BAB II

TINJAUAN PUSTAKA

A. Hasil Penelitian Terdahulu

Berhubungan dengan permasalahan penelitian yang akan dilaksanakan, berikut beberapa penelitian serupa yang telah dilaksanakan :

Pada penelitian sebelumnya yang dilakukan oleh Putri *et al.* (2022) dengan judul "Analisis Sentimen Kinerja Dewan Perwakilan Rakyat (DPR) Pada Twitter Menggunakan Metode *Naïve Bayes Classifier*". Dalam penelitian ini diperoleh hasil menurut penulis bahwa algoritma *Naïve Bayes* didapatkan *Accuracy score* sebesar 0.8 atau 80% hal ini berarti sistem mampu memprediksi 80% secara akurat dari total data *testing* sebesar 20%, dengan mendapatkan mendapatkan klasifikasi *tweet* dari Twitter mengenai DPR sebanyak 758 negatif, 693 netral dan 95 positif dari data hasil *crawling* sebanyak 1546 data, dengan nilai Precision score 75% untuk label positif, 79% netral, dan 82% label negatif, dengan nilai Recall score untuk label positif 29%, label netral 67% dan untuk label negatif sebanyak 84%, dengan nilai F1-Score untuk label positif sebanyak 43%, label netral sebanyak 70% dan label negatif sebanyak 77%.

Penelitian yang dilakukan Andika *et al.* (2019) dengan judul "Analisis Sentimen Masyarakat terhadap Hasil *Quick Count* Pemilihan Presiden Indonesia 2019 pada Media Sosial Twitter Menggunakan Metode *Naïve Bayes Classifier*". Pada penelitian ini diperoleh hasil akurasi dari model *Naïve Bayes* dengan diperoleh tingkat akurasi sebesar 82,90% dengan α = 0,05. Klasifikasi yang diperoleh masing-masing sebesar 65,5% (895) *tweet* negatif dan 34,5%

(471) *tweet* positif terhadap hasil *quick count* dengan nilai *Precision* dengan label negatif sebesar 80% dan label positif sebesar 98%, dengan nilai *Recall* yang diperoleh sebesar 100% untuk label negatif dan 46% untuk label positif, dan nilai *F1-Score* sebesar 89% untuk label negatif dan 63% untuk label positif.

Penelitian yang dilakukan Handayani & Sulistiyawati (2021) dengan judul "Analisis Sentimen Respon Masyarakat Terhadap Kabar Harian *Covid-19* Pada Twitter Kementerian Kesehatan Dengan Metode Klasifikasi *Naïve Bayes*". Penelitian ini dilakukan untuk mengetahui bagaimana hasil analisis sentimen terkait respon masyarakat dari kabar harian *Covid-19* dari Twitter Kementerian Kesehatan (Kemenkes) Republik Indonesia dan mengklassifikasikannya menjadi tiga kelas yaitu positif, negatif, dan netral menggunakan metode klasifikasi *Naïve Bayes Classifier*, dengan didapatkan *dataset* hasil *crawling* sebanyak 2397 *dataset*. Didapat hasil klasifikasi sentimen dengan tiga kelas, yaitu kelas negatif sebanyak 85%, kelas positif sebanyak 11% dan kelas netral sebanyak 4%. Data hasil klasifikasi dibagi menjadi data *training* dan *testing*, dengan proporsional jumlah data *training* sebanyak 80% dan data *testing* sebanyak 20%. Menghasilkan sentimen masyarakat pengguna Twitter mengenai respon masyarakat mengenai kabar harian *COVID-19* yang diberikan oleh Twitter Kementrian Kesehatan Republik Indonesia paling tinggi dengan presentase kelas sentimen negatif sebesaar 77%. Dengan menghasilkan nilai akurasi sebesar 78%, pengujian *Precision* sebesar 92%, nilai *Recall* sebesar 85%, *F1-Score* sebesar 88% dengan menggunakan pengujian metode *Naïve Bayes Classifier*.

Penelitian yang dilakukan Yulita *et al.* (2021) dengan judul "Analisis Sentimen Terhadap Opini Masyarakat Tentang Vaksin *Covid-19* Menggunakan Algoritma *Naïve Bayes Classifier*". Tujuan penelitian ini bertujuan untuk menganalisis opini masyarakat tentang vaksinasi *COVID-19* di Indonesia. Analisis dilakukan terhadap data 3780 *tweet* yang berkaitan vaksinasi dengan menggunakan algoritma *Naïve Bayes Classifier*. Berdasarkan analisis, dapat diamati bahwa sebagian besar *tweet* memiliki sikap positif (60,3 %), sementara jumlah *tweet* yang netral (34,4 %) dan tweet yang negatif (5,4 %). Penggunaan algoritma *Naive Bayes Classifier* dalam melakukan analisis sentimen ini sudah sangat baik ditunjukkan dengan sebuah hasil akurasi yang tinggi sebesar 93%. Kesimpulannya yang dapat diambil dari penelitian ini khusunya pengguna media sosial Twitter masyarakat Indonesia, paling banyak memberikan respon positif terkait adanya kebijakan vaksinasi *Covid - 19* di Indonesia dibuktikan dengan nilai presentase positif paling banyak sebanyak 2278 respon (60,3%) dibandingkan dengan penentangan vaksinasi hanya sebesar 203 respon (5,4%).

Penelitian yang dilakukan oleh Villavicencio *et al.* (2021) dengan judul "*Twitter Sentiment Analysis towards COVID-19 Vaccines in the Philippines Using Naïve Bayes*". Penelitian ini bertujuan untuk menganalisis sentimen terhadap vaksin *COVID-19* di Filipina menurut polaritas positif, netral, dan negatif. Berdasarkan hasil tersebut, dapat disimpulkan bahwa mayoritas atau 83% *tweet* di Filipina positif dan antusias dengan ide vaksinasi, sedangkan 9% netral dan 8% sentimen negatif. Data diolah terlebih dahulu menggunakan beberapa teknik *NLP*, dan model pengklasifikasi berhasil dikembangkan

menggunakan algoritma klasifikasi *Naïve Bayes* dengan akurasi 81,77%. Karena *Naïve Bayes* bekerja sangat baik bahkan dalam kumpulan data kecil, itu digunakan untuk penelitian ini, yang terdiri dari *tweet* untuk bulan pertama program vaksinasi di Filipina. Analisis sentimen terhadap vaksin *COVID-19* dapat membantu pemerintah Filipina membuat keputusan yang bijaksana terkait alokasi dana dan rencana peluncuran vaksinasi. Model yang dikembangkan menggunakan algoritma klasifikasi *Naïve Bayes* dapat membantu mengklasifikasikan *tweet* berdasarkan polaritasnya, terutama untuk bahasa Inggris dan Tagalog.

Penelitian yang dilakukan Hasan & Dwijayanti (2021) dengan judul "Analisis Sentimen Ulasan Pelanggan Terhadap Layanan *Grab* Indonesia Menggunakan *Multinomial Naïve Bayes Classifier*". Penelitian ini bertujuan untuk meningkatkan pelayanan dari *Grab* Indonesia dikarenakan adanya sebuah sentimen masyarakat yang mengarah kepada kepuasan dan ketidakpuasan terhadap pelayanan yang diberikan. Hasil akurasi pada pengujian Algoritma dari metode *Naïve Bayes Classifier* didapatkan hasil sebesar 92,5% dan pada proses pengujian evaluasi untuk sentimen negatif mendapatkan nilai *Precision* sebesar 57%, Recall 67% dan *F1-Score* 62%. Sedangkan untuk sentimen positif mendapatkan nilai *Precision* sebesar 97%, *Recall* 95% dan *F1-Score* 96%. Hasil analisis sentimen menunjukkan bahwa rata-rata sentimen yang diberikan mengandung arti positif, sehingga dapat diartikan bahwa pelanggan merasa puas dengan pelayanan dan fasilitas yang telah diberikan oleh *Grab* Indonesia.

Penelitian yang dilakukan oleh Djamaludin *et al.* (2022) dengan judul "Analisis Sentimen *Tweet* KRI Nanggala 402 di Twitter menggunakan Metode *Naïve Bayes Classifier*". Penelitian ini bertujuan untuk mengetahui kecenderungan respon (sentimen) pada masyarakat atas peristiwa tenggelamnya kapal selam KRI Nanggala 402. Data hasil crawling data Twitter memperoleh data sebanyak 53 data, dengan klasifikasi sebanyak 7 *tweet* bersifat positif, 7 *tweet* bersifat negatif, dan 39 *tweet* bersifat netral, dengan mendapatkan nilai *Accuracy* dari metode *Naïve Bayes Classifier* sebesar 73,00%. Kesimpulan pada penelitian ini yaitu masyarakat cenderung bersikap netral terhadap tenggelamnya kapal KRI Nanggala 402.

Penelitian yang dilakukan oleh Perdana *et al.* (2022) dengan judul "Analisis Sentimen Terhadap Isu Penundaan Pemilu di Twitter Menggunakan *Naive Bayes Classifier*". Penelitian ini dibagi menjadi tiga kelas sentimen, yaitu sentimen positif, netral, dan negatif, dengan tujuan untuk mengetahui sentimen terbesar terkait isu penundaan pemilu pada tahun 2024. Data yang diperoleh dari hasil crawling sebanyak 151.538 data, dengan klasifikasi sebanyak 60.442 bersifat positif, 32.223 bersifat netral dan 58.873 sentimen bersifat negatif. Sentimen positif dalam arti didukung dalam penundaan pemilu secara tegas jarang ditemukan, sentimen positif menunjukan sebuah pembelaan dari masyarakat kepada pemerintahan Presiden Joko Widodo terkait isu kemunculan penundaan pemilu. Untuk sentimen positif mendapatkan hasil *Precision* 96,1%, *Recall* 100%, *F1-Score* 98%. Untuk sentimen netral mendapatkan *Precision* 100%, *Recall* 94%, *F1-Score* 96,9%. Untuk sentimen negatif mendapatkan

Precision sebesar 98%, *Recall* 100%, *F1-Score* 99%. Nilai *Accuracy* menggunakan algoritma *Naïve Bayes* sebesar 98%.

Penelitian yang dilakukan oleh Syarifuddinn (2020) dengan judul "Analisis Sentimen Opini Publik Mengenai *Covid-19* Pada Twitter Menggunakan Metode *Naïve Bayes* Dan *KNN*". Penelitian ini menghasilkan nilai akurasi (nilai tingkat kedekatan antara nilai prediksi dengan nilai *actual*) didapatkan dari pengujian *Naïve Bayes* 63.21% sedangkan *KNN* 58.10%, dengan nilai *Precision* didapatkan *Naïve Bayes* dengan nilai 59.11% : *KNN* 53.10%, nilai *Recall* dari data didapatkan *Naïve Bayes* 56.80%: *KNN* 50.40%, nilai *F-Measure* didapatkan *Naïve Bayes* 57.96% : *KNN* 51.75%. Kecenderungan data *tweets* yang bersifat positif sebanyak 610 *tweets*, sedangkan negatif sebanyak 488 *tweets*, sehingga nilai *presicion*nya sebanyak 66.40% *positive* dan 58.94% *precsicion negative*, sehingga dapat ditarik kesimpulan bahwa *tweets* masyarakat umum di Twitter tentang *COVID-19* saat ini, cenderung pada opini positif. Masyarakat tidak perlu berpikir bahwa *COVID-19* hanya memberi dampak negatif, tetapi juga ada dampak positif.

Penelitian yang dilakukan oleh Pattiiha (2022) dengan judul "Perbandingan Metode *K-NN, Naïve Bayes, Decision Tree* untuk Analisis Sentimen *Tweet* Twitter Terkait Opini Terhadap PT PAL Indonesia". Penelitian ini bertujuan untuk mengetahui pendapat masyarakat mengenai pelayanan dan kinerja PT PAL Indonesia. Data dari hasil menggunakan aplikasi Rapidminer menghasilkan data sebanyak 1138 data. Data kemudian dibagi menjadi data *training* dan *testing* dengan menunjukan nilai akurasi pada metode

Naïve Bayes sebesar 84,08% dengan *class Precision* tertinggi 100%, Recall tertinggi sebesar 99,89%. Pada metode *KNN* mendapatkan akurasi sebesar 83,38, dengan *class Precision* tertinggi sebesar 96,43%, *Recall* tertinggi sebesar 99,89%. Pada metode *Decision Tree* menghasilkan akurasi sebesar 81,09%, dengan menghasilkan *class Precision* tertinggi sebesar 81,09% dan *Recall* tertinggi 100%. Disimpulkan bahwa metode *Naïve Bayes* memiliki tingkat akurasi lebih tinggi dibandingkan dengan metode *KNN* dan *Decision Tree*.

Tabel 2.1 Penelitian Terdahulu

No	Peneliti	Metode	Hasil
1.	Putri *et al.* (2022)	*Naïve Bayes Classifier*	Penelitian ini menggunakan data dari hasil *crawling* sebanyak 1546 data, dengan mendapatkan klasifikasi *tweet* dari Twitter mengenai kinerja DPR sebanyak 758 negatif, 693 netral dan 95 positif, dengan menghasilkan nilai *Accuracy score* sebesar 0.8 atau 80%. *Precision* = 78% positif, 79% netral, dan 82% negatif. *Recall* = 29% positif, 67% netral, dan 84 negatif. *F1-Score* = 43% positif, 70% netral dan 77% negatif. Membuktikan bahwa *Naïve Bayes Classifiers* dapat mengklasifikasi dengan baik.
2.	Andika *et al.* (2019)	*Naïve Bayes Classifier*	Hasil penelitian menunjukan nilai *Accuracy* sebesar 82,9%, dengan klasifikasi yang diperoleh masing-masing sebesar 34,5% (471) *tweet* positif dan 65,5% (895) *tweet* negatif terhadap hasil *quick count* pemilihan Presiden pada tahun 2019, dengan nilai *Precision* = 80% negatif, 98% positif. *Recall* = 100% negatif, 46% positif. *F1-Score* = 89% negatif dan 63% positif dengan menggunakan metode *Naïve Bayes Classifier.*
3.	Handayani & Sulistiyawati (2021)	*Naïve Bayes Classifier*	Penelitian ini bertujuan untuk mengetahui bagaimana hasil sentimen terhadap respon masyarakat dari kabar harian *Covid-19* dari Twitter Kementerian Kesehatan Republik Indonesia, dengan menggunakan *dataset* dari hasil *crawling* sebanyak 2397 *dataset*. Mendapatkan hasil klasifikasi sentimen dengan tiga kelas, yaitu kelas negatif 85%, netral 4% dan positif 11%. Data hasil klasifikasi ini dibagi menjadi 80% *training* dan 20% *testing*. Menghasilkan akurasi dengan

No	Peneliti	Metode	Hasil
			metode *Naïve Bayes Classifier* sebesar 78%, dengan nilai *Precision* sebesar 92%, *Recall* sebesar 85% dan *F1-Score* sebesar 88%.
4.	Yulita *et al.* (2021)	*Naïve Bayes Classifier*	Penelitian ini bertujuan untuk mengetahui analisis sentimen tentang program vaksinasi *Covid-19* di Indonesia. Menggunakan sebanyak 3780 data dari hasil *crawling*, dengan *tweet* memiliki sikap positif sebanyak 2278 data (60,3%), sementara *tweet* netral sebanyak 203 data (5,4%) dan *tweet* negatif sebanyak 1299 data (34,4%). Penggunaan algoritma *Naïve Bayes Classifier* dalam melakukan analisis sentimen ini sudah sangat baik ditunjukkan dengan hasil akurasi yang tinggi sebesar 93%.
5.	Villavicencio *et al*. (2021)	*Naïve Bayes Classifier*	Penelitian ini bertujuan untuk mengetahui sentimen masyarakat terhadap program vaksinasi di negara Filipina. Disimpulkan bahwa mayoritas atau 83% *tweet* di Filipina positif dan antusias dengan ide vaksinasi, sedangkan 9% netral dan 8% sentimen negatif. Data diolah terlebih dahulu menggunakan beberapa teknik NLP, dan model pengklasifikasi berhasil dikembangkan menggunakan algoritma klasifikasi *Naïve Bayes* dengan akurasi 81,77%.
6.	Hasan & Dwijayanti (2021)	*Naïve Bayes Classifier*	Penelitian ini menggunakan 1000 *dataset* yang dihasilkan dari *crawling data*. Hasil analisis sentimen masyarakat menunjuka hasil sentimen yang lebih banyak positif sebanyak 911 dibandingkan yang negatif sebanyak 89 terhadap pelayanan Grab di Indonesia. Hasil akurasi pada pengujian Algoritma dari metode *Naïve Bayes Classifier* didapatkan hasil sebesar 92,5%,

No	Peneliti	Metode	Hasil
			dengan nilai *Precision* = 57% negatif, 97% positif. Nilai *Recall* = 67% negatif, 95% positif dan *F1-Score* = 62% negatif dan 96% positif.
7.	Djamaludin *et al*. (2022)	*Naïve Bayes Classifier*	Penelitian ini menghasilkan *crawling data* sebanyak 53 data. Hasil dari klasifikasi menggunakan metode NBC dibagi menjadi tiga kelas, yaitu kelas positif sebanyak 7, kelas negatif sebanyak 7, dan kelas netral sebanyak 39, menghasilkan akurasi sebesar 73.00%. Disimpulkan bahwa sentimen masyarakat terkait tenggelamnya kapal selam KRI Nanggala 402 yaitu netral, ditunjukan dengan banyaknya *tweet* bersifat netral sebanyak 39 *tweet*.
8.	Perdana *et al.* (2022)	*Naïve Bayes Classifier*)	Data yang diperoleh dari hasil crawling sebanyak 151.538 data, dengan klasifikasi sebanyak 60.442 bersifat positif, 32.223 bersifat netral dan 58.873 sentimen bersifat negatif. Mendapatkan hasil Precision 96,1%, *Recall* 100%, F1-Score 98%. Sentimen netral mendapatkan Precision 100%, Recall 94%, F1-Score 96,9%. Sentimen negatif mendapatkan Precision sebesar 98%, Recall 100%, F1-Score 99%. Nilai *Accuracy* menggunakan algoritma *Naïve Bayes* sebesar 98%.
9.	Syarifuddinn (2020)	*Naïve Bayes Classifier* dan *K- Nearest Neighbor* (KNN)	Penelitian ini bertujuan untuk mengetahui sentimen masyarakat mengenai virus *Covid-19* di Indonesia pada media sosial twitter dengan menggunakan dua algoritma yaitu *Naïve Bayes Classifier* dan KNN dengan menghasilkan nilai akurasi untuk masing-masing metode *Naïve Bayes Classifier* sebesar 63,21% dan *KNN* sebesar 58,1%, *Precision* untuk NBC = 59.11% dan KNN = 53.10%,

No	Peneliti	Metode	Hasil
			Recall untuk *Naïve Bayes Classifier* = 56.80% dan KNN = 50.40%, nilai *f-measure* untuk algoritma *Naïve Bayes Classifier* = 57.96% dan KNN = 51.75%. Kecenderungan data bersifat negatif sebanyak 488 opini, positif sebanyak 610 opini, sehingga dapat ditarik kesimpulan dari opini atau *tweets* di Twitter tentang *Covid-19* saat ini, cenderung pada opini positif. Tingkat keakurasian algoritma *Naïve Bayes Classifier* lebih tinggi dibandingkan algoritma *K- Nearest Neighbor* (KNN).
10.	Pattiha (2022)	*Naïve Bayes Classifier, KNN, Decision Tree*	Penelitian ini bertujuan untuk mengetahui pendapat masyarakat mengenai pelayanan dan kinerja PT PAL Indonesia. Data dari hasil menggunakan aplikasi Rapidminer menghasilkan data sebanyak 1138 data. Data kemudian dibagi menjadi data *training* dan *testing* dengan menunjukan nilai akurasi pada metode *Naïve Bayes* sebesar 84,08% dengan *class Precision* tertinggi 100%, Recall tertinggi sebesar 99,89%. Pada metode *KNN* mendapatkan akurasi sebesar 83,38, dengan *class Precision* tertinggi sebesar 96,43%, *Recall* tertinggi sebesar 99,89%. Pada metode *Decision Tree* menghasilkan akurasi sebesar 81,09%, dengan menghasilkan *class Precision* tertinggi sebesar 81,09% dan *Recall* tertinggi 100%. Disimpulkan bahwa metode *Naïve Bayes* memiliki tingkat akurasi lebih tinggi dibandingkan dengan metode *KNN* dan *Decision Tree*.

B. Landasan Teori

1. *Natural Languege Processing* (NLP)

   Natural Language Processing (NLP) merupakan salah satu ilmu dari bidang ilmu komputer dari cabang *linguistic* (bahasa) dan kecerdasan buatan dengan fokus interaksi antara bahasa alami manusia, seperti Bahasa Inggris atau Bahasa Indonesia dengan komputer, dengan tujuan untuk membuat sebuah mesin yang mampu memahami dan mengerti makna dari bahasa manusia lalu memberikan respon yang sesuai (Yunefri & Fadrial, 2021).

2. Analisis Sentimen

   Analisis sentimen termasuk salah satu di dalam bidang *Natural Language Processing* (NLP) dan merupakan alat untuk membantu suatu proses identifiikasi dari sebuah *dataset*, yang didalam dataset tersebut terdapat suatu pandangan (sentimen) ataupun sebuah opini dalam bentuk teks yang mengangkat suatu isu, topik maupun sebuah kejadian yang memiliki sifat positif, netral, maupun negatif. Penerapan analisis sentimen dapat digunakan dalam berbagai hal ataupun sebuah aspek, misalnya isu politik, prediksi suatu harga saham, reputasi, maupun kepuasan terhadap suatu layanan ataupun produk (Fikri *et al.*, 2020).

3. *Text Mining*

   Text mining merupakan sebuah proses mengekstraksi sebuah informasi dari berbagai sumber data, dimana datanya belum terstruktur yang merucut pada sebuah proses penambangan data untuk dianalisis dan

memproses data tersebut. Proses *text mining* berawal dari sebuah pengambilan data kemudian data tersebut memasuki tahap *preprocessing*, yaitu tahap *case folding, cleansing, tokenization, filtering, stop removal* sebelum proses klasifikasi (Putri *et al.*, 2022).

4. Magang Bersertifikat

   Magang Bersertifikat adalah bagian dari program Kampus Merdeka yang bertujuan untuk memberikan kesempatan kepada mahasiswa belajar dan mengembangkan diri melalui aktivitas di luar kelas perkuliahan. Di program Magang Bersertifikat, mahasiswa akan mendapatkan pengalaman kerja di incustri/dunia profesi nyata selama 1-2 semester. Dengan pembelajaran langsung di tempat kerja mitra magang, mahasiswa akan mendapatkan *hard skills* maupun *soft skills* yang akan menyiapkan mahasiswa agar lebih profesional untuk memasuki dunia kerja dan kariernya (Syamsuadi *et al*., 2022)

5. Twitter

   Twitter merupakan sebuah layanan berbasis sosial media yang membantu untuk mengirim dan membaca pesan berbasis teks hingga 140 karakter penggunanya atau biasa kita sebut sebagai *tweets* atau kiacauan. Twitter juga digunakan sebagai media komunikasi untuk mengikuti *trend*, cerita, informasi dan berita dari seluruh penjuru dunia. Melalui Twitter, penggunanya dapat mengutarakan berbagai macam komentar maupun sebuah pemikiran terhadap suatu isu yang terjadi (Fikri *et al.*, 2020).

6. *Naïve Bayes Classifier*

Naïve bayes merupakan salah satu metode klasifikasi pada data mining. Teori yang mencari suatu probabilias sesuatu berdasarkan data yang telah ada sebelumnya dinamakan *Teory Bayes*. *Naïve Bayes Classifier* merupakan salah satu algoritma yang sederhana namun memiliki kemampuan dan kurasi yang tinggi dan termasuk dalam metode *machine learning* (Rosdiana *et al.*, 2019).

Dua tahapan dalam NBC dalam proses klasifikasinya, yaitu *training* dan *testing*. Pertama dilakukan pelatihan pada sentimen yang sudah diketahui kelasnya guna membangun sebuah model probalistik, kemudian tahap kedua pada proses klasifikasi sentimen yang belum diketahui kebenaran dari kelasnya (Rosdiana *et al.,* 2019).

Model *Naïve Bayes Multinomial*, model ini cocok digunakan untuk mengklasifikasikan kategori dokumen, dimana dokumen tersebut dapat dikategorikan dengan tema tertentu seperti olahraga, politik, teknologi, dan lainnya berdasarkan frekuensi kart-kata yang muncul pada dokumen. Persamaan *Naïve Bayes Multinomial* menurut Sabrani *et al.*, (2020) adalah sebagai berikut:

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

Dimana:

P(c) : Prior probability suatu dokumen berada di kelas c

Nc : Jumlah dokumen dari kelas c

N : Jumlah total keseluruhan dokumen

$$P(w|c) = \frac{count(w,c)+1}{count(c)+|V|} \quad \ldots(2)$$

Dimana:

P(w|c) : Jumlah w muncul di dokumen text yang memiliki polaritas c

Count(w,c) : Jumlah kemunculan kata w pada kelas c

Count(c) : Jumlah kemunculan kata pada dokumen dengan kategori c

|V| : jumlah atribut pada sampel.

$$P(w|c) \propto P(c) \prod_{k=1}^{n} P(wc|c) \quad \ldots(3)$$

Dimana:

P(w|c) : Probabilitas dokumen c berada dikelas w

P(c) : Prior probability suatu dokumen berada di kelas c

{t1, t2…tn} : Token dalam dokumen d yang merupakan bagian dari vocabulary dengan jumlah n

P(w|c) : Probabilitas bersyarat kata w berada di dokumen pada kelas c

Langkah-langkah dalam mengklasifikasi data dengan *Multinomial Naïve Bayes* (MNB) adalah:

1. Pertama, Menghitung probabilitas prior pada setiap kelas dengan menggunakan rumus persamaan (1).
2. Kedua, Menghitung probabilitas kata ke-n pada kelas c menggubakan rumus persmanaan (2).
3. Ketiga, Menghitung probabilitas suatu dokumen dengan menggunakan rumus persamaan (3).

4. Keempat, Menentukan kelas dokumen dengan membandingkan nilai probabilitasnya antar kelas. Nilai probabilitas yang tertinggi akan dipilih dalam menentukan menentukan kelasnya.

7. *Confussion Matrix*

Confusion Matrix merupakan tabel yang menyatakan sebuah klasifikasi jumlah pada data uji yang salah dan jumlah data uji yang benar (Normawati & Prayogi, 2021). *Confusion Matrix* untuk pengklasifikasaian biner ditunjukan pada Tabel 2. 2.

Tabel 2. 2 *Confusion Matrix*

		Kelas Prediksi	
		1	0
Kelas Sebenarnya	1	TP	FN
	0	FP	TN

Keterangan:

TP (*True Positive*) : jumlah dokumen dari kelas 1 yang benar diklasifikasikan sebagai kelas 1

TN (*True Negative*): jumlah dokumen dari kelas 0 yang benar diklasifikasikan sebagai kelas 0

FP (*False Positive*) : jumlah dokumen dari kelas 0 yang salah diklasifikasikan sebagai kelas 1

FN (*False Negative*) : jumlah dokumen dari kelas 1 yang salah diklasifikasikan sebagai kelas 0

Confusion Matrix menurut Normawati & Prayogi (2021), menyatakakan bahwa *Confusion Matrix* digunakan untuk menghitung nilai

Accuracy, Precision, Recall dan *F1-Score* dengan persamaan sebagai berikut:

$$Accuracy = \frac{TP + TN}{Total} \; x \; 100\% \quad \ldots (4)$$

$$Precision = \frac{TP}{TP + FP} \; x \; 100\% \quad \ldots (5)$$

$$Recall = \frac{TP}{TP + FN} \; x \; 100\% \quad \ldots (6)$$

$$F1 - Score = \frac{2 \; x \; Presisi \; x \; Recall}{Presisi + Recall} \; x \; 100\% \quad \ldots (7)$$

8. *Labelling Lexicon*

Labelling atau pelabelan adalah proses pemberian label pada isi *tweet* Twitter, kemudian pelabelan tersebut disimpan ke *dataset* untuk selanjutnya dimasukan ke dalam proses *splitting data*. Terdapat tiga bentuk *labelling data* pada suatu kalimat, yaitu negatif, positif maupun netral, pelabelan ini berfungsi untuk konsistensi dan menghindari adanya kesalahan pada penginputan.

Lexicon Based adalah salah satu metode dalam *labelling data*, dimana sebuah fitur kata dengan memiliki sentimen positif, netral maupun negatif berdasarkan *lexicon* (kamus). Kamus *lexicon based* berfungsi sebagai pelabelan untuk menghitung skor didalam sentimen, setelah diketahui termasuk kemana sentimen tersebut, baik ke sentimen positif, netral, maupun negatif, selanjutnya menghitung setiap kata yang mengandung sentimen didalam sebuah kalimat dengan menjumlahkan nilai

opini. Jumlah nilai untk sentimen negatif bernilai <0, netral = 0, dan sentimen positif bernilai >0 atau lebih (Mahendrajaya *et al.*, 2019)

9. *Jupyter Notebook*

Jupyter Notebook merupakan sebuah *platform notebook* untuk komputasi, yaitu dengan caramenyisipkan sebuah *natural language processing* (NLP), dimana kode sumber dan *output*nya membentuk sebuah narasi perhitungan yang mudah dipahami manusia.

Jupyter Notebook dipilih karena beberapa alasan. Pertama adalah bahwa baik *Jupyter* dan *Python* adalah *open-source*, dengan mudah distribusi untuk semua sistem operasi umum. Kedua , bahwa *python* adalah salah satu yang paling popular bahasa pemrograman di dunia, khususnya untuk data analisis dengan sejumlah besar dengan sumber daya online. Selain itu, *python* menjadi semakin populer di dalam komunitas sains, dengan banyak contoh *python* digunakan untuk simulasi dan pemodelan. Yang ketiga adalah bahwa *python* sederhana dan fleksibel, dengan sintaks langsung, sehingga relatif mudah untuk membaca dan mempelajari kodenya (Menke, 2020).

10. *Python*

Python merupakan salah satu bahasa yang mendukung pemrograman berorientasi objek dengan pemrograman tingkat tinggi. Penulisan sintaks dalam python memiliki perbedaan dengan bahasa pemrograman lain, dalam analisis data pada *python*, *python* juga menyediakan berbagai *framework* dan *library* (Putri *et al*., 2022).

11. *Web Scraping*

Web scraping adalah teknik untuk mengekstrak data tidak terstruktur dari situs *web* dan transformasi data tersebut menjadi data terstruktur yang dapat disimpan dan dianalisis dalam database. *Web scraping* juga dikenal sebagai *web data extraction*, *web data scraping*. *Web scraping* adalah salah satu bentuk *data mining*, tujuan keseluruhan dari proses *web scraping* adalah untuk mengekstrak informasi dari situs *web* dan mengubahnya menjadi struktur yang dapat dimengerti seperti *spreadsheet*, *database* atau *comma separated values* (Kurniawan & Apriliani, 2020).

12. CSV

File CSV (*Comma Separated Values*) adalah file teks biasa untuk disimpan dan ditukar dengan sederhana data tabel terstruktur. Biasanya, setiap baris dalam tabel mewakili catatan data yang terpisah. Bidang catatan dipisahkan dengan koma (terlepas dari penamaan karakter lain, misalnya titik koma atau tab, juga digunakan) dan semua catatan harus memiliki struktur yang sama, yaitu urutan bidang yang sama, ekstensi file yang umum adalah ".csv". Format file CSV terbuka, terkenal dan didukung secara luas oleh editor teks, *spreadsheet* program, *database* dan bahasa pemrograman, karena struktur berbasis teks sederhana, dapat dengan mudah diproduksi atau diedit secara manual (Mäs *et al*., 2018).