

TESIS

**ANALISIS KLASIFIKASI KEPERIBADIAN BERDASARKAN MEDIA
SOSIAL DENGAN ALGORITMA BACKPROPAGATION NEURAL
NETWORK DAN NAIVE BAYES**



Disusun oleh:

Nama : Enle Yuliani
NIM : 18.52.1062
Konsentrasi : Informatics Technopreneurship

**PROGRAM STUDI S2 TEKNIK INFORMATIKA
PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA**

2020

TESIS

**ANALISIS KLASIFIKASI KEPERIBADIAN BERDASARKAN MEDIA
SOSIAL DENGAN ALGORITMA BACKPROPAGATION NEURAL
NETWORK DAN NAIVE BAYES**

**ANALYSIS OF PERSONALITY CLASSIFICATION BASED ON SOCIAL
MEDIA WITH THE BACKPROPAGATION NEURAL NETWORK AND
NAIVE BAYES ALGORITHMS**

Diajukan untuk memenuhi salah satu syarat memperoleh derajat Magister



Disusun oleh:

Nama : Enle Yuliani
NIM : 18.52.1062
Konsentrasi : Informatika Technopreneurship

**PROGRAM STUDI S2 TEKNIK INFORMATIKA
PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA**

2020

HALAMAN PENGESAHAN

**ANALISIS KLASIFIKASI KEPERIBADIAN BERDASARKAN MEDIA SOSIAL
DENGAN ALGORITMA BACKPROPAGATION NEURAL NETWORK DAN
NAIVE BAYES**

**ANALYSIS OF PERSONALITY CLASSIFICATION BASED ON SOCIAL MEDIA
WITH THE BACKPROPAGATION NEURAL NETWORK AND NAIVE BAYES
ALGORITHMS**

Dipersiapkan dan Disusun oleh

Enie Yullani

18.52.1062

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 Teknik Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari Senin, 04 Mei 2020

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta, 04 Mei 2020

Rektor

Prof. Dr. M. Suyanto, M.M.
NIK. 190302001

HALAMAN PERSETUJUAN

**ANALISIS KLASIFIKASI KEPERIBADIAN BERDASARKAN MEDIA SOSIAL
DENGAN ALGORITMA BACKPROPAGATION NEURAL NETWORK DAN
NAIVE BAYES**

**ANALYSIS OF PERSONALITY CLASSIFICATION BASED ON SOCIAL MEDIA
WITH THE BACKPROPAGATION NEURAL NETWORK AND NAIVE BAYES
ALGORITHMS**

Dipersiapkan dan Disusun oleh

Enie Yuliani

18.52.1062

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 Teknik Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari Senin, 04 Mei 2020

Pembimbing Utama

Anggota Tim Penguji

Prof. Dr. Ema Utami, S.Si., M.Kom.
NIK. 190302037

Dr. Kusriani, M.Kom.
NIK. 190302106

Pembimbing Pendamping

Dr. Andi Sunyoto, M.Kom.
NIK. 190302052

Dr. Suwanto Raharjo, S.Si., M.kom.
NIK. 999106

Prof. Dr. Ema Utami, S.Si., M.Kom.
NIK. 190302037

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta, 04 Mei 2020

Direktur Program Pascasarjana

Dr. Kusriani, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN TESIS

Yang bertandatangan di bawah ini,

Nama mahasiswa : Enie Yuliani
NIM : 18.52.1062
Konsentrasi : Informatics Technopreneurship

Menyatakan bahwa Tesis dengan judul berikut:
**Analisis Klasifikasi Kepribadian Berdasarkan Media Sosial Dengan
Algoritma Backpropagation Neural Network Dan Naive Bayes**

Dosen Pembimbing Utama : Prof. Dr. Ema Utami, S.Si., M.Kom
Dosen Pembimbing Pendamping : Dr. Suwanto Raharjo, S.Si., M.Kom

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Tim Dosen Pembimbing
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi

Yogyakarta, 04 Mei 2020
Yang Menyatakan,



Enie Yuliani

HALAMAN PERSEMBAHAN

Pada halaman ini penulis mempersembahkan hasil tesis ini dan menyampaikan ucapan terima kasih kepada orang-orang tersayang dan pihak-pihak yang telah membantu dalam penyusunan tesis ini.

1. Pertama dan yang paling utama adalah ucapan syukur kehadiran Allah SWT, yang telah melimpahkan segala nikmat dan karunia-Nya, sehingga tesis ini dapat terselesaikan.
2. Kepada Ibu dan Bapak, berkat doa, dukungan dan motivasi beliau saya dapat menyelesaikan studi.
3. Kepada adik saya Fauzi dan Puput yang turut memberikan dukungan untuk studi saya.
4. Kepada Ibu Prof. Dr. Ema Utami, S.Si., M.Kom dan Bapak Dr. Suwanto Raharjo, S.Si., M.Kom selaku dosen pembimbing yang telah sangat membimbing dan memotivasi dalam pembuatan tesis sehingga tesis ini dapat terselesaikan.
5. Kepada teman-teman kelas MTI 20 B yang telah menjadi teman belajar, berdiskusi dan saling mendukung selama studi di MTI.
6. Kepada teman-teman seperjuangan saya di MTI yang telah saling mengingatkan, mendukung dan saling membantu selama studi S2 hingga ujian dan lulus bersama. Terima kasih, berkat kalian kita bisa lulus bersama.

HALAMAN MOTTO

Ingatlah, sesungguhnya pertolongan Allah itu dekat. (Al Baqarah: 214)

Segala sesuatu yang terjadi termasuk pertemuan kita dengan orang-orang dalam hidup kita tidak ada yang terjadi secara kebetulan, semua pasti ada hikmah/maknanya. Hanya kita saja yang sering lupa memaknai setiap kejadian.

Lakukan yang terbaik lalu pasrahkan hasilnya pada Allah, pemberian Allah selalu yang terbaik. Jika tidak sekarang, nanti pasti kita akan merasakan manfaatnya.



KATA PENGANTAR

Dengan Rahmat Allah SWT yang telah memberikan rahmat dan hidayahnya sehingga tesis ini dapat selesai tepat waktu, terima kasih kepada:

1. Prof. Dr. Suyanto selaku rektor Universitas Amikom Yogyakarta.
2. Prof. Dr. Ema Utami, S.Si., M.Kom dan Dr. Suwanto Raharjo, S.Si., M.Kom yang telah dengan sabar membimbing, memotivasi dan memberikan waktu serta pemikirannya hingga selesainya tesis ini.
3. Dr. Kusriani, M.Kom selaku direktur MTI yang telah memberikan kemudahan selama pembelajaran di MTI AMIKOM.
4. Dr. Kusriani, M.Kom., Dr. Andi Sunyoto, M.Kom dan Prof. Dr. Ema Utami, S.Si., M.Kom selaku Tim Penguji yang telah membimbing dan mengarahkan sampai tesis selesai.
5. Teman-teman MTI 20 B yang telah berjuang bersama.

Yogyakarta, 04 Mei 2020

Penulis

DAFTAR ISI

HALAMAN JUDUL	ii
HALAMAN PENGESAHAN	iii
HALAMAN PERSETUJUAN	iv
HALAMAN PERNYATAAN KEASLIAN TESIS	v
HALAMAN PERSEMBAHAN	vi
HALAMAN MOTTO	vii
KATA PENGANTAR	viii
DAFTAR ISI	ix
DAFTAR TABEL	xii
DAFTAR GAMBAR	xiii
INTISARI	xvi
<i>ABSTRACT</i>	xvii
BAB I PENDAHULUAN	1
1.1. Latar Belakang Masalah	1
1.2. Rumusan Masalah	5
1.3. Batasan Masalah	5
1.4. Tujuan Penelitian	6
1.5. Manfaat Penelitian	6
BAB II TINJAUAN PUSTAKA	8
2.1. Tinjauan Pustaka	8
2.2. Keaslian Penelitian	12

2.3. Landasan Teori.....	15
2.3.1 Pengertian Facebook	15
2.3.2 Pengertian Karakter.....	15
2.3.3 Teori Big Five.....	15
2.3.4 <i>Data Mining</i>	19
2.3.5 Tahap-Tahap <i>Data Mining</i>	21
2.3.6 Naive Bayes.....	22
2.3.7 Backpropagation Neural Network.....	25
2.3.8 Arsitektur <i>Backpropagation</i>	28
BAB III METODE PENELITIAN	31
3.1. Jenis, Sifat, dan Pendekatan Penelitian	31
3.2. Metode Pengumpulan Data.....	31
3.3. Metode Analisis Data	31
3.4. Alur Penelitian.....	32
3.4.1 Pendahuluan.....	33
3.4.2 Studi Literatur.....	33
3.4.3 Pengumpulan dan Pengolahan Data.....	33
3.4.4 Ekstraksi Fitur.....	34
3.4.5 Klasifikasi.....	34
3.4.6 Perbandingan Hasil Akurasi	40
3.4.7 Penarikan Kesimpulan.....	41

BAB IV HASIL PENELITIAN DAN PEMBAHASAN.....	42
4.1. Pengumpulan dan Pengolahan Data.....	42
4.1.1 <i>Pre-processingText</i>	42
4.1.2 <i>Data Cleaning</i>	43
4.2. Ekstraksi fitur.....	43
4.3. Klasifikasi.....	45
4.3.1 <i>Algoritma Naive Bayes</i>	46
4.3.2 <i>Algoritma Backpropagation Neural Network</i>	56
4.4. Pengujian	67
4.5. Evaluasi	67
BAB V KESIMPULAN.....	86
5.1. Kesimpulan.....	86
5.2. Saran.....	87
DAFTAR PUSTAKA	88

DAFTAR TABEL

Tabel 4 . 1. Contoh Proses <i>Case Folding</i>	42
Tabel 4 . 2. Fitur	44
Tabel 4 . 3. Data Training Perhitungan Manual Naive Bayes.....	48
Tabel 4 . 4. Data Testing Perhitungan Manual Naive Bayes	49
Tabel 4 . 5. Layer Input.....	57
Tabel 4 . 6. Layer Output	58
Tabel 4 . 7. Hasil Perhitungan Nilai y	60
Tabel 4 . 8. Tabel Nilai Target	60
Tabel 4 . 9. Nilai Error Masing-masing Neuron pada Layer Output.....	61
Tabel 4 . 10. Nilai y Baru Hasil Backward Propagation.....	61
Tabel 4 . 11. Nilai Error Setelah Backprop	62
Tabel 4 . 12. Percobaan Skenario 1 Algoritma <i>Backpropagation</i>	65
Tabel 4 . 13. Percobaan Skenario 2 Algoritma <i>Backpropagation</i>	67
Tabel 4 . 14. Komposisi Jumlah Data per Label dari Dataset.....	68
Tabel 4 . 15. Jumlah Data Tiap Label pada Dataset Awal.....	70
Tabel 4 . 16. Jumlah Status Per User (Label <i>Neuroticism</i>).....	71
Tabel 4 . 17. Jumlah Status Per User (Label <i>Extroversion</i>).....	71
Tabel 4 . 18. Jumlah Status Per User (Label <i>Agreeableness</i>)	72
Tabel 4 . 19. Jumlah Status Per User (Label <i>Consciousness</i>).....	73
Tabel 4 . 20. Jumlah Status Per User (Label <i>Openness</i>).....	75
Tabel 4 . 21. Jumlah Data Tiap Label pada Proses <i>Balancing</i> Data Skenario 3 ...	83

DAFTAR GAMBAR

Gambar 1. Alur Penelitian.....	32
Gambar 2 . 1. Contoh Form Tes Kepribadian Big Five.....	18
Gambar 2 . 2. Arsitektur <i>Backpropagation</i>	29
Gambar 3. 1. Operator Read Excel.....	35
Gambar 3. 2. Import file.....	35
Gambar 3. 3. Set role.....	36
Gambar 3. 4. Operator Replace Missing Value.....	37
Gambar 3. 5. Operator Cross Validation.....	38
Gambar 3. 6. Detail proses dari Cross Validation.....	38
Gambar 3. 7. Klasifikasi dengan Algoritma <i>Naive Bayes</i>	39
Gambar 3. 8. Klasifikasi dengan Algoritma <i>Backpropagation Neural Network</i>	39
Gambar 3. 9. Operator Apply Model.....	40
Gambar 3. 10. Operator Performance.....	40
Gambar 4 . 1. Contoh Dataset Setelah Proses Ekstraksi.....	45
Gambar 4 . 2. Contoh Dataset Setelah Proses <i>Pivot Table</i>	46
Gambar 4 . 3. Flowchart Algoritma <i>Naive Bayes</i>	47
Gambar 4 . 4. Skenario 1 Proses Klasifikasi dengan Algoritma <i>Naive Bayes</i>	55
Gambar 4 . 5. Skenario 2 Proses Klasifikasi dengan Algoritma <i>Naive Bayes</i>	55
Gambar 4 . 6. Proses <i>Naive Bayes</i> di dalam Proses <i>Cross Validation</i>	56
Gambar 4 . 7. Model / Arsitektur Sederhana Neural Network dari Perhitungan Manual.....	57

Gambar 4 . 8. Alur Forward dan Backward	58
Gambar 4 . 9. Matriks bobot.....	59
Gambar 4 . 10. Matriks Nilai Bias	59
Gambar 4 . 11. Arsitektur Backpropagation dengan 2 Dan 4 Neuron Hidden Layer	63
Gambar 4 . 12. Arsitektur Backpropagation dengan 2, 4 dan 6 Neuron Hidden Layer	63
Gambar 4 . 13. Arsitektur Backpropagation dengan 2, 4, 6 dan 8 Neuron Hidden Layer	64
Gambar 4 . 14. Skenario 1 Proses Klasifikasi dengan Algoritma <i>Backpropagation Neural Network</i>	65
Gambar 4 . 15. Skenario 2 Proses Klasifikasi dengan Algoritma <i>Backpropagation Neural Network</i>	66
Gambar 4 . 16. Persebaran Label Kepribadian Pengguna pada Data Training	69
Gambar 4 . 17. Hasil Akurasi Tiap Kelompok Kepribadian dengan Algoritma <i>Backpropagation Neural Network</i>	70
Gambar 4 . 18. Grafik jumlah status per user untuk label kepribadian Neuroticism dalam dataset	79
Gambar 4 . 19. Grafik jumlah status per user untuk label kepribadian Extroversion dalam dataset	79
Gambar 4 . 20. Grafik jumlah status per user untuk label kepribadian Agreeableness dalam dataset	80

Gambar 4 . 21. Grafik jumlah status per user untuk label kepribadian Consciousness dalam dataset.....	80
Gambar 4 . 22. Grafik jumlah status per user untuk label kepribadian Openness dalam dataset.....	81
Gambar 4 . 23. Hasil Prediksi Algoritma Naive Bayes Setelah Balancing Data dengan Skenario 3.....	84
Gambar 4 . 24. Hasil Prediksi Algoritma Backpropagation Neural Network Setelah Balancing Data dengan Skenario 3.....	85



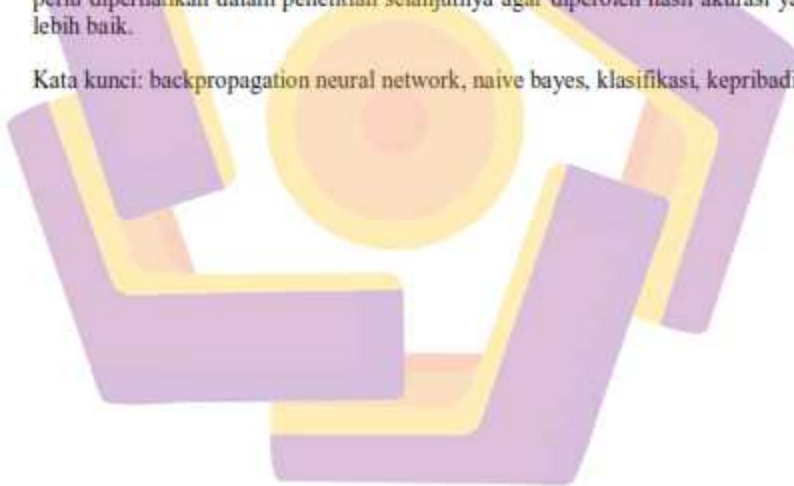
INTISARI

Penilaian kepribadian berdasarkan kategori psikologi tertentu biasanya dilakukan dengan melakukan tes kepribadian. Namun prediksi pengelompokan kepribadian dengan memanfaatkan data yang ada pada akun media sosial penggunanya memungkinkan untuk dilakukan. Penelitian ini melakukan pengelompokan kepribadian pengguna facebook ke dalam kategori *Big Five Personality*. Penulis membandingkan kinerja algoritma backpropagation neural network dan naive bayes untuk klasifikasi.

Ekstraksi fitur memanfaatkan LIWC dan menghasilkan 93 fitur yang digunakan. Dataset dalam penelitian memanfaatkan dataset dari MyPersonality. Tahapan penelitian meliputi pengumpulan data, *pre-processing*, ekstraksi fitur, klasifikasi ke dalam kedua algoritma yang dibandingkan.

Hasil akurasi dalam penelitian ini tidak sesuai dengan yang diharapkan yaitu algoritma naive bayes sebesar 37,32% dan algoritma backpropagation neural network sebesar 51,96%. Keseimbangan data yang digunakan dan pemilihan fitur perlu diperhatikan dalam penelitian selanjutnya agar diperoleh hasil akurasi yang lebih baik.

Kata kunci: backpropagation neural network, naive bayes, klasifikasi, kepribadian



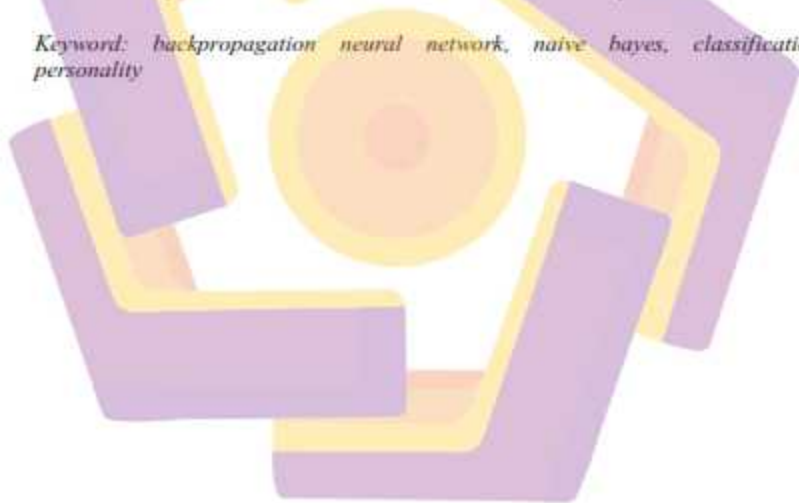
ABSTRACT

Personality assessments based on certain psychological categories are usually done by conducting personality tests. But the prediction of personality groupings by utilizing the data available on the user's social media account is possible. This study grouped Facebook users' personalities into the Big Five Personality category. The author compares the performance of the backpropagation neural network algorithm and Naive Bayes for classification.

Feature extraction utilizes LIWC and produces 93 features used. The dataset in research utilizes the dataset from MyPersonality. Stages of research include data collection, pre-processing, feature extraction, classification into the two algorithms that are compared.

Accuracy results in this study were not as expected, namely the Naive Bayes algorithm by 37.32% and the backpropagation neural network algorithm by 51.96%. The balance of the data used and the selection of features need to be considered in future studies in order to obtain better accuracy results.

Keyword: backpropagation neural network, naive bayes, classification, personality



BAB I

PENDAHULUAN

1.1. Latar Belakang Masalah

Faktor yang mempengaruhi kinerja perusahaan salah satunya adalah faktor SDM (Sumber Daya Manusia). Selain dari faktor lain seperti sumber daya fisik, sumber daya finansial dan proses bisnis perusahaan, SDM merupakan subyek yang memiliki peranan penting terhadap kinerja perusahaan. Menurut (Moran dan Brightman, 2000), SDM penting karena merupakan subyek yang akan melaksanakan proses perubahan dan hasil dari proses perubahan. Oleh karena itu penting dilakukan pengelolaan tenaga kerja yang baik.

Salah satu langkah untuk melakukan pengelolaan SDM adalah dengan mengetahui kepribadian tenaga kerja yang dimiliki perusahaan tersebut. Kepribadian tenaga kerja dalam dunia kerja mempengaruhi kinerja organisasi atau perusahaan seperti dikemukakan dalam penelitian Robertson & Callinan (2015) yang menyebutkan bahwa dalam mempertimbangkan peran kepribadian dalam kehidupan kerja pada perusahaan, perhatian dapat difokuskan pada tingkat perilaku individu, interaksi antara dua orang, dan pada kontribusi orang ke tim, kelompok, atau unit kerja yang lebih besar. Salah satu masalah utama yang dibahas dalam penelitiannya adalah kinerja orang dan / atau efektivitas di tempat kerja. Ketika mempertimbangkan sejauh mana kepribadian mungkin mempengaruhi kinerja pekerjaan, beberapa jenis variabel yang relevan.

Identifikasi kepribadian pada umumnya dilakukan dengan menggunakan kuesioner. Namun saat ini telah banyak HRD perusahaan yang mengidentifikasi kepribadian calon tenaga kerja mereka melalui media sosial. Apa yang dituliskan orang dalam media sosialnya dapat mencerminkan siapa dirinya. Media sosial merupakan tempat bagi pengguna untuk menampilkan diri kepada dunia dan banyak mengungkap detail kehidupan mereka. Sehingga data tersebut dapat dimanfaatkan untuk memprediksi kepribadian pengguna. Seperti yang dibahas pada jurnal karya Jennifer Golbeck, et al (2011).

Media sosial yang dinilai tetap eksis seiring dengan bermunculannya media sosial baru lainnya adalah Facebook. Perusahaan Facebook bahkan saat ini telah mengakuisisi media sosial lain yang memiliki pengguna dalam jumlah besar seperti Instagram dan Whatsapp. Hal tersebut membuktikan eksistensi Facebook dan media sosial Facebook hingga saat penelitian ini dilakukan, masih memiliki sejumlah besar pengguna aktif.

Beberapa kata yang tersusun pada status Facebook pengguna dapat dianalisis menggunakan beberapa tipe kepribadian, seperti DISC (Dominance, Influence, Steadiness, Compliance), MBTI (Myers Briggs Type Indicator), The Big Five (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism), RIASEC (Realistic, Investigative, Artistic, Social, Enterprising, Conventional) dan beberapa tipe kepribadian lainnya untuk mengetahui kepribadian dari pengguna Facebook tersebut. Dalam penelitian ini, penulis akan menggunakan Big Five untuk mengategorikan kepribadian pengguna Facebook dari data yang dimiliki. Dengan memahami kepribadian pengguna, maka cara ini

dapat digunakan sebagai parameter saat rekrutment seleksi tenaga kerja, sehingga pihak HRD (Human Resource Department) akan lebih efektif untuk memahami kepribadian tenaga kerja tersebut di dunia maya. Penelitian Nadeem Ahmad dan Jawaid Siddique (2017) telah berhasil mengelompokkan kepribadian ke dalam kerangka DISC dengan memanfaatkan twitter sebagai sumber datanya.

Analisis kepribadian melalui media sosial akan mudah diterapkan untuk proses *recruitment* calon karyawan jika jumlah calon karyawan sedikit. Namun akan menjadi masalah jika proses perekrutan calon karyawan dalam jumlah besar. Kepribadian hanya salah satu dari tolak ukur penilaian karyawan, jika proses identifikasi kepribadian calon karyawan memakan waktu lama tentu akan mempengaruhi efektifitas proses *recruitment*. Oleh karena itu diperlukan sistem yang secara otomatis mampu mengklasifikasikan data dari media sosial ke dalam klasifikasi kepribadian calon karyawannya.

Proses pengklasifikasian tweet dari pengguna memerlukan metode yaitu *text mining*. Terdapat beberapa algoritma yang dapat digunakan dalam melakukan *text mining* seperti Neural Network, Decision Tree, Logistic Regression, Naive Bayes, K-Nearest Neighbor (Dewi, 2016). Penelitian ini akan membandingkan dua algoritma yaitu naive bayes dan neural network karena (Dewi, 2016) telah menemukan bahwa algoritma neural network memiliki kinerja terbaik dibandingkan dengan empat algoritma lain yang memiliki akurasi sebesar 89,71% sedangkan algoritma naive bayes memiliki kinerja paling rendah tingkat akurasi sebesar 84,70% ketika diterapkan untuk memprediksi keberhasilan pemasaran produk perbankan. Algoritma neural network juga pernah diterapkan untuk

memprediksi kepribadian dari pengguna media sosial pada penelitian yang dilakukan oleh Kemas Muslim Lhaksana, Fhira Nhita, Duwi Anggraini (2017) dan memperoleh akurasi yang cukup baik yaitu sebesar 84%.

Algoritma Naive Bayes dapat diterapkan untuk mengklasifikasi kepribadian seseorang berdasarkan riwayat tweetnya seperti yang dilakukan oleh Mohamad Rozqi dan Wayan Firdaus Mahmudy (2015). Algoritma Naive Bayes memiliki akurasi 100% ketika dibandingkan dengan hasil pakar pada penelitian tersebut. Namun jumlah data yang digunakan dalam penelitian tersebut hanya 10 data sebagai data latih dan 10 data sebagai data uji, sehingga perlu dilakukan penelitian menggunakan data set yang lebih banyak untuk membandingkan dan menguji kehandalan algoritma naive bayes dalam mengklasifikasi kepribadian dari tweet. Penelitian Bayu Yudha Pratama dan Rianarto Sarno (2015) membandingkan algoritma naive bayes, K-Nearest Neighbour dan Support Vector Machine dan hasilnya adalah algoritma naive bayes sedikit mengungguli dua algoritma lain dinilai dari akurasinya dalam kasus klasifikasi kepribadian dari tweet pengguna twitter. Pada penelitian terbaru lainnya, Yusra, dkk (2018) mengklasifikasi kepribadian pengguna twitter dengan metode naive bayes mendapatkan hasil akurasi yang cukup baik yaitu sebesar 86,66%.

Diharapkan dengan membandingkan kinerja Algoritma *Backpropagation Neural Network* dengan *Naive Bayes* dapat diketahui algoritma mana yang lebih cocok untuk mengklasifikasi karakter individu dengan menggunakan kerangka Big Five. Tentunya diharapkan juga penelitian ini dapat berguna sebagai penunjang seperti pada pihak HRD perusahaan untuk mengetahui karakter

individu dari status pada linimasa akun Facebooknya, atau untuk pihak konseling pada sekolah untuk mengetahui karakter siswa sehingga dapat membantu siswa melakukan penggalan potensi yang dimiliki berdasarkan karakter yang dimiliki.

1.2. Rumusan Masalah

Berdasarkan latar belakang masalah yang telah dijelaskan diatas, maka rumusan masalah dalam penelitian ini adalah :

- a. Bagaimana Algoritma *Backpropagation Neural Network* dan *Naive Bayes* dapat mengklasifikasikan status pengguna Facebook kedalam kepribadian *Big Five*?
- b. Berapa tingkat akurasi Algoritma *Backpropagation Neural Network* dibandingkan dengan Algoritma *Naive Bayes* dalam pengklasifikasian kepribadian berdasarkan status Facebook?

1.3. Batasan Masalah

Beberapa batasan masalah dalam penelitian ini adalah sebagai berikut :

- a. Media sosial yang menjadi sumber data dalam penelitian ini adalah Facebook.
- b. Dataset yang digunakan memanfaatkan dataset *My Personality*.
- c. Hasil dari klasifikasi dengan Algoritma *Backpropagation Neural Network* dan *Naive Bayes* hanya ke dalam kepribadian *Big Five*.
- d. Penelitian ini hanya membandingkan Algoritma *Backpropagation Neural Network* dan *Naive Bayes* dalam klasifikasi karakter seseorang.
- e. Penulis melakukan beberapa variasi percobaan untuk mengoptimasi dan mencari model yang terbaik untuk mendapatkan akurasi terbaik pada

algoritma *Backpropagation Neural Network*, sedangkan untuk algoritma *Naive Bayes* penulis menggunakan model dasar algoritma naive bayes dan tidak melakukan optimasi apapun.

1.4. Tujuan Penelitian

Tujuan umum yang ingin dicapai dalam penelitian ini adalah untuk mengetahui bagaimana pemanfaatan media sosial Facebook dalam mengklasifikasi karakteristik seseorang. Tujuan khusus penelitian ini adalah sebagai berikut :

- a. Penelitian ini bertujuan untuk mengetahui bagaimana Algoritma *Backpropagation Neural Network* dan *Naive Bayes* dapat mengklasifikasikan status Facebook kedalam kepribadian *Big Five*.
- b. Penelitian ini ditujukan untuk membandingkan seberapa besar tingkat akurasi Algoritma *Backpropagation Neural Network* dengan *Naive Bayes* dalam mengklasifikasi kepribadian seseorang ke dalam kepribadian *Big Five*.
- c. Penelitian ini juga ditujukan untuk mengetahui apakah status Facebook dapat digunakan untuk mengklasifikasikan kepribadian seseorang dengan Algoritma *Backpropagation Neural Network* dan *Naive Bayes* dengan metode *Big Five*.

1.5. Manfaat Penelitian

Melalui penelitian ini diharapkan dapat diperoleh manfaat sebagai berikut :

- a. Bagi peneliti mengetahui kinerja Algoritma Backpropagation Neural Network dan Naive Bayes dalam mengklasifikasi karakter seseorang ke dalam kepribadian *Big Five* pada status seseorang.
- b. Bagi Perusahaan mempermudah dalam perekrutan tenaga kerja baru ditinjau dari kepribadian seseorang yang didapatkan dari status di akun Facebooknya.
- c. Bagi Konseling sekolah membantu penggalan potensi siswa ditinjau dari kepribadian siswa yang didapatkan dari status di akun Facebooknya.
- d. Bagi Psikolog membantu pekerjaan sebagai penunjang keputusan karakter seseorang yang didapatkan dari status di akun Facebooknya.
- e. Menjadi referensi penelitian-penelitian di masa mendatang yang berkaitan dengan klasifikasi kepribadian melalui media Facebook dengan algoritma *Backpropagation Neural Network* dan *Naive Bayes*.

BAB II

TINJAUAN PUSTAKA

2.1. Tinjauan Pustaka

Analisis kepribadian menggunakan media sosial telah banyak diterapkan. Keberhasilan melakukan penilaian kepribadian menggunakan data yang bersumber dari media sosial juga telah dibuktikan mulai dari penelitian terdahulu yang dilakukan oleh Qiu, et al (2012) yang berhasil mengklasifikasikan kepribadian dengan cara menganalisis isyarat linguistik dari tweet yang dilakukan pengguna. Dalam penelitian lain yang dilakukan oleh Farnadi, et al (2016) dalam penelitiannya mengenai *personality recognition* telah membuat pemodelan metode untuk mengklasifikasikan kepribadian dengan memanfaatkan tiga sumber data yaitu facebook, twitter dan youtube. Dari penelitian tersebut disampaikan peluang penelitian selanjutnya untuk dapat memperoleh hasil akurasi yang baik perlu dilakukan penelitian dengan memperhatikan variasi pengguna media sosial dan menerapkan metode yang tepat.

Penelitian yang dilakukan oleh Nadeem Ahmad dan Jawaid Siddique (2017) dalam jurnal yang berjudul "*Personality Assesment using Twitter Tweets*" berhasil mengelompokkan kepribadian pengguna twitter ke dalam kerangka DISC. Namun dalam jurnal tersebut tidak disebutkan algoritma apa yang digunakan untuk melakukan klasifikasi dan tidak disebutkan juga berapa jumlah data yang digunakan dalam penelitian. Dalam jurnal tersebut juga tidak dijelaskan langkah pengujian dan hasil akurasi dari pengelompokkan kepribadian yang dilakukan.

Penelitian ini akan mengelompokkan kepribadian pengguna ke dalam *Big Five* dengan membandingkan dua algoritma yaitu Backpropagation Neural Network dan Naive Bayes lalu akan dilakukan justifikasi hasil pada ahlinya dalam hal ini adalah psikolog.

Algoritma Neural network dibandingkan dengan empat algoritma lain yang antara lain Naive Bayes, Decision Tree, Logistic Regression dan K-Nearest Network pernah dilakukan pada penelitian Sari Dewi (2016) untuk tujuan prediksi keberhasilan pemasaran produk perbankan dan memperoleh hasil algoritma neural network memiliki tingkat akurasi paling tinggi yaitu sebesar 89,71% dengan nilai AUC 0,872. Penelitian ini akan membandingkan dua algoritma yang dinilai paling tinggi dan paling rendah akurasinya dalam penelitian sebelumnya yaitu neural network dan naive bayes untuk diterapkan pada klasifikasi kepribadian.

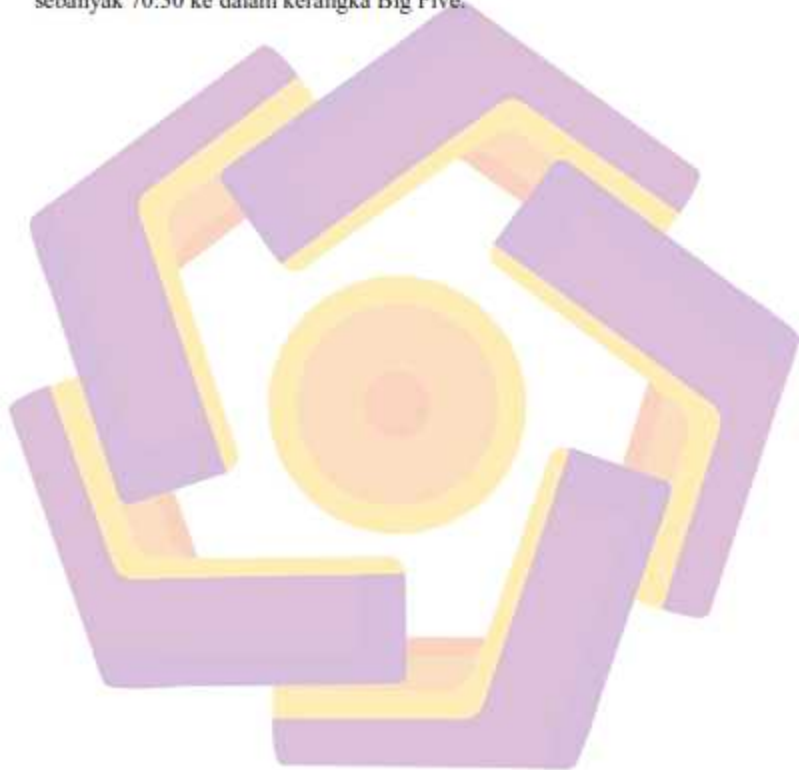
Jurnal yang berjudul "Klasifikasi Kepribadian Berdasarkan Status Facebook Menggunakan Metode Backpropagation" yang ditulis oleh Kemas Muslim Lhaksana, Fhira Nhita, Duwi Anggraini (2017) melakukan pengklasifikasian kepribadian menggunakan metode Backpropagation dengan parameter ANN menggunakan 2, 4, 6 dan 8 neuron pada hidden layer. Memperoleh hasil akurasi sebesar 84,00%. Penelitian tersebut memperoleh hasil yang baik namun media sosial yang digunakan adalah facebook. Sedangkan penelitian ini akan menggunakan twitter sebagai sumber datanya. Dalam mengekstraksi fitur, penelitian tersebut memanfaatkan tools LIWC, sedangkan dalam penelitian ini akan melakukan pembobotan fitur menggunakan TF*IDF.

Penelitian Bayu Yudha Pratama dan Rianarto Sarno (2015) Membandingkan metode Naive Bayes, KNN dan SVM untuk mengklasifikasi kepribadian berdasarkan text di Twitter. Bahasa yang digunakan dalam penelitian tersebut adalah bahasa Indonesia dan Inggris. Hasil yang diperoleh adalah metode Naive Bayes sedikit mengungguli metode lain. Namun jika dilihat dari segi akurasi dibandingkan dengan hasil dari tes menggunakan kuesioner, penelitian tersebut tidak berhasil meningkatkan akurasi dimana akurasi hanya 65%.

Penerapan algoritma naive bayes untuk pengklasifikasian kepribadian juga dilakukan pada penelitian Mohammad Zoqi Sarwani dan Wayan Firdaus Mahmudy (2015) dalam jurnal yang berjudul "Analisis Twitter untuk Mengetahui Karakter Seseorang Menggunakan Algoritma Naive Bayes Classifier". Penelitian tersebut menggunakan 10 akun twitter sebagai data latih dan 10 akun twitter sebagai data uji dan memperoleh hasil akurasi sebesar 100% setelah dibandingkan dengan hasil dari pakar. Jumlah data set yang digunakan perlu ditambah untuk bisa menarik kesimpulan akurasi algoritma tersebut sehingga dalam penelitian ini akan menambahkan jumlah data set.

Dalam penelitian lain yang lebih baru karya Yusra, Muhammad Fikry, Rinaldi Syarifanto, Reski-Mai Candra dan Elvia Budianita (2018) yang berjudul "Klasifikasi Kepribadian Big Five Pengguna Twitter dengan Metode Naive Bayes" dilakukan klasifikasi kepribadian Big Five pengguna Twitter dengan metode Naive Bayes. Data set yang digunakan dalam penelitian tersebut sebanyak 15 akun dengan jumlah tweet sebanyak 1500. Penelitian tersebut melakukan uji coba membandingkan komposisi data latih dan data uji dengan perbandingan

60:40, 70:30, 80:20 dan 90:10. Hasil yang diperoleh adalah komposisi dengan akurasi tertinggi adalah 70:30 dengan akurasi sebesar 86,66%. Penelitian ini akan mengadopsi hasil dari penelitian tersebut dengan melakukan klasifikasi menggunakan metode naive bayes dengan komposisi data training dan data testing sebanyak 70:30 ke dalam kerangka Big Five.



2.2. Keaslian Penelitian

Tabel 2.1. Matriks literatur review dan posisi penelitian Analisis Klasifikasi Kepribadian Berdasarkan Media Sosial dengan Algoritma *Backpropagation Neural Network* dan *Naive Bayes*

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
1	Komparasi 5 Metode Algoritma Klasifikasi Data Mining pada Prediksi Keberhasilan Pemasaran Produk Layanan Perbankan	Sari Dewi, Jurnal Techno Nusa Mandiri Vol. XIII, No. 1 Maret 2016	Untuk mengetahui algoritma mana yang lebih akurat antara neural network, decision tree, logistic regression, naive bayes dan KNN dalam memprediksi tingkat keberhasilan telemarketing dalam layanan produk perbankan.	Dalam kasus ini, algoritma neural network memiliki tingkat akurasi paling tinggi dibanding 4 algoritma lain yaitu sebesar 89,71% dengan nilai AUC 0,872.	Saran : Penelitian ini dapat dikembangkan pada unit bisnis atau kasus lain.	Pada penelitian yang akan dilakukan ini, akan melakukan perbandingan antara dua algoritma yaitu backpropagation neural network dan naive bayes. Penerapannya akan dilakukan untuk penilaian kepribadian.
2	Klasifikasi Kepribadian Berdasarkan Status Facebook Menggunakan Metode Backpropagation	Kemas Muslim Lhaksmana, Fhira Nhita, Duwi Anggraini, e-Proceeding of Engineering : Vol. 4, No. 3 Desember 2017	Memprediksi kepribadian pelamar kerja berdasarkan status facebooknya menggunakan metode backpropagation	Penelitian tersebut dapat mengklasifikasikan kepribadian berdasarkan status facebook menggunakan metode backpropagation dengan melakukan tiga kali pengujian dan mendapatkan hasil akurasi tertinggi sebesar 84%.	Penelitian tersebut membagi hasil klasifikasi ke dalam 2 kelas, namun tidak disebutkan dasar pelabelannya. Fitur dari LIWC yang digunakan hanya posemo, negemo dan socil word.	Penelitian ini akan menggunakan metode yang sama yaitu backpropagation dan menggunakan 93 fitur yang diekstraksi menggunakan LIWC. Pengelompokkannya akan dilakukan ke dalam kategori kepribadian Big Five.

Tabel 2.1 (Lanjutan)

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
3	Personality Assessment using Twitter Tweets	Nadeem Ahmad dan Jawaid Siddique, 21 st International Conference on Knowledge Based and Intelligent Information an Engineering Systems, KES2017, September 2017	Untuk menganalisis bagaimana twitter dapat digunakan untuk penilaian kepribadian.	Penelitian tersebut berhasil memanfaatkan data dari twitter untuk melakukan klasifikasi kepribadian ke dalam kerangka DISC.	Tidak disebutkan dengan jelas algoritma klasifikasi yang digunakan dalam penelitian ini.	Penelitian ini akan mengklasifikasikan kepribadian ke dalam kerangka Big Five menggunakan algoritma backpropagation neural network dan naive bayes.
4	Personality Classification Based on Twitter Text Using Naive Bayes, KNN and SVM.	Bayu Yudha Pratama dan Rivanarto Surno, International Conference on Data Software Engineering, 2015.	Membangun sistem kerja untuk memprediksi kepribadian dari twitter dan ingin meningkatkan akurasi dari penelitian sebelumnya.	Prediksi kepribadian berhasil dilakukan dalam penelitian tersebut. Naive bayes memiliki akurasi yang paling tinggi dibandingkan dua algoritma lain yang dibandingkan.	Saran : Perbaikan yang dapat dilakukan untuk penelitian selanjutnya adalah dengan menggunakan dataset yang lebih akurat untuk meningkatkan akurasi.	Penelitian ini menggunakan facebook sebagai sumber datanya. Menggunakan algoritma naive bayes juga dan kemudian akan dibandingkan dengan algoritma backpropagation neural network.

Tabel 2.1 (Lanjutan)

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
5	Klasifikasi Kepribadian Big Five Pengguna Twitter dengan Metode Naive Bayes	Yusra, Muhammad Fikry, Rinaldi Syarifanto, Reski Mai Candra dan Elvia Budianita, Seminar Nasional Teknologi Informasi, Komunikasi dan Industri, 2018.	Melakukan klasifikasi kepribadian pengguna twitter ke dalam kerangka big five menggunakan metode naive bayes.	Metode naive bayes dapat digunakan untuk mengklasifikasikan kepribadian big five dari pengguna twitter. Akurasi tertinggi dalam penelitian ini diperoleh pada hasil pengujian dengan komposisi data latih dan data uji 70:30 dengan tingkat akurasi sebesar 86,66%.	Saran : Perlu dilakukan normalisasi terhadap kata-kata yang ditulis tidak sesuai aturan berbahasa.	Penelitian ini akan memanfaatkan hasil penelitian sebelumnya yaitu membagi data set menjadi data latih dan data uji sebanyak 70 : 30.
6	Analisis Twitter untuk Mengetahui Karakter Seseorang Menggunakan Algoritma Naive Bayes Classifier	Mohamad Zoqi Sarwani dan Wayan Firdaus Mahmady, Seminar Nasional Sistem Informasi Indonesia, 24 November 2015	Mengklasifikasikan kepribadian pengguna twitter menggunakan algoritma naive bayes.	Twitter dapat digunakan sebagai salah satu media untuk mengetahui kepribadian seseorang. Pengklasifikasian menggunakan naive bayes memberikan tingkat akurasi yang baik dengan membandingkan hasil klasifikasi dari sistem dengan hasil dari pakar.	Data set akun twitter yang digunakan adalah 10 akun twitter sebagai data latih dan 10 akun sebagai data uji. Jumlah tersebut terlalu sedikit untuk dapat mengambil kesimpulan tingkat akurasi dari algoritma tersebut.	Penelitian ini akan menambah jumlah data set yang akan digunakan.

2.3. Landasan Teori

2.3.1 Pengertian Facebook

Facebook merupakan sebuah jejaring sosial yang diluncurkan pada bulan Februari 2004. Pengguna harus mendaftar untuk dapat menggunakan situs ini. Pengguna dapat membuat profil pribadi, menambahkan teman, bertukar pesan, saling berkomentar dan terdapat pemberitahuan otomatis untuk para penggunanya. Situs jejaring sosial Facebook ini didirikan oleh Mark Zuckerberg bersama temannya sesama mahasiswa Universitas Harvard. Facebook memiliki lebih dari 1 milyar pengguna aktif dan masih terus eksis hingga penelitian ini dilakukan.

2.3.2 Pengertian Karakter

Karakter adalah watak, tabiat, akhlak, atau juga kepribadian seseorang yang terbentuk dari hasil internalisasi berbagai kebajikan yang diyakini dan mendasari cara pandang, berpikir, sikap, dan cara bertindak orang tersebut. Kebajikan tersebut terdiri atas sejumlah nilai, moral, dan norma seperti jujur, berani bertindak, dapat dipercaya, hormat kepada orang lain.

2.3.3 Teori Big Five

Kepribadian big five merupakan salah satu metode yang terkenal dalam dunia psikologi untuk menginterpretasikan kepribadian seseorang. Kepribadian big five terdiri dari (Golbeck, 2011):

- a. *Openness to Experience*: sifat ini menilai keinginan individu untuk mencari dan menghargai pengalaman baru serta kecenderungan untuk senang mengetahui sesuatu yang tidak familiar (Pervin & John, 2001).

- b. *Conscientiousness*: sifat ini menilai kemampuan individu dalam organisasi, baik mengenai tingkat keteraturan maupun motivasi dalam mencapai tujuan (Pervin & John, 2001).
- c. *Extraversion*: sifat ini menilai intensitas interaksi interpersonal, tingkat ketergantungan dengan orang lain dan kemampuan untuk berbahagia.
- d. *Agreeableness*: Orang yang memiliki skor tinggi pada sifat ini adalah orang yang lembut hati, dapat dipercaya, suka menolong, pemaaf, penurut.
- e. *Neuroticism*: Sifat ini menggambarkan stabilitas emosional dengan cakupan perasaan negatif yang kuat termasuk kecemasan, kesedihan, iritabilitas dan ketidak percayadirian.

Terdapat beberapa macam alat ukur untuk mengukur kepribadian seseorang ke dalam kategori Big Five. Beberapa alat ukur tersebut antara lain:

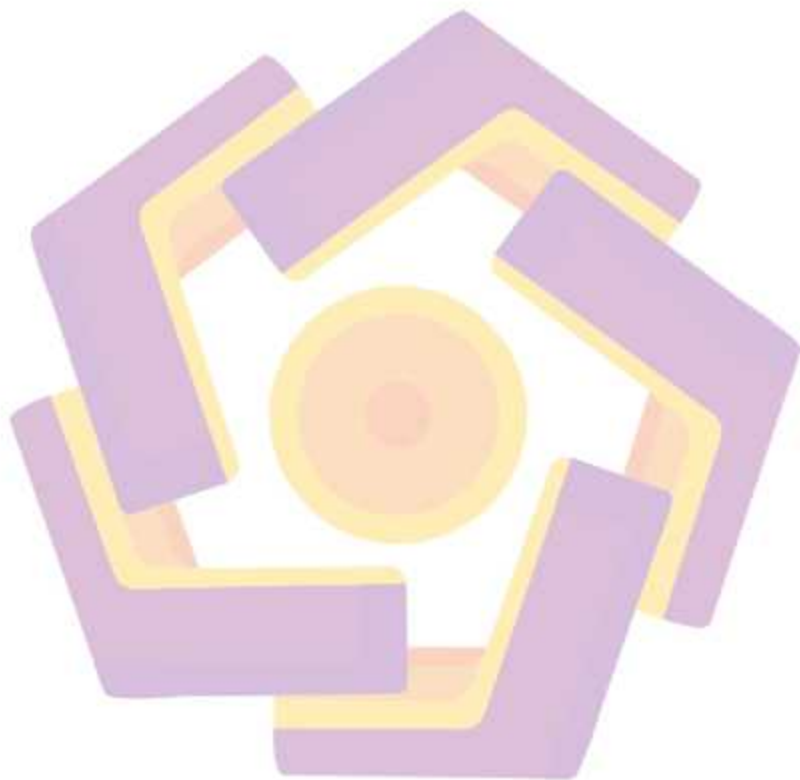
- a. *The Neuroticism Extraversion Openness – Personality Inventory – Revised* (NEO-PI-R)

Jenis tes ini termasuk ke dalam tes medis, sehingga tidak dapat di download di internet dan tidak boleh sembarang orang menginterpretasinya. Alat ukur ini terdiri dari 240 item dan merupakan alat ukur yang dikembangkan oleh Paul T. Costa dan Robert R. McCrae.

- b. *Big Five Inventory* (BFI)

Alat ukur ini dikembangkan oleh John dan Donahue tahun 1999. Alat ukur ini terdiri dari 44 item. Skalanya menggunakan skala likert dengan pilihan jawaban 1-5. Jenis alat ukur ini tidak masuk ke dalam tes medis dan

bebas dipakai selama tujuannya bukan komersil. Form tesnya dapat didownload di internet.



Scale:

The Big Five Inventory (BFI)

Here are a number of characteristics that may or may not apply to you. For example, do you agree that you are someone who likes to spend time with others? Please write a number next to each statement to indicate the extent to which you agree or disagree with that statement.

Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
1	2	3	4	5

I see myself as someone who...

- | | |
|---|--|
| ___1. Is talkative | ___24. Tends to be lazy |
| ___2. Tends to find fault with others | ___24. Is emotionally stable, not easily upset |
| ___3. Does a thorough job | ___25. Is inventive |
| ___4. Is depressed, blue | ___26. Has an assertive personality |
| ___5. Is original, comes up with new ideas | ___27. Can be cold and aloof |
| ___6. Is reserved | ___28. Perseveres until the task is finished |
| ___7. Is helpful and unselfish with others | ___29. Can be moody |
| ___8. Can be somewhat careless | ___30. Values artistic, aesthetic experiences |
| ___9. Is relaxed, handles stress well | ___31. Is sometimes shy, inhibited |
| ___10. Is curious about many different things | ___32. Is considerate and kind to almost everyone |
| ___11. Is full of energy | ___33. Does things efficiently |
| ___12. Starts quarrels with others | ___34. Remains calm in tense situations |
| ___13. Is a reliable worker | ___35. Prefers work that is routine |
| ___14. Can be tense | ___36. Is outgoing, sociable |
| ___15. Is ingenious, a deep thinker | ___37. Is sometimes rude to others |
| ___16. Generates a lot of enthusiasm | ___38. Makes plans and follows through with them |
| ___17. Has a forgiving nature | ___39. Gets nervous easily |
| ___18. Tends to be disorganized | ___40. Likes to reflect, play with ideas |
| ___19. Worries a lot | ___41. Has few artistic interests |
| ___20. Has an active imagination | ___42. Likes to cooperate with others |
| ___21. Tends to be quiet | ___43. Is easily distracted |
| ___22. Is generally trusting | ___44. Is sophisticated in art, music, or literature |

Scoring:

BFI scale scoring ('N' denotes reverse-scored items):

Extraversion: 1, 6R, 11, 16, 21R, 26, 31R, 36
 Agreeableness: 2R, 7, 12R, 17, 22, 27R, 32, 37R, 42
 Conscientiousness: 3, 6R, 13, 18R, 23R, 28, 33, 38, 43R
 Neuroticism: 4, 9R, 14, 19, 24R, 29, 34R, 39
 Openness: 5, 10, 15, 20, 25, 30, 35R, 40, 41R, 44

Gambar 2. 1. Contoh Form Tes Kepribadian Big Five

c. *International Personality Item Pool – Five Factor Inventory (IPIP-FFI)*

Jenis ini merupakan kumpulan item yang dibuat dan dikembangkan oleh peneliti dari seluruh dunia. Terdapat lebih dari 3.000 item dan 250 macam skala untuk alat ukur jenis ini. Terdapat juga skala yang terdiri dari 50 item untuk jenis IPIP ini.

d. *MINI – International Personality Item Pool (MINI-IPIP)*

Alat ukur ini mengadaptasi dari IPIP namun lebih ringkas.

e. *Neo Five-Factor Inventory (NEO-FFI)*

NEO-FFI dirancang oleh Costa dan Crae pada tahun 1992. Alat ukur ini merupakan versi pendek dari NEO-PI-R yang terdiri dari 60 item. Jenis ini menggunakan 7 skala dari 0 sampai 6 mengikuti pola skala likert.

2.3.4 Data Mining

Menurut (ACM, 2006), (Clifton, 2010), data mining didefinisikan sebagai proses penemuan pola-pola baru dari kumpulan-kumpulan data sangat besar, meliputi metode-metode yang merupakan irisan dari artificial intelligence, machine learning, statistics, dan database systems (Suyanto, 2017).

Metode learning secara garis besar terbagi ke dalam dua pendekatan yang antara lain *supervised* dan *unsupervised learning*. Metode *unsupervised learning* diterapkan tanpa adanya *learning* dan *teacher* (label data). Pengelompokan didasarkan dari yang memiliki kemiripan atau kesamaan yang tinggi. Sebaliknya, metode *Supervised learning* merupakan metode *learning* dengan adanya latihan dan pelatih (Budi Santosa, 2007 : 11-12).

Beberapa teknik dan sifat data mining antara lain :

1. Klasifikasi

Klasifikasi adalah menemukan sebuah record data baru ke salah satu dari beberapa kategori atau klas yang telah didefinisikan sebelumnya. Disebut juga "*supervised learning*" (Fajar Astuti H, 2013 : 14).

2. Regresi

Memprediksi nilai dari suatu variabel kontinyu yang diberikan berdasarkan nilai dari variabel lain, dengan mengasumsikan sebuah model ketergantungan linier atau nonlinier (Fajar Astuti H, 2013 : 15).

3. Klasterisasi (Clustering)

Mempartisi data set menjadi beberapa sub-set atau kelompok sedemikian rupa sehingga elemen-elemen dari suatu kelompok tertentu memiliki set properti yang dishare bersama, dengan tingkat similiaritas yang tinggi dalam satu kelompok dan tingkat similiaritas antar kelompok yang rendah. Disebut juga "*unsupervised learning*" (Fajar Astuti H, 2013 : 16).

4. Kaidah Asosiasi (Assosiation Rules)

Mendeteksi kumpulan atribut-atribut yang muncul bersamaan [co-occur] dalam frekuensi yang sering, dan membentuk sejumlah kaidah dari kumpulan-kumpulan tersebut (Fajar Astuti H, 2013 : 17).

5. Pencarian Pola Sekuensial (Sequence Mining)

Mencari sejumlah event yang secara umum terjadi bersama-sama (Fajar Astuti H, 2013 : 18).

2.3.5 Tahap-Tahap *Data Mining*

Sebagai suatu rangkaian proses, data mining dapat dibagi menjadi beberapa tahap, tahap-tahap data mining yaitu sebagai berikut.

1. *Data Selection*

Data yang berada pada database sering kali tidak semuanya dipakai, oleh karena itu hanya data yang sesuai yang akan diambil dari database. Sebagai contoh, sebuah kasus yang meneliti kecenderungan orang membeli dalam kasus market basket analysis, tidak perlu mengambil nama pelanggan, cukup dengan id pelanggan.

2. *Data Cleansing*

Pembersihan data merupakan proses menghilangkan noise dan data yang tidak konsisten atau tidak relevan. Pada umumnya data yang diperoleh, baik dari database suatu perusahaan maupun hasil eksperimen, memiliki isian-isian yang tidak sempurna seperti data yang hilang, data yang tidak valid atau juga hanya salah ketik. Data yang tidak relevan juga lebih baik dibuang. Pembersihan data juga mempengaruhi performa dari data mining karena data yang akan ditangani akan berkurang jumlah dan kompleksitasnya.

3. *Data Transformation*

Data diubah atau digabung ke dalam format yang sesuai untuk diproses dalam data mining. Beberapa metode data mining membutuhkan format data yang khusus sebelum bisa diaplikasikan. Sebagai contoh beberapa metode standar seperti analisis asosiasi dan clustering hanya bisa menerima input data

kategorikal. Karenanya data berupa angka numerik yang berlanjut perlu dibagi-bagi menjadi beberapa interval. Proses ini sering disebut transformasi data.

4. *Data Mining*

Data Mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam data mining sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD (Knowledge Discovery in Databases) secara keseluruhan.

5. *Interpretation / Evaluation*

Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang disebut *interpretation*. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.

2.3.6 *Naive Bayes*

Bayesian didasarkan pada *teorema bayes* yang memiliki kemampuan hampir serupa dengan *decision tree* dan *neural network*. *Teorema Bayes* adalah teorema yang digunakan dalam statistika untuk menghitung peluang suatu hipotesis. Bayes merupakan teknik prediksi berbasis *probabilistic* sederhana yang berdasar pada penerapan *teorema Bayes* (atau aturan Bayes) dengan asumsi independensi (ketidaktergantungan) yang kuat atau naif. Dengan kata lain, dalam *Naive Bayes*, model yang digunakan adalah “model fitur independen” .

Bayesian classification adalah suatu metode pengklasifikasian data dengan model statistik yang dapat digunakan untuk menghitung probabilitas keanggotaan suatu kelas. Metode *Bayesian classification* digunakan menganalisis dalam membantu tercapainya pengambilan keputusan terbaik suatu permasalahan dari sejumlah alternatif. *Bayesian classification* merupakan salah satu metode yang sederhana yang dapat digunakan untuk data yang tidak konsisten dan data bias. Metode *bayes* juga merupakan metode yang baik dalam mesin pembelajaran berdasarkan data *training* dengan berdasarkan probabilitas bersyarat.

Kaitan antara *Naïve Bayes* dengan klasifikasi, korelasi hipotesis, dan bukti dengan klasifikasi adalah hipotesis dalam *teorema Bayes* merupakan label kelas yang menjadi target pemetaan dalam klasifikasi, sedangkan bukti merupakan fitur-fitur yang menjadi masukan dalam model klasifikasi.

Salah satu tugas *data mining* adalah klasifikasi data, yaitu memetakan (mengklasifikasikan) data ke dalam satu atau beberapa kelas yang sudah didefinisikan sebelumnya. *Naïve Bayes Classifier* merupakan salah satu metoda *machine learning* yang memanfaatkan perhitungan probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi probabilitas di masa depan berdasarkan pengalaman di masa sebelumnya (Antonius Rachmat C, 2016). Dasar dari *Naïve Bayes* yang dipakai dalam pemrograman adalah rumus Bayes:

$$P(A|B) = (P(B|A) * P(A)) / P(B) \dots\dots\dots (1)$$

Peluang kejadian A sebagai B ditentukan dari peluang B saat A, peluang A, dan peluang B. Pada pengaplikasiannya nanti rumus ini berubah menjadi :

$$P(C_i|D) = (P(D|C_i) * P(C_i)) / P(D) \dots\dots\dots (2)$$

Naïve Bayes Classifier atau bisa disebut sebagai *Multinomial Naïve Bayes* merupakan model penyederhanaan dari Metode Bayes yang cocok dalam pengklasifikasian teks atau dokumen. Persamaannya adalah:

$$V_{MAP} = \arg \max_{V_j \in V} P(V_j | a_1, a_2, \dots, a_n) \dots\dots\dots (3)$$

Menurut persamaan (3), maka persamaan (1) dapat ditulis:

$$V_{MAP} = \arg \max_{V_j \in V} \frac{P(a_1, a_2, \dots, a_n | P(V_j))}{P(a_1, a_2, \dots, a_n)} \dots\dots\dots (4)$$

Untuk $P(x_1, x_2, x_3, \dots, x_n)$ nilainya konstan untuk semua kategori (V_j) sehingga persamaan dapat ditulis sebagai berikut :

$$V_{MAP} = \arg \max_{V_j \in V} P(X_1, X_2, \dots, X_n | V_j) P(V_j) \dots\dots\dots (5)$$

Persamaan diatas dapat disederhanakan menjadi sebagai berikut :

$$V_{MAP} = \arg \max_{V_j \in V} \prod_{i=1}^n P(X_i | V_j) P(V_j) \dots\dots\dots (6)$$

Keterangan :

V_j = Kategori Headline

$J = 1, 2, 3, \dots, n$. Dimana dalam penelitian ini

j_1 = kategori *headline* sentimen positif

j_2 = kategori *headline* sentimen negatif

j_3 = kategori *headline* sentiment netral

$P(x_i|V_j)$ = Probabilitas x_i pada kategori V_j

$V_j P(V_j)$ = Probabilitas dari V_j

Untuk $P(V_j)$ dan $P(x_i|V_j)$ dihitung pada saat pelatihan dimana persamaannya adalah sebagai berikut :

$$P(V_j) = \frac{|docs_j|}{|contoh|} \dots\dots\dots(7)$$

Jika $P(V_j)$ sudah ditentukan maka hitung jumlah dokumen setiap kategori j dan jumlah dokumen dari semua kategori dengan menggunakan persamaan 7. diatas.

$$P(x_i|V_j) = \frac{nk+1}{nk+|kosakata|} \dots\dots\dots(8)$$

Keterangan :

$|docs_j|$ = jumlah dokumen setiap kategori j

$|contoh|$ = jumlah dokumen dari semua kategori

nk = jumlah frekuensi kemunculan setiap kata

n = jumlah frekuensi kemunculan kata dari setiap kategori

$|kosakata|$ = jumlah semua kata dari semua kategori

Jika $P(x_i|V_j)$ sudah ditentukan maka hitung jumlah frekuensi kemunculan setiap kata ditambah 1 dan jumlah frekuensi kemunculan kata dari setiap kategori dengan menggunakan persamaan 8. diatas.

2.3.7 Backpropagation Neural Network

Jaringan syaraf tiruan dengan layer tunggal memiliki keterbatasan dalam pengenalan pola. Kelemahan ini bisa ditanggulangi dengan menambahkan satu atau beberapa layer tersembunyi di antara layer masukan dan layer keluaran. Jaringan Syaraf Tiruan Backpropagation (JST-BP) melatih jaringan mendapatkan keseimbangan antara kemampuan jaringan untuk mengenali pola yang digunakan selama pelatihan serta kemampuan untuk memberikan respon yang benar terhadap pola masukan yang serupa dengan pola yang dipakai selama pelatihan (Jumarwanto. et al, 2009).

Backpropagation merupakan algoritma pembelajaran yang terwarisi dan biasanya digunakan oleh perceptron dengan banyak lapisan untuk mengubah bobot-bobot yang terhubung dengan neuron-neuron yang ada pada lapisan tersembunyinya. Algoritma backpropagation menggunakan error output untuk mengubah nilai-nilai bobotnya dalam arah mundur (*backward*). Tahap perambatan maju (*forward propagation*) harus dikerjakan terlebih dahulu untuk mendapatkan nilai error tersebut. Saat perambatan maju neuron-neuron diaktifkan dengan menggunakan fungsi aktivasi sigmoid biner yaitu:

$$f(x) = \frac{1}{1+e^{-x}} \dots\dots\dots (9)$$

Adapun langkah-langkah algoritma backpropagation adalah sebagai berikut:

Feed forward:

1. Tiap unit input (x_i , $i=1,2,3,\dots,n$) menerima sinyal x_i dan meneruskan sinyal tersebut pada lapisan yang ada di atasnya (lapisan tersembunyi).
2. Tiap unit tersembunyi (z_j , $j=1,2,3,\dots,p$) menjumlahkan sinyal-sinyal input terbobot:

$$z_{in_j} = v_0 + \sum_{i=1}^n x_i v_{ij} \dots\dots\dots (10)$$

Gunakan fungsi aktivasi untuk menghitung sinyal outputnya:

$$z_j = f(z_{in_j}) \dots\dots\dots (11)$$

kemudian kirimkan sinyal tersebut ke semua unit dilapisan atasnya (unit-unit output).

3. Tiap unit output (y_k , $k=1,2,3,\dots,m$) menjumlahkan sinyal-sinyal input terbobot

$$y_{in_k} = w_0 + \sum_{j=1}^p z_j w_{jk} \dots\dots\dots (12)$$

gunakan fungsi aktivasi untuk menghitung sinyal outputnya:

$$y_k = f(y_{in_k}) \dots\dots\dots (13)$$

kemudian kirimkan sinyal tersebut kesemua unit dilapisan atasnya (unit-unit output)

Backpropagation:

4. Tiap-tiap unit output ($y_k, k=1,2,3,\dots,m$) menerima target pola yang berhubungan dengan pola input pembelajaran. Hitung informasi errornya.

$$\sigma_k = (t_k - y_k) f'(y_{in_k}) \dots\dots\dots (14)$$

kemudian hitung koreksi bobot (yang nantinya akan digunakan untuk memperbaiki w_{jk})

$$\Delta w_{jk} = \sigma_k z_{ij} \dots\dots\dots (15)$$

Hitung juga koreksi bias (yang nantinya akan digunakan untuk memperbaiki nilai w_{0k}):

$$\Delta w_{jk} = \sigma_k \dots\dots\dots (16)$$

Selanjutnya mengirimkan σ_k ini ke unit-unit yang ada dilapisan bawahnya.

5. Tiap-tiap unit tersembunyi ($z_j, j=1,2,3,\dots,p$) menjumlahkan delta inputnya (dari unit-unit yang berada pada lapisan di atasnya) :

$$\sigma_{in_j} = \sum_{k=1}^m \sigma_k w_{jk} \dots\dots\dots (17)$$

kalikan nilai ini dengan turunan dari fungsi aktivasinya untuk menghitung informasi error:

$$\sigma_j = \sigma_{in_j} f'(z_{in_j}) \dots\dots\dots (18)$$

kemudian hitung koreksi bobot (yang nantinya akan digunakan untuk memperbaiki nilai v_{ij}):

$$\Delta v_{jk} = \sigma \sigma_j x_k \dots\dots\dots (19)$$

Hitung juga koreksi bias (yang nantinya akan digunakan untuk memperbaiki nilai v_{0j}):

$$\Delta v_{0j} = \sigma \sigma_j \dots\dots\dots (20)$$

Tiap unit output (y_k , $k=1,2,3,\dots,m$) memperbaiki bias dan bobotnya ($j=0,1,2,3,\dots,p$).

$$w_{jk}(\text{baru}) = w_{jk}(\text{lama}) + \Delta w_{jk} \dots\dots\dots (21)$$

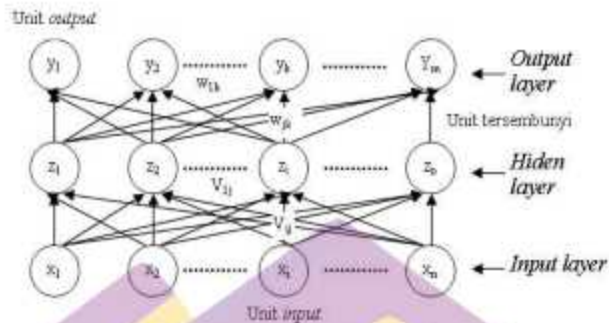
Tiap-tiap unit tersembunyi (z_j , $j=1,2,3,\dots,p$) memperbaiki bias dan bobotnya ($i=0,1,2,3,\dots,n$)

$$v_{ij}(\text{baru}) = v_{ij}(\text{lama}) + \Delta v_{ij} \dots\dots\dots (22)$$

Proses akan berulang kembali ke awal (*forward*) hingga diperoleh nilai kesalahan lebih rendah dari ambang batas maksimal kesalahan yang sudah ditentukan sebelumnya.

2.3.8 Arsitektur Backpropagation

Backpropagation memiliki beberapa unit yang ada dalam satu atau lebih layer tersembunyi. Gambar 2.1 adalah arsitektur backpropagation dengan n buah masukan (ditambah sebuah bias), sebuah lapisan tersembunyi yang terdiri dari p unit (ditambah sebuah bias), serta m buah unit keluaran.



Gambar 2. 2. Arsitektur Backpropagation

Nilai w_{jk} merupakan bobot pengali dari unit masukan x_{ji} ke unit lapisan tersembunyi. Sedangkan w_{jk} merupakan bobot pengali dari unit lapisan tersembunyi ke unit keluaran y_{jk} (Siang, JJ,2005). Pada tahap pelatihan, mula-mula nilai bobot ditentukan secara asumsi untuk menghitung nilai keluaran menggunakan perambatan maju (*feed forward propagation*) berdasarkan input yang diterima, kemudian dihitung nilai kesalahannya terhadap target. Selanjutnya berdasarkan nilai galat, dilakukan perambatan mundur (*back propagation*) untuk menghitung ulang nilai seluruh bobot dan proses perambatan maju akan berulang kembali. Sehingga pada tahapan pelatihan akan dilakukan proses *feed forward – back propagation* secara berulang (iteratif) hingga diperoleh galat minimal seperti yang diinginkan.

Pelatihan pada jaringan syaraf *backpropagation, feed-foward* (umpan maju) dilakukan dalam rangka perhitungan bobot sehingga pada akhir pelatihan akan diperoleh bobot-bobot yang baik. Selama proses pelatihan, bobot-bobot diatur secara iteratif untuk meminimumkan error (kesalahan) yang terjadi. Error (kesalahan) dihitung berdasarkan rata-rata kuadrat kesalahan (*mean square error* -

MSE). Rata-rata kuadrat kesalahan juga dijadikan dasar perhitungan unjuk kerja fungsi aktivasi. Sebagian besar pelatihan untuk jaringan feed-forward (umpan maju) menggunakan gradien dari fungsi aktivasi untuk menentukan bagaimana mengatur bobot-bobot dalam rangka meminimumkan kinerja. Gradien ini ditentukan dengan menggunakan suatu teknik yang disebut backpropagation.

Algoritma pelatihan standar backpropagation akan menggerakkan bobot dengan arah gradien negatif. Prinsip dasar dari algoritma backpropagation adalah memperbaiki bobot-bobot jaringan dengan arah yang membuat fungsi aktivasi menjadi turun dengan cepat. Pelatihan backpropagation meliputi 3 tahapan sebagai berikut:

1. Propagasi maju.

Pola masukan dihitung maju mulai dari input layer hingga output layer menggunakan fungsi aktivasi yang ditentukan.

2. Propagasi mundur.

Selisih antara keluaran jaringan dengan target yang diinginkan merupakan kesalahan yang terjadi. Kesalahan yang terjadi itu dipropagasi mundur. Dimulai dari garis yang berhubungan langsung dengan unit-unit di output layer.

3. Perubahan bobot.

Modifikasi bobot untuk menurunkan kesalahan yang terjadi. Ketiga fase tersebut diulang-ulang terus hingga kondisi penghentian dipenuhi.

BAB III

METODE PENELITIAN

3.1. Jenis, Sifat, dan Pendekatan Penelitian

Penelitian yang akan dilakukan ini termasuk ke dalam jenis penelitian kuantitatif. Penelitian ini melakukan perhitungan matematis terhadap data-data penelitian. Sifat dari penelitian ini adalah penelitian eksperimental. Penelitian ini melakukan eksperimen untuk melakukan klasifikasi kepribadian ke dalam kategori kepribadian *Big Five* menggunakan algoritma *Backpropagation Neural Network* dan *Naive Bayes*.

3.2. Metode Pengumpulan Data

Data yang dimaksud dalam penelitian ini adalah dataset yang diperoleh dari My Personality. Dataset berisi data ID user, status dan label kepribadian pengguna. Penulis memanfaatkan dataset yang telah tersedia dari My Personality. Label kepribadian dalam dataset tersebut adalah label dalam kategori kepribadian *Big Five*.

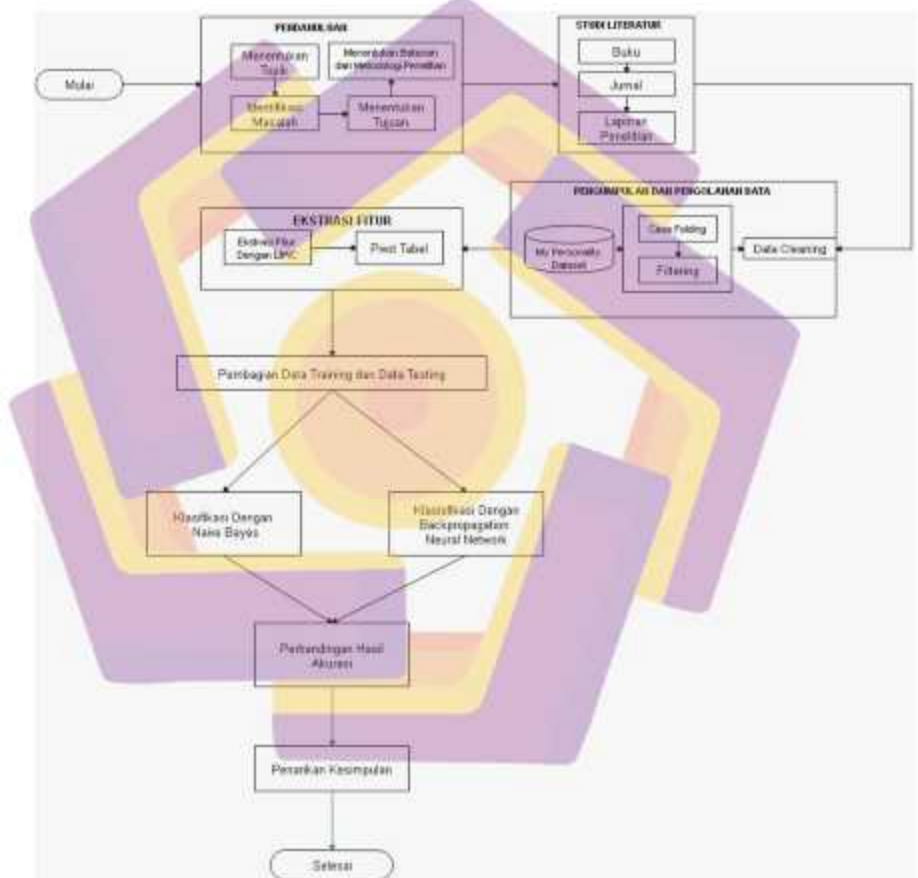
3.3. Metode Analisis Data

Setelah proses pengumpulan data, selanjutnya data yang terkumpul akan dianalisis. Metode analisis yang digunakan dalam penelitian ini adalah metode *Backpropagation Neural Network* dan *Naive Bayes* yang bertujuan untuk mengklasifikasikan kepribadian pengguna ke dalam kerangka *Big Five*. Analisa data dilakukan dengan menggunakan software Rapidminer. Selain menganalisa

hasil klasifikasi, penulis juga menganalisis kinerja akurasi dari masing-masing algoritma dengan bantuan software Rapidminer.

3.4. Alur Penelitian

Berikut adalah gambaran alur penelitian yang akan dilakukan.



Gambar 1. Alur Penelitian

3.4.1 Pendahuluan

Langkah yang dilakukan pada tahapan pendahuluan meliputi kegiatan penentuan topik, identifikasi masalah, menentukan tujuan, menentukan batasan dan metode penelitian.

3.4.2 Studi Literatur

Studi literatur dilakukan dengan mencari penelitian serupa sebelumnya untuk dipelajari. Sumber yang dipelajari pada proses studi literatur ini bersumber dari buku, jurnal dan publikasi penelitian-penelitian sebelumnya.

3.4.3 Pengumpulan dan Pengolahan Data

Data yang digunakan dalam penelitian ini merupakan dataset yang bersumber dari MyPersonality. Dataset yang diperoleh adalah sebanyak 9917 status Facebook dari 250 pengguna Facebook. Dalam dataset tersebut berisi ID pengguna, status facebook dan sudah terdapat label kepribadian dalam kelompok *Big Five*. Dataset yang diperoleh berbentuk file .csv.

Sebelum dilakukan pengolahan lebih lanjut, data melalui *text pre-processing*. Tahapan yang penulis terapkan antara lain:

- a. *Case folding* : Teks dari status yang terdapat dalam dataset tidak semuanya konsisten dalam penggunaan huruf kapital. Sehingga perlu dilakukan tahapan *case folding* ini untuk menyeragamkan semua kata ke dalam bentuk huruf kecil (*lowercase*). Contohnya dalam kalimat "SAYA TDK SUKA mendengar suara yang mengagetkan saat Saya sedang tidur" akan berubah menjadi "saya tdk suka mendengar suara yang mengagetkan saat saya sedang tidur" setelah proses *case folding*.

b. *Filtering* : Proses filtering pada umumnya menghilangkan kata yang tidak penting dan menyimpan kata penting. Pada umumnya proses ini menghilangkan tanda baca karena dianggap tidak penting. Namun dalam penelitian ini, tanda baca tidak dihilangkan karena termasuk dalam salah satu fitur untuk proses klasifikasinya.

Proses yang dilakukan selanjutnya adalah data cleaning, dalam proses ini data yang tidak dapat digunakan untuk proses klasifikasi akan dihilangkan. Contoh data yang tidak dapat digunakan antara lain data yang tidak memiliki user ID, tidak memiliki status atau tidak memiliki label.

3.4.4 Ekstraksi Fitur

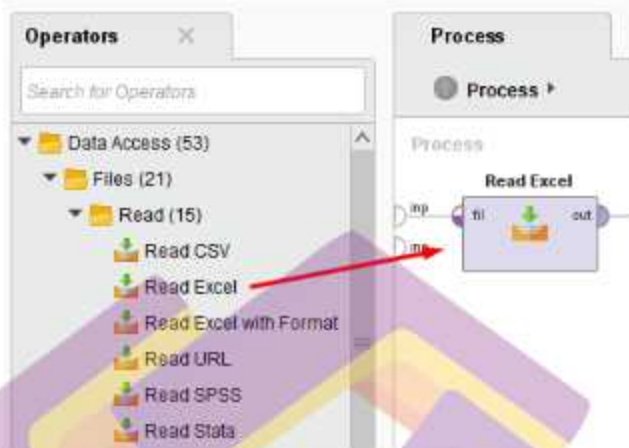
Ekstraksi fitur dalam penelitian ini memanfaatkan LIWC (*Linguistic Inquiry and Word Count*).

3.4.5 Klasifikasi

Setelah dilakukan ekstraksi fitur dan pembobotan, kemudian dilakukan proses klasifikasi. Proses klasifikasi yang dilakukan dalam penelitian ini dilakukan menggunakan dua algoritma yaitu *Backpropagation Neural Network* dan *Naive Bayes*. Proses Klasifikasi dilakukan menggunakan *tool* bernama Rapidminer. Secara umum tahapan proses klasifikasi akan melalui beberapa proses berikut.

3.4.5.1. Read Excel

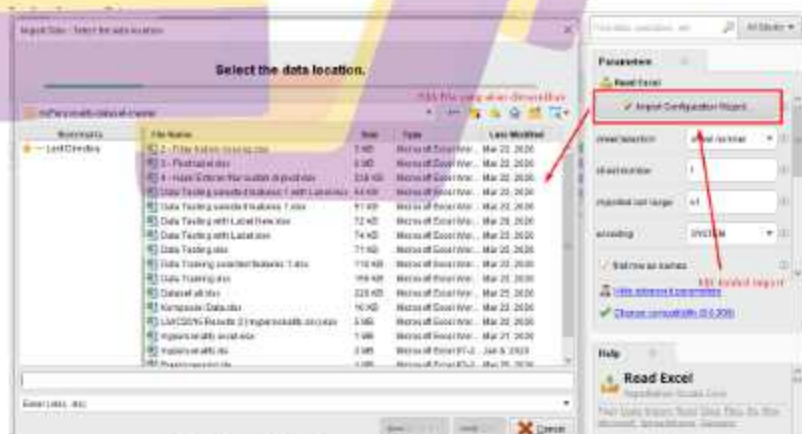
Digunakan untuk memanggil data file yang memiliki ekstensi *xlsx*, *tool* ini dapat mempermudah memanggil data.



Gambar 3. 1. Operator Read Excel

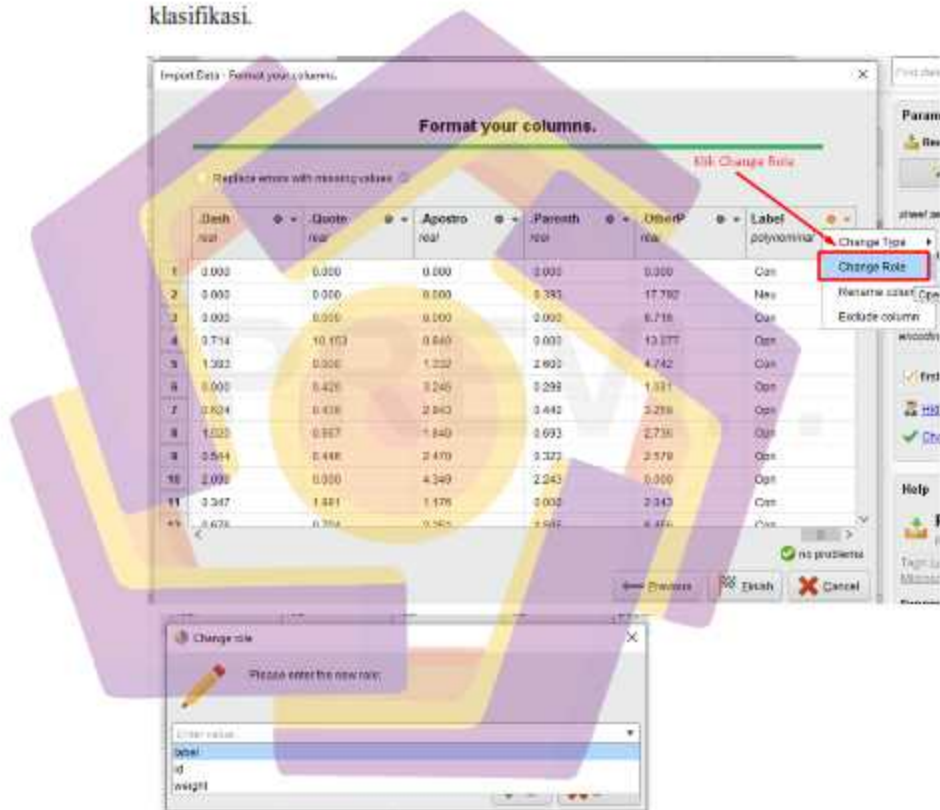
Karena dataset yang penulis miliki telah disimpan dalam format .xlsx, maka dengan operator read excel ini dataset dapat dimasukkan. Terdapat beberapa langkah yang perlu dilakukan agar dapat memasukkan data untuk diolah dan data tersebut dapat digunakan:

- a. Memilih file yang akan dimasukkan melalui tombol import



Gambar 3. 2. Import file

- b. Setting atribut label pada dataset. Proses ini dilakukan setelah memilih file, lalu klik tombol *next*. Pada bagian header data, klik icon setting, lalu pilih “Change Role”. Pada dataset yang digunakan dalam penelitian ini, kolom label akan diset sebagai label dalam proses klasifikasi.

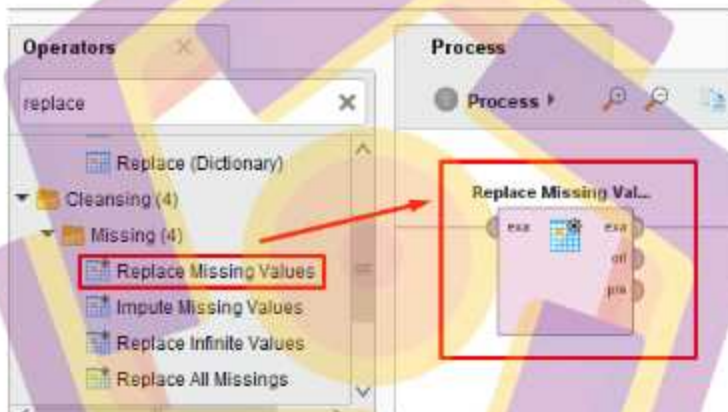


Gambar 3. 3. Set role

3.4.5.2. Replace Missing Value

Operator Replace Missing Value ini berfungsi untuk menggantikan nilai yang hilang/kosong dalam atribut. Dengan operator ini, data yang hilang/kosong

dapat digantikan oleh nilai baru. Nilai dapat berupa nilai maksimum, minimum, rata-rata atau null sesuai dengan yang ditetapkan. Dalam penelitian ini penulis telah membandingkan dengan menerapkan replace missing value dengan nilai maksimum, nilai minimum, rata-rata dan null. Namun hasil akurasi yang diperoleh tidak terdapat perbedaan. Menurut analisa penulis, hal tersebut dikarenakan sebelum memasuki proses klasifikasi, penulis telah menyaring data dan hanya menggunakan data yang memiliki nilai dalam atributnya.



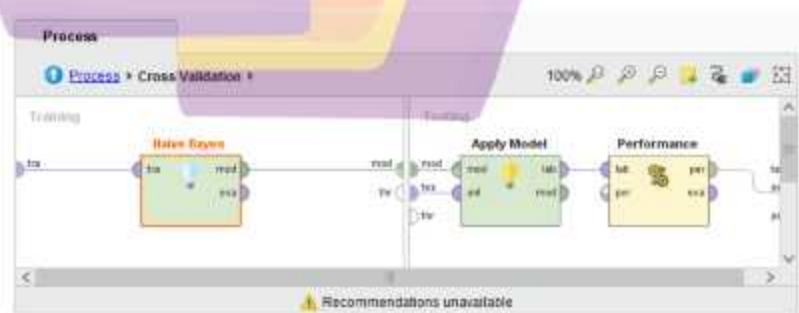
Gambar 3. 4. Operator Replace Missing Value

3.4.5.3. Cross Validation



Gambar 3. 5. Operator Cross Validation

Cross validation digunakan untuk memperkirakan seberapa akurat suatu model (dipelajari oleh Operator pembelajaran tertentu) dan akan ditampilkan dalam praktik Cross validation adalah Operator yang bersarang, memiliki dua subproses: subproses Pelatihan dan subproses Pengujian. Subproses pelatihan digunakan untuk melatih model. Model yang terlatih kemudian diterapkan dalam subproses Pengujian. Kinerja model diukur selama fase Pengujian, berikut adalah tampilan dari tahapan cross validation:



Gambar 3. 6. Detail proses dari Cross Validation

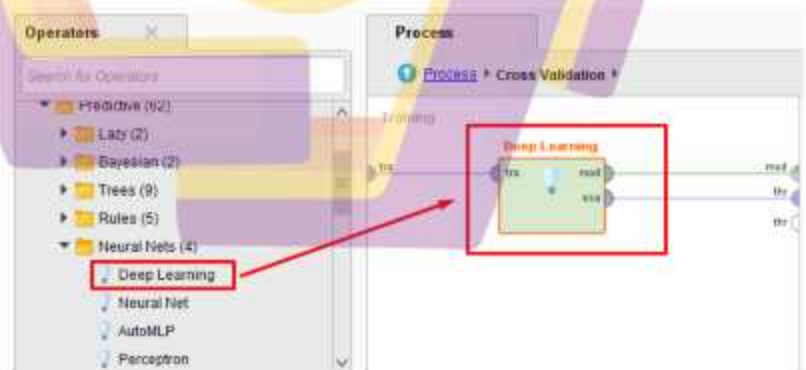
3.4.5.4. Naive Bayes Classification



Gambar 3. 7. Klasifikasi dengan Algoritma *Naive Bayes*

3.4.5.5. Backpropagation Neural Network

Proses klasifikasi dengan algoritma *Backproagation Neural Network* dalam algoritma dilakukan dengan menggunakan operator deep learning neural network, biasa disebut algoritma H2O. Sistem kerjanya adalah meminimalkan eror dari hasil iterasi neural network.



Gambar 3. 8. Klasifikasi dengan Algoritma *Backpropagation Neural Network*

3.4.5.6. Apply Model

Apply Model digunakan untuk menghubungkan model algoritma dengan *performance*, *apply model* dapat juga berperan sebagai penghubung. Pada operator *apply model* inilah data *testing* yang *unlabeled* kemudian akan diberi label berdasarkan pemodelan dari algoritma pada proses trainingnya.



Gambar 3. 9. Operator Apply Model

3.4.5.7. Performance

Operator ini digunakan untuk secara statistik mengevaluasi kekuatan dan kelemahan klasifikasi biner, setelah model yang terlatih telah diterapkan pada data berlabel.



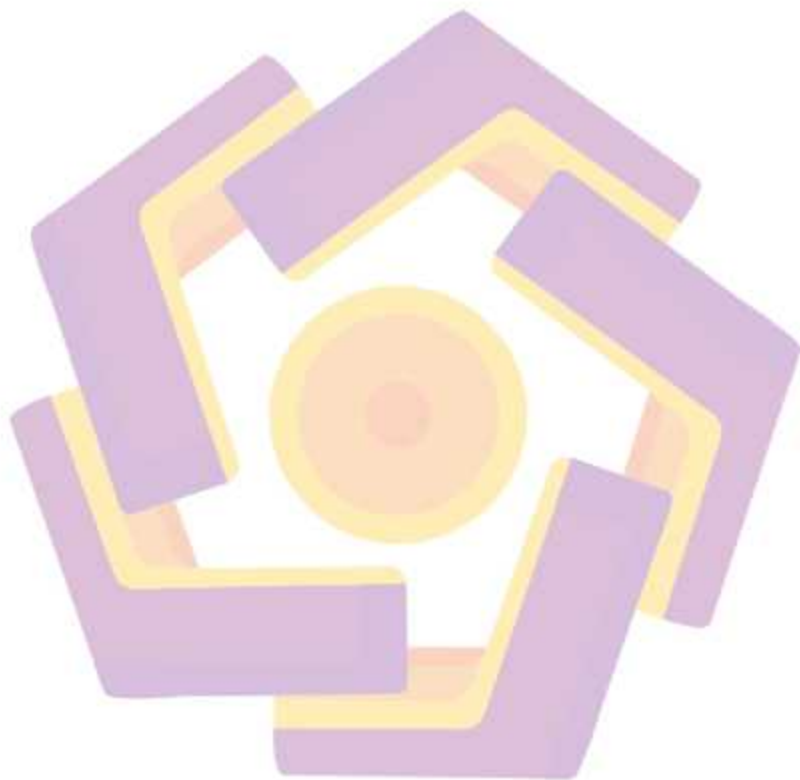
Gambar 3. 10. Operator Performance

3.4.6 Perbandingan Hasil Akurasi

Setelah proses klasifikasi dilakukan, dapat diketahui hasil akurasi masing-masing algoritma dengan cara membandingkan label sebenarnya dengan hasil prediksi. Hasil akurasi dapat diproses langsung dengan tool rapidminer dan dapat juga dilakukan manual dengan cara membandingkannya secara manual.

3.4.7 Penarikan Kesimpulan

Tahap akhir dari penelitian ini adalah penarikan kesimpulan. Tahap ini menyimpulkan hasil penelitian dengan berpatokan pada rumusan masalah yang telah dibuat sebelumnya.



BAB IV

HASIL PENELITIAN DAN PEMBAHASAN

4.1. Pengumpulan dan Pengolahan Data

Data yang digunakan dalam penelitian ini adalah dataset yang diperoleh dari My Personality yang berisi data status Facebook pengguna dan sudah terdapat label hasil klasifikasi dalam kepribadian *Big Five*. Jumlah data yang ada adalah sebanyak 9917 status Facebook dari 250 pengguna Facebook. Dari sejumlah dataset yang diperoleh tersebut tidak semuanya dapat langsung digunakan. Oleh karena itu perlu dilakukan beberapa tahap sebelum proses klasifikasi atau disebut juga *pre-processing text* yang antara lain:

4.1.1 *Pre-processing Text*

Banyak tahap yang dapat dilakukan pada tahap *pre-processing text* dalam *text mining*. Tahapan yang umum dilakukan dalam *pre-processing text* antara lain case folding, tokenizing, filtering, stemming. Dalam penelitian ini, tahapan yang penulis lakukan antara lain:

a. *Case Folding*

Tahapan ini dilakukan dengan mengubah huruf besar yang terdapat dalam status menjadi huruf kecil. Contoh dari proses ini dapat dilihat pada tabel.

Tabel 4 . 1. Contoh Proses Case Folding

Status Awal	saw Transformers, Up, and Year One this week. Good movie overload.
Diubah menjadi	saw transformers, up, and year one this week. good movie overload.

b. Filtering

Dalam penelitian ini tanda baca merupakan salah satu bagian dari fitur, sehingga proses *filtering* yang penulis lakukan tidak menghilangkan tanda baca namun menghilangkan emoticon.

4.1.2 Data Cleaning

Pada kenyataannya dari dataset yang diperoleh tidak semuanya dapat digunakan. Terdapat data yang tidak lengkap, tidak memiliki label atau tidak memiliki ID. Jika data tersebut dibiarkan ada, akan mengganggu proses klasifikasi kedepannya. Oleh karena itu setelah teks status dari dataset diolah, data dipilah lagi dan dibersihkan dalam proses ini. Hasil dari proses ini data yang tersisa adalah data yang normal dan layak untuk masuk ke tahap selanjutnya. Setelah dilakukan proses *cleaning*, dari 9917 data status tersisa 9740 data status untuk dapat diproses lebih lanjut.

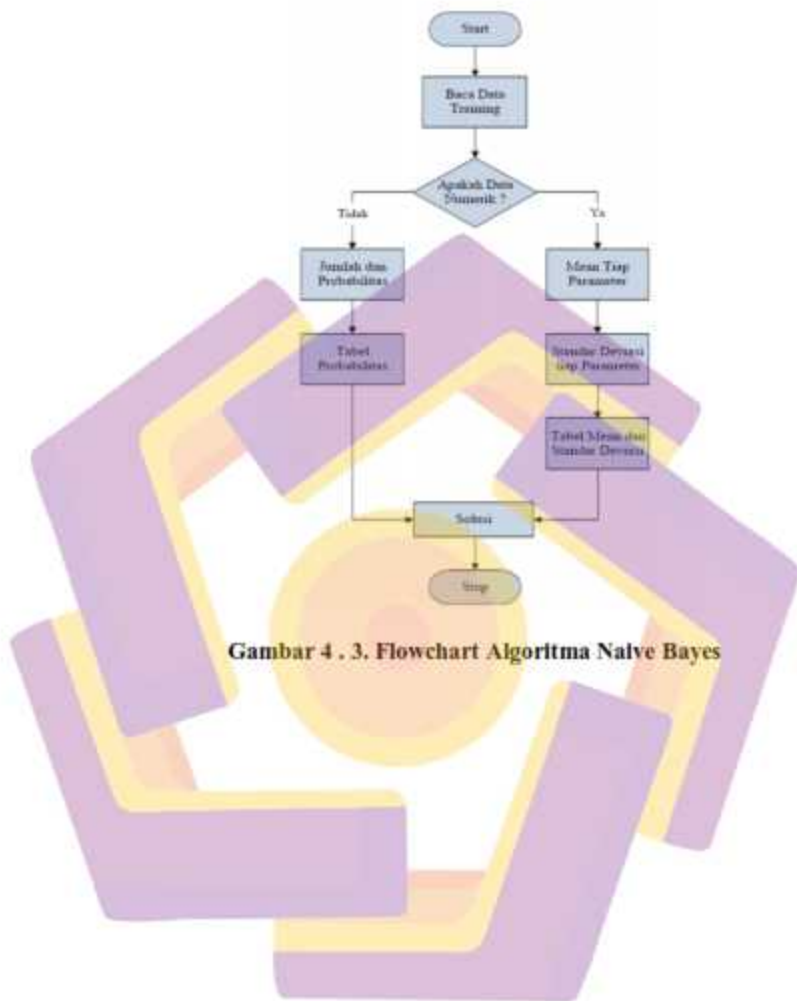
4.2. Ekstraksi fitur

Beberapa penelitian sebelumnya melakukan ekstraksi fitur linguistik dari status facebook pengguna dengan menggunakan pendekatan *bag-of-words* lalu pembobotan dilakukan dengan menggunakan TF IDF. Namun dalam penelitian ini, penulis menggunakan LIWC (*Linguistic Inquiry and Word Count*). Alat ini telah banyak digunakan untuk penelitian serupa terutama yang berhubungan dengan analisa linguistik dari teks. Jika dengan menggunakan pendekatan *bag of words* perlu dilakukan *tokenizing* dan *stemming*, dalam penelitian ini tidak dilakukan *tokenizing* dan *stemming* terlebih dahulu karena proses tersebut sudah dilakukan oleh LIWC dalam mengkategorikan kata dari kalimat yang diinputkan

berdasarkan analisis linguistik. Hasil dari proses ekstraksi fitur ini adalah diperoleh 93 fitur. Fitur linguistik yang akan digunakan antara lain dapat dilihat pada tabel.

Tabel 4 . 2. Fitur

<i>Word count</i>	<i>Common adverbs</i>	<i>Negative emotion</i>	<i>Cause</i>	<i>Sexuality</i>
<i>Analytical thinking</i>	<i>Conjunctions</i>	<i>Anxiety</i>	<i>Discrepancies</i>	<i>Ingesting</i>
<i>Clout</i>	<i>Negations</i>	<i>Anger</i>	<i>Tentativeness</i>	<i>Core Drives and Needs</i>
<i>Authentic</i>	<i>Regular verbs</i>	<i>Sadness</i>	<i>Certainty</i>	<i>Affiliation</i>
<i>Emotional tone</i>	<i>Adjectives</i>	<i>Social Words</i>	<i>Differentiation</i>	<i>Achievement</i>
<i>Words per sentence</i>	<i>Comparatives</i>	<i>Family</i>	<i>Seeing</i>	<i>Power</i>
<i>Words > 6 letters</i>	<i>interrogatives</i>	<i>Friends</i>	<i>Hearing</i>	<i>Reward focus</i>
<i>Dictionary words</i>	<i>Number</i>	<i>Female referents</i>	<i>Feeling</i>	<i>Risk/prevention focus</i>
<i>Function words</i>	<i>Quantifiers</i>	<i>Male referents</i>	<i>Biological Processes</i>	<i>Past focus</i>
<i>Total pronouns</i>	<i>Affect Words</i>	<i>Cognitive Processes</i>	<i>Body</i>	<i>Present focus</i>
<i>1st pers singular</i>	<i>Positive emotion</i>	<i>Insight</i>	<i>Health/illness</i>	<i>Future focus</i>
<i>1st pers plural</i>	<i>Relativity</i>	<i>Death</i>	<i>Commas</i>	<i>Other punctuation</i>
<i>2nd person</i>	<i>Motion</i>	<i>Informal Speech</i>	<i>Colons</i>	<i>Auxiliary verbs</i>
<i>3rd person</i>	<i>Space</i>	<i>Swear words</i>	<i>Semicolons</i>	<i>Religion</i>
<i>3rd pers singular</i>	<i>Time</i>	<i>Netspeak</i>	<i>Question marks</i>	<i>Periods</i>
<i>3rd pers plural</i>	<i>Work</i>	<i>Assent</i>	<i>Exclamation marks</i>	<i>Parentheses (pairs)</i>
<i>Impersonal pronouns</i>	<i>Leisure</i>	<i>Nonfluencies</i>	<i>Dashes</i>	
<i>Articles</i>	<i>Home</i>	<i>Fillers</i>	<i>Quotation marks</i>	
<i>Prepositions</i>	<i>Money</i>	<i>All Punctuation</i>	<i>Apostrophes</i>	



Gambar 4 . 3. Flowchart Algoritma Nalve Bayes

Tabel 4 . 3. Data Training Perhitungan Manual Naive Bayes

ID	Word Count	Positive Emotion	Negative Emotion	compare	interrogative	number	anxiety	anger	sad	Label
1	4,11	6,67	0,00	0,00	4,44	0,00	0,00	0,00	0,00	Agr
2	18,18	7,60	4,89	0,00	0,91	2,49	0,00	1,29	0,76	Agr
3	27,40	3,16	3,50	1,38	0,69	10,91	0,00	0,69	0,00	Con
4	21,17	5,66	3,80	0,41	0,21	1,88	0,24	0,23	0,00	Con
5	10,51	5,41	6,14	2,84	0,76	1,90	0,60	1,96	0,77	Ext
6	6,50	0,00	0,00	0,00	0,00	14,34	0,00	0,00	0,00	Ncu
7	13,27	5,15	1,65	1,66	1,80	2,05	0,10	0,29	0,35	Opn
8	15,74	6,45	2,15	2,40	0,84	1,44	0,08	0,98	0,76	Opn
9	21,43	1,50	1,59	1,81	3,17	2,01	0,00	1,59	0,00	Opn
10	18,36	2,93	1,46	0,21	1,10	0,00	0,00	0,00	0,21	Opn
11	13,27	5,15	1,65	1,66	1,80	2,05	0,10	0,29	0,35	Opn
12	11,84	7,84	1,66	0,75	0,73	3,33	0,13	1,03	0,09	Opn
13	15,74	6,45	2,15	2,40	0,84	1,44	0,08	0,98	0,76	Opn
14	22,20	4,05	5,48	0,57	1,05	2,38	0,07	0,56	2,95	Agr
15	12,78	8,12	6,05	1,41	2,60	4,04	0,27	0,56	1,10	Agr
16	13,00	15,38	0,00	0,00	0,00	0,00	0,00	0,00	0,00	Agr
17	16,25	6,05	2,80	2,88	1,70	2,59	0,00	0,35	0,78	Con
18	13,52	4,70	3,57	0,52	1,46	2,93	1,24	0,88	0,74	Con
19	13,00	5,89	11,39	3,32	0,32	2,50	1,75	0,32	2,50	Ext
20	13,21	7,07	3,84	0,88	0,65	0,82	0,41	0,27	0,29	Ext

Tabel 4. 4. (Lanjutan)

ID	Word Count	Positive Emotion	Negative Emotion	compare	interrogative	number	anxiety	anger	sad	Label
21	15,29	1,13	1,91	0,00	0,00	0,68	0,00	0,00	1,47	Ext
22	4,00	0,00	0,00	25,00	0,00	0,00	0,00	0,00	0,00	Neu
23	23,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	Neu

Diketahui data testing yang akan dicari labelnya adalah sebagai berikut.

Tabel 4. 5. Data Testing Perhitungan Manual Naive Bayes

ID	Word Count	Positive Emotion	Negative Emotion	compare	interrogative	number	anxiety	anger	sad	Label
24	13,49	4,27	6,52	1,71	2,15	0,70	0,40	4,68	0,83	?

Langkah perhitungan menggunakan algoritma Naive Bayes adalah sebagai berikut:

1. Menghitung probabilitas kelas
Jumlah Kelas Openness : 7
Jumlah Kelas Agreeableness : 5
Jumlah Kelas Conscientiousness : 4

Jumlah Kelas Extroversion : 4

Jumlah Kelas Neuroticism : 3

P Kepribadian Openness : $7/23 = 0,304$

P Kepribadian Agreeableness : $5/23 = 0,217$

P Kepribadian Conscientiousness : $4/23 = 0,174$

P Kepribadian Extroversion : $4/23 = 0,174$

P Kepribadian Neuroticism : $3/23 = 0,130$

2. Menghitung jumlah kasus yang sama pada kelas yang sama

Data pada tiap atribut dalam penelitian ini merupakan data numerik, sehingga pada tahapan ini dilakukan perhitungan untuk mencari *mean*, standar deviasi, kemudian menghitung probabilitas untuk tiap-tiap atribut pada masing-masing kelas.

Perhitungan Mean (μ) dilakukan menggunakan rumus berikut:

$$\mu = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

Perhitungan atribut word count

$$\begin{aligned} \mu \text{ kepribadian opn} &= \frac{13,27 + 15,74 + 21,43 + 18,36 + 13,27 + 11,84 + 15,74}{7} \\ &= 15,67 \end{aligned}$$

$$\begin{aligned} \mu \text{ kepribadian agr} &= \frac{4,11 + 18,18 + 22,20 + 12,78 + 13,00}{5} \\ &= 14,06 \end{aligned}$$

$$\begin{aligned} \mu \text{ kepribadian con} &= \frac{27,40 + 21,17 + 16,25 + 13,52}{4} \\ &= 19,59 \end{aligned}$$

$$\mu \text{ kepribadian ext} = \frac{10,51 + 13,00 + 13,21 + 15,29}{4}$$

$$= 13,00$$

$$\mu \text{ kepribadian neu} = \frac{6,50 + 4,00 + 23,00}{3}$$

$$= 11,17$$

Perhitungan dilakukan dengan langkah yang sama untuk atribut lain selain word count. Langkah selanjutnya adalah menghitung standar deviasi menggunakan rumus berikut.

$$\sigma = \left[\frac{1}{n-1} \sum_{i=1}^n (x_i - \mu)^2 \right]^{1/2}$$

Setelah itu hitung probabilitas masing-masing atribut pada masing-masing kelas dengan rumus berikut.

$$P(X | \text{Kepribadian } \text{opn}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(X - \mu)^2}{2\sigma^2}\right)$$

Dengan nilai X adalah nilai atribut dari data testing. Langkah-langkah perhitungan ini dilakukan juga untuk atribut lain sehingga diperoleh hasilnya seperti pada tabel berikut.

a. Atribut word count

Kelas Kepribadian	Mean (μ)	Standar Deviasi (σ)	Probabilitas (P)
Opn	15,67	11,10	2,02255E-07
Agr	14,06	46,25	1,65842E-08
Con	19,59	37,14	6,89247E-52
Ext	13,00	3,82	0,263728617
Neu	11,17	106,58	1,79968E+51

b. Atribut positive emotion

Kelas Kepribadian	Mean (μ)	Standar Deviasi (σ)	Probabilitas (P)
Opn	5,07	4,81	0,012068646
Agr	8,36	17,84	3,01977E-18
Con	4,89	1,66	0,143225488
Ext	4,87	6,72	0,007795489
Neu	0,00	0,00	0

c. Atribut negative emotion

Kelas Kepribadian	Mean (μ)	Standar Deviasi (σ)	Probabilitas (P)
Opn	1,76	0,08	6,261822854
Agr	3,28	9,16	120329,4408
Con	3,42	0,19	2,858797544
Ext	5,82	16,79	8,78082039
Neu	0,00	0,00	0

d. Atribut compare

Kelas Kepribadian	Mean (μ)	Standar Deviasi (σ)	Probabilitas (P)
Opn	1,55	0,66	0,636515913
Agr	0,40	0,38	1,337383228
Con	1,30	1,30	0,402763975
Ext	1,76	2,49	0,150991406
Neu	8,33	208,33	4,8633E-303

e. Atribut interrogative

Kelas Kepribadian	Mean (μ)	Standar Deviasi (σ)	Probabilitas (P)
Opn	1,47	0,77	0,676704449
Agr	1,80	3,06	0,223970143
Con	1,01	0,47	1,102716113
Ext	0,43	0,12	3,755321612
Neu	0,00	0,00	0

f. Atribut number

Kelas Kepribadian	Mean (μ)	Standar Deviasi (σ)	Probabilitas (P)
Opn	1,76	1,00	0,234034287
Agr	1,78	3,07	0,024586569
Con	4,58	18,04	1,39836E-17
Ext	1,48	0,76	0,389178342
Neu	4,78	68,55	9,8933E-64

g. Atribut anxiety

Kelas Kepribadian	Mean (μ)	Standar Deviasi (σ)	Probabilitas (P)
Opn	0,07	0,00	148,7518639
Agr	0,07	0,01	29,18407235
Con	0,37	0,35	1,14080966
Ext	0,69	0,56	0,653522581
Neu	0,00	0,00	0

h. Atribut anger

Kelas Kepribadian	Mean (μ)	Standar Deviasi (σ)	Probabilitas (P)
Opn	0,74	0,31	2,372038338
Agr	0,48	0,28	2,553583875
Con	0,54	0,09	5,28124356
Ext	0,64	0,79	2,499303199
Neu	0,00	0,00	0

i. Atribut sad

Kelas Kepribadian	Mean (μ)	Standar Deviasi (σ)	Probabilitas (P)
Opn	0,36	0,09	4,487285808
Agr	0,96	1,47	0,247950465
Con	0,38	0,19	2,168814691
Ext	1,26	0,92	0,356979754
Neu	0,00	0,00	0

3. Mengalikan semua hasil berdasarkan data testing

$$P(X | Kepribadian Opn) = 2,43957E-06$$

$$P(X | Kepribadian Agr) = 8,2006E-22$$

$$P(X | Kepribadian Con) = 2,29026E-68$$

$$P(X | Kepribadian Ext) = 0,002322773$$

$$P(X | Kepribadian Neu) = 0$$

$$P(\text{Kepribadian Opn} | X) = 7,42478E-07$$

$$P(\text{Kepribadian Agr} | X) = 1,78274E-22$$

$$P(\text{Kepribadian Con} | X) = 3,98306E-69$$

$$P(\text{Kepribadian Ext} | X) = 0,000403961$$

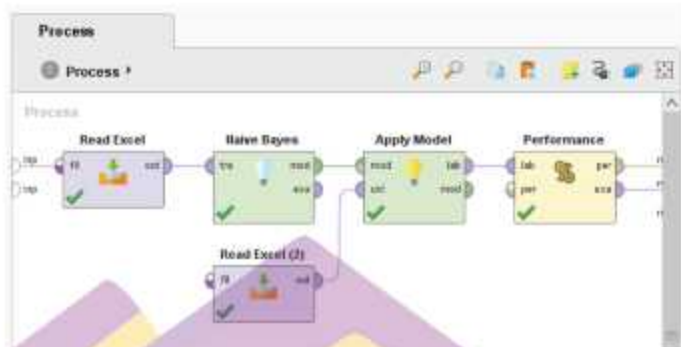
$$P(\text{Kepribadian Neu} | X) = 0$$

Berdasarkan hasil perhitungan diatas, data testing pada contoh kasus ini termasuk ke dalam kelas kepribadian Opn (Openness).

4.3.1.2. Klasifikasi dengan Algoritma *Naive Bayes* Menggunakan Rapidminer

Perhitungan klasifikasi menggunakan rapidminer ini dilakukan untuk memperoleh hasil perhitungan yang sebenarnya dari semua dataset yang dimiliki yang tidak mungkin dilakukan perhitungannya dengan cara manual. Penulis membuat 2 skenario untuk memproses perhitungan dengan algoritma *Naive Bayes*.

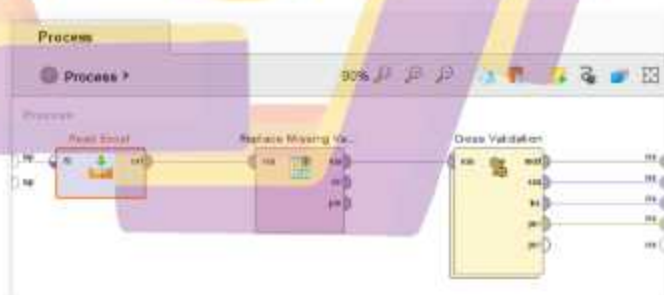
- c. Skenario 1 (Membuat model dengan 175 data sebagai *data training*, kemudian mengujinya dengan 75 data lainnya sebagai *data testing*). Pada skenario ini diperoleh hasil akurasi sebesar 30,67%. Prosesnya dapat dilihat pada gambar 4.3.



Gambar 4 . 4. Skenario 1 Proses Klasifikasi dengan Algoritma *Naive Bayes*

d. Skenario 2 (Membuat model dengan 250 data yang dimiliki kemudian menguji pemodelan dengan jumlah data yang sama).

Pada skenario ini akurasi otomatis dihitung pada proses "Performance" dalam rapidminer. Hasilnya justru mengalami penurunan akurasi dengan akurasi sebesar 29,20%. Alur prosesnya dapat dilihat pada gambar 4.4.



Gambar 4 . 5. Skenario 2 Proses Klasifikasi dengan Algoritma *Naive Bayes*

Proses pemodelan dan penerapan algoritmanya terdapat di dalam proses Cross Validation yang alur prosesnya dapat dilihat pada gambar 4.6.



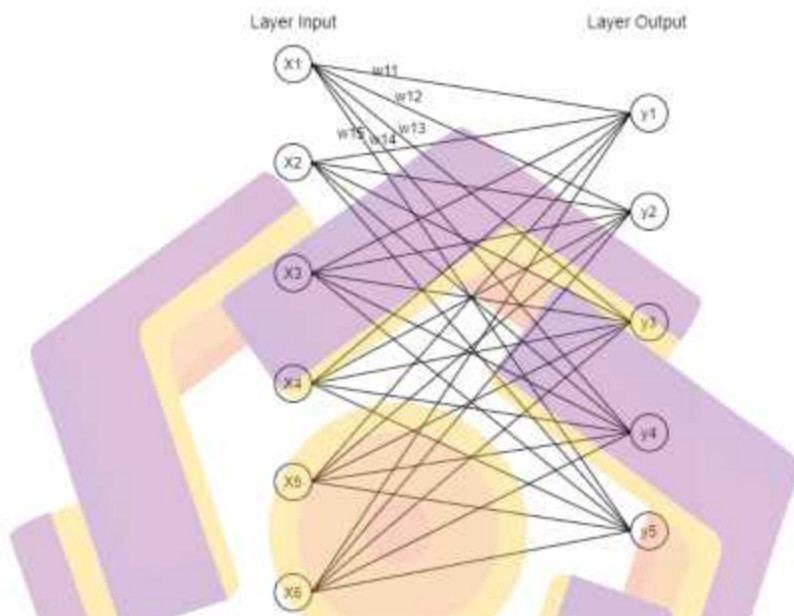
Gambar 4 . 6. Proses *Naive Bayes* di dalam Proses *Cross Validation*

4.3.2 Algoritma *Backpropagation Neural Network*

4.3.2.1. Contoh Perhitungan Klasifikasi Manual dengan Algoritma *Backpropagation Neural Network*

Perhitungan klasifikasi manual dengan algoritma *Backpropagation Neural Network* ini hanya sebagai contoh langkah yang menggambarkan proses bagaimana alur perhitungan algoritma ini. Data yang digunakan hanya sample kecil dari dataset yang dimiliki. Fitur yang digunakan dalam penelitian ini adalah sebanyak 93 fitur yang diekstraksi dari LIWC. Untuk menjelaskan alur dari algoritma *backpropagation neural network* ini penulis hanya akan menggunakan contoh 6 fitur saja. 6 fitur tersebut akan menjadi neuron dalam layer input dan layer output adalah sebanyak 5 neuron yang diperoleh dari label kepribadian.

Model/arsitektur sederhana dari contoh perhitungan ini dapat dilihat pada gambar.



Gambar 4 . 7. Model / Arsitektur Sederhana Neural Network dari Perhitungan Manual

Dari model/arsitektur diatas, diketahui layer input dan layer output adalah seperti pada tabel 4.3 dan tabel 4.4.

Tabel 4 . 6. Layer Input

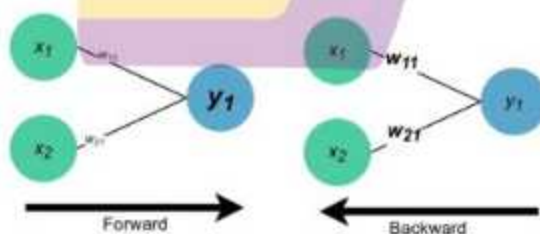
Fitur		Nilai
Simbol	Nama Fitur	
X1	Word Count	11,84
X2	Positive Emotion	7,84
X3	Negative Emotion	1,66
X4	Anxiety	0,13
X5	Anger	1,03
X6	Sad	0,09

Tabel 4 . 7. Layer Output

Kelas	Keprtbadian	Smbol
1	Openness	t1
2	Agreeableness	t2
3	Conciousness	t3
4	Extroversion	t4
5	Neuroticism	t5

Diketahui data yang akan dimodelkan ini memiliki label kepribadian openness yang disimbolkan dengan t1. Karena label data tersebut terdapat di kelas 1, maka nilai variabel t antara lain $t1=1$, $t2=0$, $t3=0$, $t4=0$, $t5=0$. Neural Network akan menghitung nilai y berdasarkan input x dan melalui proses pembelajaran, Neural network akan berusaha menghasilkan nilai y yang sama dengan nilai t.

Ada 2 proses yang akan dilakukan dalam neural network ini yaitu forward propagation dan backward propagation. Pada forward propagation, akan mencoba menghasilkan nilai y, sedangkan pada backward propagation, neural network akan memperbaiki dirinya (update bobot) sehingga pada forward propagation berikutnya diharap bisa menghasilkan nilai y yang lebih baik atau lebih mendekati label aslinya.



Gambar 4 . 8. Alur Forward dan Backward

Maka langkah yang dilakukan untuk perhitungannya adalah sebagai berikut.

a. Forward Propagation

Langkah 1

Menentukan nilai awal bobot (w) dan bias (b) dengan nilai awal ini ditentukan secara random.

	y_1	y_2	y_3	y_4	y_5
x_1	w_{11}	w_{12}	w_{13}	w_{14}	w_{15}
x_2	w_{21}	w_{22}	w_{23}	w_{24}	w_{25}
x_3	w_{31}	w_{32}	w_{33}	w_{34}	w_{35}
x_4	w_{41}	w_{42}	w_{43}	w_{44}	w_{45}
x_5	w_{51}	w_{52}	w_{53}	w_{54}	w_{55}
x_6	w_{61}	w_{62}	w_{63}	w_{64}	w_{65}

Matriks bobotnya adalah seperti berikut:

$$\begin{bmatrix} 0,01 & 0,02 & 0,03 & 0,04 & 0,05 \\ 0,02 & 0,01 & 0,03 & 0,05 & 0,04 \\ 0,03 & 0,02 & 0,04 & 0,01 & 0,05 \\ 0,04 & 0,03 & 0,05 & 0,02 & 0,01 \\ 0,05 & 0,04 & 0,01 & 0,03 & 0,02 \\ 0,01 & 0,05 & 0,02 & 0,04 & 0,03 \end{bmatrix}$$

Gambar 4 . 9. Matriks bobot

Matriks nilai bias (b) seperti berikut:

$$\{ 0,4 \quad 0,3 \quad 0,2 \quad 0,1 \quad 0 \}$$

Gambar 4 . 10. Matriks Nilai Bias

Langkah 2

Langkah selanjutnya adalah mencari nilai y . Rumus yang digunakan adalah

$$y_j = \sigma \left(\sum_{i=1}^6 w_{ij}x_i + b_j \right)$$

Nilai $\sigma(x)$ dihitung dengan rumus berikut dan nilai $e = 0,2718$.

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Cara perhitungannya adalah dengan cara berikut.

$$Y_1 = \sigma(w_{11}x_1 + w_{21}x_2 + w_{31}x_3 + w_{41}x_4 + w_{51}x_5 + w_{61}x_6 + b_1)$$

$$Y_1 = \sigma(0,01*11,84 + 0,02*7,84 + 0,03*1,66 + 0,04*0,13 + 0,05*1,03 + 0,01*0,09 + 0)$$

$$Y_1 = \sigma(0,78)$$

Hasil perhitungan semua nilai y dapat dilihat pada tabel 4.5.

Tabel 4 . 8. Hasil Perhitungan Nilai y

y1	y2	y3	y4	y5
0,27	0,29	0,24	0,21	0,21

Nilai y adalah hasil perhitungan dari neuron pada layer output.

Dalam contoh ini label outputnya adalah openness yang disimbolkan dengan t_j . Maka Nilai t adalah sebagai berikut.

Tabel 4 . 9. Tabel Nilai Target

t1	t2	t3	t4	t5
1	0	0	0	0

Langkah 3

Mencari nilai error. Penulis menghitung nilai error dengan rumus mean square error. Hasilnya dapat dilihat pada tabel.

Tabel 4 . 10. Nilai Error Masing-masing Neuron pada Layer Output

	y1	y2	y3	y4	y5
Error	0,05	-0,02	-0,02	-0,01	-0,01

b. Backward Propagation

Setelah memperoleh nilai error, neural network dapat diperbaiki dengan backpropagation.

Langkah 4

Memperbaiki bobot (w). Rumus yang digunakan adalah

$$w_{new} = w_{old} - \alpha \frac{\partial E}{\partial w}$$

nilai α dalam perhitungan ini menyatakan learning rate, penulis akan menggunakan $\alpha = 0,5$.

Hasil perhitungan bobot baru dapat dilihat pada matriks berikut:

$$\begin{bmatrix} 0,02 & 0,02 & 0,03 & 0,04 & 0,05 \\ 0,03 & 0,01 & 0,03 & 0,05 & 0,04 \\ 0,04 & 0,02 & 0,04 & 0,01 & 0,05 \\ 0,05 & 0,03 & 0,05 & 0,02 & 0,01 \\ 0,06 & 0,04 & 0,01 & 0,03 & 0,02 \\ 0,02 & 0,05 & 0,02 & 0,04 & 0,03 \end{bmatrix}$$

Langkah 5

Mencari nilai y baru dengan bobot yang telah diperbarui. Hasilnya dapat dilihat pada tabel berikut.

Tabel 4 . 11. Nilai y Baru Hasil Backward Propagation

y1	y2	y3	y4	y5
0,25	0,17	0,23	0,27	0,27

Langkah 6

Menghitung nilai error baru. Hasil perhitungan nilai errornya dapat dilihat pada tabel.

Tabel 4 . 12. Nilai Error Setelah Backprop

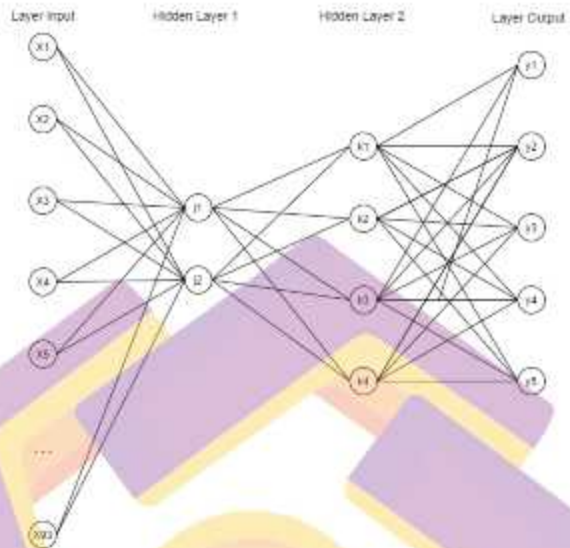
	v1	v2	v3	v4	v5
Error	0,05	-0,01	-0,01	-0,02	-0,02

Hasil perhitungan tersebut hanya sebagai contoh alur dari algoritma backpropagation neural network. Pada kenyataannya jumlah data yang dihitung berjumlah ribuan dengan jumlah neuron dalam layer input tidak hanya 5. Langkah ini juga baru merupakan iterasi pertama dan belum ditemukan nilai eror yang paling minimum.

4.3.2.2. Klasifikasi dengan Algoritma *Backpropagation Neural Network* Menggunakan Rapidminer

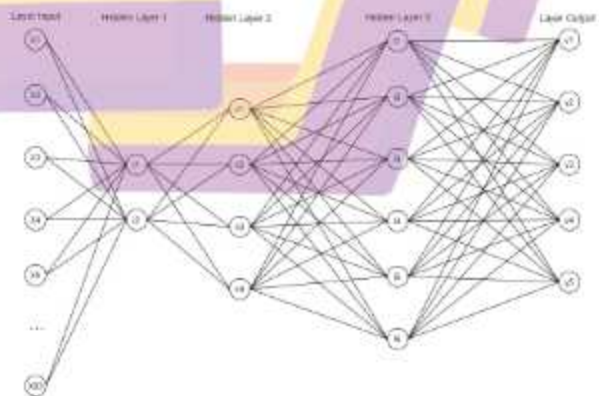
Perhitungan klasifikasi menggunakan rapidminer ini dilakukan untuk memperoleh hasil perhitungan yang sebenarnya dari semua dataset yang dimiliki yang tidak mungkin dilakukan perhitungannya dengan cara manual. Penulis membuat 2 skenario untuk proses perhitungannya yang antara lain. Masing-masing skenario akan penulis coba beberapa variasi yaitu dengan mengganti nilai epoch dan hidden layer. Model/arsitektur dari beberapa variasi tersebut antara lain.

- a. Arsitektur dengan 2 hidden layer yang terdiri dari 2 neuron pada hidden layer pertama dan 4 neuron pada hidden layer kedua.



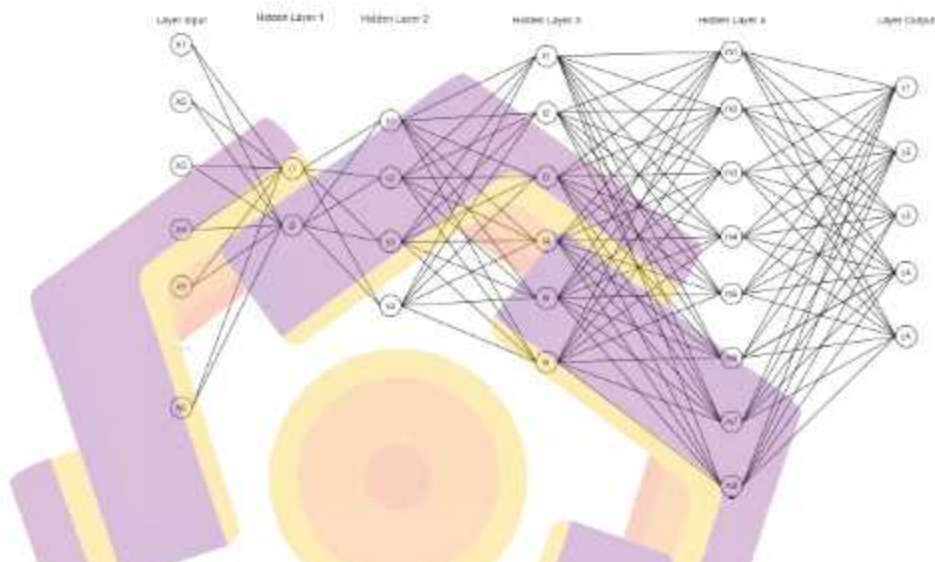
Gambar 4 . 11. Arsitektur Backpropagation dengan 2 Dan 4 Neuron Hidden Layer

- b. Arsitektur dengan 3 hidden layer yang terdiri dari 2 neuron pada hidden layer pertama, 4 neuron pada hidden layer kedua dan 6 neuron pada hidden layer ketiga.



Gambar 4 . 12. Arsitektur Backpropagation dengan 2, 4 dan 6 Neuron Hidden Layer

- c. Arsitektur dengan 3 hidden layer yang terdiri dari 2 neuron pada hidden layer pertama, 4 neuron pada hidden layer kedua, 6 neuron pada hidden layer ketiga dan 8 neuron pada hidden layer keempat.

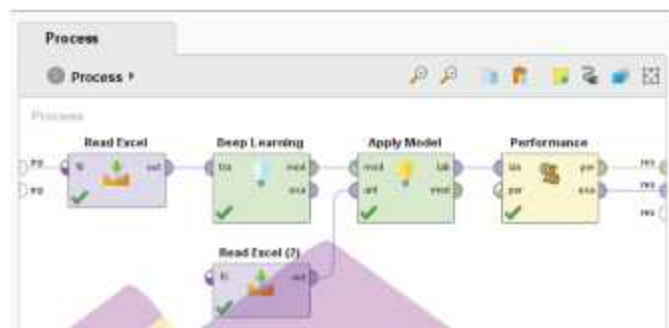


Gambar 4. 13. Arsitektur Backpropagation dengan 2, 4, 6 dan 8 Neuron Hidden Layer

Secara detail proses dan hasil dari skenario klasifikasi dengan algoritma backpropagation neural network pada penelitian ini antara lain:

- Skenario 1 (Membuat model dengan 175 data sebagai *data training*, kemudian mengujinya dengan 75 data lainnya sebagai *data testing*).

Proses dalam skenario ini dapat dilihat pada gambar 4.14.



Gambar 4 . 14. Skenario 1 Proses Klasifikasi dengan Algoritma *Backpropagation Neural Network*

Dalam skenario ini penulis melakukan beberapa percobaan dengan melakukan modifikasi pada jumlah hidden layer-nya. Berikut penulis rangkum beberapa percobaan yang berpengaruh terhadap hasil akurasi algoritma ini. Secara detail dapat dilihat pada tabel 4.12. Dari percobaan tersebut diperoleh hasil akurasi tertinggi sebesar 46,67%.

Tabel 4 . 13. Percobaan Skenario 1 Algoritma *Backpropagation*

Sub Skenario	Epoch	Size Hidden Layer	Akurasi
1.1	10	2,4	38,67%
1.2	10	2,4,6	42,67%
1.3	10	2,4,6,8	41,33%
1.4	15	2,4	45,33%
1.5	20	2,4	46,67%
1.6	25	2,4	48,00%
1.7	30	2,4	40,00%
1.8	15	2,4,6	46,67%
1.9	20	2,4,6	44,00%
1.10	25	2,4,6	44,00%
1.11	15	2,4,6,8	45,00%
1.12	20	2,4,6,8	46,67%
1.13	25	2,4,6,8	42,67%
1.14	30	2,4,6,8	42,67%

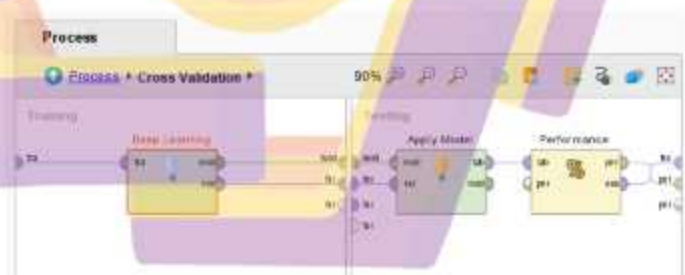
- b. Skenario 2 (Membuat model dengan 250 data yang dimiliki kemudian menguji pemodelan dengan jumlah data yang sama).

Proses dalam skenario ini dapat dilihat pada gambar 4.15.



Gambar 4. 15. Skenario 2 Proses Klasifikasi dengan Algoritma Backpropagation Neural Network

Proses pemodelan dan penerapan algoritmanya terdapat di dalam proses Cross Validation yang alur prosesnya dapat dilihat pada gambar 4.16.



Gambar 4. 16. Proses Dalam Cross Validation Algoritma Backpropagation Neural Network

Pada skenario 2 ini penulis juga melakukan beberapa variasi percobaan pada jumlah dan size hidden layer dari algoritma ini. Berikut penulis rangkum beberapa percobaan yang memiliki hasil

yang cukup berpengaruh terhadap hasil akurasinya. Secara detail dapat dilihat pada tabel. Dari percobaan tersebut diperoleh hasil akurasi tertinggi adalah dengan menggunakan 4 hidden layer dengan size 2, 4 dan 6 yaitu sebesar 49,02%.

Tabel 4 . 14. Percobaan Skenario 2 Algoritma *Backpropagation*

Sub Skenario	Epoch	Size Hidden Layer	Akurasi
1.1	10	2,4	40,00%
1.2	10	2,4,6	46,83%
1.3	10	2,4,6,8	48,50%
1.4	15	2,4	47,42%
1.5	20	2,4	43,43%
1.6	25	2,4	46,21%
1.7	30	2,4	44,02%
1.8	15	2,4,6	47,91%
1.9	20	2,4,6	47,35%
1.10	25	2,4,6	49,02%
1.11	15	2,4,6,8	45,00%
1.12	20	2,4,6,8	46,67%
1.13	25	2,4,6,8	42,67%
1.14	30	2,4,6,8	42,67%

4.4. Pengujian

Pengujian dilakukan dengan membandingkan hasil prediksi masing-masing algoritma dengan label yang telah terdapat pada dataset.

4.5. Evaluasi

Dari hasil klasifikasi dan hasil pengujian secara keseluruhan yang telah dilakukan, nilai akurasi yang diperoleh belum sesuai dengan yang diharapkan. Jika dilihat dari persebaran data pada dataset, user dengan label kepribadian *Openness* sangat dominan. Sedangkan user dengan label *Neuroticism* sangat sedikit. Hal tersebut berpengaruh terhadap hasil klasifikasi untuk data uji karena

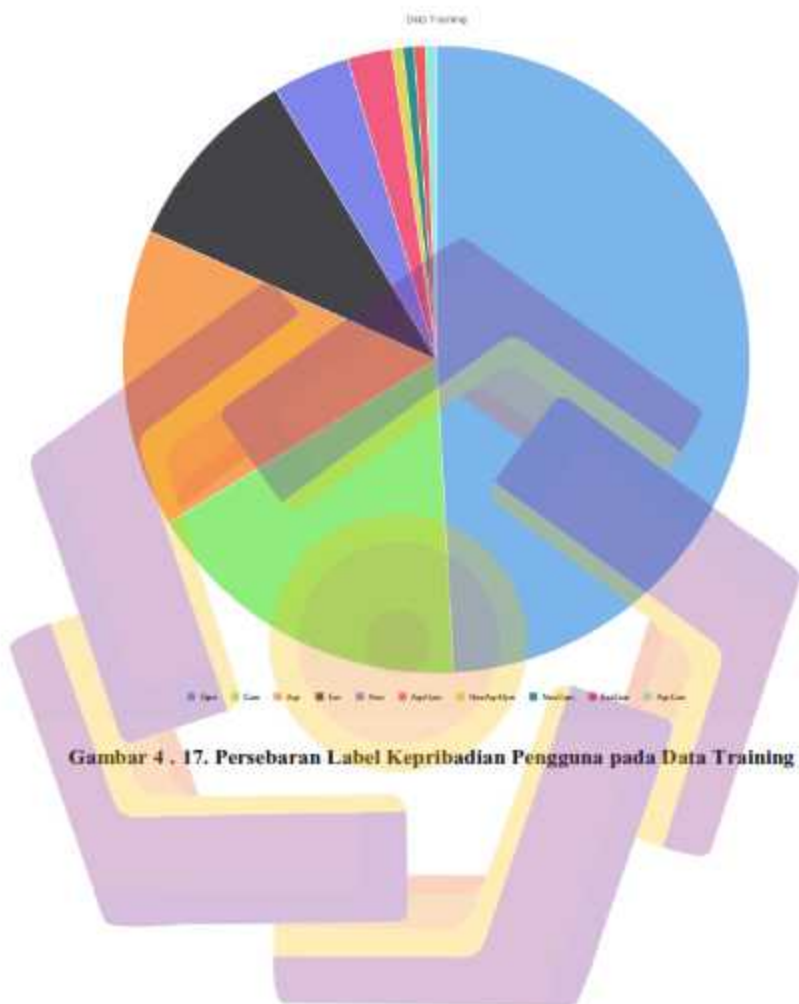
dalam pemodelannya sangat dominan dataset dengan label *Openness* sedangkan kurang cukup data untuk pemodelan label *Neuroticism*.

Komposisi dataset berdasarkan label yang ada dapat dilihat pada tabel berikut.

Tabel 4 . 15. Komposisi Jumlah Data per Label dari Dataset

Label	Jml Dataset	Data Training	Data Testing
Agr	38	27	11
Ext	27	17	10
Con	48	30	18
Opn	121	86	35
Neu	12	7	5
AgrCon	1	1	
AgrOpn	4	4	
ConOpn	2		2
ExtAgr	1		1
ExtCon	1	1	
ExtOpn	1		1
NeuAgrOpn	1	1	
NeuOpn	1	1	

Persebaran jumlah pengguna masing-masing label (kepribadian) dari data training yang digunakan untuk pemodelan dan dari prediksi berdasarkan algoritma Naive Bayes dan Backpropagation Neural Network dapat dilihat pada gambar 4.17.



Jika dilihat dari hasil akurasi klasifikasi kepribadian Big five menggunakan algoritma Naive Bayes dan Neural Network, hasil akurasi tertinggi adalah untuk kelompok kepribadian Opn (Openness). Pada algoritma *Backpropagation Neural Network* akurasi kelompok kepribadian Openness mencapai 97,67%.

accuracy: 48.59% +/- 4.08% (macro average: 48.57%)

	True Con.	True Nev.	True Opt.	True Agr.	True Ext.	True Nev.	True Nev.	True Ext.	True Agr.	True Agr.	class pre.
pred. Con.	1	0	1	0	0	0	0	0	0	0	50.00%
pred. Nev.	0	0	0	0	0	0	0	0	0	0	0.00%
pred. Opt.	26	0	84	27	16	1	1	1	4	1	50.00%
pred. Agr.	1	1	1	0	0	0	0	0	0	0	0.00%
pred. Ext.	0	0	0	0	0	0	0	0	0	0	0.00%
pred. Nev.	0	1	0	0	1	0	0	0	0	0	0.00%
pred. Nev.	0	1	0	0	0	0	0	0	0	0	0.00%
pred. Ext.	0	0	0	0	0	0	0	0	0	0	0.00%
pred. Agr.	0	0	0	0	0	0	0	0	0	0	0.00%
pred. Agr.	0	0	0	0	0	0	0	0	0	0	0.00%
data acc.	2.22%	0.00%	97.67%	0.00%	0.00%	0.00%	0.00%	0.00%	2.22%	0.00%	

Gambar 4. 18. Hasil Akurasi Tiap Kelompok Kepribadian dengan Algoritma *Backpropagation Neural Network*

Berdasarkan hasil evaluasi terhadap perhitungan dan pengujian sebelumnya, penulis mencoba menganalisis kembali dataset dan melakukan proses *balancing* dataset dan melakukan perhitungan ulang.

Berdasarkan dataset awal yang diperoleh, porsi jumlah data yang ada untuk masing-masing label dapat dilihat pada tabel 1. Jumlah yang dimaksud adalah jumlah subjek / user pada tiap-tiap label.

Tabel 4. 16. Jumlah Data Tiap Label pada Dataset Awal

Label	Jumlah Data
Neuroticism	12
Extroversion	27
Agreeableness	38
Consciousness	40
Openness	121

Detail jumlah status yang diambil dari tiap-tiap user per label kepribadian dapat dilihat pada tabel berikut.

Tabel 4 . 17. Jumlah Status Per User (Label Neuroticism)

ID User	Jumlah Status Per User
02c37028a782cfd660c7243e45244bb	6
5532642937eb3497a43e15dbb23a9d2d	1
740aa055e9dccaee874ab7ec7e499d8e	2
7f3bfec0b7228d0900b01fbc8ce9d59f	91
845bd196f34b0ee0246f94eb25ba8c3	18
96487b200f227f3f845bb8851dad2139	12
a6336ecc5e11839ae33aee01fa2163652	74
cce9e48f159ae910f38fb00810be2a68	39
d38a81dfbfbd5f2a2a0a03e9db304c6	3
d7e500ad854a1b6ced39e53a525b8a6d	171
dba59aed04c6759a955a37f021e45bb7	39
eb3eeec6966504b2c01b77f986f24523	10

Tabel 4.17 menjelaskan jumlah status tiap user yang memiliki label kepribadian *Neuroticism* pada dataset yang digunakan.

Tabel 4 . 18. Jumlah Status Per User (Label Extroversion)

ID User	Jumlah Status Per User
1695fbad7101e34d4cb4686a6770231f	92
1886cffe45fdee0240eda42e44a01e07	14
3fe44fab3eb561ae418a22182ec75fad	2
4e5bc97d95f3acu05d420bfd4e3a2639	17
5489ed38556af050d6a93e5d27b95dfb	10
567632b6b7ae8ce7297afcd7ff42c95	48
666c166767bb6645ee3dde9804faa6a9	7
80a570b74f23f56c94f639436ff92353	3
8b77d971b63c574b499cf120fca34bfd	10
8bfbfd9c91e07247d0ff83bbcacee68	61
a1a428d8cf752b3f0527c106b741a9ad	12
a286b7286b1247d4a7851709e9f31e1e	1
ab10a89d4fa299fe348a3eed1d183b3ee	4

Tabel 4 . 19. Jumlah Status Per User (Label Extroverston)(Lanjutan)

ID User	Jumlah Status Per User
acfd53e1393633ae24f8c946d79a17d4	92
aec40862b2a12be50b4d04347985b54d	3
b4a21c82de4011033c8ac67081ff939c	161
bbe1ab61508fad201213f8e8090c5a75	20
c59d2b7cca46cd85c4ac6d6ab554f170	15
c6ee79de33e47a3799c1d00e5f727b9e	29
cd99c28741e42fd9792616d3a4328f17	93
d453b26dfd348fdb99e909df7d8bec39	12
d7f1da35557607089ee2e3a707cc4a4a	36
e465fadd8b30e8669f397e32e10f6cd0	116
e4a512374eee079d2b8acc2ce69990d5	126
ecbddbfe00e0f83cfd802a7186061c7	65
ed17c3b13b9e01ed3d464149e796bcf8	7
f52fccbb0d9b299c889087accd4cc50	9

Tabel 4.18 menjelaskan jumlah status tiap user yang memiliki label kepribadian *Extroversion* pada dataset yang digunakan.

Tabel 4 . 20. Jumlah Status Per User (Label Agreeableness)

ID User	Jumlah Status Per User
13eda4b130aec87b280692cd68e3cabd	9
1d6d222bd3fb4c0af35466042cb82d78	11
1ece40ac9b657dea8c22a1fe58d3a6fe	30
200255966cabe2636535b5b93ac04497	37
22d1f7b24168528163e515b1e96a879c	1
301e1788a595203d0daf3fed10afe1e9	22
325e62f4e7e4f64a03fcf831a8d80bf1	2
37195f370c3fd7486ccedb1519b026c2	94
3b6a414b280222a29d0a95e473cf6b61	33
448084546d4ae45e47c3a83f338ade56	38
502db2fcfe26705ae16a46c5cb2ad2e5	165
6f2bebc01062eb8334dccba3e048fdb5	216
711d198ce5dcc873c3b1fd1468e84d2f	8
751e53fc64b770f7b5ea2b0964678599	56
849cd35a3dfc7200b683b67d0167598f	38

Tabel 4 . 21. Jumlah Status Per User (Label Agreeableness)(Lanjutan)

ID User	Jumlah Status Per User
8cc4acc3bb9b04d1c4a4b4d35ca514ee	7
970197105b41eac939d4fc3cec25137b	39
9e7ebc8082b0e7e92e7aedd3ba562d84	8
a4fb2531183a186f2b5362ab89d885a2	50
a755cc2250ae57df9ff4b5bf005e048	17
a764ca41dca158d7a191505dcc8ce47f	1
a7d9818fb165cd5c84b8578185ccd616	3
a8295e7c5a427770f9777a825e6daf35	18
ab38affacd75668b8eaf1df27e8b5be8	12
b4de950df13c0765018013237c6a1b3d	6
ba1463b4d7bcf9c3167df26e0eb1dc57	4
baa58fb30d937855af7bd7c22f0756b	4
c6a9a43058c8cc8398ca6e97324c0fae	37
c85845b82e705a32f674757d8912df23	91
ee1a957dabf068478d322a56c0dfa41	44
eeff8086fc8220dde9874948728787a2c	3
d39c2b0fb2e50e37795fdb3b8cd3792	113
d8b283c6d460cdc5cb95d65e1c0a71d4	29
da22dab36bb12fcf9ed1a437de278a2c	2
da8787edc39ba38d6bc9f0d10e28aa0f	3
fb67722b811d00a7a40de2ecfab636e6	17
f7456ac4e6b20911c40fdad18908a8d2	149
f83aa7290cf3ac8ed75d6aee5ebbb559	6

Tabel 4.19 menjelaskan jumlah status tiap user yang memiliki label kepribadian *Agreeableness* pada dataset yang digunakan.

Tabel 4 . 22. Jumlah Status Per User (Label Consciousness)

ID User	Jumlah Status Per User
00419a4c96b32cd63b2c7196da761274	1
03133a828cd0cf52e3752813ce5d818f	5
06b055f8e2bca96496514891057913c3	12
0ea88660e23db79e2a3c897a5490d0df	16
1187ed8a8b100eb49b8684ac30d6da29	82
11fc0a8b587b93bb9bfe957f62729d7e	42

Tabel 4.20 menjelaskan jumlah status tiap user yang memiliki label kepribadian *Consciousness* pada dataset yang digunakan.

Tabel 4 . 24. Jumlah Status Per User (Label Openness)

ID User	Jumlah Status Per User
03e6c4eca4269c183fa0e1780f73faba	7
0724fe854bd455061ba84efecdeff469	14
0737e4e4980f56c9fb1cb5743001c917	171
0abd57182992dd9167b88b198350afe4	38
0bfa3d952ffed50f25011b128e73a820	141
0e09a1e89675e72bc2c3f2fac504fced	9
12006962a8659cc890a5a66913494704	41
127d3a99f86b3ee848fd0449bec048fc	1
138ac63ec2b55b8f48fd19c300720cae	12
142635bd7f9ed8e4d8c292fa6ccd9aa4	5
172400f46880b309ca5e97d322bb8f01	72
181962441153a36333f0c60701823412	112
1bd281623fa6ad208caa394dcad75c7d	24
1c10cc0852579d2fa000e3fc3327bb98	19
259f8f9e95b214dc9924af48bcd6f0	69
28038d4b1e4bcfc3b91960f095eccc4	7
2d14a09b41a0d12b9e60377a0a8e14a8	9
2ealefd66ffc011621a4a74e06005e23	81
331a1fdb9c8304206db1b7c4c6852363	31
33e1c2ac25051eda8cc69d1379f584eb	25
35efb99775d5ee7e83cf7912591984d5	1
370a8295df2f8b9069e75422c37d2639	11
3b474abe02985674f2b64e46a939bb3d	59
3d7847b1c33b5f5811208b4aa1a7fbbd	96
43e9b5847ad2e94f53bcc9d826093a76	3
45ab97edbe4e93bc1017bcd28483a678	13
4a6ef0fc11285732bde96e8863dcb7f4	10
4bbb1f5843d9d6e8966dce58179f01de	28
4cac659f923d6f3b4605f38477a04458	76
4d035bd3fd8d9595d15cea9e388964be	40
4f48c6c8051e144c717ce143e8d77782	24
504862f12f8c000ae4358a10d657e8f5	26
521896b01c1a506dc4404e600fa99c5b	104
526ac2635cb3f5e0ee5d7424e83e9107	93

Tabel 4 . 25. Jumlah Status Per User (Label Openness)(Lanjutan)

ID User	Jumlah Status Per User
527ed53d2ba3a3bc417b8402d5b2f556	193
5299b573a6caaf4aebfb3cf6bc8acfd8	69
530e195e467f673de0917893fbf9299a	22
57c2d8496ceb8e92022bcac20cca27e1	48
5880081cd3de1619cd431a75d9052dfc	7
5c73047a8292131a9aa261cfa07932fd	7
60751629cdaa2f62ebc9a33209fcca8e	20
65266a09dc3d274bed54447cf4565392	22
6810431fc51108130be92d505ea4944b	29
690d0ead7039eccdcaeb4d9f77b168b5	46
692758eda219c4fdb6aa64e929be3ff0	69
6980ce18350d98916f56c95b4dc4496d	2
69adae32cb076b219e0d856ef233008	2
69d136c8d8dc113c3aee4835e941798d	21
6c1d913a21f2f3feb2bd15ee6044a4ac1	10
715c9eb832dc833a0b6409ddccd268b1	146
72265a88c1d9956ea6c7b49aa6e54220	18
770b44cbe67c3e68372f529acd82a3ab	25
789ce9b31990354f0a5a507347844dea	2
7938e0ab2768b67f6375eabcfeff5e1	9
7b061056ba902bf5a3f0f123fb181ad3	68
7d41c244db4ee816029991d903cc7ce1	12
7e0954e34b5af347696eb260230bccda	2
810c70fdd8c82cb4e0afc1e8cc62759a	46
83aeaafc166238ebf5aaa2530e6205341	25
85a78b7179d46b27e3f4f3a1468e2a1c	48
8736d2c84c639dc57ef984602d8d96aa	68
8974aab97d9fc4e3a53ba126b5eedd81	1
8a302f01e209fe15fa8425f2d41f633a	34
8d7faa6d7f104a6cb7c4a9e1c6310a15	110
8de8bd927b7382cb4fa28b96d9c1d539	57
8f9d4ed5d16ed1a67d734196d29d1f6b	95
906314c683222f7d232b5ea1ee82befc	50
933791b8623dadbb82664c60db91fb7e	84
94c07315276329ea99beb29b759d4442	28
9b73d5a07c079cce614932e39e2f1852	81
a1c187cbf8532b649cc3b36c01a3ca20	36
a28eb911aad029162befe9e541d9fb86	54

Tabel 4 . 26. Jumlah Status Per User (Label Openness)(Lanjutan)

ID User	Jumlah Status Per User
a40e2ef8193f5273a04cc7bbaae0789c	47
a63a43d8a197935264abe0d57d930712	39
a7637e86876374b952f4d10d0513c350	57
ac8bf16a381d07c01b11651994ecb746	1
ad30ce6078fd6ea54abca85610fba76f	14
b05f34be807a91ce0325cda601bc5856	10
b2be41464b53ffe6deae9536ddf3aee	152
b576dc0449e75306c30456902ce0c80b	44
b7b7764cfa1c523e4e93ab2a79a946c4	99
b8295738a9b021d82860887c3be41164	72
b84d2613c4cf4e73f9c230f57fac66c	43
b98b995372448f015357638c8eb6add1	54
b9a44bb356832160cacc9eb3614e0ce	42
beeb68b947eaec4a55a9d8d20a7c5b52	25
c3f4b3e345cb6b032db2e0459d179db3	152
e5898ecc4c7407e6f07cc3485073d3a6	24
e597771fab7477c2ae7b507d532130b0	47
cde9ad52e50a6a9740dcb3461deb8a0b	22
ce9e978af8369a85f4fc15137013015b	16
cf5a3f66c42918587a1428a6f7d7ecc9	24
d177a27cdf11df2e6e09de9620e9c5e	4
d298a6d98000dc07bbb7994f71d867d3	53
d2ee2871aa02698e66fde72b2218cdd8	14
d3582916ef22700cebb70f5cdedc1da9	13
d67916bbfb453cfebbd8689f7556e526	11
da56abb46f342d8e4b4bc96165cd645d	41
db39f7b2aad360b1033ec1f8fcd57f9c	10
dba5f5266d03dd6d4db084ad7dbc683c	102
ddaed24e83f0f9958336b52cf7a89373	2
deb899e426c1a5c66c24eeb0d7df6257	2
e0576e496b3bc5bd1046315ddb72646a	75
e0697da7478e23479efb723021e52606	28
e2cc76f16ed40953e5c082ff5ef9a2e9	22
e557fd2902b857797cb8ceec471f201a5	44
efcdef6f475ccc3023c5b715f8e9f110	222
e7b7f5af7d02ec06b3da2edf2053dfed	9
ea28a927cb6663480ea33ca917c3c8ba	2
ea890531d3e0547166efde52d843099e	4

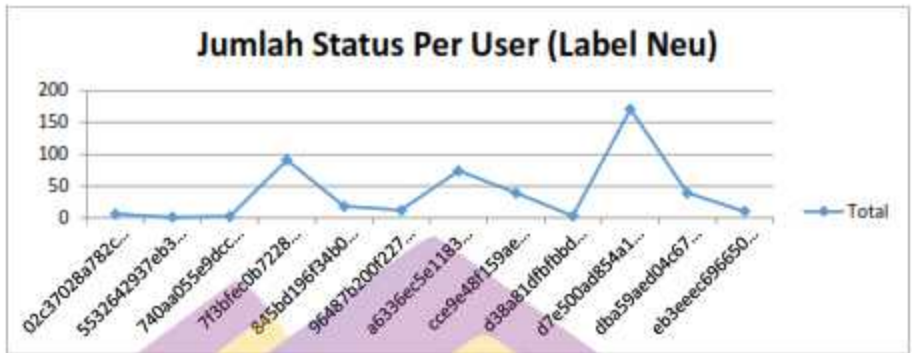
Tabel 4 . 27. Jumlah Status Per User (Label Openness)(Lanjutan)

ID User	Jumlah Status Per User
eac7f51d95da0fefde6ecc692dcf85cc	4
ed1dbabbe3ea7fd3f44efc6f6a96db68	6
ed9073376c5004080ba59ac69f2a1cc3	53
efd5faf19d608aed64447940801cbc30	36
f2026b8cb48aff9af31577ecbfda5c38	122
f65fd0385c972159743d21700c9a7887	21
f680792c707fafceeca42020292e9a	5
f725d0ec71a511886168c30cef6468d9	48
f80644859abfbc92e9a92072c2de50fa	59
fc1c9fb6c64740edcbbf8cfe9dde8b02	30
ffaife151f07a30a0ede2038a897b680	21

Tabel 4.21 menjelaskan jumlah status tiap user yang memiliki label kepribadian *Openness* pada dataset yang digunakan.

Dari tabel tersebut dapat diperoleh kesimpulan bahwa pada dataset yang digunakan, komposisi jumlah status untuk masing-masing user-nya sangat variatif. Terdapat user yang jumlah statusnya hanya 1, sebaliknya ada yang diatas 100. Dari temuan tersebut dapat diperoleh kesimpulan bahwa komposisi jumlah status tiap user dapat menjadi salah satu penyebab dari kurangnya akurasi hasil penelitian.

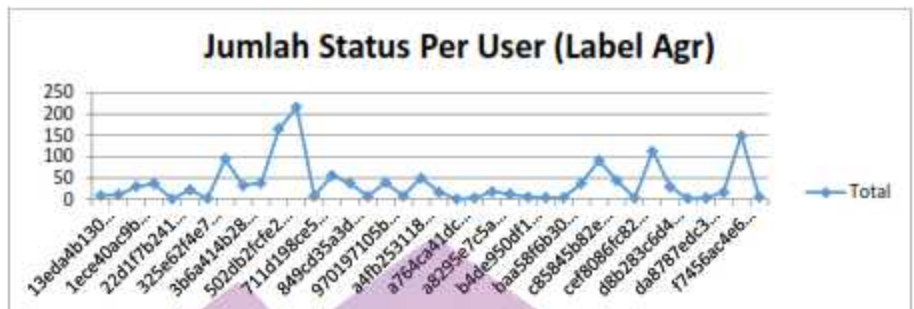
Jika disajikan dalam bentuk grafik, perbedaan jumlah status untuk masing-masing user pada tiap-tiap labelnya dapat dilihat pada grafik berikut.



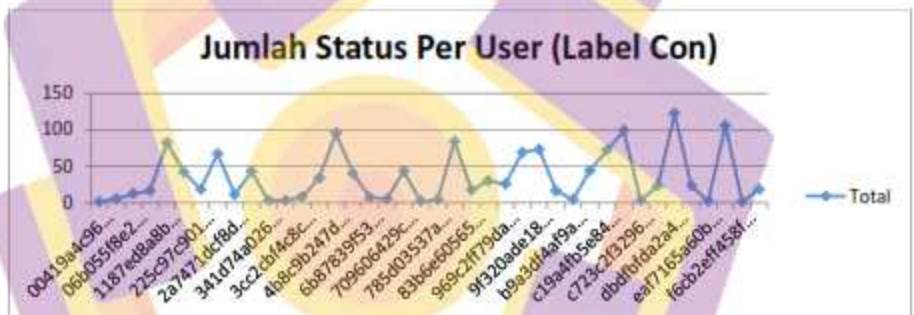
Gambar 4 . 19. Grafik jumlah status per user untuk label kepribadian Neuroticism dalam dataset



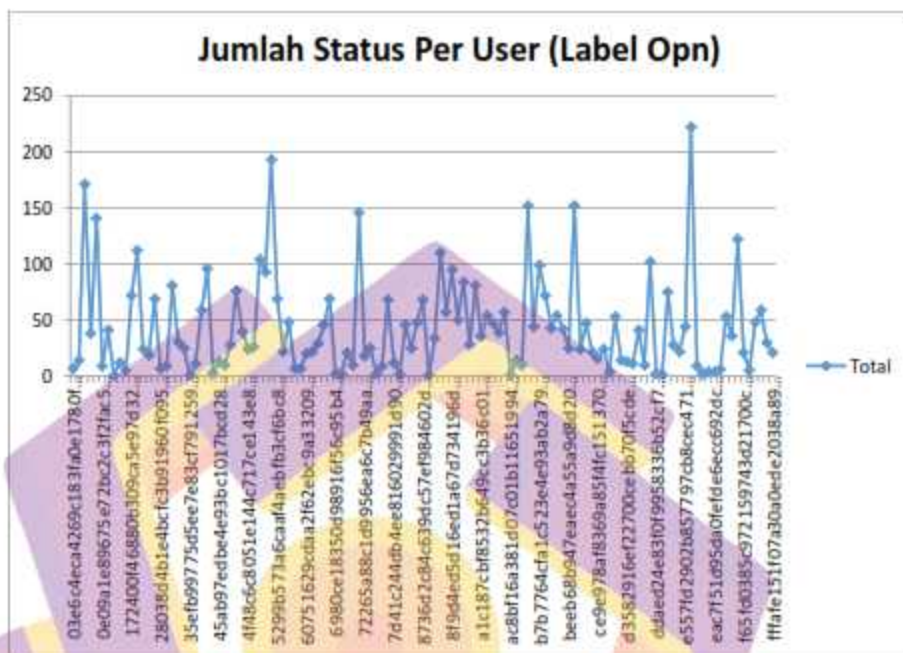
Gambar 4 . 20. Grafik jumlah status per user untuk label kepribadian Extroversion dalam dataset



Gambar 4 . 21. Grafik jumlah status per user untuk label kepribadian Agreeableness dalam dataset



Gambar 4 . 22. Grafik jumlah status per user untuk label kepribadian Conscience dalam dataset



Gambar 4. 23. Grafik jumlah status per user untuk label kepribadian Openness dalam dataset

Dari tabel dan grafik diatas, dapat disimpulkan bahwa jumlah status untuk masing-masing user tidak homogen. Berdasarkan temuan tersebut, penulis mencoba untuk melakukan *balancing* terhadap *imbalance* data. Berdasarkan pertimbangan waktu penelitian, tidak memungkinkan untuk mencari dataset baru dikarenakan perlu proses pelabelan yang dilakukan oleh psikolog dan perlu mencari responden yang memiliki label kepribadian *neuroticism* (karena label *neuroticism* yang sangat kurang jumlahnya pada dataset saat ini). Oleh karena itu penulis mencoba untuk melakukan *balancing* data. Penulis menggunakan metode *undersampling* untuk melakukan *balancing* data. Cara kerja

metode ini adalah dengan mengurangi jumlah data atau ukuran kelas yang memiliki data yang berlimpah. Metode ini penulis terapkan melalui beberapa skenario berikut:

1. Skenario 1: Mengurangi jumlah dataset untuk label yang jumlah datanya diatas 31 data. Penentuan 31 data dilakukan dengan alasan mencari jumlah data pertengahan dengan asumsi selisih dengan data tertinggi tidak terlalu besar dan selisih dengan jumlah data terendah juga tidak terlalu besar. Dengan skenario ini label Agreeableness yang awalnya sebanyak 38 data menjadi 31 data, label Conscientiousness yang awalnya 40 data menjadi 31 data dan label openness yang awalnya 121 data menjadi 31 data. Namun skenario ini tidak menjadi solusi dari permasalahan imbalance dataset yang menyebabkan rendahnya akurasi dikarenakan hasil akurasi dari skenario ini justru lebih rendah jika dibandingkan dengan hasil akurasi pada perhitungan awal.
2. Skenario 2: Mengurangi jumlah dataset yang memiliki label Openness yang awalnya 121 data menjadi 40 data. 40 data yang digunakan dalam skenario dipilih random. Namun hasil akurasi dari skenario pertama ini justru dibawah dari perhitungan awal. Kemudian penulis simpulkan skenario ini juga tidak menjadi solusi dari permasalahan.
3. Skenario 3: Mengurangi jumlah dataset dengan label Openness yang awalnya sebanyak 121 data menjadi 40 data. Penentuan 40 data dikarenakan pada label lain, jumlah dataset tertingginya sebanyak 40 data. Data dipilih dengan memilih 40 data teratas (paling banyak) dari jumlah

status per user dengan asumsi jumlah status yang banyak akan lebih akurat dalam pemodelannya. Sehingga komposisi dataset per labelnya menjadi sebagai berikut.

Tabel 4 . 28. Jumlah Data Tiap Label pada Proses *Balancing* Data Skenario 3

Label	Jumlah Data
Neuroticism	12
Extroversion	27
Agreeableness	38
Consciousness	40
Openness	40
Total	157

Hasil dari pembaruan dataset dengan skenario ini untuk masing-masing algoritma adalah sebagai berikut:

a. Algoritma Naive Bayes

Pada perhitungan dengan dataset awal, akurasi paling tinggi dari algoritma ini adalah sebesar 30,67%. Dengan menggunakan dataset yang telah diperbarui, hasil akurasi mengalami peningkatan namun tidak signifikan, akurasi saat ini yaitu sebesar 37,32%. Deainya dapat dilihat pada gambar.

accuracy: 37.32% +/- 13.43% (micro average: 37.28%)

	true Con	true Neu	true Opn	true Agr	true Est
pred Con	15	2	2	15	1
pred Neu	0	0	1	1	3
pred Opn	13	6	33	10	9
pred Agr	4	2	1	6	5
pred Est	7	2	3	5	9
pred Nie	0	0	0	0	0
pred Nie	0	0	0	0	0
pred Est	0	0	0	0	0
pred Agr	1	0	0	1	0
pred Agr	0	0	0	0	0
pred Est	0	0	0	0	0
pred Est	0	0	0	0	0
pred Co	0	0	0	0	0
class rec.	37.50%	0.00%	82.50%	15.79%	33.33%

Gambar 4 . 24. Hasil Prediksi Algoritma Naive Bayes Setelah Balancing Data dengan Skenario 3

b. Algoritma *Backpropagation Neural Network*

Pada perhitungan dengan dataset awal, akurasi paling tinggi dari algoritma ini adalah sebesar 49,20% dengan 3 hidden layer yang masing-masing layernya terdiri dari 2, 4 dan 6 neuron. Jumlah epoch 25. Hasil perhitungan dengan dataset yang telah diperbarui menggunakan skenario yang sama dengan sebelumnya diperoleh peningkatan akurasi namun tidak signifikan. Akurasi tertinggi saat ini adalah 51,96%. Detailnya dapat dilihat pada gambar.

accuracy: 51.86% +/- 4.32% (micro average: 52.88%)

	Bus Con	Bus Fleu	Bus Opt	Bus Agr	Bus Est	Bus FleuA	Bus Fleu	Bus EstCon	Bus AgrO	Bus AgrC	class pre
pred Con	1	0	0	0	2	0	0	0	0	0	25.33%
pred Fleu	0	0	0	0	0	0	0	0	0	0	0.00%
pred Opt	28	0	85	22	10	1	1	1	4	1	52.47%
pred Agr	1	1	1	4	1	0	0	0	0	0	10.00%
pred Est	0	0	0	1	1	0	0	0	0	0	10.00%
pred FleuA	0	0	0	0	0	0	0	0	0	0	0.00%
pred Fleu	0	0	0	0	0	0	0	0	0	0	0.00%
pred EstCon	0	0	0	0	0	0	0	0	0	0	0.00%
pred AgrO	0	0	0	0	0	0	0	0	0	0	0.00%
pred AgrC	0	0	0	0	0	0	0	0	0	0	0.00%
class recall	3.33%	0.00%	93.84%	14.81%	5.88%	0.00%	0.00%	0.00%	0.00%	0.00%	

Gambar 4 . 25. Hasil Prediksi Algoritma Backpropagation Neural Network Setelah Balancing Data dengan Skenario 3

BAB V

KESIMPULAN

1.1. Kesimpulan

Berdasarkan hasil perhitungan, pengujian dan perbaikan yang telah penulis lakukan, dapat ditarik kesimpulan sebagai berikut:

- a. Berdasarkan hasil klasifikasi yang telah melalui beberapa skenario dan pengujian, algoritma Naive Bayes memiliki akurasi klasifikasi tertinggi sebesar 30,67%. Sedangkan algoritma Backpropagation memiliki akurasi keseluruhan tertinggi sebesar 49,2%. Akurasi tersebut tidak sesuai dengan yang diharapkan. Berdasarkan pengujian lebih lanjut rendahnya akurasi tersebut dikarenakan imbalance dataset. Ditemukan bahwa dari dataset MyPersonality, data dengan label Openness sangat dominan dan data dengan label Neuroticism sangat sedikit. Setelah dilakukan langkah *balancing* data terjadi peningkatan untuk algoritma Naive Bayes akurasinya menjadi 37,32% dan Backpropagation Neural Network menjadi 51,96%.
- b. Hasil akurasi yang diperoleh tetap kurang dari yang diharapkan karena pada dataset jumlah data untuk label neuroticism masih sangat rendah. Untuk itu diperlukan adanya penambahan dataset pada label yang rendah untuk penelitian selanjutnya agar dapat membuktikan apakah algoritma *Naive Bayes* dan *Backpropagation Neural Network* dapat digunakan untuk pemodelan klasifikasi kepribadian Big Five Berdasarkan status facebook.

- c. Selain masalah pada data yang menyebabkan rendahnya hasil akurasi, perlu dipertimbangkan untuk melakukan analisa apakah fitur yang dipilih sudah benar-benar tepat. Sehingga penulis menyarankan untuk penelitian selanjutnya mempertimbangkan melakukan analisa untuk pemilihan fitur pada kasus klasifikasi kepribadian pengguna Facebook ke dalam kerangka Big five personality.

1.2. Saran

Berdasarkan hasil dari penelitian yang telah dilakukan, penulis mengajukan beberapa saran sebagai bahan pertimbangan bagi penelitian selanjutnya sehingga diharapkan akan memperoleh hasil yang lebih baik untuk penelitian selanjutnya, yang antara lain sebagai berikut:

- a. Keseimbangan komposisi dataset sebagai data training perlu diperhatikan untuk proses pemodelan yang lebih akurat. Sebaiknya jangan sampai ada kelas label yang sangat dominan atau sangat rendah jumlahnya.
- b. Penelitian selanjutnya dapat mencari fitur yang tepat yang benar-benar berpengaruh terhadap hasil klasifikasi kepribadian.
- c. Perlu dilakukan optimasi untuk algoritma *naive bayes* untuk mendapatkan peningkatan akurasi.
- d. Perlu adanya penelitian lebih lanjut untuk membandingkan kualitas fitur dengan pendekatan *closed vocabulary* menggunakan framework LIWC dan pendekatan *open vocabulary* menggunakan metode *bag of words*.

DAFTAR PUSTAKA

PUSTAKA BUKU

- Suyanto. (2017). *Data Mining Untuk Klasifikasi dan Klusterisasi Data*. Bandung : Penerbit Informatika.
- Santosa, Budi. (2007). *Data Mining Teknik Pemanfaatan Basis Data untuk Keperluan Bisnis*. Yogyakarta : Graha Ilmu.
- Hermawati, Fajar Astuti. (2013). *Data Mining*. Yogyakarta : Andi.

PUSTAKA MAJALAH, JURNAL ILMIAH ATAU PROSIDING

- W. Moran, John & K. Brightman, Baird. (2001). *Leading organizational change. Career Development International*. 6. 111-119.
- Dewi, Sari. (2016). *Komparasi 5 Metode Algoritma Klasifikasi Data Mining Pada Prediksi Keberhasilan Pemasaran Produk Layanan Perbankan*. Jurnal Techno Nusa Mandiri Vol.XIII, No. 1 Maret 2016.
- Xhemali, Daniela, Chris J. Hinde, and Roger G. Stone. (2009). *Naive Bayes vs. decision trees vs. neural networks in the classification of training web pages*. IJCSI, Vol. 4, No. 1.
- Lhaksana, Kemas Muslim, Nhita, Fhira., Anggraini, Duwi. *Klasifikasi Kepribadian Berdasarkan Status Facebook Menggunakan Metode Backpropagation*. e-proceeding of Engineering : Vol. 4, No. 3 Desember 2017.
- Ahmad, N and Siddique, J. 2017. *Personality Assessment using Twitter Tweets*. 21 *International Conference on Knowledge Based and Intelligent Information an Engineering Systems*. KES2017, September 2017.
- Yusra., Fikry, M., Syarifanto, R., Candra, R. M., Budiani, E. (2018). *Klasifikasi Kepribadian Big Five Pengguna Twitter dengan Metode Naive Bayes*. NTIKI-10.
- Tandera, T., Hendro, Suhartono, D., Wongso, R. And Prasetyo, Y. L. (2017). *Personality Prediction System from Facebook Users*. ICCSI 2017, 13 – 14 October 2017, Bali, Indonesia.
- Assuja, M. A., Saniati. (2016). *Analisis Sentimen Tweet Menggunakan Backpropagation Neural Network*. Jurnal TEKNOINFO, Vol. 10, No. 2, 2016.

- Farnadi G, Sitaraman G, Sushmita S, Celli F, Kosinski M, Stillwell D, et al. *Computational personality recognition in social media*. User Model User-adapt Interact. 2016.
- Qiu, L., Lin, H., Ramsay, J. And Yang, F. (2012). *You are What You Tweet : Personality Expression on Twitter*, Journal of Research in Personality 46 (2012) 710-718.
- Pratama, B. Y., Sarno, R. (2015). *Personality Classification Based on Twitter Text Using Naive Bayes, KNN and SVM*. 2015 International Conference on Data and Software Engineering.
- Sarwani, Mohamad Zoqi dan Mahmudy, Wayan Firdaus. (2015). *Analisis Twitter untuk Mengetahui Karakter Seseorang Menggunakan Algoritma Naive Bayes Classifier*. Seminar Nasional Sistem Informasi Indonesia, 24 November 2015.
- Golbeck, J., Robles, C., Edmondson, M., and Turner, K. (2011). *Predicting Personality from Twitter*. IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing.