïve Bayes secara umum dapat dilihat pada persamaan berikut ini:

$$P(c_j \mid w_i) = \frac{P(c_j)\ x\ P(w_i \mid c_j)}{P(w_i)} \quad \textbf{(2.4)}$$

**Keterangan:**

$P(c_j \mid w_i)$: Posterior merupakan peluang kategori j ketika terdapat kemunculan kata i

$P(w_i \mid c_j)$:Conditional probability merupakan peluang sebuah kata i masuk ke dalam kategori j

$P(c_j)$ : Prior merupakan peluang kemunculan sebuah kategori j

$P(w_i)$ : Peluang kemunculan sebuah kata

I : Indeks kata yang dimulai dari 1 hingga kata ke-k

J : Indeks kategori yang dimulai dari 1 hingga kategori ke-n

Peluang kemunculan kata sesungguhnya dapat dihilangkan pada proses perhitungan klasifikasi karena peluang tersebut tidak akan berpengaruh pada perbandingan hasil klasifikasi dari setiap kategori. Proses pada klasifikasi dapat disederhanakan dengan Persamaan berikut:

$$P(c_j \mid w_i) = P(c_j) x P(w_i \mid c_j) \quad \textbf{(2.5)}$$

Untuk menghitung prior atau peluang kemunculan suatu kategori pada semua dokumen dapat dilakukan dengan menggunakan Persamaan berikut:

$$P(cj) = \frac{N_{cj}}{\mathbf{N}} \quad \textbf{(2.6)}$$

**Keterangan :**

$N_{cj}$ : Dokumen yang termasuk kategori CJ

N : Jumlah keseluruhan dokumen latih yang digunkan

Pada umumnya data uji memiliki banyak kata yang diproses mulai indeks ke-1 hingga ke *k*, dalam hal *ini conditional probability* kata $w_i$ pada kategori $c_j$ dilakukan perkalian dari $i$=1 sampai $i$=k sehingga untuk mengetahui nilai posterior dapat dihitung dengan menggunakan Persamaan berikut:

$$\mathbf{P}(c_j\,|w_i) = P(c_j)xP(w_1\,|c_j)x \ldots xP(w_k|c_j) \quad \textbf{(2.7)}$$

## 2.12. Evaluasi dan Model Klasifikasi

Evaluasi kinerja model klasifikasi sangat penting dilakukan. Hal tersebut menunjukkan model klasifikasi atau prediksi yang telah dibuat sesuai dengan yang diharapkan atau tidak. *Confusion matrix* merupakan alat pengukuran yang dapat digunakan untuk menghitung kinerja atau tingkat kebenaran suatu klasifikasi. Dengan *confusion matrix* dapat dianalisa seberapa baik classifier dapat mengenali record dari kelas-kelas yang berbeda.

*Confusion matrix* ditunjukkan pada tabel 2.1 sebagai berikut:

Tabel 2.1 *Confusion Matrix*

| Confusion Matrix | Prediction Positif | Prediction Negatif |
|---|---|---|
| Actual Positif | TP | FN |
| Actual Negatif | FP | TN |

Dimana:

- TP (True Positif), banyaknya data yang kelas aktualnya adalah positif dengan kelas prediksinya merupakan positif.
- FN (False Negatif), banyaknya data yang kelas aktualnya adalah positif dengan kelas prediksinya adalah negatif.
- FP (False Positif), banyaknya data yang kelas aktualnya adalah negatif dengan kelas prediksinya adalah positif.
- TN (True Negatif), banyaknya data yang kelas aktualnya negatif dengan kelas prediksinya adalah negatif.

Hasil *confusion matrix* digunakan untuk mengukur *Accuracy*, *Precision, Recall* dan *F1-Score* untuk menganalisa kinerja algoritma dalam melakukan klasifikasi (Zulqornain and Adikara, 2021) berikut penjelasannya:

a. *Accuracy,* merupakan rasio prediksi benar (positif dan negatif) dengan keseluruhan data. *Accuracy* dapat dihitung dengan persamaan berikut:

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP + TN}}{\boldsymbol{TP + TN + FP + FN}} \quad \textbf{(2.8)}$$

b. *Precission,* merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang di prediksi positif. *Precission* dapat dihitung dengan persamaan berikut:

$$\boldsymbol{Precission} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FP}} \quad \textbf{(2.9)}$$

c. *Recall,* merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif. *Recall* dapat dihitung dengan persamaan berikut:

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FN}} \quad \textbf{(2.10)}$$

d. *F1- Score,* merupakan perbandingan rata-rata *precission* dan *recall* yang dibobotkan. *F1-Score* dapat dihitung dengan persamaan berikut:

$$\boldsymbol{F1 - Score} = \frac{\boldsymbol{2\ (Precision\ \times Recall)}}{\boldsymbol{(Precision + Recall)}} \quad \textbf{(2.11)}$$

## 2.13. Penelitian Terkait

Penelitian terkait yang dipakai sebagai acuan guna menunjang penelitian ini di paparkan pada tabel 2.2:

Tabel 2.2 Penelitian Terkait

| No | Peneliti | Judul | Dataset | Praprocessing | Fitur | Metode Klasifikasi | Teknik Pengujian |
|---|---|---|---|---|---|---|---|
| 1 | (Zulqornain and Adikara, 2021) | Analisis Sentimen Tanggapan Masyarakat Aplikasi Tiktok Menggunakan Metode Naïve Bayes Dan *Categorial Propotional Difference* (CPD) | Dataset ulasan dari *Google Play Store* mulai rating 1 sampai 5 berjumlah 1000 data dan memiliki dua kelas | *Case floding, cleaning,* tokenisasi, *filtering* dan *stemming* | *Categorial Propotional Difference* (CPD) | Naïve Bayes | *5-fold Cross Validation* |
| 2 | (Fahlapi *et al.*, 2022) | Analisia sentimen vaksinasi covid-19 dengan metode Support Vector Machine dan Naïve Bayes berbasis teknik SMOTE | Data *review* pengguna twitter diambil sejumlah 1.013 data sentimen. | *Remove Duplicate, Nominal to text, Transform Case, Filter Token and filter stop word* | | Naïve bayes dan support vector machine dengan optimasi teknik SMOTE dan PSO (Particle Swarm Optimization) | *Cross Validation (AUC)* |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 3 | Fitri, 2020 (Fitri, 2020) | Analisis Sentimen Terhadap Aplikasi Ruangguru Menggunakan Algoritma Naive Bayes, Random Forest Dan Support Vector Machine | Dataset didapat dari *review* Ruangguru di *Google Play Store* menggunakan teknik *scrapping* berjumlah 1629 data | *Tokenize, stemming, case folding,* dan *Stopword Removal* | | Naïve Bayes, Random Forest, dan Support Vactor Machine | *K-fold Cross Validation* |
| 4 | (Surohman *et al.*, 2020) | Analisa Sentimen Terhadap *Review* Fintech Dengan Metode Naive Bayes Classifier Dan K- Nearest Neighbor | 232 data *review* yang terdiri dari 116 data review positif dan 116 data review negatif | *Tokonize, Transform Case, Stopword (Dictionar)* | *Seleksi fitur data mining* | Naïve bayes, dan K-Nearest Neighbor | *Confusion Matrix + ROC/AUC* |
| 5 | (Hakim *et al.*, 2020) | Sentimen Analysis *Stay Home* Menggunakan Metode Klasifikasi Naïve Bayes, Support Vector Machine, Dan K-Nearest Neighbor | 1652 *tweets* kurun waktu 1 maret 2020- 1 april 2020. | *Regex Removal, Remove URL, Remote Annotation, Remove Duplicate Tweets, Extract Sentiment, Transform Case* | | Naïve bayes (NB), Support vector mavhine (SVM) dan K-Nearest *Neighbor (k-NN)* | *Confusion Matrix* |
| 6 | (Puspita and Widodo, 2021) | Perbandingan Metode KNN, | 1000 data *tweet* | *Data Validation, Data Integration dan* | | KNN, Decision | *Cross Validation* |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | Decision Tree, dan Naïve Bayes Terhadap Analisis Sentimen Pengguna Layanan BPJS | | *Transformation, Data Size Reduction and Dicretization* | | Tree, dan Naïve bayes | |
| 7 | (Putra and Nugroho, 2021) | Perbandingan Performa Naïve Bayes Dan KNN Pada Klasifikasi Teks Sentimen Jasa Ekspedisi | Data JNT sebanyak 46.220, JNE 5.364 dan Pos Indonesia 11.194 | *Scraping* Data, *Pre Processing (Case Folding, pembersihan data, tokenisasi, normalisasi, stopword removal, dan stemming)* Data, Pelabelan Data,Pembobotan Kata, Evaluasi Model, dan Pengeahuan | TF-IDF | Naïve Bayes, dan KNN | *Confusion Matrix* |
| 8 | (Chohan *et al.*, 2020) | Analisis Sentimen Aplikasi Duolingo Menggunakan Metode Naïve Bayes dan Synthetic Minority Over Sampling Technique | Ulasan dari *Google Play Store* | *Tokenizing,Transform Cases, Stemming, Stopword* | | Naïve Bayes dan Synthetic Minority Over Sampling Technique (SMOTE), | *10 Cross Validation* |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 9 | (Saraswati and Riminarsih, 2020) | Analisis Sentimen Terhadap Pelayanan Krl *Commuterline* Berdasarkan Data Twitter Menggunakan Algortima Bernoulli Naive Bayes | 8.130 *Tweets* | *Case Folding, Filtering, Stopword Removal dan Stemming* | *Bag Of Words* | Naïve Bayes | *Confusion Matrix* |
| 10 | (Winahyu and Suharjo, 2021) | Aplikasi Web analisis sentimen dengan algoritma multinomial Naïve Bayes | 1000 *Tweets* | Pembersihan data, *remove duplicates, select attribte, SubProcessing,* | | Naïve bayes | *Confusion matrix* |
| 11 | (Locarso, 2022) | Analisis Sentimen Review Aplikasi Peduli lindungi Pada Google Play Store Menggunakan Nbc | Ulasan dari *Google Play Store* | *Cleaning, case folding, tokenizing,stemming,* dan *filtering* | TF-IDF | Naïve bayes | *Confusion matrix* |
| 12 | (Mustopa *et al.*, 2020) | Analysis of User Reviews for the PeduliLindungi Application on Google Play Using the Support Vector Machine | 1360 data di dapat dari ulasan google play kurun waktu april 2020- juni 2020 | *Data preparations (remov duplicates, nominal to text, transform case, filter token (By Length), filter stopword (Dictionary),* | TF-IDF | *Svm* dan *Naïve Bayes* dengan optimasi PSO | *(Confusion Matrix) and AUC (Area Under Curve) value* |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | and Naive Bayes Algorithm Based on Particle Swarm Optimization | | | | | |
| 13 | (Tanggu Mara, Sediyono and Purnomo, 2021) | Penerapan Algoritma K-Nearest Neighbors Pada Analisis Sentimen Metode Pembelajaran Dalam Jaringan (DARING) Di Universitas Kristen Wira Wacana Sumba | Pengumpulan dataset: komentar pada grup facebook sebanyak 200 data dikelompokkan dengan 2 opini yaitu positif dan negatif. | *Case folding, Tokenizing, stopwords removal, dan stemming,* | TF-IDF | K-Nearest Neighbors | *10 Cross Validation* |
| 14 | (Herlinawati *et al.*, 2020) | Analisis sentimen zoom cloud meetings di play store menggunakan naïve bayes dan support vectore machine | 1.007 record dengan label positif dan label negatif. | *Regex removal, remove URL, annotion removal, remove number, tokenizing, stemming, transform case, filter stopwords, filter tokens (By Length), labelling* | | Naïve bayes dan support vector machine | *10 fold cross validation* |
| 15 | (Wisnu, Afif and Ruldevyani, 2020) | Sentiment analysis on customer | Dataset didapat dari jumlah *Tweets* | Pembersihan data proses, transformasi data *(case folding ,* | | K-Nearest Neighbors | *n-fold cross-validation* |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | satisfaction of digital payment in Indonesia: A comparative study using KNN and Naïve Bayes | @gopayindonsia adalah 2.095, 4,001 dari @ovo_id dan 31.541 dari @linkaja | *stopword list),* reduksi data *(stemming dan tokenization)* | | dan Naïve Bayes | |
| 16 | (Andreyestha, 2022) | Analisa Sentimen Kicauan Twitter Tokopedia Dengan Optimalisasi Data Tidak Seimbang Menggunakan Algoritma SMOTE | Dataset berjumlah 328 data dengan kata kunci "tokopedia" yang terdiri dari 200 ulasan positif dan 128 ulasan negatif. | *Remove numbers, Remove punctuation, Remove stop words, Strip whitespaces, Convert Lower Case, stemming* | Seleksi fitur data mining | Random Forest dan Naïve Bayes Dengan optimasi Teknik Synthetic Minority Oversampling Technique (SMOTE) | *Confusion matrix* |
| 17 | (Rezki *et al.*, 2020) | Analisis *Review* Pengguna Google Meet Dan Zoom Cloud Meeting Menggunakan Algoritma Naïve Bayes | Data diambil dalam kurun waktu 3 bulan yakni (maret 2020 sampai mei 2020) sejumlah 714 data dikelompokkan berdasarkan skor penilaian bintang yang diberikan oleh pengguna | *Gata framework (stopword, stemming, tokenization),* kemudian *transform case, filtering* | | Naïve bayes dan *Synthetic minority over-sampling technique* (SMOTE) dengan optimasi PSO | *Cross Validation (AUC)* |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | *Google Meet* dan *Zoom Cloud Meeting* | | | | |
| 18 | (Inelza and Kharisudin, 2022) | Analisis Sentimen Pengguna Aplikasi Marketplace Tokopedia Pada Situs *Google Play Store* Menggunakan Metode Support Vector Machine (SVM), Naïve Bayes, dan Logistic Regression | Jenis data yaitu data primer, berupa *review* yang terdiri dari sentimen positif dan negatif. Diperoleh 3125 *review* dengan jumah *review* positif sebanyak 2598 dan *review* negatif sebanyak 527. | *Spelling normalization, case folding, tokenizing, filtering* | TF-IDF | Support Vector Machine (SVM), Naïve Bayes, dan Logistic Regression | *Split data* dan *k-fold cross-validation* |
| 19 | (Hendra, 2021) | Analisis sentimen *review* halodoc menggunakan Naïve Bayes Classifier | Dataset berasal dari ulasan *google play store* review sebanyak 950 data sentimen. | *Cleansing (remove duplicate), tokenize, transform to lower case, filter stopword (by dictionary), filter token (by length)* | TF-IDF | Naïve Bayes Classifier | *Confusion matrix* |
| 20 | (Flores *et al.*, 2018) | An Evaluation of SVM and Naïve Bayes With SMOTE on | Dataset Tweet Administrasi Duterte dari Twitter sebagai dataset A. dan | - | - | Naïve Bayes Classifier dan Support | *k-fold cross-validation* |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | Sentiment Analysis Data Set | Dampak Program K-12 dalam pesan filipina dari Facebook sebagai dataset B. | | | Vector Machine (SVM), | |
| 21 | (Ardianto *et al.*, 2020) | Sentiment Analysis One E-Sports For Education Curriculum Using Naïve Bayes And Support Vector Machine | 8453 yang bersumber dari Data Twitter | *Transform case, remove http, remove @, remove #, tokenize, filter token (by length), filter stopwords* | | Naïve Bayes Classifier dan Support Vector Machine (SVM), dan *Synthetic minority over-sampling technique* (SMOTE) | *Confusion matrix* |

Penelitian yang akan di lakukan memiliki keterkaitan dengan penelitian-penelitian sebelumnya mengenai analisis sentimen dari media sosial khususnya dari ulasan *Google Play Store*, penelitian ini bertujuan untuk menganalisis dan memprediksi opini atau pendapat masyarakat mengenai kepuasan terhadap suatu produk yang diklasifikasikan kedalam sentimen positif dan sentimen negatif. Selain itu penelitian ini juga bertujuan untuk menguji perfomansi dari algoritma *Naïve Bayes Classifier* dengan menerapkan *Synthetic Minority Over-Sampling Technique* (SMOTE) untuk mengatasai masalah ketidakseimbangan kelas.

Penelitian yang dilakukan oleh (Pintoko and L., 2018) mengenai "Analisis Sentimen Jasa Transportasi Online Pada Twitter Menggunakan Metode Naïve Bayes Classifier" didapatkan hasil klasifikasi teks menggunakan metode *naïve bayes classifier* mendapatkan akurasi yang cukup tinggi yaitu sebesar 86,80% dimana datanya diklasifikasikan menjadi 2 kelas yaitu kelas positif dan negatif menunjukkan tingkat sentimen positif dari tweet masyarakat lebih besar dibandingkan dengan tingkat sentimen negatif.

Komparasi algoritma juga telah dilakukan oleh (Putra and Nugroho, 2021) dalam penelitiannya berjudul "Perbandingan Performa *Naïve Bayes* Dan KNN Pada Klasifikasi Teks Sentimen Jasa Ekspedisi" Dataset yang digunakan didapatkan dari ulasan pengguna Twitter pada akun *@jntexpressid*, *@JNE_ID*, dan *@posindonesia* sebanyak 4.732 dan data uji sebanyak 2037 dengan memisahkan data latih sebanyak 70% dan data uji sebanyak 30%. Tingkat akurasi pada algoritma Naïve Bayes sebesar 82% lebih tinggi dibadingkan dengan algortitma KNN dengan akurasi 71%. Pada KNN

nilai K tinggi tidak menentukan hasil akurasi, terbukti pada penelitian yang dilakukan nilai K=2 Memiliki akurasi tinggi dibanding K=4, K=6, K=7, K=8, K=10 setelah teknik resampling dilakukan. Hal tersebut membutktikan bahwa algoritma *naïve bayes classifier* memiliki performa nilai akurasi lebih baik daripada algoritma *KNN*.

Penerapan *Synthetic Minority Over-Sampling Technique* (SMOTE) dapat meningkatkan kinerja pada kelas data yang tidak seimbang agar kinerja klasifikasi lebih baik. hal tersebut dapat dibuktikan dalam penelitian (Sulistiyowati and Jajuli, 2020) pada "Integrasi Naïve Bayes Dengan Teknik *Sampling* SMOTE Untuk Menangani Data Tidak Seimbang" secara keseluruhan metode SMOTE umumnya dapat menangani permasalahan *imbalanced data.* Pada penelitian tersebut didapatkan hasil akurasi sebesar 94,015% dan *G-Mean* sebesar 0,948% dengan mengkombinasikan algorima klasifikasi Naïve Bayes dan metode SMOTE.

Penelitian mengenai analisis sentimen *riview* terhadap aplikasi Peduli Lindungi pada *Google Play Store* sudah pernah di lakukan oleh (Locarso, 2022) dengan jumlah dataset 1179 untuk *data training* dengan perbandingan 70:30 untuk *data testing* penelitian tersebut menghasilkan kesimpulan bahwa algoritma *Naïve Bayes* dapat melakukan klasifikasi dengan baik untuk memberikan label positif, negatif dan netral dengan *Accuracy* sebesar 83.3%, *Precission* sebesar 65%, dan *Recall* sebesar 63%.

Berdasarkan penelitian-penelitian yang telah dilakukan sebelumnya, penelitian ini akan menggunakan algoritma *Naïve Bayes Classifier* dengan penambahan Teknik *Synthetic Minority Over-Sampling Technique* (SMOTE) terhadap review aplikasi

Peduli Lindungi sebagai kebaruan dari penelitian yang berfungsi untuk menangani kelas data yang tidak seimbang dan meningakatkan kinerja klasifikasi.

## 2.14. Matriks Penelitian

Penelitian terdekat selanjutnya dibuatkan model matriks ruang lingkup penelitian. Adapun matriks ruang lingkup penelitian dipaparkan dalam Tabel 2.3:

Tabel 2.3 Matriks Penelitian

| No | Peneliti / Tahun | Sumber Data | Lingkup Penelitian | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Metode | | | | | | | Optimasi | | | | | |
| | | | SVM | Naïve Bayes | Bayesian Netework | KNN | Decision Tree | Random Forest | Text Mining | SMOTE | PSO | Big Of Words | Multiple Feature Selection | Unigram |
| 1 | (Mustopa *et al.*, 2020) | Google Play Store | ✓ | ✓ | - | - | - | - | - | - | ✓ | - | - | - |
| 2 | (Fahlapi *et al.*, 2022) | Twitter | ✓ | ✓ | - | - | - | - | - | ✓ | ✓ | - | - | - |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **3** | (Hakim *et al.*, 2020) | Twitter | ✓ | ✓ | - | ✓ | - | - | - | ✓ | - | - | - | - |
| **4** | (Fitri, 2020) | Google Paly Store | ✓ | ✓ | - | - | - | ✓ | - | - | - | - | - | - |
| **5** | (Mailo *et al.*, 2019) | twitter | - | ✓ | - | - | - | - | - | - | - | - | - | - |
| **6** | (Puspita and Widodo, 2021) | Twitter | - | ✓ | - | ✓ | ✓ | - | - | - | - | - | - | - |
| **7** | (Saraswati and Riminarsih, 2020) | Twitter | - | ✓ | - | - | - | - | - | - | - | ✓ | - | - |
| **8** | (Chohan *et al.*, 2020) | Google play store | - | ✓ | - | - | - | - | - | ✓ | - | - | - | - |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **9** | (Tanggu Mara, Sediyono and Purnomo, 2021) | Facebook | - | - | - | ✓ | - | - | - | - | - | - | - | - |
| **10** | (Cahyaningtyas *et al.*, 2021) | Google Play Store | - | - | - | - | ✓ | - | - | ✓ | - | - | - | - |