Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.202073622
Vol. 7, No. 4, Agustus 2020, hlm. 815-822 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 30/E/KPT/2018 e-ISSN: 2528-6579

PENINGKATAN AKURASI PADA PREDIKSI KEPRIBADIAN MBTI PENGGUNA TWITTER MENGGUNAKAN AUGMENTASI DATA

Rizki Nurhaliza Harahap*[1], Kemas Muslim[2]

[1,2]Universitas Telkom
Email: [1]lizaharahap@student.telkomuniversity.ac.id, [2]kemasmuslim@telkomuniversity.ac.id
*Penulis Korespondensi

(Naskah masuk: 9 Juni 2020, diterima untuk diterbitkan: 9 Juli 2020)

Abstrak

Kepribadian suatu individu perlu diketahui untuk membantu seseorang dalam mempertimbangkan beberapa hal, salah satunya perekrutan karier. Pada umumnya, kepribadian dapat diketahui melalui metode wawancara, observasi, maupun survei kuesioner. Akan tetapi, metode konvensional tersebut dinilai kurang praktis dari segi waktu dan materi karena dibutuhkan waktu yang lama dan biaya yang cukup besar untuk mengolah data. Selain itu, penggunaan metode konvensional juga dapat menimbulkan bias karena melibatkan orang ketiga dalam pengolahan data. Penelitian ini mencoba memberikan solusi dengan membangun model yang dapat melakukan prediksi terhadap kepribadian seseorang berdasarkan analisis data dan informasi dari media sosial Twitter. Data dan informasi tersebut akan diproses sehingga didapatkan prediksi kepribadian orang tersebut. Teori klasifikasi kepribadian yang digunakan adalah teori *Myers-Briggs Type Indicator* (MBTI). Penelitian ini juga mencoba menerapkan teknik augmentasi data untuk meningkatkan performa dari *text mining task* yang memiliki dataset sedikit. Hasil terbaik didapatkan dengan metode *Random Forest* menggunakan pembobotan *Term Frequency-Inverse Document Frequency* (TF-IDF) dan fitur yang tersedia pada Twitter. Penggunaan teknik augmentasi dapat meningkatkan akurasi hingga 30% dari akurasi awal sehingga hasil penelitian menunjukkan bahwa penggunaan teknik augmentasi data dapat meningkatkan performa pada model prediksi kepribadian MBTI.

Kata kunci: *MBTI, random forest, twitter, augmentasi data*

IMPROVEMENT OF ACCURACY IN THE TWITTER USER MBTI PERSONALITY PREDICTION USING DATA AUGMENTATION

Abstract

The personality of an individual needs to be known to help people in considering things, one of them is career recruitment. In general, personality can be known through interviews, observations, and questionnaire surveys. However, the conventional method is judged to be impractical in terms of time and material because it takes a long time and has considerable costs to process data. After all, the use of conventional methods can also cause bias because it involves a third person in data processing. The research tries to provide a solution by building a system that can predict the personality of a person based on the analysis of data and information from social media Twitter. The data and information will be processed so that the personality prediction is obtained. The personality classification theory used is the Myers-Briggs Type Indicator (MBTI) theory. The research also tries to implement data augmentation techniques to improve the performance of text mining tasks that have a slight dataset. The best results are obtained by the Random Forest method using the Term Frequency-Inverse Document Frequency (TF-IDF) weighted and the features available on Twitter. The use of augmentation techniques can increase accuracy by up to 30% from initial accuracy. So, the use of data augmentation techniques can be used to improve the performance of MBTI personality prediction models.

Keywords: *MBTI, random forest, twitter, data augmentation*

1. PENDAHULUAN

Setiap individu memiliki karakter yang bersifat unik. Sifat unik setiap karakter tersebut kerap digambarkan melalui kepribadian. Menurut ilmu psikologi, kepribadian merupakan penggabungan dari perilaku, sikap, serta pola respon emosional yang dimiliki oleh seorang individu (Bai et al., 2014). Kepribadian individu perlu diketahui untuk mengetahui potensi yang dimiliki suatu individu, kelompok yang lebih sesuai terhadap individu tersebut, pola pikir individu dalam mengambil

keputusan serta reaksi terhadap sesuatu, baik reaksi emosional maupun tindakan. Pengetahuan tentang kepribadian diri dapat digunakan sebagai bahan penilaian dalam mempertimbangkan serta memprediksi beberapa aspek kehidupan seperti keberhasilan akademis, perekrutan karier, status sosial, sikap politik, dan sebagainya (Azucar, Marengo dan Settanni, 2018).

Salah satu teori yang biasa digunakan para psikolog dalam mengklasifikasikan kepribadian adalah *Myers-Briggs Type Indicator* (MBTI). MBTI merupakan salah satu teori pengelompokan kepribadian individu tertua dan terpopuler di dunia. Teknik pengelompokan kepribadian individu MBTI merupakan teknik yang didasarkan pada teori kepribadian Jung. Teknik ini dikembangkan oleh Katharine C. Briggs dan putrinya, Isabel Briggs-Myers, selama Perang Dunia II untuk membantu menyesuaikan seseorang dengan pekerjaan. Sejak saat itu teknik ini semakin populer (Survey, 2013).

MBTI menggunakan teknik kuesioner yang berisi pertanyaan singkat. Hasil kuesioner seseorang akan diklasifikasi berdasarkan empat tipe yang dikotomis (berlawanan). Meskipun berlawanan, setiap individu sebenarnya memiliki keseluruhan dari tipe ini. Namun, kecenderungan terhadap suatu arah tipe tertentu menyebabkan terjadinya klasifikasi (Amaliyah & Noviyanto, 2013). Adapun keempat tipe dikotomis tersebut yaitu *Extrovert* (E) dan *Introvert* (I), *Sensing* (S) dan *Intuitive* (N), *Thinking* (T) dan *Feeling* (F), serta *Judging* (J) dan *Perceiving* (P).

Karakteristik MBTI dijelaskan pada Tabel 1. Perhitungan MBTI dilakukan dengan membandingkan masing-masing dikotomis seperti *Extrovert* dengan *Introvert*, *Thinking* dengan *Feeling*, *Sensing* dengan *Intuitive*, dan *Judging* dengan *Perceiving*. Setelah dilakukan perhitungan, akan dibandingkan nilai dari keduanya. Nilai terbesar akan menjadi elemen pembentuk tipe kepribadian yang diambil dari inisial masing-masing tipe dikotomis. Berdasarkan hasil perhitungan tersebut, akan didapatkan 16 kemungkinan tipe kepribadian dari MBTI. Setiap huruf mewakili satu tipe dikotomis.

Pada umumnya, penilaian kepribadian dilakukan menggunakan inventaris laporan diri kepada psikolog. Metode inventaris dapat secara tepat menentukan kepribadian seseorang dengan dasar teori mendalam. Namun, metode tersebut memiliki kelemahan, yaitu dapat menimbulkan bias karena melibatkan manusia dalam penilaiannya. Metode tersebut menghabiskan sumber tenaga dan sumber daya material yang besar karena dilakukan dengan manual. Selain itu, dibutuhkan waktu yang cukup lama untuk menyelesaikan survei dan melakukan penilaian terhadap hasil survei (Bai et al., 2014).

Tabel 1. Karakteristik Kepribadian MBTI

Extrovert Kepribadian yang berfokus pada dunia luar, mendapatkan motivasi dari berinteraksi dengan orang lain dan melakukan sesuatu.	***Introvert*** Kepribadian yang berfokus pada batin, mendapatkan motivasi dari pikiran, informasi, ide, dan konsep.
Thinking Memutuskan dengan logika dan analisis sebab akibat. Keputusan berusaha bersifat objektif tanpa melibatkan perasaan.	***Feeling*** Memutuskan dengan penekanan pada efek yang melibatkan perasaan orang lain dan diri. Keputusan mungkin didasarkan pada firasat, mencoba untuk menyesuaikan dan memuaskan orang lain.
Sensing Membuat keputusan berdasarkan fakta dan kepercayaan berdasarkan fakta, angka, dan detail.	***Intuitive*** Memutuskan berdasarkan intuisi, hubungan, dan spekulasi.
Judging Menilai dengan cepat dan memihak. Ingin menjadi pemain bukan penonton. Lebih terorganisasi daripada spontan.	***Perceiving*** Lebih senang menjadi penonton. Sangat lambat dalam menilai.

Penelitian terkait prediksi kepribadian cukup banyak dilakukan, beberapa penelitian tersebut ditunjukkan pada Tabel 2. Pada Tabel 2 terlihat bahwa sebagian besar penelitian prediksi kepribadian dilakukan dengan teori klasifikasi kepribadian Big5, karena Big5 dianggap lebih informatif sehingga lebih mudah diukur dan diprediksi. Hal tersebut menyebabkan penelitian yang menggunakan teori Big5 memiliki performa yang lebih baik dibandingkan MBTI. Namun, penelitian (Celli dan Lepri, 2018) mencoba membandingkan Big5 dan MBTI melalui sisi *personality computing*. Dari penelitian tersebut didapatkan hasil bahwa MBTI memiliki performa yang cukup baik dengan mengimplementasikan metode SVM pada model prediksi dengan nilai akurasi sebesar 64.6% dengan penggunaan fitur tertentu. Prediksi kepribadian MBTI sebelumnya dilakukan pada penelitian (Gjurković dan Šnajder, 2018) dengan menggunakan dataset yang bersumber dari Reddit. Hasil akurasi yang didapatkan cukup baik sebesar 82%.

Pada penelitian ini akan dilakukan prediksi kepribadian MBTI melalui pesan yang diunggah oleh pengguna Twitter. Penelitian ini menggunakan beberapa model pembelajaran mesin, seperti *Gradient Boosting*, *Random Forest*, *K-Nearest Neighbor*, dan *Multinomial Naïve Bayes*. Selain itu, dilakukan eksperimen terhadap beberapa skenario fitur. Enam belas tipe kepribadian dari MBTI akan dikelompokkan menjadi empat kelas. Penelitian (Lima dan De Castro, 2019) memproyeksikan 16 tipe kepribadian MBTI ke dalam empat kelas temperamen sesuai dengan model *Keirsey*. Temperamen merupakan konsep yang menyatu dengan karakteristik bawaan pada individu, karakter tersebut menentukan bagaimana seorang individu bertanggung jawab, memandang, serta berinteraksi

dengan dunia. Berdasarkan hal tersebut, dihasilkan empat jenis temperamen yang merupakan hasil gabungan dari setiap dikotomis MBTI. Empat kelas tersebut yakni *Guardian* yang merupakan kombinasi dari karakter *Sensing* dan *Judging*, *Artisan* yang merupakan kombinasi *Sensing* dan *Perceiving*, *Idealist* yang merupakan kombinasi dari *Intuitive* dan *Feeling*, dan *Rational* yang merupakan kombinasi dari *Intuitive* dan *Thinking*. Tabel 3 menggambarkan generalisasi 16 tipe kepribadian MBTI menjadi empat kelas.

Tabel 2. Daftar Literatur

Jurnal	Tahun	Teori Kepribadian	Sosial Media	Metode	Performa
Predicting Personality Traits of Microblog Users (Bai et al., 2014)	2014	Big5	Sina Microblog	*Multi-task Regression*	MAE: 0.1248 RMSE: 0.1559
Personality Prediction System from Facebook Users (Tandera et al., 2017)	2017	Big5	Facebook	LSTM + CNN	Akurasi: 74.17%
Predicting Personality Using Facebook Status Based on Semi-Supervised Learning (Zheng dan Wu, 2019)	2019	Big5	Facebook + Twitter	*Semi-Supervised Learning* + LIWC + unigram	F1: 0.71 *Precission*: 0.69 *Recall*: 0.73
Predicting Personality from Social Media Text (Golbeck, 2016)	2016	Big5	Facebook + Twitter	Receptiviti API LIWC	MAE: 15-30%
Modeling Personality Traits of Filipino Twitter Users (Tighe dan Cheng, 2018)	2018	Big5	Twitter	*Support Vector Machines* (linear SVM), *Logistic Regression* (LOG)	F1 SVM: 0.5669 F1 LOG: 0.6086
Predicting Personality Traits from Social Media using Text Semantics (Hassanein et al., 2019)	2018	Big5	Facebook	*Vector Semantic Model*	Akurasi: 64%
Personality Predictions Based on User Behaviour on the Facebook Social Media Platform (Tadesse et al., 2018)	2018	Big5	Facebook	*XGBoost*	Akurasi: 74.2%
25 Tweets to Know You: A New Model to Predict Personality with Social Media (Arnoux et al., 2017)	2017	Big5	Twitter	*Word Embedding with Gaussian Processes Regression*	33% lebih baik dari jurnal rujukan
Is Big Five better than MBTI? A personality computing challenge using Twitter data (Celli dan Lepri, 2018)	2018	Big5 & MBTI	Twitter	*Support Vector Machine*	Akurasi Big5: 61.5% Akurasi MBTI: 64.6%
Reddit: A Gold Mine for Personality Prediction (Gjurković dan Šnajder, 2018)	2018	MBTI	Reddit	*Support Vector Machine* (SVM), 2-*Regularized Logistic Regression* (LR) dan *Three-Layer Multilayer Perceptron* (MLP)	F1: 67% - 82% Akurasi: 82%

Tabel 3. Generalisasi MBTI

Guardian		***Artisan***	
ESFJ	ESTJ	ESFP	ESTP
ISFJ	ISTJ	ISFP	ISTP
Rational		***Idealist***	
ENTP	ENTJ	ENFP	ENFJ
INTP	INTJ	INFP	INFJ

Teknik augmentasi data dicoba diimplementasikan pada penelitian ini untuk meningkatkan performa dari *task text mining* yang memiliki jumlah dataset sedikit. Seperti ditunjukkan pada Tabel 2 penggunaan teknik augmentasi data masih jarang dilakukan untuk meningkatkan performa pada penelitian terkait prediksi kepribadian dengan jumlah dataset yang minim. Pada penelitian ini teknik augmentasi data diimplementasikan untuk meningkatkan akurasi dari model yang dibangun

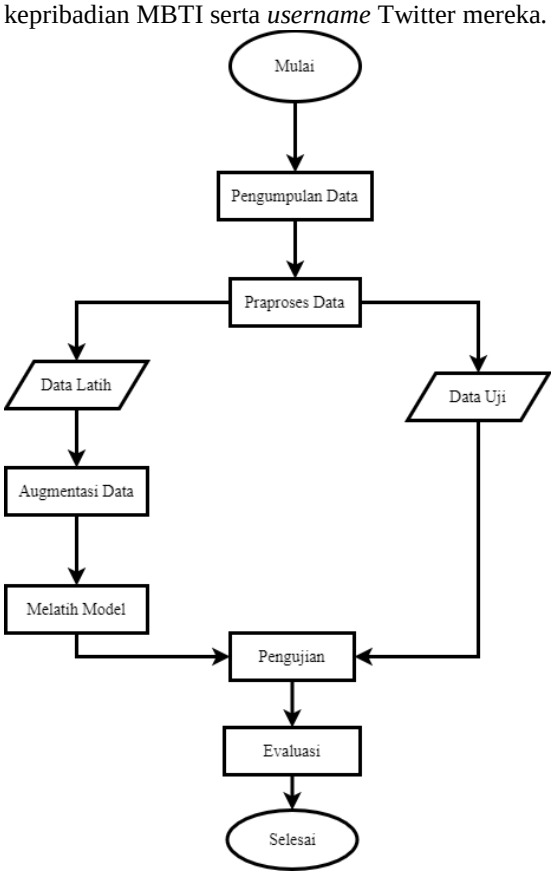
untuk memprediksi kepribadian MBTI dengan jumlah dataset yang sedikit.

2. METODE PENELITIAN

Diagram alir yang dibangun pada penelitian ini digambarkan melalui Gambar 1.

2.1. Pengumpulan Data

Pengumpulan data dilakukan dengan menyebarkan kuesioner kepada koresponden yang memiliki akun Twitter tidak di-*lock* dalam kurun waktu penyebaran kuisioner selama satu bulan. Koresponden diminta untuk mengisikan jenis kepribadian MBTI serta *username* Twitter mereka.



Gambar 1. Diagram sistem yang diimplemtasikan pada penelitian

Rata-rata koresponden adalah mahasiswa berusia 18-22 tahun. Data MBTI yang peserta isi kemudian digeneralisasi ke dalam empat kelas. Berdasarkan hasil kuesioner, didapatkan 97 akun Twitter yang dapat dilakukan *scraping* dengan jumlah keseluruhan sebanyak 671286 *tweet*. Berdasarkan 97 akun tersebut terdapat 21 akun yang memiliki kelas *Rational*, 28 akun memiliki kelas *Artisan*, 26 akun memiliki kelas *Guardian*, dan 22 akun memiliki kelas *Idealist.*

2.2. Praproses Data

Praproses data merupakan tahap yang dilakukan untuk menyajikan dokumen teks ke dalam format kata yang jelas dengan mereduksi fitur yang terdapat pada data masukan sehingga siap untuk diproses (Korde, 2012). Praproses data yang dilakukan pada penelitian ini adalah tokenisasi, *cleaning*, *case folding*, *remove punctuation*, normalisasi kata, penghapusan angka, *stemming* dan lematisasi, serta penghapusan *stopwords.*

Tokenisasi adalah proses pemecahan suatu kalimat menjadi daftar token atau kata (Allahyari et al., 2017). Contoh dari tokenisasi adalah mengubah kalimat "Saya akan pergi" menjadi "Saya", "akan", "pergi". Setelah melakukan tokenisasi, praproses yang dilakukan selanjutnya adalah *cleaning. Cleaning* merupakan proses pembersihan data dari simbol-simbol khusus yang tidak memiliki makna. Tujuan dari *cleaning* yaitu menghilangkan *noise* pada data. Adapun simbol yang dihilangkan pada step ini adalah URL, spasi berlebih, serta simbol-simbol khusus yang muncul pada data, seperti "&amp" dan "&yt". Setelah proses *cleaning* dilakukan maka selanjutnya akan dilakukan *case folding*, yakni proses konversi teks pada dokumen menjadi satu *case*. Pada penelitian ini, *case folding* yang dilakukan pada data adalah mengubah teks menjadi *lowercase*. Selanjutnya dilakukan proses *remove punctuation*, yaitu penghapusan tanda baca dan karakter lainnya, seperti "#", "@", "$", "%", "+", "-", "=", "*", "^". Penghapusan ini dilakukan karena simbol dan tanda baca tidak memberikan informasi terhadap data.

Praproses yang dilakukan selanjutnya adalah normalisasi kata. Normalisasi kata adalah proses mengubah kata yang tidak baku baik secara bahasa atau penulisan menjadi kata yang baku tanpa mengubah maknanya, seperti kata "ga" menjadi "tidak". Proses ini dilakukan dengan membuat suatu kamus yang berisi pasangan kata baku dan tidak baku, kemudian jika terdapat kata tidak baku pada data maka kata tersebut akan diubah menjadi kata baku sesuai kamus. Proses ini juga mengubah singkatan menjadi kepanjangannya seperti "DPR" menjadi "Dewan Perwakilan Rakyat". Kamus normalisasi pada penelitian ini dibentuk secara manual dengan membaca *tweet* satu per satu dan menambahkan setiap kata tidak baku yang ditemukan ke dalam kamus. Jumlah kata yang dinormalisasi yaitu sebanyak 1332 kata. Setelah itu dilakukan penghapusan angka, setiap angka yang ada pada data akan dihilangkan pada proses ini. Penghapusan angka dilakukan karena kemunculan angka tidak terlalu memiliki makna. Setelah penghapusan angka, dilakukan proses *stemming* dan lematisasi. *Stemming* merupakan proses pemotongan imbuhan sufiks dan prefiks pada kata sedangkan lematisasi adalah proses pengembalian kata ke bentuk dasarnya sesuai dengan kamus (Allahyari et al., 2017). Tahap praproses akhir yang dilakukan adalah penghapusan *stopwords. Stopwords* merupakan daftar kata yang tidak membawa informasi sehingga tidak memiliki pengaruh khusus pada teks, seperti kata "yang", "dan", "di", "ke", dan "dari" (Gaigole, Patil dan Chaudhari, 2013).

Setelah praproses data selesai dilakukan, maka selanjutnya akan dilakukan pembagian data menjadi data uji dan data latih. Berdasarkan hasil pengujian yang telah dilakukan, pembagian data 90%:10% memberikan hasil akurasi yang lebih baik dibandingkan pembagian data 70%:30% serta 80%:20%. Oleh karena itu, pada penelitian ini data dibagi menjadi 90% (87) data latih dan 10% (10) data uji.

2.4. Augmentasi Data

Penelitian (Wei dan Zou, 2019) membuktikan proses augmentasi data efektif dilakukan untuk meningkatkan performa dari klasifikasi teks yang memiliki data berukuran kecil. Beberapa operasi yang dibahas pada penelitian tersebut adalah *Synonym Replacement, Random Insertion, Random Swap,* dan *Random Deletion*.

Teknik augmentasi yang dilakukan pada penelitian ini adalah *Random Deletion*. Teknik ini dilakukan dengan menghapus kata secara acak dalam kalimat dengan probabilitas *p* kemudian menambahkan kalimat baru hasil penghapusan tersebut menjadi dataset baru. Hal ini dilakukan untuk memperkaya dataset dengan jumlah sedikit. Nilai *p* menentukan probabilitas tiap kata terhapus. Jika digunakan nilai *p* sebanyak 5% maka terdapat 5% kemungkinan suatu kata akan terhapus dari kalimat pada suatu baris data. Hasil penghapusan kalimat tersebut akan menjadi data baru yang ditambahkan pada dataset (Wei dan Zou, 2019). *Random Deletion* merupakan teknik yang paling sederhana untuk diimplementasikan dibanding teknik augmentasi data yang lain karena algoritmanya yang sederhana serta tidak membutuhkan data tambahan. Hal ini menyebabkan penelitian ini menggunakan teknik ini. Untuk dapat mengimplementasikan *Synonym Replacement*, dibutuhkan korpus sinonim dari data yang dimiliki. *Random Insertion* membutuhkan daftar kata tambahan yang sekiranya akan dimasukkan pada data, sedangkan untuk *Random Swap*, algoritma yang akan dibangun lebih kompleks untuk menggeser urutan tiap kata dan menjaga agar susunan kata tersebut tetap bermakna.

Teknik augmentasi data dilakukan terhadap data latih. Proses ini dilakukan setelah data dibagi menjadi data latih dan data uji. Penggunaan *Random Deletion* menyebabkan penambahan data latih dari jumlah awal 87 menjadi 1395 data latih.

2.3. Melatih Model dan Pengujian

Pada penelitian ini model dilatih menggunakan empat metode, yaitu *Gradient Boosting, Multinomial Naïve Bayes, K-Nearest Neighbor* dan *Random Forest*. Metode *Multinomial Naïve Bayes* dan *K-Nearest Neighbor* kerap digunakan dalam penelitian yang berkaitan dengan *text mining* (Allahyari et al., 2017). Pada penelitian ini, selain pengimplementasian dua metode yang sering digunakan dalam *text mining*, digunakan metode *Gradient Boosting* dan *Random Forest* sebagai perbandingan metode terbaik yang dapat digunakan pada kasus ini.

Proses pelatihan model masing-masing metode dilakukan dengan beberapa skenario penggunaan fitur, salah satu penambahan fitur yang digunakan adalah hasil pembobotan *Term Weighting* menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF). TF-IDF merupakan gabungan dari metode *Term Frequency* (TF) dengan *Inverse Document Frequency* (IDF) yang dihasilkan dari perkalian nilai TF dengan nilai IDF. Metode TF-IDF memberikan nilai bobot yang tinggi kepada *term* yang sering muncul pada suatu dokumen, tetapi jarang muncul dalam kumpulan dokumen (Wu dan Gu, 2014). Berikut merupakan persamaan dari TF-IDF.

$$TF * IDF(d,t) = TF(d,t) * log\ \frac{N}{df(t)} \quad (1)$$

Dengan $TF(d,t)$ merupakan frekuensi munculnya *term t* pada dokumen *d*, *N* merupakan jumlah dari dokumen, dan $df(t)$ merupakan jumlah dokumen yang mengandung *term t.*

2.4. Evaluasi Performa

Evaluasi performa digunakan untuk mengukur seberapa akurat metode yang diimplementasikan pada sistem. Pengukuran evaluasi performa dapat dilakukan menggunakan *confusion matrix*. Tabel 4 menunjukkan tabel dari *confusion matrix.*

Tabel 4. Confusion Matrix

Kategori		**Kelas Prediksi**	
		Positif	Negatif
Kelas Aktual	Ya	TP	FN
	Tidak	FP	TN

Pada penelitian ini, parameter pengukuran performa yang digunakan adalah akurasi. Akurasi digunakan untuk mengevaluasi banyaknya hasil prediksi yang sama dengan data aktual, semakin tinggi nilai akurasi maka performa dari metode yang digunakan semakin baik. Akurasi memiliki persamaan sebagai berikut.

$$akurasi = \frac{TP+TN}{TP+FP+TN+FN}\ x\ 100\% \quad (2)$$

3. HASIL DAN PEMBAHASAN

Pada penelitian ini digunakan empat metode klasifikasi dengan TF-IDF sebanyak 10000 fitur dan fitur tambahan yang tersedia pada Twitter yaitu, jumlah *following*, jumlah *followers*, dan jumlah tweet yang pengguna miliki. *Following* merupakan akun lain yang diikuti oleh akun tersebut, *followers*

adalah akun lain yang mengikuti akun tersebut, sedangkan jumlah *tweet* adalah banyaknya pesan yang pernah diunggah oleh pemilik akun tersebut.

Data didapatkan dari proses pengumpulan data berupa penyebaran kuesioner test MBTI kepada koresponden yang memiliki akun Twitter tidak diprivasi kemudian dilakukan *scrap* untuk mengambil data berupa *tweet* dari koresponden tersebut. Proses *scraping* dilakukan dengan menggunakan Twint. Twint merupakan *tools* yang dapat digunakan untuk melakukan *scraping tweet* dari Twitter sesuai persyaratan yang dibutuhkan berupa *tweet* dengan kata atau frasa tertentu, *tweet* dari akun tertentu, *tweet* yang diunggah pada waktu tertentu dan sebagainya. Pengaturan persyaratan *tweet* yang akan di-s*craping* dapat dilakukan melalui CLI atau *code module* menggunakan bahasa pemrograman python. Pada penelitian ini, proses *scraping* dilakukan berdasarkan *username* akun koresponden yang telah mengisi kuesioner. Data yang diambil pada akun koresponden yaitu *tweet* pengguna, jumlah *following*, jumlah *followers*, jumlah *tweet*, sedangkan kelas kepribadian MBTI diambil dari hasil kuesioner. Sebanyak 97 dari 124 koresponden memiliki akun yang dapat dilakukan proses *scraping*, sedangkan sisanya tidak dapat dilakukan proses *scraping* karena tidak ada *tweet* yang diunggah serta akun yang diprivasi. Hasilnya terdapat 21 akun yang memiliki kelas *Rational*, 28 akun memiliki kelas *Artisan*, 26 akun memiliki kelas *Guardian*, dan 22 akun memiliki kelas *Idealist* dengan total 671286 *tweet.*

Beberapa skenario diterapkan untuk mendapatkan model prediksi dengan hasil terbaik.

3.1. Fitur dan Metode Klasifikasi Terbaik

Skenario pengujian pertama dilakukan dengan eksperimen terhadap setiap kombinasi fitur dan model klasifikasi. Beberapa skenario fitur yang diuji yaitu fitur TF-IDF, fitur tambahan, dan campuran fitur TF-IDF dengan fitur tambahan. Skenario ini dilakukan dengan praproses penuh. Tabel 5 menunjukkan hasil dari skenario I.

Tabel 5. Nilai Akurasi Skenario I

METODE	**AKURASI**		
	TF-IDF	**Fitur Tambahan**	**TF-IDF + Fitur Tambahan**
Gradient Boosting	10.00%	10.00%	30.00%
Multinomial Naïve Bayes	20.00%	10.00%	20.00%
Random Forest	30.00%	30.00%	**40.00%**
K-Nearest Neighbor	20.00%	20.00%	20.00%

Tabel 5 menunjukkan bahwa *Random Forest* memiliki akurasi paling tinggi menggunakan fitur gabungan dari TF-IDF dan fitur tambahan dengan nilai akurasi sebesar 40%. Penggunaan fitur tambahan saja tidak memberikan hasil prediksi yang cukup baik. Penggunaan *Random Forest* secara umum memberikan nilai akurasi yang cukup baik saat digunakan bersamaan dengan penggunaan fitur TF-IDF karena penggunaan TF-IDF dapat mendukung *Random Forest* bekerja lebih baik.

3.2. Efek Praproses

Skenario pengujian kedua dilakukan dengan merekayasa praproses. Hal ini dilakukan untuk mendapatkan teknik praproses terbaik. Skenario praproses dilakukan dengan tiga percobaan yaitu praproses penuh, praproses tanpa penghapusan *stopwords*, dan praproses tanpa *stemming* dan penghapusan *stopwords*. Skenario ini dijalankan menggunakan metode dan fitur terbaik hasil dari skenario 1, yakni metode *Random Forest* dengan penggunaan fitur penuh. Hasil dari skenario pengujian kedua menggunakan fitur dan model klasifikasi terbaik ditunjukkan pada Tabel 6.

Tabel 6. Nilai Akurasi Skenario II

Praproses	**Akurasi**
Praproses penuh	**40%**
Tanpa penghapusan *stopwords*	30%
Tanpa *stemming* dan penghapusan *stopwords*	30%

Hasil akurasi terbaik didapatkan dengan melakukan praproses penuh terhadap data. Menghilangkan tahap penghapusan *stopwords* tidak efektif dilakukan pada eksperimen ini karena kata seperti “yang”, “itu”, serta “dan” tidak memiliki informasi khusus yang dapat mempengaruhi prediksi kepribadian seorang individu. Oleh karena itu, nilai akurasi praproses yang menghilangkan tahap penghapusan *stopwords* tidak cukup baik dibandingkan dengan praproses penuh.

Selain itu, penghilangan proses *stemming* juga memberikan hasil yang tidak cukup baik. Hal ini dikarenakan tidak ada kesalahan proses *stemming* pada data penelitian ini, seperti kesalahan perubahan makna akibat pemotongan kata berimbuhan.

3.3. Augmentasi Data

Proses augmentasi data dilakukan menggunakan skenario terbaik yang didapatkan dari skenario I dan skenario II, yakni menggunakan *Random Forest* sebagai *classifier* dengan fitur tambahan, TF-IDF, dan praproses penuh. Teknik augmentasi data yang digunakan adalah *Random Deletion* dengan presentasi p={5%, 10%, 20%, 40%, dan 50%}. Hasil dari proses augmentasi data ditunjukkan pada Tabel 7.

Tabel 7. Nilai Akurasi Skenario III

Probabilitas	Akurasi
5%	60%
10%	**70%**
20%	30%
40%	50%
50%	40%

Tabel 7 menunjukkan hasil bahwa penggunaan proses augmentasi data dapat meningkatkan akurasi menjadi 70% dari akurasi awal sebesar 40%. Hal ini dapat berpengaruh dikarenakan proses augmentasi data dapat menambah jumlah data latih berdasarkan nilai probabilitas dari kata yang dihapuskan pada tiap kalimat dari setiap baris data serta jumlah iterasi yang dilakukan. Peningkatan jumlah data yang dilatih dapat meningkatkan performa sistem dalam memprediksi kepribadian MBTI seseorang. Namun, untuk mendapatkan peningkatan akurasi yang dinilai cukup baik, nilai probabilitas *p* perlu diatur sedemikian rupa karena nilai akurasi tidak berbanding lurus dengan nilai *p*. Seperti pada Tabel 6, peningkatan akurasi cukup baik dengan menggunakan nilai *p* sebesar 10%. Nilai *p* yang cukup besar seperti 50% tidak berpengaruh untuk meningkatkan akurasi.

4. KESIMPULAN

Pada penelitian ini dilakukan prediksi kepribadian berdasarkan *Myers-Briggs Type Indicator* (MBTI) pengguna Twitter menggunakan metode *Gradient Boosting*, *Random Forest*, *K-Nearest Neighbor* serta *Multinomial Naïve Bayes*. Beberapa skenario diterapkan untuk mendapatkan *classifier* terbaik, fitur terbaik, serta teknik praproses terbaik. Hasil terbaik didapatkan menggunakan *classifier Random Forest* dengan fitur gabungan TF-IDF dan fitur tambahan, yakni *following*, *followers*, dan jumlah *tweet*, menggunakan teknik praproses penuh. Teknik augmentasi data dengan *Random Deletion* dilakukan untuk meningkatkan performa. Hasil eksperimen menunjukkan bahwa penggunaan teknik augmentasi data dapat meningkatkan akurasi menjadi 70% dari akurasi awal sebesar 40% dengan menggunakan nilai probabilitas penghapusan secara acak sebesar 10%. Nilai akurasi dapat berupa bilangan bulat dikarenakan data uji yang digunakan berjumlah 10 data.

Hasil dari penelitian ini menunjukkan bahwa teknik augmentasi data cukup efektif dilakukan untuk meningkatkan performa *text mining task* dengan jumlah dataset sedikit. Hal ini terlihat pada peningkatan akurasi dari model prediksi kepribadian MBTI yang dibangun.

Pengembangan penelitian selanjutnya dapat dilakukan dengan menambah jumlah dataset dan jumlah *tweet* yang seimbang pada setiap kelas untuk mendapatkan akurasi yang lebih baik. Pengembangan penelitian juga dapat dilakukan dengan mengimplementasikan *classifier*, fitur, teknik augmentasi data serta teori kepribadian yang lain.

DAFTAR PUSTAKA

ALLAHYARI, M., POURIYEH, S., ASSEFI, M., SAFAEI, S. DAN TRIPPE, E.D., 2017. *A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques*. *Cornell University*, .

AMALIYAH, M. DAN NOVIYANTO, F., 2013. Aplikasi Tes Kepribadian untuk Penempatan Karyawan Menggunakan Metode MBTI (Myers-Briggs Type Indicator) Berbasis Web (Studi Kasus : PT. Winata Putra Mandiri). *Jurnal Sarjana Teknik Informatika*, 1(2), pp.607–616.

ARNOUX, P.H., XU, A., BOYETTE, N., MAHMUD, J., AKKIRAJU, R. DAN SINHA, V., 2017. 25 tweets to know you: A new model to predict personality with social media. *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*, pp.472–475.

AZUCAR, D., MARENGO, D. DAN SETTANNI, M., 2018. Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences*, [online] 124(December 2017), pp.150–159. Available at: <https://doi.org/10.1016/j.paid.2017.12.018>.

BAI, S., YUAN, S., HAO, B. DAN ZHU, T., 2014. Predicting personality traits of microblog users. *Web Intelligence and Agent Systems*, 12(3), pp.249–265.

CELLI, F. DAN LEPRI, B., 2018. Is big five better than MBTI? A personality computing challenge using Twitter data. *CEUR Workshop Proceedings*, 2253.

GAIGOLE, P.C., PATIL, L.H. DAN CHAUDHARI, P.M., 2013. Preprocessing Techniques in Text Categorization. *National Conference on Innovative Paradigms in Engineering & Technology*, pp.1–3.

GJURKOVIĆ, M. DAN ŠNAJDER, J., 2018. Reddit: A Gold Mine for Personality Prediction. pp.87–97.

GOLBECK, J., 2016. Predicting Personality from Social Media Text. *AIS Transactions on Replication Research*, 2(September), pp.1–10.

HASSANEIN, M., HUSSEIN, W., RADY, S. DAN GHARIB, T.F., 2019. Predicting Personality Traits from Social Media using Text Semantics. *Proceedings - 2018 13th*

International Conference on Computer Engineering and Systems, ICCES 2018, pp.184–189.

KORDE, V., 2012. Text Classification and Classifiers:A Survey. *International Journal of Artificial Intelligence & Applications*, 3(2), pp.85–99.

LIMA, A.C.E.S. DAN DE CASTRO, L.N., 2019. Tecla: A temperament and psychological type prediction framework from Twitter data. *PLoS ONE*, 14(3), pp.1–18.

SURVEY, A.F., 2013. MBTI Personality Types of Project Managers and Their Success : (June).

TADESSE, M.M., LIN, H., XU, B. DAN YANG, L., 2018. Personality Predictions Based on User Behavior on the Facebook Social Media Platform. *IEEE Access*, 6(c), pp.61959–61969.

TANDERA, T., HENDRO, SUHARTONO, D., WONGSO, R. DAN PRASETIO, Y.L., 2017. Personality Prediction System from Facebook Users. *Procedia Computer Science*, [online] 116, pp.604–611. Available at: <https://doi.org/10.1016/j.procs.2017.10.016>.

TIGHE, E. DAN CHENG, C., 2018. Modeling Personality Traits of Filipino Twitter Users. pp.112–122.

WEI, J. DAN ZOU, K., 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. pp.6381–6387.

WU, H. DAN GU, X., 2014. Reducing over-weighting in supervised term weighting for sentiment analysis. *COLING 2014 - 25th International Conference on Computational Linguistics, Proceedings of COLING 2014: Technical Papers*, pp.1322–1330.

ZHENG, H. DAN WU, C., 2019. Predicting personality using facebook status based on semi-supervised learning. *ACM International Conference Proceeding Series*, Part F148150, pp.59–64.

MARCUS, B., MACHILEK, F. & SCHÜTZ, A., 2006. Personality in cyberspace: Personal Web sites as media for personality expressions & impressions. *Journal of Personality & Social Psychology*, 90(6), pp.1014–1031.

QUERCIA, D., KOSINSKI, M., STILLWELL, D. & CROWCROFT, J., 2011. Our twitter profiles, our selves: Predicting personality with twitter. *Proceedings - 2011 IEEE International Conference on Privacy, Security, Risk & Trust & IEEE International Conference on Social Computing, PASSAT/SocialCom 2011*, pp.180–185.

ZUKHRUFILLAH, I., 2018. Gejala Media Sosial Twitter Sebagai Media Sosial Alternatif. *Al-I'lam: Jurnal Komunikasi dan Penyiaran Islam*, 1(2), p.102