BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan penelitian dan implementasi yang dilakukan pada algoritma Random Forest Classifier, dapat disimpulkan bahwa penelitian telah berhasil dilakukan. Melalui proses hyperparameter dan strategi validasi silang dengan nilai K sama dengan 5 diperoleh model terbaik dengan nilai F1 score sebesar 93,49% model ini menggunakan parameter *TF-IDF vectorizer* dengan nilai *analyzer* berupa *word,* minimum document frequencies sebesar 2,5% dan nilai rentang N-gram 1 dan 2. Kemudian parameter model Random Forest terbaik diperoleh dengan nilai maksimum kedalaman pohon senilai 64, nilai minimum pembelahan sampel senilai 4, jumlah pohon senilai 250 dan sisa parameter bernilai *default* mengikuti implementasi yang digunakan.

5.2 Saran

Berdasarkan penelitian dan implementasi yang dilakukan terdapat beberapa saran untuk pengembangan lanjutan yaitu :

1. Berdasarkan implementasi dari metode TF-IDF pada dataset yang digunakan, jumlah fitur yang dihasilkan oleh algoritma TF-IDF berukuran relatif cukup besar. Padahal, berdasarkan parameter feature importances yang dimiliki model Random Forest Classifier yang telah dilatih, tidak semua fitur memiliki derajat kepentingan (importance) yang cukup signifikan.

Berdasarkan fakta yang telah dijabarkan, tentunya dapat dilakukan proses improvisasi terhadap model klasifikasi dengan mereduksi jumlah fitur yang dibutuhkan sebelum proses pelatihan model. Untuk melakukan reduksi terhadap jumlah fitur, dapat digunakan algoritma Principal Component Analysis (PCA) ataupun Factor Analysis (FA). Melalui tahapan reduksi yang ditambahkan, dengan berkurangnya jumlah fitur yang harus dipelajari oleh model saat proses pelatihan, diharapkan proses pelatihan model dapat menjadi lebih cepat. Melalui proses pelatihan model yang lebih cepat, dengan menggunakan waktu yang sama dapat dilakukan, proses eksplorasi model dengan parameter-parameter yang lebih beragam dapat dilakukan guna mencari model dengan performa yang lebih baik lagi.

2. Berdasarkan hasil evaluasi yang telah dilakukan terhadap model hasil proses pelatihan, terlihat bahwa, model terbaik menggunakan metode ekstraksi fitur berbasis kata (word-level TF-IDF). Menambahkan proses pre-preprocessing untuk pengecekan kesalah penulisan pada dataset mengingat bahwa dalam dataset yang digunakan masih terdapat kesalahan-kesalahan penulisan kata yang terjadi proses perbaikan kesalahan penulisan sebagai tahapan preprocessing tentunya diduga dapat meningkatkan performa model klasifikasi. Hal ini dikarenakan pada model TF-IDF, kata dengan kesalahan penulisan dan kata aslinya akan dianggap sebagai dua fitur yang berbeda dan memiliki bobot fitur yang berbeda pula. Hal tersebut tentunya membuat ukuran hasil ekstraksi fitur dokumen menjadi lebih besar dan membuat proses klasifikasi terhadap fitur menjadi lebih kompleks.