**BAB 3**

**Metodologi**

* 1. **Analisis Masalah**

Setiap manusia memiliki kepribadiannya masing-masing. Kepribadian ini menjadi ciri unik yang menjadi gambaran seseorang tersebut di dalam kehidupannya. Kepribadian dapat membedakan orang tersebut dalam berperilaku atau memecahkan suatu permasalahan. Saat ini, kepribadian juga menjadi sesuatu yang sangat penting untuk dijadikan pertimbangan dalam perekrutan tenaga kerja. Di sisi lain, kepribadian juga dapat menjadi faktor suatu hubungan dan relasi terhadap orang lain yang cocok kepribadiannya satu sama lain dan masih banyak lagi pengaruh kepribadian terhadap kehidupan seseorang.

Untuk itu, banyak sekali cara yang dapat dilakukan oleh seseorang untuk mengetahui kepribadiannya sendiri. Mulai dari mengisi kuesioner kepribadian atau aplikasi yang beredar di sosial media dan internet. Salah satu model kepribadian yang paling sering digunakan adalah *The Big Five Model Personality*. Model kepribadian ini terbagi menjadi 5 bagian utama yaitu, *Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness* dan *Neuroticism*. Cara mendapatkan hasil kepribadian dari model ini biasa dengan mengisi beberapa pertanyaan mengenai diri kita sendiri. Namun pada kenyataannya, hasil dari kepribadian ini biasa tidak tepat dan dapat dimanipulasi oleh orang yang mengisinya. Faktor-faktor tersebut biasa berupa untuk menutupi kepribadian dirinya yang ia anggap buruk atau pertanyaan yang terlalu banyak sehingga menyebabkan responden merasa bosan dan kesulitan untuk menjawab semuanya.

Menurut Pennebaker, Mehl, & Gosling (2006), kepribadian seseorang akan lebih akurat bila dinilai berdasarkan perilakunya sehari-hari. Perilaku seseorang ini biasa hanya dapat dilihat dan dirasakan oleh orang-orang yang berada di sekitarnya. Tetapi, akibat perkembangan teknologi zaman saat ini, dimana orang lebih banyak menghabiskan waktunya di dunia maya atau sosial media, kepribadian seseorang secara tidak langsung dapat dilihat oleh siapa saja.

Walaupun penelitian mengenai sistem prediksi kepribadian dari sosial media telah banyak dilakukan. Namun, masih sebagian kecil yang menggunakan sosial media Facebook karena tidak tersedianya *Public API* oleh Facebook untuk mendapatkan dataset dari *user*. Beberapa penelitian yang menggunakan sosial media Facebook juga masih memiliki akurasi yang tergolong rendah.

* 1. **Usulan Pemecahan Masalah**

Dari analisis permasalahan tentang penelitian sistem prediksi kepribadian, peneliti berencana mengembangkan sistem prediksi kepribadian dengan akurasi yang lebih tinggi dari penelitian sebelumnya yang menggunakan sosial media Facebook. Masalah dalam sistem prediksi ini yang paling penting adalah tingkat akurasi. Untuk itu peneliti berencana mencoba berbagai metode sebagai berikut:

1. Mencoba berbagai algoritma *Machine Learning* untuk *training* dan membandingkan akurasi yang paling tinggi untuk setiap *traits* dari *Big Five*.

2. Melakukan perbandingan berbagai *feature pre-defined* (*closed vocabulary*) yang paling akurat untuk sistem prediksi.

3. Mencoba menerapkan *Deep Learning* untuk training sistem dengan metode *open* *vocabulary*.

* 1. **Kerangka Berpikir**



**Gambar 3.1** Kerangka Berpikir

**3.3.1 *Dataset Collecting***

**1.  *myPersonality Dataset***

Sosial media Facebook tidak memberikan *Public API* untuk mendapatkan data dari *user*. Facebook hanya menyediakan *API* untuk mendapatkan data dari account kita sendiri atau account lain selama kita memiliki authentication code dari account tersebut. Maka untuk mendapatkan dataset yang lebih besar, peneliti menggunakan data user Facebook dari myPersonality (Kosinki, 2015). myPersonality adalah aplikasi Facebook yang dikembangkan oleh Michal Kosinki dimana user Facebook dapat menggunakan aplikasi itu untuk mendapatkan hasil kepribadian mereka berdasarkan model kepribadian Big Five. Data yang disediakan secara terbuka hanya berupa data 250 user dengan kurang lebih 10.000 status. Data-data ini telah ditambahkan dengan beberapa detail seperti waktu, social network dari user dan jenis kepribadiannya.

Selain itu, 250 dataset user Facebook yang disediakan ini juga telah dilabeli berdasarkan Big Five Model Personality sehingga dapat diolah secara langsung oleh para peneliti yang menggunakannya. Sebagai tambahan, myPersonality sebenarnya merupakan salah satu pemilik dataset user Facebook yang paling besar dan memiliki sekitar jutaan user Facebook beserta profil dan informasi lain mengenai user-user tersebut. Untuk membutuhkan dataset yang lebih besar itu, peneliti diharuskan untuk mendapatkan novelty yang baru dan bukan merupakan novelty yang telah dimiliki oleh para peneliti lain. Tetapi dalam kesempatan ini, peneliti belum berhasil untuk menjadi kolaborator myPersonality dan gagal mendapatkan dataset dalam jumlah yang lebih besar.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Value** | **cOPN** | **cCON** | **cEXT** | **cAGR** | **cNEU** |
| Yes | 176 | 130 | 96 | 134 | 99 |
| No | 74 | 120 | 154 | 116 | 151 |

**Tabel 3.1** Distribusi jenis kepribadian berdasarkan 250 user

dalam dataset myPersonality (Kosinki, 2015).

**2. Manual Data Gathering**

Selain mengambil data dari myPersonality, pengambilan data juga dilakukan dengan cara manual dengan memanfaatkan Graph API Facebook yang dapat memberikan akses data terhadap sebuah account selama user menyetujui untuk memberikan access tokennya. Jadi, peneliti meminta izin terhadap beberapa pengguna khususnya mahasiswa universitas Bina Nusantara untuk mendapatkan status dari akun Facebook mereka. Sebagian user memiliki status dengan bahasa Inggris dan bahasa Indonesia. Jadi user dengan status bahasa Indonesia ini selanjutnya akan dilanjutkan untuk di proses di tahap preprocessing data untuk disesuaikan dengan semua status lain di penelitian ini yang menggunakan bahasa Inggris.



**Gambar 3.2** Tampilan web aplikasi untuk Manual Data Gathering



**Gambar 3.3** Tampilan *Graph* API untuk mendapatkan

*access token user*

Dari manual data gathering ini, didapatkan user dengan jumlah ---------.

Proses yang dilakukan selanjutnya adalah pelabelan data dengan menggunakan aplikasi personality Apply Magic Sauce (<https://applymagicsauce.com/>). Setiap status dari user diprediksi kepribadiannya berdasarkan Big Five model dengan menggunakan web aplikasi ini.

User #37 dengan status:

“Everyone has an addiction, and mine happens to be you' "Best healer Hellscream EU Kappa" "woof, I guess" Hair is done Baldmongold I'm a keyboard warrior! Every day, all about 2 month WHEN PEOPLE TALK SHIT ABOUT ME I SAY THIS Love is so short, forgetting is so long”



**Gambar 3.4** Hasil kepribadian user #37 berdasarkan *apply magic sauce*

Dari gambar 3.4 dapat dilihat bahwa user #37 memiliki traits Openness yang cukup tinggi senilai +66% sehingga disimpulkan user ini dominan di Openness (Openness = y), memiliki nilai -31% di traits Conscientiousness (Conscientiousness=n), -19% di traits Extraversion (Extraversion=n), -15% di traits Agreeableness (Agreeableness=n), +85% di traits Neuroticism (Neuroticism=y). Setelah semua dataset user dilabeli, hasil akhir distribusi jumlah dataset user yang didapatkan beserta pelabelannya ditampilkan dalam tabel distribusi di bawah ini:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Value** | **cOPN** | **cCON** | **cEXT** | **cAGR** | **cNEU** |
| Yes | nn | nn | nn | nn | nn |
| No | nn | nn | nn | nn | nn |

**Tabel 3.2** Distribusi dataset user dari Manual Data Gathering

**3.3.2 *Data Preprocessing***

Dataset yang telah diperoleh dari myPersonality maupun manual data Gathering selanjutnya masuk ke dalam proses data preprocessing sebelum dilanjutkan ke tahap feature selection dan training. Proses preprocessing data ini terbagi dua karena dataset yang peneliti miliki berbeda dimana dataset yang pertama merupakan dataset dari myPersonality (bahasa inggris) dan dataset manual gathering yang sebagian merupakan bahasa Inggris dan sebagian lagi bahasa Indonesia.



**Gambar 3.5** Flow Preprocessing data

**1. Remove URLs**

Tahap Preprocessing yang pertama adalah menghilangkan urls yang ada di dalam teks atau status dari dataset user yang telah dikumpulkan.

**2. Remove Symbols**

Tahap Preprocessing yang pertama adalah menghilangkan urls yang ada di dalam teks atau status dari dataset user yang telah dikumpulkan.

**3. Remove Names**

Tahap Preprocessing yang pertama adalah menghilangkan urls yang ada di dalam teks atau status dari dataset user yang telah dikumpulkan.

**4. Remove Spaces**

Tahap Preprocessing yang pertama adalah menghilangkan urls yang ada di dalam teks atau status dari dataset user yang telah dikumpulkan.

**5. Slang Words Filter**

Tahap Preprocessing yang pertama adalah menghilangkan urls yang ada di dalam teks atau status dari dataset user yang telah dikumpulkan.

**6. Lower case all text**

Tahap Preprocessing yang pertama adalah menghilangkan urls yang ada di dalam teks atau status dari dataset user yang telah dikumpulkan.

**7. Translate to English**

Tahap Preprocessing yang pertama adalah menghilangkan urls yang ada di dalam teks atau status dari dataset user yang telah dikumpulkan

**8. Stemming**

Tahap Preprocessing yang pertama adalah menghilangkan urls yang ada di dalam teks atau status dari dataset user yang telah dikumpulkan.

**9. Remove Stopword**

Tahap Preprocessing yang pertama adalah menghilangkan urls yang ada di dalam teks atau status dari dataset user yang telah dikumpulkan

**3.3.3 *Feature Selection***

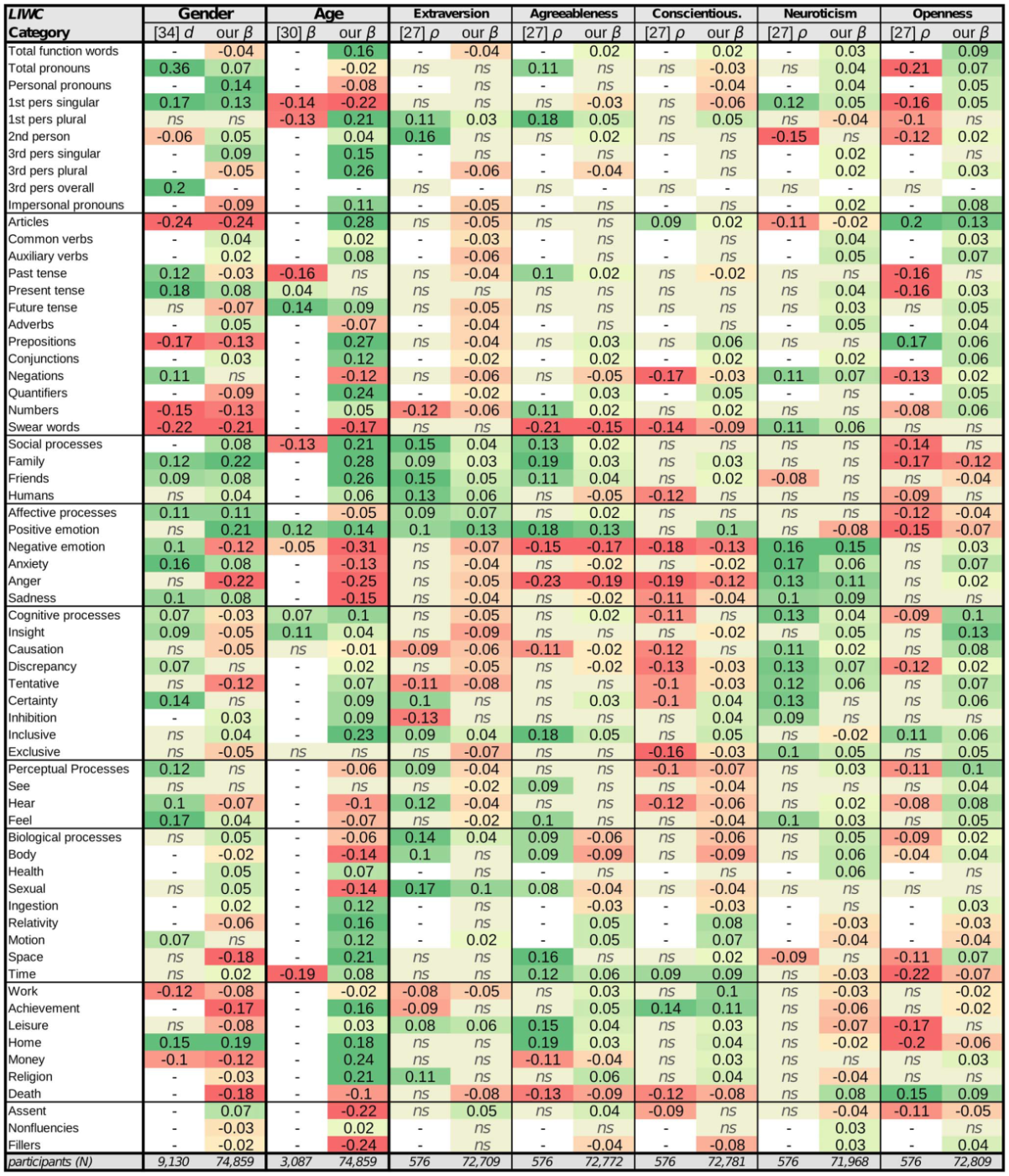
Dalam pengembangan sistem dan setelah melalui beberapa pencarian mengenai fitur yang dapat digunakan dalam melakukan proses olah bahasa. Beberapa fitur di bawah ini merupakan fitur yang peneliti coba dan kemudian peneliti bandingkan satu sama lain untuk dilihat keakuratan dan fungsionalitasnya. Berikut beberapa fitur yang ada dalam penelitian ini:

1. **LIWC**

Fitur linguistik pertama yang kita terapkan dalam penelitian adalah fitur LIWC (*Linguistic Inquiry and Word Count*). Fitur LIWC ini dapat dikatakan sebagai fitur yang paling umum dan paling sering digunakan dalam penelitian NLP karena telah lama dikembangkan dan diperbaharui hingga saat ini. Sebagai awal mula pembangunan sistem penelitian peneliti mencoba untuk mengumpulkan seluruh status dari user untuk kemudian diibagi per masing-masing user.

Setelah didapatkan susunan status per user. Kemudian peneliti menggunakan API yang disediakan LIWC melalui Receptiviti untuk mendapatkan hasil dari nilai fitur dengan status dari user.

**Tabel 3.3** Korelasi Kategori LIWC dengan jenis kelamin, umur dan *Big Five model personality* (Schwartz, et al., 2013)



1. **Splice**

Fitur linguistik kedua yang peneliti gunakan adalah SPLICE (*Structured Programming for Linguistic Cue Extraction).* Fitur ini masih cukup baru dan masih jarang digunakan, tetapi setelah melalui proses research paper dan literature. Peneliti merasa fitur ini cukup akurat dan lengkap. Proses yang dilakukan mirip dengan ketika peneliti melakukan korelasi linguistik dengan fitur LIWC. Setelah status dipilah berdasarkan masing-masing user. Kemudian API dari SPLICE dipanggil dan digunakan untuk menghasilkan nilai dari fitur-fitur yang ada di SPLICE.

1. **Social Network Analysis (SNA) Features (Only myPersonality datasets)**

Fitur selanjutnya merupakan fitur *Social Networks Analysis* dari user. *Social Networks Analysis* adalah suatu alat atau studi yang memetakan hubungan pengetahuan yang penting antar individu (Pryke, 2004). SNA dikembangkan untuk memahami hubungan –hubungan (*ties/edge*) dari aktor-aktor (*nodes/points*) yang ada dalam sebuah sistem dengan 2 fokus, yaitu aktor-aktor dan hubungan antar actor dalam konteks sosial tertentu. SNA sering diimplementasikan untuk mengidentifikasi arus informasi. Secara teori dengan mengidentifikasi arus informasi bisa membantu meningkatkan strategi yang bisa memacu para aktor untuk berbagi informasi daripada harus menciptakan strategi yang baru (Serrat, 2009).

Fitur SNA hanya akan digunakan untuk dataset yang disedikan oleh myPersonality, karena myPersonality memberikan keseluruhan fitur SNA yang telah dapat kita gunakan untuk penelitian ini. Fitur ini digunakan sebagai salah satu fitur tambahan atau opsional sebagai pembanding dari fitur lainnya sebagai konsiderasi untuk penelitian berikutnya. Fitur dari SNA yang digunakan dalam penelitian ini terbagi atas 7 fitur yaitu:

* *Network size*
* *Betweeenness*
* *nBetweenness*
* *Density*
* *Brokerage*
* *nBrokerage*
* *Transitivity*

1. **Open Vocabulary**

Metode Open Vocabulary ini berbeda dengan fitur-fitur lainnya dimana semua nilai dari fitur di define di awal sebelum proses training dilakukan. Metode Open Vocabulary tidak memiliki jumlah fitur pasti, tetapi fitur didapatkan dengan menelusuri dataset yang kita gunakan. Kemudian, kata-kata dari dataset yang kita kumpulkan dibagi dan dihitung jumlahnya (word count). Setiap kata itulah yang dikembangkan menjadi fitur dan dibandingkan dengan Big Five Model Personality.

Berdasarkan penelitian sebelumnya, masih banyak pertimbangan dalam NLP untuk menentukan metode manakah yang lebih baik, apakah metode Closed Vocabulary (LIWC, SPLICE, MRC, dan lainnya) atau dengan metode Open Vocabulary. Untuk itu dalam penelitian ini peneliti juga menerapkan metode Open Vocabulary untuk mendapatkan hasil akurasi prediksi dan dibandingkan dengan metode lainnya.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Learning method | Machine Learning | | | Deep Learning | | |
| Dataset type | myPersonality datasets | Manual datasets | Combined datasets | myPersonality datasets | Manual datasets | Combined datasets |
| LIWC | ✓ | ✓ | ✓ | 🗶 | 🗶 | 🗶 |
| SPLICE | ✓ | ✓ | ✓ | 🗶 | 🗶 | 🗶 |
| SNA | ✓ | 🗶 | 🗶 | 🗶 | 🗶 | 🗶 |
| Open Vocabulary | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

**Tabel 3.4** *Features Selection* berdasarkan *learning method*

dan *dataset* yang digunakan

**3.3.3 *Model* *Learning Process***

Proses learning dibagi menjadi dua metode, yaitu dengan menggunakan machine learning algorithm dan deep learning.

**1. Machine learning**

Setelah melalui tahap seleksi fitur, hasil-hasil tersebut kemudian dilanjutkan ke dalam proses learning dengan menerapkan beberapa algoritma classifier yaitu SVM, gradient boosting, Linear Regression dan lainnya.

**2. Deep learning**

Setelah mendapatkan hasil dari proses learning tersebut. Peneliti kemudian mencoba kembali menggunakan proses Deep Learning dengan beberapa arsitektur seperti MLP, CNN, dan LSTM. Hasil dari penggunaan fitur yang berbeda dan proses learning yang berbeda akan terus dibandingkan akurasinya untuk mendapatkan sistem prediksi yang paling akurat.

**3.3.4 *Prediction Model Testing***

Pada tahap ini data training dari sistem yang telah melewati proses learning akan dibandingkan dengan data testing. Melalui tahap ini, akan didapatkan prediksi terhadap data testing tersebut. Peneliti akan melihat tingkat akurasi dari prediksi tersebut untuk kemudian dikembangkan secara maksimal untuk mencapai tingkat akurasi yang paling maksimal berdasarkan 3 faktor paling menentukan dalam hasil sistem prediksi ini yang disebut oleh peneliti dengan FAL (*Feature extraction*, *Algorithm* *classifier*, and *Learning* *process*).

Hasil klasifikasi dari *training* ini akan menampilkan boolean value untuk setiap *trait* kepribadian dari *Big Five Model Personality*. Nilai 1 untuk *user* yang dianggap memiliki persentase tinggi di *trait* tersebut dan nilai 0 untuk *user* yang dianggap memiliki persentase rendah di *trait* tersebut.

**3.3.5 *Final result of Prediction based on Big Five model***

Tahap akhir dari penelitian akan berwujud sebuah user interface untuk menampilkan secara langsung hasil prediksi dari seorang user yang telah melakukan log in ke dalam Facebook sehingga terhubung ke API Facebook untuk mendapatkan data-data yang dibutuhkan dari user tersebut khususnya status. Hasil prediksi dalam interface akan menunjukkan kepribadian Big Five Model Personality dari user tersebut disertai dengan fitur-fitur dan informasi lainnya yang dikembangkan lebih lanjut.

* 1. **Rancangan Layar**

Hasil akhir dari aplikasi akan berupa sebuah aplikasi berbasis *web* yang memungkinkan pengguna untuk mencoba langsung sistem prediksi yang telah peneliti kembangkan dan mendapatkan hasil dari kepribadian mereka dengan terhubung ke akun Facebook mereka. Berikut adalah perancangan layar aplikasi yang akan dibuat.

**3.4.1 Rancangan Layar *Homepage***



**Gambar 3.6** Rancangan Layar *Homepage*

Rancangan layar Homepage yang akan dibuat oleh peneliti cukup sederhana. Pada layar ini terdapat logo yang akan berada tepat di tengah dan nama aplikasi di atasnya. Di bawah logo akan terdapat *button* dari Facebook yaitu *button* “Log in With Facebook” yang merupakan standar dari Facebook bagi aplikasi web yang ingin terhubung ke dalam suatu akun Facebook.

**3.4.2 Rancangan Layar *Privacy Policy***

Di bagian sudut kiri bawah dari layar *Homepage*. Terdapat *button* yang akan *redirect* ke halaman *privacy policy.*



**Gambar 3.7** Rancangan Layar *Privacy Policy*

Layar *Privacy Policy* akan berisi syarat dan ketentuan maupun segala informasi keamanan privasi di dalam aplikasi *web* yang dibuat oleh peneliti.

**3.4.2 Rancangan Layar *Result***

Layar Result akan muncul setelah user melakukan *login* ke dalam akun Facebook mereka melalui halaman *homepage.*



**Gambar 3.8** Rancangan Layar *Result* (1)



**Gambar 3.9** Rancangan Layar *Result* (2)

Layar *Result* akan menampilkan hasil dari *personality user* tersebut. Di bagian paling atas akan berisi nama aplikasi. Selanjutnya, di bagian tengah dari layar akan berisi persentase kepribadian dari user tersebut sesuai dengan model kepribadian *Big Five Traits*.

Di bagian bawah dari *web* seperti terlihat pada gambar 3.9 akan memiliki bagian-bagian tersendiri sesuai dengan model kepribadian *Big Five* yang berisi mengenai penjelasan dari hasil dari kepribadian mereka masing-masing. Setiap kotak yang berisi penjelasan mengenai kepribadian mereka berdasarkan jenis kepribadian itu akan memiliki ikon di bagian kanan bawah untuk menambah visualisasi kepada user dalam mengerti *traits* tersebut.