BAB 1

PENDAHULUAN

1.1 Latar Belakang Masalah

Dalam era revolusi industri 4.0 saat ini, data memegang peranan penting dalam berbagai bidang, khususnya bisnis. Hal ini disebabkan oleh kemampuan hasil pengolahan dan analisa data dalam memberikan *insight* baru yang bermanfaat dalam pengambilan keputusan (Harususilo, 2019).

Penggunaan internet dan maraknya istilah *Internet of Things* membuat jumlah data terus meningkat, sehingga menyulitkan proses pengolahan data secara manual. Kegiatan ini dapat lebih mudah dilakukan menggunakan *machine learning*. *Machine learning* dapat melakukan analisa data yang baik dengan dengan memadukan ilmu statistika, *artificial intelligence*, dan ilmu komputer (Müller dan Guido, 2016). Penerapannya dapat ditemukan dalam kehidupan sehari-hari, seperti rekomendasi otomatis, fitur pengenalan wajah, dan lain-lain.

Performa *machine learning* memiliki kaitan yang erat dengan ketersediaan dan kualitas data yang digunakan. Hal ini juga berlaku untuk seluruh penerapan *machine learning*, termasuk klasifikasi (Tsai, dkk., 2018). Model untuk melakukan klasifikasi akan memberikan performa yang lebih baik ketika label dari *target class* tersebar secara merata (Tsai, dkk., 2018).

Akan tetapi, data yang diperoleh dalam dunia nyata umumnya tidak memiliki distribusi yang merata, karena adanya kejadian atau perilaku yang jarang terjadi. Kejadian ini menyebabkan timbulnya perbedaan yang signifikan antara jumlah sampel dari kejadian tersebut (kelas minor) dengan kejadian lainnya (kelas mayor) (Longadge, dkk., 2013). Sebagai contoh, prediksi penyakit progeria sulit

menghasilkan performa yang baik karena penderita penyakit ini sangat langka, yaitu 1 : 4.000.000. Fakta ini mengakibatkan ketidakseimbangan yang sangat besar antara label dari *target class* anak yang didiagnosa positif progeria dengan anak yang didiagnosa negatif progeria. Keterbatasan data dari kelas minor menyebabkan pendeteksiannya sulit untuk dilakukan secara tepat, akibatnya performa akan menurun (Haixiang, dkk., 2017). Permasalahan yang dikenal dengan *class imbalance problem* atau *curse of imbalanced datasets* ini termasuk dalam 10 masalah yang paling banyak dialami dalam *data mining* dan *pattern recognition*, sehingga muncul urgensi untuk mengatasinya (Lemaˆıtre, dkk., 2017).

Terdapat 3 metode yang umumnya digunakan untuk mengatasi *class imbalance problem*, yakni *one-class classification*, *cost-sensitive learning*, dan *re-sampling* (Tsai, dkk., 2018). *One-class classification* bekerja dengan membuat model mempelajari data yang berasal dari kelas minoritas saja (Douzas dan Bacao, 2017). Sementara metode *cost-sensitive learning* bekerja dengan memberikan *cost* yang lebih besar apabila model salah mengklasifikasikan data minor dibandingkan salah mengklasifikasikan data mayor (Witten, dkk., 2016). Kemudian yang terakhir adalah *resampling*, yaitu metode yang memodifikasi jumlah data sehingga terbentuk distribusi yang merata (Douzas dan Bacao, 2017).

Resampling terbagi lagi menjadi 3, yaitu *under-sampling* kelas mayor, *oversampling* kelas minor, dan kombinasi dari keduanya (Longadge, dkk., 2013). *Under-sampling* dilakukan dengan mengurangi data dari kelas mayor secara acak (Longadge, dkk., 2013). Jenis ini baik untuk diterapkan pada data yang memiliki banyak observasi minor. Sebaliknya, *oversampling* dilakukan dengan membuat *artificial* atau *synthetic data* dari kelas minor (Lemaˆıtre, dkk., 2017).

Oversampling baik untuk diterapkan pada data yang memiliki sedikit observasi minor. Sementara kombinasi kedua teknik resampling tersebut baik untuk digunakan jika ukuran data latih terlalu besar (Haixiang, dkk., 2017).

Dari ketiga teknik *resampling* yang telah dijelaskan, teknik yang paling sering digunakan adalah *oversampling*. Penerapan *oversampling* dapat dilakukan dengan membuat data secara *random* atau menggunakan Synthetic Minority Over-sampling Technique (SMOTE) (Haixiang, dkk., 2017).

SMOTE digunakan untuk meningkatkan jumlah sampel dari kelas minor dengan membuat data sintesis baru berdasarkan interpolasi antar data dari kelas minor yang berdekatan (Fernández, dkk., 2018). Algoritma ini telah diimplementasikan dalam berbagai persoalan terkait *imbalanced data* dan terbukti dapat meningkatkan performa model klasifikasi (Fernández, dkk., 2018). Salah satu implementasinya yaitu dalam segmentasi gambar (Abeysinghe, dkk., 2018).

Penelitian yang dilakukan Wajira Abeysinghe, dkk. (2018) tersebut membandingkan 3 pendekatan dalam melakukan segmentasi gambar yaitu teknik tradisional *oversampling*, teknik tradisional *undersampling*, dan SMOTE. Hasil penelitian tersebut membuktikan penggunaan SMOTE meningkatkan akurasi paling optimal di antara ketiga teknik yang digunakan (Abeysinghe, dkk., 2018).

Di samping segmentasi gambar, SMOTE juga banyak diimplementasikan dalam analisis sentimen. Berdasarkan penelitian yang dilakukan oleh Widi Satriaji dan Retno Kusumaningrum (2018), implementasi SMOTE terbukti meningkatkan performa dalam melakukan analisis sentimen terhadap data tidak seimbang. Kombinasi SMOTE dengan Term Occurrence dan Logistic Regression dalam penelitian tersebut berhasil meningkatkan performa sebesar 12% (Satriaji dan

Kusumaningrum, 2018). Penelitian-penelitian terdahulu tersebut membuktikan bahwa SMOTE dapat mengurangi akibat dari *imbalance dataset* dan berhasil meningkatkan performa model.

Di samping masalah ketidakseimbangan data, data yang tersedia saat ini pada umumnya memiliki dimensi yang tinggi, misalnya dalam bidang *biomedical* (Blagus dan Lusa, 2013). SMOTE dalam hal ini memiliki performa yang rendah ketika diimplementasikan untuk data berdimensi tinggi. Dalam penelitian yang dilakukan oleh Rok Blagus dan Lara Lusa (2013), dibuktikan bahwa *random under-sampling* lebih efektif daripada penggunaan SMOTE dalam data berdimensi tinggi. Implementasi SMOTE dalam penelitian tersebut tidak memberikan perubahan apapun dalam performa model, kecuali untuk model KNN (Blagus dan Lusa, 2013).

Walaupun demikian, ketika mengkombinasikan SMOTE dengan algoritma untuk mengurangi dimensi, yakni Principal Component Analysis (PCA). Mehdi Naseriparsa dan Mohammad Mansour Riahi Kashani (2013) menyimpulkan tingkat performa yang dihasilkan untuk melakukan klasifikasi gambar menjadi lebih baik. Oleh sebab itu, penelitian ini menggunakan algoritma untuk mengurangi dimensi.

Terdapat beberapa teknik lain yang biasanya digunakan untuk mengurangi dimensi, yaitu t-Distributed Stochastic Neighbor Embedding (t-SNE), dan Uniform Manifold Approximation and Projection (UMAP) (Becht, dkk., 2018). PCA menggunakan teknik linear yang fokus dalam merepresentasikan data yang berbeda secara berjauhan (Maaten dan Hinton, 2008). Sebaliknya, t-SNE dan UMAP menggunakan teknik non-linear sehingga dapat merepresentasikan data yang mirip secara berdekatan dengan lebih baik dibandingkan PCA (Maaten dan Hinton, 2008). Dalam konteks visualisasi, UMAP tetap memperhitungkan struktur lokal

data seperti t-SNE (McInnes, dkk., 2018). UMAP juga terbukti membutuhkan waktu yang lebih singkat, memperhitungkan struktur data global lebih baik, dan dapat memvisualisasikan jumlah data yang lebih banyak dibandingkan t-SNE (Becht, dkk., 2018). Selain itu, UMAP dapat melakukan *mapping* untuk data baru, sehingga jika terdapat data baru yang diberikan kepada model, data dapat diubah secara langsung ke *space* yang telah dibuat model UMAP sebelumnya tanpa harus membuat *space* baru (McInnes, 2020). Atas beberapa alasan tersebut, reduksi dimensi dalam penelitian ini dilakukan menggunakan UMAP.

Berdasarkan data yang diperoleh dan analisa hasil penelitian terdahulu, maka penelitian ini mengimplementasikan SMOTE dan UMAP sebagai upaya penanggulangan *imbalance high dimensional datasets*. Kegiatan yang dilakukan adalah analisis sentimen menggunakan Embedding from Language Model (ELMo) dengan Multilayer Perceptron (MLP) sebagai model untuk melakukan klasifikasi. ELMo merupakan *contextualized word embeddings*, yaitu model yang memperhitungkan konteks kata ketika mengubah kalimat ke dalam bentuk vektor, sehingga hasilnya lebih akurat dan relevan (Reimers, dkk., 2019). Berdasarkan penelitian Matthew E. Peters, dkk. (2018), penggunaan *pre-trained* model ELMo terbukti dapat meningkatkan akurasi dalam berbagai kegiatan NLP, termasuk analisis sentimen. Model ini menerima *input* berupa kalimat dan *output* berupa hasil vektor perhitungan setiap kata dalam bentuk 1024 dimensi, sehingga penggunaannya sesuai dengan tujuan penelitian ini.

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah yang telah dijelaskan, beberapa rumusan masalah yang dibahas dalam laporan ini adalah sebagai berikut:

1. Bagaimana cara mengimplementasi algoritma SMOTE dan UMAP sebagai upaya penanggulangan *imbalance high dimensional datasets*?
2. Bagaimana perbandingan performa kombinasi SMOTE-UMAP dengan SMOTE dalam menanggulangi masalah *imbalance high dimensional datasets*?

1.3 Batasan Masalah

Batasan masalah dalam penelitian ini adalah sebagai berikut:

1. Penelitian hanya diterapkan dalam melakukan analisis sentimen, sehingga tidak mencakup seluruh cabang dari NLP. Bidang ini dipilih karena implementasi SMOTE telah terbukti baik untuk analisis sentimen dengan data berdimensi rendah.
2. Klasifikasi untuk analisis sentimen terbatas pada 2 *target class*, yaitu positif (1) dan negatif (0).
3. Penelitian menggunakan *pre-trained* ELMo berbahasa Inggris, maka tidak dilakukan *fine tuning* lagi untuk melakukan ekstraksi fitur.
4. Dataset yang digunakan adalah IMDB Dataset of 50k Movie Reviews yang tersedia di Kaggle dalam Bahasa Inggris dengan jumlah sentimen positif dan negatif masing-masing 25.000. Dataset tersebut diperoleh dari penelitian terdahulu yang dilakukan Universitas Stanford dengan judul Learning Word Vectors for Sentiment Analysis (Maas, dkk., 2011).

5. Algoritma yang digunakan untuk melakukan klasifikasi adalah Multilayer Perceptron (MLP).

1.4 Tujuan Penelitian

Tujuan dilakukannya penelitian ini adalah sebagai berikut:

1. Mengimplementasi algoritma SMOTE dan UMAP sebagai upaya penanggulangan *imbalance high dimensional datasets.*
2. Mengetahui perbandingan performa kombinasi algoritma SMOTE-UMAP dengan SMOTE sebagai upaya penanggulangan *imbalance high dimensional datasets*.

1.5 Manfaat Penelitian

Penelitian ini dilaksanakan untuk mengukur performa dalam bentuk *precision*, *recall*, dan *f-measure* dari kombinasi data asli dengan data sintesis berdimensi tinggi yang dihasilkan oleh kombinasi metode SMOTE dan UMAP sebagai upaya penanggulangan masalah *imbalance high dimensional datasets*, serta performa UMAP dalam mengurangi dimensi *text features*.

1.6 Sistematika Penulisan

Sistematika penulisan yang diterapkan dalam penyusunan skripsi ini dibagi menjadi 5 bab utama, yang dijabarkan sebagai berikut:

BAB I PENDAHULUAN

Bab pertama menjelaskan latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan sistematika penulisan laporan.

BAB II LANDASAN TEORI

Bab kedua menjabarkan teori-teori yang mendasari penelitian ini, antara lain representasi teks, ELMo, SMOTE, UMAP, MLP, PCA, Grid Search, dan metrik evaluasi.

BAB III METODOLOGI PENELITIAN

Bab ini berisi tahapan metode penelitian yang dilakukan disertai dengan diagram dalam bentuk *flowchart*.

BAB IV HASIL DAN DISKUSI

Bab keempat membahas implementasi kode, skenario yang diuji, hasil pengujian, serta evaluasi terhadap hasil yang diperoleh.

BAB V SIMPULAN DAN SARAN

Bab terakhir berisi simpulan yang menjawab tujuan penelitian dan saran untuk pengembangan ataupun penelitian yang bersangkutan di kemudian hari.