BAB II

LANDASAN TEORI

2. 1 Objek Penelitian

2.1.1 Akun

Akun adalah catatan mengenai identifikasi pengguna, sandi, dan otorisasi untuk masuk ke dalam jaringan atau sistem *online* [9]. "Akun sosial media" menggambarkan profil atau identitas pengguna pada platform-platform seperti Facebook, Twitter (X), Instagram, dan LinkedIn. Dalam konteks ini, "akun" merujuk pada informasi yang mencakup data pribadi, aktivitas, dan materi yang dibagikan oleh pengguna pada platform tersebut. Melalui akun sosial media, pengguna dapat berinteraksi dengan teman-teman, keluarga, atau individu lain secara daring.

Fitur-fitur yang umumnya terdapat dalam akun sosial media mencakup profil pengguna yang berisi foto profil, biodata, dan informasi lainnya. Terdapat juga aktivitas *feed* yang memungkinkan pengguna melihat pembaruan dan konten dari pengguna lain yang diikuti. Kemudian, ada fitur koneksi dan pertemanan yang memungkinkan pengguna menjalin hubungan sosial *online*, serta pengaturan privasi untuk mengontrol akses terhadap konten pengguna. Selain itu, akun memungkinkan pengguna untuk memposting berbagai jenis konten, seperti teks, foto, video, dan tautan. Interaksi juga dapat dilakukan melalui memberikan suka, komentar, atau berbagi konten dari pengguna lain [10].

Dengan akun sosial media, individu dapat berkomunikasi, berbagi pengalaman, dan terlibat dalam interaksi sosial di dunia maya. Meskipun demikian, penting bagi pengguna untuk mempertimbangkan privasi dan keamanan dalam penggunaan akun sosial media, serta memperhatikan kebijakan dan aturan yang berlaku di setiap platform. Contoh akun pada media sosial X dapat terlihat pada gambar 2.1 di bawah ini, yaitu atas nama *user* @jokowi dan @elonmusk.

Gambar 2.1 Dua contoh akun di X

Dalam perkembangannya, akun-akun yang ada di media sosial juga diklasifikasi menjadi akun yang dicurigai atau *suspicious account* (akun spam).

2.1.2 *Suspicious Account*

Suspicious account atau akun mencurigakan mengacu pada akun pengguna dalam *platform digital* atau daring yang menimbulkan kekhawatiran atau kecurigaan karena perilakunya yang tidak biasa atau berpotensi membahayakan. Perilaku tersebut mungkin mencakup pola *login* yang tidak teratur, aktivitas mencurigakan, atau penyimpangan dari perilaku pengguna pada umumnya [10]. *Suspicious account* mengacu pada akun di platform daring yang menimbulkan keraguan atau kecurigaan terkait keasliannya. Istilah ini sering digunakan untuk mengidentifikasi akun yang mungkin terlibat dalam perilaku melanggar aturan atau tidak jujur, seperti penipuan, penyebaran spam, penyebaran informasi palsu, atau tindakan ilegal lainnya di lingkungan daring.

Faktor-faktor yang dapat membuat sebuah akun dianggap mencurigakan termasuk pola aktivitas yang tidak wajar, inkonsistensi dalam informasi profil, jumlah pengikut yang tidak biasa, penyebaran informasi palsu, dan perubahan perilaku yang tiba-tiba. Situs web dan platform media sosial biasanya memiliki mekanisme deteksi dan respons terhadap akun-akun yang dicurigai, dan pengguna juga diharapkan untuk melaporkan

akun-akun semacam itu untuk tindakan lebih lanjut guna menjaga keamanan dan integritas platform tersebut [6]. Dalam perkembangannya, terutama pada penelitian ini *suspicious account* merupakan objek penelitian yang diperlu diteliti melalui kegiatan klasterisasi atau *clustering* dan klasifikasi (*classification*) melalui *machine learning*.

2.1.3 *Machine Learning*

Machine learning atau pembelajaran mesin adalah bagian dari kecerdasan buatan (AI) yang berfokus pada pengembangan sistem untuk meningkatkan kinerja manusia dalam melakukan tugas secara otomatis demi efisiensi yang lebih baik [11]. Dalam *machine learning*, terdapat proses *training* dan pembelajaran. Meskipun terdapat berbagai teknik dalam machine learning, secara umum terdapat dua teknik dasar pembelajaran, yaitu *supervised* dan *unsupervised learning*. *Supervised learning* adalah teknik yang diterapkan dalam pembelajaran mesin di mana model dapat memanfaatkan informasi yang sudah ada dalam data dengan memberikan label tertentu. Label ini digunakan oleh mesin untuk mengelompokkan objek berdasarkan kriteria tertentu. Sementara itu, *unsupervised learning* adalah teknik yang diterapkan pada pembelajaran mesin menggunakan data yang tidak memiliki informasi yang sudah ditetapkan. Dalam teknik ini, mesin melakukan prediksi untuk membantu mengidentifikasi struktur atau pola tersembunyi dalam data yang telah dipelajari selama proses *training*.

Machine Learning (ML) adalah paradigma kecerdasan buatan yang memberikan kemampuan komputer untuk belajar dari data dan mengembangkan keahlian tanpa perlu diatur secara eksplisit. Konsep utama di dalam machine learning melibatkan penggunaan algoritma untuk menganalisis dan memahami pola dalam data. Selama proses pelatihan, model *machine learning* disesuaikan dengan data pelatihan untuk dapat membuat prediksi atau keputusan yang akurat saat dihadapkan dengan data baru. Terdapat beberapa jenis *machine learning*, seperti *supervised learning*

yaitu mempelajari pola dari *labelled data*, *unsupervised learning* yang menemukan pola dalam data tanpa label, dan *reinforcement learning* di mana model belajar dari interaksi dengan lingkungannya melalui umpan balik berupa hadiah atau hukuman [12]

Penerapan *machine learning* melibatkan proses membangun model yang dapat digunakan untuk membuat prediksi atau keputusan berdasarkan data yang diberikan. Setelah model dilatih, itu diuji dengan menggunakan data yang tidak pernah dilihat sebelumnya untuk mengevaluasi kinerjanya. Machine learning telah diterapkan dalam berbagai bidang, termasuk pengenalan pola, analisis teks, otomasi tugas, dan pengembangan sistem cerdas. Seiring dengan kemajuan teknologi dan ketersediaan data yang melimpah, *machine learning* terus menjadi bidang penelitian dan pengembangan yang dinamis, membawa dampak signifikan dalam berbagai aspek kehidupan modern [13].

Dalam perkembangannya, machine learning dibagi menjadi dua, *unsupervised learning* dan *supervised learning*.

2.1.4 *Supervised Learning*

Supervised learning merupakan subkategori pembelajaran mesin dan kecerdasan buatan. *Supervised learning* ditandai dengan penggunaan dataset berlabel untuk melatih algoritma-algoritma yang dapat mengklasifikasikan data atau memprediksi hasil dengan akurat. Saat data masukan dimasukkan ke dalam model, bobotnya disesuaikan sampai model tersebut sesuai dengan baik, yang terjadi sebagai bagian dari proses validasi silang. Pembelajaran terawasi membantu organisasi memecahkan berbagai masalah dunia nyata dalam skala besar, seperti mengklasifikasikan spam ke folder terpisah dari *e-mail* [14].

Supervised learning menggunakan set *training* untuk mengajarkan model agar menghasilkan keluaran yang diinginkan. Dataset *training* ini

mencakup masukan dan keluaran yang benar, yang memungkinkan model untuk belajar dari waktu ke waktu. Algoritma mengukur akurasinya melalui fungsi kerugian, menyesuaikan hingga kesalahan telah cukup diminimalkan [13].

Supervised learning dapat dibagi menjadi dua jenis masalah dalam penggalian data—klasifikasi dan regresi [15]:

a. Klasifikasi menggunakan algoritma untuk secara akurat menetapkan data uji ke dalam kategori-kategori tertentu. Ini mengenali entitas-entitas tertentu dalam dataset dan berupaya menyimpulkan bagaimana entitas-entitas tersebut seharusnya diberi label atau didefinisikan. Algoritma klasifikasi umum meliputi klasifikasi linear, *k-nearest neighbors, decision tree, support vector machine* (SVM), *k-nearest neighbors*, dan *random forest*.
b. Regresi digunakan untuk memahami hubungan antara variabel dependen dan independen. Ini umumnya digunakan untuk membuat proyeksi, seperti pendapatan penjualan untuk bisnis tertentu. Regresi linear, regresi logistik, dan regresi polinomial adalah algoritma regresi yang populer.

2.1.5 Classification

Klasifikasi (*classification*) merujuk pada proses *model or algorithm development* yang mampu mengelompokkan atau melakukan prediksi label kelas terhadap data yang dipakai. Hal ini adalah metode analisis data yang memungkinkan identifikasi atau pengelompokan entitas berdasarkan atribut yang terkait [2]. Kelebihan *classification* melibatkan kapabilitasnya dalam menghasilkan keputusan yang obyektif dan konsisten tanpa dipengaruhi oleh faktor emosional atau subjektivitas. Proses algoritma ini memungkinkan diterapkan dengan tujuan menganalisis data kompleks dengan jumlah atribut yang banyak. Aplikasi klasifikasi sangat bervariasi, contohnya yaitu *email spam filtering*,

diagnosis dalam bidang medis, *credit scoring*, *image recognition*, *speech recognition*, *sentiment analysis*, *predictive maintenance*, *fraud detection*, dan *hoax detection* [16].

Jenis data dan tujuan analisis menentukan metode klasifikasi yang tepat. Dalam implementasi, elemen seperti validasi model, interpretasi hasil, dan evaluasi kinerjanya harus dipertimbangkan. [4]. Keandalan dan pemahaman hasil klasifikasi dapat dipastikan melalui teknik validasi silang dan interpretasi model. Hasil klasifikasi seperti*Support Vector Machine (SVM), Artificial Neural Networks (ANN), Naive Bayes, Decision Tree, Random Forest, dan k-Nearest Neighbors (k-NN)* termasuk dalam kategori ini. Setiap algoritma memiliki kelebihan dan kekurangan, dan jenis data dan tujuan analisis menentukan algoritma yang dipilih [17].

2.1.6 SVM *Hyperparameter*

Hyperparameter SVM adalah pengaturan yang dapat sesuaikan sebelum melatih model dan memengaruhi kesesuaian model dengan data. Biasanya, ini mencakup fungsi kernel, yang memetakan data ke ruang berdimensi lebih tinggi di mana batas keputusan linier dapat ditemukan. Ada berbagai jenis kernel, seperti linier, polinomial, fungsi basis radial (RBF), dan sigmoid. Selain itu, ada parameter regularisasi, atau C, yang menyeimbangkan antara memaksimalkan margin dan meminimalkan kesalahan *training*. Nilai C yang lebih tinggi menekankan pada penyesuaian data, sedangkan nilai C yang lebih rendah mengutamakan menghindari overfitting. Terdapat pula koefisien kernel, atau gamma, yang mempengaruhi bentuk dan kelancaran batas keputusan. Nilai gamma yang lebih tinggi membawa lebih banyak kompleksitas dan fleksibilitas; sebaliknya, nilai gamma yang lebih rendah menghasilkan kesederhanaan dan keumuman [18].

Mengoptimalkan hyperparameter SVM penting karena dapat membuat perbedaan signifikan dalam akurasi dan kemampuan generalisasi model.

Apabila hyperparameter yang dipilih salah, maka model dapat terjadi *underfitting* atau *overfitting* pada data, yang berarti model tersebut akan berperforma buruk pada data baru dan tidak terlihat. *Underfitting* artinya model terlalu sederhana dan tidak dapat menangkap kompleksitas dan pola data, sedangkan *overfitting* berarti model terlalu kompleks dan tidak dapat beradaptasi dengan variabilitas dan *noise* data [19].

Untuk menghindari *underfitting* dan *overfitting*, perlu menemukan *hyperparameter* optimal yang menyeimbangkan *trade-off bias-varians*. Bias adalah kesalahan yang disebabkan oleh asumsi dan penyederhanaan model, sedangkan varians adalah kesalahan yang disebabkan oleh sensitivitas dan ketidakstabilan model [18].

Untuk meningkatkan efisiensi dan efektivitas proses pengoptimalan, dipertimbangkan melakukan normalisasi atau standarisasi data sebelum menerapkan SVM, karena hal ini dapat meningkatkan performa dan stabilitas model serta mengurangi sensitivitas terhadap hyperparameter. Selain itu, digunakan set validasi atau teknik validasi silang untuk mengevaluasi performa model dan menghindari kesalahan yang berlebihan atau meremehkan. Setelah hyperparameter SVM dioptimalkan, dapat diterapkan pada masalah klasifikasi industri dan mendapatkan manfaat dari model yang kuat dan andal. Contoh masalah tersebut meliputi deteksi kesalahan, pengendalian kualitas, dan optimalisasi proses. Untuk menggunakan SVM dalam skenario ini, harus terlebih dahulu menentukan tujuan dan mengumpulkan data. Pemrosesan awal mungkin diperlukan, seperti mengekstraksi fitur atau menyeimbangkan kelas. Kemudian harus melatih dan menguji model dengan membagi data menjadi beberapa kumpulan dan menerapkan *hyperparameter* yang dioptimalkan. Terakhir, harus menerapkan dan memantau model, seperti mengintegrasikannya ke dalam sistem atau proses, memperbaruinya dengan data baru, dan menilai keakuratan dan keandalannya [20].

2.1.7 Akurasi

Akurasi adalah persentase klasifikasi yang benar yang dicapai oleh model pembelajaran mesin yang telah dilatih, yaitu jumlah prediksi yang benar dibagi dengan total prediksi di semua kelas. Akurasi dilaporkan sebagai nilai antara [0,1] atau [0, 100], tergantung pada skala yang dipilih. Akurasi 0 berarti klasifier selalu memprediksi label yang salah, sedangkan akurasi 1, atau 100, berarti selalu memprediksi label yang benar. Karakteristik bagus dari metrik ini adalah memiliki hubungan langsung dengan semua nilai *confusion matrix*. Ini merupakan empat pilar evaluasi supervised learning: *true positives, false positives, true negatives, dan false negatives*. Akurasi adalah ukuran proporsional dari jumlah prediksi yang benar dibandingkan dengan semua prediksi [3].

Dengan demikian, dapat didefinisikan akurasi sebagai berikut:

$$Akurasi = \frac{Jumlah\ Total\ Data}{Jumlah\ Prediksi\ Benar} \times 100\%$$

Gambar 2.2 Rumus Accuracy (sumber: [3])

Di mana:

a. Jumlah Prediksi Benar adalah jumlah data yang diklasifikasikan dengan benar oleh model.

b. Jumlah Total Data adalah total keseluruhan data yang dievaluasi oleh model.

Penting juga untuk menekankan bahwa evaluasi akurasi model sebaiknya dilakukan pada jumlah prediksi yang signifikan secara statistik seperti pada evaluasi metrik apapun.

2.1.8 *Precision*

Presisi adalah metrik yang digunakan dalam evaluasi model klasifikasi, khususnya dalam konteks masalah klasifikasi biner. Ini mengukur akurasi prediksi positif yang dibuat oleh model, menunjukkan proporsi instansi

positif yang diprediksi dengan benar di antara semua instansi yang diprediksi sebagai positif [12]. Presisi dihitung menggunakan rumus berikut:

$$Presisi = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Gambar 2.3 Rumus Presisi (sumber: [12])

Di mana:

a. Positif Benar (TP): Jumlah instansi yang diprediksi dengan benar sebagai positif.
b. Positif Salah (FP): Jumlah instansi yang diprediksi secara salah sebagai positif.

Presisi tinggi menunjukkan bahwa model membuat prediksi positif dengan tingkat akurasi yang tinggi, meminimalkan kemungkinan terjadinya positif palsu. Namun, presisi sebaiknya dipertimbangkan bersama dengan metrik lain, seperti *recall*, skor F1, dan akurasi, untuk memberikan evaluasi menyeluruh terhadap kinerja model. Pemilihan metrik mana yang harus diprioritaskan tergantung pada tujuan dan persyaratan spesifik dari tugas yang sedang dijalankan.

2.1.9 *Recall*

Recall, juga dikenal sebagai sensitivitas atau tingkat positif benar, adalah metrik yang digunakan dalam evaluasi model klasifikasi. *Recall* mengukur kemampuan model untuk mengidentifikasi dengan benar semua instansi yang relevan, khususnya proporsi positif benar yang diprediksi dengan benar dari total instansi positif sebenarnya [13]. *Recall* dihitung menggunakan rumus berikut:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Gambar 2.4 Rumus Recall (sumber: [13])

Di mana:

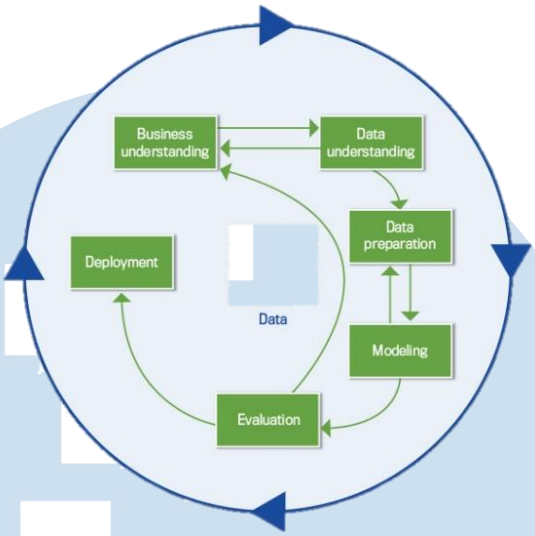
a. Positif Benar (TP): Jumlah instansi yang diprediksi dengan benar sebagai positif.
b. Negatif Salah (FN): Jumlah instansi yang diprediksi secara salah sebagai negatif.

Recall yang tinggi menunjukkan bahwa model secara efektif menangkap sebagian besar instansi positif, meminimalkan kemunculan negatif palsu. Namun, seperti halnya presisi, *recall* sebaiknya dipertimbangkan bersama dengan metrik lain seperti presisi, skor F1, dan akurasi untuk evaluasi menyeluruh terhadap kinerja model, tergantung pada tujuan dan persyaratan spesifik dari tugas yang sedang dijalankan.

2.1 Framework dan Algoritma yang digunakan

2.2.1 *Framework*: CRISP-DM

CRISP-DM, singkatan dari *Cross-Industry Standard Process for Data Mining*, adalah suatu metode standar dalam proses data mining yang digunakan sebagai panduan untuk mengatasi masalah umum dalam konteks bisnis maupun penelitian [21] . CRISP-DM memiliki 6 fase, di antaranya sebagai berikut:



Gambar 2.5 CRISP-DM

Sumber: [22]

a. *Business understanding*, fokus dalam tahap pemahaman bisnis adalah memahami tujuan dan kebutuhan dari sudut pandang bisnis. Setelah itu, informasi tersebut diekstraksi untuk merumuskan masalah dan merencanakan langkah-langkah yang diperlukan untuk mencapai tujuan bisnis yang sudah ditetapkan.
b. *Data understanding*, diawali melalui pengambilan data dan proses pemahaman serta pendekatan terhadap data yang diperoleh. Tahap ini juga bertujuan untuk mengidentifikasi masalah kualitas data, menggali wawasan dari data, dan menemukan informasi tersembunyi dalam data yang terkumpul.
c. *Data preparation*, tahap persiapan data melibatkan pengolahan dan pembersihan data mentah hingga diperoleh data final yang akan digunakan dalam analisis selanjutnya.
d. *Modeling*, tahap pemodelan mencakup penerapan teknik pemodelan dan pemilihan model terbaik dengan parameter optimal untuk mencapai kinerja yang optimal.
e. *Evaluation*, tahap evaluasi melibatkan peninjauan mendalam terhadap model yang digunakan dan penilaian lebih lanjut terhadap kemampuan model untuk mencapai tujuan penelitian.

f. *Deployment,* tahap implementasi melibatkan proses penggunaan model oleh pengguna melalui pembuatan aplikasi atau situs web yang dapat diakses oleh pengguna. Model didistribusikan dengan tujuan memberikan kemampuan kepada pengguna akhir untuk menggunakan data sebagai dasar dalam pengambilan keputusan atau sebagai pendukung dalam proses bisnis. Walaupun model ini bertujuan untuk meningkatkan pemahaman tentang data, penting bagi pemahaman tersebut diorganisir, disajikan, dan didistribusikan dengan cara yang memungkinkan pengguna akhir untuk menggunakannya. Bergantung pada kebutuhan, aktivitas penyebaran dapat berupa hal seiring sederhana seperti pembuatan laporan dengan analisis mendalam agar mudah dipahami oleh pengguna akhir.

2.2.2 Algoritma

Algoritma yang digunakan pada penelitian ini adalah SVM. SVM atau (*Support Vector Machine)* merupakan algoritma klasifikasi terkemuka dalam domain penelitian di bidang *machine learning*. SVM memiliki dasar yang solid dalam teori statistik dan memiliki efektivitas yang tinggi dalam menangani permasalahan klasifikasi, baik yang bersifat biner maupun multi-kelas. Keunggulan utama algoritma ini terletak pada kemampuannya dalam mengklasifikasikan dataset yang besar, bahkan ketika data tersebut memiliki atribut yang kompleks [23]. SVM menggunakan berbagai fungsi untuk mengoptimalkan pemisahan antar kelas data. Fungsi pemisah, atau *hyperplane*, dapat ditemukan melalui penelusuran dan dapat berbentuk dataran datar atau garis, tergantung pada dimensi ruang yang digunakan. [24] .

Support Vector Machine (SVM) memiliki beberapa kelebihan yang membuatnya menjadi pilihan yang populer dalam klasifikasi data. Pertama, SVM efektif dalam menangani dataset yang memiliki dimensi tinggi, sehingga cocok untuk masalah klasifikasi di mana jumlah atribut atau fitur sangat besar. Kelebihan lainnya adalah kemampuannya mengatasi masalah klasifikasi non-linear melalui penggunaan fungsi kernel, yang

memungkinkan SVM menangani hubungan yang kompleks antara variabel input. Fungsi kernel memungkinkan SVM untuk memetakan data ke dalam dimensi yang lebih tinggi, di mana pola klasifikasi yang kompleks dapat diidentifikasi dengan lebih baik.

Selain itu, SVM memiliki konsep margin yang membantu meningkatkan generalisasi model. Margin ini mengukur sejauh mana batas keputusan dari SVM terhadap data training, dan SVM berusaha untuk memaksimalkan margin ini. Dengan memaksimalkan margin, SVM dapat meningkatkan toleransinya terhadap data noise atau outlier, menghasilkan model yang lebih stabil dan umumnya memiliki kinerja yang baik dalam menerapkan klasifikasi pada data baru yang belum pernah dilihat sebelumnya.

Dengan memanfaatkan vektor pendukung dan teknik optimasi matematika, Support Vector Machine (SVM) memetakan batas pemisah optimal antara kelas data. SVM beroperasi dalam paradigma supervised learning, di mana model dikonstruksi berdasarkan data latih yang telah diberi label [25]. Keunggulan SVM melibatkan kemampuan generalisasi yang optimal, penanganan masalah overfitting pada dataset pelatihan, serta fleksibilitas dalam menangani berbagai tantangan klasifikasi, termasuk dataset dengan dimensi tinggi atau pola non-linear. Selain itu, SVM terbukti efektif dalam menangani data yang tidak seimbang [26].

Meskipun Support Vector Machine (SVM) memiliki banyak kelebihan, terdapat beberapa kelemahan yang perlu dipertimbangkan. Salah satu kelemahan utama SVM adalah sensitivitas terhadap pemilihan parameter, seperti parameter regulasi dan pemilihan kernel. Menentukan nilai yang optimal untuk parameter ini dapat menjadi tantangan, dan keputusan yang kurang tepat dapat menghasilkan performa model yang kurang optimal. Selain itu, SVM cenderung membutuhkan sumber daya komputasi yang signifikan, terutama ketika menangani dataset besar, sehingga dapat

menjadi kurang efisien untuk beberapa aplikasi dalam konteks perbandingan dengan beberapa algoritma klasifikasi lainnya [27].

2.2 Tools / Software yang digunakan

2.3.1 Python

Python adalah bahasa pemrograman tingkat tinggi yang serbaguna, yang didesain dengan fokus pada keterbacaan dan sintaksis yang jelas. Dikembangkan oleh Guido van Rossum, Python pertama kali diperkenalkan pada tahun 1991 dan sejak itu telah berkembang menjadi salah satu bahasa pemrograman paling populer di dunia. Keunggulan Python meliputi kemampuannya dalam mendukung paradigma pemrograman berorientasi objek, fungsional, dan prosedural, serta memiliki berbagai pustaka dan modul yang melimpah. Python digunakan secara luas dalam berbagai bidang, termasuk pengembangan perangkat lunak, analisis data, kecerdasan buatan, pengembangan web, dan rekayasa perangkat lunak, menjadikannya salah satu pilihan utama bagi para pengembang di seluruh dunia [28]. Meskipun termasuk dalam bahasa pemrograman tingkat tinggi, Python didesain agar mudah dipelajari dan dipahami. Kelebihan Python meliputi kemudahan pembelajaran, kemampuan menjalankan program kompleks dengan sedikit kode, serta kemampuan mengubah program yang kompleks menjadi lebih sederhana [29]. Berikut adalah beberapa kekuatan dan kelemahan Python [30].

Adapun kekuatan dari Python adalah sebagai berikut:

a. Python menduduki peringkat kelima sebagai bahasa pemrograman paling penting dan paling populer untuk pembelajaran mesin dan ilmu data menurut penelitian dan survei.
b. Sintaksis Python disederhanakan untuk memudahkan pemahaman, dengan fokus pada bahasa alami, membuatnya lebih mudah ditulis dan dieksekusi.
c. Python bersifat dinamis, mendukung berbagai paradigma pemrograman, termasuk fungsional, prosedural, dan OOP.

d. Python merupakan bahasa pemrograman yang powerful dan dapat diterapkan di berbagai platform, termasuk pengembangan web, aplikasi *mobile*, dan *desktop*.
e. Dukungan komunitas *opensource* yang besar menjadikan Python kuat, adaptif, dan *bug* dapat diperbaiki dengan cepat.

Di samping itu, Python juga memiliki beberapa kelemahan, di antaranya sebagai berikut:

a. Tidak ideal untuk *case* yang membutuhkan penggunaan memori tinggi karena konsumsi memori yang relatif besar dan dukungan yang kurang baik dari multiprosesor.
b. Python memiliki popularitas yang lebih rendah dibandingkan dengan Java dan Kotlin dalam konteks pengembangan aplikasi *mobile*, hal ini disebabkan oleh beberapa pembatasan desain serta performa yang cenderung lebih lambat.

Dalam pemrograman Python, modul dapat digunakan dan diorganisir dalam suatu folder atau *package*. Modul-modul ini kemudian dapat digabungkan menjadi *library*. Python memiliki lebih dari 140.000 *library*, yang terus bertambah, dan sebagian besar bersifat *open source*. Pengguna dapat dengan mudah mengakses dan menggunakan *library* ini secara gratis. Menggunakan *library* ini dapat memungkinkan penulisan kode yang lebih efisien, sistematis, dan dapat menghemat waktu. *Library* pada Python bersifat *reusable*, memungkinkan pengguna menggunakannya berulang kali secara praktis [31].

2.3.2 Google Data Studio

Google Collaboratory, atau Google Data Studio, adalah sebuah *platform* sumber terbuka yang dipersembahkan oleh Google untuk pengembangan dan pelaksanaan kode [32]. Colab memberikan manfaat agar *user* dapat *write and run* sintaks Python secara interaktif melalui browser tanpa memerlukan pengaturan atau instalasi lokal. Platform ini

menggunakan lingkungan Jupyter Notebook dan menyediakan akses gratis ke GPU (*Graphics Processing Unit*) untuk akselerasi perhitungan, memungkinkan pelatihan model machine learning yang lebih cepat. Colab juga mendukung integrasi dengan Google Drive, memudahkan penyimpanan, berbagi, dan kolaborasi pada proyek-proyek. Selain itu, Colab menyediakan akses ke berbagai pustaka populer seperti TensorFlow, PyTorch, dan scikit-learn, membuatnya menjadi alat yang sangat berguna untuk eksplorasi data, pengembangan model, dan pembelajaran mesin. Google Colab mendukung penelitian yang bergantung pada sumber daya komputasi yang signifikan, seperti pemrosesan gambar, BDA, atau *machine learning*, dengan menyediakan akses ke perangkat keras dan *cloud computing* yang bertenaga tinggi. Salah satu contoh penggunaan Google Colab dalam penelitian adalah penelitian "Pengakuan gambar yang didasarkan pada pembelajaran mendalam untuk pengemudi otonom" [33]. Penelitian ini menggunakan Google Colab sebagai platform untuk mengembangkan dan melatih model deteksi objek berbasis pembelajaran mendalam dalam konteks kendaraan otonom. Para peneliti memanfaatkan Google Colab sebagai platform untuk penulisan dan eksekusi kode Python, akses *dataset*, instruksi dan pengujian model deteksi objek dengan menggunakan arsitektur jaringan saraf konvolusi, serta analisis dan presentasi hasil eksperimen. Dengan keunggulan Google Colab dalam menyediakan lingkungan yang bersahabat dan produktif, para peneliti dapat dengan cepat dan efisien melakukan eksperimen, dan mereka dapat berkolaborasi dengan rekan penelitian dengan berbagi kode dan hasil eksperimen mereka.

2.3.3 X

Salah satu platform media sosial yang meraih popularitas global adalah X (sebelumnya dikenal sebagai Twitter). X dikarakterisasikan dalam kategori layanan *social network and microblogging*, yang memungkinkan pengguna untuk berbagi dan berinteraksi melalui pesan [34]. Sebagai

platform yang terkenal dalam ranah media sosial, X juga dapat dianggap sebagai alat pemasaran yang memiliki potensi untuk word-of-mouth secara digital [35]. Jumlah pengguna X di seluruh dunia mencapai 566 juta, dan Indonesia menempati peringkat ke-5 sebagai negara dengan pengguna Twitter terbanyak di dunia, dengan jumlah pengguna mencapai 24 juta per Januari 2023 [36].

2.3.4 Microsoft Excel

Microsoft Excel adalah perangkat lunak yang digunakan di berbagai sektor untuk keperluan pembuatan, modifikasi, dan analisis data [31]. Software ini memberikan berbagai keuntungan dan fungsionalitas, termasuk kemampuan untuk menyusun anggaran, melakukan perhitungan, merancang diagram, dan berfungsi sebagai tempat penyimpanan *database* [32].

2.3.5 *Node*XL

NodeXL merupakan sebuah perangkat lunak (*software*) yang dirancang khusus untuk analisis jaringan sosial. NodeXL memudahkan pengguna dalam mengimpor, mengorganisir, dan menganalisis data jaringan sosial dari berbagai sumber, termasuk platform media sosial seperti Twitter, Facebook, dan LinkedIn. Dikembangkan sebagai *add-in* untuk Microsoft Excel, NodeXL memanfaatkan antarmuka pengguna yang akrab bagi pengguna Excel, memungkinkan mereka untuk melakukan analisis jaringan sosial tanpa harus menguasai keterampilan pemrograman atau perangkat lunak khusus [36]. NodeXL menyediakan berbagai fitur seperti visualisasi grafik, analisis sentralitas, identifikasi kluster, dan pengukuran kepadatan jaringan, yang membantu peneliti, analis, dan profesional untuk mendapatkan wawasan mendalam tentang struktur dan pola dalam jaringan sosial mereka. NodeXL juga mendukung berbagai format ekspor data

sehingga hasil analisis dapat dengan mudah dibagikan atau disajikan dalam format yang sesuai [37].

2.3 Penelitian Terdahulu

Sejumlah penelitian sebelumnya telah mencoba pendekatan berbeda dalam mengatasi tantangan identifikasi dan pengelompokan akun-akun yang mencurigakan. Misalnya, beberapa penelitian telah menggunakan teknik *cluster*ing untuk mengelompokkan data yang kompleks menjadi kategori yang lebih terkelompok. Namun, belum banyak penelitian yang secara khusus menggabungkan metode PCA dengan K-Means dalam konteks identifikasi akun mencurigakan. Selain itu, penggunaan model klasifikasi, seperti SVM, telah terbukti efektif dalam mengidentifikasi pola dan tren dalam data yang kompleks. *Previous research* yang digunakan sebagai acuan referensi dalam penelitian ini terlihat pada Tabel 2.1

Tabel 2.1 Penelitian Terdahulu

Judul	Penulis	Nama Jurnal atau *Proceeding*	Tahun	Volume	Fenomena	Referensi	Metode	Objek	Hasil
Who is Who on Twitter–Spammer, Fake or Compromised Account? A Tool to Reveal True Identity in Real-Time	M. Singh, D. Bansal, and S. Sofat	Cybern. Syst.	2018	49	Bahayanya spammer di media sosial	[10]	Nayes Net, Logistic Regression, J48, Random Forest, dan AdaBoostM1	Akun di Twitter	Hasil percobaan menunjukkan bahwa Random Forest *classifier* mampu memprediksi *spammer* dengan akurasi sebesar 92,1%.
Recognizing Fake Headlines Using Clustering Algorithms	J. A. A. Mary Sowjanya	Math. Stat. Eng. Appl.	2022	71	Rendahnya kredibilitas sumber berita di masa pandemi COVID-19.	[8]	*K-Means* dan *SVM*	*Headlines News*	K-Means merupakan algoritma *clustering* terbaik dan SVM merupakan algoritma *classification* dengan akurasi tertinggi.
Improving fake news detection using k-means and support vector machine approaches	S. Yazdi, K. M., Yazdi, A. M., Khodayi, S., Hou, J., Zhou, W., & Saedy	Int. J. Electron. Commun. Eng	2020	14	Berita palsu dan informasi palsu yang menjadi tantangan besar di media sosial	[38]	*K-Means* dan *SVM*	*False information, fake likes, views and duplicated accounts as big social networks*	K-means untuk mengurangi fitur-fitur yang tidak relevan dan SVM untuk meningkatkan presisi dalam algoritme

Judul	Penulis	Nama Jurnal atau *Proceeding*	Tahun	Volume	Fenomena	Referensi	Metode	Objek	Hasil
									pendeteksian berita palsu.
Twitter spam account detection based on clustering and classification methods	K. S. Adewole, T. Han, W. Wu, H. Song, and A. K. Sangaiah	J. Supercomput.	2020	76	Pertumbuhan interaksi sosial Twitter meningkatkan aktivitas *spammer*	[6]	*K-means, Random Forest, SVM*	*Twitter account*	Kinerja pengklasifikasi yang dipilih berdasarkan ketidakseimbangan kelas juga mengungkapkan bahwa Random Forest mencapai akurasi, presisi, recall, dan F-measure tertinggi.
Spam Detection on Profile and Social Media Network using Principal Component Analysis (PCA) and K-means Clustering	S. A. Sanjaya and K. Surendro	Int. J. Adv. Soft Compu. Appl	2019	11	Ekosistem media sosial yang ada dipengaruhi oleh pengaruh tokoh masyarakat, *trending* topik, spam, dan *spammer*.	[39]	*K-Means*	*Profile and social media network*	Penggunaan k-means dan PCA dapat mengidentifikasi *cluster* spam yang lebih spesifik daripada metode *supervised learning*.
Classification of Fake News on Facebook a Novel Social Network with K-Means	R. Nomesh. and M. S. Saravanan	2022 3rd International Conference on Smart Electronics and	2022		Maraknya berita palsu yang tersebar di Facebook	[40]	K-Means dan PCA	Berita palsu di Facebook	*Clustering* K-Means dan algoritma PCA dan memperkirakan algoritma terbaik

Judul	Penulis	Nama Jurnal atau *Proceeding*	Tahun	Volume	Fenomena	Referensi	Metode	Objek	Hasil
Clustering Approach for Against Principal Component Analysis Method for Better Accuracy		Communication (ICOSEC)							untuk akurasi yang lebih baik.
Target specific mining of COVID-19 scholarly articles using one-class approach	S. K. Sonbhadra, S. Agarwal, and P. Nagabhushan	Chaos, Solitons Fractals Nonlinear Sci. Nonequilibrium Complex Phenom	2020	140	Banyaknya artikel mengenai COVID yang tidak sesuai dengan pencarian yang diinginkan	[41]	K-Means, DBSCAN, HAC	COVID-19 *scholarly articles*	Ditemukan bahwa algoritme pengelompokan k-means, diikuti oleh OCSVM paralel, memiliki kinerja yang lebih baik daripada metode lain, baik untuk ruang fitur asli maupun yang telah dikurangi.
Fake News Detection using Machine Learning	J. Shaikh and Rupali Patil	2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC)	2020		Perlunya mendeteksi berita palsu yang memiliki pengaruh negatif dalam kehidupan manusia	[42]	SVM, Naïve Bayes, Passive Aggressive Classifier	*Twitter news article*	SVM merupakan algoritma dengan akurasi tertinggi

Judul	Penulis	Nama Jurnal atau *Proceeding*	Tahun	Volume	Fenomena	Referensi	Metode	Objek	Hasil
Fake News detection Using Machine Learning	N. F. Baarir and A. Djeffal	2020 2nd International workshop on human-centric smart environments for health and well-being (IHSH	2021		Pentingnya mendeteksi berita palsu	[43]	TF-IDF dan SVM	Teks, penulis, sumber, tanggal, dan sentimen secara berurutan	SVM terbukti sebagai algoritma terbaik dengan tingkat pengenalan yang optimal.
Fake News Detection Using Machine Learning Approaches	Khanam, Z Alwasel, B N Sirafi, H Rashid, M	IOP Conference Series: Materials Science and Engineering, Volume 1099, International Conference on Applied Scientific Computational Intelligence using Data Science (ASCI 2020) 22nd-23rd December 2020, Jaipur, India	2021	1099	Berita bohong di media sosial dan berbagai media lainnya tersebar luas dan menjadi perhatian serius karena mampu menimbulkan banyak kerugian sosial dan nasional dengan dampak yang merusak	[44]	Naïve Bayes, Neural Network, SVM	Total unique words (types), Type/Token Ratio (TTR), Number of sentences, Average sentence length (ASL), Number of characters, Average word length (AWL), nouns, prepositions, adjectives et	*Neural Network* dan SVM merupakan algoitma dengan akurasi tertinggi
A Review of Fake News Detection	A. K. Choudhary, M., Jha, S.,	2021 2nd International Conference for	2021		Meluasnya penggunaan media sosial	[45]	*Naïve Bayes, Convolutional Neural*	*Fake news in social media*	*SVM is the best algorithm for detecting fake news*

Judul	Penulis	Nama Jurnal atau *Proceeding*	Tahun	Volume	Fenomena	Referensi	Metode	Objek	Hasil
Methods using Machine Learning	Saxena, D., & Singh	Emerging Technology (INCET)			telah membawa dampak buruk bagi masyarakat akibat penyebaran berita palsu		*Network, LSTM, Neural Network, Support Vector Machine*		
A Smart System for Fake News Detection Using Machine Learning	A. Jain, A. Shakya, H. Khatter, and A. K. Gupta	2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)	2019		Pentingnya mendeteksi berita palsu di laman berita	[46]	SVM	*Aggregated news*	SVM memeroleh hasil akurasi sebesar 93.6%
Sentiment Classification Twitter of LRT, MRT, and Transjakarta Transportation using Support Vector Machine	Dinar Ajeng Kristiyanti, Rizki Aulianita; Dwi Andini Putri, Lilyani Asri Utami, Fajar Agustini, Zulia Imami Alfianti	2022 International Conference of Science and Information Technology in Smart Administration (ICSINTESA)	2022		Pelayanan yang diberikan oleh penyedia jasa transportasi MRT, LRT dan Transjakarta berbeda-beda, seperti tanggapan positif dan negatiF	[47]	Classification using SVM	*Tweet* dari Twitter	Metode Support Vector Machine mampu mengklasifikasikan teks sentimen positif dan negatif dengan hasil akurasi sebesar 91.89%

Berbagai penelitian di atas memiliki fokus yang berbeda-beda dalam menangani isu-isu terkait deteksi spam, berita palsu, dan konten tidak diinginkan di media sosial. Penelitian Monika Singh, Divya Bansal, and Sanjeev Sofat lebih menitikberatkan pada penyebaran konten pornografi dan promosi pasar pengikut aktif di Twitter [10], sementara penelitian Arunadevi, J., & Sowjanya, A. M. lebih terfokus pada pendeteksian berita palsu melalui headline berita dengan membandingkan algoritma klustering dan klasifikasi [8]. Kasra Majbouri Yazdi, et. al, mengadopsi kombinasi K-means dan SVM untuk mendeteksi berita palsu dan spam di Twitter [48]. Perbedaan lainnya terlihat pada penelitian yang menggunakan PCA dan K-means clustering oleh Samuel Ady Sanjaya and Kridanto Surendro untuk mendeteksi spam dengan mengklasifikasikan cluster menjadi 5 kategori [39], sedangkan penelitian lain seperti *Classification of Fake News on Facebook* menggunakan K-Means *Clustering* dan PCA untuk meningkatkan akurasi deteksi berita palsu di jejaring sosial baru [40]. Di sisi lain, penelitian oleh Sanjay Kumar Sonbhadra, Sonali Agarwal, P. Nagabhushan, lebih mengeksplorasi penggunaan data penelitian terbuka COVID-19 (CORD-19) dengan fokus pada klasifikasi menggunakan SVM satu kelas paralel [41]. Sementara itu, penelitian Jasmine Shaikh dan Rupati Patil serta Nihel Fatima Baarir dan Abdelhamid Djeffal lebih mengutamakan deteksi berita palsu melalui teknik klasifikasi dengan menggunakan berbagai algoritma, seperti SVM, Naïve Bayes, dan Passive Aggressive Classifier [42]. Selain itu, penelitian oleh Z Khanam, B N Alwasel, H Sirafi, dan M Rashid menambahkan analisis tekstual POS sebagai pendekatan kuantitatif dalam deteksi berita palsu [44]. Terakhir, penelitian oleh Murari Choudhary; Shashank Jha; Prashant; Deepika Saxena membahas implementasi berbagai algoritma pembelajaran mesin, termasuk Naïve Bayes, CNN, LSTM, dan Support Vector Machine, untuk mengidentifikasi serta mengurangi penyebaran berita palsu di berbagai platform media sosial [45]. Meskipun memiliki tujuan yang serupa, setiap penelitian memberikan kontribusi uniknya dalam pengembangan teknik dan metode deteksi terkait isu-isu tersebut.

Penelitian yang dilakukan sekarang mencirikan sebuah inovasi signifikan dalam konteks penelitian deteksi berita palsu, memperkenalkan pendekatan baru yang memanfaatkan serangkaian teknik untuk meningkatkan ketepatan klasifikasi. Berbeda dengan penelitian sebelumnya, studi ini mengusung konsep kebaruan dengan mengintegrasikan metode *Principal Component Analysis* (PCA) dan K-means *clustering* untuk mengklasterisasi data media sosial terkait pemilihan presiden menjadi dua kelompok, yaitu *suspicious* dan non-*suspicious*. Poin penting penelitian ini terletak pada pengadopsian metode *Support Vector Machine* (SVM) sebagai alat utama untuk mengklasifikasikan data yang sudah diklasterisasi tersebut. Lebih lanjut, penelitian ini melibatkan langkah-langkah kritis dalam mengoptimalkan kinerja SVM dengan melakukan penyesuaian hyperparameter. Optimalisasi tersebut tidak hanya meningkatkan tingkat akurasi secara drastis, tetapi juga menggambarkan keunggulan potensial SVM dalam menangani tugas klasifikasi yang kompleks pada data hasil klasterisasi yang dilakukan sebelumnya. Hasil ini menandai sebuah langkah maju dalam upaya mendeteksi dan mengidentifikasi akun mencurigakan yang terlibat dalam penyebaran berita palsu melalui media sosial X.