BAB I
PENDAHULUAN

1.1 Latar Belakang

Teknologi merupakan salah satu faktor dalam perkembangan berbagai industri saat ini serta menjadi satu kebutuhan kepada kehidupan sehari-hari. Teknologi membuat berbagai hal berkembang secara cepat dan efisien sehingga meningkatkan produktivitas. Selain itu perkembangan teknologi juga mempengaruhi manusia dalam melakukan transaksi. Teknologi memunculkan inovasi *financial technology (fintech)* yang memungkinkan untuk melakukan kegiatan transaksi yang cepat dan efisien dengan memanfaatkan teknologi[1].

Perkembangan teknologi finansial sangat pesat di Indonesia seiring dengan penerimaan pembayaran digital di masyarakat. Pada tahun 2021, Indonesia memiliki peningkatan perusahaan teknologi finansial sebesar 32.5% dibandingkan dengan tahun sebelumnya[2]. Sepanjang tahun 2022 nilai transaksi uang elektronik tumbuh mencapai 42.06% pada triwulan pertama 2022[3]. Selain itu statistik menunjukkan bahwa sektor finansial teknologi di Indonesia memiliki nilai *Compounded Annual Growth Rate* sebesar 39%[4].

Pesatnya perkembangan teknologi finansial juga meningkatkan risiko pencucian uang dan pendanaan terorisme terutama dalam industri ini. Sifat teknologi finansial yang cepat dan dinamis membuat teknologi finansial sebagai salah satu media untuk melakukan pencucian uang dan pendanaan terorisme[5]. Hal ini menyebabkan perusahaan teknologi finansial diwajibkan dalam melakukan identifikasi, verifikasi dan analisa risiko pengguna dalam penerapan program Anti-Pencucian Uang dan Pencegahan Pendanaan Terorisme (APU-PPT)[6].

PT Kredibel Teknologi Indonesia (Kredibel) adalah perusahaan *startup* yang berfokus pada penyelesaian masalah penipuan dan kepercayaan pelanggan[7]. Perusahaan ini berfokus untuk membantu mengurangi tingkat penipuan yang ada di Indonesia. Saat ini Kredibel memiliki beberapa layanan seperti *fraud detection*

system, know your customer, dan *customer due diligence*[8]. Dalam meningkatkan layanan *customer due diligence*, Kredibel membutuhkan model *machine learning* yang digunakan untuk melakukan klasifikasi data dari media terutama dari media online yang ada di Indonesia. Proses klasifikasi data dari media akan memakan waktu lebih lama jika dilakukan secara manual dan menggunakan tenaga manusia. Oleh karena itu, dibutuhkan adanya pemilahan data dalam melakukan klasifikasi judul berita masuk ke dalam *adverse media* atau bukan *adverse media*. Setelah data diklasifikasi, Kredibel dapat menggunakan data tersebut untuk melakukan analisis seseorang maupun organisasi yang memiliki rekam jejak kriminalitas. Model *adverse media* yang dibuat akan digunakan untuk melakukan pemilahan data sebelum diproses dalam sistem *customer due diligence*. Adapun data yang digunakan dalam pembuatan model ini berasal dari *scraping* pada situs berita dan menggunakan NewsAPI sebagai sumber data dari model yang akan dibuat.

Terdapat beberapa penelitian sebelumnya dalam melakukan klasifikasi teks yang berhubungan dengan data berita dengan *dataset* AG News. Algoritma pertama dalam klasifikasi teks berita adalah XLNet oleh Zhilin Yang et al dengan nilai *error* sebesar 4.45[9]. Penelitian ini dilakukan oleh Carnegie Mellon University dan Google AI Brain Team dengan tingkat akurasi sebesar 85.4%. Penelitian lainnya dalam klasifikasi teks adalah penelitian dengan menggunakan BERT-ITPT-FiT yang dilakukan oleh Chi Sun et al dari Fudan University dengan tingkat *error* sebesar 4.8[10]. Penelitian berikutnya adalah dengan menggunakan LSTM dengan *Mixed Object Function* dengan tingkat *error* sebesar 4.95%[11]. Penelitian ini dilakukan oleh Devendra Singh Sachan dari Petuum Inc, Manzil Zaheer dari Google Research, dan Ruslan Salakhutdinov dari Carnegie Mellon University. Penelitian lainnya yaitu dengan menggunakan ULMFiT yang dilakukan oleh Jeremy Howard dari fast.ai dan Sebastian Ruder dari NUI Galway Aylien Ltd[12]. Penelitian ini menghasilkan model dengan tingkat *error* sebesar 5.01.

Terdapat penelitian terdahulu dalam proses klasifikasi teks menggunakan *machine learning*. Penelitian ini menggunakan algoritma Random Forest dengan Gradient Boosting[13] dalam melakukan proses kategori keluhan secara otomatis. Penelitian ini berhasil menerapkan Random Forest dan Gradient Boosting dalam proses klasifikasi teks multi-kelas dengan akurasi sebesar 73%. Penelitian ini memiliki keterbatasan yaitu sebagian besar data diklasifikasikan pada satu kategori yang disebabkan oleh kata yang berada di beberapa kategori.

Penelitian terdahulu lainnya menggunakan Decision Tree dan XGBoost dalam melakukan klasifikasi sentimen vaksin COVID-19 [14]. Penelitian ini menghasilkan model XGBoost dengan tingkat akurasi sebesar 66% dan Decision Tree dengan akurasi sebesar 65%. Algoritma XGBoost lebih unggul dibandingkan dengan Decision Tree pada *dataset* yang tidak seimbang. Penelitian ini memiliki keterbatasan yaitu *dataset* yang tidak seimbang. Penelitian ini masih memiliki akurasi dan *f1-score* yang rendah yaitu sebesar 66%.

Penelitian selanjutnya adalah melakukan klasifikasi berita palsu dengan menggunakan algoritma XGBoost [15]. Penelitian ini memiliki tingkat akurasi sebesar 92% dengan perbandingan dataset sebesar 80:20. Tingkat akurasi ini didapatkan berdasarkan *parameter tuning* yang dilakukan. Penelitian ini memiliki keterbatasan dalam *dataset* yaitu hanya sebesar 100 data yang dibagi secara acak. Penelitian ini mungkin hanya memiliki akurasi terbaik pada *dataset* tertentu dan tidak menghasilkan tingkat akurasi yang sama pada *dataset* yang berbeda.

Penelitian lainnya yaitu memprediksi gagal ginjal kronis dengan membandingkan metode Grid Search dan Random Search dalam *hyperparameter tuning* di algoritma XGBoost[16]. Penelitian ini memiliki tingkat akurasi sebesar 99.29% dan *f-measure* sebesar 0.99. XGBoost dengan *hyperparameter tuning*, normalisasi *z-score*, dan *random oversampling* memiliki peran penting dalam menghasilkan akurasi tersebut. Penelitian ini memiliki kekurangan yaitu proses pencarian *hyperparameter* yang tepat memakan waktu yang lama dengan

menggunakan Grid Search. Selain itu penelitian ini belum diuji pada jenis *dataset* yang berbeda sehingga tingkat *generalisasi* model tidak diketahui.

Penelitian lainnya yaitu melakukan prediksi dan diagnosis risiko diabetes di masa depan dengan menggunakan pendekatan *machine learning*. Penelitian ini menghasilkan model dengan tingkat akurasi sebesar 85% dengan menggunakan Gradient Boosting, 77% pada Naïve Bayes, dan 79% pada Logistic Regression. Penelitian ini memiliki tingkat akurasi prediksi yang tinggi serta menggunakan algoritma yang merupakan teknik yang sangat baik untuk masalah regresi seperti Gradient Boosting. Penelitian ini memiliki beberapa keterbatasan seperti data yang digunakan tidak mencakup populasi. Penelitian ini hanya menggunakan BMI dan plasma glukosa yang penting dalam memprediksi diabetes. Penelitian ini memiliki potensi pada *overfitting* karena tidak mencakup penanganan pada potensi *overfitting* yang terjadi pada model [17].

Berdasarkan hasil penelitian sebelumnya dengan menggunakan algoritma *machine learning* seperti Random Forest, *convolutional neural network, complement, multinomial naïve bayes,* Linear Regression dan XGBoost. Saat ini belum ada penelitian dalam pembuatan model untuk *adverse media* terutama dalam Bahasa Indonesia. Sebagian besar penelitian mengenai *customer due diligence* dan *adverse media* dilakukan secara tertutup. Selain itu, penelitian ini menggunakan *dataset* yang tidak tersedia secara publik dan memerlukan teknik *scraping* dan pemanggilan API dalam pengambilan datanya. Berdasarkan hal tersebut, penelitian ini akan berfokus dalam pembuatan model *machine learning* yang digunakan untuk melakukan klasifikasi sebuah berita terutama media daring apakah sebuah konten dari berita tersebut masuk ke dalam *adverse media* atau bukan *adverse media*. Penelitian ini juga melakukan percobaan dengan menggunakan TF-IDF seperti pada penelitian [18] serta menggunakan CountVectorizer dan Tokenizer yang umumnya digunakan dalam *feature extraction* data teks. Penggunaan Grid Search dalam penelitian ini bertujuan untuk menentukan *parameter* terbaik dengan lebih efisien. Grid Search memfasilitasi penemuan *hyperparameter* optimal pada algoritma yang digunakan. *Parameter*

akan mempengaruhi tingkat akurasi sesuai yang diungkapkan dalam penelitian [15], [16].

Penelitian ini menggunakan XGBoost yang memiliki kelebihan dapat menangani data yang berukuran besar dengan cepat dibandingkan algoritma lainnya. Selain itu algoritma ini juga dapat menangani *missing value* dan data *imbalance* dengan baik. Algoritma XGBoost juga memiliki kekurangan yaitu dalam proses melatih model, algoritma ini menghabiskan waktu lebih lama dibandingkan dengan algoritma lain. Selain itu juga perlu pemahaman lebih mendalam pada parameter yang ada di XGBoost untuk mendapatkan hasil yang optimal[19].

Selain menggunakan XGBoost, penelitian ini juga menggunakan Gradient Boosting dalam melakukan klasifikasi *adverse media* dan bukan *adverse media*. XGBoost adalah implementasi dari *gradient boosting machine* (*gbm)* dengan memaksimalkan sumber daya dan mengatasi keterbatasan *gradient boosting* sebelumnya[20]. Gradient Boosting pada penelitian ini untuk mengetahui apakah XGBoost lebih unggul dibandingkan pendahulunya yaitu Gradient Boosting. Gradient Boosting memiliki beberapa kelebihan seperti menggunakan pendekatan *ensemble* dalam melakukan peningkatan hasil prediksi. Selain itu Gradient Boosting memiliki kelemahan yaitu dalam proses *training* memakan waktu lebih panjang dibandingkan dengan menggunakan algoritma lain[17].

Hasil akhir dari penelitian ini adalah sebuah model yang digunakan untuk melakukan klasifikasi berita apakah masuk ke dalam *adverse media* atau bukan *adverse media.* Model yang dibuat diharapkan dapat membantu Kredibel dalam mengolah sumber data dari *customer due diligence* dengan hasil yang lebih akurat. Selain itu penelitian ini menjadi solusi keterbatasan penelitian sebelumnya dalam proses klasifikasi *adverse media* terutama dalam Bahasa Indonesia.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang sudah dijelaskan sebelumnya berikut ini adalah rumusan masalah dalam penelitian ini.

1. Bagaimana membuat model klasifikasi *adverse media* dengan algoritma XGBoost dan Gradient Boosting pada media daring?
2. Bagaimana efektivitas berdasarkan akurasi, *precision*, dan *recall* dari model yang dibuat untuk melakukan klasifikasi berita yang merupakan *adverse media* dan bukan *adverse media*?
3. Bagaimana mengimplementasikan model terbaik dari XGBoost dan Gradient Boosting klasifikasi *adverse media*?

1.3 Batasan Masalah

Dalam penelitian ini terdapat beberapa batasan masalah seperti pada point berikut ini.

1. *Dataset* yang digunakan dalam proses pengambilan data melalui NewsAPI terbatas hanya pada media Kompas, Detik, Tribunnews dan Merdeka.
2. *Dataset* yang digunakan dengan proses *scraping* memiliki keterbatasan karena setiap situs memiliki struktur situs yang beragam.
3. Penelitian ini hanya terbatas dalam pembuatan model dan tidak dijelaskan tahapan penggunaan pada *customer due diligence*.

1.4 Tujuan dan Manfaat Penelitian

1.4.1 Tujuan Penelitian

Berikut ini adalah tujuan penelitian yang dilakukan.

1. Menghasilkan model klasifikasi *adverse media* dari media daring dengan menggunakan algoritma XGBoost dan Gradient Boosting.
2. Mengevaluasi kinerja model dengan algoritma XGBoost dan Gradient Boosting dalam klasifikasi *adverse media*.
3. Menghasilkan *prototype* aplikasi website untuk klasifikasi *adverse media* dari media daring.

1.4.2 Manfaat Penelitian

Berikut ini adalah beberapa manfaat penelitian yang dilakukan.

1. Membantu PT Kredibel Teknologi Indonesia untuk memilih media yang akan digunakan dalam proses analisa *AML/CTF(Anti-Money Laundering and Counter-Terrorism Financing*) dengan model klasifikasi *adverse media* dan bukan *adverse media*.
2. Menambah pengetahuan mengenai klasifikasi teks *adverse* dengan menggunakan algoritma XGBoost dan Gradient Boosting.

1.5 Sistematika Penulisan

BAB 1 PENDAHULUAN

Dalam pendahuluan terdapat bagian latar belakang, rumusan masalah, batasan masalah, tujuan dan manfaat penelitian serta sistematika penulisan.

BAB 2 LANDASAN TEORI

Bagian ini menjelaskan mengenai teori yang digunakan sesuai dengan topik penelitian yang digunakan. Selain itu penelitian terdahulu juga terdapat pada bagian ini untuk referensi dalam penelitian ini.

BAB 3 METODOLOGI PENELITIAN

Bagian ini menjelaskan langkah dalam proses penelitian ini, dimulai dari gambaran umum dari objek penelitian, metode penelitian, teknik pengumpulan data dan teknik analisis data.

BAB 4 ANALISIS DAN HASIL PENELITIAN

Bagian ini menjabarkan mengenai implementasi dari metodologi yang ditentukan. Bab ini terdiri dari proses pengumpulan data, pemrosesan

data, seleksi data, analisis data, dan hasil dari penelitian yang diperoleh.

BAB 5 SIMPULAN DAN SARAN

Bagian ini menjelaskan menjelaskan mengenai kesimpulan dan saran dari penelitian yang dilakukan. Kesimpulan dan saran diharapkan dapat membantu penelitian berikutnya.