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

2. Mencoba menerapkan *Deep Learning* dalam penelitian.

3. Menyeimbangkan dataset yang ada dengan *preprocessing.*

4. Mencoba penerapan *resampling* karena dataset yang diperoleh cukup kecil.

* 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. Pelabelan dari setiap user tersebut dapat dilihat pada Tabel 3.1 yang merupakan distribusi jenis kepribadian. 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).

Tabel 3.1 diatas memperlihatkan penyebaran user yang diberikan oleh dataset myPersonality, terdapat 176 *user* yang dominan terhadap *traits openness* dan yang rendah di *traits openness* sebanyak 74 *user*. Ini menyebabkan *traits openness* tidak seimbang dan dapat mempengaruhi tingkat akurasi di proses *learning* selanjutnya. Penyebaran terjadi cukup merata di *traits conscientiousness*, dimana sebanyak 130 *user* tinggi di *traits conscientiousness* dan 120 *user* rendah di *traits* ini. Di *traits* *extraversion*, perbandingan juga cukup signifikan dan tidak seimbang karena hanya 96 *user* yang tinggi di *traits* ini dan sebanyak 154 *user* rendah di *traits extraversion*. Di *traits agreeableness*, 134 *user* dominan dan sebaliknya sebanyak 116 *user*. *Traits* terakhir yaitu *neuroticism* memiliki 99 *user* yang tinggi di *traits* ini dan sebanyak 151 *user* rendah sehingga juga menyebabkan sedikit ketidakseimbangan dalam *traits* ini.

**2. Manual Data Gathering**

Selain mengambil data dari myPersonality, pengambilan data juga dilakukan dengan cara manual dengan memanfaatkan Graph API Facebook (Gambar 3.3) yang dapat memberikan akses data terhadap sebuah account selama user menyetujui untuk memberikan access tokennya. Pada Gambar 3.2 terlihat sebuah interface tampilan layar yang disediakan oleh peneliti dalam proses pengambilan data manual ini. 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

Pada Gambar 3.2 di atas, tampilan tersebut muncul setelah *user* melakukan *login* ke dalam akun Facebook mereka. Hal yang selanjutnya perlu dilakukan adalah mengklik tombol “*Get your access token here*” yang akan me-redirect user ke *Graph API* Facebook (Gambar 3.3). Di *Graph API* Facebook, user hanya perlu melakukan klik tombol “*Get Token*” dan centang di *user\_post* untuk mendapatkan seluruh post dan status dari akun mereka. Setelah mendapatkan access tokennya, user tinggal meng-*copy* token tersebut dan *paste* di *text box* pada halaman utama *web* di Gambar 3.2. Hal terakhir adalah klik “*Get your status updates*” untuk mengirim semua status mereka ke dalam *database* yang telah disediakan oleh peneliti.



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

*access token user*

Dari manual data gathering ini, didapatkan user sebanyak 139 user. 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. Gambar 3.4 memperlihatkan contoh dari user #37 yang statusnya akan dilabeli.

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** Status dari user #37



**Gambar 3.5** Hasil kepribadian user #37 berdasarkan *apply magic sauce*

Dari Gambar 3.5 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 pada Tabel 3.2.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Value** | **cOPN** | **cCON** | **cEXT** | **cAGR** | **cNEU** |
| Yes | 97 | 63 | 38 | 81 | 50 |
| No | 53 | 87 | 112 | 69 | 100 |

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

Tabel 3.2 di atas menampilkan bahwa dari 150 *user* yang di dapatkan, terdapat 87 *user* yang dominan terhadap *traits openness* dan yang rendah di *traits openness* sebanyak 52. Sebanyak 56 user tinggi di *traits conscientiousness* dan 83 *user* rendah di *traits* ini. Di *traits* *extraversion*, perbandingan cukup signifikan dan tidak seimbang karena hanya 34 *user* yang tinggi di *traits* ini dan sebanyak 105 *user* rendah di *traits extraversion* ini. Di *traits agreeableness*, 75 *user* dominan dan sebaliknya sebanyak 64 *user*. *Traits* terakhir yaitu *neuroticism* memiliki 49 *user* yang tinggi di *traits* ini dan sebanyak 90 *user* rendah sehingga juga menyebabkan ketidakseimbangan dalam *traits* ini.

**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. Proses dalam tahap ini dapat dilihat pada Gambar 3.6 yang memberikan *flow* dari *preprocessing* data yang dilakukan pada penelitian ini.



**Gambar 3.6** *Flow Preprocessing* data

**1. Remove URLs**

Tahap Preprocessing yang pertama adalah menghilangkan urls dan link yang ada di dalam teks atau status dari dataset user yang telah dikumpulkan. Gambar 3.7 akan memperlihatkan status dari *user* yang memiliki url di dalamnya dan di tandai dengan *highlight* berwarna abu-abu, dan hasil setelah url dihilangkan dapat dilihat pada Gambar 3.8.

WOW!!! I just spent 20 minutes reading just this one strip... http:////xkcd.com//657//

**Gambar 3.7** Contoh status sebelum url dihilangkan

WOW!!! I just spent 20 minutes reading just this one strip...

**Gambar 3.8** Contoh status setelah url dihilangkan

**2. Remove Symbols**

Tahap Preprocessing selanjutnya adalah menghilangkan simbol-simbol-simbol yang tidak memiliki arti dan nilai dalam proses ini. Simbol-simbol ini tidak termasuk seluruh tanda baca (*punctuation*) yang memiliki nilai dan fitur dalam pengolahan teks. Simbol ini contohnya seperti simbol *copyright* atau *trademark*, dan simbol-simbol lainnya yang biasa digunakan *user* sebagai *emoticon.* Dapat dilihat pada Gambar 3.9 dimana terdapat simbol *copyright* pada status *user* dan kemudian dihilangkan pada Gambar 3.10.

©9gag,Everything is okay in the end. If it's not okay, then it's not the end.

**Gambar 3.9** Contoh status sebelum simbol dihilangkan

9gag,Everything is okay in the end. If it's not okay, then it's not the end.

**Gambar 3.10** Contoh status setelah simbol dihilangkan

**3. Remove Names**

Nama yang ada pada status tidak memiliki nilai dan fitur sehingga harus dihilangkan dari status tersebut. Dalam status Facebook umumnya nama berada dalam status karena *user* tersebut di *mention.* Oleh karena itu, nama yang ada pada status (Gambar 3.11) dihilangkan di Gambar 3.12.

Venson Wijaya Santy Chen do not say you do not know this

**Gambar 3.11** Contoh status sebelum namadihilangkan

do not say you do not know this

**Gambar 3.12** Contoh status setelah nama dihilangkan

**4. Remove Spaces**

Status yang ada di Facebook kadangkala memiliki spasi yang belebihan. Untuk itu, status-status dengan spasi yang berlebih tersebut dimasukkan dalam proses ini. Pada Gambar 3.13 terlihat bahwa terdapat jarak cukup besar di antara kata “Those” dan “who” serta antara kata “doing” dan “it”. Untuk itu penghilangan spasi dilakukan dan hasilnya terlihat seperti pada Gambar 3.14.

"Those who say it cannot be done should not interfere with those of us who are doing it"

**Gambar 3.13** Contoh status sebelum spasi dihilangkan

"Those who say it cannot be done should not interfere with those of us who are doing it"

**Gambar 3.14** Contoh status setelah spasi dihilangkan

**5. *Slang Words Filter***

Tahap *Slang Words Filter* ini dikhususkan untuk status dengan bahasa Indonesia. Status bahasa Indonesia yang belum diterjemahkan ini dilakukan *replacement*  kata-kata *slang* dengan maksud agar ketika diterjemahkan, kata-kata slang tersebut memiliki arti dan dapat diterjemah. Proses penggantian kata-kata *slang* ini dengan menggunakan sebuah list kata *slang* bahasa Indonesia sebanyak 1075 kata yang didapatkan dari penelitian Naradhipa & Purwarianti (2011). Pada Gambar 3.15 dapat terlihat bahwa *slang* dapat berupa sebuah kata yang diubah secara menyeluruh seperti “gw” yang merupakan kata informal dari kata “aku” atau “saya”. Ada juga kata yang hanya disingkat dari kata aslinya seperti “bcanda” yang merupakan kata “bercanda”. Setelah di *filter* dan di *replace*, hasil dari status seperti pada Gambar 3.16. Kumpulan kata *slang* ini dapat dilihat pada *Appendix* 2.

Ok,nih cara dapet duit gratis buat loe smua klo mo beli sejenis Mata uang game  cash(Token buat NS,Playfish cash buat game playfish(semacam Pet society)dll.)tapi gw lg coba nih cara.kira-kira gw kasih kabar selanjutnya 5 menit lg.(nih serius gk bcanda gw!gw bru nemuin nih cara dari internet.klo mo liat langsung,nanti(10 menit lg)liat aj note gw)

**Gambar 3.15** Contoh status sebelum kata-kata *slang* dihilangkan

Ok,ini cara dapat uang gratis buat kamu semua kalau mau beli sejenis Mata uang game cash (Token buat NS,Playfish cash buat game playfish(semacam Pet society)dan lain-lain.)tapi saya lagi coba ini cara.kira-kira saya kasih kabar selanjutnya 5 menit lagi.(ini serius tidak bercanda saya!saya baru nemuin ini cara dari internet.kalau mau lihat langsung,nanti(10 menit lagi)lihat saja note saya)

**Gambar 3.16** Contoh status setelah kata-kata *slang* dihilangkan

**6. Lower case all text**

Karena dalam penelitian ini semua kata yang di proses tidak membedakan antara kata *upper case* dan *lower case*, maka semua kata dinormalisasi menjadi *lower case* seperti terlihat pada Gambar 3.17 dan hasilnya pada Gambar 3.18.

THANK GOD THAT WOMAN FINALLY LEFT THE FRIGGIN' HOUSE!!!

**Gambar 3.17** Contoh status sebelum status diubah menjadi

*lower case*

thank god that woman finally left the friggin’ house!!!

**Gambar 3.18** Contoh status setelah status diubah menjadi

*lower case*

**7. Translate to English**

Tahap Preprocessing yang selanjutnya dilakukan setelah tahap *lower case all text* adalah melakukan terjemahan terhadap teks yang merupakan bahasa Indonesia. Proses translasi ini dilakukan secara manual dan menggunakan bantuan Google Translate. Status sebelum diterjemahkan dapat dilihat pada Gambar 3.19 dan hasil terjemahannya dapat dilihat pada Gambar 3.20.

aku hanya rakyat biasa jangan samakan aku dengan mu

**Gambar 3.19** Contoh status sebelum diterjemahkan

I'm just ordinary folk do not equate me with you

**Gambar 3.20** Contoh status setelah diterjemahkan

**8. Stemming**

Tahap Stemming adalah proses merubah kata menjadi kata dasar. Imbuhan yang ada pada kita dihilangkan untuk mendapatkan inti dari arti kata tersebut. Pada proses ini dapat dilihat di Gambar 3.21 dimana status yang masih memiliki kata berimbuhan disederhanakan menjadi kata dasarnya pada Gambar 3.22.

wishes to develop a super power that prevents her from needing to sleep

**Gambar 3.21** Contoh status sebelum dilakukan proses *stemming*

wish to develop a super power that prevent her from need to sleep

**Gambar 3.22** Contoh status setelah proses stemming dilakukan

**9. Remove Stopword**

Tahap Preprocessing terakhir adalah menghilangkan semua *Stopword* yang ada pada status. *Stopword* adalah kata umum (Common Words) yang biasanya muncul dalam jumlah besar dan tidak memiliki makna berarti dalam kalimat. Dalam penelitian ini, peneliti membuat *list stopwords* dengan menggunakan 153 kata dari *library* NLTK. Berdasarkan *list* *stopwords* tersebut, peneliti menghilangkan semua kata yang menjadi stopwords itu. Seperti terlihat pada Gambar 3.23 dimana kata “is”, “to” dan “at” merupakan salah satu kata dalam daftar *stopwords* sehingga dihilangkan menjadi hasil yang terlihat pada Gambar 3.24. Daftar stopwords dapat dilihat pada *Appendix* 3.

is going to bed at 9:30! Yeah!

**Gambar 3.23** Contoh status sebelum *stopwords* dihilangkan

going bed 9:30! Yeah!

**Gambar 3.24** Contoh status setelah *stopwords* dihilangkan

**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** Fitur LIWC yang digunakan dalam penelitian

|  | **Category** | **Abbrev** | **Examples** |
| --- | --- | --- | --- |
|  | **Linguistic Dimensions** |  |  |
| 1 | Total function words | funct | it, to, no, very |
| 2 | Total pronouns | pronoun | I, them, itself |
| 3 | Personal pronouns | ppron | I, them, her |
| 4 | 1st pers singular | i | I, me, mine |
| 5 | 1st pers plural | we | we, us, our |
| 6 | 2nd person | you | you, your, thou |
| 7 | 3rd pers singular | shehe | she, her, him |
| 8 | 3rd pers plural | they | they, their, they’d |
| 9 | Impersonal pronouns | ipron | it, it’s, those |
| 10 | Articles | article | a, an, the |
| 11 | Prepositions | prep | to, with, above |
| 12 | Auxiliary verbs | auxverb | am, will, have |
| 13 | Common Adverbs | adverb | very, really |
| 14 | Conjunctions | conj | and, but, whereas |
| 15 | Negations | negate | no, not, never |
|  | **Other Grammar** |  |  |
| 16 | Common verbs | verb | eat, come, carry |
| 17 | Common adjectives | adj | free, happy, long |
| 18 | Comparisons | compare | greater, best, after |
| 19 | Interrogatives | interrog | how, when, what |
| 20 | Numbers | number | second, thousand |
| 21 | Quantifiers | quant | few, many, much |
|  | **Psychological Processes** |  |  |
| 22 | Affective processes | affect | happy, cried |
| 23 | Positive emotion | posemo | love, nice, sweet |
| 24 | Negative emotion | negemo | hurt, ugly, nasty |
| 25 | Anxiety | anx | worried, fearful |
| 26 | Anger | anger | hate, kill, annoyed |
| 27 | Sadness | sad | crying, grief, sad |
| 28 | Social processes | social | mate, talk, they |
| 29 | Family | family | daughter, dad, aunt |
| 30 | Friends | friend | buddy, neighbor |
| 31 | Female references | female | girl, her, mom |
| 32 | Male references | male | boy, his, dad |
| 33 | Cognitive processes | cogproc | cause, know, ought |
| 34 | Insight | insight | think, know |
| 35 | Causation | cause | because, effect |
| 36 | Discrepancy | discrep | should, would |
| 37 | Tentative | tentat | maybe, perhaps |
| 38 | Certainty | certain | always, never |
| 39 | Differentiation | differ | hasn’t, but, else |
| 40 | Perceptual processes | percept | look, heard, feeling |
| 41 | See | see | view, saw, seen |
| 42 | Hear | hear | listen, hearing |
| 43 | Feel | feel | feels, touch |
| 44 | Biological processes | bio | eat, blood, pain |
| 45 | Body | body | cheek, hands, spit |
| 46 | Health | health | clinic, flu, pill |
| 47 | Sexual | sexual | horny, love, incest |
| 48 | Ingestion | ingest | dish, eat, pizza |
| 49 | Drives | drives |  |
| 50 | Affiliation | affiliation | ally, friend, social |
| 51 | Achievement | achieve | win, success, better |
| 52 | Power | power | superior, bully |
| 53 | *Reward* | *reward* | take, prize, benefit |
| 54 | Risk | risk | danger, doubt |
| 55 | Past focus | focuspast | ago, did, talked |
| 56 | Present focus | focuspresent | today, is, now |
| 57 | Future focus | focusfuture | may, will, soon |
| 58 | Relativity | relativ | area, bend, exit |
| 59 | Motion | motion | arrive, car, go |
| 60 | Space | space | down, in, thin |
| 61 | Time | time | end, until, season |
|  | Personal concerns |  |  |
| 62 | Work | work | job, majors, xerox |
| 63 | Leisure | leisure | cook, chat, movie |
| 64 | Home | home | kitchen, landlord |
| 65 | Money | money | audit, cash, owe |
| 66 | Religion | relig | altar, church |
| 67 | Death | death | bury, coffin, kill |
| 68 | Informal language | informal |  |
| 69 | Swear words | swear | fuck, damn, shit |
| 70 | Netspeak | netspeak | btw, lol, thx |
| 71 | Assent | assent | agree, OK, yes |
| 72 | Nonfluencies | nonflu | er, hm, umm |
| 73 | Fillers | filler | Imean, youknow |
|  | **Punctuation\*** |  |  |
| 74 | Total Punctuation | allpunc |  |
| 75 | Periods | period | **.** |
| 76 | Commas | comma | **,** |
| 77 | Colons | colon | **:** |
| 78 | Semicolons | semic | **;** |
| 79 | Question marks | qmark | **?** |
| 80 | Exclamation marks | exclam | **!** |
| 81 | Dashes | dash | **-** |
| 82 | Quotation marks | quote | **“”** |
| 83 | Apostrophes | apostro | **‘** |
| 84 | Parentheses | parenth | **( )** |
| 85 | Other punctuation | otherp |  |

Tabel 3.3 diatas menunjukkan seluruh fitur LIWC yang digunakan dalam penelitian ini. Terdapat 85 fitur dengan 5 subkategori yaitu, *Linguistic Dimensions*, *Other Grammars*, *Psychological Process*, *Personal Concerns* dan *Punctuation*. Fitur ini merupakan fitur LIWC2015 yang terbaru daripada fitur LIWC2007 yang hanya memiliki sekitar 70 fitur. Proses ekstraksi fitur langsung diimplementasikan terhadap status dari *user* yang telah di *preprocessing* seperti pada Gambar 3.25.

**Gambar 3.25** Contoh status dari user #250

Last night was amazing! Not only did I see \*PROPNAME\* and \*PROPNAME\*'s LAST performance of RENT, but we also ended the night at 3 Fires Lounge were the whole cast was celebrating!

Status diatas yang telah diekstraksi ke dalam fitur LIWC akan memiliki nilai masing-masing per fiturnya dan dapat dilihat pada Tabel 3.4 di bawah ini.

**Tabel 3.4** Hasil ekstraksi fitur LIWC dari user #250

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Features**  **Abbrev** | **Scores** | **Features**  **Abbrev** | **Scores** | **Features**  **Abbrev** | **Scores** |
| semic | 0 | adverb | 0.0606061 | leisure | 0.06061 |
| relig | 0 | space | 0.030303 | anger | 0 |
| compare | 0 | informal | 0 | verb | 0.18182 |
| family | 0 | ipron | 0 | hear | 0 |
| qmark | 0 | anx | 0 | focuspast | 0.15152 |
| feel | 0.00 | focuspresent | 0.030303 | they | 0 |
| money | 0.03 | nonflu | 0 | affect | 0.0303 |
| insight | 0.00 | power | 0 | allpunc | 0.24242 |
| assent | 0.00 | netspeak | 0 | sad | 0 |
| number | 0.03 | percept | 0.030303 | you | 0 |
| comma | 0.03 | quant | 0.030303 | tentat | 0 |
| parenth | 0.00 | posemo | 0.030303 | apostro | 0.0303 |
| time | 0.15 | certain | 0 | reward | 0 |
| affiliation | 0.060606062 | relativ | 0.1818182 | i | 0.0303 |
| cogproc | 0.060606062 | health | 0 | cause | 0 |
| otherp | 0.121212125 | exclam | 0.0606061 | work | 0.0303 |
| female | 0 | adj | 0.030303 | period | 0 |
| article | 0.060606062 | prep | 0.0606061 | ingest | 0 |
| negate | 0.030303031 | achieve | 0.030303 | dash | 0 |
| home | 0.030303031 | function | 0.4545455 | filler | 0 |
| conj | 0.09090909 | bio | 0 | swear | 0 |
| sexual | 0 | we | 0.030303 | colon | 0 |
| negemo | 0 | risk | 0 | friend | 0 |
| ppron | 0.060606062 | see | 0.030303 | focusfuture | 0 |
| motion | 0 | interrog | 0 | quote | 0 |
| differ | 0.060606062 | discrep | 0 | auxverb | 0.12121 |
| death | 0 | body | 0 | male | 0 |
| pronoun | 0.060606062 | drives | 0.0606061 | shehe | 0 |
| social | 0.060606062 |  |  |  |  |

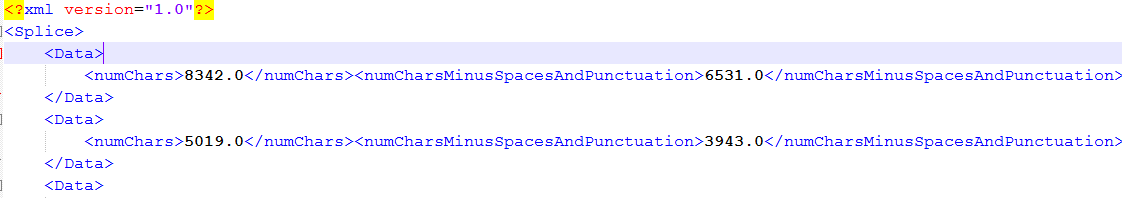
1. **Splice**

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. Fitur yang digunakan sebanyak 94 fitur dengan 14 subfitur dapat dilihat pada Tabel 3.5.

**Tabel 3.5** Fitur dari SPLICE yang digunakan dalam penelitian ini.

|  |  |
| --- | --- |
| **Quantity**  *Cues based on frequencies* [numChars](http://splice.cmi.arizona.edu/APIDetail.php?key=numChars#numChars) Total number of characters [numCharsMinusSpacesAndPunctuation](http://splice.cmi.arizona.edu/APIDetail.php?key=numCharsMinusSpacesAndPunctuation#numCharsMinusSpacesAndPunctuation) Total number of characters minus spaces and punctuation [numWords](http://splice.cmi.arizona.edu/APIDetail.php?key=numWords#numWords) Total number of words [numSentences](http://splice.cmi.arizona.edu/APIDetail.php?key=numSentences#numSentences) Total number of sentences [numPunctuation](http://splice.cmi.arizona.edu/APIDetail.php?key=numPunctuation#numPunctuation) Total number of punctuation marks | **Parts of Speech**  *All linguistic cues in the Parts of Speech category are calculated using a part of speech tagger based on the Brown corpus.* [numNouns](http://splice.cmi.arizona.edu/APIDetail.php?key=numNouns#numNouns) Total number of nouns [nounRatio](http://splice.cmi.arizona.edu/APIDetail.php?key=nounRatio#nounRatio) The ratio of nouns to the total number of words [numVerbs](http://splice.cmi.arizona.edu/APIDetail.php?key=numVerbs#numVerbs) Total number of verbs [verbRatio](http://splice.cmi.arizona.edu/APIDetail.php?key=verbRatio#verbRatio) The ratio of verbs to the total number of words [numAdjectives](http://splice.cmi.arizona.edu/APIDetail.php?key=numAdjectives#numAdjectives) Total number of adjectives [adjectiveRatio](http://splice.cmi.arizona.edu/APIDetail.php?key=adjectiveRatio#adjectiveRatio) The ratio of adjectives to the total number of words [numAdverbs](http://splice.cmi.arizona.edu/APIDetail.php?key=numAdverbs#numAdverbs) Total number of adverbs [adverbRatio](http://splice.cmi.arizona.edu/APIDetail.php?key=adverbRatio#adverbRatio) The ratio of adverbs to the total number of words |
| **Immediacy**  *Cues that indicate immediacy* [numPassiveVerbs](http://splice.cmi.arizona.edu/APIDetail.php?key=numPassiveVerbs#numPassiveVerbs) Total number of passive verbs [passiveVerbRatio](http://splice.cmi.arizona.edu/APIDetail.php?key=passiveVerbRatio#passiveVerbRatio) The ratio of passive voice verbs to the total number of verbs | **Pronouns**  *Cues that count pronoun frequencies* [firstPersonSingular](http://splice.cmi.arizona.edu/APIDetail.php?key=firstPersonSingular#firstPersonSingular) Total number of first person singular words [firstPersonSingularRatio](http://splice.cmi.arizona.edu/APIDetail.php?key=firstPersonSingularRatio#firstPersonSingularRatio) Total number of first person singular words divided by the total number of words [firstPersonPlural](http://splice.cmi.arizona.edu/APIDetail.php?key=firstPersonPlural#firstPersonPlural) Total number of first person plural words [firstPersonPluralRatio](http://splice.cmi.arizona.edu/APIDetail.php?key=firstPersonPluralRatio#firstPersonPluralRatio) Total number of first person plural words divided by the total number of words [secondPerson](http://splice.cmi.arizona.edu/APIDetail.php?key=secondPerson#secondPerson) Total number of second person words [secondPersonRatio](http://splice.cmi.arizona.edu/APIDetail.php?key=secondPersonRatio#secondPersonRatio) Total number of second person words divided by the total number of words [thirdPersonSingular](http://splice.cmi.arizona.edu/APIDetail.php?key=thirdPersonSingular#thirdPersonSingular) Total number of third person singluar words [thirdPersonSingularRatio](http://splice.cmi.arizona.edu/APIDetail.php?key=thirdPersonSingularRatio#thirdPersonSingularRatio) Total number of third person singluar words divided by the total number of words [thirdPersonPlural](http://splice.cmi.arizona.edu/APIDetail.php?key=thirdPersonPlural#thirdPersonPlural) Total number of third person plural words [thirdPersonPluralRatio](http://splice.cmi.arizona.edu/APIDetail.php?key=thirdPersonPluralRatio#thirdPersonPluralRatio) Total number of third person plural words divided by the total number of words |
| **Positive Self Evaluation**  *Cues that relate to the positive self-evaluation of a speaker.* [iCanDoIt](http://splice.cmi.arizona.edu/APIDetail.php?key=iCanDoIt#iCanDoIt) The total number of "I can" phrases in the text. [doKnow](http://splice.cmi.arizona.edu/APIDetail.php?key=doKnow#doKnow) The total number of times phrases such as "I know", "I am sure", and "I'm positive" appear in the text. [posSelfImage](http://splice.cmi.arizona.edu/APIDetail.php?key=posSelfImage#posSelfImage) The total number of times someone positively portrays themself in the text (e.g., I am a great guy or I am happy). | **Negative Self Evaluation**  *Cues that relate to the negative self-evaluation of a speaker.* [iCantDoIt](http://splice.cmi.arizona.edu/APIDetail.php?key=iCantDoIt#iCantDoIt) The total number of "I can't" phrases in the text. [dontKnow](http://splice.cmi.arizona.edu/APIDetail.php?key=dontKnow#dontKnow) The total number of times someone says they don't know something in the text. [negSelfImage](http://splice.cmi.arizona.edu/APIDetail.php?key=negSelfImage#negSelfImage) The total number of times someone portrays themself with a negative self image in the text (e.g., I am ugly or I'm stupid). |
| **Influence**  *Cues that indicate the speaker is attempting to influence someone.* [numImperatives](http://splice.cmi.arizona.edu/APIDetail.php?key=numImperatives#numImperatives) The total number of imperative statements in the text. [suggestionPhrases](http://splice.cmi.arizona.edu/APIDetail.php?key=suggestionPhrases#suggestionPhrases) The total number of suggestions in the text. [inflexibility](http://splice.cmi.arizona.edu/APIDetail.php?key=inflexibility#inflexibility) The total number of "inflexible" statements in the text. [contradict](http://splice.cmi.arizona.edu/APIDetail.php?key=contradict#contradict) The total number of times someone contradicts what has been stated. [totalDominance](http://splice.cmi.arizona.edu/APIDetail.php?key=totalDominance#totalDominance) Dominance language is a summation of inflexible language, imperative sentences, positive self evaluation, suggestions, and inflexible language. [dominanceRatio](http://splice.cmi.arizona.edu/APIDetail.php?key=dominanceRatio#dominanceRatio) The total number of sentences that include dominance language divided by the total number of sentences. [numAgreement](http://splice.cmi.arizona.edu/APIDetail.php?key=numAgreement#numAgreement) The total number of words that signify that the person accepts another's propositions. [agreementRatio](http://splice.cmi.arizona.edu/APIDetail.php?key=agreementRatio#agreementRatio) The total number of words that signify that the person accepts another's propositions divided by the total number of words. | **Deference**  *Cues that indicate the speaker is accommodating outside influence.* [askPermission](http://splice.cmi.arizona.edu/APIDetail.php?key=askPermission#askPermission) The total number of times permission is asked for in the text. [seekGuidance](http://splice.cmi.arizona.edu/APIDetail.php?key=seekGuidance#seekGuidance) The total number of times guidance is sought in the text. [totalSubmissiveness](http://splice.cmi.arizona.edu/APIDetail.php?key=totalSubmissiveness#totalSubmissiveness) The total number of instances of submissiveness language. [submissivenessRatio](http://splice.cmi.arizona.edu/APIDetail.php?key=submissivenessRatio#submissivenessRatio) The total number of sentences that include submissiveness language divided by the total number of sentences. |
| **Whissel**  *Cues that count returns scores based on the Whissel Dictionary of Affect in Language* [Imagery](http://splice.cmi.arizona.edu/APIDetail.php?key=Imagery#Imagery) Average imagery score. [Pleasantness](http://splice.cmi.arizona.edu/APIDetail.php?key=Pleasantness#Pleasantness) Average pleasantness score. [Activation](http://splice.cmi.arizona.edu/APIDetail.php?key=Activation#Activation) Average activation score. | **Complexity**  *Cues that relate to text complexity* [avgWordLength](http://splice.cmi.arizona.edu/APIDetail.php?key=avgWordLength#avgWordLength) Total number of characters minues spaces and punctuation divided by total number of words [avgSentenceLength](http://splice.cmi.arizona.edu/APIDetail.php?key=avgSentenceLength#avgSentenceLength) Total number of sentences divided by total number of words [numSyllables](http://splice.cmi.arizona.edu/APIDetail.php?key=numSyllables#numSyllables) Count of the total number of syllables [avgSyllablesPerWord](http://splice.cmi.arizona.edu/APIDetail.php?key=avgSyllablesPerWord#avgSyllablesPerWord) Total number of syllables divided by total number of words [numWordsWith3OrMoreSyllables](http://splice.cmi.arizona.edu/APIDetail.php?key=numWordsWith3OrMoreSyllables#numWordsWith3OrMoreSyllables) Total number of words with three or more syllables. [rateWordsWith3OrMoreSyllables](http://splice.cmi.arizona.edu/APIDetail.php?key=rateWordsWith3OrMoreSyllables#rateWordsWith3OrMoreSyllables) Total number of words with three or more syllables divided by total number of words. [numWordsWith6OrMoreChars](http://splice.cmi.arizona.edu/APIDetail.php?key=numWordsWith6OrMoreChars#numWordsWith6OrMoreChars) Total number of words with six or more characters. [rateWordsWith6OrMoreChars](http://splice.cmi.arizona.edu/APIDetail.php?key=rateWordsWith6OrMoreChars#rateWordsWith6OrMoreChars) Total number of words with six or more characters divided by total number of words. [numWordsWith7OrMoreChars](http://splice.cmi.arizona.edu/APIDetail.php?key=numWordsWith7OrMoreChars#numWordsWith7OrMoreChars) Total number of words with seven or more characters. [rateWordsWith7OrMoreChars](http://splice.cmi.arizona.edu/APIDetail.php?key=rateWordsWith7OrMoreChars#rateWordsWith7OrMoreChars) Total number of words with seven or more characters divided by total number of words. [LexicalDiversity](http://splice.cmi.arizona.edu/APIDetail.php?key=LexicalDiversity#LexicalDiversity) The length of a Set of all the words divided by the total number of words. [complexityComposite](http://splice.cmi.arizona.edu/APIDetail.php?key=complexityComposite#complexityComposite) Number of three syllable words + number of commas + number of conjunctions + number of singular and mass nouns + number of plural nouns + average sentance length all divided by the number of words. |
| **Spoken Word**  *Cues that relate to a speaking or writing style* [hedgeVerb](http://splice.cmi.arizona.edu/APIDetail.php?key=hedgeVerb#hedgeVerb) Total number of verbs that indicate hedging [hedgeConj](http://splice.cmi.arizona.edu/APIDetail.php?key=hedgeConj#hedgeConj) Total number of conjunctions that indicate hedging [hedgeAdj](http://splice.cmi.arizona.edu/APIDetail.php?key=hedgeAdj#hedgeAdj) Total number of adjectives that indicate hedging [hedgeModal](http://splice.cmi.arizona.edu/APIDetail.php?key=hedgeModal#hedgeModal) Total number of modal verbs that indicate hedging [hedgeAll](http://splice.cmi.arizona.edu/APIDetail.php?key=hedgeAll#hedgeAll) Total number of words that indicate hedging [numDisfluencies](http://splice.cmi.arizona.edu/APIDetail.php?key=numDisfluencies#numDisfluencies) The total number of disfluencies [disfluencyRatio](http://splice.cmi.arizona.edu/APIDetail.php?key=disfluencyRatio#disfluencyRatio) The rate at which disfluencies appear [numInterjections](http://splice.cmi.arizona.edu/APIDetail.php?key=numInterjections#numInterjections) Total number of interjections [interjectionRatio](http://splice.cmi.arizona.edu/APIDetail.php?key=interjectionRatio#interjectionRatio) The ratio of total interjections to the total number of words [numSpeculate](http://splice.cmi.arizona.edu/APIDetail.php?key=numSpeculate#numSpeculate) The number of speculative words [speculateRatio](http://splice.cmi.arizona.edu/APIDetail.php?key=speculateRatio#speculateRatio) The ratio of speculative words to total words [Expressivity](http://splice.cmi.arizona.edu/APIDetail.php?key=Expressivity#Expressivity) Total number of adverbs and adjectives divided by the total number of nouns and verbs [numIgnorance](http://splice.cmi.arizona.edu/APIDetail.php?key=numIgnorance#numIgnorance) Total number of phrases indicating ignorance [ignoranceRatio](http://splice.cmi.arizona.edu/APIDetail.php?key=ignoranceRatio#ignoranceRatio) Total number of phrases indicating ignorance divided by the total number of words [Pausality](http://splice.cmi.arizona.edu/APIDetail.php?key=Pausality#Pausality) Total number of sentences divided by the total number of punctuation marks [questionCount](http://splice.cmi.arizona.edu/APIDetail.php?key=questionCount#questionCount) Total number of question marks [questionRatio](http://splice.cmi.arizona.edu/APIDetail.php?key=questionRatio#questionRatio) Total number of question marks divided by the total number of sentences [hedgeUncertain](http://splice.cmi.arizona.edu/APIDetail.php?key=hedgeUncertain#hedgeUncertain) A combination of indefinites, demonstratives, hedgind words, uncertainty words, and the Loughran and McDonald (2011) uncertainty and weak modal dictionaries. | **Tense**  *Counts of verbs of a specific tense* [pastTense](http://splice.cmi.arizona.edu/APIDetail.php?key=pastTense#pastTense) Total number of verbs in the past tense [pastTenseRatio](http://splice.cmi.arizona.edu/APIDetail.php?key=pastTenseRatio#pastTenseRatio) Total number of verbs in the past tense divided by the total number of words [presentTense](http://splice.cmi.arizona.edu/APIDetail.php?key=presentTense#presentTense) Total number of verbs in the present tense [presentTenseRatio](http://splice.cmi.arizona.edu/APIDetail.php?key=presentTenseRatio#presentTenseRatio) Total number of verbs in the present tense divided by the total number of words. |
| **Sentiwordnet**  *Calculates positivity, negativity, and objectivity based on Sentiwordnet* [SWNpositivity](http://splice.cmi.arizona.edu/APIDetail.php?key=SWNpositivity#SWNpositivity) Average positivity scored based on Sentiwordnet [SWNnegativity](http://splice.cmi.arizona.edu/APIDetail.php?key=SWNnegativity#SWNnegativity) Average negativity scored based on Sentiwordnet [SWNobjectivity](http://splice.cmi.arizona.edu/APIDetail.php?key=SWNobjectivity#SWNobjectivity) Average objectivity scored based on Sentiwordnet | **Readability**  *Cues that return readability scores; readability scores should be calculated for text passages at least 100 words long* [ARI](http://splice.cmi.arizona.edu/APIDetail.php?key=ARI#ARI) ARI reading index score [FRE](http://splice.cmi.arizona.edu/APIDetail.php?key=FRE#FRE) FRE reading index score [FKGL](http://splice.cmi.arizona.edu/APIDetail.php?key=FKGL#FKGL) FKGL reading index score [CLI](http://splice.cmi.arizona.edu/APIDetail.php?key=CLI#CLI) CLI reading index score [LWRF](http://splice.cmi.arizona.edu/APIDetail.php?key=LWRF#LWRF) LWRF reading index score [FOG](http://splice.cmi.arizona.edu/APIDetail.php?key=FOG#FOG) FOG reading index score [SMOG](http://splice.cmi.arizona.edu/APIDetail.php?key=SMOG#SMOG) SMOG reading index score [DALE](http://splice.cmi.arizona.edu/APIDetail.php?key=DALE#DALE) DALE reading index score [LIX](http://splice.cmi.arizona.edu/APIDetail.php?key=LIX#LIX) LIX reading index score [RIX](http://splice.cmi.arizona.edu/APIDetail.php?key=RIX#RIX) RIX reading index score [FRY](http://splice.cmi.arizona.edu/APIDetail.php?key=FRY#FRY) FRY reading index score |

Untuk mengekstraksi nilai dari hasil fitur-fitur SPLICE di atas, peneliti menyiapkan code untuk meng-convert status dari user menjadi fitur SPLICE dengan API yang disediakan oleh SPLICE. Hasil dari fitur tersebut kemudian disimpan ke dalam file XML seperti pada Gambar 3.25 di bawah ini.



**Gambar 3.25** *Screenshot file* XML hasil 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*



**Gambar 3.26** Screenshot hasil fitur SNA dari salah satu *user*

yang disediakan oleh dataset myPersonality

1. **Open Vocabulary**

Metode *Open Vocabulary* ini berbeda dengan fitur-fitur lainnya dimana semua nilai dari fitur di define diawal sebelum proses training dilakukan. Metode *Open Vocabulary* tidak memiliki jumlah fitur pasti, tetapi fitur didapatkan dengan menelusuri dataset yang kita gunakan. Kata-kata dari dataset yang kita kumpulkan dikelompokkan dan dihitung jumlahnya WC (*word count*). Kemudian WC tersebut di *embed* dengan menggunakan GloVe untuk mendapatkan representasi vektor dari kata tersebut. Setiap representasi vektor itulah yang dikembangkan menjadi fitur dan diimplementasikan ke dalam sistem untuk kemudian dikorelasikan 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. Dalam penelitian ini, penggunaan Open Vocabulary hanya akan diimplementasikan sebagai fitur *deep learning.*

**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. Algoritma yang digunakan dalam penelitian ini adalah:

* SVM (*Support Vector Machine*)
* *Naïve Bayes*
* *Logistic Regression*
* *Linear Discriminant Analysis*
* *Gradient Boosting*

Algoritma diatas akan dibuat dengan menggunakan *library* scikit-learn dengan metode 10-*fold* *cross validation*.

**2. *Deep learning***

Setelah mendapatkan hasil dari proses *learning* tersebut. Peneliti kemudian mencoba kembali menggunakan proses *Deep Learning* dengan beberapa arsitektur yaitu:

* MLP (*Multilayer Perceptron*)
* LSTM (*Long Short Term Memory*)
* CNN 1D (*Convolutional Neural Network* *1-Dimension*)
* GRU (*Gated Recurrent Unit*)

Sistem *Deep Learning* akan dibangun dengan *library* Keras dan *scikit-learn*. Sedangkan *backend* yang digunakan adalah Theano.

Hasil dari penggunaan fitur yang berbeda dan proses learning yang berbeda akan terus dibandingkan dan dikembangkan untuk mendapatkan sistem prediksi yang paling akurat.

Setelah melalui proses pengambilan dataset, feature selection dan mengetahui model learning yang akan digunakan. Tabel 3.6 akan menampilkan pembagian keseluruhan metode *feature selection* yang akan digunakan di setiap dataset dan *model learning process.* Proses learning yang akan dilakukan nanti akan terbagi menjadi beberapa skenario. Masing-masing metode *learning* akan dicoba menggunakan 3 skenario dataset yang berbeda, yaitu dataset yang diperoleh dari myPersonality, dataset yang diperoleh secara manual oleh peneliti, kemudian percobaan dengan menggunakan gabungan dari kedua dataset tersebut. Ketiga skenario dari metode *machine learning* akan diuji dengan menggunakan 3 fitur yang berbeda yaitu fitur LIWC, fitur SPLICE dan fitur SNA dengan implementasi 5 algoritma *machine learning* yaitu Support Vector Machine (SVM), *Naïve Bayes*, *Logistic Regression*, *Linear Discriminant Analysis* (LDA), dan *Gradient Boosting* di tiap-tiap skenario tersebut. Sedangkan tiga skenario *deep learning* hanya akan menggunakan fitur *Open Vocabulary* dengan implementasi 4 arsitektur *deep learning* yaitu MLP, LSTM, CNN, dan GRU.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 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 3.6** *Features Selection* berdasarkan *model learning process*

dan dataset yang digunakan

**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 tinggi. Evaluasi testing akan dibagi menjadi dua tahapan yaitu evaluasi testing secara subjektif dan secara objektif. Evaluasi secara subjektif akan dilakukan dengan melakukan percobaan sistem secara langsung terhadap beberapa *user* atau responden dan meminta pendapat mereka mengenai keakuratan sistem tersebut dengan apa yang user pikirkan mengenai kepribadian itu. Sedangkan secara objektif dengan membandingkan hasil akurasi dengan penelitian sejenis yang telah dilakukan sebelumnya.

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 3.27** 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 3.28** 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 3.29** Rancangan Layar *Result* (1)



**Gambar 3.30** 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 3.30 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.