

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan mengenai klasifikasi multikategori berita olahraga berbahasa Indonesia menggunakan metode *Term Frequency Inverse Document Frequency (TF-IDF)* dengan algoritma *K-Nearest Neighbor (KNN)* dan *Support Vector Machine (SVM)*, dapat disimpulkan bahwa metode TF-IDF mampu merepresentasikan teks berita secara efektif ke dalam bentuk fitur numerik yang dapat membedakan karakteristik masing-masing kategori, yaitu Sepak Bola, Badminton, dan MotoGP. Proses *preprocessing teks* yang meliputi *case folding*, *tokenizing*, *stopword removal*, dan *stemming* terbukti berkontribusi dalam meningkatkan kualitas *representasi* teks sebelum dilakukan proses klasifikasi. Hasil pengujian pada dua skenario pembagian data, yaitu 70:30 dan 80:20, menunjukkan bahwa kedua algoritma memiliki performa yang sangat tinggi dengan tingkat akurasi di atas 99%. Pada skenario 70:30, KNN dan SVM menghasilkan performa yang identik, sedangkan pada skenario 80:20, SVM menunjukkan performa yang sedikit lebih unggul dan lebih stabil dibandingkan KNN. Validasi menggunakan *Stratified K-Fold Cross Validation* memperlihatkan bahwa kedua model memiliki tingkat konsistensi dan kemampuan generalisasi yang sangat baik, yang diperkuat oleh analisis *macro average* dan *learning curve* yang menunjukkan tidak adanya *overfitting* yang signifikan. Implementasi sistem klasifikasi berbasis Streamlit juga membuktikan bahwa model yang dikembangkan dapat digunakan untuk mengklasifikasikan berita olahraga aktual secara langsung. Dengan demikian, dapat disimpulkan bahwa algoritma *Support Vector Machine (SVM)* merupakan model yang paling optimal dalam penelitian ini, meskipun *K-Nearest Neighbor (KNN)* juga menunjukkan performa yang sangat kompetitif dalam klasifikasi berita olahraga berbahasa Indonesia.

5.2 Saran

Meskipun penelitian ini telah menghasilkan performa model yang sangat tinggi, masih terdapat beberapa hal yang dapat dikembangkan pada penelitian

selanjutnya. Penggunaan dataset yang lebih beragam dari berbagai portal berita daring dapat dilakukan untuk meningkatkan kemampuan generalisasi model terhadap variasi gaya bahasa yang lebih luas. Selain itu, penelitian berikutnya dapat mempertimbangkan penggunaan metode *representasi* teks berbasis *word embedding* atau model *transformer* seperti BERT untuk dibandingkan dengan pendekatan TF-IDF yang digunakan dalam penelitian ini. Dan Juga Pengembangan sistem ke dalam bentuk aplikasi berbasis web yang terintegrasi dengan sistem manajemen konten portal berita juga menjadi peluang pengembangan lebih lanjut agar hasil penelitian ini dapat dimanfaatkan secara lebih luas dalam praktik industri media digital.

