BAB II
LANDASAN TEORI

2.1 Penelitian Terdahulu

Pada tabel 2.1 menampilkan ringkasan penelitian sebelumnya yang fokus pada sentimen analisis dengan objek vaksin covid-19 dan proses vaksinasi.

Tabel 2. 1 Penelitian Terdahulu

Pengarang	Judul	Hasil
Cotfas et al., 2021	*The Longest Month: Analyzing COVID-19 Vaccination Opinions Dynamics From Tweet in the Month Following the First Vaccine Announcement*	Pengumpulan *tweet* dilakukan pada periode 9 November 2020 – 8 Desember 2020, dengan topik "*covid*" dan "*vaccination*" dengan kata kunci spesifik: "*covid19, covid-19, coronavirus, coronaoutbreak, coronaviruspandemic, wuhanvirus, vaccination, vaccinate, vaccinating, vaccinated*. Polaritas sentimen dibagi menjadi tiga : "*against, neutral,* dan *in favor*". Metode klasifikasi yang digunakan adalah *Multinomial Naive Bayes (MNB), Random Forest (RF), Support Vector Machine (SVM), Bidirectional Long Short-Term Memory(Bi-LSTM),* dan *Convolutional Neural Network(CNN)*.
Fitriana et al., 2021	Analisis Sentimen Opini Terhadap Vaksin Covid - 19 pada Media Sosial *Twitter* Menggunakan *Support Vector Machine (SVM)* dan *Naive Bayes (NB)*	*Crawling tweet* dilakukan dengan menggunakan kata kunci " vaksin covid-19". Metode yang digunakan SVM dan NB. SVM memiliki nilai akurasi, presisi, dan *recall* lebih baik dibandingkan dengan *Naive Bayes*.
Naraswati et al., 2021	Analisis Sentimen Publik dari *Twitter* Tentang Kebijakan Penanganan Covid-19 di Indonesia dengan *Naive Bayes Classification*	*Crawling tweet* dilakukan pada periode 5-7 Oktober 2020 dengan kata kunci "PSBB, wajib masker, dan jam malam". Polaritas dibagi dua: positif dan negatif. *Naive Bayes Classification* memiliki nilai cukup baik dengan nilai akurasi 87,34%, sensitivitas 93,43%, dan spesifisitas 71,76%.

Sattar & Arifuzzaman, 2021	*COVID-19 Vaccination Awareness and Aftermath: Public Sentiment Analysis on Twitter Data and Vaccinated Population Prediction in the USA*	Analisis sentimen menggunakan metode *Unsupervised Lexicon-Based Approaches* dan *tools* yang digunakan *TextBlob* dan *VADER*. *Crawling tweet* menggunakan kata kunci " *a safe, healthy lifestyle after vaccination*" dan 7 jenis vaksin covid-19 yaitu: *Pfizer, Moderna, Johnson & Johnson, Oxford-AstraZeneca, Sputnik V, Covavin, Sinovac*. *Crawling tweet* dilakukan pada rentang 10 April 2021 – 17 May 2021.
Hikmawan et al., 2020	Sentimen Analisis Publik Terhadap Joko Widodo terhadap wabah Covid-19 menggunakan Metode *Machine Learningpp*	*Crawling tweet* menggunakan kata kunci "Jokowi" dan "covid" dalam bahasa Indonesia. Polaritas dibagi kedalam tiga klasifikasi: positif, negatif, dan netral. Metode klasifikasi yang digunakan *Naive Bayes (NB)*, *Support Vector Machine (SVM)*, dan *K-Nearest Neighbor (KNN)*. SVM menghasilkan nilai akurasi terbaik dari ketiga metode yang dipakai yaitu sebesar 84,58%.
Nomleni, 2015	Sentiment Analysis Menggunakan *Support Vector Machine* (SVM)	*Support Vector Machine* menghasilkan nilai akurasi yang sangat baik, dengan nilai diatas 80% dari 7 kali percobaan yang dilakukan.
Kunal et al., 2018	*Textual Dissection of Live Twitter Reviews Using Naive Bayes*	Algoritma yang digunakan adalah algoritma *Naive Bayes* dengan dua *library python* yaitu *Tweepy* Dan *TextBlob*. *TextBlob* untuk menghitung nilai polaritas dan *Tweepy* untuk akses ke *live tweet* pada *Twitter*.
Villavicencio et al., 2021)	*Twitter Sentiment Analysis towards COVID-19 Vaccines in the Philippines Using Naïve Bayes*	*Tweet* yang dijadikan objek adalah *tweet* dalam bahasa Inggris dan tagalog. Polaritas diklasifikasi menjadi tiga : positif, negatif, dan netral. Hasil akurasi *Naive Bayes Classification* cukup baik yaitu 81,77%.

Nurdeni et al., 2021	*Sentiment Analysis on Covid19 Vaccines in Indonesia: From The Perspective of Sinovac and Pfizer*	*Tweet* yang digunakan adalah *tweet* dalam bahasa Indonesia yang *crawling* pada rentang Oktober – November 2020 dengan kata kunci "*Sinovac*" dan "*Pfizer*". Pelabelan polaritas dilakukan secara manual, yang terdiri dari tiga label yaitu positif, negatif, dan netral. Dengan menggunakan evaluasi kinerja model yang digunakan yaitu *Support Vector Machine* dengan menggunakan *10-fold cross-validation* nilai akurasi untuk *dataset Sinovac* sebesar 85% dan nilai akurasi untuk *dataset Pfizer* sebesar 78%.
Laurensz & Sediyono, 2021, p. 19	Analisis Sentimen Masyarakat terhadap Tindakan Vaksinasi dalam Upaya Mengatasi Pandemi Covid-19	*Tweet* di *crawling* menggunakan dua kata kunci "vaksin merah putih" dan "vaksin sinovac" yang dilakukan pada periode rentang November 2020 – Februari 2021. Polaritas dibagi menjadi dua kelas: positif dan negatif. *Naive Bayes* menghasilkan nilai akurasi terbaik sebesar 85,59% sedangkan SVM sebesar 84,41%.
Aljameel et al., 2021	*A Sentiment Analysis Approach to Predict an Individual's Awareness of the Precautionary Procedures to*	*Tweet* menggunakan bahasa Arab dengan kata kunci #covid #corona #stay_home, #wearing_mask #sterilization #distance_learning #washing_hands #social_distance. *Dataset tweet* di download dari *Twitter* menggunakan *TWINT* pada periode 23 Maret 2020 - 21 Juni 2020 di Arab Saudi. Dalam tulisan dan bahasa Arab. Polaritas diklasifikasi

Prevent COVID-19 Outbreaks in Saudi Arabia	ke dalam tiga kelas: positif, negatif, dan netral. Metode *Machine Learning* yang digunakan *Naïve Bayes, K-Nearest Neighbor*, dan *Support Vector Machine*. Menggunakan *N-Gram model (unigram, unigram TF-IDF, bigram, bigram TF-IDF)* untuk komparasi model *Natural Language Processing (NLP).* Dari ketiga metode yang digunakan, SVM, KNN, dan NB, SVM mendapatkan nilai akurasi terbaik yaitu 85%.

2.2 Machine Learning

Machine learning merupakan sub-bagian dari *Artificial Intelligence (AI)*. *Machine learning* merupakan sekumpulan metode yang secara otomatis dapat mendeteksi pola di dalam data, dan kemudian menggunakan pola yang telah ditemukan untuk memprediksi data di masa mendatang, atau untuk melakukan pengambilan keputusan dalam ketidakpastian [13]. *Machine learning* merupakan penggabungan dari sekumpulan metode pembelajaran yang telah dilatih kedalam sebuah mesin [14]. *Machine learning* dapat dikategorikan menjadi dua yaitu *supervised learning* and *unsupervised learning*.

a. Supervised Learning

Supervised learning merupakan salah satu metode yang ada di dalam *machine learning* [15]. Di dalam *supervised learning*, algoritma dibangun berdasarkan *dataset* yang sudah diberi label [16]. Tujuan dari *supervised learning* adalah memprediksi nilai label *(t)* untuk input *(x)* yang tidak ada di dalam set pelatihan. Dengan kata lain supervised learning bertujuan untuk menggeneralisasi pengamatan dalam kumpulan *data (D)* ke input baru [17] seperti pada persamaan 2.2.

$$D = \{xn, tn\}N\ n\ = 1 \tag{2.2}$$

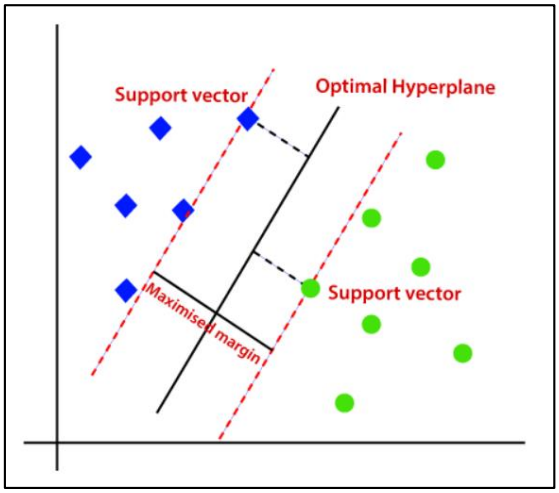
Algoritma *supervised learning* memanfaatkan *dataset* uji dari *data input* dan *output*. Algoritma ini mempelajari hubungan antara *data input* dan *data output* dari *dataset* uji dan kemudian menggunakan hubungan ini untuk memprediksi *output* pada data baru. Salah satu tujuan *supervised learning* yang paling umum adalah klasifikasi. Salah satu tujuan klasifikasi adalah untuk menggunakan informasi yang telah dipelajari untuk memprediksi keanggotaan kelas tertentu [18].

Algoritma yang masuk kedalam kelompok algoritma *supervised learning* adalah *Decision Tree*, *K-Nearest Neighbor*, *Naive Bayes Classifier, Artificial Neural Network, Support Vector Machine*, dan *Fuzzy K-Nearest Neighbor*.

Logistic Regression, *Support Vector Machine*, *Naive Bayes*, dan *Neural Network* masuk ke dalam grup klasifikasi *(classification)* algoritma *supervised learning* [19].

1. Support Vector Machine

Support Vector Machine (SVM) pertama kali dikenalkan oleh Cortes and Vapnik [17]. SVM adalah sebuah algoritma *supervised learning* yang menganalisis data, mengenali pola, yang digunakan didalam *classification* dan *regression analysis*. SVM memiliki reputasi yang baik dalam hal klasifikasi [20] dan secara efisien melakukan klasifikasi *non-linier* menggunakan apa yang disebut trik kernel, secara implisit memetakan inputnya ke dalam ruang fitur berdimensi tinggi. *Support Vector Machine* akan memaksimalkan margin diantara dua kelas [14] seperti pada gambar 2.1.



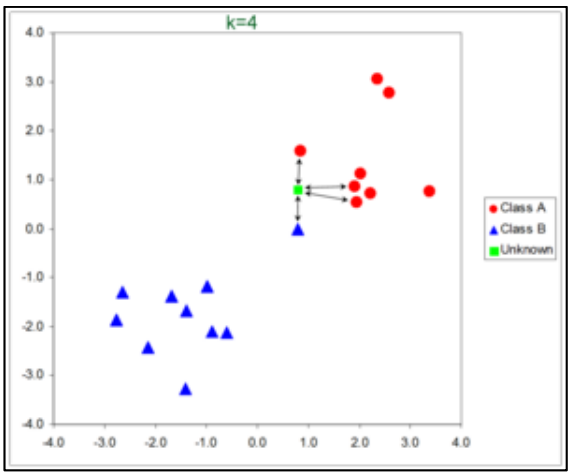
Gambar 2. 1 Support Vector Machine

2. K-Nearest Neighbor

K-Nearest Neighbor (KNN) merupakan model pengenalan pola yang dapat digunakan untuk klasifikasi *(classification)* maupun regresi *(regression)*.

K-NN didasarkan pada gagasan bahwa pola terdekat dengan pola target x_1, yang dicari labelnya, akan memberikan informasi berguna dalam pemberian

label yang diilustrasikan pada gambar 2.2. KNN memberikan label kelas dari sebagian besar informasi label yang berguna [21].



Gambar 2. 2 K –Nearest Neighbor

3. Naive Bayes Classification

Naive Bayes (NB) merupakan metode pengklasifikasian dengan metode probabilistik dan statistik yang pertama kali dikemukakan oleh ilmuwan inggris Thomas Bayes. NB masuk kedalam kelompok *supervised learning* dengan berdasarkan teori Bayes [22] seperti pada persamaan 2.3.

$$P(c|d) = p\,(c)\ \coprod_{1\leq k\leq nd^{\leq n}} \quad p(t_k\,|\mathbf{c}) \tag{2.3}$$

Naive Bayes Classification (NBC) biasanya digunakan dalam klasifikasi teks dan dokumen [23]. Dua varian algoritma *Naive Bayes* yang digunakan dalam klasifikasi sentimen adalah *Bernoulli* dan *MultinomialNB* [24].

b. Text Mining

Text mining atau *Text Analytics* merupakan sebuah penggabungan disiplin ilmu *language science* dan *computer science* dengan menggunakan ilmu statistik dan

teknik mesin *learning*. *Text mining* digunakan untuk analisis teks dan mengubahnya ke dalam bentuk yang terstruktur [16].

Identifikasi topik, pengelompokan teks, *spam filtering*, dan analisa sentimen merupakan bagian dari *text mining*. Konsep penting di dalam *text mining* adalah *Bag of Words*, yang merupakan untuk mempresentasikan dokumen dengan menyatakannya dalam bentuk vektor dari kata-kata yang terkandung di dalamnya. Bag of words merupakan representasi dokumen yang selanjutnya dapat diolah untuk berbagai tujuan seperti klasifikasi dokumen dan lain-lainnya.

c. Sentiment Analysis

Sentimen Analisis adalah studi komputasi mengenai sikap, emosi, pendapat, penilaian, pandangan dari sekumpulan teks yang difokuskan adalah mengekstraksi, mengidentifikasi atau menemukan karakteristik sentimen dalam unit *text* menggunakan metode *Natural Language Processing* , statistik, atau *machine learning* [25]. Sentimen analisis banyak digunakan di dalam dunia bisnis, kesehatan, dan politik.

Tools yang banyak menyediakan dukungan untuk berbagai tugas pada *Natural Language Processing (NLP)* termasuk analisa sentimen adalah *TextBlob*. *Textblob* menggunakan bahasa *Python*, yaitu bahasa pemrograman tingkat tinggi yang berorientasi objek dengan semantik yang dinamis.

Analisa sentimen oleh *Textblob* berdasarkan nilai *polarity* dan *subjectivity*. Nilai *polarity* terletak pada rentang [-1, 1], -1 menunjukkan sentimen negatif , 1 menunjukkan sentimen positif, dan 0 menunjukkan sentimen netral. Nilai *Subjectivity* terletak pada rentang [0, 1]. *Subjectivity* pada umumnya mengacu pada pendapat, emosi, atau penilaian pribadi, sedangkan penilaian objektif mengacu pada informasi fakta. Sentimen yang lebih objektif daripada subjektif menerima skor yang lebih rendah, yang menunjukkan kemungkinan fakta yang lebih akurat.

d. TF/IDF

Term Frequency – Inverse Document Frequency (TF-IDF) merupakan teknik dalam memberikan bobot pada kata terhadap dokumen seperti pada persamaan 2.3 berikut [16].

$$TF = f_{t,d} \tag{2.3}$$

TF adalah frekuensi (*f*) dari (*t*) di dalam dokumen (*d*).

Rumus dari *IDF* dengan penskalaan logarithmic yang paling umum dengan persamaan 2.4 dibawah ini.

$$IDF = log(N/|\{d \in D : t \in d\}|) \tag{2.4}$$

Dengan *N* jumlah total dokumen di dalam *Corpus* dan $(|\{d \in D : t \in d\}|)$ menjadi jumlah dokumen *(d)* dimana *(t)* muncul.

$$\frac{TF}{IDF} = f_{t,d}/log(N/|\{d \in D : t \in d\}|) \tag{2.5}$$

TF-IDF pada persaman 2.5 memperhitungkan semua dokumen. *IDF* memberikan gambaran tentang seberapa umum kata tersebut di seluruh *Corpus*: semakin tinggi frekuensi dokumen, semakin umum, dan semakin banyak kata yang kurang informatif. Beberapa langkah dapat dilakukan untuk memanipulasi data sebelum sampai ke tahapan *Bag of words* yaitu proses *tokenization*.

e. API Twitter

API Twitter merupakan fitur yang disediakan oleh *Twitter* sehingga pihak pengembang aplikasi dapat mengakses informasi *web Twitter* seperti *tweet*, *users, direct messages*, *lists, trends,* media, dan *places* [26].

f. WordCloud

Word Cloud merupakan teknik untuk memvisualisasikan kata yang sering muncul dalam teks di mana ukuran kata mewakili frekuensinya.