BAB II
TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Menurut penelitian yang dilakukan Raflizar Deswandi Yahya, dkk tentang analisis sentimen dengan metode SVM. Tujuan penelitian ini adalah untuk mendeteksi ujakan kebencian terkait PEMILU 2024 pada media sosial sosial dalam bahasa indonesia. Berdasarkan hasil pelatihan dan pengujian model diperoleh hasil bahwa sistem dapat mengklasifikasikan kalimat berunsur kebencian dan yang tidak berunsur kebencian. Hasil pengujian berdasarkan 100 baris data diperoleh *recall* sebesar 76%, *precision* sebesar 96% dan nilai keakuratan sebesar 81%.[12].

Menurut Maria Mega Mala Olhang, dkk dalam penelitiannya yang berjudul "Analisis Sentimen Pengguna Twitter Terhadap Covid-19 Di Indonesia Menggunakan Metode *Naïve Bayes Classifier* (NBC)" bertujuan untuk mengembangkan sistem yang dapat melakukan analisis sentimen pada data tentang Covid-19 berdasarkan *keyword* #coronavirusindonesia dan #covid19. Sistem yang dikembangkan berhasil mengelompokan sentimen positif dan negatif dengan nilai keakuratan sebesar 36% untuk total 75 data *tweet*. Namun, pada penelitian ini terkendala pada tahapan *pre-processing* yang kurang kompleks, sehingga diharapkan pada penelitian selanjutnya dapat menambahkan proses *stopword* dan *stemming* pada proses *pre-processing text* serta menambah jumlah dataset sehingga dapat meningkatkan tingkat akurasi[13].

Penelitian yang dilakukan oleh Muh. Fitra Rizki, dkk mengkaji analisis sentimen *cyberbullying* pada media sosial dengan menggunakan metode *support vector machine*. Dataset yang digunakan tidak lebih dari 100 data *tweet* dan diperoleh dengan menggunakan *keyword* yang berpotensi menimbulkan *cyberbullying* seperti #cebong atau #kadrun. Hasil pengujian model sistem menunjukkan bahwa sistem berhasil melakukan klasifikasi dengan rata-rata waktu pemrosesan sebesar 101.100,2 milidetik dan kecepatan pemrosesan sebesar 0,000989 data per milidetik serta diperoleh tingkat akurasi sebesar 70%. Hal ini

menunjukan bahwa sistem analisis sentimen dengan menggunakan metode SVM yang dikembangkan mampu mengklasifikasikan konten *tweet* menjadi sentimen *cyberbullying* dan *non-cyberbullying* dengan cukup efektif[14].

Riset yang dilakukan oleh Dedy Atmajaya, dkk mengkaji implementasi metode SVM dan *Naive Bayes* untuk analisis sentimen ChatGPT di *Twitter*. Pada penelitian ini digunakan sebanyak 1000 data tentang ChatGPT dan diberikan label sentimen positif, negatif dan netral. Berdasakan hasil pengujian, diperoleh hasil yang menunjukkan tingkat keakuratan SVM lebih baik daripada *naive bayes*. Hal ini juga dipengaruhi oleh pemilihan metode pelabelan data[9].

Menurut penelitian yang dilakukan oleh M. R. Adrian, dkk yang bejudul "Perbandingan Metode Klasifikasi *Random Forest* dan SVM Pada Analisis Sentimen PSBB", diperoleh hasil akhir yang menunjukkan *Random Forest* tidak mampu mengenali sentimen "Positif", sedangkan SVM berhasil mengenali sentimen "negatif". Namun demikian, SVM memiliki keakuratan yang lebih kecil dibandingkan *Random Forest*. Dengan demikian, dapat disimpulkan bahwa model SVM dianggap lebih efektif karena dapat mengidentifikasi *tweet* yang diberi label "Positif" dengan lebih baik[15].

Penelitian yang dilakukan oleh Nur Fitriyah, dkk mengkaji terkait "Analisis Sentimen Gojek Pada Media Sosial *Twitter* Dengan Klasifikasi *Support Vector Machine* (SVM)". Riset ini dilakukan dengan menganalisis sebanyak 1500 *tweets* dengan metode SVM (*linier kernel* dan *RBF kernel)*. Hasil penelitian menunjukan bahwa pelabelan manual dengan pelabelan sentiment *scoring* menunjukan hasil yang sama, yaitu sebesar 79,19%. Selain itu, hasil penelitian juga menunjukkan kernel RBF memiliki performa yang lebih baik daripada kernel linier ditinjau dari tingkat kecocokan sentimen pada gojek yang benar[16].

Kajian yang dilakukan oleh Putri Rahmadhan, dkk yang berjudul "Pengaruh Media Sosial *Twitter* @Greenpeace.Id Terhadap Sikap Peduli Lingkungan". Kaijan ini bertujuan untuk mengetahui pengaruh *twitter* @GreenpeaceID dalam melakukan kampanye peduli lingkungan terhadap pengguna atau pengikutnya. Hasil penelitian dengan menggunakan metode kuantitatif menunjukan bahwa akun

@GreenpeaceID berpengaruh besar pada sikap lingkungaan dengan skor 23,8%[17].

2.2 Dasar Teori

2.2.1 Text Mining

Text mining merupakan suatu proses yang berfokus pada ekstraksi data berbentuk informasi yang ada pada dokumen. Tujuannya adalah untuk menemukan informasi atau kalimat baru, sehingga dalam proses ini dibutuhkan analisis yang menghubungkan antar dokumen[18].

Text Mining sering digunakan untuk menyelesaikan masalah klasifikasi, pengambilan informasi, ekstraksi informasi, serta pengelompokan. Secara umum, proses kerjanya mirip metode dari penelitian data mining secara umum. Namun, perbedaannya terletak pada pola yang digunakan, di mana *Text Mining* mengambil pola dari sekumpulan bahasa alami yang tidak terstruktur, sementara dalam data *mining*, pola diambil dari *database* yang terstruktur.[19].

2.2.2 Analisis Sentimen

Analisis sentimen merupakan metode untuk memperoleh pemahaman, mengambil, dan memproses data teks secara otomatis guna memperoleh keterangan mengenai teks yang terdapat dalam suatu opini [10]. Proses ini penting dilakukan untuk mengidentifikasi dan mengelompokkan sentimen terkait suatu permasalahan atau isu yang sedang dibicarakan, apakah sentimen itu bersifat positif, negatif, atau netral [6].

Proses analisis sentimen juga ditujukan untuk memperoleh keterangan tentang sikap, opini bahkan emosi pada suatu teks yang dianalisis. Analisis sentimen berfokus pada pengecekan klasifikasi berdasarkan polaritas[20]. Tahapan yang penting dilakukan dalam proses analisis sentimen adalah memisahkan satu karakter menjadi token atau proses "*Encoding*". Proses tokenisasi ini merupakan proses membagi kalimat menjadi kata per kata[21].

2.2.3 Pre-processing Text

Pre-processing text adalah proses persiapan data sebelum digunakan pada tahap selanjutnya. Proses ini terdiri dari proses menghapus atau mentransformasi

data yang kurang relevan agar sistem dapat memproses data tersebut dengan mudah[22]. Tahapan ini merupakan tahapan krusial, terlebih khusus dikarenakan data di media sosial memiliki banyak item teks unik sehingga perlu dihilangkan data teks umum dan informal untuk mengurangi ruang fitur [10].

Secara umum, proses *pre-processing* melibatkan beberapa langkah, seperti membersihkan data dengan mengganti teks menjadi huruf kecil, menghilangkan angka dan tanda baca, menghilangkan kata-kata yang sering muncul, memperbaiki ejaan kata, serta menghapus imbuhan untuk mengubah kata menjadi bentuk dasar.[23].

2.2.4 Pembobotan TF-IDF

Menurut [24] Salah satu cara untuk melakukan pembobotan kata dalam proses ekstraksi kata adalah TF-IDF (*Term Frequency-Inverse Document Frequency*), yang menggunakan perhitungan kata umum dalam data retrival. Tahapan ini digunakan untuk mengukur nilai bobot suatu kata dengan menimbang frekuensi kemunculan kata tersebut. Hasil pembobotan ini dapat digunakan untuk klasifikasi, *clustering* dan pengambilan informasi dokumen[12]. Persamaan untuk menghitung nilai TF-IDF dapat dilihat pada persamaan (1)

$$W = tf \times idf \quad (1)$$

Dengan W adalah bobot TF-IDF, tf adalah hasil nilai *term frequency* dan idf adalah hasil nilai *Inverse Document Frequency*.

2.2.4.1 Term Frequency (TF)

Term Frequency (TF) menunjukkan seberapa sering kata muncul dalam dokumen. Nilai TF yang lebih tinggi, atau jumlah kata yang muncul, akan menerima bobot yang lebih besar. Persamaan menunjukkan rumus untuk menghitung frekuensi *term*(2).

$$TF = 1 + \log\left(F_{t,d}\right), F_{t,d} > 0 \quad (2)$$

Dimana TF adalah *Term Frequency*, $F_{t,d}$ adalah *Frequency Term* pada dokumen.

2.2.4.2 Inverse Document Frequency (IDF)

Inverse Document Frequency digunakan dengan tujuan menentukan jika kata yang dicari sama dengan kata kunci yang diinginkan [10]. Nilai IDF akan berkurang selama proses perhitungan ketika term yang dicari lebih cocok dengan kata kunci yang dicari. Ini disebabkan oleh proses berfokus pada kata yang unik dan relevan dengan dokumen yang dianalisis dengan mengurangi pengaruh kata yang umum. Persamaan perhitungan *Term Frequency* dapat dilihat pada persamaan (3).

$$IDF = \log\left(\frac{D}{df}\right) \quad (3)$$

Keterangan :

IDF = *Inverse Document Frequency*.

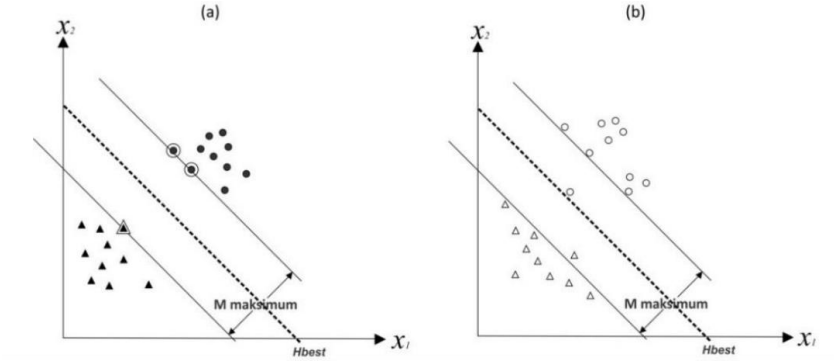
D = Jumlah keseluruhan dokumen.

Df = Jumlah dokumen yang mengandung *term*.

2.2.5 Support Vector Machine

Metode *Support Vector Machine* (SVM) didesain untuk mengidentifikasi *hyperplane* dengan margin terbesar untuk membagi kelas - kelas set data secara optimal. SVM merupakan metode *machine learning* yang memanfaatkan fungsi-fungsi linear dalam ruang fitur berdimensi tinggi[25]. Metode SVM juga dikonsepkan untuk melakukan klasifikasi data, dengan mengidentifikasi *hyperplane* terbaik lalu membentuk jarak untuk memisahkan kelas-kelas data yang telah ditentukan [26].

Support Vector dalam SVM adalah objek – objek data terdekat dengan *hyperplane* atau objek yang berada pada sisi terluar. *Support vector* akan dihitung oleh SVM untuk menentukan *hyperplane* paling optimal, seperti dilustrasikan pada gambar 2.1 berikut ini[27].



Gambar 2.1 *Hyperplane* Terbaik dan Margin Maksimum

Hyperplane terbaik didapatkan dengan memilihi margin maksimum antara dua kelas. Dengan demikian, *hyperplane* (*linear separable*) dapat memisahkan kedua kelas dengan sempurna. Namun, biasanya 2 kelas dalam ruang input tidak dapat dipisahkan dengan sempurna (*non-linear separable*). Digunakan metode *margin* agar dapat menyelesaikan permasalahan ini[14]. Untuk mengetahui *hyperplane* pada SVM dapat menggunakan persamaan (4).

$$(W.X_i) + b = 0 \quad (4)$$

Di dalam data x_i, yang termasuk dalam kelas -1 dapat dijelaskan dengan rumusan seperti yang ada pada persamaan. (5)

$$(W.X_i + b) \leq 1, y_i = -1 \quad (5)$$

Di dalam data x_i, yang termasuk pada kelas +1 dapat dijelaskan dengan rumusan seperti yang ada pada persamaan (6)

$$(W.X_i + b) \geq 1, y_i = 1 \quad (6)$$

Adapun kernel yang umum digunakan dalam pada *Support Vector Machine* (SVM) adalah seperti dibawah ini:

1. Kernel *Linier*

   $$K(x, x_i) = x.x_i \quad (7)$$

2. Kernel *Polinomial*

   $$K(x, x_i) = (1 + x.x_i)^d \quad (8)$$

3. Kernel *Radial Basis Function* (RBF)

   $$K(x, x_i) = \exp(-\gamma \|x - x_i\|^2) \quad (9)$$

4. Kernel *Sigmoid*

$$K(x, x_i) = tanh(\alpha(x.x_i) + r) \quad (10)$$

2.2.6 Evaluasi Klasifikasi

Evaluasi dilakukan untuk mengukur kualitas hasil klasifikasi dengan melakukan penilaian terhadap performanya untuk mengetahui keakuratan sistem melakukan klasifikasi[28]. Contoh evaluasinya adalah *accuracy*, *precision*, *recall*[29]. Untuk mendapatkan nilai-nilai ini, Confussion Matrix digunakan, seperti Tabel 2.1.

Tabel 2.1 Confussion Matrix

Kelas Prediksi	Kelas Sebenarnya	
	Positif	Negatif
Positif	Benar Positif (TP)	Salah Positif (FP)
Negatif	Salah Positif (FP)	Benar Negatif (TN)

Parameter-parameter pada Tabel 2.1 merupakan parameter untuk mencari nilai-nilai dari evaluasi-evaluasi berikut ini[30] :

1. *Recall* digunakan untuk mengetahui seberapa besar sistem dapat tingkat keberhasilan sistem dalam mengidentifikasi ulang suatu data. Dalam klasifikasi, recall menunjukkan proporsi data yang benar-benar positif yang berhasil diidentifikasi lalu dikomparasikan dengan seluruh jumlah data positif. Untuk mengetahui nilai tersebut dapat dihitung melalui rumusan berikut ini :

   $$Recall = \frac{TP}{TP + FP} \times 100\% \quad (11)$$

2. *Precision* digunakan untuk mencari tahu kesesuaian data yang diperoleh dengan data yang diperlukan. *Precision* berfokus pada bagian data yang telah diprediksi sebagai positif, menilai seberapa banyak dari prediksi tersebut yang benar-benar merupakan nilai positif. *Precision* menjadi tolak ukur kemampuan model menghasilkan prediksi positif yang benar. *Precision* sering digunakan bersamaan dengan *recall* untuk mendapatkan gambaran

yang lebih lengkap tentang kinerja model. *Precision* dapat dihitung melalui rumusan berikut ini :

$$Precision = \frac{TP}{TP + FN} \times 100\% \quad (12)$$

3. *Accuracy* merupakan nilai yang menjelaskan seberapa dekat nilai yang diperoleh dengan nilai asli. *Accuracy* memberikan gambaran umum tentang kinerja model secara keseluruhan, yaitu dengan mengindikasikan seberapa dekat hasil prediksi model dengan nilai sebenarnya dalam dataset. Meskipun *accuracy* adalah metrik yang banyak digunakan dan menjadi acuan utama untuk kinerja suatu model, metrik *accuracy* tetap perlu dipertimbangkan bersama dengan yang lainnya misalnya *precision* dan *recall* sehingga dapat memahami lebih dalam kinerja model yang dikembangkan. *Accuracy* dapat dihitung melalui rumusan berikut ini :

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \times 100\% \quad (13)$$

Performa untuk mengetahui hasil prediksi adalah benar atau salah dapat dilihat melalui *Confussion Matrix*. Jika model klasifikasi memiliki tiga kelas, maka akan dihasilkan *matrix* baru berordo 3x3 seperti yang ditunjukkan dalam Tabel 2.2. yang digunakan untuk mencari ketepatan, ketepatan, dan *recall* [31].

Tabel 2.2 Confussion Matrix Ordo 3X3

	Prediksi			
Aktual		Positif (P)	Negatif (N)	Netral (T)
	Positif (P)	PP	PN	PT
	Negatif (N)	NP	NN	NT
	Netral (T)	TP	TN	TT

2.2.7 Python

Python merupakan bahasa pemrograman dengan level abstraksi tinggi yang bersifat fleksibel, yang berarti kode sumber dapat langsung diterjemahkan menjadi kode mesin saat program dijalankan. *Python* juga mendukung bahasa berorientasi objek dalam pengembangan aplikasi. *Python* merupakan bahasa

yang mudah dipelajari. Selain itu, *python* juga menyediakan banyak struktur data tingkat tinggi[32].

Python dirancang untuk bekerja dalam bidang tertentu, seperti pemrograman web, tetapi juga dapat digunakan untuk CAD 3D, web, bisnis, dll. Beberapa fitur penting *Python* adalah mudah dipelajari dan digunakan, bahasa ekspresif, dan *interpreter* [33].

2.2.8 Flask

Flask merupalan *web framework* yang dikembangkan dengan bahasa *python* yang termasuk *microframework. Flask* adalah framework *web* yang mempermudah pengembangan dan pengaturan tampilan *web*. Penggunaan flask dengan *python* dapat memudahkan developer membangun aplikasi webiste yang tertata dengan baik serta mengelolah perilaku dan fungsionalitas *website* secara efisien[34].

Flask menerapkan *ToolKit* WSGI dan *Jinja Template Engine*. Kategori *Flask* dibagi menjadi 2: File Statis yang mengandung kode status yang diperlukan untuk *website*, seperti kode CSS, kode *JavaScript*, dan file gambar; dan *File Template* yang mengandung template *Jinja*, termasuk halaman HTML [35].

2.2.9 Media Sosial X

Media sosial adalah sekumpulan media berbasis internet yang membuat user membuat, menerima, dan membagikan berbagai informasi dalam waktu yang cepat dan ruang yang tidak terbatas. Media sosial biasanya didefinisikan sebagai alat, layanan, dan komunikasi yang tersedia secara online yang memungkinkan seseorang menjalin hubungan dengan orang-orang yang memiliki kepentingan tertentu dengan menggunakannya[36]. Media sosial *twitter*, instagram dan tiktok adalah paling banyak digunakan.

Seiring berjalannya waktu pertumbuhan *Twitter* yang saat ini dikenal dengan nama X mengalami peningkatan yang pesat, dan banyak diminati oleh pengguna di seluruh dunia karena dianggap mudah digunakan dan sederhana.

Menurut data yang dikumpulkan oleh *We Are Social*, *Twitter* merupakan salah satu dari lima *platform* media daring yang terpopuler di Indonesia [37].

2.2.10 Kampanye Pengurangan Plastik

Kampanye adalah upaya terencana dan berkelanjutan dalam komunikasi untuk mencapai dampak tertentu pada khalayak luas dalam periode waktu yang ditentukan[38]. Salah satu kampanye yang cukup banyak dilakukan adalah kampanye pengurangan plastik. Upaya-upaya yang digunakan sebagai upaya penanggulangan masalah plastik diantaranya adalah upaya kampanye pengurangan limbah botol plastik, bank sampah 102 yang dilakukan oleh pemerintah Kota Tangerang dengan cara dengan menyebar kotak sampah di sejumlah warung makan yang ada di sekitar Kecamatan Cibodas, Kota Tangerang[39].

Selain itu, upaya lainnya yang dilakukan untuk memperluas kampanye pengurangan sampah plastik adalah dengan menggunakan pemanfaatan teknologi digital, terutama sosial media. enurut data Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), diperkirakan jumlah pengguna internet di Indonesia pada tahun 2022-2023 mencapai 215,63 juta[40]. Angka tersebut menunjukkan bahwa media sosial dapat sangat membantu meningkatkan kesadaran masyarakat tentang pentingnya mengelola sampah dan mengurangi penggunaan barang sekali pakai.