BAB II
LANDASAN TEORI

2.1 Peneliti Terdahulu

Penelitian yang telah dilakukan sebelumnya sangat penting sebagai landasan dalam melaksanakan penelitian yang akan dilakukan. Tabel 2.1 merupakan beberapa penelitian terdahulu yang relevan.

Tabel 2. 1 Penelitian Terdahulu

1.	Nama Penulis/Tahun	Dinar Ajeng Kristiyanti, Normah, Akhmad Hairul Umam (2019).
	Nama Jurnal	*2019 5th International Conference on New Media Studies (CONMEDIA)*
	Judul Jurnal	*Prediction of Indonesia Presidential Election Results for the 2019-2024 Period Using X Sentiment Analysis*
	Permasalahan	Permasalahan utama dalam penelitian ini adalah memprediksi hasil pemilihan Presiden dan Wakil Presiden Republik Indonesia periode 2019-2024 melalui opini publik di X.
	Metode	*Support Vector Machine (SVM*) dengan teknik seleksi fitur *Particle Swarm Optimization (PSO)* dan *Genetic Algorithms (GA).*
	Hasil dan Kesimpulan	Pasangan Prabowo Subianto-Sandiaga Uno diprediksi akan terpilih sebagai Presiden dan Wakil Presiden Republik Indonesia untuk periode 2019-2024 dengan sentimen positif paling banyak, mencapai 830 dari 1000 *tweet*s yang dianalisis. Metode *Support Vector Machine (SVM)* dengan kombinasi *Particle Swarm Optimization (PSO)* mencapai akurasi prediksi sebesar **86.20%** dan nilai *AUC (Area Under the Curve)* sebesar **0.934**.
2.	Nama Penulis/Tahun	Lisyana Damayanti, Kemas Muslim Lhaksmana (2024).
	Nama Jurnal	Sinkron: Jurnal dan Penelitian Teknik Informatika
	Judul Jurnal	*Sentiment Analysis of the 2024 Indonesia Presidential Election on Twitter*
	Permasalahan	Permasalahan yang diangkat adalah bagaimana memahami dan mengelola data sentimen yang luas dengan efisien waktu dan akurasi tinggi untuk memperoleh pandangan menyeluruh tentang dukungan dan preferensi publik terhadap calon presiden.
	Metode	*Support Vector Machine (SVM)*
	Hasil dan Kesimpulan	Hasil dari penelitian ini menunjukkan skor *precision* 88,94%, *recall* 93,08%, *F1-score* 90,43%, dan akurasi **90,75%.**
3.	Nama Penulis/Tahun	Khodijah Hulliyah, Normi Sham Awang Abu Bakar, Amelia Ritahani Ismail, M. Octaviano Pratama (2020)
	Nama Jurnal	*2019 7th International Conference on Cyber and IT Service*

		Management (CITSM)
	Judul Jurnal	*A Benchmark of Modeling for Sentiment Analysis of The Indonesian Presidential Election in 2019*
	Permasalahan	Penelitian ini berfokus pada klasifikasi emosi dalam teks yang berhubungan dengan opini publik tentang Pemilihan Presiden Indonesia tahun 2019. Permasalahan utama adalah menemukan model algoritma yang paling akurat untuk mengklasifikasikan emosi dalam teks menjadi empat kategori: senang, sedih, marah, dan takut.
	Metode	*Algoritma Long Short-Term Memory (LSTM)* serta pembandingan *(benchmarking)* dengan *algoritma Random Forest dan Naive Bayes*
	Hasil dan Kesimpulan	Hasil penelitian menunjukkan bahwa model *Random Forest* mencapai akurasi sebesar 68,25%, sedangkan model *Multinomial Naïve Bayes* mencapai akurasi sebesar 66%.
4.	Nama Penulis/Tahun	Asno Azzawagama Firdaus, Anton Yudhana, Imam Riadi (2023).
	Nama Jurnal	DECODE: Jurnal Pendidikan Teknologi Informasi
	Judul Jurnal	Analisis Sentimen Pada Proyeksi Pemilihan Presiden 2024 Menggunakan Metode *Support Vector Machine*
	Permasalahan	Penelitian ini bertujuan untuk mengetahui keberpihakan masyarakat terhadap calon presiden dan wakil presiden melalui diskusi di Twitter.
	Metode	*Support Vector Machine (SVM)*
	Hasil dan Kesimpulan	hasil sentimen berdasarkan tiga dataset kandidat yang dipilih, yaitu anies baswedan 65,62%, ganjar pranowo 73,58%, dan prabowo subianto 66,34%. Hasil akurasi metode yang dimiliki oleh ketiga dataset yaitu anies baswedan 73%, ganjar pranowo 79% dan prabowo subianto 79%.
5.	Nama Penulis/Tahun	Muhammad Rizky Pribadi, Danny Manongga, Hindriyanto Dwi Purnomo, Hendry, Iwan Setyawan (2022).
	Nama Jurnal	*International Seminar on Intelligent Technology and Its Applications.*
	Judul Jurnal	*Sentiment Analysis of the PeduliLindungi on Google Play using the Random Forest Algorithm with SMOTE*
	Permasalahan	Mengetahui sentimen publik terhadap aplikasi PeduliLindungi yang ada di Google Play. Bagaimana sentimen tersebut cenderung negatif dan mengevaluasi efektivitas penggunaan kombinasi *Random Forest* dan *SMOTE* dalam klasifikasi sentimen tersebut.
	Metode	*Random Forest* dan *SMOTE*
	Hasil dan Kesimpulan	Hasil studi ini menunjukkan bahwa sentimen publik terhadap aplikasi PeduliLindungi cenderung negatif. Algoritma *Random Forest* dan *SMOTE* yang digunakan untuk analisis sentimen mencapai akurasi sebesar **71%,** dengan nilai *recall* dan presisi masing-masing 70%.

6.	Nama Penulis/Tahun	Dinar Ajeng Kristiyanti, Samuel Ady Sanjaya, Vinsencius Christio Tjokro, Jason Suhali (2024).
	Nama Jurnal	*IAES International Journal of Artificial Intelligence (IJ-AI)*
	Judul Jurnal	*Dealing imbalance dataset problem in sentiment analysis of recession in Indonesia.*
	Permasalahan	Data yang dikumpulkan dari X tidak seimbang antara sentimen positif dan negatif terhadap berita resesi global.
	Metode	Algoritma *Naïve Bayes, SVM, KNN* menggunakan teknik *SMOTE* dan *ROS*.
	Hasil dan Kesimpulan	Hasil evaluasi model menunjukkan bahwa algoritma *Bernoulli-naive Bayes*, dengan teknik pengambilan sampel *SMOTE* setelah dilakukan pemisahan data, diperoleh akurasi terbaik sebesar 84%, dan menggunakan teknik *ROS* diperoleh akurasi sebesar 81%. Sebaliknya dengan teknik *SMOTE* dan *ROS* sebelumnya pemisahan data pada algoritma *SVM* mendapatkan akurasi terbaik sebesar 93%. sebelumnya jika hanya menggunakan SVM hanya mencapai 84%
7.	Nama Penulis/Tahun	Panji Al Muqsith Prasetyo, Arief Hermawan (2023).
	Nama Jurnal	INFOTECH: Jurnal Informatika Teknologi
	Judul Jurnal	*Sentiment analysis twitter of presidential election using naïve bayes algorithm*
	Permasalahan	Masyarakat Indonesia menghadapi kesulitan dalam menavigasi opini dan reaksi di media sosial, khususnya Twitter (X), terkait calon presiden dan wakil presiden untuk Pemilihan Presiden 2024. Masyarakat yang kurang literasi kebahasaan dan literasi digital mudah terpancing oleh opini dari netizen yang lebih literate, sehingga diperlukan sistem yang dapat secara otomatis menentukan sifat komentar (netral, positif, atau negatif).
	Metode	*Naïve Bayes*
	Hasil dan Kesimpulan	Penelitian ini menghasilkan model dengan nilai akurasi sebesar 85%, precision 86.54%, recall 85%, dan f1-score 85%.
8.	Nama Penulis/Tahun	Wasim Ahmed, Josep Vidal-Alaball, Francesc Lopez Segui, Pedro A. Moreno-Sánchez (2020).
	Nama Jurnal	*International Journal of Environmental Research and Public Health*
	Judul Jurnal	*A Social Network Analysis of Tweets Related to Masks during the COVID-19 Pandemic*
	Permasalahan	Penelitian ini bertujuan untuk menganalisis konten di X yang berkaitan dengan penggunaan masker selama pandemi COVID-19.
	Metode	*Social Network Analysis*

	Hasil dan Kesimpulan	Penelitian menunjukkan bahwa jaringan X membentuk komunitas-komunitas kecil dengan berbagai pengguna, termasuk warga biasa, politisi, dan figur budaya populer, yang berdiskusi tentang masker. Hashtag yang populer mengajak publik untuk memakai masker.
9.	Nama Penulis/Tahun	Bahtiar Imran, Muh Nasirudin Karim, Nur Isna Ningsih (2024)
	Nama Jurnal	Dinamika Rekayasa
	Judul Jurnal	*Classification Of Hoax News Related to The General Election Of The President Of The Republic Of Indonesia In 2024 Using Naïve Bayes and SVM*
	Permasalahan	Pada era digital, akses berita sangat mudah diakses oleh setiap orang, yang dimanfaatkan oleh oknum tak bertanggung jawab untuk menyebarkan berita hoax mengenai isu Pilpres di Indonesia. Hal ini menyebabkan kebingungan dan potensi misinformasi di kalangan masyarakat.
	Metode	*Naive Bayes dan Support Vector Machine (SVM)*
	Hasil dan Kesimpulan	Hasil menunjukkan bahwa *Naive Bayes* mencapai akurasi sebesar 97% dengan presisi 94%, *recall* 100%, dan *F1-score* 97%. Sementara itu, *SVM* memiliki akurasi 95%, presisi 94%, *recall* 97%, serta *F1-score* 95%.
10.	Nama Penulis/Tahun	Wasim Ahmed, Josep Vidal-Alaball, Joseph Downing, Francesc López Seguí (2020).
	Nama Jurnal	*Journal Of Medical Internet Research*
	Judul Jurnal	*COVID-19 and the 5G Conspiracy Theory:Social Network Analysis of X Data*
	Permasalahan	Penyebaran teori konspirasi yang menyebabkan munculnya berita palsu terkait hubungan antara teknologi 5G dengan penyebaran COVID-19. Kebutuhan untuk memahami penyebab berita palsu dan kebijakan cepat dalam mengisolasi dan menentang informasi yang salah merupakan kunci untuk melawan penyebaran berita palsu tersebut.
	Metode	*Social Network Analysis*
	Hasil dan Kesimpulan	Hasil analisis menunjukkan dua kelompok besar dalam jaringan sosial: kelompok isolat dan kelompok penyebar. Tidak ada figur otoritas yang aktif melawan misinformasi. Dari 233 *tweet*, sebagian besar (65,2%) menolak teori konspirasi 5G-COVID-19.

Tabel 2.1 adalah tabel yang berisi referensi dari penelitian-penelitian sebelumnya yang digunakan sebagai dasar dalam penelitian ini. Pertama, penelitian terdahulu yang dilakukan oleh Dinar Ajeng Kristiyanti, Normah, dan Akhmad Hairul Umam mendapatkan hasil bahwa pasangan Prabowo Subianto-

Sandiaga Uno diprediksi akan terpilih sebagai Presiden dan Wakil Presiden Republik Indonesia untuk periode 2019-2024 dengan sentimen positif paling banyak, mencapai 830 dari 1000 *tweet*s. Akurasi *Support Vector Machine (SVM*) tanpa *Particle Swarm Optimization (PSO)* dan *Genetic Algorithms (GA)* sebesar 78.70% dan akurasi *Support Vector Machine (SVM)* dengan kombinasi *Particle Swarm Optimization (PSO)* mencapai akurasi prediksi sebesar 86.20% [10]. Kedua, penelitian terdahulu yang dilakukan oleh Lisyana Damayanti, Kemas Muslim Lhaksmana yang membahas tentang analisis sentimen pemilihan Presiden Indonesia Tahun 2024 menggunakan algoritma *Support Vector Machine (SVM)* menghasilkan nilai akurasi sebesar 90.75% [15]. Ketiga, penelitian terdahulu yang dilakukan oleh Khodijah Hulliyah, Normi Sham Awang Abu Bakar, Amelia Ritahani Ismail, M. Octaviano Pratama menggunakan metode algoritma *Long Short-Term Memory (LSTM)* serta pembandingan *(benchmarking)* dengan algoritma *Random Forest* dan *Naive Bayes* menghasilkan nilai akurasi *Random Forest* yang lebih besar dibandingkan dengan *Naïve Bayes*, yaitu *Random Forest* mencapai akurasi sebesar 68,25%, sedangkan model *Multinomial Naïve Bayes* mencapai akurasi sebesar 66% [12]. Penelitian keempat, yang membahas tentang analisis sentimen pada proyeksi pemilihan Presiden 2024 menggunakan algoritma *Support Vector Machine (SVM),* menghasilkan nilai akurasi yang dimiliki oleh ketiga dataset yaitu anies baswedan 73%, ganjar pranowo 79% dan prabowo subianto 79% [16]. Penelitian kelima, menggunakan metode *Random Forest* dan *SMOTE* menunjukkan bahwa sentimen publik terhadap aplikasi PeduliLindungi cenderung negatif. Algoritma *Random Forest* dan *SMOTE* yang digunakan untuk analisis sentimen mencapai akurasi sebesar 71% [17]. Penelitian keenam, menggunakan Algoritma *Naïve Bayes, SVM, KNN* dengan teknik *SMOTE* dan *ROS* untuk mengatasi ketidakseimbangan data, menghasilkan nilai akurasi algoritma *SVM* sebesar 93%. Jika tidak menggunakan teknik *SMOTE* dan *ROS* akurasi hanya mencapai 84% [13]. Penelitian ketujuh, yang dilakukan oleh Panji Al Muqsith Prasetyo, Arief Hermawan menggunakan algoritma *Naïve Bayes* menghasilkan nilai akurasi sebesar 85% [11]. Penelitian kesembilan, yang membahas tentang klasifikasi berita hoax terkait pemilihan Presiden Indonesia tahun 2024 menggunakan algoritma *Naive Bayes* dan *Support Vector Machine*

(SVM) menghasilkan nilai akurasi *Naive Bayes* lebih besar dibandingkan dengan *Support Vector Machine (SVM)*. Hasil menunjukkan bahwa *Naive Bayes* mencapai akurasi sebesar 97% dengan presisi 94%, *recall* 100%, dan *F1-score* 97%. Sementara itu, *SVM* memiliki akurasi 95%, presisi 94%, *recall* 97%, serta *F1-score* 95% [18]. Penelitian kedelapan dan kesepuluh, memiliki metode yang sama yaitu *Social Network Analysis*. Dari penelitian tersebut sama-sama menggunakan dataset yang berasal dari X, melakukan *content analysis*, dan dilakukan perhitungan *Betweenness centrality* [14][19].

Penelitian ini mengambil adopsi dari penelitian sebelumnya, yang serupa dengan penelitian [20][12][15], yang membahas tentang prediksi hasil Pemilihan Presiden Indonesia 2019 dengan menggunakan analisis sentimen pada *platform* X. Penelitian ini berfokus untuk memprediksi Presiden RI untuk tahun 2024-2029 menggunakan analisis sentimen dan membandingkan algoritma *Naïve Bayes*, *Random Forest*, dan *Support Vector Machine*, serta metode *Social Network Analysis* untuk mengidentifikasi akun, lokasi yang berpengaruh dan akun yang terdeteksi *buzzer*. Penelitian ini juga menggunakan teknik *SMOTE* untuk mengatasi *Imbalanced* data. Hasil penelitian akan divisualisasikan dalam bentuk *dashboard* pada sebuah *website* yang menampilkan sentimen analisis dari hasil prediksi pemilihan Presiden Republik Indonesia 2024.

2.2 Tinjauan Teori

2.2.1 Pemilu

Menurut Undang-Undang No 8 Tahun 2020 Pemilu atau pemilihan umum adalah sarana pelaksanaan kedaulatan rakyat yang dilaksanakan secara langsung, umum, bebas, rahasia, jujur, dan adil dalam Negara Kesatuan Indonesia berdasarkan Pancasila dan Undang-Undang Dasar Negara Republik Indonesia 1945 [21]. Pemilihan umum diselenggarakan untuk masyarakat Indonesia memilih seseorang untuk mengisi jabatan politik. Salah satu bentuk pemilu adalah Pilpres, yang khususnya berkaitan dengan pemilihan kepala negara atau presiden. Diselenggarakannya pemilu, diharapkan menghasilkan pemimpin dan wakil rakyat yang dapat membuat bangsa Indonesia maju,

makmur, sejahtera, dan menjadi bangsa yang besar [22]. Pengawasan dalam pemilu dibantu oleh Badan Pengawas Pemilu (BAWASLU) sebagai Lembaga yang memiliki tugas untuk pengawasan dan penegakan hukum pemilu, dalam proses serta mekanisme penegakan hukum pemilu itu sendiri [23].

2.2.2 Pemilihan Presiden RI Tahun 2024

Indonesia menggelar pemilu setiap 5 tahun sekali, dan pemilu terakhir diadakan pada tahun 2019. Pemilu akan diadakan lagi pada tahun 2024 pada tanggal 14 Februari 2024 [24]. KPU mengumumkan masa pendaftaran Capres dan Cawapres pada tanggal 16 - 18 Oktober 2023 [2]. Pada 19 Oktober 2023, Pukul 09.36 WIB, H. Anies Rasyid Baswedan, Ph.D. dan Dr. (H.C.) H. A. Muhaimin Iskandar, mendaftar sebagai kandidat Capres dan Cawapres yang diusulkan oleh Gabungan Partai Politik Nasdem, PKB (Partai Kebangkitan Bangsa), dan PKS (Partai Keadilan Sejahtera). H. Ganjar Pranowo, S.H., M.I.P. dan Prof. Dr. H. M. Mahfud MD mendaftar sebagai kandidat Capres dan Cawapres pada hari Kamis, tanggal 19 Oktober 2023, Pukul 12.20 WIB yang diusulkan oleh Gabungan Partai Politik PDIP, PERINDO, Partai Persatuan Pembangunan, dan Partai Hati Nurani Rakyat. Pada tanggal 25 Oktober 2023, Pukul 11.20 WIB, H. Prabowo Subianto dan Gibran Rakabuming Raka sebagai kandidat Capres dan Cawapres yang diusulkan oleh Gabungan Partai Politik Partai Gerakan Indonesia Raya, Parta Golongan Karya, Partai Demokrat, Partai Amanat Nasional, Partai Solidaritas Indonesia, Partai Bulan Bintang, dan Partai Garda Republik Indonesia.

Gambar 2.1 Calon Presiden dan Wakil Presiden RI Tahun 2024 -2029 [25]

Gambar 2.1 merupakan Calon Presiden dan Wakil Presiden RI Tahun 2024 -2029. Komisi Pemilihan Umum (KPU) telah menetapkan Calon Presiden dan Wakil Presiden Indonesia Tahun 2024-2029 pada tanggal 13 November 2023. Dari hasil pendaftaran pada tanggal 16 - 18 Oktober 2023, terdapat 3 pasangan Calon Presiden dan Wakil Presiden yang telah memenuhi syarat yaitu, H. Anies Rasyid Baswedan, Ph.D. dan Dr. (H.C.) H.A. Muhaimin Iskandar, H. Ganjar Pranowo, S.H., M.I.P. dan Prof. Dr. H. M. Mahfud MD, H. Prabowo Subianto dan Gibran Rakabuming Raka [4] . Setelah keputusan resmi, Komisi Pemilihan Umum (KPU) akan mengadakan proses undian dan penentuan nomor urut untuk kandidat presiden dan wakil presiden dalam Pemilu 2024. Periode kampanye dijadwalkan untuk dimulai pada tanggal 28 November 2023 dan akan berakhir pada tanggal 10 Februari 2024 [4].

2.2.3 Analisis Sentimen

Analisis sentimen, yang juga dikenal sebagai *opinion mining*, merupakan bagian dari klasifikasi teks yang berfokus pada area luas pengolahan bahasa alami, linguistik komputasional, dan penambangan teks dengan tujuan untuk memeriksa pendapat, sentimen, evaluasi, dan emosi dari seorang pembicara atau penulis terhadap topik, produk, layanan, organisasi, atau kegiatan tertentu lainnya [26]. Analisis sentimen merupakan proses otomatis dalam menginterpretasi dan memproses data teks untuk mendapatkan informasi. Tujuan analisis sentimen adalah untuk mengidentifikasi opini terhadap suatu subjek atau objek (seperti individu, organisasi, atau produk) dalam sebuah set data apakah bersifat positif atau negatif [8].

Berkembangnya penelitian dan aplikasi analisis sentimen menunjukkan dampak dan manfaat yang besar. Analisis sentimen juga digunakan untuk mengetahui siapa yang memberikan opini yang banyak direspons oleh pengguna X, dan dapat digunakan untuk mengungkap pendapat publik terhadap suatu isu [8]. Analisis sentimen bisa digunakan untuk memahami opini di berbagai bidang seperti ekonomi, politik, sosial, dan hukum. X sebagai media sosial memberikan kesempatan kepada peneliti untuk mengkaji

emosi, suasana hati, dan pandangan masyarakat dengan menggunakan analisis sentimen [27].

2.2.4 *Scraping*

Web scraping merupakan proses pengumpulan data otomatis dari situs *web* dengan struktur tertentu, yang dilakukan menggunakan aplikasi khusus atau kode pemrograman. Informasi yang berhasil diambil bisa mencakup berbagai jumlah data, mulai dari ribuan hingga miliaran entri dalam *social media* [28]. Tujuan utama dari *web scraping* adalah untuk mengambil dan mengekstraksi data dari berbagai sumber. *Scraping* memberikan sejumlah manfaat, salah satunya adalah memungkinkan informasi yang diperoleh menjadi lebih terfokus, sehingga mempermudah proses pencarian informasi yang dibutuhkan [29].

2.2.5 *Text Preprocessing*

Text Preprocessing adalah langkah yang harus dilakukan dalam *Natural Language Process (NLP). Text Preprocessing* adalah proses pembersihan teks ke dalam bentuk yang dapat diprediksi dan dianalisis untuk tugas tertentu. Tujuan utama dari *preprocessing* teks adalah untuk memecah teks menjadi bentuk yang dapat dicerna oleh algoritma pembelajaran mesin [30]. Untuk menyiapkan data teks untuk pembuatan model, harus dilakukan *preprocessing* teks. Beberapa langkah preprocessing adalah [30][31]:

1. *Removing punctuations*

   Pada tahapan ini akan dilakukan penghapusan karakter seperti, ! $( ) * % @ yang ada didalam dataset yang akan digunakan.
2. *Removing URLs*

   Removing URL adalah tahapan yang akan dilakukan penghapusan URL pada dataset yang akan digunakan.
3. *Removing Stop words*

   Stop-words adalah kata-kata yang umum digunakan dalam suatu bahasa. *Stop-word* dihilangkan dari teks sehingga kita dapat berkonsentrasi pada kata-kata yang lebih penting dan mencegah stop-word dianalisis. Jika

kami menelusuri 'apa itu pemrosesan awal teks', kami ingin lebih fokus pada 'pemrosesan awal teks' daripada 'apa itu'.

4. *Lower casing*

   Lower casing adalah tahapan pengubahan data yang memiliki huruf besar atau kapital diubah menjadi huruf kecil. Langkah ini tidak perlu dilakukan setiap kali Anda mengerjakan masalah *NLP* karena pada beberapa masalah, huruf besar yang lebih rendah dapat menyebabkan hilangnya informasi.

5. *Tokenization*

   Tokenisasi merupakan proses untuk membagi teks menjadi unit yang lebih kecil, yaitu, token, mungkin pada saat yang sama membuang karakter tertentu, seperti tanda baca. Token dapat berupa kata, angka, simbol, *n-gram,* atau karakter. *N-gram* adalah kombinasi dari n kata atau karakter bersama-sama. Tokenisasi melakukan tugas ini dengan menemukan batas kata.

6. *Stemming*

   Stemming adalah proses dasar berbasis aturan untuk menghilangkan bentuk infleksional dari sebuah token. Token diubah menjadi bentuk akarnya. Misalnya kata ‘bermasalah’ diubah menjadi ‘masalah’ setelah dilakukan *stemming*.

7. *Lemmatization*

   Lemmatization mirip dengan stemming, perbedaannya adalah *lemmatization* mengacu pada melakukan sesuatu dengan benar menggunakan kosa kata dan analisis morfologi kata, yang bertujuan untuk menghilangkan infleksi dari kata tersebut dan mengembalikan bentuk dasar atau kamus dari kata tersebut, yang juga dikenal sebagai kata tersebut. kata pengantar singkat.

2.2.6 *TF-IDF*

TF-IDF (Term Frequency-Inverse Document Frequency) adalah metode ekstraksi fitur yang umum digunakan dalam klasifikasi teks berbasis ruang vektor. *TF-IDF* memberikan bobot pada setiap kata yang muncul dan menghitung nilai invers di dalam kalimat. Kata tersebut mewakili setiap

fitur dalam dokumen. Representasi *TF-IDF* disajikan dalam bentuk array, di mana setiap baris dari array berisi data dan kolom-kolom dari array berisi kata-kata atau fitur-fitur. Bobot yang diperoleh digunakan sebagai input untuk proses klasifikasi.

Dalam metode pembobotan *TF-IDF*, terdapat beberapa variasi yang telah dikembangkan untuk meningkatkan hasil klasifikasi, antara lain *TF-IDF-CF (Term Frequency-Inverse Document Frequency-Class Frequency), TF-IGM (Term Frequency-Inverse Gravity Moment), dan TF-RF (Frequency-Relevance Frequency).* Formulasi *TF-IDF* adalah sebagai berikut [32]:

$$a_{ij} = tf_{ij} * \log\left(\frac{N}{nij}\right) \quad (1)$$

Rumus 2. 1 Formula *TF-IDF*

2.2.7 *SMOTE*

SMOTE adalah metode untuk mengatasi ketidakseimbangan dalam distribusi sampel data pada kelas minoritas dengan memperluas sampel-sampel tersebut sehingga jumlahnya menjadi seimbang dengan jumlah sampel pada kelas mayoritas. Penerapan *SMOTE* bisa menyebabkan overfitting karena duplikasi data di kelas minoritas, yang berpotensi menghasilkan redundansi dalam data latih. Langkah-langkah dalam proses *SMOTE* dimulai dengan mengukur jarak antara data pada kelas minoritas, kemudian menentukan persentase *SMOTE* yang diinginkan, menentukan jumlah tetangga terdekat *k*, dan terakhir adalah menciptakan data sintetis [33].

2.2.8 *Confusion Matrix*

Confusion matrix adalah sebuah matriks dua dimensi di mana barisnya menunjukkan label sebenarnya dan kolomnya menunjukkan label yang diprediksi oleh sebuah pengklasifikasi. Tabel 2.2 merupakan label dari *confusion matrix. True Positive (TP)* merupakan kasus positif yang diklasifikasikan dengan benar sebagai positif, *False Positive (FP)* sebagai

kasus negatif yang diklasifikasikan salah sebagai positif, *True Negative (TN*) sebagai kasus negatif yang diklasifikasikan dengan benar sebagai negatif, *False Negative (FN)* sebagai kasus positif yang diklasifikasikan salah sebagai negatif [34], [35].

Tabel 2. 2 *Confusion Matrix*

	Prediction Positive	Prediction Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

2.2.9 *Akurasi*

Akurasi didefinisikan sebagai seberapa dekat nilai prediksi dengan nilai aktual. Ini merupakan proporsi dari prediksi yang benar dibandingkan dengan jumlah total data. Rumus akurasi adalah sebagai berikut [34], [35]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

Rumus 2. 2 Rumus Akurasi

2.2.10 *Presisi*

Presisi adalah proporsi dari prediksi yang benar dibandingkan dengan total prediksi yang benar. Ini juga berfungsi sebagai indikator untuk mengukur efektivitas sistem temu balik informasi. Dalam konteks penelusuran, presisi digunakan untuk mengukur jumlah dokumen relevan yang berhasil diambil. Rumus presisi adalah sebagai berikut [34], [35]:

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

Rumus 2. 3 Rumus Presisi

2.2.11 *Recall*

Recall adalah proporsi prediksi yang benar dibandingkan dengan total data yang benar. Konsep ini merujuk pada jumlah dokumen relevan yang berhasil dipanggil oleh sistem informasi sesuai dengan query yang dimasukkan oleh pengguna. Rumus *recall* adalah sebagai berikut [35][34]:

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

Rumus 2. 4 Rumus *Recall*

2.2.12 *F-Measure*

F-measure sering juga disebut sebagai *F1-score*. F-measure adalah metode evaluasi dalam sistem informasi yang menggabungkan *recall* dan presisi. Dalam suatu konteks, *recall* dan presisi dapat memiliki bobot yang berbeda. F-measure adalah ukuran yang mencerminkan keseimbangan antara *recall* dan presisi, dihitung sebagai rata-rata harmonik dari keduanya. Rumus F-measure adalah sebagai berikut [35]:

$$F1 = 2 \cdot \frac{Precision * Recall}{Precision + Recall} \quad (5)$$

Rumus 2. 5 Rumus F-Measure

2.3 Algoritma dan *Framework*

2.3.1 *CRISP-DM*

Cross-Industry Standard Process for Data Mining (*CRISP-DM*) merupakan sebuah standarisasi yang diterapkan untuk melakukan sebuah analisis terhadap data-data organisasi atau perusahaan [36]. *CRISP-DM* memiliki enam tahapan dalam proses *data mining* yaitu, *Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment*[37].

1. *Business Understanding*

   Pada tahap ini dilakukan sebuah tujuan dari projek yang akan dilakukan dan melihat kebutuhan yang diperlukan dalam lingkup organisasi atau perusahaan, serta menyiapkan strategi dan solusi untuk mencapai tujuan yang diinginkan oleh sebuah organisasi atau perusahaan.

2. *Data Understanding*

   Pada tahap ini dilakukan sebuah proses pemahaman data untuk melakukan sebuah analisis terhadap data yang akan digunakan. Pemahaman umum yang biasa dilakukan adalah memahami data tersebut, dan atribut dari data.

3. *Data Preparation*

   Pada tahap ini dilakukan proses untuk mempersiapkan data agar mempermudah dalam proses modeling. Pada data preparation biasanya dilakukan *data cleansing* dan split data. Pada data *cleansing* dilakukan pengecekan terhadap data yang memiliki *missing value,* dan menghapus data-data yang tidak dibutuhkan. Pada split data, dilakukan pembagian antara *training* dan *testing*.

4. *Modeling*

   Pada tahap ini dilakukan pemodelan data dengan model prediktif atau deskriptif. Pada pemodelan data dibantu dengan *tools* untuk menerapkan algoritma *data mining.*

5. *Evaluation*

   Pada tahap ini dilakukan evaluasi terhadap hasil dan performa yang dihasilkan dari model yang telah dibuat pada tahap sebelumnya.

6. *Deployment*

   Pada tahap ini dilakukan perencanaan penggunaan model yang telah dihasilkan serta terdapat saran untuk diimplementasikan.

2.3.2 *Machine Learning*

Pengertian *machine learning* pertama kali diartikan oleh Arthur Samuel pada tahun 1959. Menurutnya, *machine learning* merupakan sebuah ilmu *computer* untuk mengetahui permasalahan. *Machine learning* menggunakan metode statistic, untuk membuat klasifikasi atau prediksi, kemudian dapat menghasilkan sebuah keputusan untuk membantu organisasi atau perusahaan [38]. *Machine learning* memiliki 3 jenis [39] yaitu:

1. *Supervised Learning*

   Supervised Learning merupakan teknik machine learning yang telah diberi label. Dari label yang telah terbuat maka akan dibuat prediksi. Pada *Supervised Learning* memiliki kelebihan yaitu proses yang mudah dipahami, tetapi memerlukan waktu yang cukup lama.

2. *Unsupervised Learning*

Unsupervised Learning merupakan teknik machine learning yang datanya tidak memiliki label. Pada *Unsupervised Learning* mendeteksi model deskriptif yang tidak membutuhkan label atau kategori. Algoritma ini cocok digunakan untuk *association rule* dan *clustering*.

3. *Reinforcement learning*

   Reinforcement learning melakukan pembelajaran dengan menentukan tindakan. Metode ini cocok untuk pembelajaran dependen yang memberi label pada semua keputusan.

2.3.3 *Social Network Analysis*

Social Network Analysis (SNA) adalah metode analisis yang digunakan untuk mempelajari struktur sosial melalui penggunaan jaringan dan teori grafik. *SNA* mengidentifikasi hubungan antara individu, organisasi, atau entitas lain dan memeriksa pola dan implikasi dari hubungan ini [9]. Dalam *SNA* terdapat dua bagian, yaitu *nodes* dan *edges*. *Nodes* dan *edges* merupakan kunci dalam melakukan *SNA*. *Node* dapat mewakili berbagai aktor. Misalnya, dalam jaringan internet *node* dapat mewakili halaman *web* sementara di jaringan sosial *node* dapat mewakili orang. *Edges* dapat mewakili berbagai hubungan. Dalam jaringan internet, tepi dapat mewakili *hyperlink* dan di jejaring sosial tepi dapat mewakili koneksi [40]. Beberapa karakteristik utama dalam *Social Network Analysis (SNA)* meliputi: *node*s, *edges*, *average degree*, diameter, dan *average path length* [41].

Selain itu, untuk menilai *node*s dan *edges*, model khusus diperlukan guna mengestimasi level keputusan dan struktur lingkaran dalam jaringan sosial, dikenal sebagai *centrality*. Metode pengukuran *centrality* ini digunakan untuk mengidentifikasi individu yang memegang peranan utama dalam jaringan sosial, yang mencerminkan posisi sentral seseorang dalam jaringan tersebut [41]. Beberapa metode pengukuran *centrality* [42]:

1. *Degree centrality*

   Degree centrality adalah metode pengukuran *centrality* yang paling sederhana untuk dihitung. Jumlah *edges* yang terkoneksi langsung ke sebuah *node* menentukan besaran *degree* dari *node* tersebut, yang

kemudian dikenal sebagai *degree centrality*. Dalam jaringan yang bersifat *directional*, *degree centrality* dapat dihitung dengan dua pendekatan: *in-degree* dan *out-degree*, yang masing-masing memiliki interpretasi yang berbeda, seperti dalam kasus aliran donasi dari pemerintah ke masyarakat atau sebaliknya.

2. *Betweenness Centrality*

   Betweenness Centrality diartikan sebagai indikator yang mengukur pentingnya seseorang berdasarkan kapasitasnya untuk menjembatani banyak individu lain dalam jaringan sosialnya. Secara esensial, *node*s dengan nilai *betweenness Centrality* yang tinggi berperan signifikan dalam menghubungkan jaringan tersebut. Berikut adalah rumus untuk menghitung *Betweenness Centrality.*

3. *Closeness Centrality*

   Closeness Centrality diartikan sebagai indikator yang mengukur seberapa pentingnya seseorang dalam sebuah jaringan sosial, berdasarkan jarak rata-rata mereka ke semua individu lain dalam jaringan tersebut. Konsep ini erat kaitannya dengan cara memilih jalur dalam jaringan sosial, di mana jalur yang paling efisien dan sering digunakan untuk menghubungkan dua *node*s adalah jalur dengan jarak terpendek atau shortest path. Individu yang memiliki skor *closeness centrality* yang paling tinggi dianggap sebagai orang yang paling sentral, karena mereka memiliki jumlah *shortest* path yang paling rendah untuk terhubung dengan semua orang lain di dalam jaringan sosial tersebut.

4. *Eigenvector Centrality*

   Eigenvector Centrality diartikan sebagai indikator seberapa pentingnya seseorang berdasarkan jumlah pentingnya orang-orang di sekitarnya. Dalam kalkulasi *eigenvector*, setiap interaksi seseorang dengan orang lain meningkatkan nilai *centrality* mereka sebanding dengan nilai *centrality* dari *node* yang mereka interaksikan. Contohnya, seseorang dianggap penting jika dia memiliki hubungan dekat dengan individu penting lainnya.

2.3.4 *Naïve Bayes*

Naïve bayes termasuk kedalam algoritma klasifikasi, teorema ini diungkapkan oleh ilmuwan inggris untuk memprediksi peluang masa depan berdasarkan pengalaman di masa sebelumnya [43]. *Naïve bayes* adalah metode klasifikasi yang akan menghitung probabilitas dengan menjumlahkan frekuensi nilai dataset yang digunakan. Algoritma ini juga dikatakan sebagai salah satu algoritma yang sederhana [44]. *Naïve bayes* memiliki ciri utama yaitu, memiliki asumsi independensi yang sangat kuat dari masing-masing kejadian. Kegunaan *naïve bayes* yaitu, untuk klasifikasi dokumen teks, membuat diagnosis medis, mendeteksi, metode *machine learning* yang menggunakan probabilitas [45]. Pada algoritma ini mengasumsikan bahwa suatu fitur pada kelas tidak memiliki hubungan dengan fitur lain pada kelas yang sama. *Naïve bayes* menghasilkan label kategori yang paling tinggi probabilitasnya. Persamaan teorema Bayes [46]:

$$\mathrm{P}(C_i|X) = \frac{P(X|C_i)\,P(C_i)}{P(X)} \qquad (6)$$

Rumus 2. 6 Persamaan *teorema Bayes*

Keterangan:

$\mathrm{P}(C_i|X)$: Probabilitas hipotesis Ci

$\mathrm{P}(X|C_i)$: Mencari nilai parameter yang memberikan kemungkinan yang paling besar

P(Ci): Probability X (*Prior prability*)

P(X): Jumlah *probability* yang muncul

2.3.5 *Random Forest*

Random Forest merupakan salah satu metode yang telah dikembangkan oleh CART (*Classification and Regression Trees*). Selain itu, *Random Forest* menjadi kombinasi dari teknik *decision tree* yang ada dan digabungkan

kedalam satu model [47]. Langkah-langkah untuk metode *Random Forest* yaitu, menghasilkan set pelatihan baru dengan sampel acak, membangun sebuah *tree* untuk pemilihan fitur acak pada setiap simpul *tree* tanpa melakukan pemotongan, memprediksi data baru dengan menggabungkan hasil semua *tree* [48]. Algoritma *Random Forest* memiliki kelebihan, yaitu bisa meningkatkan akurasi meskipun memiliki *missing value* serta untuk *resisting outliers*. Selain itu, *Random Forest* memiliki proses seleksi fitur yang mampu mengambil fitur yang terbaik [49]. *Random Forest* dapat dikembangkan melalui teknik bagging yang memilih atribut secara acak. Penggunaan metode *CART (Classification and Regression Tree)* memungkinkan pembentukan pohon keputusan yang berkembang sampai mencapai ukuran penuh tanpa proses pemangkasan, membentuk suatu kumpulan pohon yang dikenal sebagai forest [50].

$$f_{ij} = \frac{\sum \text{j:node j splits on feature i nij}}{\sum k \in all\ nodes\ ni} \quad (7)$$

Rumus 2. 7 Rumus *Random Forest*

Keunggulan dari metode *Random Forest* adalah [51]:

1. Tingkat akurasi yang tinggi,
2. Ketahanan yang baik terhadap *outliers* dan *noise*,
3. Kecepatan yang lebih unggul dibandingkan teknik *bagging* dan *boosting*, serta
4. Kemudahan dalam implementasi dan potensi untuk diparalelkan.

2.3.6 *Support Vector Machine*

Support Vector Machine merupakan algoritma *machine learning* yang dapat digunakan untuk regresi dan klasifikasi [52]. Cara kerja algoritma ini dengan mendefinisikan batasan antara kelas dengan jarak maksimal dari data terdekat. Membentuk *hyperplane* (garis pemisah) menjadi cara untuk mendapatkan batas maksimal antar kelas. Selain itu, *Support Vector Machine* dapat melakukan klasifikasi secara *linier* dan *non-linear* [53]. Berikut merupakan rumus untuk mencari hyperlane terbaik [13].

$$w \cdot x + b = 0 \quad (8)$$

Rumus 2. 8 Rumus *SVM*

Hyperplane yang dihasilkan oleh *SVM* berada tepat di tengah antara dua kelas. Ini berarti bahwa jarak dari *hyperplane* ke titik data terdekat dari kedua kelas yang berlawanan, yang ditandai dengan lingkaran kosong dan simbol positif. Dalam model *SVM,* data paling tepi yang berlokasi paling dekat dengan *hyperplane* disebut sebagai *support vector. Support vector* ini merupakan data yang paling kompleks untuk diklasifikasi karena letaknya yang nyaris bersinggungan atau beririsan dengan kelas yang berbeda [54].

2.4 *Software* dan *Tools* yang digunakan

2.4.1 *Google Colab*

Google Colab atau *Google Colaboratory* adalah *software open source* berbasis *cloud* dari layanan *Jupyter Notebook*. Layanan ini disediakan oleh Google dan dirancang untuk memudahkan penelitian *machine learning* dan *data science*. *Google Colab* menawarkan akses gratis ke sumber daya komputasi, termasuk *GPU (Graphics Processing Unit)* dan *TPU (Tensor Processing Unit)*, yang memungkinkan pengguna untuk menjalankan kode yang memerlukan kekuatan komputasi tinggi [55].

Gambar 2.2 Logo Google Colab

Pada Gambar 2.2 merupakan logo dari *google colab. Google colab* sering digunakan untuk mengeksekusi kode *Python*. *Google colab* memiliki fitur yang dapat digunakan secara bersamaan dengan user lain. Pada penelitian ini *google colab* digunakan untuk mengeksekusi dan menjalankan *code* algoritma *Machine Learning* dengan menggunakan bahasa *python* untuk menganalisis dataset yang telah dikumpulkan.

2.4.2 *Python*

Menurut survei pengembang Stack Overflow tahun 2022, *Python* menempati peringkat keempat sebagai bahasa pemrograman paling populer. Hampir setengah dari responden menyatakan bahwa mereka menggunakan bahasa pemrograman ini dalam hampir setengah dari waktu kerja mereka [31]. *Python* adalah bahasa pemrograman yang bersifat open source dan memiliki dukungan untuk pengelolaan data yang beragam implementasi [56]. *Python* memiliki kepustakaan yang luas dan telah disediakan modul-modul 'siap pakai' untuk berbagai keperluan. *Python* juga memiliki tata bahasa yang jernih dan mudah dipelajari serta memiliki aturan layout kode sumber yang jelas [57].

2.4.3 *Platform X*

X merupakan *platform* yang memungkinkan teman, keluarga, dan rekan kerja untuk tetap terhubung dan berkomunikasi melalui pertukaran pesan yang cepat dan sering. Pengguna dapat berbagi post yang mengandung foto, video, tautan, dan teks. Post ini dipublikasikan di profil pengguna, dikirimkan kepada pengikut, dan dapat dicari melalui fungsi pencarian di X[58]. Berbeda dari *platform* media sosial lainnya, X membatasi jumlah karakter dalam penulisan pesan hingga 280 karakter, sementara *platform* lain tidak memiliki batasan seperti itu. X memiliki keunggulan berupa jangkauan yang ekstensif, kemampuan untuk menghubungi tokoh publik, media promosi yang lebih menyeluruh, jaringan yang luas, serta kemudahan dalam mengukur efektivitasnya [59].

2.4.4 *Stremlit*

Streamlit adalah sebuah perpustakaan (*library*) *Python* yang bersifat *open source* dan dirancang untuk penggunaan yang mudah, sehingga memungkinkan pengembang aplikasi *Python* untuk mengonversi skrip data menjadi aplikasi web yang interaktif dengan lebih sederhana [60]. *Stremlit* tidak memerlukan pengetahuan mendalam tentang pengembangan *web*. Ada beberapa *library* yang harus di *install* untuk menjalankan stremlit, yaitu *library* streamlit, pandas, numpy [61].