BAB V
KESIMPULAN DAN SARAN

5.1. Kesimpulan

Kesimpulan yang diperoleh dari penelitian penerapan NLP dan algoritma klasifikasi terhadap data *post* media sosial "X" dari pengguna MRT Jakarta adalah sebagai berikut:

1. Pada penelitian ini dilakukan perancangan terhadap tiga model klasifikasi menggunakan algoritma *Logistic Regression*, algoritma *Random Forest Classifier*, dan algoritma *support vector machine* terhadap data *post* pengguna transportasi umum MRT Jakarta yang diperoleh dari media sosial "X" menggunakan fitur *advanced search* yang disediakan. Data tersebut berjumlah sebanyak 525 teks, yang terdiri atas data kategori 'positif' dengan jumlah 222 data, kategori 'negatif' dengan jumlah 185 data, serta kategori 'netral' dengan jumlah 118 data. Data tersebut dibersihkan dengan teknik *pre-processing* yang terdiri dari *case-folding*, penghapusan karakter non-alfabetik atau *noise reduction*, proses *tokenization*, menghapus kata-kata umum dengan *stopword removal*, proses normalisasi pengejaan, serta penghapusan imbuhan kata dengan stemming. Kemudian dilakukan pembagian terhadap dataset tersebut dengan perbandingan 70:30, yaitu 70% dari total data untuk data latih, dan 30% dari total data untuk data uji, serta perbandingan 80:20, yaitu 80% dari total data untuk data latih, dan 20% dari total data untuk data uji, yang kemudian ditransformasikan menggunakan teknik *TF-IDF*. Untuk nilai dari setiap parameter pada ketiga model, dilakukan pencarian nilai terbaik dengan teknik *hyperparameter tuning* menggunakan "*GridSearchCV"* terhadap segala kemungkinan pada daftar parameter yang penulis tentukan. Sebagai evaluasi, dilakukan analisis terhadap *confusion matrix* dan *classification report* untuk nilai nilai *accuracy*, nilai *precision*, nilai *recall*, dan nilai *f1-score*. Terakhir, dilakukan uji coba menggunakan data baru sebanyak 4 sampel data terhadap ketiga model klasifikasi tersebut.
2. Hal yang dilakukan *Natural Language Processing* (NLP) dalam proses pengenalan terhadap bahasa manusia yang terhimpun pada teks beropini adalah dengan mengubah teks yang mengandung tata bahasa yang mudah dipahami

oleh manusia, dengan struktur atau bentuk yang kompleks dan sarat akan makna, menjadi bentuk penggalan kata yang di mana setiap kata yang tersisa diidentifikasi sebagai inti topik pembicaraan dari kalimat asalnya. Kumpulan kata dasar tersebut kemudian ditransformasikan menjadi bentuk bilangan desimal yang disimpan dalam suatu matriks yang mudah dipahami oleh komputer, namun sulit dipahami jika dibaca secara langsung oleh manusia.

3. Di antara hasil perbandingan dari ketiga pengujian model *machine learning* untuk proses klasifikasi terhadap data *post* pengguna transportasi umum MRT Jakarta dari media sosial "X", disimpulkan bahwa model klasifikasi dengan hasil paling akurat diperoleh melalui rasio pembagian dataset sebesar 80:20 dan dengan menggunakan algoritma *Random Forest Classifier*. Model klasifikasi *Support Vector Machine* dengan rasio pembagian dataset terbaik sebesar 80:20, dan parameter terbaik yang diperoleh menggunakan *hyperparameter tuning* metode *grid search*, yaitu nilai *cost* (*C*) = 1, nilai *gamma* = 'scale', dan penggunaan *kernel* '*linear*', menghasilkan nilai *accuracy* sebesar 78%, nilai *precision* sebesar 77.8%, nilai *recall* sebesar 78%, dan nilai *f1-score* sebesar 77.7%, serta berhasil memprediksi 3 dari 4 sampel data baru berdasarkan target kelasnya. Untuk model *Logistic Regression* multinomial dengan rasio pembagian dataset terbaik sebesar 80:20, dan parameter terbaik menurut hasil *hyperparameter tuning*, yaitu nilai *hyperparameter* (*C*) = 1.6238 dan *penalty* L2, menghasilkan nilai *accuracy* sebesar 74%, nilai *precision* sebesar 72.8%, nilai *recall* sebesar 74%, dan nilai *f1-score* sebesar 72.6%, dan berhasil dengan sesuai memprediksi 4 dari 4 sampel data baru sesuai dengan target kelasnya. Model klasifikasi dengan nilai akurasi paling tinggi pada penelitian ini, yaitu model dengan algoritma *Random Forest Classifier*, menggunakan rasio pembagian dataset terbaik sebesar 80:20, dan parameter terbaik yang diperoleh melalui teknik *hyperparameter tuning* dengan nilai *class_weight='balanced*', nilai *max_depth*=350, nilai *min_samples_split*=5, serta nilai *n_estimators*=200, menghasilkan nilai *accuracy* sebesar 81%, nilai *precision* sebesar 82.3%, nilai *recall* sebesar 81%, dan nilai *f1-score* sebesar 81.5%, serta berhasil dengan akurat memprediksi 4 dari 4 sampel data baru berdasarkan target kelasnya.

5.2. Saran

Selama proses penelitian berlangsung, penulis menyadari adanya kekurangan serta hal yang dapat ditingkatkan lagi, adapun hal-hal tersebut penulis tuliskan sebagai saran di bawah ini:

1. Menggunakan *dataset* dengan jumlah data lebih banyak dan dengan kalimat atau kata yang lebih bervariasi. Serta pelabelan dataset secara otomatis agar mempermudah pengolahan terhadap data dengan skala yang lebih besar.
2. Mengotomatisasi proses normalisasi pengejaan atau *spelling correction* dengan metode tertentu, seperti metode Peter Norvig, N-Gram, serta dapat juga menggunakan library *SymSpell* untuk mempermudah pemrosesan data baru atau kalimat dengan kata yang lebih bervariasi.
3. Dikarenakan proses otomatisasi yang disebutkan di atas memerlukan sumber daya komputasi yang cukup besar, untuk penelitian kedepannya diperlukan sistem dengan kapasitas performa yang lebih memadai, atau penggunaan Google Colab dengan tingkat *plan* yang lebih tinggi.
4. Menambahkan variasi dari penggunaan algoritma dan parameter sebagai pembanding hasil penelitian.