|  |  |  |
| --- | --- | --- |
| A Objective Meningkatkan minat dan retensi membaca masyarakat Indonesia adalah tujuan yang penting. Diperlukan upaya kolaboratif dari pemerintah, lembaga pendidikan, dan masyarakat secara luas. Pengembangan program literasi yang menarik dan mudah diakses dapat menjadi langkah yang efektif dalam mencapai tujuan tersebut. KELOMPOK Kelompok 4:   * Klasik S.K.M.Taidi – 210711185 * Bima Wahyu L – 21071133 * Valentinus Wilbert - 210711340 | |  | | --- | | kLASIFIKASI GENRE BUKUMENGKLASIFIKASI GENRE BUKU BERDASARKAN SINOPSIS DAN DAN JUDUL BUKU |  Latar belakaNg Klasifikasi genre buku memiliki peranan penting dalam industri literatur, tidak hanya membantu pembaca menemukan buku sesuai minat mereka tetapi juga mendukung penerbit dalam memahami pasar target dan merancang strategi pemasaran. Namun, tantangan muncul karena banyak buku yang mencampur unsur-unsur genre, dan batasan antara genre seringkali menjadi kabur. Oleh karena itu, penting untuk mengembangkan metode klasifikasi yang akurat dan efektif guna memahami kompleksitas sastra modern dan memfasilitasi navigasi yang lebih baik di antara berbagai jenis buku, sekaligus mendukung pemahaman yang lebih baik tentang preferensi pembaca dan tren pasar, serta terus memperbarui sistem klasifikasi ini sesuai dengan perkembangan industri literatur. Deskripsi DatasetSumber Projek yang kami buat kali ini memanfaatkan dataset dari Google Book API yang kami akses melalui kode Python yang kami buat sendiri, khususnya untuk mengakses buku-buku berbahasa Indonesia. Dengan memanfaatkan API ini, kami dapat mengumpulkan informasi yang relevan untuk analisis dan pengembangan projek kami. Penggunaan sumber data ini menjadi langkah awal yang penting dalam proses merancang dan melaksanakan projek kami, karena kami dapat memastikan keakuratan dan kelengkapan informasi yang kami gunakan serta melakukan pengolahan data secara efisien melalui kode Python. karakteristik Dataset yang kami gunakan memiliki beberapa unsur, seperti Title (judul), Author(Penulis), Categories(Genre), dan Description(Sinopsis). pada dataset kami juga memiliki jumlah data sebanyak 344 baris yang kami pilih untuk mendapatkan dataset yang seimbang dan memenuhi aspek yang di butuhkan dalam proses learning projek. Rancangan an Implementasi Pemrosesan Dataset Langkah pertama dalam persiapan data adalah impor pustaka yang diperlukan serta dataset yang akan digunakan. Dalam kode yang diberikan, kita dapat melihat impor pustaka seperti numpy, pandas, dan nltk, yang penting untuk pengolahan data dan pemodelan teks. Selain itu, dataset yang digunakan diimpor dari file Excel menggunakan modul pandas. Setelah itu, eksplorasi awal terhadap dataset dilakukan dengan menampilkan dimensi dataset dan distribusi kategori menggunakan plot grafik. Langkah ini penting untuk memahami struktur dataset dan mendapatkan gambaran awal tentang distribusi kategori yang akan menjadi dasar untuk analisis lebih lanjut.  Langkah berikutnya melibatkan pemrosesan teks menggunakan berbagai fungsi normalisasi yang telah didefinisikan. Dalam kode, terdapat fungsi-fungsi seperti normalisasi karakter berulang, pemeriksaan ejaan, tokenisasi, ekspansi kontraksi, stemming, penghapusan karakter khusus, dan penghapusan stopwords. Fungsi-fungsi ini digunakan untuk membersihkan teks dalam dataset dari noise dan standarisasi format teks agar sesuai dengan kebutuhan analisis lebih lanjut. Setelah normalisasi, dataset dibagi menjadi data pelatihan dan data uji menggunakan metode train\_test\_split, yang merupakan praktik umum dalam pemodelan untuk mengevaluasi kinerja model.  Langkah terakhir melibatkan ekstraksi fitur menggunakan model TF-IDF (Term Frequency-Inverse Document Frequency). Dengan menggunakan TF-IDF, teks dalam dataset diubah menjadi representasi numerik yang dapat digunakan dalam pemodelan. Setelah ekstraksi fitur dilakukan pada data pelatihan, model TF-IDF yang sama diterapkan pada data uji untuk memastikan konsistensi dalam representasi teks. Selanjutnya, data pelatihan dan data uji disusun ke dalam kerangka data pandas untuk kemudian digunakan dalam pemodelan. Dengan langkah-langkah ini, data yang terkumpul telah diproses dan disiapkan untuk digunakan dalam pemodelan lanjutan seperti klasifikasi atau analisis teks. Rancangan dan Implementasi Model Pembelajaran Mesin Rancangan pemodelan algoritma melibatkan tiga jenis model: SGD Classifier (SGDC), Linear Support Vector Classifier (LinearSVC), dan Multinomial Naive Bayes (NB). Setiap model diimplementasikan dalam sebuah pipeline dengan dua langkah utama: pemilihan fitur menggunakan metode SelectFromModel dengan LinearSVC sebagai estimator, dan klasifikasi menggunakan model yang sesuai dengan jenis algoritma yang dipilih. Parameter-parameter yang akan dioptimalkan untuk masing-masing model mencakup berbagai variabel seperti 'loss', 'alpha', 'penalty', dan 'max\_iter' untuk SGDC, 'C', 'loss', dan 'max\_iter' untuk LinearSVC, dan 'alpha' dan 'fit\_prior' untuk NB. Dengan menggunakan GridSearchCV, pencarian parameter terbaik dilakukan untuk setiap model, yang kemudian menghasilkan parameter terbaik yang memberikan skor validasi terbaik untuk model yang digunakan.  Setelah proses pencarian parameter terbaik selesai, hasilnya dicetak untuk masing-masing model, termasuk nilai skor terbaik, parameter terbaik, dan estimator terbaik. Dengan demikian, kita mendapatkan model-model yang dioptimalkan siap untuk digunakan dalam analisis teks lebih lanjut atau klasifikasi. Implementasi ini memungkinkan penggunaan algoritma dengan parameter-parameter yang telah dioptimalkan untuk meningkatkan kinerja model secara keseluruhan. Dengan mengintegrasikan pemilihan fitur dan klasifikasi dalam pipeline, serta menggunakan pencarian parameter terbaik dengan GridSearchCV, proses pemodelan menjadi lebih efisien dan efektif dalam mencapai hasil yang optimal. Rancangan dan Implementasi Evaluasi Model Proses evaluasi model dimulai dengan perhitungan metrik evaluasi seperti akurasi, presisi, recall, dan F1 Score untuk masing-masing model yang dievaluasi. Metrik-metrik ini memberikan gambaran komprehensif tentang kinerja model dalam melakukan klasifikasi data uji. Selain itu, matriks kebingungan (confusion matrix) juga dihasilkan untuk masing-masing model, yang memberikan informasi lebih rinci tentang seberapa baik model dapat mengklasifikasikan sampel data ke dalam kelas yang benar, serta seberapa sering ia melakukan kesalahan klasifikasi.  Setelah menghitung metrik evaluasi dan mendapatkan matriks kebingungan, langkah selanjutnya adalah melakukan visualisasi confusion matrix menggunakan ConfusionMatrixDisplay dari pustaka sklearn.metrics. Visualisasi ini memberikan pemahaman yang lebih intuitif tentang sebaran prediksi yang benar dan salah dari masing-masing kelas. Dengan demikian, proses evaluasi model ini memberikan wawasan yang mendalam tentang kinerja masing-masing model, yang penting untuk pemilihan model yang tepat dan pengambilan keputusan yang akurat dalam konteks analisis atau klasifikasi data. KESIMPULAN Dalam proyek klasifikasi genre buku, kesimpulan utama dapat diambil dari pemilihan fitur yang relevan dan pemilihan algoritme yang sesuai. Memilih fitur yang paling informatif dan mengidentifikasi algoritma pembelajaran mesin yang sesuai untuk data adalah langkah pertama yang penting. Selain itu, penyetelan parameter yang optimal juga merupakan kunci untuk meningkatkan kinerja model. Evaluasi kinerja model dengan menggunakan metrik yang tepat, serta interpretasi mendalam terhadap hasil, membantu kita memahami kekuatan dan kelemahan model yang dikembangkan. Dengan merangkum temuan-temuan dari proyek ini, kami dapat mengidentifikasi peluang pengembangan lebih lanjut dalam industri penerbitan dan memberikan wawasan yang berharga bagi para pengambil keputusan dalam konteks bisnis. |