

BAB II

TINJAUAN PUSTAKA

2.1. Studi Sebelumnya

Bagian ini menyajikan tentang penelitian-penelitian terdahulu yang relevan sebagai dasar acuan dalam penelitian ini. Penelitian-penelitian sebelumnya memberikan wawasan yang berharga dan metodologi yang dapat diadaptasi untuk analisis yang lebih mendalam dalam konteks yang sedang diteliti. Prastyo, dkk [16] pada tahun 2020 melakukan penelitian mengenai analisis sentimen terhadap teks berbahasa Indonesia. Penelitian ini bertujuan untuk membandingkan performa empat fungsi kernel (Polynomial, Sigmoid, RBF, dan Linear) pada algoritme SVM menggunakan TF-IDF. Metode yang digunakan adalah Support Vector Machine. Hasil penelitian ini menunjukkan bahwa kernel Linear memberikan hasil terbaik dengan tingkat akurasi yang konsisten sebesar 96% pada jumlah fitur yang tinggi (4.000 fitur). Sedangkan kernel Radial Basis Function (RBF) memberikan hasil terbaik dengan tingkat akurasi yang konsisten sebesar 96% pada jumlah fitur sebanyak 1.000, 2.000, dan 3.000.

Penelitian selanjutnya dilakukan oleh Husada & Paramita [10] pada tahun 2021 yang membahas analisis sentimen terhadap maskapai penerbangan di platform Twitter. Tujuan pada penelitian ini adalah menerapkan analisis sentimen untuk menilai opini publik mengenai maskapai penerbangan di Twitter. Metode yang digunakan adalah Support Vector Machine. Akurasi yang dihasilkan mencapai 84.37%.

Selanjutnya penelitian dilakukan oleh Cahyani & Patasik [17] pada tahun 2021 melakukan studi mengenai klasifikasi emosi pada teks, khususnya pada data tweet Commuter Line dan Transjakarta. Penelitian ini bertujuan membandingkan

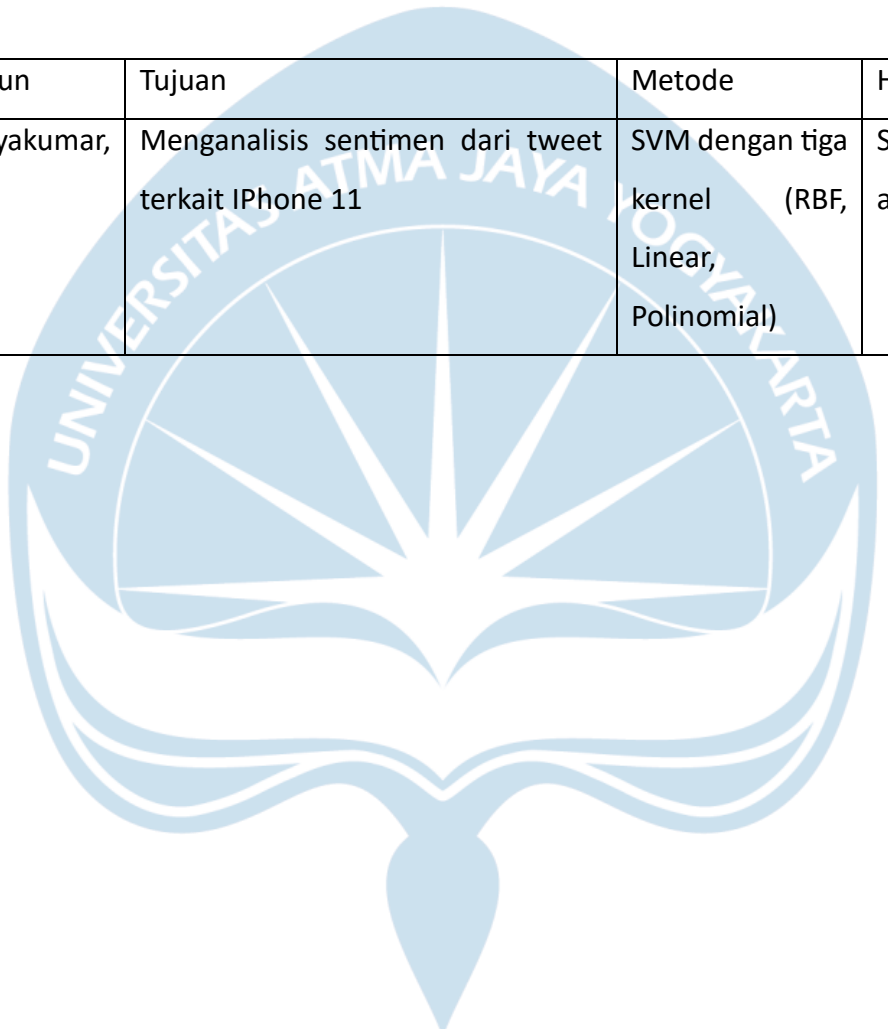
kinerja model TF-IDF dan Word2Vec dalam merepresentasikan fitur untuk klasifikasi emosi teks. Metode klasifikasi yang digunakan adalah Support Vector Machine (SVM) dan Multinomial Naïve Bayes (MNB). Hasilnya menunjukkan bahwa SVM dengan TF-IDF memberikan akurasi tertinggi dalam klasifikasi, diikuti oleh MNB dengan TF-IDF, dan kemudian SVM dengan Word2Vec. Penelitian ini menyimpulkan bahwa TF-IDF memiliki performa yang lebih baik dibandingkan Word2Vec dalam klasifikasi emosi teks serta meningkatkan hasil klasifikasi dibandingkan penelitian sebelumnya yang menggunakan MNB dengan TF-IDF.

Styawati, dkk [18] pada tahun 2021, melakukan penelitian analisis sentimen terhadap program Kartu Prakerja. Penelitian ini bertujuan mengukur tingkat akurasi algoritme SVM dalam menganalisis sentimen masyarakat Indonesia terhadap program tersebut. Metode yang digunakan adalah Support Vector Machine, dan hasilnya menunjukkan bahwa SVM dengan kernel linear mencapai akurasi 98,67%, sedangkan kernel RBF mencapai akurasi 98,34%.

Driyani & Jeyakumar [19] pada tahun 2021 meneliti analisis sentimen ulasan smartphone iPhone 11. Penelitian ini bertujuan untuk menganalisis sentimen tweet mengenai iPhone 11. Metode yang digunakan adalah Support Vector Machine. Hasil penelitian menunjukkan SVM dengan kernel RBF memiliki akurasi tertinggi sebesar 91,87%.

Tabel 2. 1. Perbandingan Studi Sebelumnya

No.	Peneliti & Tahun	Tujuan	Metode	Hasil
1.	Prastyo dkk., 2020 [16]	Membandingkan performa empat fungsi kernel (Polynomial, Sigmoid, RBF, dan Linear) pada algoritme SVM menggunakan TF-IDF	SVM	Kernel Linear memberikan hasil terbaik dengan tingkat akurasi yang konsisten sebesar 96% pada jumlah fitur yang tinggi.
2.	Husada & Paramita, 2021 [10]	Menerapkan analisis sentimen untuk menilai opini publik mengenai maskapai penerbangan di Twitter	SVM	Akurasi tertinggi yang dihasilkan adalah SVM dengan kernel RBF sebesar 84.37%.
3.	Cahyani & Patasik, 2021 [17]	Membandingkan kinerja model TF-IDF dan Word2Vec dalam merepresentasikan fitur untuk klasifikasi emosi teks	SVM dan Multinomial Naïve Bayes (MNB).	SVM dengan TF-IDF memberikan akurasi tertinggi diikuti oleh MNB dengan TF-IDF, dan terakhir SVM dengan Word2Vec.
4.	Styawati dkk., 2021 [18]	Mengukur tingkat akurasi algoritme SVM dalam menganalisis sentimen masyarakat Indonesia terhadap program Prakerja	SVM	Algoritme SVM dengan kernel Linear menghasilkan akurasi sebesar 98.67% dan kernel RBF memiliki akurasi sebesar 98.34%.



No.	Peneliti & Tahun	Tujuan	Metode	Hasil
5.	Driyani & Jeyakumar, 2021 [19]	Menganalisis sentimen dari tweet terkait iPhone 11	SVM dengan tiga kernel (RBF, Linear, Polinomial)	SVM dengan kernel RBF memiliki akurasi tertinggi sebesar 91,87%.

2.2. Dasar Teori

2.2.1. Anggaran Pendidikan

Anggaran Pendidikan adalah dana yang dikeluarkan oleh pemerintah untuk memenuhi kebutuhan pendidikan. Anggaran dana pendidikan sudah diatur dalam Undang-Undang Dasar tahun 1945 Pasal 31 [6]. Pemerintah menetapkan alokasi anggaran dana pendidikan sebesar 20% dari total APBN [7]. Anggaran ini digunakan untuk berbagai program seperti peningkatan akses pendidikan, kualitas tenaga pendidik, dan penyediaan beasiswa [8].

2.2.2. Media Sosial

Media sosial sarana digital yang dimanfaatkan oleh masyarakat untuk bertukar informasi, berdiskusi, serta menjalin komunikasi [10]. Hal ini terlihat dari banyaknya platform media sosial yang tersedia dan digemari oleh pengguna. Beberapa platform yang umum digunakan antara lain X, YouTube, Instagram, Threads, Facebook, dan TikTok.

2.2.3. X

X adalah salah satu platform media sosial yang sering digunakan oleh masyarakat [11]. Berdasarkan laporan Data Reportal, pengguna aktif X di Indonesia mencapai 24,69 juta dan menempatkannya sebagai media sosial terpopuler keenam di Indonesia [12]. Selain itu, X dikenal sebagai ruang diskusi terbuka di mana opini publik tentang isu-isu sosial, termasuk kebijakan pemerintah, sering kali dibahas secara luas [10]. Fitur seperti *retweet* dan *trending topic* juga memungkinkan penyebaran opini dengan cepat, sehingga memberikan data yang kaya untuk dianalisis.

2.2.4. *Artificial Intelligence*

Artificial Intelligence atau kecerdasan buatan merupakan teknologi yang memungkinkan mesin atau komputer untuk mensimulasi kemampuan manusia seperti pemahaman, pemecahan masalah dan pengambilan keputusan menggunakan algoritme dan model matematika [20] .

2.2.5. *Machine learning*

Machine learning adalah salah satu cabang dalam *artificial intelligence* yang memungkinkan komputer belajar mandiri tanpa perlu diprogram secara eksplisit. Teknik ini memanfaatkan data *historis* untuk mengidentifikasi pola atau tren guna menghasilkan prediksi di masa depan [20]. *Machine learning* terbagi menjadi dua jenis utama, yaitu *supervised learning* dan *unsupervised learning* [21]. *Supervised learning* menggunakan data berlabel untuk melatih model dalam melakukan klasifikasi atau regresi [21], sementara *unsupervised learning* bekerja dengan data tanpa label untuk menemukan pola tersembunyi dalam *dataset* [21].

2.2.6. *Text Mining*

Text mining adalah proses eksplorasi dan ekstraksi informasi dari teks yang tidak terstruktur menggunakan teknik analisis, pemrosesan bahasa alami (NLP), dan *machine learning* [22]. *Text mining* bertujuan untuk menemukan pola, tren dan informasi dalam kumpulan teks berskala besar [23].

2.2.7. Analisis Sentimen

Analisis sentimen adalah proses menganalisis teks untuk menentukan apakah suatu opini bersifat positif, negatif, atau netral [24]. Namun, analisis sentimen secara manual memiliki beberapa tantangan, seperti subjektivitas dan inkonsistensi akibat bias kognitif dan emosional [25]. Untuk mengatasi hal

ini, pendekatan berbasis *machine learning* atau *artificial intelligence* (AI) semakin banyak diterapkan guna meningkatkan akurasi dan efisiensi analisis sentimen [26]. Studi oleh Haleem, dkk [27] menerangkan bahwa dengan menganalisis data historis dengan menggunakan algoritme *machine learning* dapat mengidentifikasi pola perilaku pelanggan yang membantu perusahaan dalam mengantisipasi kebutuhan dan preferensi pelanggan di masa depan.

2.2.8. *Data Crawling*

Data crawling adalah proses otomatis untuk mengumpulkan data dari internet menggunakan teknik pemrograman [13]. Proses ini sering dilakukan dengan bantuan *web scraping* tools atau API yang disediakan oleh platform tertentu. Namun, dalam pelaksanaannya, terdapat batasan tertentu yang harus diperhatikan, seperti pembatasan kecepatan crawling serta jumlah data yang dapat diambil dalam satu hari [28].

2.2.9. *Data Preprocessing*

Data preprocessing langkah awal dalam analisis data yang bertujuan untuk membersihkan serta menyiapkan data agar siap digunakan dalam proses pemodelan [13].

2.2.10. Pelabelan Data

Pelabelan data adalah proses memberikan kategori atau tag pada data sesuai dengan karakteristik yang ingin dianalisis [29]. Dalam analisis sentimen, data biasanya diberi label sebagai positif, negatif, atau netral [10]. Pelabelan dapat dilakukan secara manual oleh anotator atau menggunakan metode semi-otomatis dengan bantuan model *machine learning* yang sudah dilatih sebelumnya [29].

2.2.11. Term Frequency Inverse Document Frequency (TF-IDF)

TF-IDF merupakan teknik pemberian bobot pada kata yang mempertimbangkan seberapa sering kata tersebut muncul dalam suatu dokumen serta seberapa umum kata tersebut di seluruh dokumen yang ada [17]. *Term Frequency (TF)* menunjukkan frekuensi kemunculan kata dalam satu dokumen, sedangkan *Inverse Document Frequency (IDF)* menunjukkan tingkat kepentingan kata dengan melihat seberapa jarang kata tersebut muncul di keseluruhan dokumen dalam korpus [10]. Formula TF-IDF dihitung sebagai berikut [17]:

Rumus *term frequency (TF)*:

$$tf_t = 1 + \log (tf_t)$$

Di mana:

tf_t adalah jumlah kemunculan istilah t

Rumus *inverse document frequency (IDF)*:

$$idf_t = \log\left(\frac{D}{df_t}\right)$$

Di mana:

idf_t adalah *inverse document frequency* dari istilah t .

D adalah jumlah total dokumen dalam korpus.

df_t adalah jumlah dokumen yang mengandung istilah t .

Sehingga TF-IDF dirumuskan sebagai berikut:

$$w_{t,d} = tf_t \times idf_t$$

Di mana:

$w_{t,d}$ adalah bobot istilah t dalam dokumen d .

tf_t adalah frekuensi istilah t dalam dokumen d (yang sudah dihitung menggunakan logaritma).

idf_t adalah inverse document frequency dari istilah t .

2.2.12. Penanganan *Imbalanced Data*

Imbalanced data adalah situasi ketika proporsi antar kelas dalam sebuah dataset tidak merata [30], dengan salah satu kelas memiliki jumlah sampel yang jauh lebih sedikit dibandingkan kelas lainnya. Ketimpangan ini dapat membuat model machine learning cenderung bias terhadap kelas mayoritas, sehingga performa prediksi pada kelas minoritas menjadi kurang akurat. Pendekatan untuk menangani *imbalanced data* dalam *machine learning* dapat dibagi menjadi tiga kategori utama, yaitu metode *oversampling* yang menambah sampel pada kelas minoritas (seperti SMOTE), metode *undersampling* yang mengurangi sampel pada kelas mayoritas, serta metode berbasis algoritme yang menyesuaikan bobot atau merancang model khusus untuk menangani ketidakseimbangan data.

2.2.13. Klasifikasi

Klasifikasi merupakan salah satu teknik dalam machine learning yang bertujuan untuk mengelompokkan data ke dalam label atau kategori tertentu berdasarkan pola yang dipelajari dari data pelatihan [31]. Model yang digunakan dalam klasifikasi harus dilatih dengan *dataset* berlabel untuk dapat memprediksi kategori dari data baru [32]. Beberapa algoritme yang umum digunakan dalam proses klasifikasi antara lain Naïve Bayes, Decision Tree, dan Support Vector Machine (SVM) [31]. Naïve Bayes adalah metode klasifikasi berbasis probabilitas yang mengacu pada Teorema Bayes dengan asumsi

bahwa setiap fitur bersifat independen. Decision Tree membentuk model berbentuk pohon, di mana setiap simpul merepresentasikan atribut dan cabang menunjukkan keputusan berdasarkan nilai atribut tersebut, sehingga model ini mudah dipahami. Sementara itu, SVM bekerja dengan menentukan hyperplane optimal yang memisahkan kelas data dengan margin maksimum, serta mampu menangani klasifikasi non-linear dengan bantuan teknik kernel [31].

2.2.14. Support Vector Machine

Support Vector Machine (SVM) adalah algoritme *supervised learning* yang digunakan untuk klasifikasi dan regresi [10]. Pada awalnya, SVM dikembangkan untuk menyelesaikan masalah klasifikasi dua kelas atau *binary classification*. Namun seiring dengan perkembangannya, algoritme ini telah dikembangkan untuk menangani klasifikasi multi kelas [18]. Dalam ruang dua dimensi, SVM memisahkan data menggunakan garis lurus sedangkan dalam tiga dimensi berupa bidang datar atau *plane* dan pada dimensi yang lebih tinggi (lebih dari tiga) disebut *hyperplane*. SVM bekerja dengan cara mencari hyperplane yang paling optimal untuk memisahkan data ke dalam berbagai kelas, tanpa memerlukan asumsi tertentu mengenai distribusi data [33]. Untuk mengatasi masalah klasifikasi multi kelas, diperlukan pendekatan yang berbeda dibandingkan dengan klasifikasi dua kelas. Salah satu metode yang umum digunakan adalah One vs One (OvO) [18]. Dalam metode One vs One, dibangun pasangan model SVM biner untuk setiap kombinasi dua kelas. Jika terdapat k kelas, maka diperlukan $k(k-1)/2$ model SVM biner. Setiap model akan membandingkan dua kelas dan menentukan pemisah terbaik (optimal *hyperplane*) antara keduanya [18]. Hasil akhir klasifikasi ditentukan berdasarkan suara terbanyak (*voting*) dari semua model yang dibangun.

Keunggulan utama SVM adalah kemampuannya dalam mengatasi overfitting dengan menggunakan *margin maximization*, sehingga model tetap akurat meskipun terdapat data yang bervariasi [34][35]. Dibandingkan dengan algoritme lain seperti Naïve Bayes yang mengasumsikan independensi fitur [36], SVM lebih efisien dalam memberikan hasil klasifikasi yang akurat pada *dataset* analisis sentimen. SVM pada umumnya memiliki 4 jenis kernel yaitu Polynomial, Linear, Sigmoid, dan Radial Basis Function (RBF) [16].

1. Kernel polynomial memungkinkan pembelajaran terhadap pola non linier dengan menggunakan transformasi polinomial dan cocok untuk data yang sudah di normalisasi.
2. Kernel linear digunakan ketika data dapat dipisahkan secara linier dan efektif untuk data berdimensi tinggi seperti dalam klasifikasi teks.
3. Kernel sigmoid berasal dari konsep jaringan saraf tiruan dan menggunakan fungsi aktivasi sigmoid.
4. Kernel RBF merupakan kernel yang paling populer karena fleksibilitas dan kemampuannya dalam menangani data non linier secara umum.

2.2.15. Confusion Matrix

Confusion matrix merupakan sebuah tabel yang digunakan untuk menilai performa model klasifikasi dengan membandingkan hasil prediksi model terhadap label sebenarnya [16]. Tabel ini terdiri dari empat elemen utama [37]:

- *True Positive (TP)*: Jumlah data yang benar-benar positif dan berhasil diklasifikasikan dengan benar sebagai positif oleh model.

- *True Negative (TN)*: Jumlah data yang benar-benar negatif dan tepat diprediksi sebagai negatif oleh model.
- *False Positive (FP)*: Jumlah data negatif yang keliru diprediksi sebagai positif oleh model (juga dikenal sebagai Kesalahan Tipe I).
- *False Negative (FN)*: Jumlah data positif yang salah diklasifikasikan sebagai negatif oleh model (dikenal juga sebagai Kesalahan Tipe II).

Dari *confusion matrix*, berbagai metrik evaluasi dapat dihitung [37] [38], seperti:

- *Accuracy*: Seberapa banyak prediksi yang benar dari seluruh data. *Accuracy* dihitung dengan rumus sebagai berikut:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- *Precision*: Ketepatan model dalam memprediksi kelas positif. *Precision* dihitung dengan rumus sebagai berikut:

$$Precision = \frac{TP}{TP + FP}$$

- *Recall (Sensitivity)*: Kemampuan model dalam menangkap data positif secara benar. *Recall* dihitung dengan rumus sebagai berikut:

$$Recall = \frac{TP}{TP + FN}$$

- *F1-Score*: *Harmonic mean* antara *precision* dan *recall* untuk keseimbangan evaluasi. *F1-Score* dihitung dengan rumus sebagai berikut:

$$F1\text{-Score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Matriks ini sering digunakan dalam *machine learning* terutama dalam klasifikasi biner dan multi kelas untuk memahami seberapa baik model bekerja dalam mengkategorikan data.

