Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.2024117989
Vol. 11, No. 5, Oktober 2024, hlm. 1033-1042 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 36/E/KPT/2019 e-ISSN: 2528-6579

PENERAPAN SMOTE UNTUK MENGATASI *IMBALANCE CLASS* DALAM KLASIFIKASI KEPRIBADIAN MBTI MENGGUNAKAN *NAIVE BAYES CLASSIFIER*

Mutiara Persada Pulungan[1], Andi Purnomo[2], Aliyah Kurniasih[3]

[1,2,3]Sekolah Tinggi Ilmu Manajemen dan Ilmu Komputer ESQ, Jakarta
Email: [1]m.persada.p@students.esqbs.ac.id, [2]andi.purnomo@esqbs.ac.id, [3]aliyah.kurniasih@esqbs.ac.id
[*]Penulis Korespondensi

(Naskah masuk: 06 November 2023, diterima untuk diterbitkan: 30 Oktober 2024)

Abstrak

Kepribadian *Myers-Briggs Type Indicator* (MBTI) telah menjadi topik populer dalam memahami karakteristik individu dan dampaknya pada interaksi sosial, karir, dan pengambilan keputusan. Model *Machine Learning* dengan algoritma *Naive Bayes Classifier* sering digunakan untuk memprediksi kepribadian MBTI berdasarkan data Twitter. Namun, seringkali terjadi ketidakseimbangan kelas, dengan beberapa jenis kepribadian yang memiliki sampel lebih sedikit. Untuk mengatasi hal ini, penelitian ini menggunakan teknik *Synthetic Minority Over-sampling Technique* (SMOTE) untuk meningkatkan jumlah sampel pada kelas minoritas. Selain itu, metode *Term Frequency-Inverse Document Frequency* (TF-IDF) digunakan untuk mengekstraksi fitur penting dari teks. Penelitian ini bertujuan menerapkan teknik SMOTE untuk mengatasi ketidakseimbangan kelas dalam klasifikasi kepribadian MBTI menggunakan beberapa algoritma *Naive Bayes Classifier*, termasuk *Gaussian, Multinomial*, *Bernoulli*, *Complement,* dan *Logistic Regression* berdasarkan model *Keirsey*: *Artisan*, *Guardian, Rational*, dan *Idealist*. Evaluasi menggunakan metode *Hold-Out-Validation* dengan membagi data menjadi 90% data latih dan 10% data uji. Hasil evaluasi menunjukkan performa rendah algoritma *Naive Bayes Classifier* untuk kelas *Artisan* dan *Guardian*, tetapi baik untuk kelas *Rational* dan *Idealist*. Algoritma *Logistic Regression* memiliki akurasi tertinggi 80% dan performa yang lebih baik secara keseluruhan, meskipun masih rendah untuk kelas *Artisan* dan *Guardian*. Dengan demikian, penelitian ini memberikan pemahaman tentang penggunaan algoritma *Naive Bayes Classifier* dan teknik SMOTE dalam prediksi kepribadian MBTI, dengan potensi peningkatan kinerja melalui penggunaan algoritma *Logistic Regression*.

Kata kunci: *Myers-Briggs Type Indicator*(MBTI), *Imbalance Class*, *Synthetic Minority Over-sampling Technique* (SMOTE), *Term Frequency-Inverse Document Frequency* (TF-IDF), *Naive Bayes Classifier*.

APPLICATION OF SMOTE TO OVERCOME CLASS IMBALANCE IN THE MBTI PERSONALITY CLASSIFICATION USING THE NAÏVE BAYES CLASSIFIER

Abstract

Myers-Briggs Type Indicator (MBTI) personality is becoming a popular topic in understanding individual characteristics and their impact on social interaction, career, and decision-making. Machine Learning models with Naive Bayes Classifier algorithms are often used to predict MBTI personalities from Twitter data. However, there is often a class imbalance, with some personality types having a smaller sample. To overcome this, this study used the Synthetic Minority Over-sampling Technique (SMOTE) technique to increase the number of samples in minority classes. Additionally, the Term Frequency-Inverse Document Frequency (TF-IDF) method is used to extract important features from text. This study aims to apply SMOTE techniques to address class imbalances in MBTI personality classification using several Naïve Bayes Classifier algorithms, including Gaussian, Multinomial, Bernoulli, Complement, and Logistic Regression based on Keirsey's model: Artisan, Guardian, Rational, and Idealist. Evaluation using the Hold-Out-Validation method by dividing the data into 90% training data and 10% test data. The evaluation results showed low performance of the Naive Bayes Classifier algorithm for the Artisan and Guardian classes, but both for the Rational and Idealist classes. The Logistic Regression algorithm has the highest accuracy of 79% and better performance overall, although it is still low for the Artisan and Guardian classes. Thus, this study provides insight into the use of Naive Bayes Classifier algorithm and SMOTE technique in MBTI personality prediction, with potential performance improvement through the use of Logistic Regression algorithm.

Keywords: *Myers-Briggs Type Indicator(MBTI), Imbalance Class, Synthetic Minority Over-sampling Technique (SMOTE), Term Frequency-Inverse Document Frequency (TF-IDF), Naive Bayes Classifier.*

1. PENDAHULUAN

Kepribadian adalah keseluruhan perilaku yang dimiliki seseorang dalam berinteraksi dengan orang lain (Haq and Budi, 2019). Kepribadian individu memengaruhi perilaku sosial, gaya hidup, dan kesehatan mental individu tersebut (Mahajan et al., 2022). Kepribadian menjadi faktor penting dalam berbagai aspek kehidupan seperti akademik, pekerjaan, hubungan, dan penggunaan media sosial (Bharadwaj et al., 2018). Kepribadian diukur melalui tes psikotes grafis dan kuesioner. Tes grafis melibatkan gambar yang dibuat individu, sementara tes kuesioner menggunakan jawaban terhadap pertanyaan (Wijaya and Cendana, 2020). Penilaian kepribadian dilakukan menggunakan inventaris laporan diri kepada psikolog. Metode inventaris laporan diri dapat akurat, namun rentan terhadap bias dan membutuhkan banyak sumber daya serta waktu yang cukup lama (Harahap, Muslim and Korespondensi, 2020).

Saat ini, telah tersedia berbagai jenis psikotes yang dapat diakses secara online untuk mengetahui kepribadian seseorang, diantaranya *Big Five Factor Personality*, *Myers-Briggs Type Indicator* (MBTI) dan DISC (Sher Khan et al., 2020). Namun, MBTI adalah salah satu yang populer dan teruji keakuratannya, serta digunakan dalam berbagai bidang seperti pendidikan, karir, organisasi, dan kelompok. MBTI juga mudah dipahami, diaplikasikan, dan sering digunakan dalam penelitian kepribadian (Claudy, Setya Perdana and Fauzi, 2018). Kepribadian MBTI menggunakan teknik kuesioner yang berisi pertanyaan singkat (Utami and Bathiar, 2020).

Tes MBTI penting dalam berbagai skenario seperti toleransi risiko, konflik manajemen, dan kesuksesan karir (Bharadwaj et al., 2018). Untuk mengklasifikasikan kepribadian MBTI, model *machine learning* seperti SVM, *Random Forest*, *K-Nearest Neighbor*, *Convolutional Neural Network*, dan *Decision Tree* dapat digunakan. Dalam melakukan klasifikasi kepribadian MBTI, algoritma *Naïve Bayes Classifier* merupakan pilihan terbaik dalam klasifikasi kepribadian MBTI karena memiliki tingkat akurasi yang lebih baik dibandingkan dengan model *classifier* lainnya, serta dapat digunakan dalam klasifikasi teks, filter spam, analisis sentimen, dan sistem rekomendasi (Felicia Watratan et al., 2020).

Klasifikasi kepribadian umunya telah banyak dilakukan menggunakan data teks Twitter. Twitter digunakan oleh banyak orang untuk berinteraksi dan menyampaikan pendapat. Data dari aktivitas pengguna Twitter dapat digunakan sebagai sumber informasi potensial untuk klasifikasi kepribadian (Hasri and Alita, 2022). Penelitian terkait prediksi kepribadian MBTI terhadap data twitter telah dilakukan Fikry, (2018) menggunakan algoritma SVM dan menunjukkan tingkat akurasi 88,89% pada perbandingan data training dan testing 80:20, namun mengalami penurunan menjadi 68,29% pada perbandingan 90:10 Fikry, (2018). Penelitian kepribadian MBTI terhadap data twitter juga dilakukan oleh Wijaya and Cendana, (2020) dengan menggunakan algoritma *Naive Bayes Classifier* dan menggunakan data training sebanyak 8.675 baris yang telah diberi label kelas sesuai dengan 16 tipe kepribadian MBTI, penelitian tersebut mencapai tingkat akurasi sebesar 71%.

Masalah ketidakseimbangan kelas dalam klasifikasi teks menjadi faktor yang perlu diperhatikan dalam konteks penelitian ini. Masalah ketidakseimbangan kelas terjadi saat jumlah data pada kelas target tidak seimbang (Iskandar and Nataliani, 2021). Dalam prediksi kepribadian MBTI, ketidakseimbangan data pada setiap tipe kepribadian dapat mengurangi presisi dan recall pada kelas minoritas. Akibatnya, model memiliki kecenderungan untuk lebih baik dalam memprediksi kelas mayoritas dan kurang efektif dalam memprediksi kelas minoritas (Sulistiyowati and Jajuli, 2020). Untuk mengatasi masalah ini, pendekatan oversampling, yaitu mensintesis data sampel dari kelas minoritas, digunakan pada penelitian ini. Teknik *oversampling* dipilih karena tidak mengurangi dataset dan memberikan hasil yang lebih baik dalam menangani ketidakseimbangan kelas pada dataset dengan jumlah sampel kelas minoritas yang sedikit (Qadrini, Hikmah and Megasari, 2022).

Teknik *oversampling* mencakup metode seperti *Random Over Sampling* (ROS), *Synthetic Minority Oversampling Technique* (SMOTE), *Borderline SMOTE*, *k-Means SMOTE*, *Support Vector Machine SMOTE* (SVM-SMOTE), dan *Adaptive Synthetic* (ADASYN) (Indrawati, 2021). Pada penelitian ini, digunakan model SMOTE karena efektif dalam menangani ketidakseimbangan kelas dan mengurangi *overfitting*, serta menghasilkan akurasi yang baik (Sulistiyono et al., 2021). Penelitian mengenai Teknik SMOTE telah dilakukan oleh Khasana, Muladi and Pujianto Utomo, (2019) untuk mengatasi *imbalance class* dalam klasifikasi objektivitas berita *online* dengan algoritma k-NN. Hasil penelitian menunjukkan bahwa SMOTE meningkatkan performa klasifikasi dengan nilai akurasi 87.5% dan presisi, recall, serta F-measure yang baik.

Maka dari beberapa penelitian yang dijabarkan diatas, penelitian ini akan menerapkan teknik SMOTE untuk mengatasi *imbalance class* dalam klasifikasi kepribadian MBTI menggunakan berbagai jenis algoritma *Naive Bayes Classifier*, seperti *Multinomial Naive Bayes*, *Gaussian Naive Bayes*, *Bernoulli Naive Bayes*, *Complement Naive Bayes*, dan *Logistic Regression* yang akan dibandingkan dalam klasifikasi teks menggunakan pembobotan kata TF-IDF. Evaluasi model akan dilakukan dengan *Hold-Out Validation* dan

menggunakan metrik akurasi, presisi, recall, dan F1-score. Data yang digunakan berjumlah 8.675 dengan label sesuai dengan tipe kepribadian MBTI, yang diambil dari situs *Kaggle*. Tujuan penelitian ini adalah untuk mengatasi data *imbalance class* pada kepribadian MBTI menggunakan teknik SMOTE dan membandingkan kinerja algoritma *Naive Bayes Classifier* dalam klasifikasi kepribadian MBTI.

2. METODE PENELITIAN

Penelitian ini dimulai dengan menentukan topik penelitian dengan cara melakukan pencarian permasalahan yang ada disekitar dan yang dapat diselesaikan menggunakan metode yang telah dipelajari, dilanjutkan dengan mengidentifikasi, merumuskan dan memeberikan batasan masalah hingga menentukan tujuan dan manfaat yang diperoleh dari hasil penelitian ini. Tahap berikutnya adalah melakukan tinjauan pustaka, mengolah dan mengumpulkan data, melakukan pengujian dan menganalisis hasil serta penarikan kesimpulan dan saran.

2.1 Pengumpulan Data

Data dikumpulkan dari situs resmi *Kaggle* dengan pemilik himpunan data adalah Mitchell J yang terakhir di update pada tahun 2017. Adapun karakteristik yang terdapat pada dataset tersebut terdiri dari 2 atribut yaitu atribut type dan post yang terdiri dari 16 jenis tipe kepribadian dan 8765 baris data teks narasi cuitan Twitter. Dataset tersebut dapat diunduh pada link https://www.kaggle.com/datasets/datasnaek/MBTI-type. Adapun contoh dari beberapa baris data dari dataset yang digunakan dapat dilihat pada table dibawah ini.

Tabel 1. Contoh Dataset

No	Type	Posts
1	ENFJ	friends don't seem them that way usually. I don't really know why it is.....
2	INFJ	friend posted on his facebook before committing suicide the next.....
3	INTJ	I enjoyed our conversation the other day. Esoteric gabbing about ……..
4	ENTJ	'You're fired.||That's another silly misconception. That approaching is.....
5	ENTP	I've posted this thread on the philosophy board too, but I am really ENFPs.....
6	INFP	'I think we do agree. I personally don't consider myself Alpha.....
7	INTP	'https://www.youtube.com/watch?v=w8-egj0y8Qs|||I'm in this …….
8	ENTP	'I'm finding the lack of me in these posts very alarming.|||Sex can be boring ……..
9	ESFP	Hi all, if you've got some spare time and MORE TALENT IN….....
10	ISFP	'ok so i'm moving and my new housemate said i could drop.....
11	ISTP	'Sir, are you high?|||Wow, you are pretty sucessfull person, being CEO….....
12	ESTP	'It can be difficult to tell sometimes. Both would seem to have.....
13	ESFJ	'Lol that's what I figured it meant but then I was like, no way|||I don't speak ….....
14	ISFJ	'I dislike how I always seem to manage to step on INFP's toes by.....
15	ISTJ	'When I'm not at a busy part in my life, I'll keep everything.,...…….....
16	ESTJ	'People have their priorities. It sounds like you know what his are. I'd.....

2.2 Pengelompokan Data

Dataset yang telah dikumpulkan akan dikelompokkan kedalam 4 kategori kelas di mana masing-masing kelas tersebut berisi 4 data tipe kepribadian MBTI. Kelas tersebut adalah kelas *Artisan*, *Guardian*, *Rational* dan *Idealist*. Rincian frekuensi pada dataset (MBTI) *Myers-Briggs Type Indicator* yang telah di kelompokkan ke dalam 4 kelas tersebut dapat dilihat pada table berikut.

Tabel 2. Model *Keirsey*

No	Kelas	Frekuensi
1	*Artisan* (ISFP, ISTP, ESFP, ESTP)	678
2	*Guardian* (ISTJ, ISFJ, ESFJ, ESTJ)	404
3	*Rational* (ENFP, INFJ, INFP, ENFJ)	3758
4	*Idealist* (INTP, ENTJ, ENTP, INTJ)	2967

2.3 *Pre-processing* Data

Konsep *pre-processing* akan diterapkan ke seluruh dataset yang digunakan pada penelitian ini. Adapaun proses *pre-processing* yang digunakan meliputi proses *cleaning data*, *case folding*, *remove number*, *punctuation*, *remove single char*, *remove lead, trail* dan *duplicate whitespace*, *tokenizing*, *filtering stopword*, dan *lemmatization*. Contoh dataset yang telah dilakukan *pre-processing* dapat dilihat pada table berikut.

Tabel 3 Dataset Setelah *Preprocessing*

No	Data Sebelum *preprocessing*	Data Sesudah *preprocessing*
1	[intj, moments, sportscenter, top, ten, plays,.....	[intj, moments, sportscenter, top, ten, play, ...
2	[finding, lack, posts, alarming, sex, boring, ...	[find, lack, post, alarm, sex, bore, position,.....
3	[good, one, course, say, know, blessing, curse.....	[good, one, course, say, know, bless, curse, a.....

Dengan melakukan *pre-processing*, data akan menjadi lebih siap dan sempurna untuk digunakan dalam analisis dan pemodelan. *Pre-processing* juga membantu memastikan bahwa data yang digunakan adalah data yang berkualitas, lebih mudah diinterpretasikan, dan menghasilkan hasil yang lebih akurat dan bermakna.

2.4 Pembagian Data

Setelah dataset melewati proses *pre-processing* data, maka selanjutnya data tersebut dibagi menjadi data training dan data testing. Hal ini dilakukan untuk membagi data menjadi dua bagian yang terpisah, di mana set pelatihan digunakan untuk melatih model dan set pengujian digunakan untuk menguji performa model yang telah dilatih. Pada proses pembagian data ini ditentukan proporsi dari data yang akan dialokasikan untuk set pengujian sebesar 10% dari data keseluruhan dan 90% dialokasikan untuk data training. Adapun

frekuensi data training dan data testing yang digunakan adalah sebagai berikut.

Tabel 4 Pembagian Dataset

Data Set	Jumlah Data
Data Training	7807
Data Testing	868

2.5 Pembobotan TF-IDF

Setelah data dibagi menjadi data training dan data testing, selanjutnya data tersebut akan dibobotkan menggunakan teknik pembobotan TF-IDF. Data training dan data testing akan dihitung nilai TF dan IDF yang nantinya akan diubah menjadi representasi vektor dan dikonversi menjadi array. Untuk menghitung nilai TF-IDF, Proses pertama yang dilakukan adalah dengan mencari nilai TF *(Term Frequency)* dengan rumus sebagai berikut:

$$Tf_{t,d} = \frac{f_d(t)}{\max f_t(d)} \quad (1)$$

Dimana:
$Tf_{t,d}$:Jumlah frekuensi kata t pada dokumen d
$f_d(t)$: Kemunculan kata t pada dokumen d
$\max f_t(d)$: Total seluruh kata pada dokumen d

Selanjutnya, mencari nilai IDF *(Invers Document Frequency)* dengan rumus sebagai berikut:

$$Idf_{(t,D)} = \log\left(\frac{N}{df_t}\right) \quad (2)$$

Dimana:
$Idf_{t,D}$: Invers Frekuensi seluruh dokumen
N : Jumlah semua dokumen
df_t : Jumlah dokumen yang mengandung kata t

Kemudian nilai TF dan IDF akan dikalikan, dengan rumus sebagai berikut:

$$W_{t,d} = Tf_{t,d} \times Idf_{(t,D)} \quad (3)$$

Dimana:
$W_{t,d}$: Bobot kata t pada dokumen d
$Tf_{t,d}$: Jumlah frekuensi kata t pada dokumen d
$Idf_{t,d}$: Jumlah frekuensi kemunculan kata t pada seluruh dokumen

Bobot TF-IDF memberikan nilai penting bagi kata-kata dalam dokumen berdasarkan frekuensi kemunculan mereka dalam dokumen dan sejauh mana kata-kata tersebut eksklusif dalam dokumen tersebut.

2.6 Implementasi Teknik SMOTE

SMOTE (*Synthetic Minority Over-sampling Technique*) adalah sebuah teknik *oversampling* yang digunakan untuk menangani ketidakseimbangan kelas dalam data. Ketidakseimbangan kelas terjadi jika jumlah setiap kelas yang ada memiliki kelas mayoritas dan minoritas. Jika ketidakseimbangan kelas tersebut tidak ditangani atau diabaikan maka dapat menyebabkan model memiliki bias yang sangat signifikan terhadap kelas mayoritas. Hal ini dapat mengakibatkan model yang tidak peka terhadap kasus-kasus dalam kelas minoritas yang mungkin memiliki nilai prediktif yang penting. Akurasi model dapat sangat tinggi karena dominasi kelas mayoritas, tetapi hasil prediksi pada kelas minoritas akan menjadi sangat rendah.

Pada Tabel 2 dapat dilihat bahwa terdapat ketidak seimbangan data pada setiap kelas. Maka, dilakukan teknik SMOTE untuk menangani ketidakseimbangan kelas pada data training yang digunakan pada penelitian ini. Teknik ini bertujuan untuk meningkatkan jumlah sampel dalam kelas minoritas dengan menciptakan sampel sintetis berdasarkan data yang ada. Adapun rumus teknik SMOTE adalah sebagai berikut:

$$Xsyn = Xi + (Xknn - Xi) \times \delta \quad (4)$$

Dimana:
$Xsyn$: Data Sintesis
Xi : Data yang akan direplikasi
$Xknn$: Data yang memiliki jarak terdekat dengan data Xi
δ : Nilai random antara 0 dan 1

Teknik SMOTE bekerja dengan cara memilih sebuah sampel dari kelas minoritas Xi yang akan di-*oversampling*. Kemudian menemukan *K-Nearest Neighbors* (tetangga terdekat) dari Xi dalam kelas minoritas. Nilai k harus ditentukan terlebih dahulu secara acak. Setelah itu akan dihitung perbedaan antara Xi dan $Xknn$ dan menghitung rasio acak (δ) antara 0 dan 1 dan dilakukan perkalian. Langkah terakhir adalah dengan menambahkan hasil perkalian pada Xi untuk mendapatkan hasil $Xsyn$. Hasil yang didapatkan setelah Teknik SMOTE digunakan pada data training adalah sebagai berikut.

Tabel 5 Implementasi Teknik SMOTE

No	Kelas	Frekuensi Sebelum	Frekuensi Sesudah
1	*Artisan*	678	3758
2	*Guardian*	404	3758
3	*Rational*	3758	3758
4	*Idealist*	2967	3758

2.7 Implementasi Algoritma Naïve Bayes Classifier

Model yang akan dibangun pada penelitian ini adalah menggunakan algoritma *Naive Bayes Classifier*. Algoritma ini akan digunakan untuk membandingkan tingkat akurasi yang diperoleh dari setiap algoritma yang digunakan. Adapun algoritma yang di gunakan adalah *Multinomial Naive Bayes Classifier, Gaussian Naive Bayes Classifier, Bernouli Naive Bayes Classifier, Complement Naive Bayes Classifier*, dan *Logistic*

Regresion. Berikut adalah penjelasan untuk pengimplementasian masing-masing algoritma:

2.7.1 Algoritma *Multinomial Naïve Bayes*

Algoritma *Multinomial Naive Bayes* (MNB) adalah algoritma yang digunakan dalam klasifikasi teks untuk memprediksi kelas atau kategori dari sebuah dokumen berdasarkan kemunculan kata-kata dalam dokumen tersebut. *Multinomial Naive Bayes* dapat diformulasikan pada rumus persamaan berikut ini.

$$P(p|n) \propto P(p) \prod 1 \leq k \leq nd \; P(tk|p) \quad (5)$$

Dimana:

P(p|n)	:	Probabilitas dokumen n berada di kelas p
P(p)	:	Prior ptobabilitas suatu dokumen berada di kelas p
P(tk|p)	:	Probabilitas bahwa data tk benar jika hipotesis p benar.
$\prod 1 \leq k \leq nd \; P(tk\|p)$	:	Hasil perkalian dari probabilitas likelihood P(tk|p) untuk setiap elemen data dalam n, yang menunjukkan seberapa baik hipotesis p cocok dengan seluruh dataset.

2.7.2 Algoritma *Bernouli Naive Bayes*

Algoritma *Bernoulli Naive Bayes* (BNB) adalah model yang menentukan bahwa sebuah dokumen diwakili oleh vektor atribut biner yang menunjukkan kata mana yang muncul dan mana yang tidak muncul dalam dokumen. Model ini juga digunakan untuk data kategorikal, tetapi dengan asumsi bahwa atributnya adalah variabel biner (dengan hanya dua kemungkinan nilai: 0 atau 1). *Bernouli* dapat diformulasikan pada rumus persamaan berikut ini.

$$P(y|X) = P(y) \prod (1 \leq i \leq n) \, P(xi|y)) + \prod (1 \leq i \leq n) \, (1 - P(xi \, |y)) \quad (6)$$

Dimana:

P(y|X)	:	Probabilitas posterior X berada dalam kelas y
P(y)	:	Probabilitas prior kelas y
P(xi|y)	:	Probabilitas atribut biner ke-i adalah muncul dalam kelas y
n	:	Jumlah atribut biner dalam instance X

2.7.3 Algoritma *Gaussian Naive Bayes*

Algoritma *Gaussian Naive Bayes* adalah varian dari algoritma *Naive Bayes Classifier* yang dihitung menggunakan distribusi normal. *Gaussian* sendiri memungkinkan klasifikasi data numerik dengan distribusi *Gaussian* dan data kategorikal.

$$P(C|Z) = \frac{p(Z|C) \times p(c)}{p(z)} \quad (7)$$

Dimana:

P(C|Z)	:	Probabilitas posterior kelas C pada data Z
P(Z|C)	:	Probabilitas data Z muncul pada kelas C
P(c)	:	Probabilitas prior
P(Z)	:	Probabilitas evidence

2.7.4 Algoritma *Complement Naïve Bayes*

Algoritma *Complement Naive Bayes*, yang merupakan transformasi dari algoritma *Multinomial Naive Bayes* untuk mengatasi kumpulan data yang tidak seimbang. Model ini dikembangkan untuk memecahkan masalah kelas minoritas yang tidak seimbang dalam data. Algoritma *Complement* memperhitungkan distribusi atribut di kelas mayoritas dan mengabaikan atribut yang terjadi di kelas minoritas. *Complement Naive Bayes* dapat diformulasikan pada persamaan berikut ini.

$$P(C|W_i) = argmin \, P(c) \prod i = 1 \frac{1}{P(w|\hat{c})^f i} \quad (8)$$

Dimana:

$P(C\|W_i)$	:	Probabilitas posterior kelas C berdasarkan atribut W_i yang di amati
argmin P(c)	:	Mencari kelas c yang memiliki probabilitas prior terendah.
$\prod i = 1 \frac{1}{P(w\|\hat{c})^f i}$	:	Hasil perkalian dari invers probabilitas atribut W_i terhadap kelas yang diprediksi $\hat{c}$, dengan bobot ^{f} i yang mungkin diberikan pada setiapatribut.

2.7.5 Algoritma *Logistic Regression*

Logistic Regression (Regresi Logistik) adalah teknik pembelajaran statistik dan mesin yang digunakan untuk memodelkan hubungan antara variabel dependen biner dan satu atau lebih variabel independen. Tujuan dari Regresi Logistik adalah untuk memprediksi probabilitas kejadian suatu peristiwa atau kategori. *Logistic Regression* dapat diformulasikan pada persamaan berikut ini.

$$P = \frac{e^{a+bX}}{1 + e^{a+bX}} \quad (9)$$

Dimana:

P	:	Probabilitas suatu instance/data
a	:	Konstanta atau bias
bX	:	Koefisien yang menggambarkan hubungan antara atribut X dan Probabilitas.

3. HASIL DAN PEMBAHASAN

Hasil yang di dapatkan pada penelitian yang dilakukan dalam membangun sebuah model machine learning adalah sebagai berikut.

3.1 Hasil Pengujian

Pengujian model pada penelitian ini berfokus pada tingkat akurasi, *precision*, *recall* dan *F1-score* yang dihasilkan beberapa algoritma yang disebutkan sebelumnya. Data yang digunakan pada proses pengujian

ini adalah 10% dari keseluruhan data yaitu sebanyak 868 baris data. Adapun ringkasan analisis pengujian model adalah sebagai berikut:

- Pengujian Menggunakan *Gaussian Naive Bayes*

Tabel 6 *Gaussian Naive Bayes*

Class	*Precission*	*Recall*	*F1-Score*
Artisan	0.33	0.39	0.35
Guardian	0.28	0.44	0.34
Idealist	0.73	0.60	0.70
Rational	0.79	0.78	0.78
Acuracy		**0.68**	

Tabel 6 menunjukkan bahwa model memiliki kinerja yang baik dalam mengklasifikasikan kelas '*Rational*' dan '*Idealist*', tetapi kinerja yang terbatas dalam mengklasifikasikan kelas '*Artisan*' dan '*Guardian*'.

- Pengujian Menggunakan *Multinomial Naive Bayes*

Tabel 7 *Multinomial Naive Bayes*

Class	*Precission*	*Recall*	*F1-Score*
Artisan	1.00	0.00	0.00
Guardian	1.00	0.00	0.00
Idealist	0.81	0.68	0.74
Rational	0.67	0.94	0.78
Acuracy		**0.71**	

Tabel 7 menunjukkan bahwa model memiliki kinerja yang baik dalam mengklasifikasikan kelas '*Rational*' dengan presisi yang lebih rendah, tetapi memiliki performa yang buruk untuk kelas '*Artisan*' dan '*Guardian*'.

- Pengujian Menggunakan *Bernouli Naive Bayes*

Tabel 8 *Bernoulli Naive Bayes*

Class	*Precission*	*Recall*	*F1-Score*
Artisan	0.25	0.45	0.34
Guardian	0.36	0.08	0.14
Idealist	0.73	0.71	0.72
Rational	0.76	0.77	0.77
Acuracy		**0.68**	

Tabel 8 menunjukkan bahwa model memiliki performa yang baik dalam mengklasifikasikan kelas '*Idealist*' dan '*Rational*', dengan presisi dan recall yang cukup tinggi. Namun, model memiliki kinerja yang rendah dalam mengklasifikasikan kelas '*Artisan*' dan '*Guardian*'.

- Pengujian Menggunakan *Complement Naive Bayes*

Tabel 9 *Complement Naive Bayes*

Class	*Precission*	*Recall*	*F1-Score*
Artisan	0.44	0.54	0.48
Guardian	0.42	0.48	0.45
Idealist	0.81	0.73	0.77
Rational	0.82	0.84	0.83
Acuracy		**0.75**	

Tabel 9 menunjukkan bahwa model memiliki kinerja yang baik dalam mengklasifikasikan kelas '*Idealist*' dan '*Rational*', dengan presisi yang tinggi dan recall yang baik. Model memiliki performa yang sedang dalam mengklasifikasikan kelas '*Artisan*' dan '*Guardian*', dengan presisi dan recall yang cukup seimbang

- Pengujian Menggunakan *Logistic Regresion*

Tabel 10 *Logistic Regression*

Class	*Precission*	*Recall*	*F1-Score*
Artisan	0.49	0.66	0.56
Guardian	0.44	0.58	0.50
Idealist	0.87	0.77	0.82
Rational	0.86	0.86	0.86
Acuracy		**0.80**	

Tabel 10 menunjukkan bahwa secara keseluruhan, model memiliki performa yang baik dalam mengklasifikasikan sampel-sampel sebagai '*Idealist*' dan '*Rational*', namun memiliki performa yang lebih rendah dalam mengklasifikasikan sampel-sampel sebagai '*Artisan*' dan '*Guardian*'.

3.2 Hasil Evaluasi Model

Berdasarkan hasil pengujian yang dilakukan, terlihat bahwa algoritma *Naive Bayes Classifier* dalam berbagai jenisnya memiliki performa yang berbeda dalam mengklasifikasikan kelas-kelas yang berbeda pula. Hasil evaluasi dari pengujian yang dilakukan dapat dilihat pada Tabel berikut:

Tabel 11 Hasil Evaluasi Tanpa SMOTE

Evaluasi		Model				
		GNB	MNB	BNB	CNB	LR
Akurasi		68%	71%	68%	75%	79%
Artisan	Presisi	33%	100%	26%	46%	48%
	Recall	39%	0%	43%	55%	69%
	F1-Score	35%	0%	32%	50%	57%
Guardian	Presisi	28%	100%	34%	41%	43%
	Recall	44%	0%	12%	49%	61%
	F1-Score	34%	0%	18%	45%	51%
Rational	Presisi	79%	67%	75%	81%	87%
	Recall	78%	94%	75%	80%	81%
	F1-Score	78%	78%	75%	80%	84%
Idealist	Presisi	73%	81%	71%	78%	83%
	Recall	66%	68%	68%	74%	78%
	F1-Score	70%	74%	70%	76%	81%

Tabel 122 Hasil Evaluasi Menggunakan SMOTE

Evaluasi		Model				
		GNB	MNB	BNB	CNB	LR
Akurasi		66%	75%	58%	75%	80%
Artisan	Presisi	33%	100%	28%	44%	49%
	Recall	39%	100%	45%	54%	66%
	F1-Score	35%	100%	34%	48%	56%
Guardian	Presisi	28%	100%	36%	42%	44%
	Recall	44%	100%	8%	48%	58%
	F1-Score	34%	100%	14%	45%	50%
Rational	Presisi	79%	81%	76%	82%	86%
	Recall	78%	68%	77%	84%	86%
	F1-Score	78%	74%	77%	83%	86%
Idealist	Presisi	34%	67%	73%	81%	87%
	Recall	73%	94%	71%	73%	77%
	F1-Score	66%	78%	72%	77%	82%

Keterangan:
GNB : *Gaussian Naive Bayes Classifier*
MNB : *Multinomial Naive Bayes Classifier*

BNB : *Bernoulli Naive Bayes Classifier*
CNB : *Complement Naive Bayes Classifier*
LR : *Logistic Regression*

Pada Tabel 12 dapat dilihat bahwa algoritma *Naive Bayes Gaussian, Multinomial, Bernoulli* memiliki performa yang rendah dalam mengklasifikasikan kelas *'Artisan' dan 'Guardian'*. Nilai presisi, recall, dan F1-*Score* untuk kedua kelas tersebut cenderung rendah, yang menunjukkan kesulitan dalam memprediksi dengan tepat dan menemukan kembali sampel-sampel yang sebenarnya termasuk dalam kelas tersebut. Namun, algoritma *Naive Bayes Classifier* dalam berbagai jenisnya memiliki performa yang baik dalam mengklasifikasikan kelas *'Rational' dan 'Idealist'*. Presisi, recall, dan F1-score untuk kedua kelas tersebut relatif tinggi, menunjukkan kemampuan algoritma dalam memprediksi dengan tepat dan menemukan kembali sampel-sampel yang sebenarnya termasuk dalam kelas tersebut.

Pada pengujian menggunakan algoritma *Logistic Regression*, terlihat bahwa performa untuk kelas *'Artisan' dan 'Guardian'* juga masih rendah. Presisi dan recall untuk kedua kelas tersebut cenderung rendah, menunjukkan kesulitan dalam memprediksi dengan tepat dan menemukan kembali sampel-sampel yang sebenarnya termasuk dalam kelas tersebut. Namun, performa untuk kelas *'Rational' dan 'Idealist'* cukup baik, dengan presisi, recall, dan F1-score yang tinggi. Jadi, hal ini menunjukkan bahwa algoritma *Naive Bayes Classifier*, terutama pada jenis *Gaussian, Multinomial,* dan *Bernoulli*, cenderung memiliki performa yang lebih baik dalam mengklasifikasikan kelas *'Rational'* dan *'Idealist'* dibandingkan dengan kelas '*Artisan'* dan *'Guardian'*.

Model dengan *Logistic Regression* memiliki akurasi tertinggi, yaitu sebesar 80%. Model ini memiliki performa yang relatif lebih baik dalam mengklasifikasikan sampel secara keseluruhan. Namun, meskipun akurasi secara keseluruhan cukup tinggi, model masih memiliki kinerja yang rendah dalam mengklasifikasikan kelas *'Artisan'* dan *'Guardian'*. Presisi, recall, dan F1-score untuk kedua kelas tersebut cenderung rendah, menunjukkan kesulitan dalam memprediksi dan menemukan kembali sampel-sampel yang termasuk dalam kelas-kelas tersebut. Sebaliknya, model memiliki performa yang lebih baik dalam mengklasifikasikan kelas '*Rational*' dan '*Idealist*', dengan presisi, recall, dan F1-*Score* yang lebih tinggi.

3.3 Pembahasan

Berdasarkan hasil dari pengujian yang dilakukan terhadap model machine learning dengan menerapkan teknik SMOTE terhadap klasifikasi kepribadian MBTI menggunakan *Naive Bayes Classifier* didapatkan hasil bahwa algoritma logistic regresi memiliki nilai akurasi yang cukup baik dibandingkan algoritma lainnya. Namun, performanya menurun saat mengklasifikasikan kelas *Artisan* dan *Guardian*. Adapun rincian pembahasan untuk masing-masing pengujian adalah sebagai berikut:

1. Algoritma *Logistic Regression* mendapatkan hasil yang lebih baik yaitu sebesar 79% dimana nilai presisi, recall dan f1-score pada kelas *Rational* dan *Idealist* lebih tinggi dibandingkan pada kelas *Artisan* dan *Guardian*.
2. Hasil dari penelitian yang dilakukan telah menunjukkan peningkatan dari penelitian-penelitian sebelumnya, yaitu peningkatan akurasi yang sebelumnya. Selain itu, terdapat pembaharuan terhadap preprocessing data yang dilakukan dan penambahan teknik SMOTE.
3. Ada beberapa faktor yang mendukung mengapa hasil dari model dengan menggunakan algoritma logistic regression lebih baik disbanding algoritma lainnya adalah sebagai berikut:
   - *Logistic Regression* merupakan metode yang kuat dalam mengatasi masalah klasifikasi dengan data yang memiliki atribut numerik atau kontinu dan data diskrit.
   - Dilakukannya proses *pre-processing* secara penuh, sehingga hasil yang lebih baik dapat dicapai.
   - Kualitas dan kuantitas data pelatihan yang digunakan dalam model dapat sangat mempengaruhi hasilnya. Data yang tidak representatif, memiliki kecacatan, atau tidak lengkap dapat menyebabkan model yang buruk. Oleh karena itu, penting untuk memastikan data pelatihan berkualitas tinggi dan cukup representatif untuk mencakup variasi yang ada dalam data yang akan diprediksi.
4. Algoritma *Naive Bayes Classifier* yaitu *Gaussian, Multinomial,* dan *Bernoulli* memiliki performa yang rendah dalam mengklasifikasikan kelas *Artisan* dan *Guardian*. Nilai presisi, recall, dan F1-score untuk kedua kelas tersebut cenderung rendah. Namun, algoritma *Naive Bayes Classifier* dalam berbagai jenisnya memiliki performa yang baik dalam mengklasifikasikan kelas *Rational* dan Idealis'. Presisi, recall, dan F1-score untuk kedua kelas tersebut relatif tinggi.
5. Adapun hal-hal yang menjadi penyebab model machine learning yang dibangun menggunakan jenis *Naive Bayes Classifier Multinomial, Gaussian* dan *Bernoulli* mendapatkan nilai akurasi yang rendah adalah sebagai berikut:
   - Ketidakcocokan fitur. Jika kelas *Artisan* dan *Guardian* memiliki hubungan antar fitur yang kompleks atau ketergantungan yang tidak memenuhi asumsi independensi, maka algoritma *Naive Bayes Classifier* akan mengalami kesulitan dalam mengklasifikasikan kedua kelas tersebut.
   - Ketidakseimbangan kelas, Jika kelas *Artisan* dan *Guardian* memiliki jumlah sampel yang jauh lebih sedikit dibandingkan dengan kelas *Rational* dan *Idealist*, algoritma *Naive Bayes Classifier* cenderung memberikan lebih sedikit perhatian dan pemodelan terhadap kedua kelas tersebut.
   - Ketidakseimbangan distribusi data
   - Kurangnya informasi fitur, Jika terdapat fitur-fitur yang memiliki informasi yang kurang relevan atau

tidak signifikan dalam membedakan antara kelas *Artisan* dan *Guardian*, maka performa algoritma *Naive Bayes Classifier* dapat menurun karena tidak dapat memperhitungkan faktor-faktor penting yang membedakan kedua kelas tersebut.

6. Di samping dari alasan-alasan yang telah dipaparkan, pembagian data training dan data testing juga dapat mempengaruhi performa algoritma klasifikasi, termasuk algoritma *Naive Bayes Classifier*. Jika pembagian data tidak dilakukan dengan proporsi yang tepat atau terjadi ketimpangan dalam jumlah sampel untuk setiap kelas antara data training dan data testing, hasil evaluasi performa algoritma bisa menjadi bias atau tidak representatif.

4 KESIMPULAN

Berdasarkan hasil dari analisis dan eksperimen yang telah dilakukan, dapat diperoleh simpulan sebagai berikut:

1. Penerapan Algoritma *Naive Bayes Classifier* untuk mengklasifikasikan kepribadian MBTI berhasil dilakukan dengan memanfatkan *library python* khususnya pada *text mining*, yaitu NLTK (Natural Language Toolkit), *Scikit-learn*, *numphy* dan pandas. Jenis algoritma *Naive Bayes Classifier* yang memiliki performa paling baik adalah logistik regresi dengan dengan perbandingan data 90:10 yaitu sebesar 80%.
2. Teknik SMOTE dapat mengatasi ketidak seimbangan kelas dan dapat meningkatkan performa prediksi kelas minoritas, sehingga memberikan hasil yang lebih akurat dan relevan dalam proses klasifikasi.

DAFTAR PUSTAKA

BHARADWAJ, S., SRIDHAR, S., CHOUDHARY, R. AND SRINARTH, R., 2018a. Persona Traits Identification Based On Myers-Briggs Type Indicator(Mbti) - A Text Classification Approach.

BHARADWAJ, S., SRIDHAR, S., CHOUDHARY, R. AND SRINARTH, R., 2018b. Persona Traits Identification Based On Myers-Briggs Type Indicator(Mbti) - A Text Classification Approach.

CLAUDY, Y.I., SETYA PERDANA, R. AND FAUZI, M.A., 2018. Klasifikasi Dokumen Twitter Untuk Mengetahui Karakter Calon Karyawan Menggunakan Algoritme K-Nearest Neighbor (Knn). [Online] Available At: <Http://J-Ptiik.Ub.Ac.Id>.

FELICIA WATRATAN, A., PUSPITA, A.B., MOEIS, D., INFORMASI, S. AND PROFESIONAL MAKASSAR, S., 2020. Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid-19 Di Indonesia. [Online] Journal Of Applied Computer Science And Technology (Jacost), Available At: <Http://Journal.Isas.Or.Id/Index.Php/Jacost>.

FIKRY, M., 2018a. Ekstrover Atau Introver : Klasifikasi Kepribadian Pengguna Twitter Dengan Menggunakan Metode Support Vector Machine. Jurnal Sains, Teknologi Dan Industri, 16(1), Pp.72–76.

FIKRY, M., 2018b. Ekstrover Atau Introver : Klasifikasi Kepribadian Pengguna Twitter Dengan Menggunakan Metode Support Vector Machine. Jurnal Sains, Teknologi Dan Industri, 16(1), Pp.72–76.

HAQ, F. AND BUDI, E., 2019. Implementasi Naive Bayes Classifier Untuk Prediksi Kepribadian Big Five Pada Twitter Menggunakan Term Frequency-Inverse Document Frequency (Tf-Idf) Dan Term Frequency-Relevance Frequency (Tf-Rf).

HARAHAP, R.N., MUSLIM, K. AND KORESPONDENSI, P., 2020. Peningkatan Akurasi Pada Prediksi Kepribadian Mbti Pengguna Twitter Menggunakan Augmentasi Data. 07, Pp.815–822. Https://Doi.Org/10.25126/Jtiik.202073622.

HASRI, C.F. AND ALITA, D., 2022. Penerapan Metode Naïve Bayes Classifier Dan Support Vector Machine Pada Analisis Sentimen Terhadap Dampak Virus Corona Di Twitter. Jurnal Informatika Dan Rekayasa Perangkat Lunak (Jatika), [Online] 3(2), Pp.145–160. Available At: <Http://Jim.Teknokrat.Ac.Id/Index.Php/Informatika>.

INDRAWATI, A., 2021. Penerapan Teknik Kombinasi Oversampling Dan Undersampling Untuk Mengatasi Permasalahan Imbalanced Dataset. Jurnal Informatika Dan Komputer) Akreditasi Kemenristekdikti, [Online] 4(1). Https://Doi.Org/10.33387/Jiko.

ISKANDAR, J.W. AND NATALIANI, Y., 2021. Perbandingan Naïve Bayes, Svm, Dan K-Nn Untuk Analisis Sentimen Gadget Berbasis Aspek. Jurnal Resti (Rekayasa Sistem Dan Teknologi Informasi), 5(6), Pp.1120–1126. Https://Doi.Org/10.29207/Resti.V5i6.3588.

KHASANA, A., MULADI AND PUJIANTO UTOMO, 2019. Penerapan Teknik Smote Untuk Mengatasi Imbalance Class Dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma Knn. Jurnal Resti (Rekayasa Sistem Dan Teknologi Informasi), 1(3), Pp.196–201.

MAHAJAN, R., MAHAJAN, R., SHARMA, E. AND MANSOTRA, V., 2022a. “Are We Tweeting Our Real Selves?” Personality Prediction Of Indian Twitter Users Using Deep Learning Ensemble Model. Computers In Human Behavior, 128. Https://Doi.Org/10.1016/J.Chb.2021.107101.

MAHAJAN, R., MAHAJAN, R., SHARMA, E. AND MANSOTRA, V., 2022b. “Are We Tweeting Our Real Selves?” Personality Prediction Of Indian Twitter Users Using Deep Learning Ensemble Model. Computers In Human Behavior, 128. Https://Doi.Org/10.1016/J.Chb.2021.107101.

QADRINI, L., HIKMAH, H. AND MEGASARI, M., 2022. Oversampling, Undersampling, Smote Svm Dan Random Forest Pada Klasifikasi Penerima Bidikmisi Sejawa Timur Tahun 2017. Journal Of

Computer System And Informatics (Josyc), 3(4), Pp.386–391. Https://Doi.Org/10.47065/Josyc.V3i4.2154.

SHER KHAN, A., AHMAD, H., ZUBAIR ASGHAR, M., KHAN SADDOZAI, F., ARIF, A. AND ALI KHALID, H., 2020. Personality Classification From Online Text Using Machine Learning Approach. [Online] Ijacsa) International Journal Of Advanced Computer Science And Applications, Available At: <Www.Ijacsa.Thesai.Org>.

SULISTIYONO, M., PRISTYANTO, Y., ADI, S. AND GUMELAR, G., 2021. Implementasi Algoritma Synthetic Minority Over-Sampling Technique Untuk Menangani Ketidakseimbangan Kelas Pada Dataset Klasifikasi. Sistemasi: Jurnal Sistem Informasi, [Online] 10, Pp.445–459. Available At: <Http://Sistemasi.Ftik.Unisi.Ac.Id>.

SULISTIYOWATI, N. AND JAJULI, M., 2020. Integrasi Naïve Bayes Dengan Teknik Sampling Smote Untuk Menangani Data Tidak Seimbang. Jurnal Nuansa Informatika, [Online] 14(1). Available At: <Https://Journal.Uniku.Ac.Id/Index.Php/Ilkom>.

UTAMI, G. AND BATHIAR, N., 2020. Aplikasi Pengenalan Kepribadian Tipe Myers Briggs Menggunakan Metode Fuzzy Saw Berbasis Android. Jurnal Masyarakat Informatika, Volume 11, Nomor 1, Issn 2086 – 4930.

WIJAYA, A. AND CENDANA, M., 2020. Klasifikasi Kepribadian Myres-Briggs Type Indicator Berdasarkan Cuitan Di Twitter Menggunakan Metode Tf-Idf Dan Naive Bayes Classifier. Jurnal Linguistik Komputasional, 3.

Halaman ini sengaja dikosongkan.