BAB II
TINJAUAN PUSTAKA

2.1 Penelitian terdahulu

Sebagai bahan untuk menyempurnakan dan memberikan kontribusi dalam kajian tugas akhir ini, peneliti telah merangkum kajian-kajian yang relevan dengan kajian ini seperti terlihat pada Tabel 2.1.

Pembaruan dalam penelitian ini adalah menggunakan data komentar dari YouTube dan akan membahas PPKM di DKI Jakarta. Oleh karena itu peneliti akan membuat analisis sentimen terhadap PPKM di DKI Jakarta dengan data komentar pada YouTube pada tanggal 5-6 juli 2022 menggunakan metode Naïve Bayes. Pendekatan Naive Bayes digunakan karena menunjukkan akurasi yang cukup baik untuk analisis sentimen kategorikal.

2.2 Dasar Teori

2.2.1 Data Mining

Data mining adalah metode untuk menemukan pola tersembunyi dari data yang ada dengan menggunakan teknik pengolahan data [2]. Metode penambangan data penting untuk membuat keputusan berdasarkan analisis volume besar data klinis [26]. Analisis terhadap komputasi sikap, tindakan, dan perasaan orang terhadap berbagai hal dikenal sebagai analisis sentimen (SA) atau penambangan opini (OM). Entitas dapat mewakili individu, peristiwa, atau topik. Topik ini mungkin akan dibahas dalam ulasan. Kedua ekspresi SA atau OM dapat dipertukarkan. Keduanya mengekspresikan makna bersama [11].

Tabel 2. 1 Penelitian Terkait Sebelumnya

No	Judul Penelitian	Peneliti	Objek Penelitian	Metode Pelitian	Hasil
1	Analisis Sentimen Terhadap Opini Masyarakat Tentang Vaksin Covid-19 Menggunakan Algoritma *Naïve Bayes Classifier* (2021) [13].	Winda Yulita, Eko Dwi Nugroho, Muhammad Habib Algifari	Membuat Analisis Sentimen terhadap Opini Masyarakat Tentang Vaksin Covid-19.	Algoritma *Naïve Bayes Classifier*	Melalui eksperimen tersebut, tingkat kinerja model *Naïve Bayes* telah mencapai Akurasi sebesar 93%
2	*A sentiment analysis approach to predict an individual's awareness of the precautionary procedures to prevent covid-19 outbreaks in Saudi Arabia* (2021) [19].	Sumayh S. Aljameel , Dina A. Alabbad, Norah A. Alzahrani, Shouq M. Alqarni, Fatimah A. Alamoudi, Lana M. Babili, Somiah K. Aljaafary and Fatima M. Alshamrani.	Melakukan analisis pada kesadaran individu tentang pecegahan Covid-19 di Arab Saudi dengan mengunakan metode KNN, SVM dan Naïve Bayes.	Menggunakan metode KNN, SVM dan Naïve Bayes.	Melalui eksperimen tersebut, tingkat akurasi tertinggi yaitu SVM dengan akurasi tertinggi 85%, Naïve Bayes dengan akurasi tertinggi 80% dan KNN dengan akurasi tertinggi 64%.
3	Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter [20].	Dwi Normawati, Surya Allit Prayogi	Membuat sistem yang dapat mengolah data dengan menerapkan analisis sentiment.	Menggunakan Naïve Bayes dan Confusion Matrix.	hasil implementasi NBC dan pengujian performa menggunakan confusion matrix yang didapatkan akurasi sebesar 82%, presisi 93%, dan recall sebesar 52%.

No	Judul Penelitian	Peneliti	Objek Penelitian	Metode Pelitian	Hasil
4	Penerapan Algoritma *K-Means* Dalam Menentukan Tingkat Kepuasan Mahasiswa Terhadap Pembelajaran Online (2020) [4].	Kristin D R Sianipar, Septri Wanti Siahaan, Marina Siregar, P.P.P.A.N.W Fikrul Ilmi R.H Zer, Dedy Hartama	Menentukan tingkat kepuasan mahasiswa terhadap pembelajaran online.	Algoritma K-Means.	Melalui eksperimen tersebut, tingkat kinerja model *K-Means* dapat menentukan 3 *cluster*.
5	*Sentiment and Emotion in Social Media COVID–19 Conversations: SAB-LSTM Approach* (2020) [21].	Ashok Kumar D, Anandan Chinnalagu	Membuat deteksi sentiment dan emosi di media sosial yang diakibatkan oleh Covid-19.	Algoritma *SAB-LSTM Approach.*	Melalui eksperimen tersebut SAB-LSTM memang lebih baik, namun metode ini membutuhkan *dataset* berbagai Bahasa agar melakukan sentiment.
6	*COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via BERT Model* (2020) [22].	TIANYI WANG, KE LU, KAM PUI CHOW, dan QING ZHU.	Melakukan analis terhadap media media sina weibo menggunakan metode BERT dan di sempurnakan dengan TF-IDF.	Menggunakan algoritma TF-IDF.	Melalui eksperimen tersebut, dengan menggunakan TF-IDF bisa mendapatkan akurasi sebesar 78.65%.
7	*Sentiment Identification in COVID-19 Specific Tweets* (2020) [23].	Manoj Sethi, Sarthak Pandey, Prashant Trar , dan Prateek Soni.	Melakuakan analisis terdapat komentar Twitter dalam pengaturan bi-class dan multi-class.	Menggunakn bi-class dan multi-class.	Melalui eksperimen tersebut, tingkat kinerja pengaturan bi-class dan multi-class dapat menghasilkan akurasi 93%.

No	Judul Penelitian	Peneliti	Objek Penelitian	Metode Pelitian	Hasil
8	*COVID-19 public sentiment insights and machine learning for tweets classification* (2020) [18].	Jim Samuel , G. G. Md. Nawaz Ali , Md. Mokhlesur Rahman, Ek Esawi dan Yana Samuel	Melakukan (1) sentimen publik terkait dengan perkembangan virus Corona dan COVID-19, (2) penggunaan data Twitter, yaitu Tweet, untuk analisis sentimen, (3) analisis tekstual deskriptif dan visualisasi data tekstual, dan (4) perbandingan mekanisme klasifikasi tekstual yang digunakan dalam kecerdasan buatan (AI).	menggunakan metode *Naïve Bayes* dan *Logistic Regression.*	Melalui eksperimen tersebut, tingkat prediksi Naïve bayes untuk kalimat pendek mendapat akurasi 91% dan dengan *Logistic Regression mendapatkan akurasi 74%*.
9	Aspect Based Sentiment Analysis pada Layanan Umpan Balik Universitas dengan Menggunakan Metode Naïve Bayes dan Latent Semantic Analysis (2019) [24].	Gunawan Setiawan, Henry Novianus Palit, Endang Setyati Program	Membuat analisis sentiment pada layanan umpan balik universitas kristen petra.	Metode Naïve Bayes dan Latent Semantic Analysis.	Melalui eksperimen tersebut, tingkat kinerja model *Naïve Bayes* lebih baik dengan akurasi 87,87%. Sedangkan LSA mendapatkan 86,2%

No	Judul Penelitian	Peneliti	Objek Penelitian	Metode Pelitian	Hasil
10	*Sentiment Analysis using Sentiwordnet and Machine Learning Approach (Indonesia general election opinion from the twitter content)* (2019) [25].	Eka Miranda, Mediana Aryuni, Ricky Hariyanto dan Edwin Satya Surya	Membuat analisis sentimen opini pemilu dari Twitter.	Menggunakan *SentiWordnet* dan klasifikasi *Naïve Bayes*.	Melalui eksperimen tersebut, tingkat prediksi dengan Naïve bayes menghasilkan akurasi tertinggi 74,94%.

2.2.2 Clustering

Menggunakan premis memaksimalkan tren di antara anggota kelas dan membatasi kesamaan antara kelas atau kelompok, pengelompokan adalah tindakan mengkategorikan item berdasarkan informasi yang dikumpulkan dari data yang menunjukkan hubungan antar objek [26]. Perspektif lain mengklaim bahwa clustering adalah teknologi data mining yang berupaya mengelompokkan item ke dalam cluster atau pengelompokan [27].

2.2.3 API v3

API (Application Programming Interface) merupakan suatu metode yang bermanfaat untuk menghubungkan sehingga bisa berkomunikasi antar perangkat lunak atau website.Pada penelitian ini, API digunakan untuk mengambil data komentar pada konten YouTube.

2.2.4 Klasifikasi

Klasifikasi merupakan proses kategorisasi yang memakai serangkaian data training buat mengenali, membedakan, serta memahami objek. Selain itu *klasifikasi* dapat mengkelompokan data terhadap keterikatan data sample. Klasifikasi merupakan Teknik dalam *data maining* [28]. Untuk mengklasifikasi sebuah data, terdapat banyak algoritma klasifikasi seperti: SVM, Naïve Bayes, CNN dan lain-lain. Penelitian ini melakukan Klasifikasi dengan metode Naïve Bayes.

2.2.5 Sastrawi

Sastrawi adalah *library* pada bahasa pyton yang dapat digunakan untuk mendapatkan kata dasar pada data inputan. Sastrawi menggunakan algoritma Nazief dan Andriani yang sangat disukai untuk membendung kata-kata bahasa Indonesia. *Library* ini bergantung pada kata dasar dari www.kateglo.com [24].

2.2.6 SentiWordNet

SentiWordNet adalah sumber leksikal di mana setiap kata di Wordnet terkait dengan tiga skor numerik, yaitu *Netral, Positive* dan *Negative*. Skor ini mendefinisikan objektivitas, kepositifan, dan tingkat negatif dalam istilah yang terkandung dalam kata-kata. Setiap nilai skor berkisar dari 0,0 hingga 1,0, dan jumlahnya adalah 1,0 untuk setiap synset (synset adalah kumpulan dari satu atau lebih sinonim). SentiWordNet menggunakan algoritma random walk. Teori untuk tahap random walk yaitu, jika 2 synset atau kata memiliki konteks yang sama, mereka cenderung memiliki sentimen yang sama. Setiap synset akan dikaitkan dengan positif atau negative konteks. Semakin banyak hubungan dengan konten positif, semakin semakin besar nilai positifnya, dan sebaliknya [25].

2.2.7 TF-IDF

Metode yang paling sering digunakan untuk menentukan bobot setiap kata disebut TF-IDF (Term Frequency - Inverse Document Frequency). Pendekatan ini terkenal karena lebih efektif, berguna, dan menghasilkan hasil yang tepat. Rumus berikut akan digunakan untuk menghitung setiap token (istilah) dalam dokumen:

$$W_{dt} = tf_{dt} \times IDF_t \quad (2.1)$$

d = dokumen ke-d. t = kata ke-t dari kata kunci. w = bobot dokumen ke-d terhadap kata ke-t. tf = banyaknya kata yang di cari. Sedangkan untuk mencari IDF:

$$IDF = hasil\ Log\ 2\ \left(\frac{D}{df}\right) \quad (2.2)$$

D = total dokumen

df = banyaknya dokumen yang mengandung kata yang dicari [29].

2.2.8 K-Means

K-means adalah algoritma pembelajaran mesin populer yang digunakan untuk pengelompokan data di mana jumlah *cluster* (K)

diketahui, sebelum dijumlahkan atau ditunjukkan sebelumnya [30]. K-Means dapat digunakan untuk mengidentifikasi pusat *cluster* sedemikian rupa sehingga jarak pengamatan ke pusat *cluster* terdekat mereka diminimalkan. Dengan demikian, semua pengamatan yang paling dekat dengan pusat *cluster* yang sama dipandang sebagai milik kelompok yang sama [31]. Sedangkan untuk menentukan *Euclidean Distance* sebagai berikut:

$$D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2} \quad (2.3)$$

D_e adalah *Euclidean Distance*. i adalah banyak data. (x,y) adalah kordinat objek. (s,t) adalah kordinat *centroid.*

Pada tahap ini peneliti melihat hasil menggunakan *Silhouette*. *Silhouette* adalah salah satu indikator validitas yang digunakan untuk menilai hasil dari algoritma pengelompokan. Nilai rata-rata setiap titik data ditentukan oleh *Silhouette*. Perhitungan ini dilakukan dengan cara membagi nilai terbesar antara nilai *separation* dan *compactness* dengan selisih antara keduanya. Hasil Nilai *Silhouette* yang mendekati rentang nilai 1 menunjukkan cluster yang baik [32].

2.2.9 Label Encorder

Label Encorder mrupakan algoritma untuk mengubah nilai pada kolom menjadi angka. Blok komputasi *transformator* berdasarkan diri sendiri digunakan untuk mengkodekan urutan label secara independent [33].

2.2.10 CountVectorizer

CountVectorizer adalah algoritma untuk melakukan perhitungan frekuensi kata dalam data. Tujuan countvectorizer adalah untuk mengekstrak kata-kata dari setiap kalimat dan membangun kosakata dari setiap kata yang unik untuk kalimat itu. Jumlah kata apa pun dapat digunakan sebagai vektor fitur dalam kosakata. Sehingga dengan menambahkan Tfidf Vectorizer pada

tahap TF-IDF dapat memudahkan membuat daftar file kata-kata yang berperngaruh pada kelas [34].

2.2.11 Tokenisasi

Tokenisasi adalah pemisahan atau penguraian dari data tekstual menjadi token, yang lebih kecil artinya komponen. Prinsip dari tokenisasi yaitu memisahkan setiap kata yang Menyusun kalimat pada dokumen. Ini dapat diklasifikasikan ke dalam tokenization kalimat dan tokenisasi kata [35]. Selain itu, prosedur ini akan menghilangkan angka, tanda baca, dan karakter alfabet. Sebagai contoh misalnya terdapat kalimat "Itu motor saya" saat di tokenisasi akan menjadi "itu, motor, saya". Dalam tokenisasi terdapat bermacam algoritma seperti *RegexpTokenizer*. *RegexpTokenizer* adalah algoritma untuk mengeksekusi tokenisasi kata yang tersedia di NLTK.

2.2.12 Naïve Bayes

Naïve Bayes adalah teknik klasifikasi yang memanfaatkan teknik probabilistik dan statistik. Teknik pembelajaran terawasi yang cukup mudah digunakan dalam metode ini. Multinomial Naive Bayes, Bernoulli Naive Bayes, dan Gaussian *Naive Bayes* adalah tiga bentuk teknik Naive Bayes yang berbeda. Untuk menghitung dengan Teorema Naïve Bayes:

$$P(c) = \frac{N_c}{N} \tag{2.2}$$

$$P(w|\hat{c}) = \frac{count(\hat{c},w)+\alpha_w}{count(\hat{c})+\alpha} \tag{2.3}$$

$$P(c|w_i) = \ argmin\, P(c)\prod_{i=1} \frac{1}{P(W|\hat{c})^{fi}} \tag{2.4}$$

w = kata, serta c =kelas sasaran. Nc =jumlah kelas c. N = jumlah semua kelas. $Count(\hat{c},w)$ = berapa kali kata w ada selain dalam dokumen selain kelas c. $Count(\hat{c})$ = jumlah semua istilah yang muncul pada kelas selain c. αw = smoothing parameter. α = jumlah holistik dari αw. fi = jumlah frekuensi istilah i pada

dokumen wi. $P(w|\hat{c})$ adalah likelihood yang bisa diartikan probabilitas dari istilah w yang timbul selain pada kelas c. $P(c)$ adalah prior probability yang bisa diartikan probabilitas kelas c yang timbul dalam semua kelas. $P(c|wi$) biasa dianggap posterior probability yang bisa diartikan nilai terendah berasal probabilitas kelas c yang timbul dalam dokumen wi .

Complement Naïve Bayes yaitu algoritma yang mengadopsi *Multinomial Naïve Bayes* serta didesain buat data yang tidak seimbang. Sistem di *Complement Naïve Bayes* dengan menghitung probabilitas kata yang timbul dari luar kelas. lalu menghitung probabilitas setiap kelas dan dipilih nilai terendah. Nilai probabilitas terendah tadi dipilih sebab bukan didapatkan dari kelas. sebagai akibatnya kelas tadi mempunyai probabilitas tertinggi [36].

2.2.13 Evaluasi

Untuk melihat performa klasifikasi ada beberapa cara, tetapi di penelitian ini akan memakai Confusion Matrix untuk menghitung Akurasi, Presisi, Recall, dan F1. Confusion matrix merupakan tabel yang menampilkan jumlah data atau hasil klasifikasi yang benar dan salah [20]. Hasil dari confusion matrix ini merukan tabel yang berisi *True positive* (data yang berkelas 1 di klasifikasi menjadi 1), *True negative* (data yang berkelas 0 di klasifikasi menjadi 0), *False positive* (data yang berkelas 0 di klasifikasi menjadi 1) dan *False negative* (data yang berkelas 1 di klasifikasi menjadi 0).

Akurasi adalah persentase yang diperoleh dari total sentimen yang diidentifikasi dengan benar [37]. Akurasi dihitung dengan membagi jumlah data sentimen yang benar dengan total dan data uji, sebagai berikut:

$$Akurasi = \frac{True\ positive + True\ negative}{True\ positve + False\ negative + True\ negative + False\ negative} \text{ X } 100\% \quad (2.5)$$

Precision adalah proporsi data relevan yang ditemukan terhadap data total. [37]. Untuk menghitung Precision sebagai berikut:

$$\text{Precision} = \frac{True\ positive}{True\ positive + false\ positive} \quad (2.6)$$

Recall adalah perbandingan kuantitas material relevan yang ditemukan dengan kuantitas material relevan [37]. Untuk menghitung Recal sebagai berikut:

$$\text{Recall} = \frac{True\ positive}{True\ positive + false\ negative} \quad (2.7)$$

F1 adalah parameter tunggal untuk mengukur keberhasilan pengambilan, menggabungkan daya ingat dan presisi[37]. Untuk menghitung F-measure sebagai berikut:

$$\text{F} - \text{measure} = 2 * \frac{Precision * recall}{Precision + recall} \quad (2.8)$$

Untuk membuat kategori pada hasil akan dikategorikan berdasarkan akurasi yang di dapat. Untuk range akurasi yang digunakan sebagai berikut:

1. 0.90 – 1.00 = *Excellent classification*
2. 0.80 – 0.90 = *Good classification*
3. 0.70 – 0.80 = *Fair classification*
4. 0.60 – 0.70 = *Poor classification*
5. 0.50 – 0.60 = *Failure*