BAB V
PENUTUP

5.1. Kesimpulan

Berdasarkan hasil penelitian dan pembahasan yang telah diuraikan pada Bab IV, dapat ditarik kesimpulan sebagai berikut, yang disusun berdasarkan rumusan masalah dan tujuan penelitian yang telah ditetapkan:

1. Data penelitian berhasil dikumpulkan melalui proses *web scraping* menggunakan browser extension Data Scrapper pada halaman ulasan enam judul buku dari platform *Goodreads* dengan genre yang beragam (*Literary Fiction, Gothic Horror, Science Fiction, Dystopian, Romance, dan Romance Fantasy*). Proses ini menghasilkan data ulasan teks berbahasa Inggris, yang setelah melalui tahap pembersihan (penghapusan duplikat, nilai kosong, filtering bahasa, dan filtering panjang kata minimal 200 kata) serta penyeimbangan jumlah sampel antar buku (150 ulasan per buku), menghasilkan *dataset* akhir sebanyak 900 baris ulasan yang siap digunakan untuk tahap selanjutnya.
2. Data ulasan melalui serangkaian tahap *text preprocessing* yang meliputi *text cleaning* (penghapusan URL, simbol, dan spasi berlebih), *case folding*, *tokenizing*, *stopword removal*, dan *lemmatization* menggunakan WordNetLemmatizer. Tahapan ini terbukti efektif dalam mempersiapkan data teks agar dapat diproses lebih lanjut, meskipun ditemukan pula keterbatasan pada pendekatan *stopword removal* standar yang berpotensi menghapus kata-kata negasi, sehingga perlu ditinjau lebih lanjut pada penelitian selanjutnya. Data yang telah melalui *Preprocessing* kemudian direpresentasikan secara numerik menggunakan metode BM25 (*Best Matching 25*), yang terbukti mampu memberikan bobot lebih tinggi pada kata-kata yang diskriminatif terhadap suatu kelas dibanding kata-kata umum yang sering muncul di seluruh korpus.
3. Model klasifikasi dibangun menggunakan pendekatan *stacking ensemble* dua lapis, dengan *Naive Bayes* dan *XGBoost* sebagai *Base learner* pada lapisan pertama, serta *Maximum Entropy* (Logistic Regression) sebagai *meta-learner*

pada lapisan kedua. Skema *5-fold cross-validation* dengan pendekatan *out-of-fold prediction* diterapkan untuk menghasilkan fitur meta yang bebas dari kebocoran data, sementara teknik *Class Weighting* (`class_weight="balanced"` dan `sample_weight`) diterapkan secara konsisten di seluruh komponen pipeline untuk menangani ketidakseimbangan kelas antara ulasan *recommended* dan *not recommended* pada kedua *dataset* yang digunakan.

4. Model *stacking ensemble* menunjukkan performa yang baik dan konsisten pada metrik yang relevan untuk data *imbalanced*, yaitu *macro F1-score* dan *recall* kelas minoritas, dengan hasil terbaik dicapai pada kombinasi dataset *manual-labeled* dan rasio split 70:30 (*macro* F1 0,76; *recall* kelas *not recommended* 0,73). Temuan ini turut didukung oleh nilai ROC AUC yang tergolong *acceptable* hingga *good* pada seluruh kombinasi eksperimen (0,7620 hingga 0,8491), dengan nilai tertinggi juga diperoleh pada dataset *manual-labeled*, menunjukkan bahwa model stacking memiliki kemampuan yang cukup baik dalam membedakan kelas rekomendasi dan tidak rekomendasi secara menyeluruh di berbagai *threshold* klasifikasi. Pada seluruh rasio split yang diuji (60:40, 70:30, 80:20), *dataset* hasil anotasi manual secara konsisten menghasilkan performa yang lebih baik dibanding *dataset* hasil pelabelan otomatis berbasis VADER, baik dari sisi *macro* F1-*score, recall* kelas minoritas, maupun ROC AUC, menegaskan bahwa kualitas metode pelabelan data berpengaruh signifikan terhadap performa akhir model secara menyeluruh. Penerapan teknik *class balancing* secara konsisten di seluruh komponen pipeline juga terbukti berkontribusi menjaga kemampuan model dalam mengenali kelas minoritas pada rentang yang wajar (*recall* 0,62 hingga 0,73), alih-alih mengabaikannya sepenuhnya sebagaimana lazim terjadi pada model yang tidak menerapkan teknik penyeimbangan kelas. Meskipun demikian, perbandingan lebih lanjut dengan model *standalone*, khususnya *Maximum Entropy*, pada rasio 80:20 menunjukkan bahwa model *stacking* belum sepenuhnya optimal dalam memanfaatkan kombinasi *base learner*-nya,

sebagaimana tercermin dari nilai ROC AUC stacking yang masih berada di bawah model *standalone* terbaiknya pada kedua skema pelabelan.

5. Penelitian ini berhasil mengimplementasikan hasil model ke dalam sebuah sistem *Dashboard* interaktif berbasis web bernama "*The Reading Room*"**,** yang menampilkan katalog buku beserta ulasannya, dilengkapi fitur *Insight* untuk menampilkan hasil klasifikasi sentimen (*recommended*/*not recommended*) dari setiap ulasan menggunakan salah satu dari dua model yang telah dilatih (*dataset* VADER atau manual, keduanya dengan rasio split 80:20, pemilihan rasio ini didasari pertimbangan konsistensi metodologi antar kedua model serta memaksimalkan jumlah data latih untuk model yang digunakan dalam produksi). Sistem ini juga dilengkapi halaman "*Behind the Screen*" yang menjelaskan metode dan data yang digunakan, fitur permintaan penambahan buku bagi pengguna terdaftar, serta panel khusus admin untuk mengelola katalog buku, *dataset*, dan permintaan pengguna.

Secara keseluruhan, penelitian ini berhasil menjawab seluruh rumusan masalah dan mencapai tujuan penelitian yang telah ditetapkan, yaitu membangun model analisis sentimen berbasis *stacking ensemble* dengan *Maximum Entropy* sebagai *meta-learner* pada ulasan buku *Goodreads*, mengevaluasi performanya secara menyeluruh, serta mengimplementasikannya ke dalam sebuah sistem *Dashboard* yang fungsional.

5.2. Saran Pengembangan

Berdasarkan hasil, pembahasan, serta beberapa keterbatasan yang ditemukan selama penelitian ini berlangsung, beberapa saran yang dapat dipertimbangkan untuk penelitian atau pengembangan selanjutnya adalah sebagai berikut:

1. Ditemukan bahwa proses *stopword removal* menggunakan daftar kata baku berpotensi menghapus kata negasi yang secara semantik penting untuk membedakan sentimen.
2. *Dataset* penelitian ini terbatas pada 900 ulasan dari enam judul buku. Penelitian selanjutnya dapat memperluas jumlah data serta variasi judul dan genre buku yang digunakan, untuk menguji apakah pola yang ditemukan pada penelitian

ini tetap konsisten pada skala data yang lebih besar, serta untuk mengurangi variansi estimasi metrik evaluasi pada kelas minoritas yang timbul akibat ukuran data uji yang relatif kecil.

3. Penelitian ini menggunakan satu skema anotasi manual. Penelitian selanjutnya disarankan melibatkan lebih dari satu anotator dan mengukur tingkat kesepakatan untuk memastikan konsistensi dan reliabilitas label yang dihasilkan.
4. Penelitian selanjutnya dapat membandingkan performa BM25 dengan teknik representasi teks lain, seperti *word embeddings* (Word2Vec, GloVe) atau model berbasis *transformer* (BERT, IndoBERT untuk kasus multibahasa), serta mengeksplorasi kombinasi *Base learner* dan *meta-learner* lain dalam arsitektur *stacking ensemble*, guna mengetahui apakah performa pada kelas minoritas dapat ditingkatkan lebih lanjut.
5. Selain *Class Weighting* yang digunakan pada penelitian ini, penelitian selanjutnya dapat membandingkan performa dengan teknik *resampling* lain seperti SMOTE (*Synthetic Minority Oversampling Technique*) atau kombinasi *oversampling-undersampling*, untuk mengevaluasi pendekatan mana yang paling efektif menangani ketidakseimbangan kelas pada kasus klasifikasi sentimen ulasan buku.
6. Penelitian ini terbatas pada ulasan berbahasa Inggris. Penelitian selanjutnya dapat memperluas cakupan ke ulasan multibahasa, termasuk ulasan berbahasa Indonesia, untuk memperluas manfaat dan aplikabilitas model pada konteks pengguna yang lebih luas.
7. Sistem "The *Read*ing Room" yang dikembangkan pada penelitian ini belum melalui pengujian *usability* formal dengan pengguna akhir. Penelitian atau pengembangan selanjutnya disarankan melakukan pengujian penerimaan pengguna (*user acceptance testing*) untuk mengevaluasi efektivitas antarmuka serta kejelasan fitur *Insight* dalam membantu pengguna memahami hasil analisis sentimen.
8. Penelitian ini menggunakan parameter yang sebagian besar merupakan nilai *default* pustaka (misalnya pada *Naive Bayes* dan BM25) atau parameter yang ditentukan berdasarkan konvensi umum (*XGBoost*). Penelitian selanjutnya

dapat menerapkan teknik pencarian hyperparameter yang lebih sistematis, seperti *grid search* atau *Bayesian optimization*, untuk mengetahui potensi peningkatan performa lebih lanjut pada masing-masing komponen model.