

BAB II

LANDASAN TEORI

2.1 Penelitian Terdahulu

Peninjauan terhadap literatur ilmiah sebelumnya dilakukan guna memperkuat landasan teori dan pemilihan metodologi dalam mendeteksi serta mengoreksi kesalahan penulisan klitik. Kajian ini berfokus pada penelitian-penelitian terdahulu yang menerapkan kerangka kerja CRISP-DM, algoritma klasifikasi data teks, hingga implementasi framework Flask. Rangkuman mengenai landasan riset terdahulu tersebut secara terstruktur dapat dilihat pada tabel berikut.

Table 2.1 Penelitian Terdahulu

No.	Judul dan Nama Peneliti Artikel	Metode	Hasil Penelitian
1.	Enhancing Sentiment Analysis Accuracy Through Intelligent Spelling Correction Using Damerau-Levenshtein Distance and N-Gram with Random Forest Classifier[8]	<i>Random Forest Classifier, Damerau-Levenshtein Distance, dan N-Gram</i>	a) Penggunaan algoritma Random Forest efektif untuk klasifikasi teks. b) Kombinasi Damerau-Levenshtein Distance dan N-Gram mampu membantu proses deteksi kesalahan penulisan. c) Integrasi metode NLP dapat meningkatkan akurasi sistem deteksi kesalahan bahasa.
2.	Pengklasifikasian Dokumen Teks Bahasa Indonesia berbasis Vector Space Model dengan menggunakan Metode k-Nearest Neighbor (k-NN) dan Euclidean Distance[10]	Metodologi <i>CRISP-DM, Vector Space Model, dan k-NN</i>	a) Metode CRISP-DM memberikan alur penelitian yang sistematis dalam pengembangan sistem berbasis data mining. b) Tahapan preprocessing dan klasifikasi teks sangat penting dalam pengolahan data bahasa Indonesia. c) Algoritma k-NN dan Vector Space Model dapat digunakan dalam proses klasifikasi dokumen teks.
3.	Deteksi Berita Palsu menggunakan Algoritma Random Forest[9]	<i>Random Forest, TF-IDF, dan Confusion Matrix Evaluation</i>	a) Algoritma Random Forest mampu melakukan klasifikasi teks dengan performa yang baik. b) Penggunaan TF-IDF efektif untuk ekstraksi fitur teks. c) Evaluasi menggunakan Confusion Matrix membantu mengukur performa model klasifikasi.
4.	Aplikasi Pendeteksi Kesalahan Ejaan Bahasa Indonesia pada Karya Ilmiah Bidang Ilmu Komputer Menggunakan KEBI 1.0 Checker[7]	<i>Natural Language Processing (NLP) dan Rule-Based Tokenization</i>	a) Pendekatan Natural Language Processing (NLP) dan Rule-Based efektif digunakan untuk deteksi kesalahan bahasa.

			b) Sistem berbasis web dapat digunakan untuk proses pengecekan bahasa secara otomatis. c) Pendekatan tokenisasi membantu proses analisis teks bahasa Indonesia.
5.	Penggunaan Random Forest dalam Sistem Klasifikasi Kecemasan pada Generasi Z[11]	<i>Random Forest</i> dan <i>Text Preprocessing</i> Baku (Cleansing, Case Folding, Normalization, Tokenizing)	a) Tahapan text preprocessing sangat berpengaruh terhadap hasil klasifikasi machine learning. b) Proses seperti cleansing, case folding, normalization, dan tokenizing penting dalam pengolahan teks bahasa Indonesia. c) Algoritma Random Forest dapat digunakan dalam proses klasifikasi teks setelah ekstraksi fitur dilakukan.
6.	Analisis Deteksi Kesalahan Penulisan (Typo) Pada Tugas Mahasiswa Menggunakan Algoritma Levenshtein Distance[12]	<i>Rule-Based Approach</i> menggunakan algoritma <i>Levenshtein Distance</i> dan pencocokan berbasis kamus kata baku.	a) Algoritma Levenshtein Distance terbukti efektif dalam mengukur jarak kemiripan antar-string pada teks bahasa Indonesia. b) Sistem mampu mengidentifikasi kesalahan ketik (typo) secara mendetail melalui perhitungan operasi karakter yang meliputi penyisipan (insertion), penghapusan (deletion), dan penggantian (substitution). c) Penggunaan algoritma jarak teks ini memberikan fondasi yang kuat untuk penyaringan kesalahan mekanis pada tingkat kata sebelum teks diproses lebih lanjut.
7.	Klasifikasi Berita Hoaks Menggunakan Metode Random Forest Berbasis Tf-Idf[13]	Kombinasi pembobotan kata menggunakan metode <i>TF-IDF</i> dan algoritma <i>Random Forest Classifier</i> .	a) Penggabungan ekstraksi fitur TF-IDF dengan algoritma Random Forest terbukti menghasilkan kinerja klasifikasi yang sangat baik pada dataset teks berbahasa Indonesia. b) Model yang dibangun menunjukkan tingkat stabilitas yang tinggi dan akurasi yang optimal saat mengolah struktur penulisan formal pada dokumen teks berita. c) Pengujian membuktikan bahwa karakteristik pohon keputusan (decision trees) pada Random Forest mampu memproses korpus berita yang dinamis dengan risiko overfitting yang rendah.
8.	Deteksi Kesalahan Ejaan dan Penentuan Rekomendasi Koreksi Kata yang Tepat Pada Dokumen Jurnal JTIK Menggunakan Dictionary Lookup dan Damerau-Levenshtein Distance[14]	<i>Rule-Based Approach</i> menggunakan teknik parsing morfologi dan pencocokan pola kata berbasis korpus KBBI.	a) Metode rule-based terbukti sangat efektif untuk mendeteksi kesalahan ejaan yang memiliki pola struktural baku dan kaku, seperti kekeliruan pemisahan kata depan (preposition) dan peletakan imbuhan (affixation). b) Sistem mampu memetakan variasi string kesalahan penulisan tanpa memerlukan proses komputasi yang berat (low computational cost).

			c) Penelitian ini menegaskan bahwa pendekatan berbasis aturan sangat ideal ditempatkan sebagai first-line filter (penyaring tahap awal) untuk membersihkan kesalahan struktural dasar sebelum teks masuk ke pemrosesan kecerdasan buatan yang lebih kompleks.
--	--	--	---

Berdasarkan hasil kajian terhadap penelitian terdahulu, dapat diketahui bahwa berbagai metode telah diterapkan untuk menyelesaikan permasalahan klasifikasi teks maupun deteksi kesalahan bahasa Indonesia. Penelitian oleh Doni Abdul Fatah dkk.[8] menunjukkan bahwa kombinasi algoritma Random Forest, Damerau-Levenshtein Distance, dan N-Gram mampu meningkatkan akurasi klasifikasi serta deteksi kesalahan penulisan. Penelitian tersebut menjadi referensi dalam pemilihan algoritma Random Forest sebagai model klasifikasi pada penelitian ini. Namun, penelitian tersebut lebih berfokus pada peningkatan performa algoritma dan belum mengimplementasikannya ke dalam sistem penapisan berbasis web yang dapat digunakan oleh pengguna secara langsung.

Penelitian Setiawan dkk.[10] mengadopsi metodologi CRISP-DM sebagai kerangka kerja dalam pengembangan sistem klasifikasi dokumen teks bahasa Indonesia. Kedua penelitian tersebut menunjukkan bahwa CRISP-DM mampu memberikan alur pengembangan yang sistematis mulai dari tahap business understanding hingga deployment. Meskipun demikian, fokus penelitian masih terbatas pada proses klasifikasi dokumen dan belum diarahkan pada penyelesaian permasalahan kesalahan penulisan klitik maupun implementasi dalam proses editorial berita.

Penelitian Rahmat Dipo Setyadin dkk.[9] membuktikan bahwa algoritma Random Forest dengan ekstraksi fitur TF-IDF mampu menghasilkan performa klasifikasi teks yang baik. Sementara itu, penelitian Raina Rahmawati Fitri dkk.[11] juga menunjukkan pentingnya tahapan text preprocessing dalam meningkatkan kualitas klasifikasi menggunakan Random Forest. Kedua penelitian tersebut menjadi dasar dalam pemilihan proses preprocessing, ekstraksi fitur TF-IDF, dan penggunaan Random Forest pada penelitian ini. Akan tetapi, kedua penelitian masih berorientasi

pada tugas klasifikasi teks dan belum mengembangkan sistem yang mampu mendeteksi kesalahan penulisan klitik secara spesifik.

Di sisi lain, penelitian Tresna Maulana Fahrudin dkk.[7] dan Analisis Deteksi Kesalahan Penulisan (Typo) Menggunakan Algoritma Levenshtein Distance[12] menunjukkan bahwa pendekatan Natural Language Processing (NLP), rule-based, serta algoritma Levenshtein Distance efektif digunakan dalam mendeteksi kesalahan penulisan. Namun, pendekatan tersebut lebih difokuskan pada pendeteksian kesalahan ejaan atau typo secara umum, sehingga belum secara khusus menangani kesalahan penulisan klitik yang memiliki pola linguistik tertentu sesuai kaidah bahasa Indonesia.

Selain itu, penelitian mengenai klasifikasi berita hoaks menggunakan TF-IDF dan Random Forest[13] memperlihatkan bahwa kombinasi ekstraksi fitur TF-IDF dengan algoritma Random Forest mampu menghasilkan model klasifikasi yang stabil dan memiliki risiko overfitting yang rendah. Penelitian tersebut memperkuat alasan penggunaan Random Forest dalam penelitian ini. Meskipun demikian, objek penelitian masih berupa klasifikasi berita hoaks sehingga belum membahas deteksi kesalahan kebahasaan maupun integrasi dengan sistem berbasis web.

Berdasarkan hasil analisis penelitian terdahulu tersebut, terdapat beberapa celah penelitian (research gap), yaitu belum adanya penelitian yang secara khusus menggabungkan pendekatan rule-based untuk mendeteksi pola kesalahan penulisan klitik dengan algoritma Random Forest sebagai model klasifikasi dalam sebuah sistem penapisan teks berita berbasis web. Selain itu, sebagian besar penelitian terdahulu masih berfokus pada pengembangan model klasifikasi atau deteksi kesalahan bahasa secara terpisah tanpa mengimplementasikan alur kerja yang mendukung proses editorial. Oleh karena itu, penelitian ini berkontribusi dengan mengembangkan sistem berbasis web yang mengintegrasikan pendekatan rule-based, ekstraksi fitur TF-IDF, algoritma Random Forest, serta konsep multi-user yang melibatkan reporter, editor, dan administrator sehingga dapat digunakan sebagai alat bantu penapisan kesalahan penulisan klitik pada teks berita berbahasa Indonesia.

2.2 Teori yang berkaitan

2.2.1 Klitik dalam Bahasa Indonesia

Klitik merupakan bentuk morfem terikat yang secara fonologis tidak dapat berdiri sendiri dan harus melekat pada kata lain[15]. Dalam bahasa Indonesia, klitik yang umum digunakan antara lain “-ku”, “-mu”, dan “-nya”. Klitik biasanya digunakan untuk menunjukkan kepemilikan, penegasan, atau hubungan tertentu dalam sebuah kalimat. Penulisan klitik harus mengikuti kaidah bahasa Indonesia yang berlaku agar struktur kata tetap sesuai dengan aturan kebahasaan[16].

Dalam penulisan bahasa Indonesia, klitik ditulis melekat dengan kata yang mengikutinya tanpa menggunakan spasi[17]. Sebagai contoh, penulisan yang benar adalah “bukuku”, “rumahmu”, dan “mobilnya”. Namun dalam praktiknya, masih banyak ditemukan kesalahan penulisan seperti “buku ku”, “rumah mu”, atau “mobil nya”. Kesalahan tersebut dapat memengaruhi kualitas penulisan dan menunjukkan ketidaksesuaian dengan Pedoman Umum Ejaan Bahasa Indonesia (PUEBI).

Kesalahan penggunaan klitik sering ditemukan dalam penulisan informal maupun formal, termasuk pada teks berita[18]. Pada dunia jurnalistik, ketepatan penggunaan bahasa menjadi hal penting karena berita yang dipublikasikan harus memiliki kualitas bahasa yang baik dan mudah dipahami pembaca. Oleh karena itu, diperlukan mekanisme atau sistem yang mampu membantu mendeteksi kesalahan penggunaan klitik secara otomatis sehingga proses penapisan bahasa dapat dilakukan dengan lebih efektif dan efisien.

2.2.2 NLP

Natural Language Processing (NLP) merupakan salah satu subbidang dari Artificial Intelligence (AI) yang berfokus pada pengolahan dan pemahaman bahasa manusia oleh komputer[19]. NLP memungkinkan sistem komputer untuk mengenali, memproses, menganalisis, serta memahami bahasa yang digunakan manusia, baik yang disajikan dalam bentuk tulisan maupun tuturan[20]. Teknologi NLP banyak digunakan dalam berbagai bidang, di antaranya adalah agen percakapan virtual, layanan penerjemahan otomatis, dan pengolahan sentimen, sistem rekomendasi, hingga deteksi kesalahan bahasa.

Pada dasarnya, NLP bekerja dengan mengubah bahasa manusia menjadi bentuk data yang dapat dipahami oleh komputer. Dalam prosesnya, sistem akan melakukan analisis terhadap struktur teks, pola kata, serta hubungan antar kata untuk memperoleh informasi tertentu[21]. Dengan menggunakan NLP, komputer dapat mengenali pola penulisan, memahami konteks sederhana, serta melakukan klasifikasi terhadap teks berdasarkan data yang telah dipelajari sebelumnya.

Dalam penelitian ini, NLP digunakan untuk memproses teks berita dan mendeteksi pola kesalahan penggunaan klitik dalam bahasa Indonesia. Proses NLP yang digunakan meliputi preprocessing teks, tokenisasi, ekstraksi fitur, dan klasifikasi teks. Tahapan tersebut bertujuan guna menyiapkan data sebelum diolah menggunakan algoritma machine learning sehingga sistem mampu mengenali pola penulisan klitik yang benar maupun salah secara otomatis.

2.2.3 Machine Learning

Machine Learning merupakan salah satu subbidang di bidang kecerdasan buatan yang memungkinkan sistem komputer mempelajari pola dari sekumpulan data sehingga dapat menghasilkan prediksi maupun menentukan keputusan secara otomatis tanpa harus diprogram secara khusus untuk setiap kondisi[22]. Machine Learning bekerja dengan memanfaatkan data sebagai bahan pembelajaran sehingga sistem dapat mengenali pola tertentu dan menghasilkan prediksi berdasarkan pola tersebut[23]. Penerapan teknologi ini mencakup berbagai bidang, antara lain klasifikasi teks, analisis sentimen, pengenalan gambar, hingga sistem rekomendasi.

Dalam proses machine learning, model akan dilatih menggunakan kumpulan data yang telah dipersiapkan terlebih dahulu. Data tersebut biasanya terdiri dari data input dan label atau kategori tertentu[24]. Selama proses pelatihan, model akan mempelajari keterkaitan antara data dan label sehingga dapat mengenali karakteristik dari masing-masing kategori. Model yang telah menjalani tahap pelatihan selanjutnya dapat dimanfaatkan untuk menghasilkan prediksi pada data baru yang sebelumnya tidak pernah dianalisis oleh sistem[24].

Dalam penelitian ini, machine learning digunakan untuk membantu proses klasifikasi teks yang mengandung kesalahan klitik pada bahasa Indonesia. Model

dilatih menggunakan dataset teks yang telah diberi label benar dan salah berdasarkan pola penulisan klitik. Setelah model selesai dilatih, sistem dapat melakukan prediksi terhadap teks berita baru untuk membantu mendeteksi kemungkinan kesalahan penggunaan klitik secara otomatis.

2.2.4 Sistem Berbasis Web

Sistem berbasis web adalah aplikasi yang dapat dijalankan dan diakses melalui browser menggunakan jaringan internet maupun jaringan lokal tanpa perlu melakukan pemasangan perangkat lunak tambahan pada perangkat pengguna[25]. Sistem ini menggunakan konsep client-server, yang mana browser berfungsi sebagai client untuk menampilkan antarmuka pengguna, sedangkan proses pengolahan data dan logika sistem dijalankan pada server[26]. Penggunaan sistem berbasis web saat ini banyak diterapkan karena lebih fleksibel, mudah dikembangkan, dan dapat diakses melalui berbagai perangkat, termasuk komputer dan laptop.

Dibandingkan aplikasi desktop, sistem berbasis web menawarkan berbagai keunggulan, antara lain kemudahan dalam akses, proses pembaruan sistem yang lebih sederhana, serta kemampuan mendukung banyak pengguna secara bersamaan (multi-user)[27]. Selain itu, sistem berbasis web juga mempermudah proses penyimpanan dan pengelolaan data secara terpusat melalui database. Dengan adanya sistem berbasis web, pengguna dapat melakukan aktivitas secara lebih efisien tanpa dibatasi oleh perangkat tertentu[27].

Dalam penelitian ini, sistem berbasis web dimanfaatkan untuk menunjang kegiatan penapisan berita serta pengelolaan data pengguna. Sistem dirancang untuk mendukung alur kerja antara reporter dan editor dalam proses penapisan kesalahan klitik pada teks berita. Reporter bertugas menginput berita ke dalam sistem, sedangkan editor melakukan proses penapisan dan melihat hasil deteksi kesalahan. Melalui penerapan sistem berbasis web, proses penapisan dapat dilakukan secara lebih terstruktur, fleksibel, dan mudah diakses oleh beberapa pengguna sesuai hak akses masing-masing.

2.3 Framework/Algoritma yang digunakan

2.3.1 CRISP-DM

Cross-Industry Standard Process for Data Mining (CRISP-DM) merupakan salah satu kerangka kerja yang diterapkan dalam proses data mining dan machine learning[28]. Pendekatan ini menyediakan tahapan kerja yang sistematis dan terstruktur mulai dari pemahaman masalah hingga implementasi model. CRISP-DM banyak digunakan karena bersifat fleksibel dan dapat diterapkan pada berbagai jenis penelitian berbasis data[29].

CRISP-DM memiliki enam tahapan utama yang terdiri dari Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, dan Deployment yang digunakan sebagai alur sistematis dalam proses pengembangan data mining dan machine learning.

1) Business Understanding

Business Understanding adalah tahapan awal untuk memahami permasalahan penelitian dan menetapkan tujuan yang hendak dicapai. Pada tahap ini dilakukan identifikasi kebutuhan sistem dan analisis permasalahan yang akan diselesaikan[29].

2) Data Understanding

Data Understanding merupakan tahap pengumpulan serta analisis data yang akan dimanfaatkan dalam penelitian. Tahap ini bertujuan untuk memahami karakteristik data, pola data, serta kualitas data yang tersedia[30].

3) Data Preparation

Data Preparation merupakan tahap persiapan data sebelum dimanfaatkan pada proses modeling. Tahapan ini meliputi pembersihan data, preprocessing, transformasi data, serta pemilihan fitur yang relevan[31].

4) Modeling

Modeling merupakan tahap pembangunan model atau algoritma berdasarkan data yang telah dipersiapkan sebelumnya. Pada tahap ini dilakukan pelatihan model dengan menerapkan metode atau algoritma tertentu[32].

5) Evaluation

Evaluation merupakan tahap pengujian dan evaluasi model yang telah dibuat. Tahapan ini dilakukan untuk mengetahui performa model dan memastikan hasil yang diperoleh sesuai dengan tujuan penelitian[33].

6) Deployment

Deployment merupakan tahap implementasi model agar dapat digunakan dalam sistem atau lingkungan nyata. Pada tahap ini, model yang sudah selesai dikembangkan dapat digunakan oleh pengguna sesuai kebutuhan penelitian atau aplikasi tertentu[33].

2.3.2 Pendekatan Rule-Based

Pendekatan rule-based adalah metode yang beroperasi berdasarkan aturan atau pola tertentu untuk mendeteksi suatu kondisi pada data. Metode ini sering diterapkan pada sistem pengolahan teks karena mampu mengenali pola yang memiliki struktur tetap dan jelas[34]. Pada pendekatan rule-based, sistem tidak mempelajari data secara otomatis seperti machine learning, melainkan menggunakan aturan yang telah ditentukan sebelumnya oleh pengembang sistem. Aturan tersebut biasanya dibuat berdasarkan kaidah bahasa, pola penulisan, atau karakteristik tertentu yang ingin dideteksi[35]. Dengan demikian, keberhasilan metode ini sangat bergantung pada ketepatan aturan yang dirancang sesuai dengan karakteristik permasalahan yang akan diselesaikan.

Dalam penelitian ini, pendekatan rule-based digunakan untuk mendeteksi kesalahan penulisan klitik berdasarkan pola spasi yang tidak sesuai dengan kaidah bahasa Indonesia. Aturan (rules) yang digunakan disusun berdasarkan Pedoman Umum Ejaan Bahasa Indonesia (PUEBI) yang menyatakan bahwa klitik "-ku", "-

mu", dan "-nya" sebagai pronomina enklitik harus ditulis serangkai dengan kata yang mendahuluinya[36]. Berdasarkan kaidah tersebut, sistem menyusun aturan untuk mengidentifikasi penulisan klitik yang masih dipisahkan oleh spasi. Aturan dibuat untuk mencari pola penulisan seperti “kata + spasi + ku”, “kata + spasi + mu”, dan “kata + spasi + nya”. Jika sistem menemukan pola tersebut, maka bagian teks akan ditandai sebagai kesalahan penulisan klitik. Contoh kesalahan yang dapat dideteksi antara lain “buku ku”, “rumah mu”, dan “mobil nya” yang seharusnya ditulis menjadi “bukuku”, “rumahmu”, dan “mobilnya”. Pada implementasinya, aturan tersebut diterapkan menggunakan Regular Expression (Regex) untuk melakukan proses pattern matching terhadap teks masukan. Apabila sistem menemukan pola yang sesuai dengan aturan yang telah ditetapkan, maka bagian teks tersebut akan diberi tanda (highlight) sebagai indikasi adanya kesalahan penulisan klitik[37].

Pendekatan rule-based dipilih karena kesalahan klitik memiliki pola yang cukup spesifik dan mudah dikenali menggunakan aturan sederhana. Selain itu, metode ini memiliki proses deteksi yang cepat dan mudah diimplementasikan pada sistem berbasis web. Dengan menggunakan aturan yang terstruktur, sistem dapat membantu proses penapisan bahasa secara otomatis sehingga editor lebih mudah menemukan kesalahan penulisan klitik pada teks berita.

Table 2.2 Aturan Rule-Based

No	Aturan	Contoh Salah	Contoh Benar
1	kata + spasi + ku	buku ku	bukuku
2	kata + spasi + mu	rumah mu	rumahmu
3	kata + spasi + nya	mobil nya	mobilnya

2.3.3 Machine Learning Classification

Machine Learning Classification adalah salah satu metode pada machine learning yang digunakan untuk mengklasifikasikan data ke kategori tertentu berdasarkan pola yang telah dipelajari dari data sebelumnya[38]. Dalam klasifikasi teks, sistem akan mempelajari karakteristik dari data latih untuk menentukan apakah suatu teks termasuk ke dalam kategori tertentu. Metode klasifikasi banyak digunakan dalam

bidang Natural Language Processing (NLP), seperti analisis sentimen, spam detection, klasifikasi dokumen, dan deteksi kesalahan bahasa[39].

Dalam penelitian ini, metode machine learning classification digunakan untuk membantu proses identifikasi teks yang mengandung kesalahan penulisan klitik. Model machine learning dilatih dengan dataset yang memuat contoh penulisan benar dan salah. Sebelum pelatihan model dilakukan, dataset terlebih dahulu melewati tahap preprocessing untuk membersihkan dan menyesuaikan format teks agar lebih mudah diproses oleh sistem. Selanjutnya dilakukan tahap ekstraksi fitur dengan metode TF-IDF (Term Frequency–Inverse Document Frequency) untuk mengonversi data teks ke dalam bentuk numerik agar dapat diproses dan dipahami oleh model machine learning.

Tahapan berikutnya adalah proses pelatihan model dengan memanfaatkan data yang telah diekstraksi menjadi fitur numerik. Model kemudian mempelajari pola penulisan yang terdapat pada dataset sehingga dapat membedakan antara teks yang benar dan salah. Setelah pelatihan selesai dilakukan, model dapat dimanfaatkan untuk menghasilkan prediksi pada teks baru yang dimasukkan ke dalam sistem. Dengan adanya pendekatan machine learning classification, sistem diharapkan mampu membantu proses deteksi kesalahan klitik secara lebih otomatis dan konsisten.

2.3.4 Random Forest

Random Forest merupakan salah satu algoritma machine learning yang termasuk dalam metode ensemble learning. Algoritma ini bekerja dengan membangun sejumlah decision tree berdasarkan sampel data yang dipilih secara acak (bootstrap sampling), kemudian menghasilkan keputusan akhir melalui mekanisme majority voting pada kasus klasifikasi atau nilai rata-rata (averaging) pada kasus regresi[40]. Pendekatan tersebut memungkinkan Random Forest menghasilkan model yang lebih stabil dan memiliki tingkat akurasi yang lebih baik dibandingkan penggunaan satu decision tree tunggal.

Pada proses pembentukan model, setiap decision tree dibangun menggunakan kombinasi data pelatihan dan subset fitur yang dipilih secara acak. Strategi ini

bertujuan untuk mengurangi korelasi antar pohon sehingga meningkatkan kemampuan generalisasi model terhadap data baru[41]. Selain itu, penggunaan banyak pohon keputusan mampu meminimalkan risiko overfitting, yaitu kondisi ketika model memiliki performa yang sangat baik pada data pelatihan tetapi kurang mampu melakukan prediksi terhadap data yang belum pernah dilihat sebelumnya.

Dalam penelitian ini, Random Forest digunakan sebagai model klasifikasi untuk membantu proses identifikasi kesalahan penulisan klitik pada teks berita berbahasa Indonesia. Sebelum proses klasifikasi dilakukan, data teks terlebih dahulu melalui tahapan preprocessing dan ekstraksi fitur menggunakan metode TF-IDF sehingga setiap dokumen direpresentasikan dalam bentuk vektor numerik[42]. Selanjutnya, model Random Forest mempelajari pola dari data pelatihan untuk mengklasifikasikan teks berdasarkan kategori yang telah ditentukan. Pemilihan algoritma ini didasarkan pada kemampuannya dalam menangani data berdimensi tinggi, menghasilkan performa klasifikasi yang baik pada data teks, serta memiliki ketahanan terhadap overfitting. Beberapa penelitian terdahulu juga menunjukkan bahwa kombinasi TF-IDF dan Random Forest mampu memberikan hasil klasifikasi yang baik pada berbagai permasalahan pengolahan teks berbahasa Indonesia.

2.3.5 Flask Framework

Flask merupakan framework berbasis Python yang dimanfaatkan untuk mengembangkan aplikasi web secara sederhana, ringan, dan fleksibel[43]. Flask termasuk jenis microframework karena hanya menyediakan komponen inti untuk pengembangan web tanpa fitur tambahan yang terlalu kompleks. Dengan sifatnya yang ringan, Flask banyak digunakan dalam pengembangan aplikasi yang menerapkan machine learning dan NLP karena dapat diintegrasikan dengan berbagai library Python seperti Scikit-learn, Pandas, dan NumPy[43].

Flask bekerja dengan konsep routing yang menghubungkan URL dengan fungsi tertentu pada program. Selain itu, Flask juga mendukung pengelolaan template HTML, pemrosesan form input, penggunaan session login, serta integrasi database[44]. Framework ini memungkinkan pengembang membuat sistem berbasis web dengan struktur yang lebih sederhana namun tetap mampu

mendukung berbagai kebutuhan aplikasi. Karena menggunakan bahasa Python, Flask menjadi pilihan yang sesuai untuk penelitian yang melibatkan proses pengolahan data dan machine learning.

Dalam penelitian ini, Flask digunakan sebagai framework utama dalam pembangunan website penapisan kesalahan klitik berbasis web. Flask digunakan untuk mengelola routing halaman, memproses input pengguna, menghubungkan sistem dengan database SQLite, serta menampilkan hasil penapisan pada halaman website. Selain itu, Flask juga digunakan untuk mendukung sistem multi-user dengan pembagian hak akses antara reporter, editor, dan admin sehingga alur kerja penapisan berita dapat berjalan secara terstruktur.

2.3.6 SQLite Database

SQLite merupakan salah satu sistem pengelolaan basis data relasional (Relational Database Management System/RDBMS) yang memiliki karakteristik ringan, sederhana, dan menggunakan penyimpanan berbasis file[45]. Berbeda dengan database server seperti MySQL atau PostgreSQL, SQLite tidak memerlukan instalasi server terpisah karena seluruh data disimpan langsung dalam satu file database. SQLite banyak digunakan dalam pengembangan aplikasi skala kecil hingga menengah karena memiliki proses konfigurasi yang sederhana, penggunaan memori yang rendah, serta mudah diintegrasikan dengan berbagai bahasa pemrograman termasuk Python[46].

SQLite dipilih dalam penelitian ini karena sesuai untuk kebutuhan pengembangan sistem berbasis web yang masih berada pada tahap penelitian dan pengujian. Selain mudah digunakan, SQLite juga mampu mendukung proses penyimpanan dan pengelolaan data secara terstruktur. Dengan menggunakan SQLite, sistem dapat melakukan operasi database meliputi penambahan, pembaruan, penghapusan, dan pengambilan data secara lebih efisien. Integrasi SQLite dengan Flask juga dapat dilakukan secara langsung menggunakan library bawaan Python, yaitu sqlite3.

Dalam penelitian ini, database SQLite digunakan untuk menyimpan berbagai data yang dibutuhkan oleh sistem website. Data tersebut meliputi akun pengguna seperti reporter, editor, dan admin, data berita yang diinput oleh reporter, status proses

penapisan berita, serta hasil deteksi kesalahan klitik yang dilakukan oleh sistem. Penggunaan database memungkinkan seluruh data tersimpan secara terpusat sehingga proses pengelolaan informasi pada website dapat dilakukan dengan lebih mudah dan terorganisir.

2.3.7 Metrik Evaluasi Model

Evaluasi model merupakan tahapan yang dilakukan untuk mengetahui tingkat kinerja algoritma machine learning dalam melakukan proses klasifikasi[47]. Pada penelitian ini, performa model Random Forest dievaluasi menggunakan Confusion Matrix beserta metrik Accuracy, Precision, Recall, dan F1-Score[48]. Penggunaan beberapa metrik tersebut bertujuan untuk memberikan gambaran yang lebih komprehensif mengenai kemampuan model dalam mengklasifikasikan data secara benar.

1) Confusion Matrix

Confusion Matrix merupakan tabel yang digunakan untuk membandingkan hasil prediksi model dengan label sebenarnya[49]. Melalui Confusion Matrix dapat diketahui jumlah prediksi yang benar maupun prediksi yang salah sehingga dapat digunakan sebagai dasar perhitungan berbagai metrik evaluasi[49].

Confusion Matrix terdiri atas empat komponen utama, yaitu:

- a) True Positive (TP), yaitu jumlah data positif yang berhasil diprediksi sebagai positif.
- b) True Negative (TN), yaitu jumlah data negatif yang berhasil diprediksi sebagai negatif.
- c) False Positive (FP), yaitu jumlah data negatif yang diprediksi sebagai positif.
- d) False Negative (FN), yaitu jumlah data positif yang diprediksi sebagai negatif.

2) Accuracy

Accuracy merupakan metrik yang menunjukkan persentase keseluruhan prediksi yang dilakukan dengan benar oleh model dibandingkan dengan

seluruh data yang diuji. Semakin tinggi nilai accuracy, semakin baik kemampuan model dalam melakukan klasifikasi[50].

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

Rumus 2.1 Accuracy

3) Precision

Precision digunakan untuk mengukur tingkat ketepatan model dalam memberikan prediksi terhadap kelas positif. Nilai precision menunjukkan seberapa banyak prediksi positif yang benar dibandingkan dengan seluruh prediksi positif yang dihasilkan model[51].

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

Rumus 2.2 Precision

Semakin tinggi nilai precision, semakin kecil kemungkinan model menghasilkan prediksi positif yang salah[52].

4) Recall

Recall menunjukkan kemampuan model dalam menemukan seluruh data yang benar-benar termasuk ke dalam kelas positif. Metrik ini menggambarkan seberapa banyak data positif yang berhasil dikenali oleh model[53].

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

Rumus 2.3 Recall

Nilai recall yang tinggi menunjukkan bahwa model mampu mendeteksi sebagian besar data positif yang tersedia[51].

5) F1-Score

F1-Score merupakan rata-rata harmonik antara nilai precision dan recall. Metrik ini digunakan ketika kedua nilai tersebut sama-sama penting dalam proses evaluasi model[51].

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Rumus 2.4 F1-Score

Nilai F1-Score yang tinggi menunjukkan bahwa model memiliki keseimbangan yang baik antara ketepatan klasifikasi dan kemampuan mendeteksi seluruh data positif[54].

2.4 Tools/software yang digunakan

Dalam proses pengembangan sistem penapisan kesalahan klitik berbasis web ini, digunakan beberapa tools dan software yang memiliki fungsi berbeda sesuai kebutuhan pengembangan sistem, pengolahan data, implementasi algoritma, hingga pembuatan antarmuka website. Penggunaan berbagai tools tersebut bertujuan untuk mendukung proses penelitian agar berjalan lebih efektif, terstruktur, dan mudah dikembangkan.

2.4.1 Python

Python merupakan bahasa pemrograman yang menjadi dasar pengembangan dalam penelitian ini. Bahasa pemrograman ini dipilih karena memiliki sintaks yang sederhana, mudah dipelajari, serta mendukung pengembangan aplikasi berbasis machine learning dan NLP[55]. Dibandingkan beberapa bahasa pemrograman lain, Python memiliki struktur kode yang lebih ringkas yang dapat mendukung proses pengembangan serta pemeliharaan sistem. Selain itu, Python juga memiliki komunitas pengguna yang sangat besar sehingga dokumentasi, referensi, dan library pendukung dapat diperoleh dengan mudah[55].

Dalam bidang pengolahan data dan kecerdasan buatan, Python menjadi salah satu bahasa pemrograman yang banyak dimanfaatkan karena menyediakan beragam library yang lengkap serta mudah diintegrasikan dalam proses pengembangan sistem[56]. Pada penelitian ini, Python digunakan untuk membangun logika sistem, melakukan pemrosesan teks berita, menjalankan algoritma deteksi kesalahan klitik, serta menghubungkan model machine learning dengan aplikasi berbasis web. Python juga digunakan dalam proses preprocessing dataset, pelatihan model machine learning, penyimpanan model, hingga implementasi backend website menggunakan framework Flask.

Python memiliki ekosistem library yang sangat luas sehingga mendukung seluruh kebutuhan penelitian. Beberapa library yang dimanfaatkan dalam penelitian ini

meliputi Pandas untuk proses pengolahan dan manipulasi dataset, Scikit-learn untuk proses machine learning dan klasifikasi teks, Pickle untuk menyimpan model machine learning yang sudah melalui proses pelatihan, Flask untuk membangun aplikasi website berbasis Python, serta Regex (re) untuk proses pencocokan pola kesalahan klitik menggunakan regular expression. Selain itu, Python juga mendukung integrasi dengan database SQLite sehingga memudahkan proses penyimpanan data pengguna dan berita pada sistem website.

Penggunaan Python memberikan fleksibilitas yang tinggi karena seluruh proses penelitian dapat dilakukan menggunakan satu bahasa pemrograman yang sama. Peneliti tidak perlu menggunakan bahasa yang berbeda antara proses pengolahan data, machine learning, dan pengembangan website sehingga proses implementasi sistem menjadi lebih sederhana dan terintegrasi. Dengan dukungan library yang lengkap serta kemudahan penggunaan, Python menjadi pilihan yang sesuai untuk pengembangan sistem deteksi kesalahan klitik berbasis web pada penelitian ini.

2.4.2 Flask

Flask merupakan framework web berbasis Python yang digunakan untuk mengembangkan aplikasi website pada penelitian ini[57]. Flask termasuk jenis microframework, yaitu framework yang memiliki struktur ringan dan sederhana namun tetap mampu mendukung pengembangan aplikasi web secara lengkap[58]. Flask dipilih karena mudah digunakan, memiliki dokumentasi yang lengkap, serta dapat diintegrasikan dengan berbagai library Python, terutama library machine learning dan NLP.

Dalam penelitian ini, Flask digunakan sebagai backend website yang mengatur komunikasi antara pengguna, database, dan algoritma deteksi kesalahan klitik. Framework ini bertanggung jawab dalam menangani proses routing halaman, pengolahan input pengguna, autentikasi akun, pengelolaan data berita, hingga proses penapisan teks berita. Setiap permintaan dari pengguna melalui browser akan diproses oleh Flask, kemudian sistem akan memberikan respon berupa tampilan halaman atau hasil proses deteksi kesalahan klitik.

Selain mengelola backend sistem, Flask juga digunakan untuk menampilkan hasil deteksi kesalahan dalam bentuk highlight pada teks berita. Sistem akan mengirimkan hasil proses deteksi ke halaman website sehingga pengguna dapat melihat bagian teks yang terdeteksi sebagai kesalahan secara langsung. Dengan menggunakan Flask, aplikasi dapat dijalankan secara lokal melalui browser tanpa memerlukan server yang kompleks sehingga memudahkan proses pengembangan dan pengujian sistem selama penelitian berlangsung.

Framework Flask dipilih karena memiliki beberapa kelebihan yang sesuai dengan kebutuhan penelitian. Flask mudah dipelajari dan dikembangkan, kompatibel dengan Python serta library machine learning seperti Scikit-learn, memiliki proses konfigurasi yang sederhana, dan ringan untuk dijalankan pada perangkat lokal. Selain itu, Flask juga cocok digunakan dalam pengembangan sistem penelitian dan prototipe karena mendukung proses integrasi backend, database, dan antarmuka web secara lebih fleksibel dan efisien.

2.4.3 Visual Studio Code

Visual Studio Code merupakan code editor yang dimanfaatkan sebagai sarana utama untuk pengembangan sistem dalam penelitian ini[59]. Visual Studio Code dikembangkan oleh Microsoft serta menjadi salah satu code editor yang banyak digunakan karena memiliki antarmuka yang sederhana, ringan, dan mendukung beragam bahasa pemrograman. Selain itu, Visual Studio Code juga menyediakan banyak fitur dan ekstensi yang membantu proses pengembangan aplikasi berbasis web maupun machine learning[60].

Dalam penelitian ini, Visual Studio Code dimanfaatkan untuk mengembangkan dan mengelola seluruh kode program yang berkaitan dengan sistem penapisan kesalahan klitik berbasis web. Pengembangan backend Flask, template HTML, CSS Bootstrap, pengelolaan database SQLite, hingga implementasi algoritma deteksi dilakukan menggunakan aplikasi ini. Visual Studio Code juga digunakan untuk mengatur struktur folder project, mengelola file program, serta menjalankan terminal untuk mengeksekusi perintah Python dan Flask selama proses pengembangan sistem berlangsung.

Visual Studio Code memiliki berbagai fitur yang membantu meningkatkan efisiensi proses pemrograman. Beberapa fitur yang dimanfaatkan dalam penelitian ini meliputi syntax highlighting untuk mempermudah pembacaan kode program, auto completion untuk membantu penulisan sintaks secara otomatis, integrated terminal untuk menjalankan perintah langsung di dalam aplikasi, serta debugging tools untuk membantu proses pengecekan kesalahan program. Selain itu, penggunaan extension support seperti Python Extension dan Flask Extension juga membantu proses pengembangan backend website menjadi lebih mudah dan terstruktur.

Penggunaan Visual Studio Code memberikan kemudahan karena seluruh proses coding dapat dilakukan dalam satu aplikasi tanpa perlu berpindah software. Dengan tampilan yang ringan dan fitur yang lengkap, Visual Studio Code mendukung proses pengembangan sistem menjadi lebih cepat, rapi, dan efisien. Oleh karena itu, Visual Studio Code menjadi salah satu tools yang sesuai untuk mendukung pembangunan sistem website penapisan kesalahan klitik pada penelitian ini..

2.4.4 SQLite

SQLite merupakan perangkat lunak pengelola basis data relasional (Relational Database Management System/RDBMS) yang digunakan sebagai database lokal pada penelitian ini. SQLite dipilih karena memiliki struktur yang ringan, sederhana, dan tidak memerlukan server database terpisah seperti MySQL atau PostgreSQL[61]. Seluruh data pada SQLite disimpan dalam satu file database sehingga proses penggunaan dan pengelolaan database menjadi lebih praktis. SQLite banyak digunakan pada pengembangan aplikasi skala kecil hingga menengah, terutama pada tahap penelitian, prototipe, dan pengembangan sistem lokal[46].

Dalam penelitian ini, SQLite digunakan untuk menyimpan berbagai data yang dibutuhkan oleh sistem website penapisan kesalahan klitik. Data yang disimpan meliputi data akun pengguna seperti reporter, editor, dan admin, data berita yang diinput oleh reporter, status penapisan berita, serta hasil deteksi kesalahan klitik yang dihasilkan sistem. Dengan menggunakan database, seluruh data dapat

tersimpan secara terstruktur sehingga mempermudah proses pengelolaan informasi dan komunikasi antar pengguna dalam sistem website.

SQLite diintegrasikan langsung dengan framework Flask menggunakan library bawaan Python, yaitu sqlite3. Integrasi ini memungkinkan sistem melakukan proses tambah, ubah, hapus, dan tampil data secara lebih mudah. Selain itu, SQLite juga membantu proses autentikasi akun pengguna dan penyimpanan riwayat berita yang telah diproses oleh editor. Penggunaan SQLite pada penelitian ini dinilai sesuai karena sistem masih berada pada tahap pengembangan dan implementasi prototipe sehingga tidak memerlukan database server dengan konfigurasi yang kompleks.

Keunggulan SQLite dalam penelitian ini meliputi proses instalasi yang sederhana, tidak membutuhkan konfigurasi server tambahan, memiliki ukuran yang ringan, serta mampu berjalan dengan cepat pada aplikasi berbasis lokal. Selain itu, SQLite juga mudah digunakan dalam pengembangan aplikasi berbasis Python dan Flask sehingga mempercepat proses implementasi sistem. Dengan berbagai kelebihan tersebut, SQLite menjadi pilihan yang tepat untuk mendukung kebutuhan database pada penelitian sistem penapisan kesalahan klitik berbasis web ini.

2.4.5 Bootstrap

Bootstrap merupakan framework frontend yang digunakan untuk membantu proses pembuatan antarmuka website pada penelitian ini[59]. Bootstrap dikembangkan menggunakan kombinasi HTML, CSS, dan JavaScript serta menyediakan berbagai komponen antarmuka yang siap digunakan[62]. Framework ini dipilih karena mampu membantu pengembangan tampilan website menjadi lebih modern, responsif, dan terstruktur tanpa perlu membuat desain dari awal secara manual. Selain itu, Bootstrap juga memiliki dokumentasi yang lengkap sehingga mempermudah proses implementasi pada aplikasi berbasis web.

Dalam penelitian ini, Bootstrap digunakan untuk membangun berbagai tampilan antarmuka pada sistem website penapisan kesalahan klitik. Komponen Bootstrap digunakan dalam pembuatan halaman login, sidebar navigasi, tabel data berita, form input berita, tampilan hasil penapisan, hingga halaman manajemen pengguna. Penggunaan class dan komponen bawaan Bootstrap membantu proses desain

halaman menjadi lebih cepat dan konsisten. Selain itu, Bootstrap juga digunakan untuk memberikan tampilan highlight warna pada hasil deteksi kesalahan klitik sehingga bagian teks yang salah dapat terlihat lebih jelas oleh pengguna.

Bootstrap menyediakan berbagai fitur antarmuka seperti navbar, button, card, form, tabel, alert, dan grid system yang membantu proses pengembangan website menjadi lebih efisien. Framework ini juga mendukung desain responsif sehingga tampilan website dapat menyesuaikan ukuran layar pengguna secara otomatis. Dengan adanya fitur responsif, website tetap dapat digunakan dengan baik pada berbagai ukuran perangkat seperti laptop maupun komputer dengan resolusi berbeda.

Penggunaan Bootstrap dalam penelitian ini membantu meningkatkan kualitas tampilan website dan pengalaman pengguna. Antarmuka website menjadi lebih rapi, mudah dipahami, dan nyaman digunakan oleh reporter, editor, maupun admin. Selain mempercepat proses pengembangan frontend, Bootstrap juga membantu menciptakan tampilan website yang lebih profesional dan terstruktur sehingga mendukung implementasi sistem penapisan kesalahan klitik berbasis web secara lebih optimal.

2.4.6 Scikit-learn

Scikit-learn merupakan library machine learning berbasis Python yang digunakan dalam penelitian ini untuk proses klasifikasi teks dan ekstraksi fitur. Scikit-learn merupakan salah satu pustaka machine learning yang paling banyak digunakan karena memiliki struktur penggunaan yang sederhana, dokumentasi yang lengkap, serta mendukung berbagai metode pengolahan data dan algoritma klasifikasi[63]. Library ini dirancang untuk mempermudah proses implementasi machine learning pada berbagai jenis penelitian, termasuk penelitian berbasis pengolahan teks dan NLP[64].

Dalam penelitian ini, Scikit-learn digunakan untuk melakukan ekstraksi fitur pada teks dengan metode TF-IDF, melatih model machine learning, serta melakukan prediksi terhadap data teks baru. Proses ekstraksi fitur dilakukan untuk mengubah teks menjadi representasi numerik agar dapat diolah menggunakan algoritma machine learning secara lebih efektif. Setelah data diubah menjadi fitur numerik,

model machine learning dilatih menggunakan dataset yang telah diberi label benar dan salah berdasarkan pola penulisan klitik.

Selain digunakan untuk proses klasifikasi, Scikit-learn juga membantu proses pengelolaan data machine learning secara lebih efisien. Library ini menyediakan berbagai fungsi siap pakai untuk preprocessing data, pembagian dataset training dan testing, pelatihan model, hingga evaluasi performa model. Penggunaan fungsi yang telah tersedia membantu mempercepat proses implementasi sistem sehingga peneliti tidak perlu membangun algoritma machine learning dari awal secara manual.

Scikit-learn dipilih karena mudah digunakan dan memiliki kompatibilitas yang baik dengan library Python lainnya seperti Pandas dan NumPy. Integrasi tersebut mempermudah proses pengolahan dataset sebelum digunakan dalam pelatihan model machine learning. Selain itu, Scikit-learn juga mendukung berbagai metode klasifikasi dan pengolahan data sehingga sangat sesuai digunakan dalam penelitian pengembangan sistem deteksi kesalahan klitik berbasis web ini.

2.4.7 Pandas

Pandas merupakan library Python yang dimanfaatkan untuk mengolah dan memanipulasi data dalam bentuk tabel maupun dataset terstruktur[65]. Library ini menjadi salah satu tools yang paling sering dimanfaatkan dalam bidang data science dan machine learning karena menyediakan berbagai fungsi yang memudahkan proses pengelolaan serta analisis data[66]. Library ini mampu menangani data dalam jumlah besar dengan struktur yang rapi sehingga sangat membantu proses analisis dan preprocessing data sebelum dimanfaatkan pada proses pelatihan model machine learning.

Dalam penelitian ini, Pandas digunakan untuk membaca dataset, membersihkan data, mengelola struktur data, serta melakukan preprocessing teks sebelum digunakan dalam proses pelatihan model. Dataset yang digunakan disimpan dalam format CSV sehingga Pandas membantu proses pembacaan dan pengelolaan data menjadi lebih mudah dan terstruktur. Dengan menggunakan Pandas, peneliti dapat melihat isi dataset, melakukan pengecekan data kosong, memfilter data tertentu,

serta mempersiapkan data sebelum diproses menggunakan algoritma machine learning.

Beberapa fungsi Pandas yang digunakan dalam penelitian ini antara lain membaca file dataset menggunakan `read_csv()`, memfilter data berdasarkan kategori tertentu, menghapus data kosong atau duplikat, mengelompokkan data, serta menyimpan kembali hasil pengolahan dataset. Selain itu, Pandas juga digunakan untuk membantu proses pengelolaan label data benar dan salah pada dataset klitik sehingga proses training model dapat dilakukan dengan lebih terorganisir.

Penggunaan Pandas membantu mempercepat proses pengolahan dataset dalam jumlah besar karena berbagai fungsi manipulasi data telah tersedia secara siap pakai. Integrasi Pandas dengan library lain seperti NumPy dan Scikit-learn juga mempermudah proses preprocessing dan machine learning pada penelitian ini. Dengan kemampuan pengolahan data yang fleksibel dan efisien, Pandas menjadi salah satu tools penting dalam pembangunan sistem deteksi kesalahan klitik berbasis web.

2.4.8 Pickle

Pickle merupakan library bawaan pada Python yang digunakan untuk menyimpan dan memuat kembali objek Python ke dalam bentuk file[65]. Serialisasi adalah proses mengubah objek Python menjadi bentuk file sehingga dapat disimpan dan digunakan kembali di kemudian hari. Dalam bidang machine learning, Pickle sering digunakan untuk menyimpan model yang telah dilatih agar tidak perlu melakukan proses training ulang setiap kali sistem dijalankan[67]. Dengan menggunakan Pickle, proses implementasi model machine learning pada aplikasi menjadi lebih praktis dan efisien.

Dalam penelitian ini, Pickle digunakan untuk menyimpan model klasifikasi machine learning dan vectorizer TF-IDF yang telah melalui proses pelatihan. Setelah model selesai dilatih menggunakan dataset klitik, model dan vectorizer disimpan dalam file berekstensi `.pkl`. File tersebut kemudian dapat dipanggil kembali pada aplikasi website berbasis Flask untuk melakukan proses prediksi terhadap teks berita yang dimasukkan pengguna. Dengan cara ini, sistem dapat

langsung menggunakan model yang telah dilatih tanpa harus mengulang seluruh proses training.

Penggunaan Pickle membantu integrasi antara proses machine learning dan website menjadi lebih mudah. Pada saat website dijalankan, sistem hanya perlu melakukan proses load model menggunakan fungsi `pickle.load()` sehingga model siap digunakan untuk mendeteksi kesalahan klitik pada teks berita. Pendekatan ini membuat proses implementasi sistem menjadi lebih ringan dan cepat karena model tidak perlu dibuat ulang setiap kali aplikasi digunakan.

Selain meningkatkan efisiensi sistem, penggunaan Pickle juga membantu mempercepat proses pengembangan dan pengujian aplikasi. Peneliti dapat melakukan pelatihan model secara terpisah di lingkungan pengembangan seperti Jupyter Notebook, kemudian menggunakan hasil model tersebut langsung pada website Flask. Dengan demikian, Pickle menjadi salah satu tools penting yang mendukung proses deployment model machine learning pada sistem penapisan kesalahan klitik berbasis web.

2.4.9 Jupyter Notebook

Jupyter Notebook merupakan aplikasi berbasis web yang digunakan untuk menjalankan, menyusun, serta mendokumentasikan kode program secara interaktif[68]. Aplikasi ini banyak dimanfaatkan pada bidang data science, machine learning, dan kegiatan penelitian karena memungkinkan pengguna menjalankan kode per bagian (cell) sehingga proses eksperimen dan analisis data menjadi lebih mudah dilakukan[68]. Selain menampilkan kode program, Jupyter Notebook juga dapat digunakan untuk menampilkan output, tabel, grafik, serta penjelasan teks dalam satu dokumen yang terstruktur.

Dalam penelitian ini, Jupyter Notebook digunakan pada tahap eksperimen dan pengembangan algoritma machine learning sebelum sistem diimplementasikan ke dalam website. Beberapa proses yang dilakukan menggunakan Jupyter Notebook antara lain preprocessing dataset, pengujian algoritma, pelatihan model machine learning, evaluasi hasil prediksi, serta pengujian dataset klitik. Dengan menggunakan Jupyter Notebook, peneliti dapat mencoba berbagai proses

pengolahan data dan melihat hasilnya secara langsung tanpa harus menjalankan keseluruhan program dari awal.

Jupyter Notebook dipilih karena mendukung proses eksekusi kode secara interaktif sehingga sangat membantu dalam proses eksperimen machine learning dan pengolahan teks. Peneliti dapat menjalankan kode per cell, memperbaiki kesalahan program dengan lebih cepat, serta mengamati hasil output secara langsung. Selain itu, struktur notebook yang tersusun rapi juga mempermudah proses dokumentasi penelitian karena kode program, hasil output, dan penjelasan dapat ditampilkan dalam satu halaman yang sama.

Penggunaan Jupyter Notebook membantu mempercepat proses pengembangan model machine learning sebelum diintegrasikan dengan website Flask. Dengan lingkungan kerja yang interaktif dan fleksibel, peneliti dapat melakukan pengujian dataset, evaluasi model, serta analisis hasil prediksi secara lebih efektif. Oleh karena itu, Jupyter Notebook menjadi salah satu tools penting yang mendukung proses penelitian sistem deteksi kesalahan klitik berbasis web ini.

2.4.10 HTML dan CSS

HTML (HyperText Markup Language) dan CSS (Cascading Style Sheets) merupakan teknologi dasar yang digunakan dalam pembangunan antarmuka website. HTML berfungsi untuk membangun struktur serta elemen pada halaman website, sedangkan CSS dimanfaatkan untuk mengatur aspek visual seperti warna, ukuran, tata letak, dan desain antarmuka website[69]. Kombinasi HTML dan CSS memungkinkan pengembang membangun website yang tidak hanya berjalan dengan baik, tetapi juga memiliki antarmuka yang lebih rapi dan mudah digunakan oleh pengguna[69].

Dalam penelitian ini, HTML digunakan bersama template engine dari Flask untuk membangun berbagai halaman pada sistem website penapisan kesalahan klitik. Halaman yang dibuat menggunakan HTML meliputi halaman login, halaman reporter, halaman editor, halaman admin, serta halaman hasil penapisan. HTML digunakan untuk membuat elemen-elemen website seperti form input berita, tabel daftar berita, tombol aksi, navigasi halaman, dan tampilan hasil deteksi kesalahan.

Struktur halaman yang dibuat menggunakan HTML membantu sistem menampilkan data secara terorganisir dan mudah dipahami oleh pengguna.

CSS digunakan untuk mempercantik tampilan website dan mengatur tata letak antarmuka agar lebih nyaman digunakan. Dengan CSS, tampilan halaman dapat disesuaikan mulai dari warna, ukuran teks, posisi elemen, jarak antar komponen, hingga efek visual tertentu. Pada penelitian ini, CSS juga digunakan untuk mendukung tampilan highlight hasil deteksi kesalahan klitik sehingga bagian teks yang dianggap salah dapat ditampilkan dengan warna tertentu agar lebih mudah dikenali oleh editor.

Penggunaan HTML dan CSS membantu menciptakan antarmuka website yang lebih interaktif, rapi, dan responsif. Selain meningkatkan kenyamanan pengguna, tampilan website yang terstruktur juga membantu proses penggunaan sistem oleh reporter, editor, maupun admin. Dengan kombinasi HTML, CSS, dan framework Bootstrap, website pada penelitian ini dapat memiliki tampilan yang lebih modern dan mendukung proses penapisan berita secara lebih efektif.

