

## 1. Pendahuluan

### Latar Belakang

Pada era digital saat ini *microblog* berkembang sangat pesat dan menjadi sangat populer, dimana orang-orang dapat membagikan tentang keseharian mereka dan mengekspresikan emosi mereka. Salah satu *microblog* terbesar adalah twitter. Twitter menjadi salah satu ranah peneliti dalam mengumpulkan informasi yang berkaitan dengan social media dikarenakan akses dengan API (*Application Programming Interface*) yang lengkap sehingga memudahkan peneliti untuk menyambungkan *twitter* kedalam aplikasi mereka. Hal ini didukung bahwa twitter masih memiliki tingkat popularitas yang tinggi dan pengguna aktif yang banyak [1][16]. Informasi yang melimpah di twitter jika dimanfaatkan dengan baik dapat berguna dalam bidang perusahaan seperti *Talent Management* untuk membantu mengambil keputusan dalam rekrutasi karyawan baru [2]. Salah satu cara pemanfaatannya dengan memprediksi kepribadian calon karyawan yang akan direkrut. Informasi seperti ini dapat menjadi pertimbangan bagi HR (*Human Resource*) dimana divisi ini yang bertanggung jawab dalam merekrut karyawan baru dalam sebuah perusahaan [2]. Informasi tadi bisa berguna untuk membantu mengambil keputusan terhadap calon karyawan sehingga lebih selektif untuk memilih karyawan yang dianggap pantas dalam perusahaan tersebut.

Penelitian terdahulu melakukan klasifikasi dengan basis 5 faktor kepribadian manusia atau the *big five model of personality* yang terdiri dari *extroversion*, *agreeableness*, *conscientiousness*, *openness* dan *neuroticism* dengan menggunakan algoritma *machine learning* [6][10]. Penelitian tersebut menggunakan dataset pengguna dengan bahasa utama seperti *chinese* dan *english* dalam melakukan proses klasifikasi. Bahasa tersebut memiliki korpus yang besar dan mendapat dukungan oleh perangkat lunak yang bisa melakukan ekstraksi fitur seperti LIWC (*Linguistic Inquiry Word Count*) [10]. Dengan menggunakan LIWC maka dimungkinkan untuk menghasilkan fitur unik sebagai *input* untuk algoritma *machine learning* yang dipakai dari *tweet* yang dikumpulkan dari tiap-tiap *user* [10]. Hal tersebut membuat penelitian yang berjudul *Personality Prediction Based on Twitter Information in Bahasa Indonesia* [11] tidak menyertakan LIWC dalam penelitiannya dikarenakan bahasa dan *user* yang berbeda. Penelitian tersebut mendefinisikan fitur-fitur yang mereka nilai cocok sebagai ciri yang digunakan untuk klasifikasi berbasis *machine learning* mereka [11]. Hasil evaluasi yang diperoleh tidak setinggi akurasi dengan menggunakan LIWC, namun dengan fitur yang terbatas penelitian ini cukup bisa memberikan hasil dengan akurasi yang lumayan stabil diangka 70 sampai 75% [11]. Penelitian lain yang mengimplementasikan hal serupa tanpa menggunakan LIWC adalah *Personality-based refinement for sentiment classification in microblog* [1]. Penelitian tersebut membangun aturan yang secara khusus didesign untuk melakukan klasifikasi kepribadian dari setiap *user* tanpa menggunakan LIWC dan bisa dipakai untuk bahasa apapun. Sistem berbasis aturan tersebut diberi nama PbSC (*Personality based Sentiment Classification*). Algoritma PbSC hanya menggunakan *tweet* yang ditulis pengguna sebagai fitur klasifikasi, namun berdasarkan hasil evaluasi yang dilakukan penelitian tersebut memiliki tingkat akurasi yang sangat tinggi dalam melakukan klasifikasi kepribadian dan mengungguli algoritma berbasis *machine learning* lainnya [1].

Berdasarkan pertimbangan diatas, penelitian ini akan mengimplementasikan algoritma PbSC sebagai metode utama dalam melakukan tugas klasifikasi kepribadian berbasis *big five model personality*. Klasifikasi kepribadian bertujuan untuk memprediksi 3 dari 5 faktor kepribadian yang ada dalam *big-five model personality* dimana faktor tersebut adalah *Extroversion*, *Agreeableness*, dan *Conscientiousness*. Metode ini hanya terbatas untuk 3 faktor kepribadian ini saja sehingga faktor lain seperti *openness* dan *neuroticism* tidak dilakukan dikarenakan langkanya informasi yang bisa dikumpulkan lewat interaksi *user* dengan *social media* yang harus dipakai sebagai faktor penentu 2 kepribadian tersebut [1]. Penelitian ini juga akan mengevaluasi metode tersebut dan melakukan analisis terkait faktor-faktor apa yang mempengaruhi hasil dari evaluasi dan memberikan saran untuk penelitian kedepannya.

### Topik dan Batasannya

Penelitian ini bertujuan untuk mengklasifikasikan kepribadian pengguna dengan data yang didapatkan dari *social media twitter* mereka. Data yang dimaksud adalah *tweet* yang ditulis dari masing-masing pengguna. Data *tweet* yang diperoleh merupakan tulisan yang dimana kebanyakan berisi bahasa indonesia. Data tersebut kemudian diklasifikasikan dengan menggunakan metode PbSC dan melakukan evaluasi dan analisis terhadap metode tersebut. Kelas kepribadian dibagi menjadi 3 dimensi kelas/faktor kepribadian yaitu *Extroversion*, *Agreeableness*, *Conscientiousness*. Metode PbSC akan memprediksi level kepribadian tersebut apakah *high*, *low* atau netral. Hasil dari tiap-tiap klasifikasi yaitu berupa vektor/data yang merepresentasikan level kepribadian dari tiap-tiap pengguna, sebagai contoh, @jooshpn = [1,Netral,0]. Hasil diatas menunjukkan bahwa pengguna dengan username @jooshpn,

memiliki level kepribadian *extrovert* yang tinggi (index 0 dalam *array*), *agreeableness* yang netral (index 1) dan *consciouness* yang rendah (index 2). Sebelum masuk kedalam klasifikasi terdapat beberapa proses yang harus dilakukan seperti pengumpulan data pengguna lewat kuesioner, *crawling tweet*, *preprocessing*, yang akan dijelaskan di bab berikutnya. Adapun batasan penelitian adalah mengabaikan faktor *imbalance data* yang didapatkan dan Penggunaan parameter dan korpus yang tidak dijelaskan pada penelitian [1] akan ditentukan oleh peneliti.

### **Tujuan**

Tujuan dari penelitian ini adalah melakukan klasifikasi kepribadian dengan menggunakan metode PbSC (*Personality based Sentiment Classification*). Hasil klasifikasi yang didapatkan akan dievaluasi dan dilakukan analisis terkait kedua metode tersebut.

### **Organisasi Tulisan**

Penelitian ini terdiri dari 4 bab, dimana setelah bab pendahuluan terdapat bab 2 yaitu studi terkait dimana pada bagian ini akan menjelaskan terkait teori-teori yang berhubungan dengan penelitian ini. Bab 3 berisi tentang penjelasan terkait detail dari perancangan sistem mulai dari awal sampai sistem mampu menghasilkan sebuah output dari tiap-tiap dimensi kepribadian. Bab 4 berisi hasil evaluasi dan analisis dari sistem yang dibangun beserta kesimpulan dan saran untuk penelitian berikutnya.