BAB 5

SIMPULAN DAN SARAN

5.1 Simpulan

Berdasarkan penelitian yang telah dilakukan, diperoleh kesimpulan bahwa algoritma SMOTE dan UMAP sebagai upaya penanggulangan *imbalance high dimensional datasets* telah berhasil diimplementasikan. Algoritma SMOTE memberikan performa yang lebih baik dibandingkan dengan kombinasi algoritma UMAP-SMOTE untuk seluruh skenario yang telah dijalankan.

Implementasi UMAP-SMOTE-MLP menurunkan performa SMOTE-MLP dengan persentase rata-rata *f-measure* terkecil 27% untuk jumlah data kelas mayor 1000 dan diawali reduksi dimensi menggunakan PCA, serta persentase rata-rata *f-measure* maksimum 34% untuk jumlah data kelas mayor 1000 tanpa reduksi dimensi menggunakan PCA. Ketika dibandingkan dengan performa MLP, *resampling* menggunakan SMOTE berhasil meningkatkan nilai *f-measure* model hingga 4.8% untuk dataset dengan jumlah data kelas mayor 1000 yang diawali reduksi dimensi dengan PCA, disertai dengan nilai *precision* dan *recall* yang terbagi secara lebih merata antara data sentimen positif dan negatif. Sebaliknya, implementasi UMAP-SMOTE menurunkan nilai *f-measure* dengan persentase minimum 24.8% untuk dataset dengan jumlah data kelas mayor 1000 dan diawali reduksi dimensi menggunakan PCA.

Terdapat beberapa informasi lain yang dihasilkan penelitian ini, yakni sebagai berikut:

1. Jumlah dataset yang lebih besar meningkatkan performa model MLP dan SMOTE-MLP.

2. Reduksi menggunakan PCA tidak memberikan hasil analisa sentimen yang serupa ataupun lebih baik jika dibandingkan hasil analisa sentimen dengan dimensi asli, berbeda dengan landasan teori yang menjelaskan PCA dapat mempertahankan fitur-fitur utama dari dataset dan mengindikasikan klasifikasi yang lebih baik. Kedua hasil analisa sentiment memiliki selisih maksimal rata-rata *f-measure* sebesar 0.035.
3. Hasil *sentence embedding* dari layer *semantic* tidak berperan dalam meningkatkan performa model secara umum.
4. Tidak terdapat korelasi antara perbandingan jumlah data kelas mayor dan minor terhadap performa SMOTE ataupun UMAP-SMOTE.
5. Berdasarkan hasil visualisasi, performa UMAP tidak memiliki korelasi dengan keadaan data yang tidak seimbang ataupun jumlah data yang diberikan.

5.2 Saran

Saran yang dapat diberikan untuk penelitian selanjutnya adalah sebagai berikut.

1. Mengganti ELMo dengan *token* atau *sentence embedding* lainnya untuk meneliti pengaruh *low feature* dan *high feature input* terhadap performa UMAP. Berdasarkan penelitian yang telah dilakukan, saran ini merupakan salah satu alasan yang berpotensi memengaruhi performa UMAP-SMOTE sehingga perlu diteliti lebih lanjut. Beberapa contoh model yang dapat digunakan adalah Bag of Words dan TF-IDF yang lebih berfokus pada aspek *syntactic*.

2. Melakukan *preprocessing* yang diperlukan sebelum mengubah kalimat atau kata-kata menjadi vektor, seperti *lemmatization*, *stop-words removal*, penghapusan HTML *tag*, dan penghapusan simbol-simbol yang tidak diperlukan.
3. Menggunakan algoritma lain untuk melakukan analisa sentimen, karena berdasarkan penelitian sebelumnya, *oversampling* dengan SMOTE untuk data berdimensi tinggi memberikan performa yang lebih baik ketika dikombinasikan dengan KNN, oleh sebab itu perlu dilakukan klasifikasi analisa sentimen menggunakan beberapa model lain untuk melihat model yang memberikan performa optimal bagi kombinasi UMAP-SMOTE, kemudian membandingkan hasil performa tersebut dengan performa SMOTE saja.
4. Menambahkan *hyperparameter* dari UMAP dan SMOTE agar UMAP-SMOTE dapat mereduksi dimensi dataset dengan lebih baik dan meningkatkan performa analisa sentimen.
5. Mengganti UMAP dengan PCA untuk melakukan reduksi dimensi sebelum melakukan *resampling* dengan SMOTE.
6. Mengganti UMAP dengan *autoencoder* sesuai dengan saran yang diberikan pada *warning* ketika mereduksi dimensi dataset menjadi 5 ke atas. Pesan *warning* memberikan keterangan bahwa UMAP tidak dapat digunakan atau tidak dapat memberikan performa yang baik untuk mencari *latent space* dari data jika *middle layer* terdiri dari dimensi yang tinggi, oleh sebab itu disarankan menggunakan *autoencoder* untuk menggantikan peran UMAP, karena *autoencoder* memang dibuat untuk tujuan tersebut.