# BAB IITINJAUAN PUSTAKA

## Penelitian Terdahulu

Penelitian yang dilakukan oleh (Laurensz & Eko Sediyono, 2021) dengan judul “Analisis Sentimen Masyarakat terhadap Tindakan Vaksinasi dalam Upaya Mengatasi Pandemi Covid-19”. Penelitian ini bertujuan untuk mengungkap sentimen publik tentang vaksin. Sebanyak 845 *tweet* data yang digunakan, 2 kata kunci yaitu “vaksin merahputih” dan “vaksin sinovac”. Metode yang digunakan dalam penelitian ini adalah metode SVM dan Naïve Bayes. Rata-rata akurasi Naive Bayes sebesar 85,59% dan SVM sebesar 84,41%. Sentimen Metode Naïve Bayes dengan *keyword* “vaksin sinovac” menghasilkan sentimen positif 66% dan negatif 34%, sedangkan “vaksin merahputih” menghasilkan sentimen yang positif 89% dan negatif 11%. Metode SVM dengan kata kunci “vaksin sinovac” mendapatkan sentimen yang positif 96% dan negatif 4%, sedangkan “vaksin merahputih” menghasilkan sentimen yang positif 98% dan negatif 2%.

Penelitian yang dilakukan oleh (Septiana et al., 2021) dengan judul “Analisis Sentimen Vaksinasi Covid-19 Pada Twitter Menggunakan Naïve Bayes Classifier Dengan Feature Selection Chi-Squared Statistic Dan Particle Swarm Optimization”. Penelitian ini bertujuan untuk menganalisis opini publik terhadap upaya pemerintah terhadap vaksinasi Covid-19 untuk mencapai hasil prediksi dengan tingkat akurasi yang paling optimal. Proses *crawling* dilakukan tools Rapidminer untuk mengambil data tweets yang mengandung 5 *keyword*, yaitu “Vaksin Sinovac”, “Vaksin Astrazeneca”, “Vaksin Moderna”, “Vaksin Merah Putih”, dan “Vaksinasi Covid-19”. Dataset menerima 2.060 *tweet* yang ditandai secara manual. Jumlah kicauan yang diterima sebanyak 1193 kicauan positif, 73 kicauan negatif, dan 794 kicauan netral. Hasil pengujian menunjukan bahwa Algoritma Naïve Bayes Classifier (NBC) tanpa Feature Selection mencapai akurasi 63,69%. Hasil penelitian menunjukkan akurasi sebesar 69,13% untuk Algoritma Naïve Bayes Classifier (NBC) dengan Feature Selection Chi-Squared Statistic. Sedangkan hasil pengujian algoritma Naïve Bayes Classifier (NBC) dengan Particle Swarm Optimization mempunyai tingkat akurasi 66,02%. Oleh karena itu, hasil pemilihan fungsi statistik chi-square memperoleh nilai akurasi yang lebih baik dengan selisih akurasi 3,11% dibandingkan dengan optimasi swarm partikel proses klasifikasi algoritma Naive Bayes Classifier (NBC).

Penelitian yang dilakukan oleh (Lestari & Saepudin, 2021) dengan judul “Analisis Sentimen Vaksin Sinovac pada Twitter Menggunakan Algoritma Naïve Bayes”. Penelitian ini bertujuan untuk memberikan solusi penentuan presentase pendapat positif atau negatif *user* akun media sosial Twitter mengenai vaksin Sinovac. Persentase dan tanggapan menjadi masukan dan penilaian bagi perusahaan dan pemerintah terkait pengadaan vaksin Sinovac. Penelitian ini menggunakan metode Naive Bayes dengan alat RapidMiner untuk mengurutkan opini positif atau negatif tentang vaksin Sinovac dalam tweet berbahasa Inggris. Menurut hasil penelitian ini, hingga 86% tweet memiliki sentimen positif, sedangkan 14% tweet memiliki sentimen negatif. Nilai akurasi untuk perhitungan menggunakan algoritma Naive Bayes adalah 92,96%.

Penelitian yang dilakukan oleh (Aldisa & Maulana, 2022) dengan judul “Analisis Sentimen Opini Masyarakat Terhadap Vaksinasi Booster COVID-19 Dengan Perbandingan Metode Naive Bayes, Decision Tree dan SVM”. Penelitian ini bertujuan untuk melakukan analisis sentimen terkait opini masyarakat terhadap vaksinasi booster COVID-19 di Indonesia menggunakan model Naive Bayes. Penelitian ini menggunakan metode Naive Bayes. Hasil dari peneltian ini menunjukkan skor AUC terbesar jatuh kepada model SVM (75.40%), namun untuk presisi yang lebih akurat jatuh kepada model Naive Bayes (83.81%). Selain itu, terdapat confusion matrix yang menunjukkan bahwa uji coba model Naive Bayes yang dilakukan berjalan dengan baik.

Penelitian yang dilakukan oleh (Fathonah & Herliana, 2021) dengan judul “Penerapan Text Mining Analisis Sentimen Mengenai Vaksin Covid - 19 Menggunakan Metode Naïve Bayes”. Penelitian ini bertujuan untuk melakukan analisis sentimen terhadap Vaksin COVID-19 dengan menggunakan metode Naïve Bayes. Penelitian ini menggunakan metode Naive Bayes. Naïve Bayes dinilai berpotensi baik dalam mengklasifikasi dokumen dibandingkan metode pengklasifikasian yang lain dalam hal akurasi dan efisiensi Berdasarkan hasil pengujian terhadap 100 data latih yang kemudian dipilih kembali dengan menggunakan teknik data crawling menjadi 34 data, didapatkan bahwa analisis sentimen dari pengguna twitter untuk Vaksin COVID-19 ini didapatkan persentase accuracy sebesar 100%.

## Teori Terkait

### Text Mining

Text mining adalah teori pemrosesan kumpulan teks dengan tujuan mengetahui dan mengekstraksi informasi yang berguna dari kumpulan teks ini. Diinformasikan dengan mengidentifikasi dan memeriksa pola yang menarik dari sumber data (Sabrani et al., 2020).

Penambangan teks dapat didefinisikan secara luas sebagai proses interaksi pengguna dengan kumpulan dokumen menggunakan alat analisis. Ketika datang ke penambangan data, penambangan teks adalah proses mengekstraksi informasi yang berguna dari sumber data dengan mengidentifikasi dan mempelajari pola tertentu. Penambangan teks adalah proses mengekstraksi informasi yang berguna dari sumber data dengan mengidentifikasi dan menyelidiki pola tertentu. (Rahman & Doewes, 2017).

Tujuan dari text mining adalah untuk memproses informasi tekstual yang tidak terstruktur, mengekstrak indeks numerik yang bermakna dari teks, dan membuat informasi yang terkandung dalam teks dapat diakses menggunakan berbagai algoritma data mining. (Rahman & Doewes, 2017).

### Text Preprocessing

Data teks media sosial Twitter dapat menghadirkan *chalenge* unik karena memiliki kata tidak baku, bahasa daerah yang tidak termasuk dalam KBBI, atau penggunaan singkatan. Preprocessing diperlukan agar data lebih terstruktur. Proses ini dilakukan untuk membakukan semua bahasa non-standar dan membersihkan data dari gangguan yang ada, membuat data terstruktur dan agar siap dianalisa. (Laurensz & Eko Sediyono, 2021).

Tahapan dalam *text preprocessing*, yaitu sebagai berikut.

1. Transform Cases: Kata-kata yang tidak berhubungan akan diubah. Misalnya, kata-kata yang mengandung huruf besar akan diubah menjadi huruf kecil atau *lower case*.
2. Tokenize: karakter khusus, tanda baca, simbol, atau karakter yang bukan termasuk huruf akan dihapus.
3. Filter Token (*By Length*): Kata-kata yang lebih pendek dari 4 atau lebih panjang dari 25 karakter akan dihapus, misalnya ‘tdk’, ‘yg’, dan ‘gan’.
4. Stopword Removal: Kata yang tidak relevan, misalnya tetapi, dengan, untuk, yang, dan kata sambung lainnya, akan dihapus.
5. Stemming: Kata yang berimbuhan, seperti mem-, me-, meny-, meng-, per-, dan ber-, akan diseleksi dan diubah menjadi kata dasar.

### Analisis Sentimen

Analisis sentimen adalah proses menentukan sentimen dan mengklasifikasikan polaritas teks suatu dokumen atau kalimat sehingga dapat ditentukan kategori sentimen positif dan negatif. Analisis sentimen sekarang menjadi bidang penelitian ilmu komputer yang populer bagi para peneliti. Jejaring sosial seperti Twitter sering digunakan untuk analisis sentimen untuk menentukan persepsi publik.

Analisis sentimen juga dapat dibandingkan dengan penambangan opini karena berfokus pada opini positif atau negatif. Analisis sentimen melibatkan melakukan penambangan data untuk menganalisis, memproses, dan mengekstrak data teks dalam entitas. B. Layanan, produk, orang, fenomena, atau topik tertentu. Proses analitis mungkin melibatkan teks ulasan, forum, tweet, atau blog, termasuk data yang telah diproses sebelumnya seperti tokenization, stopwords, penghapusan, stemming, identifikasi sentimen, dan klasifikasi sentimen (Samsir et al., 2021).

### Naïve Bayes Multinomial Text

Naive Bayes adalah metode stokastik yang pertama kali diperkenalkan oleh ilmuwan Inggris Thomas Bayes. Naive Bayes digunakan untuk memprediksi peluang masa depan berdasarkan pengalaman masa lalu. Menerapkan Naive Bayes ke analisis sentimen membutuhkan dua proses penting: pelatihan dan pengujian. Keuntungan dari Naive Bayes adalah menggunakan lebih sedikit data pelatihan, sehingga komputasi dapat dilakukan lebih cepat dan lebih efisien. Kelemahan Naive Bayes adalah pemilihan fitur yang buruk mengurangi akurasi dan meningkatkan waktu komputasi (Laurensz & Eko Sediyono, 2021).

Langkah pertama yang dilakukan dalam algoritma Naive Bayes Multinomial Text ialah menghitung probabilitas kelas pada dokumen dengan rumus (Purwiantono & Aditya, 2020):

$P(c)= \frac{N(c)}{N}$ (1)

Yang mana :

- P(c) : Probabilitas kelas terhadap dokumen.

- N(c) : Jumlah tweet pada kelas.

- N : Jumlah tweet.

Setelah itu kata unik untuk semua kelas dihitung untuk mencari probabilitas kata dengan rumus:

$p(w|c)=\frac{count\left(w,c\right)+1}{count\left(c\right)+|V|}$ (2)

Yang mana :

 - P(w|c) : Probabilitas kata terhadap class.

- count(w,c) : Jumlah kemunculan kata pada class.

- count(c) : Jumlah kata pada class.

- V : Jumlah kata unik pada dokumen.

- || : Nilai mutlak.

Kemudian probabilitas *tweet* pada kelas dihitung dengan rumus:

$P\left(d(n)\right)=P\left(c\right) x π\prod\_{}^{}P(w|c)$ (3)

Dimana:

- P(c|d(n)) : Probabilitas kalimat terhadap class.

- P(c) : Probabilitas class terhadap dokumen.

- Π : Product (perkalian beruntun).

- P(w|c) : Probabilitas kata terhadap class.

Guna memastikan perhitungan manual dari probabilitas nantinya peneliti akan menguji perhitungan dengan memakai software Weka dengan rumus :

$np\left(c\right)=\frac{P\left(c\right\}d(n))}{∑P(c|d(n))}$ (4)

Dimana:

- np(c) : Nilai prediksi class.

- P(c|d(n)) : Probabilitas kalimat terhadap class.

- Σ : Menjumlahkan seluruh bilangan.

Guna hitung semua akurasi *confusion matrix*, dengan rumus seperti dibawah ini:

$A=\frac{TP}{N}$ (5)

Dimana:

- A : Akurasi.

- TP : True Positive.

- N : Jumlah dokumen/tweet.

### Unified Modeling Language (UML)

UML adalah salah satu alat/model untuk merancang pengembangan perangkat lunak berorientasi objek. UML sendiri juga menyediakan standar untuk membuat sistem cetak biru yang berisi konsep proses bisnis, pembuatan kelas dalam bahasa pemrograman tertentu, skema database, dan komponen yang diperlukan untuk perangkat lunak sistem (Sonata & Sari, 2019).

UML adalah salah satu standar bahasa yang banyak digunakan dalam industri untuk mendefinisikan persyaratan, melakukan analisis dan desain, dan menggambarkan arsitektur untuk pemrograman berorientasi objek. UML muncul dari kebutuhan pemodelan visual untuk menentukan, menggambarkan, membuat, dan mendokumentasikan sistem perangkat lunak (Fatimah & Nuryaningsih, 2018).