BAB I
PENDAHULUAN

1.1 Latar Belakang

Perkembangan internet sangat berperan dalam memfasilitasi kemajuan *electronic Word-of-Mount* (e-WOM). e-WOM merupakan suatu media komunikasi yang berisi kumpulan opini (berbentuk ulasan, forum, maupun mikro blog) terkait suatu entitas tertentu seperti produk, jasa, isu sosial bahkan tokoh (Li dkk., 2018). Di dunia bisnis, keberadaan e-WOM memberikan kemudahan bagi penggunanya dalam memperoleh wawasan terkait entitas tertentu serta dapat membantu pengguna dalam proses pengambilan keputusan yang lebih baik (Prayustika, 2017). Sebelum memutuskan untuk membeli suatu barang, selain melihat deskripsi pada produk, biasanya konsumen akan melihat review konsumen lain terhadap produk tersebut. Bagi pelaku bisnis, e-WOM berupa ulasan dapat membantu mereka dalam melakukan riset pasar baik terkait produk maupun terkait karakteristik kebiasaan konsumen.

Sebuah ulasan biasanya berupa deskripsi teks yang juga direpresentasikan dalam bentuk skala peringkat 1 sampai 5 untuk memudahkan pengguna dalam menerjemahkan ulasan. Namun, perlu disadari bahwa mengandalkan peringkat sebagai acuan dalam menentukan sentimen tidaklah cukup sebab seringkali terjadinya bias di antara keduanya (Tama dkk., 2019). Mengingat hal tersebut, maka diperlukan adanya suatu teknik analisis sentimen untuk mengidentifikasi dan mengklasifikasi ulasan-ulasan tersebut apakah termasuk dalam kategori positif atau negatif. Analisis sentimen suatu proses mengekstrak dan mengolah data tekstual untuk memperoleh informasi sentimen di dalam sebuah opini secara otomatis (Sharma dan Dutta, 2021; Li dkk., 2018).

Klasifikasi sentimen merupakan salah satu tugas dalam teknik analisis sentimen. Klasifikasi sentimen terdiri dari tiga tingkatan yakni tingkat dokumen, tingkat kalimat, hingga tingkat aspek (Liu, 2015). Di dalam penelitian ini, analisis tingkat dokumen akan dibahas. Analisis tingkat dokumen akan digunakan pada tipe dataset review produk pada penelitian ini. Klasifikasi sentimen pada tingkat

dokumen membahas mengenai bagaimana menampilkan sebuah sentimen pada seluruh dokumen menggunakan model yang tepat. Terdapat berbagai jenis pendekatan dalam klasifikasi sentimen antara lain pendekatan *lexicon-based, learning-based,* dan *hybrid* (Chiarello et al., 2020). Pendekatan *lexicon-based* menggunakan instrument kamus kata yang telah diberi label berdasarkan nilai polaritas. Pendekatan sederhana ini memiliki performa yang cepat pada kumpulan data. Hanya saja, *lexicon-based* memiliki kelemahan dalam tingkat akurasi, presisi dan sensitivitas yang rendah sedangkan pendekatan pembelajaran mesin memiliki performa yang lebih baik (Mumtaz dan Ahuja, 2018). Dari ketiga pendekatan tersebut, penelitian terkait dengan pendekatan pembelajaran mesin juga telah banyak dibahas dalam literatur (Parikh dan Shah, 2020; Ruz dkk., 2020).

Satu hal yang juga patut menjadi perhatian dalam melakukan klasifikasi sentimen yakni pelabelan. Tahapan ini dilakukan ketika data sentimen atau ulasan yang ditambang belum memiliki kelas label. Adapun metode pelabelan yang biasa digunakan antara lain pelabelan secara manual dan pelabelan secara otomatis. Pelabelan manual mampu meberikan hasil yang lebih akurat sebab teknik ini dilakukan dengan adanya bantuan dari ahli bahasa (Rachmat dan Lukito, 2016). Namun metode tersebut akan sangat banyak memakan waktu jika dilakukan pada kumpulan data yang sangat banyak. Penggunaan metode lain yakni pelabelan secara otomatis ialah pelabelan yang dilakukan menggunakan referensi *binary labeling* (angka 1 untuk positif, angka 0 untuk negatif) atau *average labeling* yakni dengan membagi label berdasarkan rating (Tama dkk., 2019).

Berikut beberapa penelitian terdahulu terkait klasifikasi sentimen dengan pendekatan pembelajaran mesin yang telah dilakukan. Shaheen (2019) dalam penelitiannya melakukan analisis sentimen pada platform ulasan di situs Amazon.com. Teknik yang digunakan dalam membangun model antara lain melakukan *exploratory analysis* serta menerapkan pendekatan pembelajaran mesin. Hasil dari studi ini menyatakan bahwa, *Random forest* menjadi algoritma yang mengungguli 7 algoritma lainnya. Xu, dkk (2020) menggunakan algoritma *Continuous Naïve Bayes* pada klasifikasi sentimen ulasan produk pada *ecommerce* Amazon. Model yang dihasilkan berupa model klasifikasi yang mampu

mempelajari pengetahuan dari domain sebelumnya dan menerapkan model tersebut pada domain yang baru dan berbeda (F. Xu dkk., 2020). Neogi, dkk (2021) di dalam penelitiannya melakukan teknik analisis sentimen pada kumpulan data protes para petani India di sosial media Twitter. Sebanyak 20.000 tweet, studi ini menganalisis data dengan menggunakan metode *Bag of Word* dan TF-IDF serta melakukan klasifikasi menggunakan empat algoritma di mana *Random forest* keluar sebagai algoritma yang menghasilkan akurasi tertinggi dari algoritma yang lainnya (Neogi dkk., 2021).

Dari beberapa algoritma pembelajaran mesin yang telah banyak diterapkan pada penelitian sebelumnya seperti *Naïve bayes*, SVM, SGD *classifier*, *Random forest* untuk klasifikasi sentimen, Pada penelitian ini, jenis pendekatan klasifikasi sentimen yang digunakan adalah pendekatan pembelajaran mesin. *Random Forest* (RF) menjadi salah salah satu algoritma yang memiliki performa terbaik untuk membangun model klasifikasi (Parmar dkk., 2019). Algoritma RF merupakan hasil dari pengembangan algoritma *Decision Tree*. Di mana RF memiliki keunggulan antara lain algoritma ini memiliki performa akurasi yang lebih baik serta mampu menghadapi *overfitting* (Bahwari, 2019). Selain itu, RF memiliki skalabilitas yang baik untuk kumpulan sampel data yang besar karena strukturnya yang menyerupai pohon (Al Ajrawi dkk., 2021).

Meskipun klasifikasi sentimen dengan pendekatan pembelajaran mesin terbukti dalam segi performanya, namun sebuah model klasifikasi tidak lepas dari masalah *imbalanced class*. Seiring berjalannya waktu, pembahasan terkait kelas data tidak seimbang juga dipelajari pada klasifikasi sentimen (Li dkk., 2018). Kasus kelas data tidak seimbang akan sangat mempengaruhi kinerja model klasifikasi yang akan dibangun. Hal tersebut disebabkan akibat dominasi kelas mayor yang dihasilkan menjadikan model lebih memprioritaskan pembelajaran pada kelas mayor saja (Mustaqim dkk., 2019).

Rumusan masalah dalam penelitian ini adalah mengetahui seberapa baik tingkat performa model *Random Forest* dan implementasi teknik kelas *imbalanced* pada sistem klasifikasi ulasan konsumen. Berdasarkan hal tersebut, maka dibangun sebuah sistem klasifikasi ulasan konsumen dengan menerapkan teknik klasifikasi

sentimen tingkat dokumen. Penelitian ini juga menghasilkan kebaharuan terkait performa model klasifikasi sentimen yang menggunakan teknik kelas *imbalanced*. Pendekatan klasifikasi sentimen yang digunakan menggunakan algoritma *Random Forest* sebagai metode klasifikasinya. Metode kelas data tidak seimbang yakni *Synthetic Minority Oversampling Technique* (SMOTE) dan *Random Under-sampling* (RUS) juga ditambahkan untuk menghadapi kelas data tidak seimbang serta meningkatkan kinerja pada *pre-processing* model pelatihan sistem klasifikasi ulasan.

1.2 Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk menerapkan salah satu teknik analisis sentimen yakni klasifikasi sentimen dengan pendekatan pembelajaran mesin terawasi menggunakan algoritma *Random forest classifier* dalam membangun sistem klasifikasi ulasan konsumen pada keseluruhan dokumen secara otomatis. Selain itu, *Synthetic minority oversampling technique* (SMOTE) dengan kolaborasi *Random Undersampling* (RUS) diterapkan untuk meningkatkan performa klasifikasi pada model klasifikasi menggunakan *Random forest.*

1.3 Manfaat Penelitian

Manfaat dari penelitian ini adalah dapat memperoleh model klasifikasi sentimen terbaik dari pendekatan *Machine learning* menggunakan algoritma *Random forest* dan kolaborasi teknik *Synthetic minority oversampling technique* (SMOTE) dan *Random Under-sampling* (RUS) pada kasus klasifikasi ulasan. Sementara dari sisi manfaat bagi pengguna, sistem dari hasil penelitian ini dapat membantu dalam mengetahui sentimen pada kumpulan ulasan konsumen.