

## BAB 3

### METODOLOGI

#### 3.1. 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

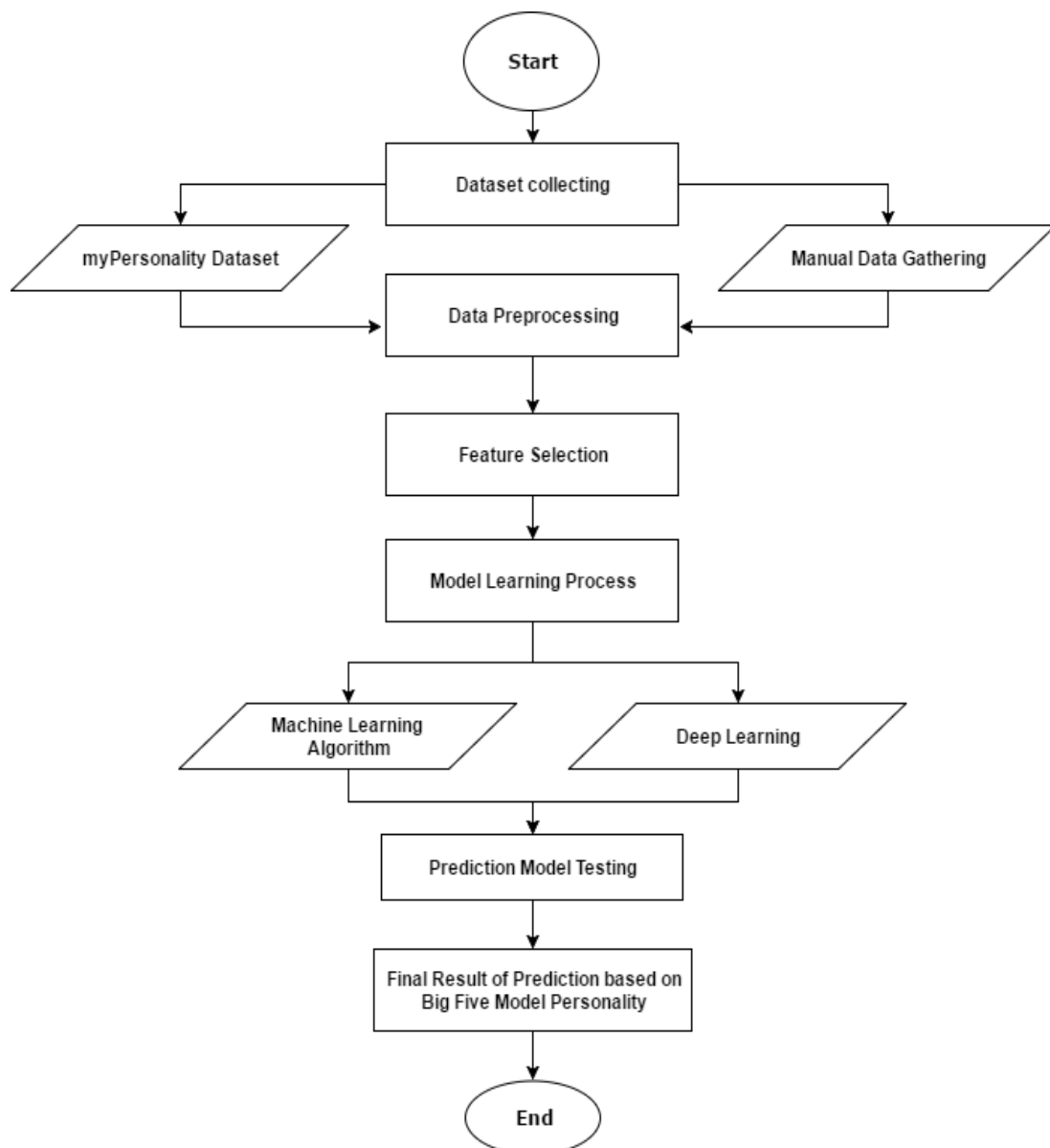
mendapatkan dataset dari *user*. Beberapa penelitian yang menggunakan sosial media Facebook juga masih memiliki akurasi yang tergolong rendah.

### 3.2. 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 untuk melihat perbandingan hasil dengan *Machine Learning*.
3. Melakukan *preprocessing* dan normalisasi pada dataset *training*.
4. Mencoba penerapan *Resampling* untuk menyeimbangkan ketidakseimbangan distribusi *traits* dari dataset.

### 3.3. 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 akun kita sendiri atau akun lain selama kita memiliki *authentication code* dari akun tersebut. Maka untuk mendapatkan dataset yang lebih besar, peneliti menggunakan data *user* Facebook dari myPersonality (Kosinski, 2015). myPersonality adalah aplikasi Facebook yang dikembangkan oleh Michal Kosinski 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

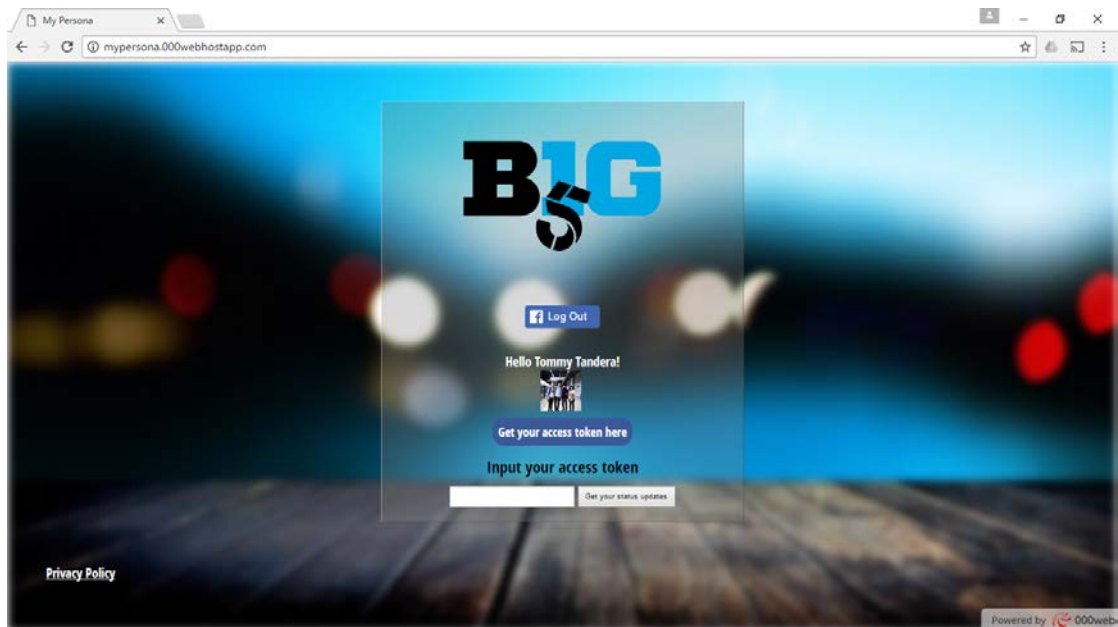
**Tabel 3.1** Distribusi jenis kepribadian berdasarkan 250 *user* dalam dataset myPersonality (Kosinski, 2015).

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*

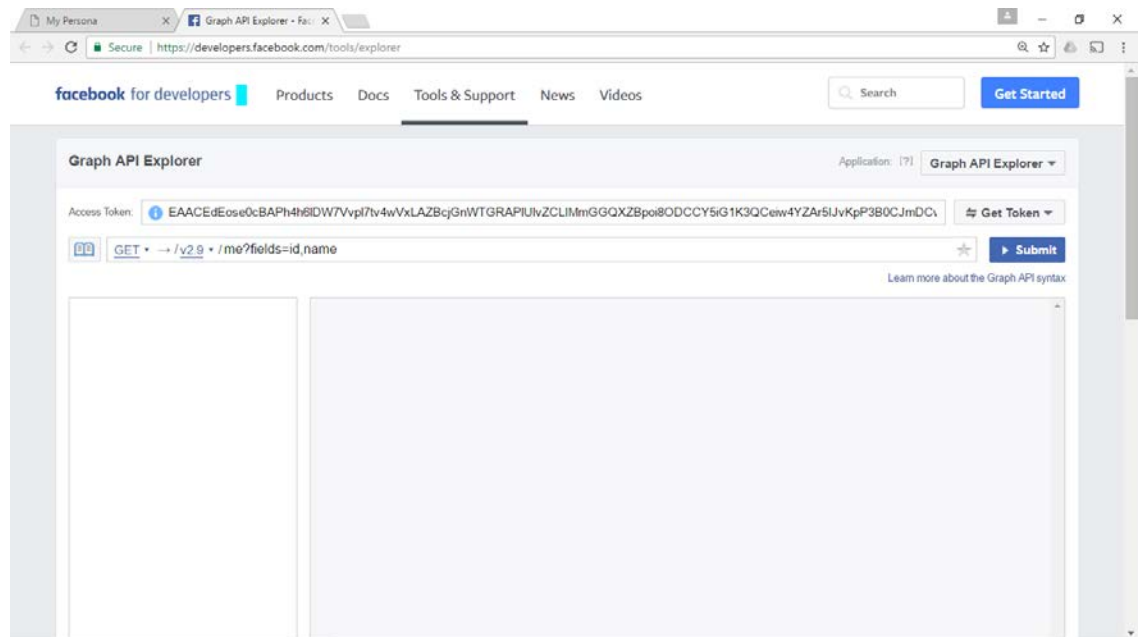
Selain mengambil data dari myPersonality, pengambilan data juga dilakukan dengan cara manual dengan memanfaatkan Graph API Facebook (Gambar 3.3) yang dapat memberikan akses data terhadap sebuah account selama *user* menyetujui untuk memberikan *access* tokennya. Pada Gambar 3.2 terlihat sebuah *interface*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

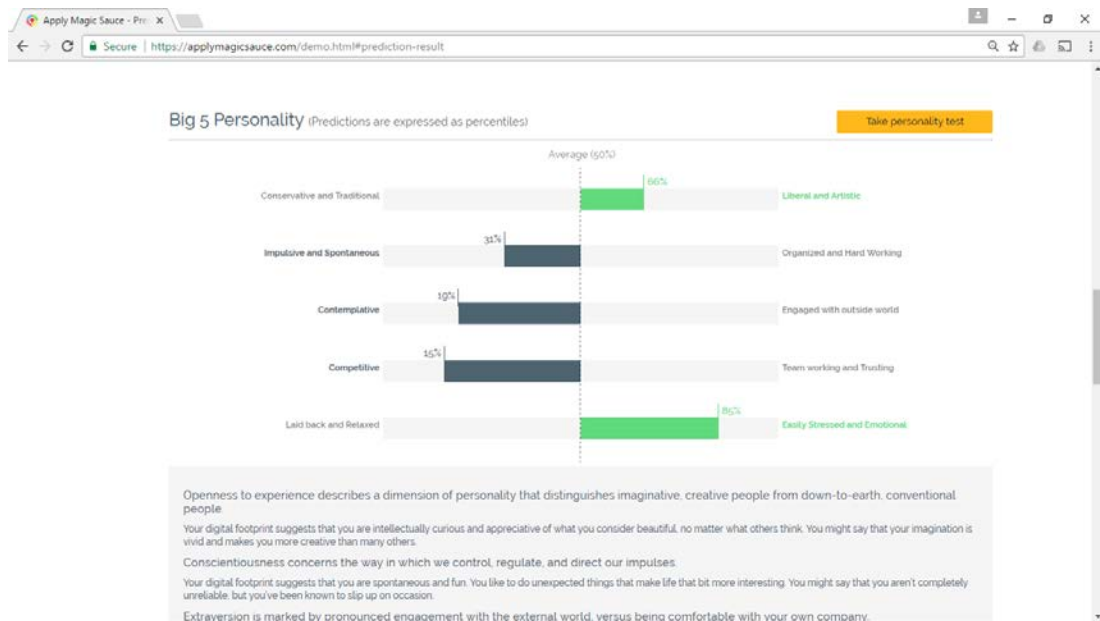


**Gambar 3.3** Tampilan *Graph API* untuk mendapatkan *access token* *user*

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*

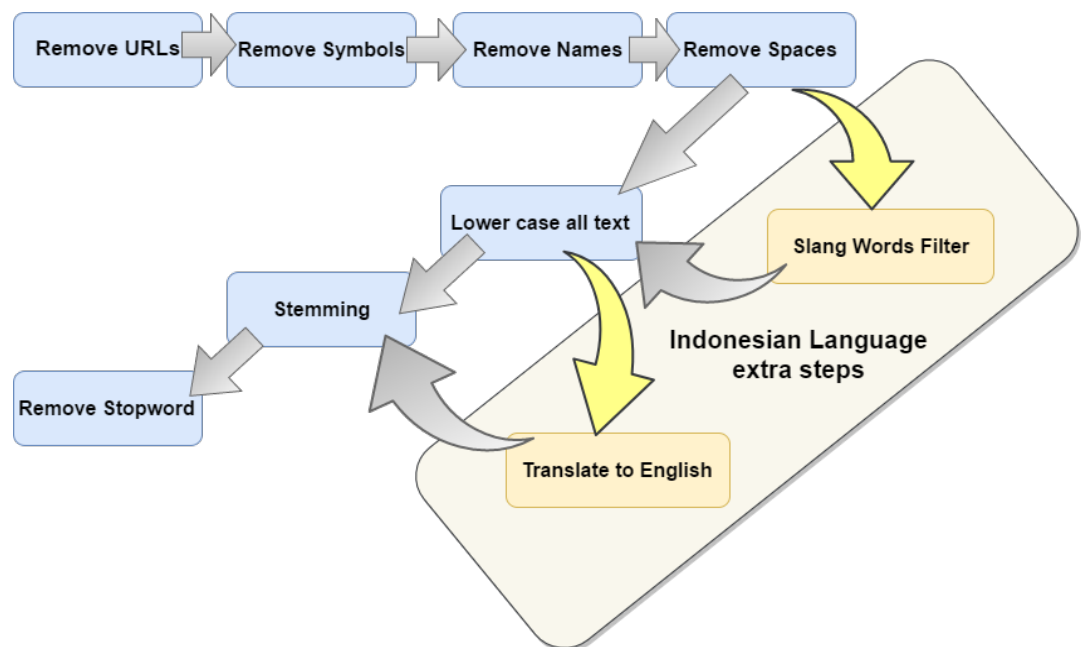
Tabel 3.2 di atas menampilkan bahwa dari 150 *user* yang di dapatkan, terdapat 87 *user* yang dominan terhadap *traits openness* dan yang rendah di *traits openness* sebanyak 52. Sebanyak 56 *user* tinggi di *traits conscientiousness* dan 83 *user* rendah di *traits* ini. Di *traits extraversion*, perbandingan 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

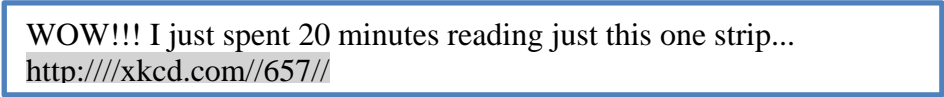
Dataset yang telah diperoleh dari myPersonality maupun *manual data gathering* selanjutnya masuk ke dalam proses data *preprocessing* sebelum dilanjutkan ke tahap feature selection dan training. Proses *preprocessing* data ini terbagi dua karena dataset yang peneliti miliki berbeda dimana dataset yang 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 Data Preprocessing*

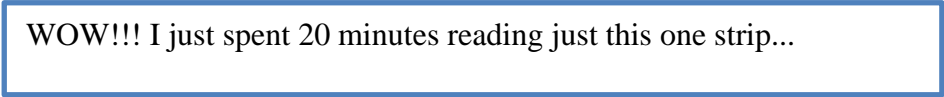
### 1. *Remove URLs*

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

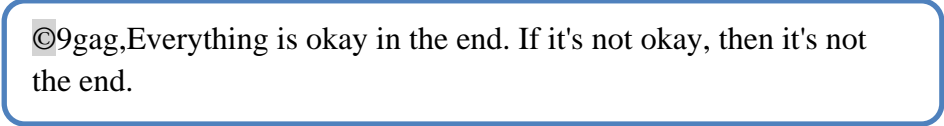


WOW!!! I just spent 20 minutes reading just this one strip...

**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 berlebihan. 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 Lampiran 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 kata 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 Lampiran 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
	<b>Linguistic Dimensions</b>		
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
	<b>Other Grammar</b>		
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
	<b>Psychological Processes</b>		
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	<i>Reward</i>	<i>reward</i>	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
	<b>Personal concerns</b>		
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	I mean, you know
	<b>Punctuation*</b>		
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.

likes the sound of thunder.  
 is so sleepy it's not even funny that's she can't get to sleep.  
 is sore and wants the knot of muscles at the base of her neck to stop hurting. On the  
 other hand, YAY I'M IN ILLINOIS! <3

**Gambar 3.25** Contoh penggalan status dari *user* #1

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* #1

Features Abbrev	Scores	Features Abbrev	Scores	Features Abbrev	Scores
semic	0.0	adverb	0.05570118	leisure	0.017038008
relig	0.001310616	space	0.066841416	anger	0.008519004
compare	0.019003931	informal	0.01965924	verb	0.1802097

family	0.001965924	ipron	0.035386633	hear	0.007208388
qmark	0.008519004	anx	0.00327654	focuspast	0.031454783
feel	0.005897772	focuspresent	0.13237222	they	0.001965924
money	0.003931848	nonflu	0.001965924	affect	0.06815203
insight	0.022280471	power	0.020969857	allpunc	0.23984273
assent	0.007863696	netspeak	0.008519004	sad	0.001965924
number	0.030799476	percept	0.02948886	you	0.0091743115
comma	0.044560943	quant	0.018348623	tentat	0.0163827
parenth	0.005897772	posemo	0.046526868	apostro	0.02293578
time	0.0740498	certain	0.015072084	reward	0.0124508515
affiliation	0.0163827	relativ	0.15661861	i	0.0163827
cogproc	0.106159896	health	0.007208388	cause	0.015072084
otherp	0.035386633	exclam	0.00982962	work	0.024901703
female	0.04521625	adj	0.05242464	period	0.0989515
article	0.061598953	prep	0.13630407	ingest	0.005242464
negate	0.0163827	achieve	0.011795544	dash	0.007863696
home	0.011140236	function	0.48492792	filler	0.000655308
conj	0.056356486	bio	0.034076016	swear	0.001965924
sexual	0.000655308	we	0.001965924	colon	0.005897772
negemo	0.01965924	risk	0.007863696	friend	0.001310616
ppron	0.07339449	see	0.013761468	focusfuture	0.022280471
motion	0.017693317	interrog	0.010484928	quote	0.0
differ	0.02293578	discrep	0.02293578	auxverb	0.09633028
death	0.00327654	body	0.018348623	male	0.001965924
pronoun	0.10878113	drives	0.06422018	shehe	0.043905634
social	0.085845344				

## 2. 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.

<p><b>Quantity</b></p> <p>Fitur Berdasarkan frekuensi atau jumlah.</p> <p><b><u>numChars</u></b> Jumlah karakter</p> <p><b><u>numCharsMinusSpacesAndPunctuation</u></b> Jumlah karakter dikurangi spasi dan tanda baca</p> <p><b><u>numWords</u></b> Jumlah total kata</p> <p><b><u>numSentences</u></b> Jumlah kalimat</p> <p><b><u>numPunctuation</u></b> Jumlah tanda baca</p>	<p><b>Parts of Speech</b></p> <p>Semua fitur linguistik dalam kategori <i>Part of Speech</i> (POS) yang dikalkulasikan dengan POS <i>tagger</i> berdasarkan <i>Brown corpus</i>.</p> <p><b><u>numNouns</u></b> Jumlah kata benda</p> <p><b><u>nounRatio</u></b> Rasio jumlah kata benda dari total kata</p> <p><b><u>numVerbs</u></b> Jumlah kata kerja</p> <p><b><u>verbRatio</u></b> Rasio jumlah kata kerja dari total kata</p> <p><b><u>numAdjectives</u></b> Jumlah kata sifat</p> <p><b><u>adjectiveRatio</u></b> Rasio jumlah kata sifat dari total kata</p> <p><b><u>numAdverbs</u></b> Jumlah kata keterangan</p> <p><b><u>adverbRatio</u></b> Rasio jumlah kata keterangan dari total kata</p>
<p><b>Immediacy</b></p> <p>Fitur yang mengindikasikan kesiapan.</p> <p><b><u>numPassiveVerbs</u></b> Jumlah kata kerja pasif</p> <p><b><u>passiveVerbRatio</u></b> Rasio jumlah kata kerja pasif dari seluruh kata</p>	<p><b>Pronouns</b></p> <p>Fitur berdasarkan jumlah kata ganti.</p> <p><b><u>firstPersonSingular</u></b> Jumlah kata ganti orang pertama</p> <p><b><u>firstPersonSingularRatio</u></b> Jumlah kata ganti orang pertama dibagi jumlah seluruh kata</p> <p><b><u>firstPersonPlural</u></b></p>

	<p>Jumlah kata ganti orang pertama jamak  <u><b>firstPersonPluralRatio</b></u>          Jumlah kata ganti orang pertama jamak dibagi jumlah seluruh kata</p> <p><u><b>secondPerson</b></u>          Jumlah kata ganti orang kedua</p> <p><u><b>secondPersonRatio</b></u>          Jumlah kata ganti orang kedua dibagi jumlah seluruh kata</p> <p><u><b>thirdPersonSingular</b></u>          Jumlah kata ganti orang ketiga</p> <p><u><b>thirdPersonSingularRatio</b></u>          Jumlah kata ganti orang ketiga dibagi jumlah seluruh kata</p> <p><u><b>thirdPersonPlural</b></u>          Jumlah kata ganti orang ketiga jamak</p> <p><u><b>thirdPersonPluralRatio</b></u>          Jumlah kata ganti orang ketiga jamak dibagi jumlah seluruh kata</p>
<p><b>Positive Self Evaluation</b></p> <hr/> <p>Fitur yang berhubungan dengan evaluasi positif pembicara.</p> <p><u><b>iCanDoIt</b></u>          Jumlah dari kata "I can" dalam teks.</p> <p><u><b>doKnow</b></u>          Jumlah kata "I know", "I am sure", dan "I'm positive" yang muncul dalam teks.</p> <p><u><b>posSelfImage</b></u>          Jumlah berapa kali seseorang berpendapat positif mengenai dirinya sendiri (contohnya "I am a great guy" atau "I am happy").</p>	<p><b>Negative Self Evaluation</b></p> <hr/> <p>Fitur yang berhubungan dengan evaluasi negatif pembicara.</p> <p><u><b>iCantDoIt</b></u>          Jumlah dari kata "I can't" dalam teks.</p> <p><u><b>dontKnow</b></u>          Jumlah berapa kali seseorang berkata tidak tahu dalam teks</p> <p><u><b>negSelfImage</b></u>          Jumlah berapa kali seseorang berpendapat negatif mengenai dirinya sendiri (contohnya "I am ugly" atau "I'm stupid").</p>
<p><b>Influence</b></p> <hr/> <p>Fitur yang mengindikasikan pembicara mempengaruhi seseorang.</p> <p><u><b>numImperatives</b></u>          Jumlah pernyataan dalam teks</p> <p><u><b>suggestionPhrases</b></u>          Jumlah saran dalam teks.</p> <p><u><b>inflexibility</b></u></p>	<p><b>Deference</b></p> <hr/> <p>Fitur yang menghitung hasil <i>Whissel Dictionary of Affect</i> dalam bahasa.</p> <p><u><b>askPermission</b></u>          Jumlah kata meminta izin dalam teks</p> <p><u><b>seekGuidance</b></u>          Jumlah kata mencari perhatian dalam teks</p> <p><u><b>totalSubmissiveness</b></u></p>

<p>Jumlah pernyataan yang tidak fleksibel dalam teks</p> <p><b><u>contradict</u></b> Jumlah ketidaksetujuan seseorang dalam teks</p> <p><b><u>totalDominance</u></b> Jumlah bahasa dominasi</p> <p><b><u>dominanceRatio</u></b> Jumlah kalimat bahasa dominasi dibagi jumlah seluruh kalimat</p> <p><b><u>numAgreement</u></b> Jumlah pernyataan persetujuan terhadap pendapat</p> <p><b><u>agreementRatio</u></b> Jumlah kata yang menyatakan persetujuan terhadap pendapat dibagi dengan jumlah seluruh kata</p>	<p>Jumlah kata yang menurut terhadap orang lain</p> <p><b><u>submissivenessRatio</u></b> Jumlah kalimat yang mengandung kata yang menurut dibagi dengan jumlah seluruh kalimat</p>
<p><b><u>Whissel</u></b> Fitur yang menghitung hasil <i>Whissel Dictionary of Affect</i> dalam bahasa.</p> <p><b><u>Imagery</u></b> Nilai rata-rata pencitraan.</p> <p><b><u>Pleasantness</u></b> Nilai rata-rata kepuasan.</p> <p><b><u>Activation</u></b> Nilai rata-rata aktivasi.</p>	<p><b><u>Complexity</u></b> Fitur kompleksitas teks.</p> <p><b><u>avgWordLength</u></b> Jumlah karakter dikurang spasi dan tanda baca dibagi dengan jumlah seluruh kata</p> <p><b><u>avgSentenceLength</u></b> Jumlah seluruh kalimat dibagi jumlah seluruh kata</p> <p><b><u>numSyllables</u></b> Jumlah suku kata</p> <p><b><u>avgSyllablesPerWord</u></b> Jumlah suku kata dibagi jumlah seluruh kata</p> <p><b><u>numWordsWith3OrMoreSyllables</u></b> Jumlah kata-kata dengan lebih dari 3 suku kata</p> <p><b><u>rateWordsWith3OrMoreSyllables</u></b> Jumlah kata-kata dengan lebih dari 3 suku kata dibagi jumlah seluruh kata</p> <p><b><u>numWordsWith6OrMoreChars</u></b> Jumlah kata yang terdiri lebih dari 6 karakter</p> <p><b><u>rateWordsWith6OrMoreChars</u></b> Jumlah kata yang terdiri lebih dari 6 karakter dibagi jumlah seluruh kata.</p>

	<p><b><u>numWordsWith7OrMoreChars</u></b> Jumlah kata yang terdiri lebih dari 7 karakter.</p> <p><b><u>rateWordsWith7OrMoreChars</u></b> Jumlah kata yang terdiri lebih dari 7 karakter dibagi jumlah seluruh kata.</p> <p><b><u>LexicalDiversity</u></b> Panjang sebuah set dari semua kata dibagi jumlah seluruh kata</p> <p><b><u>complexityComposite</u></b> Jumlah kata 3 suku kata + jumlah koma + jumlah kata sambung + jumlah kata benda tunggal + jumlah kata benda jamak + rata-rata panjang kalimat dibagi jumlah seluruh kata</p>
<p><b><u>Spoken Word</u></b> Fitur yang berhubungan dengan cara menulis atau berbicara.</p> <p><b><u>hedgeVerb</u></b> Jumlah kata kerja yang menunjukkan penghindaran</p> <p><b><u>hedgeConj</u></b> Jumlah kata hubung yang menunjukkan penghindaran</p> <p><b><u>hedgeAdj</u></b> Jumlah kata sifat yang menunjukkan penghindaran</p> <p><b><u>hedgeModal</u></b> Jumlah kata bantu yang menunjukkan penghindaran</p> <p><b><u>hedgeAll</u></b> Jumlah semua kata yang menunjukkan penghindaran</p> <p><b><u>numDisfluencies</u></b> Jumlah kata yang menunjukkan kegagapan</p> <p><b><u>disfluencyRatio</u></b> Rasio kata kegagapan</p> <p><b><u>numInterjections</u></b> Jumlah kata seru</p> <p><b><u>interjectionRatio</u></b> Jumlah kata seru dibagi jumlah seluruh kata</p>	<p><b><u>Tense</u></b> Jumlah kata kerja untuk menunjukkan waktu kejadian suatu peristiwa.</p> <p><b><u>pastTense</u></b> Jumlah kata kerja di kejadian lampau</p> <p><b><u>pastTenseRatio</u></b> Jumlah kata kerja di kejadian lampau dibagi jumlah seluruh kata</p> <p><b><u>presentTense</u></b> Jumlah kata kerja di kejadian saat ini</p> <p><b><u>presentTenseRatio</u></b> Jumlah kata kerja di kejadian saat ini dibagi jumlah seluruh kata</p>

<p><b><u>numSpeculate</u></b> Jumlah kata spekulatif</p> <p><b><u>speculateRatio</u></b> Rasio jumlah kata spekulatif</p> <p><b><u>Expressivity</u></b> Jumlah kata keterangan dan kata sifat dibagi dengan jumlah kata benda dan kata kerja</p> <p><b><u>numIgnorance</u></b> Jumlah frasa ketidakpedulian</p> <p><b><u>ignoranceRatio</u></b> Rasio jumlah frasa ketidakpedulian</p> <p><b><u>Pausality</u></b> Jumlah kalimat dibagi jumlah tanda baca</p> <p><b><u>questionCount</u></b> Jumlah tanda tanya</p> <p><b><u>questionRatio</u></b> Jumlah tanda tanya dibagi jumlah seluruh kata</p> <p><b><u>hedgeUncertain</u></b> Kombinasi kata tak terbatas, kata demonstratif, kata lindung, kata ketidakpastian, dan ketidakpastian Loughran dan McDonald (2011) dan kamus kata bantu</p>	
<p><b>Sentiwordnet</b> <i>Kalkulasi positivity, negativity, and objectivity berdasarkan Sentiwordnet.</i></p> <p><b><u>SWNpositivity</u></b> Rata-rata nilai positif berdasarkan Sentiwordnet</p> <p><b><u>SWNnegativity</u></b> Rata-rata nilai negatif berdasarkan Sentiwordnet</p> <p><b><u>SWNobjectivity</u></b> Rata-rata nilai objektifitas berdasarkan Sentiwordnet</p>	<p><b>Readability</b> Fitur yang memberikan hasil <i>readability</i>.</p> <p><b><u>ARI</u></b> Skor indeks membaca ARI</p> <p><b><u>FRE</u></b> Skor indeks membaca FRE</p> <p><b><u>FKGL</u></b> Skor indeks membaca FKGL</p> <p><b><u>CLI</u></b> Skor indeks membaca CLI</p> <p><b><u>LWRF</u></b> Skor indeks membaca LWRF</p> <p><b><u>FOG</u></b> Skor indeks membaca FOG</p>



	<p><b><u>SMOG</u></b> Skor indeks membaca SMOG</p> <p><b><u>DALE</u></b> Skor indeks membaca DALE</p> <p><b><u>LIX</u></b> Skor indeks membaca LIX</p> <p><b><u>RIX</u></b> Skor indeks membaca RIX</p> <p><b><u>FRY</u></b> Skor indeks membaca FRY</p>
--	--

Untuk mengekstraksi nilai dari hasil fitur-fitur SPLICE di atas, peneliti menyiapkan *code* untuk mengubah status dari *user* menjadi fitur SPLICE dengan API yang disediakan oleh SPLICE. Status pada Gambar 3.25 yang telah diekstraksi ke dalam fitur SPLICE akan memiliki nilai masing-masing per fiturnya dan dapat dilihat pada Tabel 3.6 di bawah ini.

**Tabel 3.6** Hasil ekstraksi fitur SPLICE dari Gambar 3.25

Features Abbrev	Scores	Features Abbrev	Scores
numChars	8342.0	numWordsWith3Or MoreSyllables	154.0
numCharsMinusSpaces AndPunctuation	6531.0	rateWordsWith3Or MoreSyllables	0.102122015915
numWords	1508	numWordsWith6Or MoreChars	428.0
numSentences	169	rateWordsWith6Or MoreChars	0.283819628647
numPunctuation	300	numWordsWith7Or MoreChars	313.0
numNouns	646	rateWordsWith7Or MoreChars	0.207559681698
numVerbs	278	lexicalDiversity	0.45

numAdjectives	84	complexityComposite	0.499285860029
numAdverbs	75	hedgeVerb	1
numPassiveVerbs	3	hedgeConj	0
firstPersonSingular	0	hedgeAdj	0
firstPersonPlural	0	hedgeModal	6
secondPerson	0	hedgeAll	7
thirdPersonSingular	0	numDisfluencies	2
thirdPersonPlural	0	numInterjections	10
iCanDoIt	0	numSpeculate	0
doKnow	0	expressivity	0.00309119010819
posSelfImage	0	numIgnorance	0
iCantDoIt	0	pausality	0.563333333333
dontKnow	0	questionCount	13
negSelfImage	0	hedgeUncertain	0
numImperatives	4	pastTense	52
suggestionPhrases	0	presentTense	232
inflexibility	0	swnPositivity	0.0594207723036
contradict	0	swnNegativity	0.0525965379494
totalDominance	4	swnObjectivity	0.887982689747
numAgreement	4	ARI	3.4300862069
askPermission	0	FRE	78.0028779841
seekGuidance	0	FKGL	4.59623342175
totalSubmissiveness	0	CLI	-257.983295756
imagery	1.56049822064	LWRF	4.37278106509
pleasantness	1.86424501779	FOG	7.65411140584
activation	1.66899145907	SMOG	8.58243206864
avgWordLength	4.33090185676	DALE	4.0952096817
avgSentenceLength	8.92307692308	LIX	37.3050397878
numSyllables	2135	RIX	1.85207100592
avgSyllablesPerWord	1.41578249337	FRY	12.6331360947

### 3. *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 aktor 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 disediakan 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 pembandingan dari fitur lainnya sebagai konsiderasi untuk penelitian berikutnya. Fitur dari SNA yang digunakan dalam penelitian ini terbagi atas 7 fitur yaitu:

- *Network size*
- *Betweenness*
- *nBetweenness*
- *Density*
- *Brokerage*
- *nBrokerage*
- *Transitivity*

#AUTHID	STATUS	cEXT	cNEU	cAGR	cCON	cOPN	NETWORKSIZE	BETWEENNESS	NBETWEENNESS	DENSITY	BROKERAGE	NBROKERAGE	TRANSITIVITY
b7b7764cf	likes the sound of thunder.	n	y	n	n	y	180	14861.6	93.29	0.03	15661	0.49	0.1
b7b7764cf	likes how the day sounds in this new song.	n	y	n	n	y	180	14861.6	93.29	0.03	15661	0.49	0.1
b7b7764cf	is home. <3	n	y	n	n	y	180	14861.6	93.29	0.03	15661	0.49	0.1

**Gambar 3.26** *Screenshot* hasil fitur SNA dari salah satu *user* yang disediakan oleh dataset myPersonality

#### 4. 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*.

#### 3.3.4 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 3.7 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 3.7** *Features Selection* berdasarkan *model learning process* dan dataset yang digunakan

### 3.3.5 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.

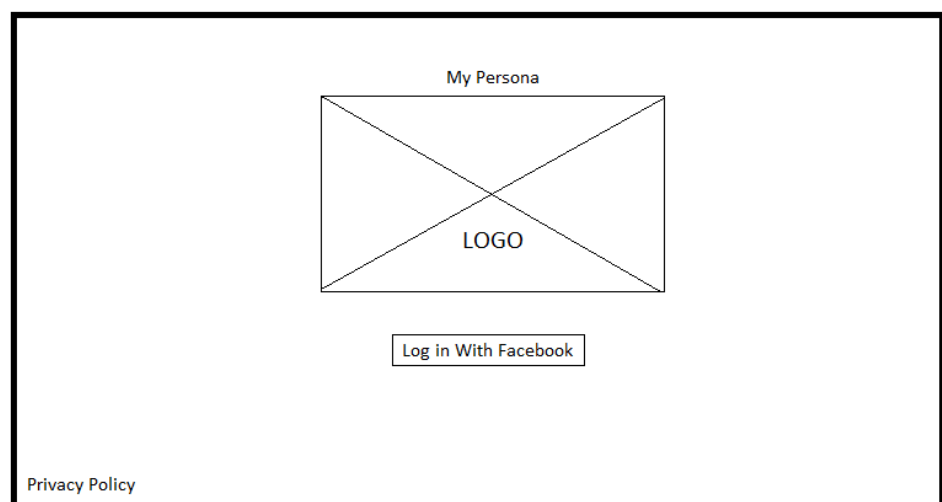
### 3.3.6 *Final Result of Prediction Based on Big Five Model*

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.

## 3.4. 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*



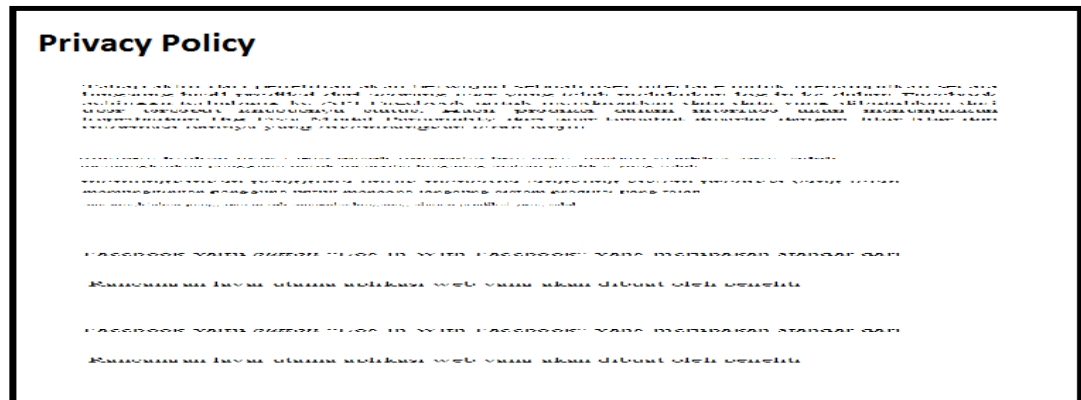
**Gambar 3.27** Rancangan Layar *Homepage*

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*.

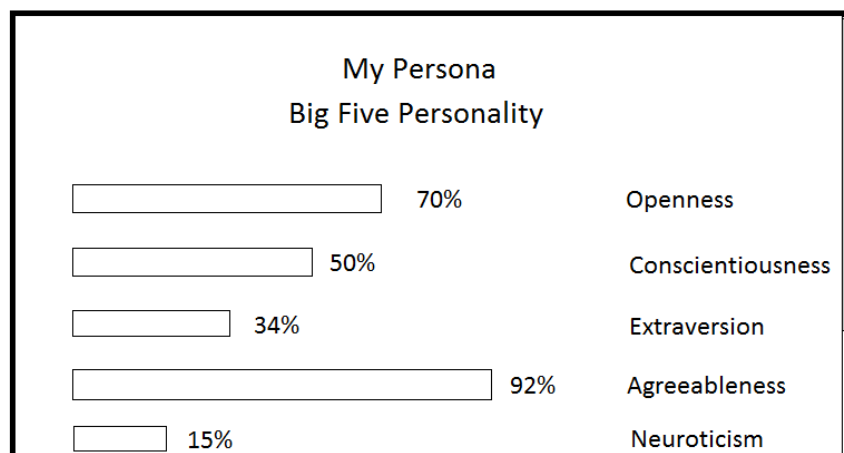


**Gambar 3.28** Rancangan Layar *Privacy Policy*

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

### 3.4.3 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.