Personality Prediction Based on Twitter Information in Bahasa

Penelitian tentang prediksi kepribadian telah banyak dilakukan dengan menggunakan berbagai sosial media. Penelitian yang dilakukan oleh Ong, Rahmanto, Williem, & Suhartono (2017) ini menggunakan sosial media Twitter dan berfokus pada bahasa Indonesia. Beberapa statistik mengemukakan bahwa Indonesia merupakan salah satu pengguna sosial media Twitter tertinggi, sekitar 2,4% dari tweets di dunia berasal dari user yang berlokasi di Jakarta, Indonesia (Carley et al., 2015).

Sistem kepribadian dalam penelitian ini dibangun berdasarkan model kepribadian Big Five oleh McCrae dan Costa yang membagi kepribadian seseorang menjadi 5 jenis yaitu, Agreeableness, Conscientiousness, Emotional Stability, Extraversion, dan Openness. Tujuan penelitian ini dibuat adalah untuk membangun sistem prediksi kepribadian dengan bahasa Indonesia, melakukan beberapa skenario percobaan untuk meningkatkan akurasi, dan perbandingan 2 algoritma machine learning yang diimplementasikan ke model prediksi.

Dataset yang yang digunakan pada penelitian ini sebanyak 359 dimana 329 data sebagai data training dan sisa 30 data sebagai data testing. Dataset user yang dipilih berdasarkan kriteria:

1. User post di Twitter minimal sekali setiap bulan
2. User menggunakan bahasa Indonesia

Informasi users diekstraksi menjadi 12 fitur yaitu:

1. Jumlah tweets
2. Jumlah followers
3. Jumlah following
4. Jumlah favorit
5. Jumlah retweets
6. Jumlah tweet yang di retweet
7. Jumlah tweet dengan kutip
8. Jumlah mentions
9. Jumlah replies
10. Jumlah hashtags
11. Jumlah URLs
12. Rata-rata perbedaan waktu setiap tweet

Pelabelan dataset itu dilakukan dengan bantuan ahli psikologi dengan table distribusi seperti di bawah ini.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Agreeableness** | **Conscientiousness** | **Emotional Stability** | **Extraversion** | **Openness** |
| High | 134 | 92 | 150 | 202 | 163 |
| Low | 195 | 237 | 179 | 127 | 166 |

Tabel distribusi dataset training

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Agreeableness** | **Conscientiousness** | **Emotional Stability** | **Extraversion** | **Openness** |
| High | 19 | 16 | 21 | 23 | 16 |
| Low | 11 | 14 | 9 | 6 | 14 |

Tabel distribusi dataset testing

Selanjutnya, tahap yang dilakukan adalah preprocessing data untuk menghilangkan beberapa elemen secara otomatis dan manual seperti, penghilangan emoji, penghilangan URLs/hyperlinks, penghilangan stop words dan elemen lainnya. Setelah melalui tahap preprocessing, sistem prediksi dibangun dengan menggunakan Support Vector Machine dan XGBoost, dimana SVM dijalankan dengan WEKA dan XGBoost dengan R.

Setelah proses training, sistem di evaluasi dengan 10-fold cross validation dan memasukkan 30 data testing ke dalam sistem. Sistem prediksi ini dibangun dengan beberapa skenario seperti terlihat pada Tabel.

Tabel Skenario evaluasi

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Scenario** | **Minimum occurrence of n- gram** | | **n-gram weighting scheme** | | **LDA topic features** | | **Stop words omission** | |
| **1** | **2** | **Boolean** | **TF** | **Use LDA** | **Don’t use LDA** | **Omit** | **Don’t omit** |
| 1 |  |  |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |  |  |
| 3 |  |  |  |  |  |  |  |  |
| 4 |  |  |  |  |  |  |  |  |
| 5 |  |  |  |  |  |  |  |  |
| 6 |  |  |  |  |  |  |  |  |
| 7 |  |  |  |  |  |  |  |  |
| 8 |  |  |  |  |  |  |  |  |
| 9 |  |  |  |  |  |  |  |  |
| 10 |  |  |  |  |  |  |  |  |
| 11 |  |  |  |  |  |  |  |  |
| 12 |  |  |  |  |  |  |  |  |
| 13 |  |  |  |  |  |  |  |  |
| 14 |  |  |  |  |  |  |  |  |
| 15 |  |  |  |  |  |  |  |  |
| 16 |  |  |  |  |  |  |  |  |



Hasil akurasi SVM dengan 10-fold cross validation



Hasil akurasi SVM dengan menggunakan dataset testing



Hasil akurasi XGBoost dengan menggunakan 10-fold cross validation



Hasil akurasi XGBoost dengan menggunakan dataset testing

Dari penelitian dan hasil akurasi pada tabel diatas menunjukkan bahwa evaluasi dengan menggunakan 10-fold cross validation pada SVM menghasilkan akurasi tertinggi dengan 76.2310%, sedangkan XGBoost mendapatkan 97.9962%. Hasil evaluasi juga menunjukkan bahwa penggunaan fitur topik LDA dan frekuensi TF dapat meningkatkan akurasi sitem prediksi.

Hasil akhir dari penelitian menunjukkan penggunaan XGBoost pada sistem jauh lebih baik daripada menggunakan Support Vector Machine.