Klasifikasi Kepribadian Berbasis Sentimen di Sosial Media Twitter Menggunakan Metode PbSC

Tugas Akhir

diajukan untuk memenuhi salah satu syarat

memperoleh gelar sarjana

dari Program Studi Informatika

Fakultas Informatika

Universitas Telkom

1301164044

Joshua Panjaitan

Program Studi Sarjana Informatika

Fakultas Informatika

Universitas Telkom

Bandung

2020

LEMBAR PENGESAHAN

Klasifikasi Kepribadian Berbasis Sentiment di Sosial Media Twitter Menggunakan Metode PbSC

Personality Classification based on Sentiment in Twitter Social Media using PbSC Methodology

NIM :1301164044

Joshua Panjaitan

Tugas akhir ini telah diterima dan disahkan untuk memenuhi sebagian syarat memperoleh gelar pada Program Studi Sarjana Informatika

Fakultas Informatika

Universitas Telkom

Bandung, 4 Agustus 2020

Menyetujui

Pembimbing I,

Dr.Warih Maharani,S.T.,M.T.

NIP : 01780020

Ketua Program Studi Sarjana Informatika,

Niken Dwi Wahyu Cahyani, Ph.D

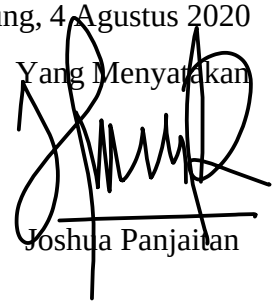
NIP: 00750052

LEMBAR PERNYATAAN

Dengan ini saya, Joshua Panjaitan, menyatakan sesungguhnya bahwa Tugas Akhir saya dengan judul **Klasifikasi Kepribadian Berbasis Sentiment di Sosial Media Twitter Menggunakan Metode PbSC** beserta dengan seluruh isinya adalah merupakan hasil karya sendiri, dan saya tidak melakukan penjiplakan yang tidak sesuai dengan etika keilmuan yang belaku dalam masyarakat keilmuan. Saya siap menanggung resiko/sanksi yang diberikan jika di kemudian hari ditemukan pelanggaran terhadap etika keilmuan dalam Laporan TA atau jika ada klaim dari pihak lain terhadap keaslian karya,

Bandung, 4 Agustus 2020

Yang Menyatakan



Joshua Panjaitan

Klasifikasi Kepribadian Berbasis Sentiment di Sosial Media Twitter Menggunakan Metode PbSC

Joshua Panjaitan[1], Warih Maharani[2]

[1,2]Fakultas Informatika, Universitas Telkom, Bandung
[1]joshuapanjaitan@students.telkomuniversity.ac.id,
[2]wmaharani@telkomuniversity.ac.id,

Abstrak

Pada era digital saat ini *sosial media* di indonesia menjadi sebuah kebutuhan sosial untuk orang-orang saling bertukar informasi. Hal tersebut membuat beberapa perusahaan mulai mencoba memanfaatkan informasi dari sosial media mereka seperti *twitter* untuk membantu mereka menggambil keputusan dalam melakukan perekrutan karyawan baru. Dengan menggunakan pendekatan yang sesuai, informasi seperti kepribadian pengguna bisa diperoleh dengan memanfaatkan data dari pengguna *sosial media twitter*. Informasi seperti ini dapat membantu divisi *HR (Human Resource) / Talent Management* dalam membantu pengambil keputusan dalam rekruitasi karyawan. Untuk memperoleh informasi seperti kepribadian calon karyawan maka penelitian ini menggunakan metode klasifikasi berbasis aturan yang diberi nama PbSC (*Personality Refirement for Sentiment Classification*) untuk melakukan klasifikasi kepribadian dengan menggunakan dataset yang diperoleh dari informasi yang terkandung dalam *sosial media twitter* seseorang. Penggunaan PbSC dipilih dikarenakan metode ini bisa diimplementasikan untuk semua jenis dataset tidak terkecuali *user twitter* indonesia. Dalam penggunaannya setiap *user* yang akan diambil 450 *tweet* terbarunya dan dilakukan proses *preprocessing*, setelah itu dilakukan proses klasifikasi dengan metode PbSC. Pengujian dilakukan dengan menggunakan 2 skenario dimana 1 menggunakan *dataset* yang dikumpulkan dan dilabeli oleh peneliti sebanyak 122 *data user*. Skenario ke 2 dengan menggunakan 295 dataset yang berbeda yang labelnya diperoleh dari hasil kuesioner *user* dengan BFI (*Big Five Inventory*). Hasil evaluasi berupa akurasi prediksi dari algoritma PbSC dan dimana rata-rata akurasi yang didapatkan dari data hasil dari responden sebesar 26.3% dan 11.67% untuk *user* dengan label BFI untuk semua dimensi kepribadian.

Kata kunci: *sosial media*, kepribadian, *talent management, twitter, PbSC.*

Abstract

In the digital era, social media in Indonesia has become a social necessity for people who want to exchange information. This caused some companies start to using information from their social media such as twitter to help them make a decisions about recruiting new employees. By using the appropriate method, information such as user personality can be obtained using data from user's social media. Information like this can help the HR (Human Resources) / Talent Management division in helping get decisions on employee recruitment. In order to get the result of personality from employees this research will try to apply the classification method that called PBSC (Personality Refirement for Sentiment Classification) to classify personality by using a dataset obtained from information contained in someone's twitter social media. The use of PBSC was chosen because this method can be applied to all types of datasets including Indonesian twitter users. For implementation, 450 tweets will be taken and the bring them into preprocessing process, after that the classification process will be carried out using the PbSC method. Testing is done by using 2 scenarios, the first one uses a dataset collected and labeled by researchers as many as 122 data. The second scenario uses 295 different datasets whose labels were obtained from the results of a user questionnaire with BFI (Big Five Inventory). The evaluation results consist of predictions from the PbSC algorithm and where the average accuracy obtained from the user dataset is 26.3% and 11.67% for users with the BFI label for all dimensions of personality.

Keywords: sosial media, personality, talent management, twitter, PbSC.

1. Pendahuluan

Latar Belakang

Pada era digital saat ini *microblog* berkembang sangat pesat dan menjadi sangat popular, dimana orang-orang dapat membagikan tentang keseharian mereka dan mengekspresikan emosi mereka. Salah satu *microblog* terbesar adalah twitter. Twitter menjadi salah satu ranah peneliti dalam mengumpulkan informasi yang berkaitan dengan social media dikarenakan akses dengan API (*Application Programming Interface*)

yang lengkap sehingga memudahkan peneliti untuk menyambungkan *twitter* kedalam aplikasi mereka. Hal ini didukung bahwa twitter masih memiliki tingkat popularitas yang tinggi dan pengguna aktif yang banyak [1][16]. Informasi yang melimpah di twitter jika dimanfaatkan dengan baik dapat berguna dalam bidang perusahaan seperti *Talent Management* untuk membantu mengambil keputusan dalam rekruitasi karyawan baru [2]. Salah satu cara pemanfaatannya dengan memprediksi kepribadian calon karyawan yang akan direkrut. Informasi seperti ini dapat menjadi pertimbangan bagi HR (*Human Resource*) dimana divisi ini yang bertanggung jawab dalam merekrut karyawan baru dalam sebuah perusahaan [2]. Informasi tadi bisa berguna untuk membantu mengambil keputusan terhadap calon karyawan sehingga lebih selektif untuk memilih karyawan yang dianggap pantas dalam perusahaan tersebut.

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

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

Topik dan Batasannya

Penelitian ini bertujuan untuk mengklasifikasikan kepribadian pengguna dengan data yang didapatkan dari *sosial media twitter* mereka. Data yang dimaksud adalah *tweet* yang ditulis dari masing-masing pengguna. Data *tweet* yang diperoleh merupakan tulisan yang dimana kebanyakan berisi bahasa indonesia. Data tersebut kemudian diklasifikasikan dengan menggunakan metode PbSC dan melakukan evaluasi dan analisis terhadap metode tersebut. Kelas kepribadian dibagi menjadi 3 dimensi kelas/faktor kepribadian yaitu *Extroversion, Agreeableness, Conscientiousness*. Metode PbSC akan memprediksi level kepribadian tersebut apakah *high, low* atau netral. Hasil dari tiap-tiap klasifikasi yaitu berupa vektor/data yang merepresentasikan level kepripadian dari tiap-tiap pengguna, sebagai contoh, @jooshpn = [1,Netral,0]. Hasil diatas menunjukkan bahwa pengguna dengan username @jooshpn, memiliki level kepribadian *extrovert* yang tinggi (index 0 dalam *array*), *agreeableness* yang netral (index 1) dan *consciouness* yang rendah (index 2). Sebelum masuk kedalam klasifikasi terdapat beberapa proses yang harus dilakukan seperti pengumpulan data pengguna lewat kuesioner, *crawling tweet, preprocessing*, yang akan dijelaskan di bab berikutnya. Adapapun batasan penelitian adalah mengabaikan faktor *imbalance data* yang didapatkan dan Penggunaan parameter dan korpus yang tidak dijelaskan pada penelitian [1] akan ditentukan oleh peneliti.

Tujuan

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

Organisasi Tulisan

Penelitian ini terdiri dari 4 bab, dimana setelah bab pendahuluan terdapat bab 2 yaitu studi terkait dimana pada bagian ini akan menjelaskan terkait teori-teori yang berhubungan dengan penelitian ini. Bab 3 berisi tentang penjelasan terkait detail dari perancangan sistem mulai dari awal sampai sistem mampu

menghasilkan sebuah output dari tiap-tiap dimensi kepribadian. Bab 4 berisi hasil evaluasi dan analisis dari sistem yang dibangun beserta kesimpulan dan saran untuk penelitian berikutnya.

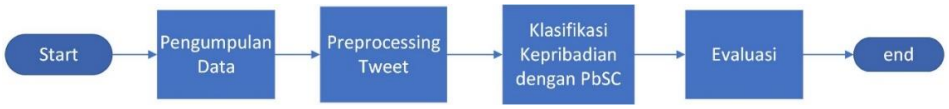
2. Studi Terkait

Big Five Model Personality merupakan sebuah teori psikologi yang menjelaskan tentang kecendrungan dari kepribadian manusia [11]. 5 Faktor/Model tersebut adalah *Extroversion, Agreeableness, conscientiousness, openness* dan *neuroticism*. Dalam kaitannya dengan studi komputasi, 5 faktor kepribadian ini bisa diprediksi dengan menganalisis sifat-sifat yang mereka lakukan lewat interaksinya dengan komputer salah satunya lewat sosial media [10]. Orang dengan kecenderungan *high extrovert* cenderung memiliki banyak teman dan sering mengucapkan kata-kata yang berhubungan dengan ”orang” [17][19]. Sementara orang yang memiliki kecenderungan *conscientiousness* yang tinggi memiliki kecenderungan untuk memposting sesuatu yang berkaitan dengan ”kerja keras” atau tujuan hidup mereka [1][11]. Dalam kaitannya dengan divisi HR (*Human Resource*), dikutip dari artikel [3], salah satu kecakapan yang dicari oleh perusahaan *bonafide* adalah antusiasme/*passion* dan memiliki ambisi yang kuat dalam bekerja. Kecakapan tersebut bisa dideteksi dari *user* dengan kepribadian *high conscientiousness*, dikarenakan orang yang memiliki dimensi *high conscientiousness* pada teorinya dijelaskan adalah orang-orang yang memiliki kesungguhan hati dalam meraih tujuannya [1]. Informasi seperti ini tentunya bisa dimanfaatkan sebagai alat bantu pengambilan keputusan dalam divisi HR tersebut.

Penelitian terdahulu yang mengaplikasikan *big five model* adalah *Personality based refirement for semtiment classification in microblogs* [1]. Penelitian ini bertujuan untuk mengklasifikasikan sentiment yang diperoleh dari *user* twitter kedalam positif dan negatif dengan terlebih dahulu mengelompokkan *user* kedalam 3 dari 5 dimensi *big five*. *User* dikelompokkan dengan membangun sistem berbasis aturan yang diberi nama PbSC untuk melakukan klasifikasi kepribadian [1]. Penelitian tersebut memilih membangun sebuah aturan daripada menggunakan menggunakan algoritma *machine learning* dikarenakan algoritma *machine learning* dinilai sangat kompleks dan tidak praktis untuk dipakai [1]. Sistem tersebut memanfaatkan penggunaan kata yang cenderung diucapkan oleh masing-masing kepribadian. Setelah mengelompokkan user kedalam kelompok-kelompok kepribadian peneliti kemudian memanfaatkan *machine learning* algoritma untuk mengklasifikasikan sentimen tersebut kedalam kelas positif dan negatif. Kelebihan dari penelitian ini dengan mengelompokkan *user* terlebih dahulu hasil akurasi yang diperoleh jauh lebih bagus dibandingkan tanpa melakukan pengelompokan. Kelemahan dalam penelitian ini dikarenakan sistem klasifikasi kepribadian yang digunakan dengan melakukan pendekatan berbasis aturan sehingga aturan sehingga akan ada *user* yang tidak terklasifikasi kedalam sebuah kepribadian karena tidak menyentuh angka *treshold* pada sebuah aturan tertentu. Peneliti menyiasati dengan membahkan kelas netral pada penelitiannya.

3. Sistem yang Dibangun

Penelitian ini terdiri dari 3 tahapan utama, yaitu pengumpulan data, *preprocessing* dan klasifikasi. Rincian kegiatan akan dijelaskan pada gambar 3.1



Gambar 3.1. Gambaran Umum Sistem

3.1 Pengumpulan Data

Pengumpulan data dilakukan untuk mendapatkan responden dengan menyebar kuesioner. Kuesioner disebar untuk mendapatkan informasi seperti *username twitter*, bahasa utama yang digunakan dan informasi lainnya, dimana fokus utama calon responden adalah pengguna aktif dengan bahasa utama indonesia. Total responden yang sudah dikumpulkan sebanyak 122 pengguna. Setelah *username* didapatkan maka kemudian akan dilakukan proses *crawling* untuk mengekstrak 450 *tweet* terbaru dari setiap *user*. *Crawling* merupakan proses pencarian & pengumpulan informasi dari sebuah halaman *website* [4], dimana halaman yang digunakan adalah halaman *profile twitter* dari masing-masing *user*. Berikut contoh *tweet* hasil crawling dari seorang *user*,

	tweet
0	b'Weekend produktif'
1	b'ayam,rubah,pelabuhan,tidur https://t.co/quH5pydVFb'
2	b'udah tidor2, kumpulkan tenaga dlu'
3	b'hari ini jg aku akan menang'
4	b'@CoachJustinL beneren masuk dong minimano \\U0001f602'
...	...
446	b'cuan dipagi hari.'
447	b'@CoachJustinL blok aja coc'
448	b'@CoachJustinL everton goalnya semua gol jatuh dari langit coc, yakin babak 2 menang'
449	b'@FirminoTjk @mrardein @CoachJustinL kalau vpn gratisan speed internetnya bakal nurun , kecuali lu punya vpn premium.'
450	b'@mrardein @CoachJustinL Kalau udh konek cpn nya matiin'

Gambar 3.2. Hasil *Crawling tweet* dari seorang *user twitter*

3.2 *Preprocessing*

Preprocessing merupakan tahap mempersiapkan data sebelum nantinya diolah oleh algoritma utama [4]. *Preprocessing* yang dilakukan seperti membersihkan *tweet*, *replace string/special char*/URL, *decode* dan *encode emoticon*. *Tweet* yang dikumpulkan terdiri dari 450 *tweet* terbaru dari setiap pengguna, dan keluaran dari tahapan ini adalah 450 *tweet* yang sudah bersih dan siap untuk digunakan pada metode *ruled-based*. Berikut berupakan contoh tahapan *preprocessing* yang dilakukan pada satu *tweet* dari seorang *user*.

Table 3.1 Contoh *Preprocessing* Data

Proses	Input	Output
Menghilangkan URL/Link dari setiap tweet	”b'Seru rame keren dan diberkati sekali. Thankyou Jesus thankyou eXcellent Generation! //U000256 #LIVELOVEDREAMS’ “https://t.co/quH5pydVFb'	”b'Seru rame keren dan diberkati sekali. Thankyou Jesus thankyou eXcellent Generation! //U000F256 #LIVELOVEDREAMS’ “
Menghilangkan *Special Char* dari setiap tweet	”b'Seru rame keren dan diberkati sekali. Thankyou Jesus thankyou eXcellent Generation! //U000F256 #LIVELOVEDREAMS’ “	”Seru rame keren dan diberkati sekali. Thankyou Jesus thankyou eXcellent Generation! //U000F256 #LIVELOVEDREAMS’ “
Mengubah tulisan menjadi huruf kecil	“Seru rame keren dan diberkati sekali. Thankyou Jesus thankyou eXcellent Generation! //U000F256 #LIVELOVEDREAMS’ “	”seru rame keren dan diberkati sekali. thankyou jesus thankyou excellent generation! //U000F256 #livelovedreams“
Menyederhaka-kan *escape code* menjadi unicode	”seru rame keren dan diberkati sekali. thankyou jesus thankyou excellent generation! //U000F256 #livelovedreams“	”seru rame keren dan diberkati sekali. thankyou jesus thankyou excellent generation! //u256 #livelovedreams“
Mengubah *unicode* menjadi menjadi bahasa yang bisa dipahami manusia	”seru rame keren dan diberkati sekali. thankyou jesus thankyou excellent generation! //u256 #livelovedreams“	”seru rame keren dan diberkati sekali. thankyou jesus thankyou excellent generation! #blackheart# #livelovedreams“

Pada tahapan pertama dalam *preprocessing* dilakukan proses menghilangkan *url/link* yang terkandung dalam setiap cuitan, *link* ini biasanya dihasilkan jika *tweet* tersebut mengarah kesebuah gambar atau video sehingga hal ini tidak dibutuhkan. Pada tahapan kedua dilakukan proses menghilangkan *special character*. *Special character* biasanya dihasilkan saat melakukan *crawling* dan dikonversi otomatis oleh *twitter*. Contoh *special character* adalah b’, rt’, bot’. *Character* b’ muncul secara *default* saat proses *crawling* yang menandakan bahwa *tweet* tersebut adalah *tweet* yang *original* ditulis oleh *user*, sementara rt’ menandakan *tweet* tersebut adalah hasil *reetweet* dari *user* lain, sedangkan bot’ menandakan bahwa *tweet* tersebut ditulis oleh bot. Tahapan ketiga adalah merubah semua *tweet* menjadi huruf kecil. Tahapan ke empat dan kelima berhubungan dengan *emoticon.* Tujuan dari 2 tahapan ini adalah mengubah *emoticon* yang ditulis oleh *user* menjadi sebuah kata-kata yang bisa dimengerti manusia. Prosesnya dengan mengubah setiap *emoticon* kedalam *universal code*, bisa dilihat dalam tahapan 4 dimana kode U000f256 dirubah menjadi kode yang lebih ringkas yaitu //u256. U000f256 diperoleh saat melakukan *crawling*, dimana hal ini dilakukan otomatis oleh *twitter* API untuk mengganti *emoticon* yang ditulis lewat *virtual keyboard* menjadi sebuah *escape code* untuk menghindari *error*. Pada dasarnya *escape code* yang sudah dikonversi otomatis tersebut memiliki referensi yang sudah terdaftar secara internasional yang disebut sebagai *unicode* [8]. Setelah itu kode ini diterjemahkan kedalam kalimat dengan mencocoknya dengan kamus *unicode* yang tersedia di *website unicode* [8] sehingga dihasilkan #blackheart#. *Emoticon* yang sudah dikonversi menjadi kata ditandai dengan awalan # (*hashtag*) dan diakhiri dengan #. Hasil dari tahapan ini adalah *tweet* yang sudah bersih dari special character,*link*, *emoticon* dan siap untuk digunakan dalam proses klasifikasi dengan PbSC. Dibawah merupakan contoh hasil *preprocessing* dari *tweet* yang sudah *dicrawl* dalam gambar 3.2.

	tweet
0	weekend produktif
1	ayam,rubah,pelabuhan,tidur
2	udah tidor2, kumpulkan tenaga dlu
3	hari ini jg aku akan menang
4	@coachjustinl beneren masuk dong minimano #face with tears of joy#
...	...
446	cuan dipagi hari.
447	@coachjustinl blok aja coc
448	@coachjustinl everton goalnya semua gol jatuh dari langit coc, yakin babak 2 menang
449	@firminotjk @mrardein @coachjustinl kalau vpn gratisan speed internetnya bakal nurun , kecuali lu punya vpn premium.
450	@mrardein @coachjustinl kalau udh konek cpn nya matiin

Gambar 3.3. Hasil *Preprocessing dari tweet* 3.2

3.3 Klasifikasi

Algoritma Metode *PbSC* yang digunakan merujuk pada penelitian [1]. Aturan yang digunakan berbasis frekuensi dengan ide utama menghitung kata-kata yang relevan dengan sebuah kepribadian dan jika memenuhi *threshold* yang sudah ditentukan maka pengguna akan teridentifikasi menjadi sebuah kepribadian [1]. Tiap kepribadian memiliki basis aturan yang berbeda namun dengan ide yang sama, hanya berbeda dari segi *treshold* dan variabel yang dibandingkan. Berikut adalah beberapa contoh aturan yang digunakan untuk memprediksi sebuah kelas kepribadian.

Rule HE [1]:
IF *#HE_tweet* ≥ p1 ^ *#LE_tweet* ≤ p2 ^ *Ratio(HE_tweet)* ≥q1 ^ *Ratio(LE_tweet)* ≤ q2 THEN = ***high***.

Rule LE [1] :
IF *#LE_tweet* ≥ p3 ^ *#HE_tweet* ≤ p4 ^ *Ratio(LE_tweet)* ≥q3 ^ *Ratio(HE_tweet)* ≤ q4 THEN = ***low***

Diatas merupakan contoh klasifikasi untuk menentukan user masuk kedalam kelas *high extrovert* dan *low extrovert* dengan kata lain untuk dimensi *extrovert* yang nilai labelnya adalah 1 dan 0, 1 untuk *high* dan 0 untuk *low*. Cara kerjanya adalah dengan menyediakan korpus acuan yang berisi kata-kata yang cenderung sering digunakan orang-orang dengan tipe kepribadian *high extrovert* atau *low extrovert* dan menghitung jumlah kata-kata dalam setiap tweet untuk mengisi variabel *#HE_Tweet* atau *#LE_tweet*. Sebagai contoh korpus untuk HE (*High Extrovert*) adalah kumpulan kata-kata yang diyakini sering diucapkan oleh orang *extrovert* seperti ”*guys”, ”ayo”,”yuk” ,”main*” [1][18]. Sedangkan LE_*Tweet* (*Low Extrovert Tweet*) adalah *tweet* yang mengandung unsur *low extrovert* seperti “*alone”, “sendiri”, “bosan*” [1]. Nilai dari variabel tersebut akan dibandingkan dengan *treshhold* yaitu p1,p2,q1 dan q2. Nilai *hyperparameter (threshold)* mengacu pada penelitian [18]. Nilai *hyperparameter* yang dipilih merupakan ambang batas minumin sebuah dimensi kepribadian berkorelasi/berhubungan dengan sebuah kata-kata yang mengandung unsur-unsur *big-five*, seperti *extroversion* terhubung dengan kata-kata yang berkaitan dengan ”orang”,”keluarga”,”sosial” [18]. Jika melakukan modifikasi tanpa alasan dan referensi yang jelas seperti menurunkan nilai *hyperparameter*, ditakutkan hasil klasifikasi yang sudah dilakukan tidak mencerminkan sebuah dimensi kepribadian dikarenakan ambang batas korelasinya diturunkan. Hasil dari tahapan ini berupa vektor dengan 3 nilai yaitu *high, low* atau normal seperti pada bagian Tujuan (Bab 1). Nilai normal didapatkan jika tidak ada aturan yang terpenuhi dari hasil extraksi informasi yang didapatkan dari user.

Table 3.2 *Threshold* Setiap Dimensi Kepribadian

High Extrovert	Low Extrovert	High Agreeableness	Low Agreeableness	High Conscientiousness	Low Conscientiousness
0.23	0.21	0.22	0.21	0.21	0.21

Berikut adalah contoh aturan dengan studi kasus seorang user dengan 450 *tweet* yang sudah di *crawl* dan dilakukan *preprocessing*. Untuk memprediksi dimensi *extroversion* dari *user* tersebut maka dibutuhkan 2 aturan PbSC yaitu Rule HE (*High Extrovert*) dan Rule LE (*Low Extrovert*), yang dimana fungsinya untuk menentukan apakah user tersebut memiliki kepribadian *low extrovert* atau *high extrovert*. Dalam rule HE seperti yang bisa dilihat diatas terdapat 4 *hyperparameter* yaitu p1,p2,q1 dan q2, dimana p1 dan q1 merupakan *hyperparameter* pembanding untuk HE_tweet sedangkan p2 dan q2 adalah *hyper parameter* pembanding untuk LE Tweet. Nilai tersebut bisa diperoleh dari tabel 3.2 dimana untuk nilai p1 = 0.23 * 450, sedangkan nilai q1 = 0.23. Nilai p2 = 0.21 * 450 dan nilai q2 = 0.21. Nilai yang tertera dalam tabel 3.2 merupakan rata-rata nilai korelasi antara penggunaan kata-kata yang berkolerasi dengan sebuah dimensi kepribadian. Sebagai contoh untuk nilai *High Extrovert* 0.23 berarti korelasi antara dimensi *high extrovert*

dengan korpus yang berisi kata-kata yang berhubungan dengan ”orang”, ”keluarga” itu sebesar 23% [18]. Berdasarkan nilai tersebut, nilai *hyperparameter* p yang berisi nilai pembanding untuk jumlah tweet dihitung dengan 0.23 * total *tweet* untuk mendapatkan nilai *tweet* sebanyak 23% dari total keseluruhan *tweet* tiap-tiap *user*. Berdasarkan hasil perhitungan diatas maka diperoleh *rule* yang sesuai untuk menghitung kepribadian *extroversion* dari *user* tersebut adalah :

***Rule Extroversion* [1]**:

IF *#HE_tweet* ≥ 103 ^ *#LE_tweet* ≤ 95 ^ *Ratio(HE_tweet)* ≥0.23 ^ *Ratio(LE_tweet)* ≤ 0.21 **THEN** = ***high***.

ELSE IF *#LE_tweet* ≥ 95 ^ *#HE_tweet* ≤ 103 ^ *Ratio(LE_tweet)* ≥0.21 ^ *Ratio(HE_tweet)* ≤ 0.23 **THEN** = ***low***

Setelah nilai *hyperparameter* sudah ditentukan selanjutnya adalah menghitung nilai #HE_tweet dan #LE_tweet. Nilai tersebut akan dihitung dari pencocokan kata-kata dari setiap *tweet* dengan korpus yang sudah dikumpulkan. Korpus untuk dimensi *High Extrovert* berisi kata-kata yang cenderung diucapkan oleh orang dengan kecenderungan *high extrovert* seperti kata-kata yang berhubungan dengan ”*orang*”, ”*gais*”, ”*kumpul yuk*” dan sebagainya. Setiap kelas dimensi memiliki korpus masing-masing sehingga total terdapat 6 Korpus. Sehingga jika *user* tersebut memiliki salah satu *tweet* yang berbunyi : “*yuk gais kita kumpul*”, maka *tweet* tersebut akan membuat variabel HE_Tweet bertambah nilainya 1 dikarenakan kata ”gais” terdapat dalam korpus. Hal tersebut dilakukan sampai semua *tweet* dari setiap *user* selesai dihitung. Hasil dari perhitungannya disimpan dan dimasukkan kedalam algoritma PbSC untuk mengklasifikasikan kelas kepribadian dari *user* tersebut. Untuk kelas *Agreeableness*, dan *Conscientiousness* kurang lebih caranya sama, hanya berbeda dari nilai *hyperparameter* dan isi korpusnya saja.

Hasil klasifikasi yang didapatkan adalah 3 nilai yang dirangkai dalam satu *array* seperti :

@jooshpn = [1,Netral,0]

Jika seorang *user* tidak mendapatkan kelas 1 atau 0 dalam sebuah dimensi kepribadian maka sistem akan mengkasifikasikannya sebagai netral. Dimana pada contoh diatas *user* @jooshpn memiliki kelas netral di index ke-1 dimana index tersebeut adalah tempat untuk dimensi *agreeableness*. Label netral dipengaruhi oleh nilai *tweet* dari *user* tersebut yang mengandung unsur *high agreeableness* atau *low agreeableness* tidak menyentuh angka *threshold*. Salah satu faktor yang mempengaruhinya adalah jumlah korpus yang bertanggung jawab dalam proses perhitungan variabel #HA_Tweet atau #LA_Tweet tersebut sebagai pembanding untuk *hyperparameter* (*threshold*). Korpus dikumpulkan oleh peneliti berdasarkan referensi pada penelitian [1][18]. Namun tidak menutup kemungkinan korpus yang sudah dikumpulkan belum lengkap untuk meng-*cover* semua kata-kata yang diucapkan oleh sebuah dimensi kepribadian. Hal ini yang menyebabkan banyaknya label netral yang didapatkan dalam proses klasifikasi dikarenakan ruang lingkup dari korpus yang dipakai tidak mampu men*cover* semua kata-kata yang diucapkan oleh setiap *user* dan tidak semua *user* juga menggunakan kata-kata yang dinilai merepresentasikan sebuah dimensi kepribadian sebagai kata-kata sehari-hari mereka.

4. Evaluasi

Proses evaluasi dibagi menjadi 2 tahap, yang pertama adalah mengevaluasi hasil klasifikasi algoritma PbSC dengan menggunakan 122 *user* (Data Responden) yang sudah dikumpulkan lewat kuesioner yang sudah disebarkan pada bab 3.1. Hasil evaluasi dinilai berdasarkan akurasi dari data yang sudah diklasifikasikan. Tahapan kedua dilakukan dengan membandingkan hasil akurasi data responden dengan 295 dataset lain yang dikumpulkan dengan label BFI *questionnaire* (*Big Five Inventory*). BFI *questionnaire* merupakan sebuah *questionnaire* pertanyaan sebanyak 44 pertanyaan yang diajukan kepada responden, dimana 44 pertanyaan tersebut didesign untuk menentukan tingkat kepribadian bagi orang yang menjawabnya kedalam *big* 5 *model personality* [9]. Tujuan menggunakan 295 dataset BFI ini sebagai validasi akurasi terhadap data responden untuk melihat pengaruh metode PbSC dengan label yang lebih bisa dipertanggungjawabkan [9].

4.1 Hasil Pengujian

Tahapan pertama dalam evaluasi ini adalah mengevaluasi hasil klasifikasi algoritma PbSC dengan menggunakan 122 *userdata* (data respondedn) yang sudah dikumpulkan pada bab 3.1. Hasil prediksi akan disandingkan dengan label yang sudah ditentukan oleh peneliti. Peneliti melakukan pelabelan dengan seluruh dataset dengan mempertimbangkan faktor-faktor kepribadian yang bisa dianalisis lewat interaksi setiap *user* dengan *sosial media* mereka. Faktor-faktor pertimbangan yang dilakukan peneliti dalam menentukan label dapat dilihat dalam Tabel 4.1.

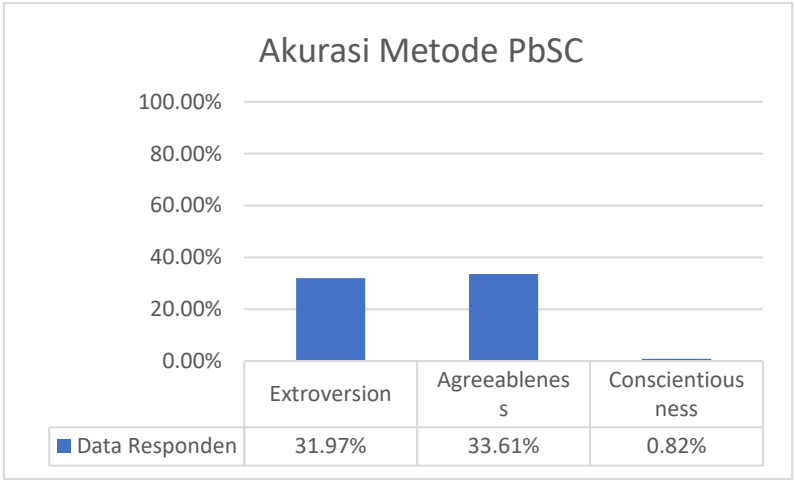
Dalam Tabel 4.1 dapat dilihat bahwa setiap dimensi kepribadian memiliki faktor yang dipertimbangkan dalam mentukan label kepribadiannya. Sebagai contoh dimensi *extroversion* memiliki 5 faktor pelabelan, dimana jika seorang *user* memenuhi kelimanya maka *user* tersebut akan dilabeli *high extrovert*. Semua faktor pelabelan bernilai ganjil sehingga setiap user pasti memiliki label antara *high* atau *low* tergantung seberapa banyak faktor yang terpenuhi dalam setiap dimensi kepribadiannya. Label netral tidak dihitung dalam perhitungan akurasi dikarenakan user yang dilabeli netral tidak memiliki kecenderungan apa-apa, dan pada penelitian [1] metode PbSC hanya menggunakan label *high* dan *low* dalam melakukan perhitungan akurasi, yang dimana hal ini juga diadopsi dalam penelitian ini. Jika faktor kepribadian lebih kecil dari jumlah faktor maka *user* alan dilabeli *low*. Sebagai contoh seorang *user* memenuhi hanya 1 faktor dalam

dimensi *agreeableness* sehingga label *agreeableness*nya adalah *low*, dan sebaliknya untuk menentukan label *high*. Setiap data akan dilabeli berdasarkan faktor yang tertera dalam tabel 4.1 untuk skenario uji pertama ini.

Table 4.3 Faktor Pertimbangan Pelabelan 122 Dataset

No	*Extroversion*	*Agreeableness*	*Conscientiousness*
1	Jumlah teman yang banyak/ Following, *Mean* = 150. [19]	Berhubungan dengan kata-kata *relationship*/hubungan [18]	Jumlah teman yang banyak, Mean = 147 [19]
2	Berpartisipasi dalam obloran/Group, *Mean* 20%+(mention/reply/reetweet) = 150 [19]	*Selective* memilih teman namun masih tetap banyak *followers* > 20% *following* [19][10]	*Upload Picture / link* tinggi /[19]
3	Jumlah *Retweet/Quote* tinggi [19] , mulai dari diatas 100	Tendensi berinteraksi dengan media/ upload picture [19]	Berhubungan dengan kata-kata kerja keras dan kesuksesan [1]
4	Memiliki hubungan dengan *emoticon* positif / kata-kata sosial r = 0.21[1]		
5	Memiliki jumlah avg *tweet* yang tinggi Mean = 10 [17]		

Setelah dataset sudah dilabeli, selanjutnya dilakukan perhitungan akurasi. Hasil akurasi yang diperoleh dapat dilihat dalam gambar 4.1. Dari gambar dibawah dapat dilihat bahwa akurasi yang didapatkan dari data responden tersebut dengan label yang ditentukan peneliti memiliki akurasi yang rendah, dimana untuk dimensi *extroversion* memiliki akurasi sebesar 31.97%, *agreeableness* memiliki akurasi sebesar 33.61% sementara dimensi *conscientiousness* yang terendah dari ketiganya sebesar 0.82%.

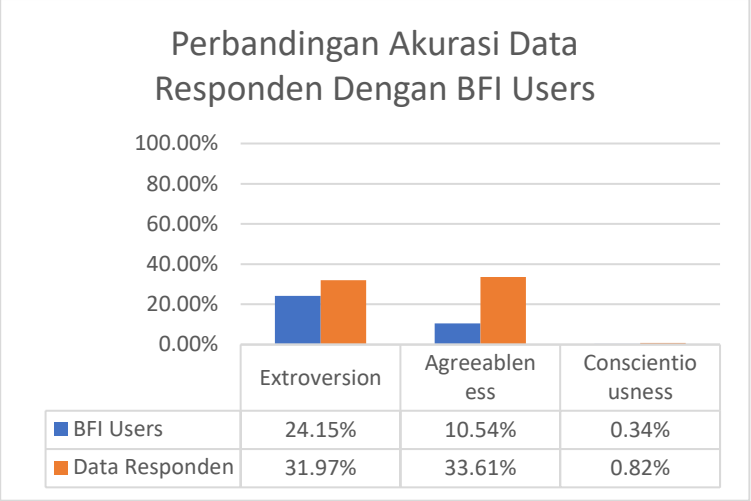


Gambar 4.1 Hasil Akurasi Metode PbSC

Salah satu faktor yang menyebabkan rendahnya akurasi yang didapatkan ini adalah metode ini terlalu banyak melakukan prediksi netral kepada *user*/responden yang didapat. Hal yang menyebabkan prediksi netral ini adalah terbatasnya isi korpus yang dikumpulkan dan tidak bisa mencakup kata-kata yang dinilai cenderung akan digunakan bagi masing-masing *user* yang merepresentasikan dimensi kepribadian mereka. Topik atau kata-kata yang berhubungan dengan ”cuaca”, ”politik”, ”olahraga” dan *Trending* Topik sering kali dicuit oleh responden namun kata/topik tersebut tidak terdaftar dalam referensi yang terkait dengan sebuah kepribadian, dengan kata lain menjadi kata-kata netral [18]. Sementara itu korpus yang dikumpulkan berdasarkan referensi penelitian sebelumnya [18][1] yang dimana tidak menjamin *user* akan menggunakannya terus-menerus didalam keseharian mereka dalam interaksinya dengan *sosial media* mereka. Sehingga keterbatasan isi/kata-kata didalam korpus menjadi salah satu kelemahan dalam metode ini dalam menentukan label prediksi yang sesuai. Untuk memvalidasi hal tersebut dilakukan evaluasi tahap kedua dimana menggunakan *dataset* terpisah sebanyak 295 dataset yang sudah dilabeli dengan BFI *questionnaire* yang sudah dikumpulkan oleh peneliti lain, hasilnya dapat dilihat dalam gambar 4.2.

Hasil yang diperoleh dalam evaluasi tahap 2 dengan membandingkan hasil akurasi dengan *user* yang sudah dilabeli dengan BFI *questionnaire*. Dapat dilihat terdapat perbedaan akurasi yang dimana akurasi yang didapatkan oleh BFI *users* lebih rendah dibandingkan akurasi yang didapatkan data responden. Hal ini

bisa disebabkan oleh beberapa hal salah satunya adalah jumlah data yang digunakan hampir 2 kali lipat dari data responden sehingga faktor pembangi untuk rata-rata akurasinya semakin besar. Namun dari hasil tersebut dapat dilihat akurasi yang rendah juga didapatkan dari dataset dengan label BFI, hasil ini didapat dikarenakan metode ini terlalu banyak memprediksi *user* dengan kelas netral, terutama pada dimensi *conscientiousness*. Label netral tidak terdapat dalam label pembanding akurasi, hal ini yang menyebabkan rendahnya tingkat akurasi karena tidak ada label yang sesuai dengan label pembanding akurasi.



Gambar 4.2 Hasil Akurasi Data Responden dengan BFI User

4.2 Analisis Hasil Pengujian

Berdasarkan hasil pengujian yang telah dilakukan, akurasi yang diperoleh baik dari data responden maupun dataset yang sudah dilabeli BFI *questionnaire* semuanya dibawah 40%. Salah satu faktor penyebab rendahnya akurasi metode ini adalah hasil prediksi yang dilakukan metode PbSC ini kepada setiap *user* banyak mendapatkan hasil netral. Penyebab hal ini adalah terbatasnya jumlah isi korpus yang dikumpulkan yang dimana tidak bisa mencakup kata-kata yang dinilai cenderung akan digunakan bagi masing-masing *user* yang merepresentasikan dimensi kepribadian mereka. Korpus dikumpulkan berdasarkan referensi penelitian sebelumnya [18][1], yang dimana tidak menjamin *user* akan menggunakannya terus-menerus didalam keseharian mereka dalam interaksinya dengan sosial media mereka. Seperti pada dimensi *ekstroversion* referensi kata/topiknya adalah kata yang berhubungan dengan ”orang”,”keluarga”,”keramaian”, ”kesenangan” [1]. Sehingga topik/kata lain diluar itu sulit untuk dimasukkan kedalam korpus karena tidak ada bukti/referensi yang jelas yang membuktikan bahwasanya kata/topik tersebut memiliki keterkaitan dengan dimensi *extroversion*. Keterbatasan inilah yang menyebabkan sedikitnya jumlah korpus yang terkumpul.

Faktor lain yang menyebabkan rendahnya akurasi prediksi ini adalah penggunaan *hyperparameter* yang ditentukan pada penelitian ini. *Hypyerparameter* yang digunakan berdasarkan pada penelitian [18]. Dikarenakan langkanya informasi terkait dengan *hyperparameter* acuan maka penelitian ini hanya menggunakan 1 referensi kepada *hyperparameter*. Label netral diakibatkan nilai variabel *tweet* tidak menyentuh angka pada *hyperparamter* yang menjadi *threshold* sehingga untuk kedepannya nilai ini bisa dimodifikasi tentunya dengan mempertimbangkan referensi yang sesuai untuk *hyperparameter*. Jika hasil akurasi dibandingkan dengan penelitian [1] tempat metode ini diciptakan akurasi rata-rata yang didapatkan dari masing-masing dimensi mencapai 80%. Hal ini bisa diraih dikarenakan mereka hanya menggunakan *dataset* yang memiliki tingkat *evidence* yang tinggi terkait sebuah dimensi kepribadian. Hal ini bisa mereka lihat dari *tweet* yang mereka kumpulkan, dimana hanya *user* dengan *tweet* yang punya evidensi yang tinggi terhadap sebuah dimensi kepribadian tertentu yang dijadikan sebagai *dataset*. Berbeda dengan penelitian ini yang menggunakan *user* yang *random* tanpa memperhatikan tingkat evidensi *tweet* dari setiap user sehingga hasil yang didapat tidak terlalu bagus.

4.2 Kesimpulan dan Saran

Berdasarkan hasil pengujian dan analisis diatas, disimpulkan bahwa sistem klasifikasi kepribadian dengan menggunakan metode PbSC memiliki akurasi sebesar 31.97% untuk dimensi *extroversion*, dimensi *agreeableness* memiliki akurasi sebesar 33.61% sementara dimensi *conscientiousness* yang terendah dari ketiganya sebesar 0.82%. Sementara validasi akurasi yang didapatkan dengan *user* label BFI sebesar 24.15% untuk dimensi *extroversion*, 10.54% untuk dimensi *agreeableness* dan 0.34% untuk dimensi *conscientiousness*. Dimana hasil ini dipengaruhi oleh terbatasnya isi korpus yang dikumpulkan yang dinilai belum bisa mencakup mayoritas penggunaan kata dari setiap *user*. Saran untuk penelitian berikutnya adalah agar melakukan beberapa modifikasi/perbaikan terhadap *treshhold* dan aturan yang dipakai, seperti menambahkan faktor-faktor lain seperti jumlah teman, rata-rata *tweet* dan faktor-faktor lainnya sebagai aturan baru pada proses klasifikasinya sehingga aturan tidak hanya tergantung dari faktor *tweet* saja serta meningkatkan isi dalam korpus.

Daftar Pustaka

[1] Lin,Mao,D.Zeng. (2017). Personality-based refinement for sentiment classification in microblog. *Knowledge-Based System*. 132, 204-214.

[2] Jeske, Shultz. (2015). Using social media content for screening in recruitment and selection: pros and cons.

[3] Agus Dewi, *10 Hal yang Dicari Perusahaan Bonefid saat Merekrut Karyawan,* Diakses pada 26 Juli 2020 dari : https://www.duniakaryawan.com/hal-yang-dicari-perusahaan-bonafid-saat-merekrut-karyawan

[4] B. Pang , L. Lee , S. Vaithyanathan , Thumbs up? Sentiment classification using machine learning techniques, in: Proceedings of the 2002 Conference on Em- pirical Methods in Natural Language Processing, 2002, pp. 79–86

[5] S.M. Mohammad , S. Kiritchenko , X. Zhu , NRC-Canada: building the state-of-the-art in sentiment analysis of tweets, Comput. Sci. (2013)

[6] L. Qiu , H. Lin , J. Ramsay , F. Yang , You are what you tweet: personality expres- sion and perception on Twitter, J. Res. Personal. 46 (6) (2012) 710–718 .

[7] J. Golbeck , C. Robles , M. Edmondson , K. Turner , Predicting personality from Twitter, in: Proceedings of the Third IEEE International Conference on Social Computing, 2011, pp. 149–156 .

[8] Unicode , *Full Emoji List*, Diakses 25 Juli 2020 dari : https://unicode.org/emoji/charts/full-emoji-list.html.

[9] John, O. P., & Srivastava, S. (1999). The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (Vol. 2, pp. 102–138). New York: Guilford Press.

[10] J. Golbeck , C. Robles , M. Edmondson , K. Turner , Predicting personality from Twitter, in: Proceedings of the Third IEEE International Conference on Social Computing, 2011, pp. 149–156 .

[11] V.Ong, A.D.S. Rahmanto , Williem , D.Suhartono, A.E. Nugroho, E.W Andangsari, M.N Suprayogi, Personality Precidtion Based on Twitter Information in Bahasa Indonesia, in: Proceedings of the Federated Conference on Computer Science and Information Systems, 2017, pp. 367-372.

[12] Hanani Nabilah (6/10/2019), *Pengertian Twitter Beserta Sejarah dan Manfaat Twitter yang Dibahas Secara Lengkap*, Diakses 22 Oktober 2019 dari : https://www.nesabamedia.com/pengertian-twitter/

[13] Emilove, Jalaludin (18 Juni 2015), *Membuat web crawling*, Diakses pada 22 Oktober 2019 dari: https://www.kompasiana.com/jalaludin.ax/54f67c7fa333112b758b4ebf/membuat-web-crawling?page=all

[14] G. Angiani, L.Ferrari, L.Fontanini, P. Fornacciari, E. Iotti, F. Magliani, S. Manicardi, A Comparisson Between Preprocessing Techniques for Sentiment Analysis in Twitter, 2016, Dipartimento di Ingegneria dell'Informazione Universit_a degli Studi di Parma

[15] Suyanto, Machine Learning Tingkat Dasar dan Lanjut, INFORMATIKA: Bandung, 2018

[16] Twitter Investor Relations, "Q414 Selected Company Metrics and Financials," 2014.

[17]J. Golbeck, Sibel Adali, "Predicting Personality with Social Behaviour",International Conference on Advances in Social Networks Analysis and Mining, 2012.

[18] J.B. Hirsh , J.B. Peterson ,Personality and language use in self-narratives, J. Res. Personal. 43 (3) (2009) 524–527

[19] Y. Amichai-Hamburger and G. Vinitzky, "Social network use and personality," *Comput. Human Behav.*, vol. 26, no. 6, pp. 1289– 1295, 2010.

[20] Zulick,Joseph (9 Aug 2019), *How Machine Learning is Transforming Industrial Production*, Diakses pada 16 Juni 2020 dari :https://www.machinedesign.com/automation-iiot/article/21838038/how-machine-learning-is-transforming-industrial-production

Lampiran

1. Aturan PbSC

a. Untuk Dimensi *Extroversion*

```
IF:
   #HE_tweet ≥ p1 ^ #LE_tweet ≤ p2 ^ Ratio(HE_tweet) ≥q1 ^ Ratio(LE_tweet)
   ≤ q2  THEN = high
Else IF:
   #LE_tweet ≥ p3 ^ #HE_tweet ≤ p4 ^ Ratio(LE_tweet) ≥q3 ^ Ratio(HE_tweet)
   ≤ q4  THEN = low
```

b. Untuk Dimensi *Agreeableness*

```
IF :
   #HA_tweet ≥ p5 ^ #LA_tweet ≤ p6 ^ Ratio(HA_tweet) ≥q5 ^ Ratio(LA_tweet)
   ≤ q6  THEN = high
Else IF :
    #LA_tweet ≥ p7 ^ #HA_tweet ≤ p8 ^ Ratio(LA_tweet) ≥q7 ^ Ratio(HA_tweet)
    ≤ q8  THEN = low
```

c. Untuk Dimensi *Consciouness*

```
IF :
     #HC_tweet ≥ p9 ^ #LC_tweet ≤ p10 ^ Ratio(HC_tweet) ≥q9 ^ Ratio(LC_tweet)
     ≤ q10  THEN = high
Else IF :
     #LC_tweet  ≥  p11  ^  #HC_tweet  ≤  p12  ^  Ratio(LC_tweet)  ≥q11  ^
     Ratio(HC_tweet) ≤ q12  THEN = low
```

2. Korpus

```
# kamus
highExtrovert = ['wkawka', 'haha', 'yuk', 'kuy', 'kami', 'yeah', 'orang', 'handshake', 'lot of laught', 'hore',
                 'guys', 'gais', 'kumpul', 'joy', 'wkwk', 'main', 'kumpul', 'tears of joy', 'ayo', 'hehe', 'siapa', 'kita']

lowExtrovert = ['hmmm', 'unhappy', 'ahh', 'disapointed', 'huft', 'face without mouth', 'maaf',
                'maap', 'sorry', 'gua', 'gw', 'sendiri', 'alone', 'rumah', 'home', 'aku', 'weary face']

highAgreeableness = ['keren', 'wow', 'wah', 'bagus', 'puji tuhan', 'alhamdulillah', 'bless', 'selamat',
                     'semangat', 'love', 'terima kasih', 'thanks', 'makasi', 'hai', 'sayang', 'mantap', 'heart', 'kasian', 'good',
                     'thank', 'thank you', 'rasa', 'cinta', 'aman']

lowAgreeableness = ['jangan', 'keluar', 'diam', 'benci', 'mampus', 'mampos', 'rasain', 'anjir', 'cok', 'jelek',
                    'lo', 'elo', 'ngantuk', 'mager', 'smirking face', 'kesal', 'tidur', 'males', 'malas', 'marah', 'rasa']

highConscientiousnes = ['belajar', 'kampus', 'work hard', 'sibuk', 'stres', 'tugas', 'sekolah', 'ujian', 'kuat', 'raih', 'enjoy',
                        'ulangan', 'sukses', 'done', 'berhasil', 'sks', 'setres', 'bersih', 'rapi', 'baca', 'sempro', 'skripsi']

lowConscientiousnes = ['mabok', 'fight', 'brantem', 'mati',
                       'fuck', 'males', 'malas', 'shit', 'bangke', 'goblok', 'cok', 'telek']
```