

Hasil dan pembahasan

Hasil dan pembahasan merupakan penjelasan hasil dari penelitian yang telah dilakukan mengenai perbandingan performa SVM dengan Naïve Bayes dan analisis sentimen terhadap data Clash of Clans. Pada tahap ini, hasil akan dijelaskan dari Pengambilan Data, Preprocessing Data, Ekstraksi dan Seleksi Fitur, Machine Learning Model, Hasil dari Klasifikasi Sentimen Analisis menggunakan SVM dan Naïve Bayes dan Evaluasi terhadap hasil yang telah diperoleh dalam penelitian. Berikut penjelasan mengenai hasil dan pembahasan dari setiap langkah penelitian.

Pengambilan Data Dataset yang peneliti gunakan pada penelitian ini merupakan dataset ulasan dari pengguna aplikasi Clash of Clans dari hasil scrapping pada Google Play Store. Data ini memiliki 10 atribut dan 1000 data ulasan. Langkah pertama yang harus dilakukan adalah menginstall google play scraper, kemudian install library pandas dan numpy. Arahkan google play scraper ke “com.supercell.clashofclans” untuk melakukan scrapping data ulasan pada aplikasi Clash of Clans. Data yang diambil berjumlah 1000 data ulasan newest atau terbaru dengan bahasa dan negara Indonesia pada 5 Januari 2023 sampai 8 Januari 2023. Tabel 1 menunjukkan dataset ulasan yang berisi score, at dan content.

Preprocessing Data Tahap preprocessing data adalah tahap dimana data yang diambil dari google play store menggunakan teknik scrapping akan dibersihkan sebelum dianalisa lebih lanjut. Data mentah tersebut akan melalui beberapa proses, yang pertama adalah Case Folding yaitu mengubah seluruh kata menjadi lower case, menghapus URL, angka dan karakter tanda baca termasuk emoji. Filtering (Stopword Removal) yaitu menghapus kata atau karakter tanda baca dengan informasi rendah atau ganda. Stemming yaitu menghapus imbuhan. Proses terakhir adalah Word Normalization yaitu mengubah kata slang menjadi kata baku atau kata dasar. Tabel 3 merupakan hasil preprocessing data pada salah satu ulasan pengguna yaitu rawdata pada baris ke 771. Keempat proses tersebut dilakukan untuk mendapatkan data yang bersih sehingga memudahkan proses pembobotan kata dan evaluasi model.

Ekstraksi Fitur dan Seleksi Fitur Ekstraksi fitur adalah tahapan di mana fitur-fitur baru diekstraksi dari fitur-fitur asli melalui suatu pemetaan fungsional, dan nantinya nilai-nilai yang dihasilkan akan dianalisis untuk tahapan berikutnya. Seleksi fitur digunakan untuk mengurangi atribut atau fitur yang tidak relevan dalam dataset dan mempercepat pemrosesan data. Dalam penelitian ini, dilakukan ekstraksi fitur menggunakan metode TF-IDF dan N-Gram, serta seleksi fitur menggunakan metode Chi-Square dan SelectKBest.

Penerapan optimasi ekstraksi fitur dan seleksi fitur dengan cara melakukan 6 kali uji kombinasi pada fitur. Tujuan pengujian ini untuk mendapatkan model klasifikasi yang paling optimal. Selain itu, perbandingan pengujian untuk mendapatkan hasil akurasi yang lebih akurat. Enam kombinasi uji fitur yang dilakukan yaitu TF-IDF, TF-IDF dan N-Gram, TF-IDF dan Chi-Square, TF-IDF N-Gram dan Chi-Square, TF-IDF Chi-Square dan SelectKBest, TF-IDF

N-Gram Chi-Square dan SelectKBest. TF-IDF dan N-Gram merupakan kombinasi dari metode ekstraksi fitur. Kombinasi tersebut digunakan untuk mengatasi kelemahan yang ada pada TF-IDF.

Machine Learning Model Machine learning model adalah hasil dari fase pelatihan (training phase) di mana sistem belajar untuk menemukan pola di dalam data. Train-test split adalah salah satu metode evaluasi model machine learning. Dalam metode ini, dataset dibagi menjadi dua bagian yaitu training data dan testing data, masing-masing bagian digunakan untuk tujuan yang berbeda. Training data digunakan untuk melatih model machine learning, sementara testing data digunakan untuk mengevaluasi performa model tersebut. Proporsi pembagian antara training data dan testing data pada penelitian ini adalah 70:30, 80:20, dan 90:10.

Hasil dan Evaluasi Setelah mendapatkan data uji dan data latih model akan dievaluasi menggunakan algoritma klasifikasi Support Vector Machine dan Naïve Bayes. Pada penelitian ini evaluasi model dilakukan dengan bantuan pustaka sk-learn. Naïve Bayes pada penelitian ini menggunakan variasi yang sering digunakan untuk klasifikasi teks yaitu Gaussian Naïve Bayes, sementara Support Vector Machine menggunakan variasi SVC. Setelah model diujikan akan dilakukan perhitungan akurasi untuk mengetahui performa pada tiap algoritma menggunakan tabel confusion matrix. Gambar 3 merupakan tabel confusion matrix yang memiliki dua kelas untuk memudahkan dalam proses perhitungan performa dari suatu model algoritma klasifikasi.