1. Pendahuluan

1.1 Latar Belakang

Tingkat penggunaan media sosial berbanding lurus dengan perkembangan teknologi. Saat ini banyak manusia menggunakan media sosial untuk memenuhi kebutuhan sosialnya. Salah satunya adalah Twitter dengan total pengguna aktif di Indonesia sebanyak 11.2 juta [1]. Twitter memungkinkan penggunanya untuk beropini dan bercerita yang disebut dengan cuitan. Cuitan yang dikumpulkan dapat digunakan untuk merepresentasikan kepribadian dengan pendekatan pembelajaran mesin [2].

Kepribadian adalah status karakteristik dan kebiasaan manusia dalam kesehariannya. Mendeteksi kepribadian pada twitter menjadi penting karena data yang dibutuhkan untuk analisis banyak dan terbuka, serta proses pendeteksian kepribadian tradisional seperti wawancara dengan psikolog membutuhkan biaya dan waktu yang besar [3]. Hasil dari pendeteksian kepribadian dapat digunakan untuk meningkatkan akurasi sistem rekomendasi, serta mempermudah perekrut dalam menempatkan posisi kerja yang sesuai. Kepribadian dapat dikelompokkan menjadi beberapa model seperti *Big Five*, *Myers Briggs Type Indicator* (MBTI), STIFin [4][5][6]. Model pengelompokan kepribadian yang digunakan dalam penelitian ini adalah *Big Five*. Menurut penelitian Michael [4], kepribadian Big Five dipelajari dengan baik dan memiliki hubungan erat terhadap kehidupan dan kepuasan bekerja. Hal tersebut memberikan manfaat untuk perekrut menempatkan posisi pekerja yang tepat sesuai kepribadian Big Five.

Prediksi kepribadian dengan pendekatan pembelajaran mesin terbagi menjadi 3 kelompok, yaitu *Supervised Learning*, *Semi-Supervised Learning*, dan *Unsupervised Learning*. Nadia [7] menjelaskan kelebihan pendekatan Semi-Supervised Learning adalah dapat menggunakan data yang tidak berlabel, tetapi pendekatan *Semi-Supervised Learning* didesain untuk klasifikasi 2 kelas. Berdasarkan penjelasan Nadia, Penelitian ini tidak bisa menggunakan pendekatan *Semi-Supervised Learning* karena pada penelitian ini akan mengklasifikasikan ke dalam 5 kelas kepribadian *Big Five* (*Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness*, *Neuroticism*). Ritu [8] menjelaskan kelebihan *Unsupervised Learning* adalah dapat menggunakan data yang tidak berlabel dan melakukan analisis *real-time*, tetapi performansi yang dihasilkan *Unsupervised Learning* hanya cukup dibandingkan dengan *Supervised Learning* meskipun membutuhkan data berlabel. Maka dari itu, penelitian ini akan berfokus pada pendekatan *Supervised Learning*.

Terdapat banyak penelitian sebelumnya yang membahas tentang prediksi kepribadian di media sosial dengan pendekatan *Supervised Learning*. Monica [9] menggunakan metode *Naïve Bayes* terhadap ekstraksi fitur emosi, fitur sentimen, dan fitur sosial. Penelitian tersebut menghasilkan rata-rata akurasi tertinggi pada fitur sentimen sebesar 38.95%. Bayu [3] menghasilkan rata-rata akurasi tertinggi sebesar 60% dengan metode Naïve Bayes dibandingkan dengan metode KNN, dan SVM. Alaa [5] menggunakan metode *Naïve Bayes*, KNN, dan *Random Forest* untuk prediksi kepribadian MBTI. Metode *Random Forest* menghasilkan kinerja lebih baik dibandingkan dengan *Naïve Bayes* dan KNN dengan akurasi 100%. *Sensitivity* 92.29%. *Specificity* 19%, dan *Precision* 64.35%. Berdasarkan penelitian sebelumnya metode *Naïve Bayes* memiliki kekurangan yaitu memproses setiap fiturnya secara independen [9]. Metode KNN memiliki kekurangan yaitu sulit dalam menentukan nilai K [3]. Metode SVM memiliki kekurangan yaitu sulit dalam memisahkan kelas kata dengan data yang kurang akurat [3]. Metode *Random Forest* mendapatkan akurasi tertinggi untuk prediksi kepribadian MBTI karena dapat menangani data yang kurang akurat dengan baik [5][10][11]. Metode *Random Forest* memiliki banyak parameter yang bisa diatur sesuai dengan studi kasus. Contoh parameter pada *Random Forest* adalah "*n_estimator*" untuk menentukan jumlah *tree* yang dibangun [12]. Banyaknya parameter pada *Random Forest* juga menjadi kekurangannya yaitu sulit dalam mengatur parameter yang benar [10][13]. Data yang digunakan pada penelitian ini masih kurang akurat dan memiliki banyak fitur yang digunakan untuk memprediksi kepribadian. Oleh karena itu, penelitian ini membangun sistem prediksi tipe Kepribadian *Big Five* menggunakan metode *Random Forest*.

1.2 Topik dan Batasannya

Berdasarkan latar belakang, penelitian ini akan membangun sistem prediksi tipe kepribadian *Big Five* dengan memanfaatkan data twitter berupa cuitan dan informasi profil sebagai data masukan, Data cuitan akan digunakan sebagai fitur sentimen dan fitur emosi. Data informasi akun digunakan sebagai fitur sosial. Kemudian sistem akan melakukan klasifikasi data masukan menggunakan metode *Random Forest*. Adapun rumusan masalah dalam penelitian ini adalah bagaimana performansi sistem prediksi Tipe Kepribadian *Big Five* menggunakan metode *Random Forest*? Bagaimana memaksimalkan parameter pada metode R*andom Forest*? Bagaimana korelasi antara setiap fitur dengan setiap kepribadian *Big Five*?

Karena keterbatasan sumber daya, penelitian ini terdapat beberapa batasan masalah antara lain, pengguna twitter yang dipilih menjadi data set adalah pengguna yang sudah mengisi kuesioner kepribadian *Big Five*, pengguna twitter mempunyai minimal 300 cuitan dengan akun bertipe publik, eksplorasi pengaturan parameter hanya menggunakan *n_estimators*, *random_state*, dan *min_sample_split*.

1.3 Tujuan

Tujuan dari penelitian ini adalah membangun sistem prediksi Tipe Kepribadian *Big Five* dengan menggunakan metode *Random Forest*, mengobservasi parameter yang mempengaruhi kinerja *Random Forest*, serta menganalisis korelasi setiap fitur dengan setiap tipe kepribadian *Big Five*.

1.4 Organisasi Tulisan

Organisasi tulisan pada jurnal ini terbagi menjadi lima bagian. Bagian 1 adalah latar belakang masalah yang mendasari penelitian. Bagian 2 adalah studi terkait yang berhubungan dengan penelitian. Bagian 3 adalah penjelasan dari setiap tahap pada sistem yang dibangun. Bagian 4 adalah hasil dan evaluasi dari sistem yang dibangun. Bagian 5 adalah kesimpulan dari penelitian dan saran untuk penelitian berikutnya.