TESIS

ANALISA KLASIFIKASI KEPRIBADIAN MYERS-BRIGGS INDICATOR TYPE (MBTI) MENGGUNAKAN NAIVE BAYES DAN K-NEAREST NEIGHBOR

Disusun oleh:

Nama	**: Ahmad Fikri Iskandar**
NIM	**: 19.77.1170**
Konsentrasi	**: Business Intelligence**

PROGRAM STUDI S2 TEKNIK INFORMATIKA

PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2021

TESIS

ANALISA KLASIFIKASI KEPRIBADIAN MYERS-BRIGGS INDICATOR TYPE (MBTI) MENGGUNAKAN NAIVE BAYES DAN K-NEAREST NEIGHBOR

ANALYSIS OF CLASSIFICATION MYERS-BRIGGS INDICATOR TYPE (MBTI) PERSONALITY TRAIT USING NAIVE BAYES AND K-NEAREST NEIGHBOR

Diajukan untuk memenuhi salah satu syarat memperoleh derajat Magister

Disusun oleh:

Nama	**: Ahmad Fikri Iskandar**
NIM	**: 19.77.1170**
Konsentrasi	**: Business Intelligence**

PROGRAM STUDI S2 TEKNIK INFORMATIKA

PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2021

HALAMAN PENGESAHAN

ANALISA KLASIFIKASI KEPRIBADIAN MYERS-BRIGGS INDICATOR TYPE (MBTI) MENGGUNAKAN NAIVE BAYES DAN K-NEAREST NEIGHBOR

ANALYSIS OF CLASSIFICATION MYERS-BRIGGS INDICATOR TYPE (MBTI) PERSONALITY TRAIT USING NAIVE BAYES AND K-NEAREST NEIGHBOR

Dipersiapkan dan Disusun oleh

Ahmad Fikri Iskandar

19.77.1170

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 Teknik Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari Selasa, 03 Agustus 2021

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

Yogyakarta, 03 Agustus 2021
Rektor

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

HALAMAN PERSETUJUAN

ANALISA KLASIFIKASI KEPRIBADIAN MYERS-BRIGGS INDICATOR TYPE (MBTI) MENGGUNAKAN NAIVE BAYES DAN K-NEAREST NEIGHBOR

ANALYSIS OF CLASSIFICATION MYERS-BRIGGS INDICATOR TYPE (MBTI) PERSONALITY TRAIT USING NAIVE BAYES AND K-NEAREST NEIGHBOR

Dipersiapkan dan Disusun oleh

Ahmad Fikri Iskandar

19.77.1170

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 Teknik Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari NamaHari, tanggal ujian tesis

Pembimbing Utama	**Anggota Tim Penguji**
Prof. Dr. Ema Utami, S.Si., M.Kom	**Dr. Wing Wahyu Winarno, MAFIS, Ak.**
NIDN. 0521027501	**NIDN. 0525016201**
	Alva Hendi Muhammad, S.T., M.Eng. , Ph.D
Pembimbing Pendamping	**NIDN. 0518078401**
Agung Budi Prasetio, S.T., M.Eng	**Prof. Dr. Ema Utami, S.Si., M.Kom**
NIDN. 0507048502	**NIDN. 0521027501**

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

Yogyakarta, 03 Agustus 2021
Direktur Program Pascasarjana

Dr. Kusrini, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN TESIS

Yang bertandatangan di bawah ini,

Nama mahasiswa : Ahmad Fikri Iskandar
NIM : 19.77.1170
Konsentrasi : Business Intelligence

Menyatakan bahwa Tesis dengan judul berikut:
Analisa Klasifikasi Kepribadian Myers-Briggs Indicator Type (MBTI) Menggunakan Naïve Bayes Dan K-Nearest Neighbor

Dosen Pembimbing Utama : Prof. Dr. Ema Utami, S.Si., M.Kom
Dosen Pembimbing Pendamping : Agung Budi Prasetio, S.T., M.Eng

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, 03 Agustus 2021
Yang Menyatakan,

Ahmad Fikri Iskandar

HALAMAN PERSEMBAHAN

Alhamdulillah dengan rasa syukur yang mendalam dengan telah diselesaikannya tesis ini berkat segala nikmat dan kasih sayang yang telah diberikan oleh Allah Subhanahu Wa Ta'ala, maka penulis mempersembahkan tesis ini kepada semua pihak yang terlibat secara langsung ataupun tidak langsung dalam proses pembuatan tesis.

1. Orang tua dan saudara kakak-kakak, yang selalu mendoakan, memberikan semangat untuk menjalani perkuliahan.
2. Prof. Dr. M. Suyanto, M.M selaku Rekor Universitas AMIKOM Yogyakarta yang telah memberikan kesempatan kepada penulis untuk melanjutkan Studi jenjang Strata 2 Program Studi Magister Teknik Informatika di Universitas Amikom Yogyakarta.
3. Prof. Dr. Ema Utami, S.Si., M.Kom dan Bapak Agung Budi Prasetio, S.T., M.Eng yang telah membimbing penulis dari awal sampai akhir proses pembuatan tesis.
4. Semua pihak yang tidak bisa disebutkan satu persatu yang sudah memberi semua ilmu pengetahuan, informasi dan segalanya sehingga penulis bisa menyelesaikan tesis ini.

KATA PENGANTAR

Puji dan syukur penulis panjatkan kehadirat Allah Subhanahu Wa Ta'ala atas nikmat dan karunia-Nya sehingga penulis dapat menyelesaikan tesis yang berjudul "Analisa Klasifikasi Kepribadian Myers-Briggs Indicator Type (MBTI) Menggunakan Naïve Bayes Dan K-Nearest Neighbor". Penulis menyadari tidak akan dapat menyelesaikan tesis ini dengan baik tanpa bimbingan, saran dan motivasi dari berbagai pihak. Peneliti mengucapkan terimakasih yang sebesar-besarnya kepada:

1. Prof. Dr. M. Suyanto, MM. selaku Rektor Universitas AMIKOM Yogyakarta.
2. Prof. Dr. Ema Utami, S.Si., M.Kom dan Agung Budi Prasetio, S.T., selaku pembimbing 1 dan 2.
3. Dr. Wing Wahyu Winarno, MAFIS, Ak. dan Alva Hendi Muhammad, S.T., M.Eng., Ph.D selaku Penguji 1 dan 2.
4. Orang tua serta saudara selaku wali yang telah memberikan dukungan dan motivasi.

Semoga Allah Subhanahu Wa Ta'ala memberikan balasan yang lebih kepada pihak yang telah membantu penulis menyelesaikan tesis ini. Semoga tesis ini dapat bermanfaat bagi masyarakat.

Yogyakarta, 20 September 2021

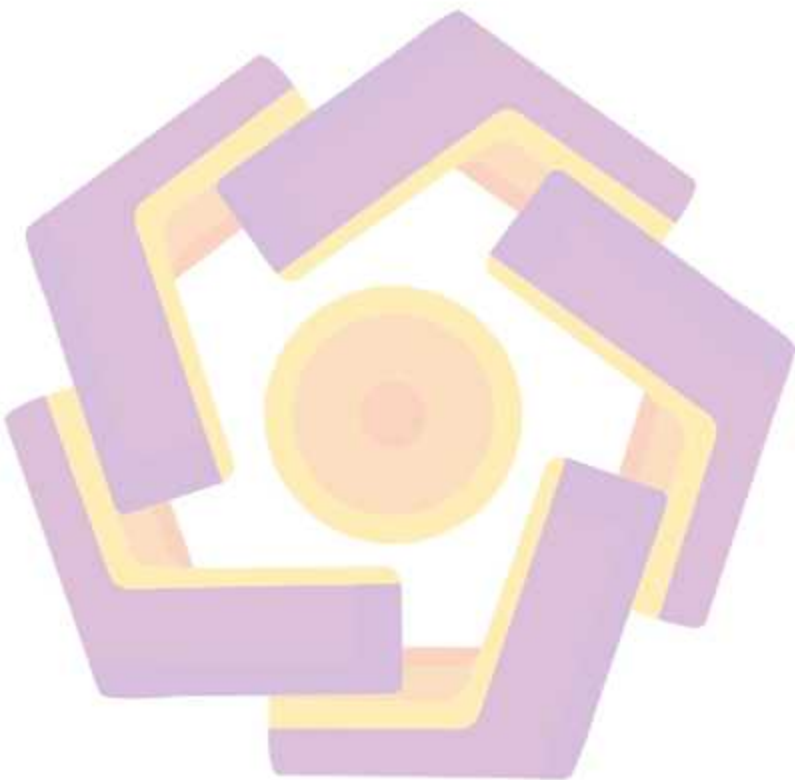
Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN



INTISARI

Tujuan penelitian yang berjudul "Analisa Klasifikasi Kepribadian Myers-Briggs Indicator Type (MBTI) menggunakan Naïve Bayes dan K-Nearest Neighbor (KNN)" ini adalah melakukan perbandingan tingkat akurasi klasifikasi menggunakan model NBC dengan model K-NN serta mengetahui perubahan tingkat akurasi dengan menggunakan fitur ekstraksi dan seleksi fitur dalam mengklasifikasikan preferensi kepribadian berdasarkan kepribadian MBTI. Model yang digunakan adalah NBC dan K-NN, nilai parameter K pada model K-NN adalah 1,3,5,7, dan 9 serta pembobotan kata yang digunakan berdasarkan BoW dan TF-IDF. Terdapat empat skenario dalam uji coba klasifikasi yaitu skenario pertama adalah BoW atau TF-IDF, skenario kedua adalah penambahan ekstraksi fitur skenario ketiga adalah menyeleksi fitur dengan chi-square dan skenario keempat adalah skenario *balancing*.

Hasil dari rata-rata akurasi klasifikasi terbaik adalah yaitu untuk model NBC untuk dimensi IE 71,216%, dimensi NS, 77,455%, dimensi TF 76,174% dan dimensi JP 74,93%. Sedangkan model K-NN untuk dimensi IE 74,191%, dimensi NS 78,300%, dimensi TF 76,113%, dan dimensi JP 75,521%. Kenaikan akurasi dari skenario pertama ke skenario kedua 1,356%. Kenaikan akurasi dari skenario kedua ke skenario 4,441%, sedangkan pada skenario ketiga ke skenario keempat mengalami penurunan 3,385%.

Kata kunci: NB, K-NN, MBTI, Chi-Square, Ekstraksi Fitur, *Undersampling*

ABSTRACT

The purpose of this research, entitled "Analysis of Classification Myers-Briggs Indicator Type (MBTI) Personality Trait Using Naïve Bayes and K-Nearest Neighbor (K-NN)" is to compare the level of classification accuracy using the NBC model with the K-NN model and determine increasing accuracy by using feature extraction and feature selection in classifying personality preferences based on MBTI personality.Models used are NBC and K-NN, parameter value K in the K-NN model are 1,3,5,7, and 9 with word weighting used is based on the frequency and TF-IDF. There are four scenarios in the classification trial, namely the first scenario is frequency or TF-IDF, the second scenario is the addition of feature extraction, the third scenario is selecting features with chi-square and the fourth scenario is the balancing data.

The results of the best average classification accuracy are for NBC model for IE dimensions 71.216%, NS dimensions, 77.455%, TF dimensions 76.174% and JP dimensions 74.93%. While, K-NN model for IE dimensions 74.191%, NS dimensions 78.300%, TF dimensions 76.113%, and JP dimensions 75.521%. The increase in accuracy from the first scenario to the second scenario is 1.356%. The increase in accuracy from the second scenario to the scenario was 4.441%, while in the third scenario to the fourth scenario it decreased by 3.385%.

Keyword: NBC, K-NN, MBTI, Chi-Square, *feature extraction, Undersampling*

BAB I

PENDAHULUAN

1.1. Latar Belakang Masalah

Kepribadian merupakan ciri khas yang membuat setiap manusia memiliki keunikan dan perbedaannya masing-masing. Kepribadian dapat dijadikan sebagai ukuran sebagai dasar pemikiran, perasaan, serta pola-pola berperilaku dalam berbagai aspek kehidupan sehari-hari baik berhubungan langsung pada diri sendiri maupun orang lain (Utami, dkk 2019).

Salah satu implementasi kepribadian dalam kehidupan sehari-hari dapat ditemukan pada perusahaan bagian sumber daya manusia dalam merekrut pelamar yang berkompetensi. Tes seleksi dilakukan untuk mengukur tingkat kemampuan pelamar dalam menguasai bidangnya, kepribadiannya, serta prestasi apa yang dimiliki oleh pelamar tersebut (Utami, dkk, 2019). Preferensi kepribadian diperlukan dalam proses awal merekrut pelamar, di mana tujuannya adalah untuk mendeteksi masalah psikologis pelamar (Yuan, dkk, 2018). Kepribadian pelamar dan karyawan juga dapat menunjukkan kemampuan bekerja sama dan berkolaborasi pada tim (Claudy, dkk 2018).

Kepribadian juga dapat dijadikan sebagai strategi pemasaran dan deteksi kesehatan mental seseorang. Kepribadian disesuaikan dengan preferensi orang mengenai produk tertentu sehingga dapat memberikan promosi yang berbeda pada setiap orangnya (Moreno, dkk 2019), (Lukito, dkk 2016). Deteksi secara dini kepribadian seseorang juga bermanfaat pada kesehatan seseorang tersebut. Menurut

Preotiuc-Pietro, dkk (2015) menganalisis linguistik pada penggunaan media sosial dapat menjadi salah satu alternatif yang dapat digunakan untuk melihat penyakit mental pengguna. Permasalahan untuk mengukur kepribadian secara konvensional membutuhkan waktu yang lama tergantung kepada jumlah item dan konten kuesioner serta fokus para peneliti. Responden enggan memberikan ketersediaannya untuk mengisi dan merasa bosan jika konten pada kuesioner tersebut jika terlalu banyak dan memuat pernyataan yang berulang (Moreno, dkk, 2019).

Terdapat 3 preferensi kepribadian yang sering digunakan yaitu *Dominant Influence Steadiness Compliant* (DISC), Big-Five atau *Openness Conscientiousness Extroversion Agreeableness Neuroticism* (OCEAN) dan Myers-Briggs Indicator Type (MBTI) (Verhoeven, dkk, 2016). Detail dari perbandingan dari ketiga preferensi kepribadian ini dapat dilihat pada Tabel 1.1.

Tabel 1.1. Perbandingan Preferensi Kepribadian

	MBTI	**DISC**	**Big-Five**
Karya	Myress, Brigss dan Carl Jung	William Moulton Marston	Lewis R. Goldberg
Measured Trait	4 *binary* atau 16 *trait*	12 *trait*	30 *spectrum*
Pendekatan	Berlawanan	Pendekatan	Pendekatan
Jumlah Instrumen	64 -100 item	24 item	180 item
Pergunakan untuk	Tujuan pribadi atau Akademik	Lingkungan tempat kerja	Analisis Faktor
Data tersedia di media sosial	Ya, (Hasil 16personalities.com)	Tidak	Tidak

Big-Five digunakan untuk memahami, menanyakan kebutuhan serta mengoptimalkan diri sendiri yang digunakan untuk prediksi kepribadian yang akan datang. DISC sering digunakan pada dunia kerja untuk mengetahui individu

menanggapi aturan, lingkungan, serta masalah dan tantangan. MBTI merupakan instrumen yang sangat mudah dipahami serta digunakan oleh kelompok muda atau remaja untuk tujuan pribadi atau akademik. Pengguna media sosial yang didominasi oleh remaja menggunakan preferensi MBTI membicarakan hasil tes kepribadian di media sosial seperti Twitter, sehingga data hasil tes kepribadian MBTI mereka tersedia dan *Open Access* (Celli dan Lepri, 2018). Berbeda dengan DISC dan Big-Five yang jarang dibicarakan oleh pengguna di media sosial.

MBTI didefenisikan 4 dimensi yaitu : *Introvert-Extrovert* (IE), *Intuition-Sensing* (NS), *Thinking-Feeling* (TF) dan *Judging-Perceiving* (JP). Dimensi IE merupakan perhatian untuk melihat arah energi kita ke dalam atau ke luar seperti bergaul, menyukai interaksi sosial dan sebaliknya. Dimensi NS memahami informasi dari bagaimana memproses informasi yang akan masuk dengan menggunakan fakta atau melihat pola pengetahuan serta hubungannya. Dimensi TF menarik kesimpulan dan keputusan berdasarkan analisa dan logika atau melibatkan perasaan dan empati. Terakhir, dimensi JP pola hidup yang selalu berfokus pada rencana yang sistematis atau beradaptasi terhadap perubahan yang mendadak. Berdasarkan keempat dimensi dasar ini akan dihasilkan 16 tipe kepribadian seperti ESFJ (*Extrovert, Sensing, Feeling, Judging*), ENTP (*Extrovert, Intuition, Thinking, Perceiving*), dan seterusnya (Briggs & Myres, 1995).

Pengunaan media sosial berkembang s semakin cepat, diikuti dengan orang yang ingin saling terhubung untuk mengekspresikan pendapat, emosi, perilaku lainnya di jejaring sosial. Dikutip dari Jayani (2020) menjelaskan pengguna media sosial di Indonesia menghabiskan rata-rata waktu sebesar 3 jam 26 menit setiap

hari. Pengunaan media sosial di Indonesia tidak hanya digunakan sebagai media penyampaian saran dan kritik, isu terkait topik tertentu, tetapi adu argumen terkait pilihan masing-masing, sehingga penggunaan media sosial telah mendorong peningkatan informasi teks tanpa batas yang bisa diakses oleh siapa saja tanpa mengurangi nilai informasinya (Mihuandayani, dkk, 2018), (Suryono, dkk 2018).

Penggunaan media sosial Twitter saat ini sangat menarik untuk mengetahui isi cuitan yang dapat berupa opini atau informasi tentang kebiasaan pengguna tesebut berdasarkan kata-kata yang sering digunakan (Utami,dkk 2019). Perilaku yang ditampilkan pada isi cuitan oleh pengguna dipengaruhi berdasarkan kepribadian mereka masing-masing, sehingga memudahkan para peneliti untuk menjadikan sebagai data (Yuan, dkk 2018). Penelitian Qiu, dkk (2012) menjelaskan bahwa cuitan seseorang mengandung isyarat linguistik untuk kepribadian di media sosial dengan korelasi lemah hingga sedang antara isyarat linguistik dan ciri-ciri kepribadian. Tipe kepribadian ekstrovert berkorelasi positif dengan kata-kata emosi positif dan sosial (Yarkoni, 2010). Selain itu, Peterka-Bonetta, dkk (2021) mengungkapkan bahwa asosiasi kepribadian hanya muncul kuat pada kelompok pengguna yang menunjukkan minat terbesar pada topik tertentu.

Prediksi kepribadian MBTI berbahasa Indonesia sudah pernah dilakukan sebelumnya seperti yang dilakukan oleh Claudy (2018). Penelitian tersebut bertujuan untuk mengklasifikasi dokumen twitter untuk mengetahui karakter dari calon karyawan yang dimana menggunakan data yaitu 160 data dengan label *Artisan, Guardian, Idealist,* dan *Rational* dengan model K-NN mendapat akurasi

66%. Penelitian yang dilakukan Fikry (2018) juga melakukan penelitian terkait preferensi kepribadian pada bagian *Introvert* atau *Extrovert* dengan menggunakan ekstraksi fitur pada teks. Hasil yang didapatkan dengan menggunakan model Support Vector Machine (SVM) sebesar 88%. Penelitian Ong, dkk (2017) melakukan klasifikasi kepribadian dan membandingan hasil 12 skenario dengan kombinasi seleksi fitur dan pembobotan kata. Hasil akurasi model SVM mencapai rata-rata tertinggi 76,23%. Terakhir, penelitian Lukito, dkk (2016) juga melakukan penelitian untuk mengklasifikasi preferensi kepribadian pengguna Twitter. Hasil yang didapatkan dengan menggunakan model Naïve Bayes yaitu IE: 80%, NS: 60%, TF: 60 % dan JP: 60 %.

Terdapat dua model klasifikasi yang popular digunakan, yaitu Naïve Bayes Classifier (NBC) dan K-Nearest Neighbor (K-NN). NBC memiliki asumsi *naïve* yaitu semua fitur penting dan bebas. Penentuan kelas pada model NBC yaitu membandingkan nilai probalitas satu sampel yang berada di kelas yang satu dengan nilai probalitas suatu sampel berapa di kelas lainnya. Model NBC biasanya digunakan untuk mengklasifikasi teks dan NBC juga memiliki akurasi yang tinggi dengan perhitungan sederhana (Lukito dkk, 2016),(Dinov dkk, 2018). Berbeda dengan NBC, K-NN merupakan model klasifikasi yang menentukan label dari suatu objek baru berdasarkan kelas yang mayoritas dari jarak terdekat (*neighbor*). K-NN menghitung berdasarkan jarak antara objek dengan kelas, sehingga klasifikasi menggunakan model K-NN akan mendapatkan hasil akurasi yang lebih baik (Claudy dkk, 2018). Berdasarkan masalah diatas, penulis tertarik untuk melakukan analisis preferensi kepribadian MBTI dengan menggunakan NBC dan K-NN.

1.2. Rumusan Masalah

Adapun rumusan masalah pada penelitian ini sebagai berikut:

a. Berapa perbandingan tingkat akurasi klasifikasi menggunakan model NBC dengan model K-NN dalam mengklasifikasi preferensi kepribadian berdasarkan kepribadian MBTI?
b. Berapa perubahan tingkat akurasi dengan menggunakan fitur ekstraksi dan seleksi fitur dalam mengklasifikasikan preferensi kepribadian berdasarkan kepribadian MBTI?

1.3. Batasan Masalah

Batasan masalah pada penelitian ini adalah:

a. Sumber data diperoleh dari media sosial Twitter yang berbahasa Indonesia.
b. Model yang digunakan adalah NBC dan K-NN.
c. Nilai parameter K pada model K-NN adalah 3,5,7 dan 9.
d. Pembobotan kata yang akan berdasarkan BoW dan TF-IDF.
e. Pembagian data latih dan uji yaitu 70:30 dan 80:20.
f. Teknik *sampling* yang digunakan adalah *Undersampling*

1.4. Tujuan Penelitian

Tujuan dari penelitian ini adalah

a. Melakukan perbandingan tingkat akurasi klasifikasi menggunakan model NBC dengan model K-NN dalam mengklasifikasi preferensi kepribadian berdasarkan kepribadian MBTI.
b. Mengetahui perubahan tingkat akurasi dengan menggunakan fitur ekstraksi dan seleksi fitur dalam mengklasifikasikan preferensi kepribadian berdasarkan kepribadian MBTI.

1.5. Manfaat Penelitian

Manfaat yang dapat diberikan oleh penelitian ini sebagai berikut:

a. Memberikan model yang dapat melakukan klasifikasi kepribadian MBTI berdasarkan cuitan di media sosial Twitter.
b. Sebagai alat bantu pengguna Twitter dalam mendeteksi secara dini kepribadian sesuai dengan preferensi MBTI.
c. Dapat menjadi rekomendasi perusahaan dalam sistem seleksi pelamar baik secara langsung maupun tidak langsung.

BAB II

TINJAUAN PUSTAKA

2.1. Tinjauan Pustaka

Penelitian yang dilakukan oleh Claudy dkk, (2018) mengimplementasikan K-NN untuk melakukan klasifikasi karakter calon karyawan dari cuitan menjadi 4 kelas yaitu *Artisan, Guardian, Idealist,* dan *Rasional*. Data yang didapatkan adalah sebesar 160 data. Hasil implementasi model K-NN pada didapatkan hasil yang benar sebesar 53 data dan salah sebesar 27 data, maka hasil akurasi sebesar 66%.

Terdapat 3 pendekatan yang dilakukan oleh Lukito, dkk (2016) dalam mengembangkan serta membandingkan hasil klasifikasi Kepribadian MBTI berbahasa Indonesia yaitu, Naive Bayes, *Lexicon Based,* dan *Grammatical Rule*. Data yang didapatkan adalah 97 pengguna Twitter. Pembobotan kata yang digunakan yaitu TF-IDF. Pembagian data yaitu 84,5% data latih dan 15,5% data uji. Model Naive Bayes adalah model yang tingkat *perfomance* lebih baik daripada 2 pendekatan lainnya. Akurasi IE sebesar 72,5%, NS sebesar 60%, TF sebesar 61,2% dan JP sebesar 55,4%.

Ekstrasi fitur menjadi salah satu alternatif untuk meningkatkan data, seperti yang dilakukan oleh Fikry, dkk (2018) menggunakan ekstraksi fitur sebanyak 17 fitur yang dibentuk dari cuitan untuk klasifikasi kepribadian ekstrover dan introver yaitu jumlah cuitan, jumlah tautan, jumlah tagar, jumlah cuitan disukai, jumlah pengikut, jumlah mengikuti dan lain sebagainya. Proses pembagian data latih dan

data uji sebanyak 3 kali yaitu 70:30, 80:20, dan 90:10. Hasil pengujian terbaik dengan model SVM, memperoleh akurasi sebesar 88,89%.

Penelitian Ong, dkk (2017) juga membangun sistem klasifikasi kepribadian Big-Five berdasarkan bahasa Indonesia serta membandingan 12 skenario dengan parameter pembobotan kata, *topic modelling*, *stopword* dan n-gram. Data yang didapatkan adalah 359 pengguna Twitter. Pembagian data untuk data latih sebanyak 98% dari data awal dan untuk data uji sebanyak 8% dari dari data awal. Evaluasi menggunakan validasi silang 10 pada SVM berhasil mencapai akurasi rata-rata tertinggi 76,23%, sedangkan XGBoost mencapai 97,99%.

Iskandar, dkk (2020) dan Utami, dkk (2019) melakukan eksplorasi terhadap kata-kata pada kalimat yang digunakan dalam isi oleh pengguna. Iskandar, dkk (2020) melakukan beberapa hipotesis terhadap kata yang mengaruhi setiap dimensi. Dimana fitur ini akan menjadi *lexicon* untuk melakukan klasifikasi kepribadian MBTI. Parameter skenario yang digunakan adalah BoW sebagai pembobotan kata, Chi-Square sebagai fitur seleksi dan SMOTE. Hasil klasifikasi pengguna, didapatkan IE= 75.80%, NS = 55.52%, TF = 95.02%, dan JP = 88.26%. Penelitian Utami, dkk (2019) yang melakukan pendakatan *open-vocabulary analysis* untuk mengklasifikasan kepribadian DISC. Data yang didapatkan adalah 139 pengguna yang sudah divalidasi oleh pakar. Pembobotan terhadap sinonim kata juga dilakukan dalam penelitian ini, yaitu 0.85 untuk sinonim pertama, dan 0.35 untuk sinonim kedua. Hasil klasifikasi berdasarkan kata dengan *not stemmed-not weighted, stemmed-not weighted, not stemmed-weighted,* dan *stemmed-weighted*

key-word vocabularies, didapatkan akurasi sebesar 37,41%, 30,21%, 35,97%, dan 30,93%.

Penelitian ini akan melakukan klasifikasi kepribadian pengguna berdasarkan konsep MBTI menggunakan pendekatan yang hampir sama dengan Ong, dkk (2017). Perbedaan yang akan dilakukan dari fitur ekstraksi dan model pembelajaran mesin yang berbeda. Fitur yang akan digunakan mengadaptasi dari Fikry, dkk (2018) serta beberapa fitur tambahan lainnya dari peneliti. Seleksi fitur dengan pendekatan Chi-Square dan *balancing method* akan dilakukan pada penelitian. Model pembelajaran mesin yang akan digunakan adalah model yang sama dengan Lukito, dkk (2016) dan Claudy dkk, (2018) yaitu NBC dan K-NN dengan variasi parameter K yaitu 3, 5, 7, dan 9.

2.2. Keaslian Penelitian

Tabel 2.1. Matriks *literature review* dan posisi penelitian

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
1	Klasifikasi Dokumen Twitter Untuk Mengetahui Karakter Calon Karyawan Menggunakan Algoritme K-Nearest Neighbor (K-NN)	Claudy, Y. dkk, (J-PTIIK, 2018)	Mengimplementasikan metode K-NN untuk melakukan klasifikasi karakter calon karyawan dari tweet menjadi 4 kelompok besar sesuai konsep MBTI dengan data kuesioner	K-NN tetap dapat melakukan klasifikasi kepribadian atau karakter dengan baik dengan nilai akurasi sebear 66%	Peneliti: 1.hanya menggunakan satu parameter K yaitu 4 2. Tidak melakukan perbaikan kata tidak baku 3. Tidak membandingkan Pembobotan kata lainnya 4. Tidak membandingan model machine learning lainnya	Tindak Lanjut: 1. membandingkan parameter K 3,5,7 dan 9 2. Membandingkan pembobotan kata BoW dan TF-IDF.
2	Personality Prediction Based on Twitter Information in Bahasa Indonesia	Ong, dkk (CCSIS, 2017)	Membangun sistem klasifikasi kepribadian berdasarkan bahasa indonesia, membandingkan skenario yang berkontribusi terhadap peningkatan akurasi dan membandingan 2 model machine learning dalam proses uji	Evaluasi menggunakan cross Validation 10fold menunjukkan bahwa sistem prediksi kepribadian yang dibangun pada SVM berhasil mencapai akurasi rata-rata tertinggi 76,2310%, sedangkan XGBoost mencapai 97,9962%.	Penulis: 1. Karena hanya sekitar 35 ribu data tweet yang diambil dari 359 user, maka penulis menyaran penelitian selanjutnya untuk menambah data Peneliti: 1. Pada data terdapat Imblance Data schingga perlu dilakukan Sampling 2. pembobotan kata tidak membandingan dengan TF-IDF	Tindak lanjut: Membandingkan pembobotan kata BoW dan TF-IDF.

Tabel 2.1. Matriks *literature review* dan posisi penelitian (Lanjutan)

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
3	What You Say Define Who You Are? Word Exploration and Automatic Personality Detection	Iskandar, A.F., dkk (IJSTR, 2020)	Melakukan analis kata yang digunakan, serta melakukan klasifikasi berdasarkan kepribadian MBTI dengan *machine learning*	Hasil klasifikasi pengguna, didapatkan IE= 75.80%, NS = 55.52%, TF = 95.02% JP = 88.26%	Penulis: Dapat menambah ekstrasi fitur pada kata dalam melakukan klasifikasi. Peneliti: Tidak membandingkan pembobotan kata TF-IDF	Tindak Lanjut: 1. Membandingkan pembobotan kata BoW dan TF-IDF. 2. Menggunakan ekstraksi fitur seperti jumlah karakter, jumlah kata, dan lainnya.
4	"Ekstrover atau Introver: Klasifikasi Kepribadian Pengguna Twitter dengan Menggunakan Metode Support Vector Machine"	Fikry, M. & Yusra (SITEKIN, 2017)	Melakukan Klasifikasi Kepribadian Ekstrover dan Introvert dengan menggunakan SVM dari Twitter dengan menggunakan feature extraction sebanyak 16 dari tweets	Hasil pengujian terbaik dengan Model SVM, memperoleh akurasi 88,89%, peneliti juga membandingan dengan penelitian sebelumnya yaitu NBC dengan akurasi 83.33 %	Penulis: 1. jumlah data yang digunakan disarankan untuk diperbanyak 2. Menggunakan seleksi fitur penting pada data Peneliti: 1. Tidak menjelaskan secara detail praproses yang dilakukan. 2. Tidak melakukan pendekatan N-Gram dalam praproses data	Tindak lanjut: 1. Melakukan seleksi fitur dengan Chi-Square 2. Membandingkan pembobotan kata BoW dan TF-IDF.

Tabel 2.1. Matriks *literature review* dan posisi penelitian (Lanjutan)

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
5	Social Media Pengguna Personality Classification Using Computational Linguistic	Lukito, P. H., dkk (ICITEE, 2016)	Mengembangkan dan Membandingkan hasil Klasifikasi Kepribadian MBTI berbahasa Indonesia dengan menggunakan 3 pendekatan yaitu, Naive Bayes, Lexicon Based, dan Grammatical Rule	Akurasi Naive Bayes I-E sebesar 72,5%, S-N sebesar 60 %, T-F sebesar 61,2%, J-P sebesar 55,4%	Penulis: Menambahkan feature atau parameter lainya supaya dapat meningkatan akurasi Peneliti: 1.Penulis tidak membandingkan pembobotan kata 2. Model machine learning yang digunakan hanya 1 yaitu NB	Tindak Lanjut: 1. Membandingkan hasil klasifikasi Naïve Bayes dengan K-NN 2. Membandingkan pembobotan kata BoW dan TF-IDF. 3. Menggunakan ekstraksi fitur seperti jumlah karakter, jumlah kata, dan lainnya.
6.	Profiling analysis of DISC personality traits based on Twitter posts in Bahasa Indonesia.	Utami, E., dkk (Journal of King Saud University – Computer and Information Sciences, 2019)	Melakukan pendakatan *open-vocabulary analysis* untuk mengklasifikasan kepribadian DISC	Hasil Klasifikasi berdasarkan kata dengan not stemmed-not weighted = 37,41%, stemmed-not weighted = 30,21%, not stemmed-weighted =35,97% stemmed-weighted key-word vocabularies= 30,93%.	Peneliti: 1. menggunakan fitur *hashtags*, demografi, emoticon, *behaviour*, dan lokasi, sebagai fitur tambahan 2. dapat menggunakan Chi Square Feature Selection Penulis: 1. dapat menggunakan fitur linguistic seperti jumlah karakter, jumlah simbol dan lainnya. 2. terdapat imbalance class	Tindak Lanjut: 1. Menggunakan ekstraksi fitur seperti jumlah karakter, jumlah kata, dan lainnya. 2. Melakukan seleksi fitur dengan Chi-Square 3. Melakukan *Imbalance aprroach* jika terjadi *imbalance data*

2.3. Landasan Teori

Terdapat beberapa landasan teori yang dibutuhkan dalam penelitian ini, mulai dari landasan teori untuk preferensi kepribadian MBTI, penambangan kata, klasifikasi, model NBC dan K-NN.

2.3.1. Kepribadian MBTI

Salah satu preferensi kepribadian yang sudah dikenal luas oleh masyarakat adalah MBTI. Kepribadian MBTI pertama kali dikembangkan tahun 1962 oleh Katherine Cook Briggs dan Isabel Briggs Myers (Briggs & Myers, 1995). MBTI memiliki empat model faktor yang memungkinkan seseorang untuk menggambarkan diri mereka sendiri, seperti ISFP, INTJ, ENFJ, dan lainya. Menurut Briggs & Myers (1995), terdapat 4 dimensi preferensi yaitu: dimensi *Introvert-Extrovert* (IE) menggambarkan bagaimana melihat energi ke dalam atau ke luar, dimensi *Intuition-Sensing* (NS) menjelaskan bagaimana seseorang menerima informasi, dimensi *Thinking-Feeling* (TF) menggambarkan cara yang digunakan seseorang untuk membuat keputusan, dan *Judging-Perceiving (JP)* menggambarkan kecepatan seseorang mengambil keputusan.

Dimensi IE atau dimensi pemusatan perhatian dimana sikap introver cenderung akan asyik dengan dunia mereka sendiri. Mereka lebih suka melakukan apa yang mereka mau dengan sendiri seperti menonton, membaca, menulis, dan lainnya Berbeda sikap seorang ekstrover selalu melibatkan lingkungan dunia luar pada kesehariannya. Seseorang bertipe ekstrover akan suka bergaul, mudah berinteraksi dan berteman pada banyak orang. (Briggs & Myers, 1995).

Dimensi NS atau dimensi memahami informasi dari luar dimana sikap seorang *intuition* akan tertarik pada masa depan, makna tersirat, dan pola simbolik atau teoritis yang disarankan oleh wawasan. Mereka cenderung kreatif, inovatif, dan memiliki sejuta ide-ide yang unik. Sedangkan sikap seorang *sensing* lebih cendurung akan mengandalkan presepsi sehinga akan tertarik kepada apa yang nyata dan langsung yang dilihat oleh panca indera. Seorang *sensing* akan bersikat logika dan berdasarkan fakta yang dia temuakan pada pengalaman terdahulu (Briggs & Myers, 1995).

Dimensi TF atau dimensi menarik kesimpulan dan keputusan, dimana sikap seorang *thinking* akan secara rasional memutuskan melalui proses analisis logis dari sebab dan akibat. Mereka juga cenderung lebih berfokus pada prosedur atau standar yang telah ditetapkan sehingga konsisten dan berfokus pada tugas yang telah diberikan dan objektif. Sikap seorang *feeling* menggunakan penilaian pada perasaan sehingga akan bersikap berorientasi pada subjektif dan sangat terkesan berpihak. Mereka lebih mementingkan untuk menjaga hubungan dan keharmonisan (Briggs & Myers, 1995).

Dimensi JP atau dimensi pola hidup dimana sikap seorang *judging* akan menikmati bergerak cepat menuju keputusan dan menikmati pengorganisasian, perencanaan, dan penataan. Mereka bukan bersifat langsung mengakimi sesuatu, tetapi sesantiasa berpikir dan bertindak secara teratur sesuai dengan perencanaan awal. Sedangkan sikap orang *perceiving* akan menikmati rasa ingin tahu dan terbuka terhadap perubahan, lebih suka membiarkan opsi terbuka jika ada sesuatu yang lebih baik muncul. Mereka sangat beradaptasi terhadap perubahan yang

mendadak dan adaptif terhadap perubahan yang terjadi secara mendadak (Briggs & Myers, 1995).

2.3.2. Penambangan Teks

Penambangan teks merupakan bagian dari penambangan data, yang dapat didefinisikan sebagai proses mengekstraksi pengetahuan dari teks atau dokumen (Jo, 2019). Menurut Kantardzic (2020) menjelaskan bahwa bentuk paling dasar dari informasi yaitu teks dan hampir 80% dari informasi perusahaan termuat dalam dokumen teks, sehingga tantangan terkait penambangan teks menjadi isu saat ini. Penambangan teks dapat diyakini memiliki potensi komersial yang lebih tinggi daripada penambangan data tradisional dengan data yang terstruktur.

Penambangan teks dapat mengekstraksi bagian penting dalam dokumen teks, mengelompokkan dokumen, dan membuat ringkasan pada dokumen yang tidak mungkin dilakukan secara manual oleh manuasia (Jo, 2019). Beberapa tujuan dari penambangan teks, dijelaskan oleh Kantardzic, (2020) adalah memberikan gambaran umum terkait isi topik apa yang ada pada dokumen-dokumen, meningkatkan efesiensi dan efektivitas pencarian informasi pada dokumen, dan mendeteksi informasi atau dokumen yang duplikat. Menurut Jo (2019), perbedaan antara penambangan teks dengan sistem temu balik yaitu dari hasil *output*. Penambangan teks digunakan sebagai pengetahuan yang diperlukan dan digunakan secara langsung untuk membuat keputusan, sedangkan sistem temu balik adalah beberapa item data dari pengambilan informasi.

Sumber data pada penambangan teks dapat berupa teks pada suatu dokumen berita, jurnal, maupun surat (Jo, 2019). Pada sebuah paragraf pada suatu dokumen dapat berupa kombinasi dari kalimat yang diurutkan untuk menjaga konsistensi pada sub topik tertentu dengan tujuan untuk mencari kata-kata yang dapat mewakili dokumen-dokumen tersebut (Amanullah, 2017). Sehingga hasil dari penambangan teks tersebut didapatkan suatu pengetahuan yang digunakan pada kebijakan yang tepat pada suatu perusahan. Penambangan teks dapat dilakukan dengan klasifikasi atau melihat jumlah kata yang sering muncul atau *word cloud* (Mihuandayani, 2018).

2.4. Klasifikasi

Klasifikasi merupakan salah satu dari bidang machine learning, yang dimana kalasifikasi masuk kedalam *supervised learning* yaitu suatu pembelajaran yang terarah dimana data dipelajari oleh mesin sudah diberikan label (Putra, 2019). *Supervised learning* termasuk kedalam bagian dari *predictive data mining* yang dimana tujuannya untuk mengklasifikasikan item data menjadi salah satu dari beberapa kelas yang telah ditentukan (Kantardzic, 2019). Klasifikasi didefinisikan sebagai proses menetapkan kategori atau beberapa kategori di antara yang telah ditentukan untuk setiap item data. Klasifikasi dipandang sebagai kotak hitam yang input dan outputnya masing-masing merupakan item data dan kategorinya. Dengan menerapkan algoritma dari data sampel, model klasifikasi dibangun dalam berbagai bentuk simbolik, persamaan matematika, dan probabilitas (Jo, 2019).

Terdapat 2 jenis klasifikasi yang sering digunakan yaitu klasifikasi biner dan klasifikasi multi-kelas. Klasifikasi biner adalah klasifkasi dilakukan dengan hanya ada dua kategori (Putra 2019). Klasifikasi biner mengacu pada tugas klasifikasi paling sederhana di mana setiap item diklasifikasikan ke dalam salah satu dari dua kategori. Asumsi yang mendasari dalam klasifikasi biner adalah bahwa setiap item harus masuk ke salah satu dari dua kelas (Jo, 2019). Klasifikasi multi-kelas mengacu pada tugas klasifikasi, di mana setiap item diklasifikasikan ke dalam setidaknya satu dari tiga kategori (Putra 2019). Klasifikasi multi-kelas dapat didekomposisi menjadi tugas klasifikasi biner sebanyak kategori seperti diberikan empat kelas yaitu kelas 1, kelas 2, kelas 3, dan kelas 4 (Jo, 2019).

Terdapat algoritma pembelajaran kalsifkasi yang membedakan kelas, baik menggunakan klasifikasi biner maupun klasifikasi multi-kelas yang sering digunakan, dua diantaranya dalah Naïve Bayes dan K-NN (Jo, 2019), (Putra, 2019), (Kantardzic, 2020)

2.4.1. Naïve Bayes

Model Naïve Bayes adalah bentuk model klasifikasi data menggunakan metode probabilitas dan statistik (Kantardzic, 2020). Model probabilitas Naïve Bayes ini didasarkan pada teorema Bayes, dengan asumsi bahwa fitur-fitur dalam dataset saling independen, sehingga model Naive Bayes dapat diartikan sebagai model yang tidak memiliki aturan (Putra, 2019). Ada dua langkah proses pembelajaran dalam model ini, langkah pertama adalah mendefinisikan hubungan

kausal antara atribut dan langkah kedua adalah menghitung kemungkinannya (Jo 2019).

Teorema Bayes menjelaskan kemungkinan suatu peristiwa berdasarkan pengetahuan sebelumnya tentang kondisi yang mungkin terkait dengan peristiwa tersebut. Misalkan B adalah sampel data dan A adalah beberapa hipotesis, sehingga sampel data B milik kelas tertentu "k". P(A|B) adalah probabilitas yang dimiliki hipotesis A mengingat sampel data yang diamati hipotesis B. P(A|B) adalah *posterior probability* yang mewakili pada hipotesis B. Sebaliknya, P(A) adalah probabilitas A sebelumnya untuk setiap sampel, terlepas dari bagaimana data dalam sampel terlihat. P(A|B) didasarkan pada informasi yang lebih banyak daripada probabilitas sebelumnya P(A) (Dinov, 2018,) (Kantardzic, 2020). Secara sederhana, dapat melihat Persamaan (1) berikut:

$$Posterior\ Probability = \frac{likelihood\ x\ Prior\ Probability}{Marginal\ Likelihood}$$

(Dinov, 2018)

Dalam bentuk variabel menjadi:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \quad \text{..... Persamaan (1)}$$

Dimana:

$P(A|B)$ = probabilitas dari sample A pada data B disebut juga *posterior probability*.

$P(B|A)$ = probabilitas data B yang dimana A benar disebut juga *likelihood*

$P(A)$ = probabilitas hipotesis A adalah benar (terlepas dari data), disebut juga *Prior Probability*

$P(B)$ = probabilitas B (terlepas dari hipotesis), disebut juga *Marginal Likelihood*

Dari persamaan (1) dapat diperluas menjadi persamaan (2):

$$P(A = k|B) = \frac{P(B|A = k)\,P(A = k)}{P(B)}$$ ..... Persamaan (2)

Persamaan (2) dapat diartikan sebagai probabilitas untuk klasifikasi target dengan kelas k diberikan fitur matriks B, dan diberikan oleh probabilitas klasifikasi fitur matriks B diberikan kelas tertentu A dikali probabilitas pada kelas k (Dinov, 2018). Hasil dari probabilitas dari NBC berada pada range [0,1]. Jumlah likehood mungkin lebih dari 1, sehingga nilai likelihood perlu diubah menajdi bentuk probalitas dengan menggunakan teknik *softmax* (Putra, 2019)

Dalam penerapan penambangan teks, Misalkan teks pada suatu dokumen berkata "Kota Medan panas"dan ingin tahu apakah teks tersebut negatif atau positif sehingga dengan menggunalan teorema Bayes, didapatkan Persamaan (3).

$$P(negatif|\text{"Kota Medan panas"}) = \frac{P(\text{"Kota Medan panas"}|negatif)\,P(negatif)}{P(\text{"Kota Medan panas"})}$$

..... Persamaan (3)
(Wilmott, 2019)

Kemudahan menggunakan model NBC yaitu sederhana dan mudah dimengerti, perhitungannya cepat dan efisien saat waktu pelatihan model, meningkatkan kinerja klasifikasi dengan menghilangakn atribut yang tidak sesuai,

bisa digunakan klasifikasi biner dan klasifikasi multi-kelas, dan NBC banyak digunakan untuk klasifikasi teks (Jo 2019), (Putra, 2019) (Wilmott, 2019).

2.4.2. K-Nearest Neighbor

Model K-Nearest Neighbor merupakan model yang melakukan klasifikasi terhadap objek berdasarkan koleksi suara terbanyak atau *voting* berdasarkan koleksi yang diberikan Wilmott (2019). K-NN disebut juga algoritma yang malas karena tidak mempelajari data, melainkan hanya mengingat data yang sudah ada (Putra, 2019). Algoritma klasifikasi ini juga bekerja dengan awal menentukan nilai parameter K selanjutnya menentukan jarak pada persamaan (4) antara setiap data pengujian dan pelatihan yang sudah tetapkan, mengurutkan berdasarkan jarak K terdekat, dan menggunakan mayoritas pada nilai parameter K dari kategori tetangga terdekat sebagai nilai prediksi dari klasifikasi sampel pengujian (Kantardric, 2019),(Zuhdi dkk, 2019).

$$D(X,Y) = \sqrt{\sum_{i=1}^{n} d_i(x_i, y_i)} \quad \text{..... Persamaan (4)}$$

Persamaan (4) menampilkan rumus perhitungan jarak yang digunakan pada K-NN. Kesamaan didefinisikan menurut metrik jarak antara dua titik data. Sebelum melakukan perhitungan jarak, terlebih dahulu harus menentukan data latih dan data uji. Metode jarak yang dapat digunakan adalah metode jarak Manhattan, Minkowski, Hamming dan Euclidean (Wilmott, 2019). Salah satu yang sering

digunakan adalah metode jarak Euclidean. Perhitungan nilai jarak terdekat dapat mengunakan jarak Euclidean yang dapat dilihat pada Persamaan (5).

$$D(a,b) = \sqrt{\sum_{k=1}^{d}(a_k - b_k)^2} \quad \text{.....Persamaan (5)}$$

Dimana:

$D(a,b)$ = scalar dari dua buah vector data a dan b

(Claudy dkk, 2018)

Selanjutnya, sebelum mencari jarak data ke tetangga adalah menentukan nilai parameter K. Klasifikasi K-NN dapat berpengaruh terhadap pemilihan nilai paramater K. Menurut Kantardric (2020), nilai parameter K pada K-NN sering dipilih berdasarkan pengalaman *try-error* atau pengetahuan tentang masalah klasifikasi yang dihadapi. Beberapa ide yang bisa yaitu, memilih nilai terlalu kecil akan mengakibatkan *noise* terhadap hasilnya, nilai terlalu lebih besar dari K akan bias dan nilai parameter K harus ganjil (Willmot, 2019). Model K-NN pada penelitian ini akan dipilih parameter K bernilai ganjil yaitu 3,5,7 dan 9.

Pada penambangan teks, teks pada suatu dokumen dianggap sebagai item data yang harus dikonversi ke dalam bentuk vektor numerik (Jo, 2019). Teks tersebut, kemudian akan dihitung jaraknya pada vektor dokumen yang lain, baik pada kelas yang sama ataupun yang berbeda.

Kelebihan menggunakan model K-NN menurut Putra (2019) dan Wilmott (2019) yaitu mudah diimplementasikan karena besifat malas yaitu hanya mencari jarak terdekat tanpa membuat model probabilitas seperti Naïve Bayes, dapat menangani kasus multi-kelas, dan efektif jika terdapat jumlah data cukup banyak.

BAB III

METODE PENELITIAN

3.1. Dataset

Dataset yang digunakan pada penelitian ini adalah data yang didapatkan dari Iskandar, dkk (2020). *Dataset* tersebutkan merupakan data yang didapatkan dari media sosial Twitter, yang terdiri label kepribadian MBTI dan cuitan sebanyak 281 pengguna. Persentase data pengguna berdasarkan kepribadian MBTI dapat dilihat pada Gambar 3.1. Sampel *dataset* yang akan digunakan dapat dilihat pada Lampiran 1.

Tipe Kepribadian	Jumlah Pengguna
INFP	50
ENFP	49
INFJ	37
ENFJ	22
INTJ	17
ISTJ	10
INTP	10
ISFJ	10
ESFP	8
ESFJ	8
ENTJ	8
ENTP	8
ISTJ	6
ISFJ	6
ESTP	3
ESTP	3

Gambar 3.1. Persentase Pengguna berdasarkan Kepribadian MBTI

Data didominasi oleh pengguna bertipe INFP sebesar 50 pengguna, ENFP sebesar 49 pengguna, INFJ sebesar 37 pengguna dan ENFJ sebesar 22 pengguna. Jumlah cuitan didominasi oleh tipe kepribadian INFP sebanyak 3060 cuitan, tipe ENFP sebanyak 2677 cuitan, tipe INTP sebanyak 1526 cuitan dan tipe INFJ sebanyak 1470 cuitan. Distribusi pengguna berdasarkan dimensi IE yaitu I sebesar

161 pengguna (57,30%) sedangkan E sebesar 120 pengguna (42,70%). Selanjutnya, untuk dimensi NS yaitu N sebesar 206 pengguna (73,31%) sedangkan S sebesar 75 pengguna (26,69%). Untuk dimensi TF yaitu T sebesar 68 pengguna (24,20%) sedangkan F sebesar 213 pengguna (75,80%). Terakhir dimensi JP yaitu J sebesar 122 pengguna (43,42%) sedangkan P sebesar 159 pengguna (56,58%).

Isi cuitan pada dataset yang digunakan, sudah dilakukan seleksi untuk memastikan bahwa akun pengguna tersebut adalah asli dengan mengacu pada:

a. Melakukan postingan tweet seperti "Open Order", "Jual ..." atau kata kunci lainnya untuk mempromosikan sesuatu atau berjualan.
b. Penggunaan tagar yang tidak jelas, seperti #murmer, #hargamati, dan lainnya.

3.2. *Preprocessing* Data

Pada bagian ini akan dilakukan beberapa tahapan, yaitu menyeleksi tweet yang menggunakan bahasa Indonesia, mengubah teks menjadi huruf kecil atau *case folding*, menghapus link, fitur twitter seperti #, RT, @, angka dan simbol, normalisasi kata slang atau kata lainnya, *stemming* dan tokenisasi (Utami, dkk, 2019).

Gambar 3.2. Tahapan *Preprocessing*

Tahapan *preprocessing* pada Gambar 3.2. dilakukan dengan bahasa pemograman Python dengan menggunakan perangkat lunak Anaconda. *Case Folding* merupakan tahapan yang mengubah kata-kata pada teks menjadi huruf kecil. *Number, Link, and Puncuation Removal* merupakan tahapan untuk menghapus angka, URL atau link serta simbol-simbol yang tidak dibutuhkan. *Slang Normalization* merupakan tahapan untuk memperbaiki kata slang menjadi kata normal yang sesuai kamus KBBI. Proses *slang normalization* dilakukan secara berulang-ulang. Jika masih terdapat kata yang tidak terdapat pada KBBI, maka kata tersebut tetap akan dilakukan proses selanjutnya dikarenakan bisa menjadi fitur unik pada analisis kepribadian MBTI. *Stemming* merupakan tahapan untuk melakukan konversi kata menjadi kata dasar berdasarkan imbuhan ("berbaik" menjadi "berbaik"). Terakhir, *Lemmatization* merupakan tahapan melakukan konversi kata berdasarkan morfologi kata tersebut seperti ("lebih baik" menjadi "baik"). *Source Code* yang menggunakan dengan bahasa pemograman Python yang digunakan pada tahapan preprocessing dapat dilihat pada Lampiran 2. Sampel perubahan data pada tahapan preprocessing data dapat dilihat pada Tabel 3.1.

Tabel 3.1. Sampel Tahapan *Preprocessing Data*

Kode Pengguna	Tipe MBTI	Text	*Case Folding*	*Number, Link, Punctuation Removal*	*Slang Normalization*	*Text Stemming*	*Text Lemmatization*
USER_0	ESFP	cape bat dah idup	cape bat dah idup	cape bat dah idup	capek banget deh hidup	capek banget deh hidup	capek banget deh hidup
USER_1	ENFP	Baru pulang reunian bareng temen" SD :D gilaa seruu bangett :D	baru pulang reunian bareng temen" sd :d gilaa seruu bangett :d	baru pulang reunian bareng temen sd d gilaa seruu bangett d	baru pulang reunian bareng teman sd ada gila seru banget ada	baru pulang reuni bareng teman sd ada gila seru banget ada	pulang reuni bareng teman sd gila seru banget
USER_2	ISFJ	Yaudah ntar makan2 aja kita di P2 wkwkwk	yaudah ntar makan2 aja kita di p2 wkwkwk	yaudah ntar makan aja kita di p wkwkwk	audah antar makan saja kita di px wkwkwk	yaudah entar makan saja kita di p wkwkwk	ya sudah entar makan
USER_2	ISFJ	Alhamdulillah udah sabar dr kemarin kok mas wkwkwk	alhamdulillah udah sabar dr kemarin kok mas wkwkwk	alhamdulillah udah sabar dr kemarin kok mas wkwkwk	alhamdulillah sudah sabar dari kemarin kok mas wkwkwk	alhamdulillah sudah sabar dari kemarin kok mas wkwkwk	alhamdulillah sabar kemarin mas
USER_2	ISFJ	Yaudah selow yak Ditunggu foto2 erichan nya hehehe	yauduh selow yak ditunggu foto2 erichan nya hehehe	yaudah selow yak ditunggu foto erichan nya hehehe	audah selow ya ditunggu foto herichan nya hehehe	yaudah selow ya tunggu foto erichan nya hehehe	ya sudah selow ya tunggu foto erichan nya
USER_2	ISFJ	Yuk ntar makan2 kita di P2	yuk ntar makan2 kita di p2	yuk ntar makan kita di p	yuk antar makan kita di px	yuk entar makan kita di p	yuk entar makan
USER_3	ISTP	Semoga cepet sembuh ya untuk sepupu kamu	semoga cepet sembuh ya untuk sepupu kamu	semoga cepet sembuh ya untuk sepupu kamu	semoga cepat sembuh ya untuk sepupu kamu	moga cepat sembuh ya untuk sepupu kamu	semoga cepat sembuh ya sepupu

3.3. Ekstraksi Fitur

Isi cuitan pada Twitter merupakan fitur yang menghubungkan antara kata yang digunakan dan kepribadian pengguna. Dalam pendekatan linguistik, kata dari cuitan akan dilakukan pendekatan *Bag of Word* (BoW) atau frekuensi kata yang muncul dan pendekatan TF-IDF untuk dilakukan ekstraksi fitur sebelum dilakukan pelatihan model (Willy, dkk 2019). Pendekatan TF-IDF dapat dilihat pada Persamaan (6):

$$W_{ij} = TF_{ij} \times \log\left(\frac{D_i}{df_i}\right) \quad \text{..... Persamaan (6)}$$

(Willy, dkk 2019)

Dimana

TF_{ij} = jumlah kata ke-i yang muncul pada dokumen ke-j.

D_i = jumlah dokumen atau data,

df_i = jumlah dokumen yang mengandung kata ke-i

Dataset juga akan dilakukan ekstraksi fitur sebanyak 20 fitur. Detail ekstraksi 20 fitur dapat dilihat pada Tabel 3.2

Tabel 3.2. Ekstraksi Fitur

No	Kode	Keterangan
1	TW01	Jumlah cuitan yang disukai
2	TW02	Jumlah cuitan yang dibalas
3	TW03	Jumlag cuitan yang dipost ulang
4	TXT01	Jumlah karakter unik
5	TXT02	Jumlah kata pada teks cuitan
6	TXT03	Jumlah kalimat
7	TXT04	Jumlah url/link

Tabel 3.2. Ekstraksi Fitur (Lanjutan)

No	Kode	Keterangan
8	TXT05	Jumlah media
9	TXT06	Jumlah pertanyaan
10	TXT07	Jumlah kalimat perintah
11	TXT08	Jumlah tagar yang digunakan
12	TXT09	Jumlah postingan dipost ulang
13	TXT010	Jumlah orang yang di-*mention*
14	TXT011	Jumlah *quotes*
15	USR01	Jumlah *follower*
16	USR02	Jumlah *following*
17	USR03	Jumlah cuitan yang disukai pengguna
18	USR04	Jumlah cuitan yang mengandung foto atau video
19	USR05	Jumlah karakter pada bio pengguna
20	USR06	Jumlah kata pada bio pengguna

Pada Tabel 3.2. terdapat 20 ekstraksi fitur yang terdiri dari 3 fitur spesial dari Twitter dengan awalan kode TW, 11 fitur untuk dari ekstrasi teks pada tweet dengan awalan kode TXT dan 6 fitur dari ekstraksi biografi pengguna dengan awalan kode USR. Ekstraksi fitur BoW dan TF-IDF serta ekstraksi fitur dilakukan menggunakan bahasa pemograman Python dengan library dari Sklearn dan Regex. Source Code dari Ekstraksi Fitur ini dapat dilihat pada Lampiran 2.

3.4. Seleksi Fitur

Seleksi fitur adalah bagian penting dari pemrosesan teks. Pemilihan kata yang unik dalam seleksi fitur akan dapat mengoptimalkan kinerja dari model, sehingga dapat meminimalkan fitur yang tidak penting dalam data (Sulistiani, 2017). Seleksi fitur akan berdampak positif yang ditandai dengan peningkatan nilai akurasi (Pratama, dkk, 2019). Seleksi fitur yang akan digunakan adalah Chi Square.

Seleksi fitur hanya dilakukan terhadap fitur yang dibentuk BoW dan TF-IDF. Pendekatan Chi Square dapat dilihat pada Persamaan (7) berikut ini:

$$\chi^2(D,t,c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}}$$

.... Persamaan (7)

Dimana

N adalah frekuensi yang diamati dalam dan *E* frekuensi yang diharapkan

t mengambil nilai 1 jika dokumen berisi istilah t dan 0 sebaliknya

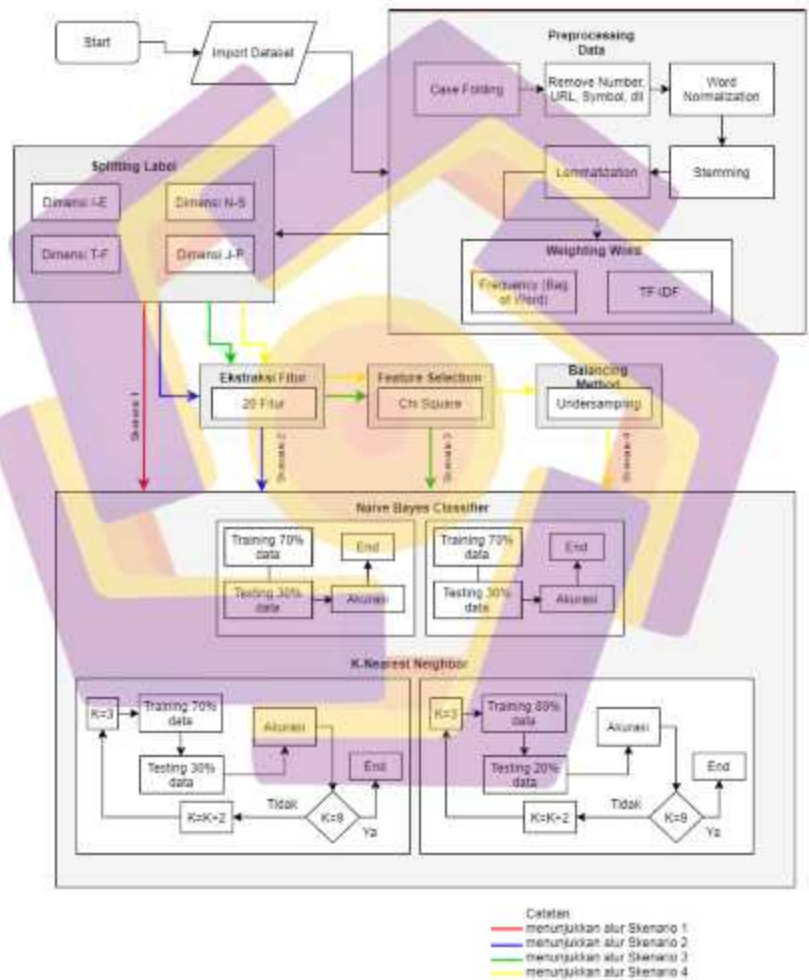
c mengambil nilai 1 jika dokumen ada di kelas c dan 0 sebaliknya

3.5. Evaluasi Model

Evaluasi model NBC dan K-NN dilakukan dengan akurasi. Akurasi adalah ukuran kinerja dengan rasio pengamatan yang diprediksi dengan benar terhadap pengamatan total (Amanullah, dkk, 2019). Penelitian ini berfokus kepada model NBC yang digunakan oleh Lukito (2016) dengan akurasi antara 55,4% - 72,5% dan model K-NN yang dilakukan oleh Claudy (2018) yang mendapatkan akurasi sebesar 66%, maka penelitian berfokus kepada peningkatan akurasi, sehingga dengan mengimplementasikan fitur ekstraksi dan seleksi fitur Chi Square serta *undersampling method* yang diharapkan hasil akurasi yang didapatkan lebih baik dari penelitian sebelumnya.

3.6. Alur Penelitian

Alur penelitian ini dapat dilihat pada Gambar 3.3, dimulai dari pengumpulan data, *preprocessing* data hingga pelatihan model dan pengujian model sehingga didapatkan hasil akurasi.



Gambar 3.3. Alur Penelitian

Alur penelitian ini dimulai dengan *import dataset*, lalu akan dilakukan *preprocessing data* seperti *case folding*, menghapus angka, link, hingga pembobotan kata BoW dan TF-IDF. Selanjutnya proses akan dibentuknya fitur ekstraksi sebanyak 20 fitur. Proses pemisahan label dilakukan untuk memperkecil kelas pada label, sehingga menjadi 8 label. Proses pemisahan data dilakukan sebanyak 2 kali yaitu 70:30 dan 80:20, proporsi itu dipilih karena merupakan angka yang ideal. Terakhir, proses pelatihan dan pengujian model NBC dan K-NN untuk mendapatkan akurasi.

3.6.1. Dasar Skenario

Terdapat 4 skenario yang akan dilakukan yaitu skenario dasar yaitu hanya fitur BoW dan TF-IDF, skenario dengan penambahan esktraksi fitur, skenario dengan seleksi fitur, dan sekenario dengan *balancing data*. Keempat skenario akan tetap di uji coba dengan model NBC dan K-NN dengan K yaitu 3, 5, 7 dan 9. Hasil keluaran yang diharapkan adalah akurasi setiap model dan skenario.

Skenario ke-1 merupakan adaptasi dari skenario yang dilakukan oleh Lukito, dkk (2016), Claudy,dkk (2018), dan Utami,dkk (2019). Pada penelitian Claudy, dkk (2018) menggunakan model K-NN mendapatkan akurasi 66%, Lukito, dkk (2016) menggunakan model Naïve Bayes mendapatkan akurasi IE, NS, TF dan JP sebesar 72,5% ;60%; 61,2%; 55,4% dan Utami, dkk (2019) menggunakan SVM 37,41%. Kata-kata yang sudah diekstrasi dengan menjadi fitur BoW dan TF-IDF akan dilakukan pelatihan model dengan partisi data 80:20 dan 70:30, serta melakukan perbandingan antara model NBC dan K-NN. Berdasarkan referensi

akurasi tertinggi diatas, didapatkan acuan akurasi pada skenario 66% - 72,5% pada setiap dimensi pada MBTI.

Skenario ke-2 merupakan skenario yang diadaptasi dari Fikry, dkk (2018). Fitur ekstasi yang digunakan pada Fikry,dkk (2018) yaitu jumlah cuitan, jumlah tautan, jumlah tagar, jumlah cuitan disukai, jumlah cuitan yang disukai, jumlah *mention*, jumlah unik pengguna di-*mention*, jumlah pengikut, jumlah mengikuti, keaktifan, jumlah cuitan diawali RT atau @, jumlah balasan cuitan, jumlah kata di profil, rata-rata jumlah kata, rata-rata jumlah karakter, jumlah emoticon dan emoji, dan jumlah foto atau video. Pengujian skenario dari Fikry,dkk (2018) mendapatkan akurasi sebesar 88,89% dengan menggunakan model SVM. Pada penelitian ini melakukan kombinasi dari kedua fitur ekstrasi referensi diatas, sehingga didapatkan fitur ekstrasi sebanyak 20 yang dapat dilihat Tabel 3.1. Berdasarkan referensi skenario, maka acuan akurasi skenario 72,5% - 80% pada setiap dimensi MBTI.

Skenario ke-3 merupakan skenario yang diadaptasi dari Ong,dkk (2017) dengan menggunakan fitur ekstrasi yang digunakan yaitu fitur ekstasi yaitu jumlah tweet, jumlah pengikut, jumlah mengikuti, jumlah *favorite*, jumlah retweet jumlah diretweet, jumlah quote dari cuitan, jumlah *mention*, jumlah membalas, jumlah tagar, jumlah link dan jumlah perbedaan waktu setiap cuitan, dan fitur seleksi dari LDA dengan SVM mendapatkan akurasi sebesar 76,23%. Pada skenario ke-3 pada penelitin ini, akan melakukan pengujian pada fitur ekstrasi BoW atau TF-IDF dikombinasi dengan 20 fitur ekstrasi dan fitur seleksi Chi-square. Berdasarkan referensi diatas, acuan akurasi yang diharapkan pada skenario ini antara 77% – 85%

Skenario 4 merupakan skenario yang diadaptasi dari skenario Iskandar, dkk (2020). Pada skenario ini menggunakan menggunakan seleksi fitur Chi-Square dan sampling SMOTE. Pengujian skenario tersebut mendapatkan akurasi pengguna, yaitu IE=75,80%, NS=55,52%, TF=95,02% dan JP=88,26%. Pada penelitian sebagai akan menggunakan parameter yang yang berbeda dimana seleksi fitur Chi-Square, fitur ekstraksi dan model NBC dan K-NN. Referensi acuan akurasi yang diharapkan pada skenario ini yaitu 80%-95%.

3.6.1. Skenario Pembagian Data

Pembagian antara data latih dan data uji menjadi faktor dalam evaluasi perfomansi klasifikasi. Data latih digunakan untuk melatih model agar representasi terhadap data, sedangkan data uji digunakan untuk mengevaluasi model tersebut. Pada penelitian Fikry,dkk (2018) dan Iskandar, dkk (2020) melakukan proporsi data 70:30 dan 80:20, yang dimana mendapatakn akurasi yang cukup baik.

Pertimbangan kelebihan dan kekurangan menggunakan proporsi data 70:30 dan 80:20 adalah data latih dapat merepresentasikan data, evaluasi model yang cukup cepat, dapat mengambarkan distribusi yang sama pada data latih dan data uji, dan mendapatkan akurasi terbaik (Adi, dkk 2018), (Baradwaj, dkk 2018). Kekurangan menggunakan proporsi data 70:30 dan 80:20 adalah model belum dapat merepresentasikan data sehingga bisa terjadinya *underfitting* dan *overfitting* data jika datanya tidak seimbang dan komputasi yang dilakukan cukup lama dalam melatih data, terutama pada model K-NN (Claudy, dkk 2018), (Iskandar, dkk 2020)

BAB IV

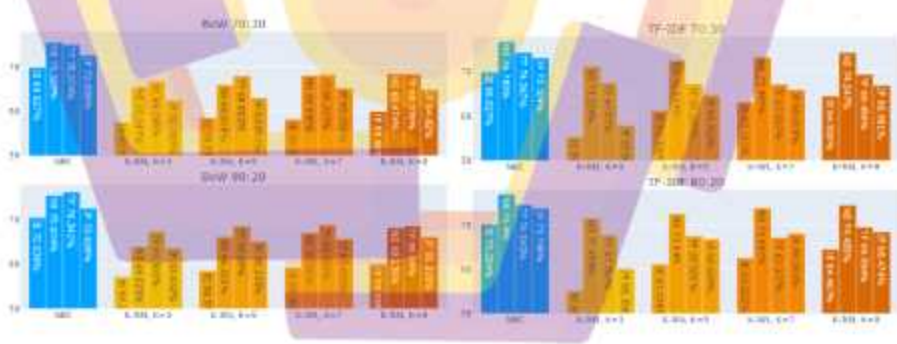
HASIL PENELITIAN DAN PEMBAHASAN

4.1. Hasil Klasifikasi

Pada klasifikasi kepribadian MBTI telah dilakukan empat skenario pada empat dimensi MBTI. Setiap skenario menggunakan parameter pembobotan kata (BoW dan TF-IDF), model (NBC dan K-NN), serta ukuran data (70:30 dan 80:20).

4.1.1. Hasil Klasifikasi Skenario Pertama

Skenario pertama merupakan skenario dasar pada penelitian ini. Skenario ini melakukan ektraksi kata-kata dengan BoW dan TF-IDF, kemudian dilakukan klasifikasi dengan model NBC dan K-NN. Hasil klasifikasi skenario pertama dapat dilihat Gambar 4.1.



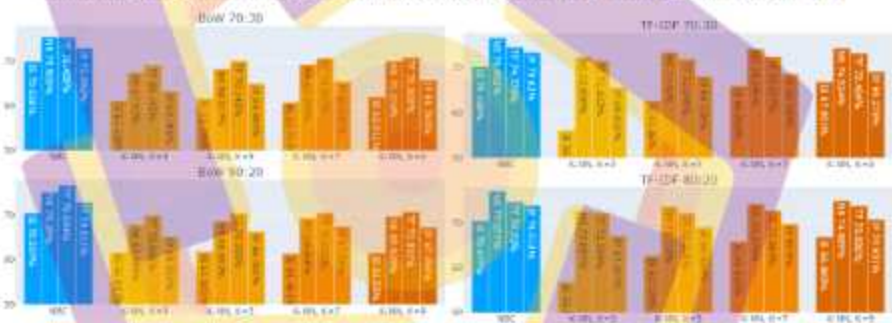
Gambar 4.1. Hasil Klasifikasi Skenario Pertama

Akurasi terbaik model NBC pada skenario pertama untuk dimensi IE adalah 70,536% (BoW dan 80:20), untuk dimensi NS adalah 76,9% (TF-IDF dan 80:20), untuk dimensi TF adalah 76,341% (BoW dan 80:20), dan untuk dimensi JP adalah 73,748 % (TF-IDF dan 80:20). Akurasi terbaik pada model K-NN pada skenario

pertama untuk dimensi IE adalah 64,506% (K=9, TF-IDF dan 70:30), untuk dimensi NS adalah 74,455% (K=9, TF-IDF dan 80:20), untuk dimensi TF adalah 69,534% (K=9, TF-IDF dan 80:20), dan untuk dimensi JP adalah 68,474% (K=9, TF-IDF dan 80:20).

4.1.2. Hasil Klasifikasi Skenario Kedua

Skenario kedua merupakan skenario klasifikasi yang dilakukan tidak jauh berbeda dengan skenario pertama, perbedaan yang dapat dilihat dari penambahan ekstraksi fitur pada Tabel 3.2. Hasil skenario kedua dapat dilihat pada Gambar 4.2.



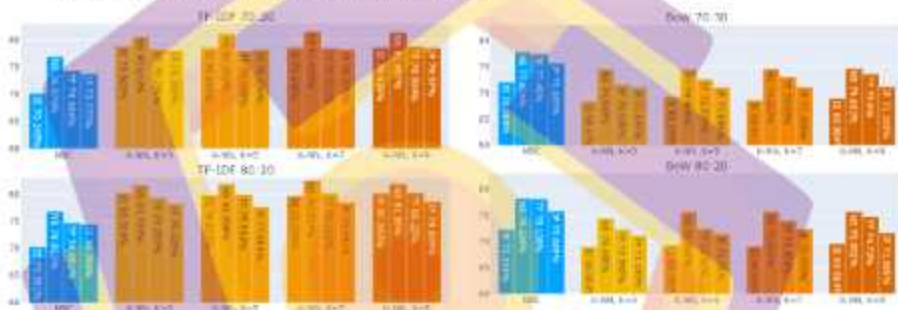
Gambar 4.2. Hasil Klasifikasi Skenario Kedua

Akurasi terbaik pada model NBC pada skenario kedua untuk dimensi IE adalah 70,507% (BoW dan 80:20), untuk dimensi NS adalah 77,077% (TF-IDF dan 80:20), untuk dimensi TF adalah 76,694% (BoW dan80:20) dan untuk dimensi JP adalah 74,013% (BoW dan80:20). Akurasi terbaik pada model K-NN di skenario kedua untuk dimensi IE adalah 67,001% (K=9, TF-IDF dan 70:30), untuk dimensi NS adalah 74,985% (K=9, TF-IDF dan 80:20), untuk dimensi TF adalah 73,836%

(K=9, TF-IDF dan 80:20) dan untuk dimensi JP adalah 70,831% (K=9, TF-IDF dan 80:20).

4.1.3. Hasil Klasifikasi Skenario Ketiga

Skenario ketiga merupakan skenario yang mengurangi dimensi dari fitur pada skenario kedua dengan mengguanakan seleksi fitur Chi-square. Hasil dari skenario ketiga dapat dilihat pada Gambar 4.3.

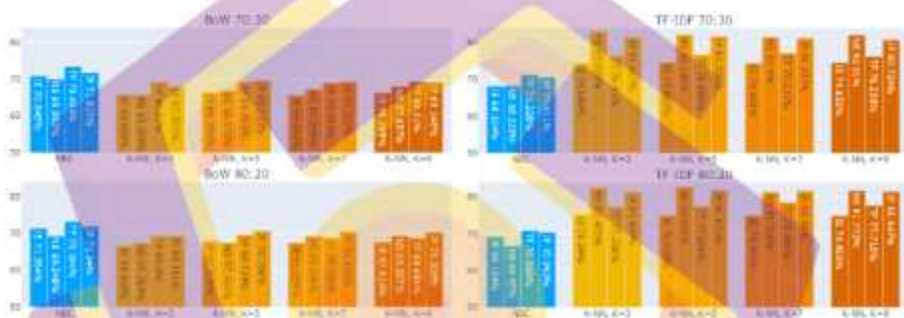


Gambar 4.3. Hasil Klasifikasi Skenario Ketiga

Akurasi terbaik pada model NBC pada skenario ketiga untuk dimensi IE adalah 72,334% (BoW dan 80:20), untuk dimensi NS adalah 78,226% (BoW dan 80:20), untuk dimensi TF adalah 78,108% (BoW dan 80:20) dan untuk dimensi JP adalah 75,869% (BoW dan 80:20). Akurasi terbaik pada model K-NN pada skenario ketiga untuk dimensi IE adalah 80,082% (K=9, TF-IDF dan 80:20), untuk dimensi NS adalah 82,557% (K=7, TF-IDF dan 80:20), untuk dimensi TF adalah 80,230% (K=9, TF-IDF dan 80:20) dan untuk dimensi JP adalah 78,609% (K=9, TF-IDF dan 80:20).

4.1.4. Hasil Klasifikasi Skenario Keempat

Pada skenario keempat akan dilakukan keseimbangan data pada setiap dimensi. Metode keseimbangan data yang digunakan penelitian adalah *undersampling* dengan menyetarakan data berdasarkan kelas dengan jumlah terendah dari kelas lainnya. Metode *undersampling* menggunakan fitur yang sama dengan skenario ketiga. Hasil dari skenario keempat dapat dilihat pada Gambar 4.4.



Gambar 4.4. Hasil Klasifikasi Skenario Keempat

Akurasi terbaik pada model NBC pada skenario keempat untuk dimensi IE adalah 71,364% (BoW dan 80:20), untuk dimensi NS adalah 69,993% (BoW dan70:30), untuk dimensi TF adalah 73,404% (BoW dan70:30) dan untuk dimensi JP adalah 71,940% (BoW dan80:20). Akurasi terbaik pada model K-NN pada skenario keempat untuk dimensi IE adalah 75,290% (K=3, TF-IDF dan 80:20), untuk dimensi NS adalah 83,028% (K=3, TF-IDF dan 70:30) untuk dimensi TF adalah 78,489% (K=7, TF-IDF dan 80:20), dan untuk dimensi JP adalah 81,871% (K=5, TF-IDF dan 80:20).

4.1.5. Akurasi Terbaik Setiap Skenario

Rangkuman akurasi hasil klasifikasi terbaik setiap dimensi dan skenario dapat dilihat pada Tabel 4.1.

Tabel 4.1. Akurasi Terbaik Setiap Dimensi

Skenario	NBC				K-NN			
	IE	NS	TF	JP	IE	NS	TF	JP
1	70,536%	76,900%	76,341%	73,748%	70,536%	76,900%	76,341%	73,748%
2	70,507%	77,077%	76,694%	74,013%	70,507%	77,077%	76,694%	74,013%
3	72,334%	78,226%	78,108%	75,869%	72,334%	78,226%	78,108%	75,869%
4	71,364%	69,993%	73,404%	71,940%	71,364%	69,993%	73,404%	71,940%

Penjelasan parameter hasil akurasi pada Tabel 4.1. sebagai berikut:

1. Akurasi terbaik pada dimensi IE didapatkan pada skenario ketiga sebesar 80,082% dengan model K-NN K=9 dan paramater TF-IDF dan 80:20.
2. Akurasi terbaik untuk dimensi NS didapatkan pada skenario keempat sebesar 83.028% dengan model K-NN K=3, parameter TF-IDF dan 70:30.
3. Akurasi terbaik pada dimensi TF didapatkan pada skenario ketiga sebesar 80,230% dengan model K-NN K=9 dan parameter TF-IDF 80:20.
4. Akurasi terbaik dimensi dimensi JP didapatkan pada skenario keempat sebesar 81,871 % dengan model K-NN K=5, parameter TF-IDF dan 80:20.

Model klasifikasi mendapatkan akurasi terbaik pada model K-NN dengan parameter pembobotan kata TF-IDF dan pembagian data 80:20. Parameter K pada model K-NN yang terbaik ada K=9, dimana parameter ini mendapatkan 68,75% atau 11 dari 16 klasifikasi yang sudah dilakukan. Pembobotan kata TF-IDF mendapatkan akurasi terbaik 12 klasifikasi, sedangkan BoW mendapatkan 4 klasifikasi. Terakhir, proporsi data latih dan uji terdapat perbedaan signifikan antara

70:30 dan 80:20, dimana 70:30 mendapatkan akurasi terbaik sebanyak 1 klasifikasi, sedangkan 80:20 sebanyak 15 klasifikasi.

4.1.6. Klasifikasi Pengguna

Klasifikasi pengguna merupakan klasifikasi yang dilakukan untuk mengetahui kepribadian MBTI berdasarkan pengujian skenario dengan model dan paramter terbaik dari setiap dimensi. Parameter model dan hasil klasifikasi pengguna dapat dilihat pada Tabel 4.2.

Tabel 4.2. Parameter dan Hasil Klasifikasi Pengguna

Dimensi	Model	Bobot Kata	Ekstraksi Fitur	Fitur Seleksi	Data Balancing	Akurasi Cross Validation (STD)	Akurasi Klasifikasi Pengguna
IE	K-NN, K=9	TF-IDF	V	V	X	79,505% (0.5)	95,018%
NS	K-NN, K=3	TF-IDF	V	V	V	82,138% (0.6)	62,993%
TF	KN-N, K=9	TF-IDF	V	V	X	79,499% (0.7)	84,178%
JP	K-NN, K=5	TF-IDF	V	V	V	81,973% (0.8)	90,747%

Pada dimensi IE, parameter yang digunakan untuk model K-NN yaitu K=9, TF-IDF, ekstraksi fitur, dan fitur seleksi dengan akurasi sebesar 79,505% dan akurasi klasifikasi pengguna sebesar 95,018%. Pada dimensi NS, parameter yang digunakan untuk model K-NN yaitu K=3, TF-IDF, ekstraksi fitur, fitur seleksi Chi-Square dan *undersampling* dengan akurasi sebesar 82,138% dan akurasi klasifikasi pengguna sebesar 62,993%. Pada dimensi TF, parameter yang digunakan untuk model K-NN yaitu K=9, TF-IDF, ekstraksi fitur, dan fitur seleksi dengan akurasi sebesar 79,499 % dan akurasi klasifikasi pengguna sebesar 84,178%. Terakhir, dimensi JP, parameter yang digunakan untuk model K-NN yaitu K=5, TF-IDF,

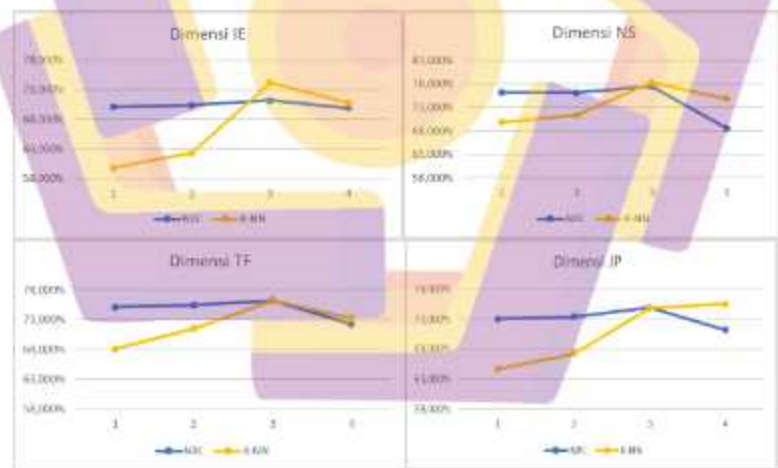
ekstraksi fitur, fitur seleksi, dan *undersampling* dengan akurasi sebesar 81,973 % dan akurasi klasifikasi pengguna sebesar 90,747%.

4.2. Pembahasan

Bagian ini akan membahas perbandingan hasil klasifikasi antara model NBC dengan K-NN, pengaruh ekstraksi fitur dan seleksi fitur dan perbandingan hasil penelitian dengan penelitian rujukan

4.2.1. Perbandingan NBC dengan K-NN

Perbandingan hasil performansi NBC dan K-NN dari setiap skenario dan dimensi dilakukan berdasarkan rata-rata dari setiap klasifikasi. Rata-rata hasil klasifikasi berdasarkan dimensi dan skenario dapat dilihat pada Gambar 4.5.



Gambar 4.5. Rata-rata Akurasi NBC dan K-NN setiap dimensi

Pada skenario pertama, dimana skenario pertama hanya melakukan ekstraksi teks ke dalam pembobotan kata BoW atau TF-IDF. Model NBC melakukan klasifikasi cukup baik daripada K-NN pada skenario pertama. Pada skenario kedua melakukan penambahan ekstraksi fitur sebanyak 20 fitur, model NBC melakukan klasifikasi cukup baik daripada K-NN pada skenario pertama. Rata-rata kenaikan akurasi yang didapatkan untuk NBC sebesar 0,220% sedangkan untuk K-NN sebesar 2,493%, sehingga didapatkan rata-rata penambahan akurasi dari skenario pertama ke skenario kedua sebesar 1,356%. Pada skenario ketiga ekstraksi fitur dan bobot kata akan diseleksi fitur menggunkan Chi-square. Model KNN melakukan klasifikasi cukup baik daripada NBC pada skenario ketiga. Jika dibandingkan dengan skenario kedua rata-rata kenaikan akurasi yang didapatkan untuk NBC sebesar 1,114% sedangkan untuk K-NN, mendapatkan peningkatan signifikan sebesar 7,768%. Pada skenario 4 hanya menerapkan *undersampling method* pada data. Hasil sampling didapatkan untuk dimensi IE sebesar 6687, untuk dimensi NS sebesar 4910, untuk dimensi TF sebesar 5821 dan untuk dimensi JP sebesar 6493. Perbandingan hasil klasifikasi skenario keempat dengan skenario ketiga, didapatkan bahwa terdapat penurunan akurasi dengan rata-rata dimensi dan model sebesar 3,385%. Penerapan *undersampling method* ternyata belum berhasil meningkatkan akurasi.

Berdasarkan penjelasan 4 skenario, didapatkan bahwa performansi model K-NN lebih baik jika dibandingkan dengan performansi model NBC. Hal ini juga menjawab rumusan masalah nomor 1 yaitu "Berapa perbandingan tingkat akurasi klasifikasi menggunakan model NBC dengan model K-NN dalam mengklasifikasi

preferensi kepribadian berdasarkan kepribadian MBTI?", bahwa tingkat akurasi klasifikasi kepribadian menggunakan model NBC lebih rendah dari model K-NN dalam mengklasifikasi preferensi kepribadian berdasarkan kepribadian MBTI. Perbedaan signifikan dapat ditunjukan dengan rata-rata akurasi pada skenario 3 (skenario terbaik pada penelitian ini) untuk NBC mendapat rata-rata akurasi sebesar 1,114%, sedangkan untuk K-NN sebesar 7,768% atau berbeda 6,654%.

Perbedaan akurasi antara NBC dan K-NN dikarenakan model K-NN memiliki banyak parameter dalam melakukan *tranning*. Parameter yang digunakan pada model K-NN yaitu nilai K, pembobotan kata (BoW dan TF-IDF), pembagian data (70:30 dan 80:20), fitur ekstraksi, seleksi fitur serta jarak Euclidean yang digunakan, sedangkan parameter yang digunakan pada model NBC yaitu pembobotan kata (BoW dan TF-IDF), pembagian data (70:30 dan 80:20), fitur ekstraksi dan seleksi fitur. Terdapat 2 perbedaan parameter yaitu nilai K dan jarak Euclidean. Dua parameter ini merupakan bagian yang wajib ada pada model K-NN. Berdasarkan uji coba klasifikasi dengan kombinasi parameter didapatkan parameter nilai K=9 merupakan nilai parameter yang mendapatkan akurasi terbaik 11 dari 16 klasifikasi. Kombinasi K=9 dengan jarak Euclidean memberikan *performance* hasil klasifikasi yang cukup baik.

4.2.2. Pengaruh Ekstrasi Fitur dan Fitur Seleksi

Pengaruh ekstrasi fitur dapat dilihat dari hasil rata-rata akurasi setiap dimensi dan model pada skenario 1 ke skenario 2. Rata-rata kenaikan akurasi dari skenario 1 ke skenario 2 untuk NBC sebesar 0,220% sedangkan untuk K-NN sebesar 2,493%. Perubahan tingkat akurasi dengan menggunakan fitur ekstraksi dalam mengklasifikasikan preferensi kepribadian berdasarkan kepribadian MBTI sebesar 1,356%.

Perubahan tingkat akurasi yang terjadi dengan penambahan ekstraksi fitur belum terjadi signifikan. Ektraksi fitur yang digunakan merupakan hasil ekstraksi dari teks yang yang digunakan oleh pengguna, fitur dan demografi pengguna di Twitter (Fikry, 2018). Perubahan belum signifikan terjadi karena korelasi antar ekstraksi fitur yang dilihat pada Gambar 4.6.

Gambar 4.6. Korelasi Antar Setiap Ekstraksi Fitur

Jika korelasi warna merah maroon mengambarkan korelasi positif sangat kuat sedangkan korelasi berwarna biru tidak mempunyai hubungan antar kedua variabel. Korelasi kuat antar fitur dengan batas 0.6 yang dapat diartikan sebagai korelasi yang kuat. Dari Gambar 4.6 dapat dilihat yang terdapat hanya 5 fitur yang

saling berkorelasi kuat yaitu TXT01 dan TXT04, USR01 dan USR02, USR01 dan USR04, USR02 dan USR04 serta USR05 ke USR06. Jika dibandingkan dengan jumlah fitur seharusnya terdapat 20 fitur tersebut harus berkorelasi, tetapi pada data klasifikasi kepribadian MBTI ini tidak semua berkorelasi kuat.

Pengaruh ekstrasi fitur dan seleksi fitur dapat dilihat dari hasil rata-rata akurasi setiap dimensi dan model pada skenario 2 ke skenario 3. Rata-rata kenaikan akurasi dari skenario 2 ke skenario 3 untuk NBC sebesar 1,114% sedangkan untuk K-NN sebesar 7,768%. Perubahan tingkat akurasi dengan menggunakan fitur ekstraksi dalam mengklasifikasikan preferensi kepribadian berdasarkan kepribadian MBTI sebesar 6.412%. Hal ini juga menjawab rumusan masalah nomor 2 yaitu "Berapa perubahan tingkat akurasi dengan menggunakan fitur ekstraksi dan seleksi fitur dalam mengklasifikasikan preferensi kepribadian berdasarkan kepribadian MBTI?", bahwa perubahan tingkat akurasi dengan menggunakan fitur ekstraksi dan seleksi fitur dalam dalam mengklasifikasi preferensi kepribadian berdasarkan kepribadian MBTI sebesar 6.412% dengan menggunakan model NBC dan K-NN.

Ekstraksi fitur dan ekstraksi kata seperti BoW atau TF-IDF tidak bisa dipisahkan dari *noise* di data, sehingga dibutukan pemilihan fitur yang memiliki pengaruh signifikan (Adi, dkk., 2018). Fitur seleksi Chi-Square pada model K-NN mendapatkan peningkatan akurasi lebih tinggi daripada NBC. Hal ini dikarenakan model NBC memiliki asumsi bahwa semua fitur saling *independent*, sedangkan model K-NN tidak memiliki asumsi (hanya menghitung jarak antar dua data). Pada fitur seleksi Chi-Square, semakin banyak jumlah sampel cuitan yang digunakan,

semakin banyak pula kata-kata yang digunakan sehingga korelasi di antara kedua fitur tersebut positif. Pemilihan kata menggunakan Chi-Square menggunakan nilai probabilitas α *(alpha)* di bawah 0,05 berdasarkan tingkat kesalahan dalam statistik (Rahmad, 2015). Semakin kecil nilai α *(alpha)*, data akan mengurangi *noise* pada jumlah fitur dan kinerja waktu pelatihan model. Hal tersebut sangat berbeda dengan asumsi NBC, dimana mengharuskan semua fitur saling *independent*, sehingga menyebabkan pengaruh tingkat akurasi menggunakan NBC hanya sebesar 1,114%.

4.2.3. Karakteristik Dimensi pada Kepribadian MBTI di Twitter

Pemilihan fitur dengan *Chi-Square* sangat baik dalam mengurangi *noise* data berdasarkan label data. *Chi-Square* merupakan salah satu metode pemilihan fitur yang bekerja dengan menghilangkan beberapa fitur tanpa mengurangi akurasi kinerja (Sulistiani, 2017). Fitur-fitur yang telah berhasil diseleksi oleh Chi-Square memberikan karakteristik penggunaan kata-kata terhadap dimensi kepribadian MBTI di Twitter yang dapat dilihat pada Gambar 4.7.

Dimensi IE Dimensi NS Dimensi TF Dimensi JP

Gambar 4.7. *Wordcloud* Fitur Seleksi Chi-Square

a) Dimensi IE sangat didominasi oleh "aku", "gue", "sama", "kita" dan "kamu" dikuti dengan ekstraski fitur yaitu USER_WORDBIO dan USR_FOLLOWING. Terdapat perbedaan penggunaan kata terkait orang seperti "kita" antara Ekstrovert dan Introvert. Kata terkait orang jarang digunakan untuk orang Introvert daripada orang Ekstrovert karena kegiatan Ekstrovert lebih sering menjalin komunikasi dengan orang lain.

b) Dimensi NS sangat didominasi oleh "ada", "jadi", "sama", "mau", "sudah", "banget" dan "foto". dikuti dengan ekstraski fitur yaitu USER_WORDBIO, TXT_MEDIA dan USER_CUNQ_BIO. Terdapat perbedaan kata yang sudah terjadi ("terjadi" atau "jadi", "sudah", "ada") antara Intuisi dan Sensing. Sensing merasakan lebih banyak perhatian pada informasi yang masuk melalui panca indera, sedangkan intuisi lebih memperhatikan pola dan kemungkinan yang lihat dalam informasi.

c) Dimensi TF sangat didominasi oleh "saja", "dan", "jadi", "suka", "apa" dan "ke" dikuti dengan ekstraski fitur yaitu USR_WORDBIO, USR_FOLLOWING, USR_MEDIA, dan TXT_URL. Terdapat kata objektif ("suka") yang berbeda antara Thinking dan Feeling. Thinking lebih memiliki prinsip-prinsip objektif, sedangkan Feeling lebih kepada masalah pribadi.

d) Dimensi JP sangat didominasi oleh "sinis", "ambisi", "santun", "sekarat" dan "sunyi" dikuti dengan ekstraski fitur yaitu USR_WORD_BIO, USR_FOLLOWING, TXT_MENTION dan TXT_MEDIA. Terdapat perbedaan kata *rigid* ("santun", "ambisi") antara Judging dan Perceiving.

Judjing lebih memilih gaya hidup yang lebih terstruktur dan tegas, sedangkan Perceiving memilih gaya hidup *flexible* dan mudah beradaptasi.

4.2.2. Perbandingan Hasil Penelitian

Bagian ini akan membahas hasil klasifikasi yang dilakukan perbandingan rata-rata akurasi terbaik penelitian ini dengan referensi rujukan pada Tabel 2.1. yang dapat dilihat pada Tabel 4.3 berikut:

Tabel 4.3. Perbandingan Hasil Penelitian

Variabel	Akurasi Skenario Rujukan	Akurasi Skenario Dasar	Modifikasi Skenario	Keterangan
Klasifikasi Tweet				
K-NN	66% (Claudy,dkk 2018)	65,598% (Rata-rata Skenario 1 semua dimensi MBTI pada model K-NN)	**78,589%** (Rata-rata Skenario 3 semua dimensi MBTI pada model K-NN)	Terjadi peningkatan akurasi dengan memodifikasi Skenario 1 menjadi Skenario 3 sebesar 10,261%.
NBC	65% Rata-rata NB pada setiap dimensi MBTI (Lukito, dkk 2016)	73,610% (Rata-rata Skenario 1 semua dimensi MBTI pada model NBC)	**74,944%** (Rata-rata Skenario 3 semua dimensi MBTI pada model NBC)	Terjadi peningkatan akurasi dengan memodifikasi Skenario 1 menjadi Skenario 3 sebesar 1,334 %.
Bobot Kata (BoW)	37,41% (Utami, dkk 2019)	64,517% (Rata-rata Skenario 1 semua dimensi, BoW pada model KNN)	**72,115%** (Rata-rata Skenario 3 semua dimensi, BoW, Ekstraksi fitur, dan seleksi fitur, pada model KNN)	Terjadi peningkatan akurasi dengan memodifikasi Skenario 1 menjadi Skenario 3 sebesar 7.598 %.
Klasifikasi Pengguna				
IE	75,80% (Iskandar,dkk 2020)	-	**95,018%.**	Terjadi peningkatan akurasi sebesar 19,218%
NS	55.52% (Iskandar, dkk 2020)	-	**62,993%.**	Terjadi peningkatan akurasi sebesar 7,473%
TF	**95.02%** (Iskandar, dkk 2020)	-	84,178%.	Terjadi penurunan akurasi sebesar 10,842%
JP	88.26% (Iskandar, dkk 2020)	-	**90,747%.**	Terjadi peningkatan akurasi sebesar 2,487%

Keterangan:

- Skenario rujukan adalah skenario yang mendapatkan akurasi terbaik yang digunakan oleh penulis pada referensi penelitian rujukan.
- Skenario dasar adalah skenario yang memiliki parameter yang sama dengan skenario rujukan.
- Modifikasi skenario adalah modifikasi skenario dasar yang digunakan untuk meningkatkan hasil dari variabel terkait.

Hasil performansi model KNN membandingkan dengan penelitian yang dilakukan Claudy, dkk (2018) yang skenario yang sama pada skenario pertama sehingga perbandingannya akan *apple to apple*. Dari hasil perbandingan didapatkan pada skenario Claudy, dkk (2018) sebesar 66%, sedangkan skenario dasar pada penelitian ini yaitu skenario 1, mendapatkan akurasi sebesar 65.598%. setelah dilakukan modifikasi skenario dengan skenario 3 didapatkan akurasi sebesar 78,589% sehingga pada penelitian ini dapat merekomendasi parameter chi-square dan fitur seleksi dapat meningkatan akurasi model K-NN sebesar 10.261%.

Selanjutnya hasil rata-rata model Naïve Bayes pada penelitian Lukito dkk (2016) mendapatkan akurasi 65%. Jika membandingkan *apple to apple*, pada rata-rata skeniaro 1 pada semua dimensi MBTI didapaktan sebesar 73,610%. Setelah dilakukan modifikasi skenario yaitu skenerio 3 didapatkan akurasi sebesar 74,944%. Pada penelitian ini merekomendasi untuk meningkatan akurasi dapat menggunakan parameter ekstraksi fitur dan fitur seleksi chi-square dapat meningkatan akurasi model NBC sebesar 1,334%.

Penelitian Utami, dkk (2019) menggunakan *not stemmed-not weighted* mendapatkan akurasi sebsar 37,41%, sedangkan pada penelitian skenario yang hampir mirip dengan yang dilakukan oleh Utami, dkk (2019) adalah skenairo pertama pada bobot kata BoW mendapatkan rata-rata akurasi sebesar 64,517%. Setelah dilakukan modifikasi skenario yaitu skenario 3, didapatkan akurasi sebesar 72.115%. Sehingga dengan menerapkan fitur seleksi chi-square didapatkan akurasi meningkant sebesar 7,598%.

Perbandingan yang didapatkan dengan data yang sama, yang digunakan oleh Iskandar, dkk (2020) dengan menggunakan parameter yang hampir sama yaitu bobot kata = BoW, fitur seleksi Chi-Square taraf kesalahan 0,05 dan SMOTE *sampling method*. Hasil klasifikasi pengguna pada penelitian ini berdampak baik, didapatkan akurasi untuk klasifikasi kepribadian untuk dimensi IE sebesar 75,80% sedangkan penelitian ini mendapatkan 95,018%. Untuk dimensi NS sebesar 55.52% sedangkan penelitian ini mendapatkan 62,993%. Untuk dimensi TF sebesar 95.02% sedangkan penelitian ini mendapatkan 84,178%. Terakhir, dimensi JP sebesar 88.26% sedangkan penelitian ini mendapatkan 90,747%. Tiga dimensi yang mengalami peningkatan, dan 1 dimensi yag mengalami penurunan. Penelitian ini memberikan perbandingan atau kontribusi terhadap peningkatan akurasi pada data yang sama pada penelitian Iskandar,dkk (2020).

4.2.5. Kelebihan dan Kekurangan Penelitian

Selama proses pelatihan dan pengujian dalam melakukan klasifikasi kepribadian berdasarkan prefensi kepribadian MBTI, terdapat dengan beberapa kelebihan dan kelamahan. Adapun kelebihan yang didapatkan sebagai berikut:

a. Proses *training* menggunakan model NBC penelitian ini dilakukan dengan waktu kurang dari 3 menit dan penambahan ruang *memory* sebanyak 1 GB.
b. Proses *traning* dengan menggunakan seleksi fitur Chi-Square pada model K-NN mengalami penurunan dari rata-rata waktu pada skenario ke 1 selama 10 menit menjadi 4 menit pada rata-rata waktu pada skenario ke 3.
c. Ekstrasi fitur pada Tabel 3.1. dan seleksi fitur Chi-Square meningkatkan akurasi, hal ini dilihat dari peningkatan akurasi sebesar dari rata-rata skenario 1 ke skenario 2 sebesar 1,356% dan Skenario 2 ke Skenario 3 sebesar 4,441% pada model NBC dan K-NN.

Kelemahan yang didapatkan selama menerapkan skenario sehingga mendapatkan perbedaan akurasi baik meningkat atau menurun yaitu:

a. Normalisasi fitur kata yang digunakan hanya pada kekurangan 1 karakter ("kaka" menjadi "kakak") dan kelebihan 2 karakter ("akuuu" menjadi "aku"). Penelitian ini belum bisa mengatasi fitur kata tidak baku yang merupakan gabungan dari 2 kata atau imbuhan ("gapenting", "satusatu", "gemesiin", dst)
b. Model K-NN memiliki 4 kali perhitungan atau 15 menit lebih lama dari model NBC. Hal ini dikarenakan model K-NN melakukan perhitungan jarak pada jumlah tetangga terdekat untuk 4 paramter nilai K yaitu 3, 5, 7 dan 9.

c. Metode *sampling* data yang digunakan pada penelitian ini (*undersampling*) dapat mengurangi hasil akurasi, hal ini dilihat dari peningkatan akurasi sebesar dari rata-rata skenario 3 ke skenario 4 sebesar 3,385% pada model NBC dan K-NN.

d. Model NBC memiliki kekurangan pada fitur atau variabel yang bernilai 0. Hal ini dapat menyebabkan hasil probabilitas dari model Naive Bayes menjadi 0.

4.2. Profiling Kepribadian MBTI

Pada bagian ini akan menjelaskan *profiling* secara langsung berbasis web dengan menggunakan model yang sudah dilakukan *tranning*. Hasil tampilan *user interface* web dan hasil dari prediksi MBTI dengan menggunakan data Twitter dapat dilihat pada Gambar 4.8.

Gambar 4.8. *Profiling* Kepribadian dengan data Twitter

Gambar pada bagian kiri menunjukan tampilan awal, dimana terdapat *form* untuk *input public username*, sedangkan pada bagian kanan menampilkan hasil klasifikasi kepribadian MBTI dengan model KNN. Pada simulasi diatas, *username* yang digunakan adalah "Himynameisagung", setalah dilakukan prediksi, maka hasil kepribadian MBTI dengan *username* "Himynameisagung" adalah ENFP.

Prediksi kepribadian diatas merupakan klasifikasi yang berdasarkan isyarat linguistik dari cuitan yang lebih cenderung mencerminkan kepribadian yang dirasakan oleh orang lain pada saat menuliskan cuitan. Misalnya, mengeluarkan ekspresi pendapat Twitter mendorong orang untuk mengungkapkan perasaan batin dan berbagi aktivitas sosial mereka dengan orang lain, yang berarti bahwa semua pengguna akan tampak seperti kepribadian Extrovert. Demikian pula, kebanyakan orang cenderung meuliskan tweet tentang pengalaman atau penemuan baru mereka yang memberi kesan kepada orang lain bahwa mereka Perceiving untuk pengalaman baru (Qiu, dkk 2012).

BAB V

PENUTUP

5.1. Kesimpulan

Berdasarkan rumusan masalah dan penjelasan yang telah peneliti deskripsikan pada bab sebelumnya, maka peneliti dapat menarik kesimpulan sebagai berikut:

a. Model K-NN memiliki hasil akurasi yang lebih baik dari pada model NBC dalam mengklasifikasi preferensi kepribadian berdasarkan kepribadian MBTI. Pendekatan model K-NN lebih tepat digunakan dengan fitur ekstraksi dan seleksi fitur Chi-Square daripada model NBC dalam mengklasifikasi preferensi kepribadian berdasarkan kepribadian MBTI karena parameter nilia K dan perhitungan jarak Euclidean pada K-NN dapat meminimalkan *noise* pada data.
b. Pengaruh fitur ekstraksi dan seleksi dalam mengklasifikasi preferensi kepribadian berdasarkan kepribadian MBTI memberikan dampak peningkatan akurasi yang signifikan. Pemilihan fitur dengan Chi-Square sangat baik dalam mengurangi noise data berdasarkan label data. Fitur dimensi IE lebih berfokus kata-kata terkait orang ("aku", "gue", "kita"). Fitur dimensi NS lebih berfokus kata-kata terkait fakta ("ada", "jadi", "sama", "mau" dan "sudah". Fitur dimensi TF lebih berfokus kepada fitur "saja", "dan", "jadi" dan "suka". Terakhir, Fitur dimensi JP lebih berfokus kepada fitur "sinis", "ambisi" dan "santun".

5.2. Saran

Peneliti penyadari bahwa pendekatan klasifikasi masih memiliki beberapa kekurangan serta kterbatasan. Oleh karena itu, ada beberapa hal yang perlu dipertimbangkna untuk mengembangkan pendekatan model klasifikasi kepribadian MBTI menjadi lebih baik yaitu:

a. Untuk mengatasi stemming kata tidak baku, dapat menggunakan modifikasi stemming non-formal affix pada saat text preprocessing
b. Untuk mengurangi komputasi yang tinggi pada model K-NN, dapat menggunakan fitur seleksi lain atau dapat menggunakan kombinasi beberapa fitur seleksi lainnya.
c. Untuk menghindari penyebaran distribusi data yang tidak merata, dapat melakukan proses *prototional sampling* dari setiap fitur pada data.
d. Untuk menghindari probabilitas bernilai nol, dapat melakukan analisis penyimpangan data berdasarkan rata-rata.

DAFTAR PUSTAKA

PUSTAKA BUKU

Briggs, M. I. & Myers, P.B, 1995, *Gifts Differing: Understanding Personality Type.* Palo Alto, Calif.: Davies-Black Pub

PUSTAKA ELEKTRONIK

Adi, G.Y.N.N; Tandio, M.H.; Ong, V.; Suhartono, D. 2018, *Optimization for Automatic Personality Recognition on Twitter in Bahasa Indonesia*, 3rd International Conference on Computer Science and Computational Intelligence https://www.sciencedirect.com/science/article/pii/S1877050918314893

Amanullah, R. F., Utami, E., & Sunyoto, A. (2019). *Citation Detection on Scientific Journal Using Support Vector Machine*. 2019 International Conference on Information and Communications Technology (ICOIACT). doi:10.1109/icoiact46704.2019.8938522

Bharadwaj, S.; Sridhar, S.; Choudhary, R.; Srinath, R., 2018, *Persona Traits Identification based on Myers-Briggs Type Indicator(MBTI) - A Text Classification Approach*, Conference: 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI). https://10.1109/ICACCI.2018.8554828

Celli, F., & Lepri, B. (2018). *Is Big Five Better than MBTI? A Personality Computing Challenge Using Twitter Data*. In CLiC-it. http://ceur-ws.org/Vol-2253/paper04.pdf

Claudy, Y.; Perdana, R.; Fauzi, M., 2018, *Klasifikasi Dokumen Twitter Untuk Mengetahui Karakter Calon Karyawan Menggunakan Algoritme K-Nearest Neighbor (K-NN)*. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer Vol. 2, No. 8, Agustus 2018, hlm. 2761-2765. https://www.researchgate.net/publication/322959490

Dinov, I. D., 2018, *Probabilistic Learning: Classification Using Naive Bayes. In: Data Science and Predictive Analytics Chapter 8.* Springer, Cham. https://link.springer.com/chapter/10.1007/978-3-319-72347-1_8

Fikry, M. dan Yusra, 2018, *Ekstrover Atau Introver: Klasifikasi Kepribadian Pengguna Twitter Dengan Menggunakan Metode Support Vector Machine,* Jurnal Sains Dan Teknologi Industri Vol 16, No 1 http://ejournal.uin-suska.ac.id/index.php/sitekin/article/view/5326

Iskandar, A. F., Utami, E. and Budi, A.P., 2020, *Impact of Feature Extraction and Feature Selection Using Naïve Bayes on Indonesian Personality Trait*, 2020 3rd International Conference on Information and Communications Technology (ICOIACT), Pending Publication.

Iskandar, A. F., Utami, E. and Budi, A.P., 2020, *What You Say Define Who You Are? Word Exploration and Automatic Personality Detection*, IJSTR, Vol 9: 2. https://www.ijstr.org/final-print/dec2020/What-You-Say-Define-Who-You-Are-Word-Exploration-And-Automatic-Personality-Detection.pdf.

Iskandar, A. F., Utami, E. and Budi, A.P., 2020, *Word Analysis of Indonesian Keirsey Temperament*, IJCCS, Vol 9: 2. doi:10.22146/ijccs.58595

Jo, T. (2019). *Text Mining: Concepts, Implementation, and Big Data Challenge*. Springer International Publishing. doi: 10.1007/978-3-319-91815-0

Kantardzic, M. (2020). *Data Mining: Concepts, Models, Methods, and Algorithms, 3nd Edition*. Wiley-IEEE Press. http://93.174.95.29/main/38960843B7D6FD8B71F3CE97F7D5B5BB

Lukito, L.C.; Erwin, A.; Purnama, J.; & Danoekoesoemo, W., 2016, *Social media pengguna personality classification using computational linguistic*, 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE), Yogyakarta, 2016, pp. 1-6., doi:10.1109/ICITEED.2016.7863313

Mihuandayani, Utami, E., & Luthfi, E. T. (2018). *Text Mining Based on Tax Comments as Big Data Analysis Using SVM and Feature Selection*. International Conference on Information and Communications Technology (ICOIACT). doi:10.1109/icoiact.2018.8350743

Moreno, D. R. J., Gomez, J. C., Almanza-Ojeda, D. L., & Ibarra-Manzano, M. A. (2019). *Prediction of Personality Traits in Twitter Penggunas with Latent Features*. In 2019 International Conference on Electronics, Communications and Computers (CONIELECOMP) (pp. 176-181). IEEE. doi: 10.1109/CONIELECOMP.2019.8673242

Ong, V., dkk, 2017, *Personality Prediction Based on Twitter Information in Bahasa Indonesia*, Federated Conference on Computer Science and Information Systems. https://10.15439/2017F359

Peterka-Bonetta, J., Sindermann, C., Elhai, J. D., & Montag, C. (2021). *How objectively measured Twitter and Instagram use relate to self-reported personality and tendencies toward Internet/Smartphone Use Disorder*. Human Behavior and Emerging Technologies. doi.org/10.1002/hbe2.243

Preot iuc-Pietro, D., Eichstaedt, J., Park, G., Sap, M., Smith, L., Tobolsky, V., Schwartz, H. A., and Ungar, L. H. (2015). *The role of personality, age and gender in tweeting about mental illnesses*. In Proceedings of the Workshop on Computational Linguistics and Clinical Psy- chology: From Linguistic Signal to Clinical Reality. ACL. https://www.aclweb.org/anthology/W15-1203.pdf

Pratama, B. T., Utami, E., & Sunyoto, A. (2019). *The Impact of Using Domain Specific Features on Lexicon Based Sentiment Analysis on Indonesian App Review*. 2019 International Conference on Information and Communications Technology (ICOIACT). doi:10.1109/icoiact46704.2019.8938419

Pratiwi, N. I., Budi, I., & Alfina, I. (2018). *Hate Speech Detection on Indonesian Instagram Comments using FastText Approach*. 2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS). doi:10.1109/icacsis.2018.8618182

Putra, J. W. G. (2019). *Pengenalan Pembelajaran Mesin dan Deep Learning*. https://wiragotama.github.io/ebook_machine_learning.html

Putra, R. B. S., & Utami, E. (2018). *Non-formal affixed word stemming in Indonesian language*. 2018 International Conference on Information and Communications Technology (ICOIACT). doi:10.1109/icoiact.2018.8350735

Qiu, L., Lin, H., Ramsay, J., & Yang, F. (2012). *You are what you tweet: Personality expression and perception on Twitter*. Journal of Research in Personality, 46(6), 710–718. doi:10.1016/j.jrp.2012.08.008

Rohman, A. N., Utami, E., & Raharjo, S. (2019). *Deteksi Kondisi Emosi pada Media Sosial Menggunakan Pendekatan Leksikon dan Natural Language* Processing. Jurnal Eksplora Informatika, 9(1), 70-76. https://eksplora.stikom-bali.ac.id/index.php/eksplora/article/view/277

Suhendra, I. R. Naive Bayes Algorithm with Chi Square and NGram Feature for Reviewing Laptop Product on Amazon Site. International Research Journal of Computer Science, (12), 28-33. 10.26562/IRJCS.2017.DCCS10087

Sulistiani, H., & Tjahyanto, A. (2017). *Comparative Analysis of Feature Selection Method to Predict Customer Loyalty*. Journal of Engineering, Vol. 3, No. 1. 2017. http://iptek.its.ac.id/index.php/joe/article/view/2257/

Suryono, S., Utami, E., & Luthfi, E. T. (2018). *Klasifikasi Sentimen pada Twitter Dengan Naive Bayes Classifier*. Angkasa: Jurnal Ilmiah Bidang Teknologi, 10(1), 89-96. http://dx.doi.org/10.28989/angkasa.v10i1.218

Utami, E., Hartanto, A. D., Adi, S., Oyong, I., & Raharjo, S. (2019). *Profiling analysis of DISC personality traits based on Twitter posts in Bahasa Indonesia.*

Journal of King Saud University-Computer and Information Sciences. https://doi.org/10.1016/j.jksuci.2019.10.008

Verhoeven, B., Daelemans, W., & Plank, B. (2016). *Twisty: a multilingual twitter stylometry corpus for gender and personality profiling*. In Proceedings of the 10th Annual Conference on Language Resources and Evaluation (LREC 2016)/Calzolari https://www.aclweb.org/anthology/L16-1258.pdf

Willy, Setiawan, E. B., & Nugraha, F. N. (2019). *Implementation of Decision Tree C4.5 for Big Five Personality Predictions with TF-RF and TF-CHI2 on Social Media Twitter*. 2019 International Conference on Computer, Control, Informatics and Its Applications (IC3INA). doi: 10.1109/ic3ina48034.2019.8949601

Wilmott, P. (2019). *Machine Learning: An Applied Mathematics Introduction*. Panda Ohana Publishing. http://93.174.95.29/main/2634B5FA5DFD49A90B6A22ADD1622BB6

Yarkoni, T. (2010). *Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers*. Journal of Research in Personality, 44(3), 363–373. doi:10.1016/j.jrp.2010.04.001

Yuan, C., Wu, J., Li, H., & Wang, L. (2018). *Personality Recognition Based on Pengguna Generated Content*. 2018 15th International Conference on Service Systems and Service Management (ICSSSM). doi: 10.1109/icsssm.2018.8465006

Zuhdi, A. M., Utami, E., & Raharjo, S. (2019). Analisis Sentiment Twitter Terhadap Capres Indonesia 2019 dengan Metode K-NN. Jurnal Informa, 5(2), 1-7.

LAMPIRAN

Lampiran 1. Dataset

Justifikasi terhadap dataset pada penelitian ini adalah bahwasanya setiap pengguna sudah melakukan test kepribadian secara online. Sampel hasil dari *screenshot* tes kepribadian secaraonline tersebut dapat dilihat sebagai berikut:

Tabel 1. *Sample Data*

No	Username	Tipe	Tweet	IE	NS	TF	JP
1	user_0	ESFP	dia colong punya lu ya	1	1	1	1
2	user_0	ESFP	ya lu minta kenal dari doi	1	1	1	1
3	user_4	INFJ	salah copas link wkwk	0	0	1	0
4	user_10	INFJ	iya terima kasih atas doa	0	0	1	0
5	user_62	ISTJ	kak aku mau check out apakah masih bisa	0	1	0	0

Terdapat 7 kolom yaitu "Username", "Tipe", "Tweet", "IE", "NS", "TF" dan "JP". Kolom "Username" menunjukan *username* pengguna yang terlah diganti atau disamarkan dengan kode untuk menjaga privasi dari pengguna. Kolom "Tipe" menunjukkan tipe kepribadian berdasarkan hasil dari tes kepribadian secara online. Kolom "Tweet" menunjukkan ekspresi pengguna atau cuitan di meida sosial Twitter. Kolom "IE", "NS", "TF" dan "JP" merupakan kolom yang diekstrak berdasarkan kolom "Tipe". Pada kolom "IE", label 0 menunjukan *Introvert* dan label 1 menunjukkan Extrovert. Kolom "NS" , label 0 menunjukkan *Intuition* dan label 1 menunjukan *Sensing*. Kolom "TF" , label 0 menunjukkan *Thinking* dan label 1 menunjukkan *Feeling*. Terakhir, kolom "JP" , label 0 menunjukan *Judging* dan label 1 meunjukkan *Perceiving*.

Lampiran 2. *Source Code* pada *Preprocessing* Data

Load DataFrame

```
In [2]: import pandas as pd
        pd.set_option('display.max_rows', 150)
        pd.set_option('display.max_columns', 500)
        pd.set_option('display.width', 1000)
```

```
In [46]: data=pd.read_csv('987.csv',sep=';')
         print(len(data))
         data.head()
```

Preprocessing

```
In [ ]: data['text']=data['text'].str.lower()
```

Number, Link and Punctuation Removal

```
In [ ]: data['text']=data['text'].apply(lambda x: re.sub(r'\d+', "", x))
        data['text']=data['text'].apply(lambda x: x.translate(str.maketrans("","",string.punctuation)))
```

```
In [ ]: def replace_kata(kalimat):
            user = re.compile('@')
            hastag = re.compile('#')
            gambar = re.compile('pic')
            link = re.compile('http[a-zA-Z]')
            katai = []
            for kata in kalimat.split():
                if re.search(user, kata):
                    kata=kata.replace(kata, 'username') ### untuk user
                if re.search(hastag, kata):
                    kata=kata.replace(kata, 'hastag') ### untuk hastag
                if re.search(gambar, kata):
                    kata=kata.replace(kata, 'gambar') ### untuk gambar
                if re.search(link, kata):
                    kata=kata.replace(kata, 'link') ### untuk link

                katai.append(kata)
            return ' '.join(katai)
```

```
In [ ]: data['text']=data['text'].apply(lambda x: replace_kata(x))
```

Word Normalization

```
In [ ]: from nltk.tokenize.treebank import TreebankWordDetokenizer
```

```
In [ ]: slangword=pd.read_csv('../../Kamus Normalisasi Data.csv',sep=';',error_bad_lines=False)
        slangword=slangword[['slang','format']]
        slangword.drop_duplicates(inplace = True)
        print(len(slangword))
        slangword.head()
```

```
In [1]: def perbaiki_slang(katacoba):
            kalimat=[]
            for katai in katacoba.split():
                try:
                    kataperbaikan=slangword[slangword['slang']==katai].reset_index()
                    kataperbaikan=kataperbaikan['format'][0]
                    kalimat.append(kataperbaikan)
                except:
                    kalimat.append(katai)
                girlnxt =' '
                hasil=TreebankWordDetokenizer().detokenize(kalimat)
            return hasil
```

```
In [ ]: %%time
        data['text_slang']=data['text'].apply(lambda x: perbaiki_slang(x))
```

Word Normalization 1

```
In [69]: import gensim
         # Load pre-trained Word2Vec model.
         #model = gensim.models.KeyedVectors.load("tdt4841_word2vec_plus_200.txt")
         model = gensim.models.KeyedVectors.load_word2vec_format("../../../Kunnskapsgraf ITTEST/tdt4841_word2vec_plus_200.txt", binary=False)

In [60]: words = list(model.wv.vocab)

         w_rank = {}
         for i, word in enumerate(words):
             w_rank[word]=i

         WORDS = w_rank

         C:\Users\lenovo\Anaconda3\lib\site-packages\ipykernel_launcher.py:1: DeprecationWarning: Call to deprecated `wv` (Attribute will be removed in 4.0.0, use self instead).
           """Entry point for launching an IPython kernel.

In [3]: import itertools

In [68]: import re
         from collections import Counter

         def words(text): return re.findall(r'\w+', text.lower())

         def P(word):
             "Probability of `word`."
             # use inverse of rank as proby
             # return 0 if the word isn't in the dictionary
             return - WORDS.get(word, 0)

         def correct(word):
             "Most probable spelling correction for word."
             return max(candidates(word), key=P)

         def candidates(word):
             "Generate possible spelling corrections for word."
             return (known([word]) or known(edits1(word)) or known(edits2(word)) or [word])

         def known(words):
             "The subset of `words` that appear in the dictionary of WORDS."
             return set(w for w in words if w in WORDS)

         def edits1(word):
             "All edits that are one edit away from `word`."
             letters    = 'abcdefghijklmnopqrstuvwxyz'
             splits     = [(word[:i], word[i:])    for i in range(len(word) + 1)]
             deletes    = [L + R[1:]               for L, R in splits if R]
             transposes = [L + R[1] + R[0] + R[2:] for L, R in splits if len(R)>1]
             replaces   = [L + c + R[1:]           for L, R in splits if R for c in letters]
             inserts    = [L + c + R               for L, R in splits for c in letters]
             return set(deletes + transposes + replaces + inserts)

         def edits2(word):
             "All edits that are two edits away from `word`."
             return (e2 for e1 in edits1(word) for e2 in edits1(e1))

In [70]: def korreksi_kalimat(text):
             kalimat=[]
             tokens=text.split()
             for token in tokens:
                 kalimat.append(correction(token))

             hasil=TreebankWordDetokenizer().detokenize(kalimat)
             return hasil

In [70]: korreksi_kalimat('saya')

Out[70]: 'saya'

In [71]: %%time
         data['text1_sbr']=data['text1_clang'].apply(lambda x: korreksi_kalimat(x))

         Wall time: 11min 42s
```

Stemming

```
In [73]: !pip install Pysastrawi

Requirement already satisfied: Pysastrawi in c:\users\lenovo\anaconda3\lib\site-packages (1.2.0)

Error processing line 7 of c:\users\lenovo\anaconda3\lib\site-packages\pysastrawi.pth:

  Traceback (most recent call last):
    File "c:\users\lenovo\anaconda3\lib\site.py", line 168, in addpackage
      exec(line)
    File "<string>", line 1, in <module>
  ModuleNotFoundError: No module named 'pysastrawi.monitrmap'

Remainder of file ignored
```

```
In [74]: from Sastrawi.Stemmer.StemmerFactory import StemmerFactory

factory = StemmerFactory()
stemmer = factory.create_stemmer()
```

```
In [76]: %%time
data['tanti_stem'] = data['tanti_clean'].apply(lambda x: stemmer.stem(x))

Wall time: 3.74 s
```

Lemmatization

```
In [77]: from spacy.lang.id import Indonesian
import nltk
from Sastrawi.StopWordRemover.StopWordRemoverFactory import StopWordRemoverFactory
```

```
In [78]: nlp = Indonesian()  # blank object
nlp = spacy.blank('id')  # blank instance

def lemma_word(sentence):
    text = nlp(sentence)
    token_data = [token.lemma_ for token in text]
    resultWithSpace = ' '.join(token_data)
    return resultWithSpace.lower()
```

```
In [79]: %%time
data['tanti_lemma'] = data['tanti_stem'].apply(lambda x: lemma_word(x))

Wall time: 3.04 s
```

Ekstraksi Fitur

```
In [80]: #menghitung jumlah karakter
listkarak=['a','b','c','d','e','f','g','h','i','j','k','l','m','n','o','p','q','r','s','t','u','v','w','x','y','z',
           'A','B','C','D','E','F','G','H','I','J','K','L','M','N','O','P','Q','R','S','T','U','V','W','X','Y','Z']

for i in listkarak:
    a=data['tanti'].str.count(i)
    j+=a
data['character']=j
```

```
In [82]: data['words'] = data['text'].apply(lambda x: len(x.split()))data['countwords'].astype(np.int64)
data['words'] = data.words.astype(np.int64)
data['Like'] = data['text'].apply(lambda x: x.count('#Like'))
data['pic.twi'] = data['text'].apply(lambda x: x.count('pic.twitter.com'))
data['Sentence'] = data['text'].apply(lambda x: x.count('.'))
data['Question'] = data['text'].apply(lambda x: x.count('?'))
data['Imperative'] = data['text'].apply(lambda x: x.count('!'))
data['Hastag'] = data['text'].apply(lambda x: x.count('#'))
data['We Retweet'] = data['text'].apply(lambda x: x.count('RT '))
data['Mentulis_User'] = data['text'].apply(lambda x: x.count('@'))
```