BAB II
TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Penelitian terdahulu menjadi unsur yang penting dalam penelitian dikarenakan menjadi acuan oleh peneliti dalam melakukan dan menyusun penelitian. Dengan adanya penelitian terdahulu, maka peneliti dapat mengetahui keterkaitan dengan penelitian yang akan dilakukan nantinya. Selain itu, juga membantu untuk menghindari adanya duplikasi dari penelitian yang akan dilakukan.

Dalam penelitian tahun 2021 oleh Shania Gabriela Kaparang, Daniel R. Kaparang, dan Vivi P. Rantung berjudul Menganalisis New Normal Sentiment Selama Covid-19 Menggunakan Algoritma Naive Bayes [9]. Penelitian ini bertujuan untuk memodelkan analisis opini publik terkait kebijakan new normal pemerintah pada masa pandemi Covid-19 di Indonesia. Hasil kinerja classifier meliputi akurasi 80,37%, presisi 87,38%, recall 82,57%, dan ukuran F 84,91%, dan hasil klasifikasi dari 5194 tweet adalah sentimen positif, 2908 tweet dikategorikan sebagai tweet berkategori sentimen negatif.

Penelitian berikutnya oleh Fajar Ratnawati yang berjudul "Implementasi Algoritma Naive Bayes Terhadap Analisis Sentimen Opini Film Pada Twitter" pada tahun 2018 [10]. Penelitian ini menggunakan metode Naive Bayes untuk menganalisis pendapat pengguna Twitter tentang film dengan akurasi 90%, recall 90%, dan presision 90% dalam melaporkan skor f1 90%. Implikasi dari penelitian lanjutan ini adalah semakin banyak data latih yang digunakan maka semakin akurat hasilnya dan semakin akurat sistem dalam mengenali data uji.

Diana Ikasari, Yuliana Fajarwati, dan Widiastuti melakukan studi tahun 2020 berjudul Menganalisis dan Mengklasifikasikan Sentimen Tweet Berbahasa Indonesia Angkutan Umum MRT Jakarta Menggunakan Naive Bayes Classifiers [11]. Dalam penelitian ini, berdasarkan hasil tweet dari Twitter MRT Jakarta, kami akan menjelaskan apakah ada tren positif atau negatif dalam tren reaksi pengguna di MRT Jakarta. Akurasi sistem untuk menganalisis sentimen tweet yang terdapat pada Twitter MRT Jakarta adalah 95,88%.

Sebuah penelitian berjudul "Analisis Sentimen terhadap Joko Widodo di Media Sosial Twitter Menggunakan Algoritma Klasifikasi Naive Bayesian" dilakukan oleh Yonathan Sari Mahardhika dan Eri Zuliarso pada tahun 2018 [12]. Pengklasifikasi naif Bayes yang digunakan dalam penelitian ini mampu mengklasifikasikan tweet dengan sentimen negatif dan positif dengan akurasi 97%. Proposal penelitian lebih lanjut harus menentukan kinerja algoritma classifier naive Bayes dibandingkan dengan algoritma klasifikasi lainnya.

Penelitian oleh M. Fitra Rizki, Karina Auliasari, Renaldi Primaswara Prasetya, 2021, "Analisis Sentimen Cyberbullying di Media Sosial Twitter Menggunakan Metode Support Vector Machine" [13]. Pada penelitian ini, proses analisis sentimen pengguna yang melakukan cyberbullying di media sosial Twitter dilakukan dengan mengembangkan sistem berbasis web yang mengklasifikasikan sentiment tersebut menggunakan metode support vector machine. Dari hasil pengujian dengan menggunakan 100 data tweet, diperoleh hasil evaluasi klasifikasi dengan nilai pengenalan 64%, akurasi 58%, dan tingkat akurasi 70% dengan menggunakan metode matriks konfusi. Dengan menggunakan 200 data dokumen tweet, didapatkan 71% recall, 63% akurasi, dan tingkat akurasi 75%.

Fadhilah Dwi Ananda dan Yoga Pristianto melakukan penelitian pada tahun 2021 berjudul “Menganalisis Sentimen Pengguna Twitter terhadap Penyedia Layanan Internet Menggunakan Algoritma Support Vector Machine” [14]. Penelitian ini menerapkan algoritma support vector machine yang digunakan untuk menentukan peringkat sentimen pengguna Twitter terhadap layanan internet Biznet. Kernel yang digunakan adalah kernel linier dan RBF. Tes dijalankan dalam tiga skenario. 800 data dengan akurasi kernel linear 76,25% dan RBF 76,25% untuk skenario 1, akurasi linear kernel 84,44% dan RBF 85,55% untuk skenario 2, 1000 data dengan akurasi kernel linear 90% dan 88% RBF untuk skenario 3 . Berdasarkan hasil pengujian yang dilakukan oleh algoritma SVM dengan kernel linear dan RBF, diperoleh hasil evaluasi kinerja yang hampir sama dalam hal akurasi, presisi dan recall.

Kajian tahun 2019 oleh Arsya Monica Pravina, Imam Cholissodin, dan Putra Pandu Adikara berjudul “Analisis Sentimen Opini Maskapai pada Dokumen Twitter Menggunakan Algoritma Support Vector Machine (SVM)” [15]. Dengan menggunakan metode support vector machine, penelitian ini menghasilkan skor akurasi sebesar 40%, presisi sebesar 40%, recall sebesar 100%, dan fMeasure sebesar 57,14%. Data dapat diklasifikasikan untuk studi lebih lanjut klasifikasikan menjadi tiga kelas: positif, negatif, dan netral, dan implementasikan metode klasifikasi tambahan.

Dery Anjas Ramadhan dan Erwin Budi Setiawan S.Si., MT melakukan penelitian pada tahun 2019 berjudul “Menganalisis Program Sctv Sentimen Twitter Menggunakan Metode Naive Bayes dan Support Vector Machines”[16] . Dalam penelitian ini, kami menggunakan dua metode untuk menemukan metode yang memberikan hasil paling akurat, yaitu Support Vector Machines untuk Semua Program Acara mencapai akurasi 88,57%, sedangkan Naive Bayes mencapai akurasi 66,09%, yang telah terbukti mencapai akurasi yang baik.

Tabel 2. 1 Penelitian Terdahulu Terkait Analisis Sentimen

No.	Penulis	Judul Penelitian	Metode Penelitian	Hasil Penelitian
1	Shania Gabriela Kaparang, Daniel R. Kaparang, Vivi P. Rantung (2021)	Analisis Sentimen New Normal Pada Masa Covid-19 Menggunakan Algoritma Naïve Bayes	*naive bayes*	Hasil perhitungan akurasi 80,37%, presisi 87,38%, recall 82,57%, dan F-measure 84,91%. Berdasarkan hasil klasifikasi, 5194 tweet diklasifikasikan sebagai sentimen positif dan 2908 tweet diklasifikasikan sebagai sentimen negatif.
2	Fajar Ratnawati (2018)	Implementasi Algoritma Naive Bayes Terhadap Analisis Sentimen Opini Film Pada Twitter	*naive bayes*	Hasil pengujian nilai akurasi 90%, presisi 92%, recall 90%, dan nilai F skor 90%.
3	Diana Ikasari, Yuliana Fajarwati dan Widiastuti (2020)	Analisis Sentimen Dan Klasifikasi Tweets Berbahasa Indonesia Terhadap Transportasi Umum Mrt Jakarta	*naive bayes*	Analisis yang dilakukan memiliki akurasi sistem 95,88%, presisi positif 70%, dan presisi negatif 30%.

No.	Penulis	Judul Penelitian	Metode Penelitian	Hasil Penelitian
		Menggunakan Naive Bayes Classifier		
4	Yonathan Sari Mahardhika, Eri Zuliarso (2018)	Analisis Sentimen Terhadap Pemerintahan Joko Widodo pada Media Sosial Twitter Menggunakan Algoritma Naives Bayes Classifier	*naive bayes*	Metode Naive Bayes Classifier yang mengklasifikasikan tweet dengan sentimen negatif dan positif dengan akurasi 97%.
5	Muh. Fitra Rizki, Karina Auliasari, Renaldi Primaswara Prasetya (2021)	Analisis Sentiment Cyberbullying Pada Sosial Media Twitter Menggunakan Metode Support Vector Machine	*support vector machine*	Penggunaan 100 data tweet menghasilkan nilai recall 64%, precision 58% dan tingkat accuracy sebesar 70%. Penggunaan 200 data tweet menghasilkan nilai recall 71%, precision 63% dan tingkat accuracy sebesar 75%.
6	Fadhilah Dwi Ananda, Yoga Pristyanto (2021)	Analisis Sentimen Pengguna Twitter Terhadap Layanan Internet Provider	*support vector machine*	Untuk penelitian ini menjalankan tiga skenario pengujian dengan jumlah data yang berbeda. Algoritma SVM dengan

No.	Penulis	Judul Penelitian	Metode Penelitian	Hasil Penelitian
		Menggunakan Algoritma Support Vector Machine		RBF dan kernel linier memberikan nilai akurasi yang sangat baik. Skenario 1 Kernel linier 76,25% dan RBF 76,25% Skenario 2 Kernel linier 84,44% dan RBF 85,55% Skenario 3 Linear Kernel 90% dan RBF 88%
7	Arsya Monica Pravina, Imam Cholissodin dan Putra Pandu Adikara (2019)	Analisis Sentimen Tentang Opini Maskapai Penerbangan pada Dokumen Twitter Menggunakan Algoritme Support Vector Machine (SVM)	*support vector machine*	Hasil akurasi 40%, precision 40%, 100% recall, dan f- measure sebesar 57,14%.

No.	Penulis	Judul Penelitian	Metode Penelitian	Hasil Penelitian
8	Dery Anjas Ramadhan dan Erwin Budi Setiawan S.Si., M.T (2019)	Analisis Sentimen Program Acara di Sctv pada Twitte Menggunakan Metode Naïve Bayes dan Support Vector Machine	*naive bayes dan support vector machine*	Labelling terbaik didapatkan dengan cara Labelling Otomatis sedangkan untuk metode dengan hasil akurasi terbaik yaitu Support Vector Machine dari seluruh Program Acara didapatkan hasil akurasi 88,57% sedangkan Naïve Bayes menghasilkan akurasi 66,09%

Berdasarkan penelitian terdahulu pada Tabel 2.1 diketahui bahwa metode *naive bayes classifier* dan *support vector machine* dapat digunakan dalam analisis sentimen yang kemudian menghasilkan nilai akurasi yang cukup bagus. Penelitian ini menerapkan metode *naive bayes classifier* dan *support vector machine* pada analisis sentimen dan sebagai perbedaan penelitian tugas akhir ini terletak pada studi kasus yang diterapkan.

2.2 Dasar Teori

Berikut ini terdapat beberapa dasar teori yang menjadi landasan dari penelitian yang dilakukan:

2.2.1 Bullying

Bullying ialah perilaku atau tindakan yang dilakukan secara sengaja bertujuan untuk memberikan gangguan secara fisik maupun psikologis kepada orang lain [17]. Tindakan *bullying* dapat menimbulkan efek negatif bagi korban maupun pelaku. Perilaku *bullying* dapat menyebabkan depresi, masalah kesehatan bahkan sampai menyakiti diri sendiri.

2.2.2 Twitter

Twitter yaitu media sosial yang menjadi salah satu dari beberapa bagian perkembangan teknologi terutama pada media komunikasi yang bertujuan agar pengguna bisa memberikan pendapat, kritik, ekspresi, aspirasi, dan dapat bertukar informasi tentang informasi yang sedang ramai menjadi perbincangan, tanpa terbatas waktu dan ruang [18]. Sehingga, pendapat atau informasi tersebut dapat tersampaikan secara cepat dan langsung, memposting pendapat dalam twitter ini disebut sebagai istilah tweet dengan memuat aturan satu tweet berisi maksimal 280 karakter [19]. Hal ini menjadikan *twitter* menjadi salah satu sumber data teks yang dapat diambil dan dimanfaatkan dalam berbagai keperluan penelitian terutama dalam bidang teknologi informasi.

2.2.3 Text Mining

Text Mining merupakan sebuah teknik untuk memecahkan masalah klasifikasi, pengelompokan, ekstraksi informasi dan pencarian informasi. Sumber data yang digunakan dalam *text mining*

yaitu kumpulan dari teks yang terstruktur atau tidak terstruktur [20]. *Text mining* bertujuan untuk mengekstraksi informasi yang berguna dari dokumen. Tahapan dari text mining meliputi pemrosesan teks awal *(text preprocessing),* transformasi teks *(text transformation)* pemilihan fitur, dan penemuan pola teks atau penambangan data *(pattern discovery).*

2.2.4 Klasifikasi

Klasifikasi adalah proses pengelompokan objek memiliki sifat atau karakteristik yang sama di beberapa kelas [21]. Klasifikasi bertujuan untuk dapat menentukan kelas dari sebuah objek yang labelnya tidak diketahui. Untuk memperoleh tujuan, maka dibentuk model yang dapat membedakan data ke dalam kelas berbeda berdasarkan aturan atau fungsi tertentu.

2.2.5 Analisis Sentimen

Analisis sentimen yaitu bagian dari bidang text mining yang berguna mengolah data opini atau pandangan publik yang menghasilkan kesimpulan berupa persentase sentimen positif, negatif, atau netral [22]. Tujuan utama dari *sentiment analysis* yaitu untuk menganalisis suatu ulasan yang berisi opini ataupun sudut pandang berdasarkan penilaian terhadap suatu objek apakah setiap ulasan tersebut bersifat baik ataupun tidak. Ada nilai sentimen positif, negatif dan netral yang banyak digunakan dalam analisis sentimen. Nilai ini dapat digunakan sebagai parameter keputusan.

2.2.6 Text Preprocessing

Text preprocessing adalah tahapan transformasi data dari dokumen teks yang tidak terstruktur menjadi data terstruktur.

Preprocessing adalah proses pertama teks yang mempersiapkannya menjadi data yang dapat diproses lebih lanjut [23].

a. *Cleaning*

Cleaning adalah proses menghilangkan kata-kata yang tidak diinginkan untuk mengurangi *noise*. Kata-kata yang dihapus pada langkah ini meliputi tagar (#), URL, nama pengguna (@nama pengguna), titik (.), koma (,), dan tanda baca lainnya.

b. *Case Folding*

Case folding merupakan sebuah langkah untuk membentuk kata sehingga mereka memiliki bentuk yang sama. Sebuah tweet biasanya dituliskan dengan huruf kapital dan huruf kecil sekaligus. Sebuah kata atau kalimat dapat dirubah huruf kecil *(lower case)* semua melalui tahap case folding.

c. *Tokenizing*

Tokenizing merupakan tahap *preprocessing* proses penguraian data teks menjadi potongan yang lebih kecil (token) seperti kata dan frasa. Pada langkah ini dilakukan pemecahan kata dalam teks dibagi jadi beberapa kata yang dipisahkan spasi.

Contoh penerapan dari tokenizing “kecewa kalo kim garam beneran pelaku bullying.” menjadi [kecewa], [kalo], [kim], [garam], [beneran], [pelaku], [bullying].

d. Normalization

Normalisasi bahasa dilakukan untuk kata yang tidak standar atau baku. Langkah ini menunjukkan gaya penulisan setiap kata sesuai KBBI. Proses ini dilakukan dengan cara

mencocokkan setiap kata dalam dokumen dengan kata dalam kamus.

e. *Stopword Removal*

Stopword Removal adalah langkah menghilangkan kata-kata yang tidak diinginkan dalam teks menggunakan *stopwords* yang telah dibuat. Contoh kata umum yaitu 'dan', 'yang' dan 'itu' dalam daftar stopword.

f. *Stemming*

Stemming adalah langkah untuk mengembalikan kata yang berimbuhan kembali ke bentuk aslinya. Tahap ini menghapus kata-kata imbuhan pada teks untuk memperoleh kata dasarnya. Adapun contoh kata “berkurang” setelah *stemming* berubah ke kata “kurang”.

2.2.7 *Term frequency inverse document frequency (TF-IDF)*

Term frequency inverse document frequency (TF-IDF) ialah adalah metode pembobotan untuk menentukan ukuran kata *(term)* dalam suatu dokumen dengan memberikan bobot pada setiap kata [24].

Perhitungan bobot dalam setiap dokumen dapat dilakukan menggunakan persamaan (2.1).

$$\boldsymbol{W_{i,j} = tf_{i,j}\, x\, log\, (\frac{N}{df_i})} \qquad \textbf{(2.1)}$$

Keterangan :

$W_{i,j}$: Bobot kata i pada dokumen j

$tf_{i,j}$: Frekuensi kata i pada dokumen j

N : Jumlah dokumen

df_i : Jumlah dokumen yang terdapat kata i

2.2.7 Naive Bayes

Naive Bayes adalah metode pengelompokan data untuk memprediksi probabilitas kelas [25]. Metode ini banyak digunakan dalam penelitian karena sangat mudah diatur, mudah diinterpretasikan, dan tidak memerlukan skema estimasi parameter iteratif yang kompleks. Pada penelitian ini, penerapan metode naive bayes classifier yaitu dilakukan perhitungan nilai probabilitas dari label class atau variabel dari dataset terhadap masukan yang diberikan. Kemudian bandingkan nilai probabilitas untuk mencari nilai probabilitas tertinggi. Label kelas dengan nilai probabilitas tertinggi digunakan sebagai label kelas untuk data masukan. Adapun Persamaan Teorema Bayes pada persamaan (2.2) berikut:

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad \textbf{(2.2)}$$

Keterangan:

X : data yang belum diketahui kelasnya

H : hipotesis data suatu kelas spesifik

$(H|X)$: probabilitas hipotesis H diberikan kondisi X (probabilitas posterior)

(H) : probabilitas hipotesis H (probabilitas prior)

$(X|H)$: probabilitas X berdasarkan kondisi pada hipotesis

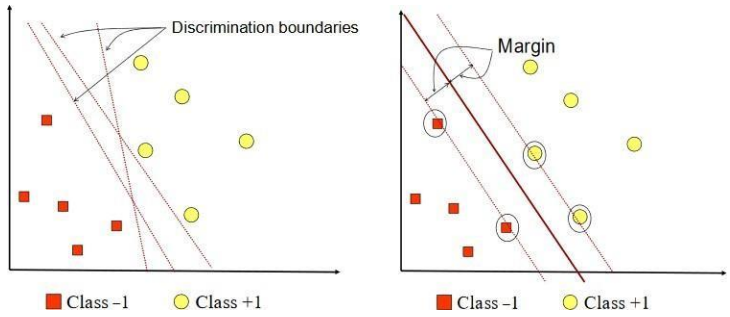
H (X) : probabilitas

2.2.8 Support Vector Machine

Support vector machine (SVM) termasuk dalam kategori pembelajaran mesin yang diawasi dan dapat memprediksi kelas

berdasarkan hasil proses pelatihan. Metode *support vector machine* memungkinkan perhitungan masalah linier dengan menerapkan transformasi matematis menggunakan fungsi kernel [26]. Cara kerja *support vector machine* adalah mencari *hyperplane* terbaik yang memberikan jarak atau batas antara dua kelas. Jarak antara titik data yang paling dekat dengan hyperplane disebut margin. *Support vector* adalah titik yang paling dekat dengan *hyperplane.*

Sebuah *hyperplane* bertindak sebagai pemisah antara dua kelas data yang berbeda sebagai contoh positif (+1) dan negatif (-1). Seperti yang terlihat pada Gambar 2.1, data positif (+1) dilambangkan dengan warna kuning dan data negatif (-1) dengan warna merah. Proses *support vector machine* ditunjukkan pada Gambar 2.1. Diagram di sebelah kiri menunjukkan bahwa ada beberapa kemungkinan garis pemisah (discrimination boundaries) pada *Support Vector Machine* melakukan set data. Di sisi lain, grafik di sebelah kanan menunjukkan batas diskriminasi dengan margin terbesar. Batas atau pemisah adalah jarak antara dua kelas data terdekat dalam bidang *hyperplane.* Sebuah *hyperplane* dengan margin optimal yang memberikan generalisasi untuk hasil klasifikasi yang lebih baik.



Gambar 2. 1 Proses SVM dalam menemukan hyperplane

2.2.9 Confusion Matrix

Confusion matrix atau *error matrix* adalah metode yang digunakan untuk memberikan informasi tentang hasil perhitungan akurasi menggunakan konsep data mining [27]. *Confusion matrix*, dalam bentuk tabel matriks seperti yang ditunjukkan pada Tabel 2.2, menunjukkan kinerja model klasifikasi pada sekumpulan data uji dengan nilai sebenarnya yang diketahui.

Tabel 2. 2 *Confusion matrix*

confusion matrix		***actual***	
		positive (0)	***negative (1)***
predicted	***positive (0)***	*TP*	*FP*
	negative (1)	*FN*	*TN*

Keterangan:

TP *(true positive)* : data positif diprediksi benar sebagai positif

TN *(true negative)* : data negatif diprediksi benar sebagai negatif

FP *(false positive)* : data dengan nilai negatif salah diprediksi sebagai positif

FN *(false negative)* : data dengan nilai positif salah diprediksi sebagai negatif

Terdapat beberapa pesamaan rumus untuk menghitung performa klasifikasi. Beberapa performa klasifikasi dari *confusion*

matrix yaitu *accuracy*, *recall*, *precision* dan *f1 score.* Akurasi menghitung semua istilah yang diprediksi menggunakan nilai yang benar untuk semua istilah yang diprediksi. Akurasi menghitung seberapa baik model dapat mengklasifikasikan data dengan benar. Nilai akurasi diperoleh dari persamaan (2.3).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} X\ 100\% \qquad \textbf{(2.3)}$$

Perhitungan pada kondisi benar, yaitu kelas aktual dan kelas prediksi yang sama (positif) untuk semua kondisi yang diprediksi positif disebut dengan *precision*. Nilai *precision* diperoleh dengan persamaan **(2.4)**.

$$Precision = \frac{TP}{TP+FP} X\ 100\% \qquad \textbf{(2.4)}$$

Recall yaitu perhitungan pada kondisi benar, kelas data positif untuk semua kondisi aktual dengan nilai positif. Nilai *recall* didapat dengan persamaan **(2.5)**.

$$Recall = \frac{TP}{TP+FN} X\ 100\% \qquad \textbf{(2.5)}$$

Skor f1 adalah perhitungan untuk menemukan nilai rata-rata untuk penilaian *recall* dan *precision*. Skor f1 didapatkan oleh persamaan **(2.6)**.

$$F1\ Score = 2X \frac{Recall\ X\ Precision}{Recall+\ Precision} \qquad \textbf{(2.6)}$$