JURNAL INFORMATIKA DAN TEKONOLOGI KOMPUTER
Halaman Jurnal: https://journal.amikveteran.ac.id/index.php/jitek
Halaman UTAMA Jurnal : https://journal.amikveteran.ac.id/index.php

KLASIFIKASI KEPRIBADIAN MENGGUNAKAN ALGORITMA MACHINE LEARNING

Mawadatul Maulidah
Program Studi Teknologi Komputer Kampus Kota Tegal Komputer, Universitas Bina Sarana Informatika
mawadatul.mwm@bsi.ac.id

ABSTRACT

Myers-Briggs Personality Type (MBTI) is a popular personality metric that uses four dichotomies as indicators of personality traits. This study uses a public dataset from Kaggle, namely the Myers-Briggs Personality Type Dataset, the model tested is several machine learning classification models with the help of imlearn under-over sampling techniques for classifying MBTI personality types. This study aims to classify the Myers-Briggs Type Indicator (MBTI) personality type based on text from user posts on the social media platform Reddit. The dataset used in this study consists of around 8,000 posts collected from the MBTI subreddit. Several text processing methods such as tokenization, punctuation removal, and stemming are used to process the raw data before it is entered into the model. The experimental results show that the LSTM model using Adam's optimizer and a learning rate of 0.01 produces good performance with an accuracy of 80.73 compared to other machine learning models. In addition to the LSTM model, XG Boost is also a classification model with the highest accuracy based on 16 personality types producing an accuracy of 60.09 and Logistic Regression with the NS dimension as the best accuracy value of 87.21%.

***Keywords**: classification algorithm, personality type, mbti*

ABSTRAK

Myers-Briggs Personality Type (MBTI) adalah metrik kepribadian populer yang menggunakan empat dikotomi sebagai indikator sifat kepribadian. Penelitian ini menggunakan dataset publik dari Kaggle yaitu Myers-Briggs Personality Type Dataset, model yang diujikan adalah beberapa model klasifikasi machine learning dengan bantuan teknik pengambilan sampel imlearn under-over sampling untuk klasifikasi tipe kepribadian MBTI. Penelitian ini bertujuan untuk klasifikasi jenis kepribadian Myers-Briggs Type Indicator (MBTI) berdasarkan teks dari postingan pengguna di platform sosial media Reddit. Dataset yang digunakan dalam penelitian ini terdiri dari sekitar 8.000 postingan yang dikumpulkan dari subreddit MBTI. Beberapa metode pengolahan teks seperti tokenisasi, penghapusan tanda baca, dan stemming digunakan untuk memproses data mentah sebelum dimasukkan ke dalam model . Hasil eksperimen menunjukkan bahwa model LSTM dengan menggunakan Adam's optimizer dan learning rate 0,01 menghasilkan kinerja yang baik dengan akurasi 80,73 dibandingkan dengan model machine learning lainnya. Selain model LSTM, XG Boost juga merupakan model klasifikasi dengan akurasi tertinggi berdasarkan 16 tipe kepribadian menghasilkan akurasi sebesar 60,09 dan Regresi Logistik dengan dimensi NS sebagai nilai akurasi terbaik sebesar 87,21%.

Kata kunci: algoritma klasifikasi, tipe kepribadian, mbti

1. PENDAHULUAN

Kepribadian memainkan peran penting dalam memprediksi banyak faktor individu seperti kesehatan mental dan fisik, kebugaran, dan kesejahteraan karier. Oleh karena itu, mendapatkan wawasan mendalam tentang tipe kepribadian seseorang adalah kuncinya. Indikator Tipe Myers-Briggs (MBTI) adalah indikator persepsi individu tentang dunia dan proses pengambilan keputusan. Katherine Cook dan putrinya, Isabel Briggs Myers, adalah orang pertama yang menemukan metode ini, yang merupakan pengembangan dari teori kepribadian yang sebelumnya dikemukakan oleh Carl Gustav Jung. MBTI telah berfungsi sebagai instrumen yang digunakan orang dalam upaya untuk lebih memahami keyakinan dan motivasi mereka sendiri. Meskipun reliabilitas dan validitas MBTI sering dikritik, MBTI tetap menjadi ukuran kepribadian yang paling banyak digunakan.[1].

Received Januari 11, 2023; Revised Maret 2, 2023; Accepted Maret 31, 2023

Ada segudang penelitian yang bertujuan merangkum berbagai aspek penelitian kepribadian. Mohammad Hossein Amirhosse melakukan studi tentang Pendekatan Pembelajaran Mesin untuk Prediksi Tipe Kepribadian Berdasarkan Indikator Tipe Myers–Briggs pada tahun 2020. Studi ini mengembangkan metode pembelajaran mesin baru untuk mengotomatiskan proses deteksi program meta dan prediksi tipe kepribadian berdasarkan indikator tipe kepribadian MBTI . Toolkit pemrosesan bahasa alami (NLTK) dan XGBoost, yang didasarkan pada pustaka Gradient Boosting dengan Python, digunakan untuk mengimplementasikan algoritme pembelajaran mesin. Abidin, et., al., (2020) membandingkan kinerja metode yang diusulkan dengan algoritma pembelajaran mesin populer lainnya. Evaluasi eksperimental menunjukkan bahwa pengklasifikasi Hutan Acak bekerja lebih baik daripada tiga algoritme pembelajaran mesin yang berbeda dalam hal akurasi, sehingga membantu pemberi kerja mengidentifikasi tipe kepribadian untuk memilih kandidat yang sesuai. Agastya et., al., (2019) dan Mehta et., al., (2019) telah mensurvei studi prediksi kepribadian berdasarkan beberapa sumber data dan modalitas. Keduanya memberikan perhatian khusus pada pendekatan berbasis pembelajaran yang mendalam. Jacques et., al., (2018) dan Escalera et al., (2018) telah membatasi ruang lingkup tinjauan untuk prediksi kepribadian yang terlihat. Majumder et., al., (2017) menggunakan deep CNN untuk tugas deteksi kepribadian tingkat dokumen. CNN membantu mengekstrak fitur monogram, bigram, dan trigram dari teks. Telah diamati bahwa penghapusan kalimat netral memberikan peningkatan yang nyata dalam akurasi prediksi. Setiap kata direpresentasikan dalam input sebagai vektor fitur dengan panjang tetap menggunakan Word2Vec dan kalimat direpresentasikan sebagai jumlah variabel dari vektor kata. Pada akhirnya, fitur tingkat dokumen Mairesse (LIWC, MRC, dll.), Sebanyak 84 fitur, digabungkan dengan vektor fitur yang diekstraksi dari CNN dalam. Akhirnya, vektor gabungan ini kemudian dimasukkan ke dalam lapisan yang terhubung sepenuhnya untuk prediksi sifat kepribadian akhir.

Selanjutnya pada penelitian lain, Hernandez dan Knight (2017) menggunakan berbagai jenis recurrent neural network (RNN) seperti simple RNN, gated recurrent unit (GRU) yang merupakan mekanisme gating pada recurrent neural network, long short-term memory (LSTM) yang merupakan arsitektur jaringan saraf. pengulangan buatan yang digunakan dalam pembelajaran mendalam, dan Bidirectional LSTM untuk membangun kelas yang mampu memprediksi tipe kepribadian MBTI seseorang berdasarkan sampel teks dari postingan media sosial mereka. Kumpulan Data Tipe Kepribadian Myers–Briggs dari Kaggle digunakan dalam penelitian mereka. Mereka membandingkan hasilnya dan menemukan bahwa LSTM memberikan hasil terbaik. Studi terbaru oleh Cui dan Qi (2017) menggunakan Baseline, Logistic Regression, Naïve Bayes, dan SVM untuk memprediksi tipe kepribadian MBTI seseorang dari salah satu postingan media sosial mereka. Mereka membandingkan hasil dari semua metode ini dan menemukan bahwa kinerja SVM lebih baik. Mereka menggunakan database yang sama yang digunakan dalam penelitian sebelumnya, Myers–Briggs Personality Type Data Set of Kaggle.

Dari beberapa penelitian terkait yang telah diuraikan di atas, hanya menerapkan teknik evaluasi kinerja berupa nilai akurasi tanpa memperhatikan kondisi data dalam keadaan seimbang atau tidak. Kondisi data yang tidak seimbang terjadi pada dataset kepribadian MBTI dimana beberapa tipe kepribadian memiliki data yang lebih banyak dari yang lain. Masalah ketidakseimbangan merupakan masalah yang muncul yaitu nilai performance klasifikasi menunjukkan nilai akurasi yang tinggi karena jumlah kelas mayor yang sangat banyak. Namun, ini sebenarnya memiliki performa klasifikasi yang sangat buruk saat mengklasifikasikan data dari kelas minor. Terkait masalah ketidakseimbangan data, salah satu strategi yang dapat diterapkan adalah teknik pengambilan sampel, baik oversampling maupun undersampling.

Berdasarkan penjelasan di atas, permasalahan utamanya adalah dataset yang tidak seimbang. Dengan demikian proses evaluasi model yang dihasilkan menjadi bias. Pada penelitian ini diusulkan penerapan teknik sampling untuk mengatasi masalah ketidakseimbangan data, baik oversampling maupun undersampling dalam proses prediksi kepribadian MBTI menggunakan LSTM. Long-Short Term Memory (LSTM) adalah salah satu jenis Machine Learning berdasarkan pendekatan Recurrent Neural Network yang dapat memprediksi tipe kepribadian MBTI saat ini. Dong et.al menunjukkan hal itu LSTM dapat dianggap sebagai solusi yang valid dan lebih baik daripada teknik lainnya. Jaringan LSTM menjadi pilihan terbaik berkat kemampuannya menyimpan memori dalam waktu yang lama dalam waktu yang bersamaan, korelasi kompleks antar data memberikan informasi yang sangat berguna dalam menentukan prediksi [9].

2. TINJAUAN PUSTAKA

2.1. Myers-Briggs Type Indicator (MBTI)

Teori Kepribadian Carl Jung didasarkan pada pengamatan klinis. Jung mendalilkan serangkaian empat fungsi kognitif yaitu berpikir, merasa, sensasi, dan intuisi, masing-masing memiliki satu atau dua kutub, totalnya adalah delapan fungsi dominan. Myers-Briggs Type Indicator (MBTI) adalah penilaian populer dalam pendidikan yang diusulkan untuk tujuan menunjukkan preferensi psikologis manusia yang berbeda yang mendasari minat, kebutuhan, nilai, dan motivasi.

Ada empat pasangan preferensi yang berlawanan: Introversion (I) dan Extraversion (E), Intuition (N) dan Sensing (S), Feeling (F) dan Thinking (T), dan Perceiving (P) dan Judging (J).

2.2. Data Mining

Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Interpretation / Evaluation pola informasi yang dihasilkan dari proses data mining diterjemahkan menjadi bentuk yang lebih mudah dimengerti oleh pihak yang berkepentingan [11].

Data mining memiliki fungsi utama untuk menerapkan berbagai teknik dan algoritma untuk mendeteksi dan mengekstrak pola dari set data yang disimpan dan diberikan [8]. *Data mining* diklasifikasikan ke dalam berbagai algoritma dan teknik, yaitu klasifikasi, *clustering*, regresi, aturan asosiasi, dll, yang digunakan untuk penemuan pengetahuan dari *database* [8].

2.3. Machine Learning

Machine Learning sering digambarkan sebagai pembelajaran dari pengalaman atau tanpa pengawasan dari manusia. Dalam masalah belajar yang diawasi, sebuah program memprediksi output untuk input dengan belajar dari pasangan input dan output berlabel; yaitu, program belajar dari contoh jawaban yang benar. Dalam tanpa pengawasan belajar, suatu program tidak belajar dari data berlabel. Machine Learning mengeksplorasi studi dan konstruksi algoritma yang dapat belajar dari dan membuat prediksi pada data. Algoritma tersebut beroperasi dengan membangun model dari input contoh untuk membuat prediksi berbasis data atau keputusan, daripada mengikuti instruksi program yang benar-benar statis [12].

2.4. Python

Python merupakan bahasa pemrograman yang berorientasi objek dinamis, dapat digunakan untuk bermacam-macam pengembangan perangkat lunak [12]. Phyton menyediakan dukungan yang kuat untuk integrasi dengan Bahasa pemrograman lain dan alat-alat bantu lainnya. Python hadir dengan pustaka-pustaka standar yang dapat diperluas serta dapat dipelajari hanya dalam beberapa hari.

3. METODOLOGI PENELITIAN

3.1. Dataset

Dataset yang digunakan dalam penelitian ini diambil dari forum kepribadian online di personalitycafe.com dan tersedia gratis di Kaggle [1], sebuah komunitas ilmu data online. Kumpulan data ini berisi postingan dari 8675 pengguna forum. Setiap pengguna memiliki 50 contoh teks yang dipisahkan dengan urutan '|||', dengan total 433.750 komentar pengguna. Dataset MBTI hanya memiliki 2 kolom yaitu type dan posts yang memiliki 16 tipe kepribadian MBTI (misalnya INTJ, ESFP). Setiap pengguna memiliki tipe MBTI yang diberi label dengan keempat dimensi: introversi (I) - ekstraversi (E), intuisi (N) - penginderaan (S), pemikiran (T) - perasaan (F), dan penilaian (J) - persepsi (P). Tidak ada nilai null dalam kumpulan data dan semua nilai bersifat tekstual, sehingga harus diubah menjadi numerik.

3.2. Pemrosesan Data

Preprocessing pada penelitian ini menggunakan beberapa fungsi NLTK. Tahap preprocessing pada penelitian ini terdiri dari tiga bagian utama yaitu data cleaning, tokenization, dan stopword removal. Semua itu merupakan komponen penting yang harus dilakukan pada dataset ini untuk mencapai kinerja yang optimal.

3.2.1. Pembersihan Data

Pembersihan data adalah komponen penting dari pembelajaran mesin dalam masalah klasifikasi dan merupakan teknik penting dalam preprocessing

fase untuk data mentah. Pembersihan data memungkinkan pengurangan kebisingan, ketidakkonsistenan, dan kesalahan dalam kumpulan data.

Kumpulan data yang digunakan dalam penelitian ini berisi 50 komentar dari setiap pengguna individu, dan karena memungkinkan hampir semua teks untuk diposting oleh pengguna, ini berisi sejumlah besar data yang berisi informasi yang tidak signifikan untuk tujuan kami.

a. Tautan dihapus dari kumpulan data ini karena sering berisi informasi tidak berarti yang dapat dilihat tanpa menjelajahi konten tautan (mis. http://www.youtube.com/watch?v=EY21CYqGaMw).
b. Semua teks dikapitalisasi karena informasi penting yang ingin kami periksa dalam kumpulan data ini melibatkan konteks dan kami memiliki sedikit kegunaan untuk mengklasifikasi permulaan kalimat. Selain itu, karena sifat sumber data yang informal, kapitalisasi merupakan sumber informasi yang tidak dapat diandalkan.
c. Informasi dalam tanda kurung siku dihilangkan karena biasanya berisi informasi yang diparafrasekan dalam teks yang sudah tertulis atau teks tidak masuk akal yang hanya akan mengaburkan data.
d. Tanda baca dan kata berhenti dihilangkan karena ini adalah praktik NLP standar karena kedua komponen tersebut umumnya tidak terlalu berpengaruh pada keseluruhan makna data tekstual untuk tujuan ini. Stopwords dalam bahasa Inggris mengacu pada kata-kata yang umumnya tidak mempengaruhi arti kalimat, seperti "the", "and", dan "was".
e. Kata-kata dengan angka telah dihapus dari kumpulan data ini karena jumlah urutan alfanumerik dan karakter unicode yang sangat tinggi (misalnya 'x93a', '").

3.2.2. Lemmitisasi

Kami preprocess posting dengan menggunakan teknik Lemmitization. Lemmatization adalah proses pengelompokan bentuk infleksi yang berbeda dari sebuah kata sehingga mereka dapat dianalisis sebagai satu item. Lemmatization mirip dengan stemming tetapi membawa konteks pada kata-kata, oleh karena itu kami menggunakan ini sebagai gantinya dalam model kami. Jadi itu menghubungkan kata-kata dengan arti yang mirip dengan satu kata.

3.2.3. Tokenisasi

Menggunakan tokenizer kata Keras, kami menandai 2500 kata paling umum dari teks lemmatisasi. Artinya, kata yang paling umum menjadi 1, kata yang paling umum kedua menjadi 2, dst. hingga 2500. Kata-kata lain dalam teks lemmat telah dihapus, sehingga pada titik ini teks tersebut berupa daftar dari bilangan bulat (dengan kosakata 1-2500).

3.2.4. Term Frequency-Inverse Document Frequency (TF-IDF)

Tf–idf untuk rekayasa fitur mengevaluasi seberapa relevan/penting sebuah kata bagi dokumen dalam kumpulan dokumen atau korpus. Saat kami melatih pengklasifikasi individu di sini, ini sangat berguna untuk menilai kata dalam algoritme pembelajaran mesin untuk Pemrosesan Bahasa Alami.
Untuk model kami, kami membuat vektor menggunakan count vectorizer dan tf-idf vectorizer agar kata-kata tetap muncul antara 10 hingga 50 posting.

3.3. Data Ketidakseimbangan (Data Imbalance)

Beberapa teknik berbeda untuk menangani kumpulan data yang tidak seimbang. Kelas teknik yang paling naif adalah pengambilan sampel: mengubah data yang disajikan ke model dengan mengecilkan kelas umum, oversampling (menduplikat) kelas langka, atau keduanya.

Sekelompok peneliti menerapkan berbagai teknik pengambilan sampel data modern yang komprehensif dengan modul kontribusi pembelajaran yang tidak seimbang untuk sklearn. Submodul ini diinstal sebagai bagian dari instalasi sklearn dasar secara default, imblearn mengimplementasikan over-sampling dan under-sampling menggunakan kelas khusus.

3.4. Model

3.4.1. Long Short Term Memory (LSTM)

LSTM pertama kali diperkenalkan pada tahun 1997 oleh Hochreiter dan Schmidhuber [10]. LSTM adalah jenis Recurrent Neural Network (RNN). LSTM sendiri dapat menemukan lapisan tersembunyi dari setiap sel

dan dirancang untuk menyimpan informasi sel sebelumnya. Metode LSTM digunakan dengan mengklasifikasikan data jangka panjang dengan menyimpannya di sel memori. Sampai penelitian ini banyak dilakukan oleh para peneliti dalam rangka mengembangkan metode LSTM, metode LSTM sendiri memiliki empat komponen utama yaitu: Gerbang Input, koneksi berulang, gerbang lupa dan gerbang keluaran [2].

Ini memungkinkan model untuk mengingat informasi untuk jangka waktu yang lama dan akibatnya memahami konteks dengan lebih baik. Fitur-fitur ini ideal untuk masalah NLP seperti ini karena konteks kata dalam kalimat dan kalimat dalam paragraf penting [2].

3.3.2. Model Klasifikasi

Selain menggunakan LSTM sebagai model prediksi, penelitian ini juga akan melatih beberapa model dengan mengklasifikasikan berbagai algoritma seperti Naive Bayes, Logistic Regression, K-Neighbours Classifier, Support Vector Machine, Random Forest Classifier, Gradient Boosting Classifier dan Extreme Gradient Boosting untuk setiap 16 jenis dan 4 dimensi target. Dari keenam model tersebut akan dianalisis dengan nilai akurasi sebagai nilai keluarannya.

Model dalam penelitian ini akan diimplementasikan dengan Python Library versi 3.0 dengan Google Colab sebagai toolsnya

4. HASIL DAN PEMBAHASAN

4.1. Dekripsi Data

Pada penelitian ini dataset yang digunakan memiliki 8675 data dan 2 atribut (type dan posts). Tipe-tipe dalam dataset ini memiliki 16 tipe kepribadian MBTI. Selain menggunakan 16 tipe kepribadian, penelitian ini juga memilih untuk membuat 4 dimensi kelas, satu untuk setiap kategori. Dan pada penelitian ini data akan dibagi menjadi 90% data latih dan 10% data uji. Selain itu, terdapat data yang digunakan sebagai validasi data sebesar 10%.

Distribusi nilai variabel unik dari 16 jenis:

INFP	: 1832	ENTJ	: 231
INFJ	: 1470	ISTJ	: 205
INTP	: 1304	: ENFJ	: 190
INTJ	: 1091	ISFJ	: 166
ENTP	: 685	ESTP	: 89
ENFP	: 675	ESFP	: 48
ISTP	: 337	ESFJ	: 42
ISFP	: 271	ESTJ	: 39

Distribusi jumlah postingan pada kelas 4 dimensi :
Introversi (I) / Ekstroversi (E): 1999 / 6676
Intuisi (N) / Penginderaan (S):1197 / 7478
Berpikir (T) / Merasa (F) : 4694 / 3981
Judging (J) / Perceiving (P) : 5241 / 3434

3.2. Long Short Term Memory

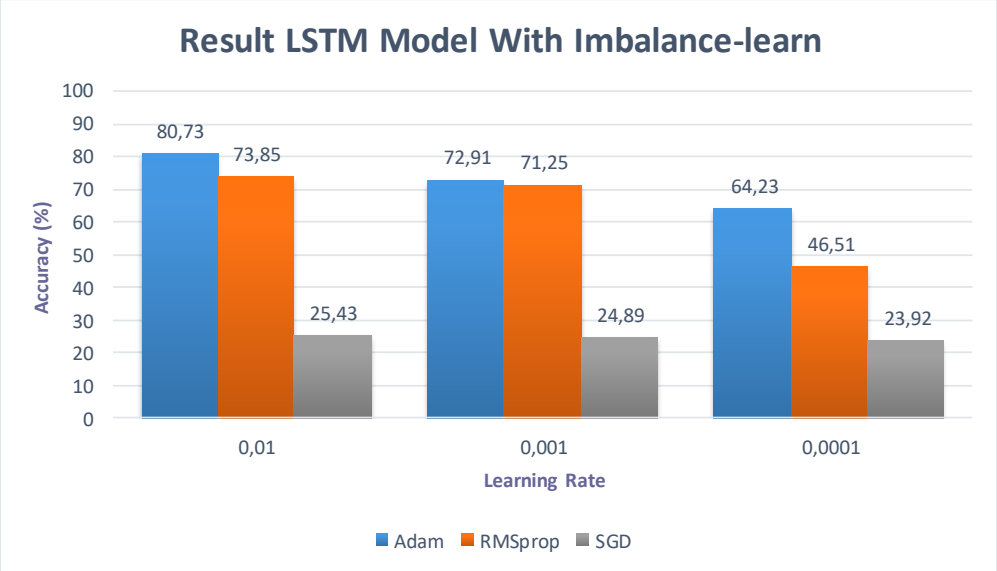
Pada model LSTM yang digunakan dalam penelitian ini, beberapa konfigurasi hyperparameter untuk pengujian menggunakan tiga varian learning rate (0.01, 0.001, 0.0001) dan tiga varian optimizer (Adam, RMSprop, SGD) dengan ukuran batch (128), epochs 10, max sequence length 2500, Panjang vektor input 640, batch 128, dropout 0,2, LSTM Unit 100 dan Aktivasi Softmax.

Berdasarkan hyperparameter yang telah dijelaskan, dilakukan pengujian menggunakan model LSTM dengan bantuan balanced-learn random over-under sampling dan hasilnya dapat dilihat pada tabel 1.

Tabel 1. Hasil LSTM dengan ROS

Tingkat Pembelajaran	Akurasi (%)		
	Adam	RMSprop	SGD
0,01	80,73	73,85	25.43
0,001	72.91	71.25	24.89
0,0001	64.23	46.51	23.92

Berdasarkan tabel 1 hasil penelitian model LSTM dengan pengujian hyperparameter yang telah dijelaskan pada tabel 1 dan dievaluasi dengan 4 dimensi kategori kelas, model LSTM yang memiliki kinerja baik adalah Adam Optimizer dengan learning rate 0,01.



Gambar 2. Diagram Hasil Model LSTM

Model Klasifikasi

Model klasifikasi menggunakan algoritma Naive Bayes, Logistic Regression, K-Neighbors Classifier, Support Vector Machine, Random Forest Classifier, Gradient Boosting Classifier dan Extreme Gradient Boosting. Beberapa metode klasifikasi yang digunakan dalam percobaan ini memiliki kinerja yang agak mirip ketika parameter dioptimalkan, seperti yang ditunjukkan pada Tabel 2.

Tabel 2. Hasil Model Klasifikasi

Penggolong	Akurasi (%)				
	Keseluruhan	I/N	N/S	F/T	J/P
XGB	60.09	78.80	87.10	69,59	64.29
LR	57.96	77.53	87.21	74.19	64.63
GD	48.87	78.23	86.98	73.50	65.09
RF	40.25	77.88	86.98	67.86	62.44
SVM	37.30	78.23	86.98	74.88	66.01
KNN	17.82	69.47	86.64	54.03	47.70
XGB ** parameter	-	78.46	86.98	71.08	63.36

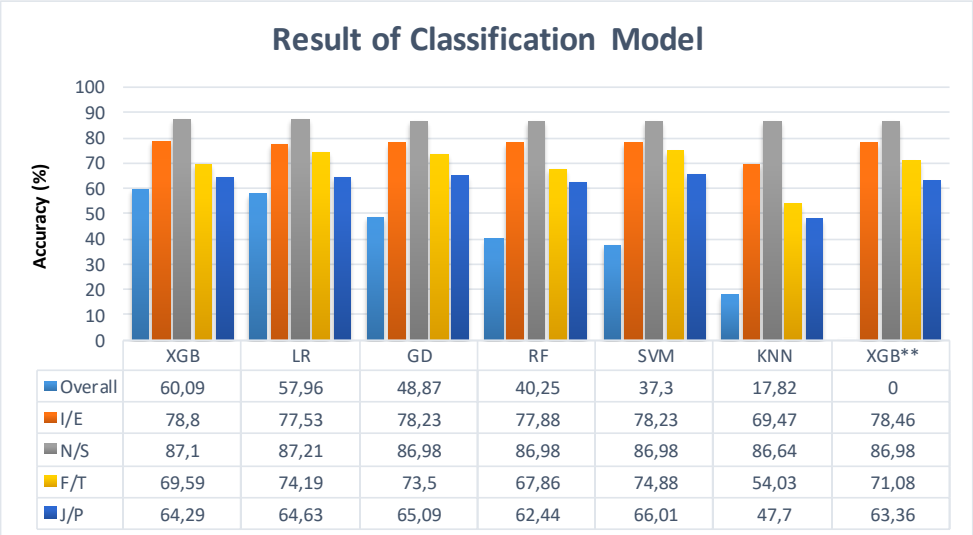
Berdasarkan tabel 2 hasil penelitian model klasifikasi dapat dijelaskan bahwa dari semua model penelitian yang telah dievaluasi berdasarkan 16 tipe kepribadian, XG Boost memberikan kinerja yang relatif baik yaitu 60,09, oleh karena itu kami memilihnya untuk dibangun. model prediksi Kepribadian kami. Ini akan berguna

karena model XGBoost [2] bahkan dapat digunakan untuk mengevaluasi dan melaporkan performa pada set pengujian untuk model selama pelatihan.

Konfigurasikan beberapa parameter untuk meningkatkan performa model XGBoost.

a. learning_rate di XGBoost harus disetel ke 0,1 atau lebih rendah, dan nilai yang lebih kecil akan memerlukan penambahan lebih banyak pohon.
b. tree_depth di XGBoost harus dikonfigurasi dalam rentang 2 hingga 8, di mana tidak banyak manfaat terlihat dengan pohon yang lebih dalam.

Sedangkan untuk kelas 4 dimensi diperoleh hasil terbaik pada Logistic Regression NS sebesar 87,21.



	XGB	LR	GD	RF	SVM	KNN	XGB**
Overall	60,09	57,96	48,87	40,25	37,3	17,82	0
I/E	78,8	77,53	78,23	77,88	78,23	69,47	78,46
N/S	87,1	87,21	86,98	86,98	86,98	86,64	86,98
F/T	69,59	74,19	73,5	67,86	74,88	54,03	71,08
J/P	64,29	64,63	65,09	62,44	66,01	47,7	63,36

Gambar 3. Diagram Hasil Model Klasifikasi

Berikut untuk hasil perbandingan yang diperoleh dari hasil implementasi beberapa model algoritma machine learning.

Tabel 3. Diagram Model Perbandingan

Algoritma	Akurasi (%)
XGB	60.09
LR	57.96
GD	48.87
RF	40.25
SVM	37.30
KNN	17.82
LSTM	80.73%

5. KESIMPULAN DAN SARAN

Penelitian ini bertujuan untuk membuat model dalam mengklasifikasikan tipe kepribadian MBTI. Dataset yang digunakan adalah dari Kaggle. Pemodelan yang digunakan dalam penelitian ini adalah Long Short Term Memory (LSTM), menggunakan tiga varian pengoptimal (Adam, RMSprop, SGD) dan learning rate (0.01, 0.001, 0.0001) dibantu dengan teknik pengambilan sampel data. Pemodelan LSTM dengan Adam optimizer dan learning rate 0,01 menjadi model LSTM yang baik dengan nilai akurasi 80,73%. Selain model LSTM, XG Boost juga merupakan model klasifikasi dengan akurasi tertinggi berdasarkan 16 tipe kepribadian menghasilkan akurasi sebesar 60,09 dan Regresi Logistik dengan dimensi NS sebagai nilai akurasi terbaik sebesar 87,21%.

DAFTAR PUSTAKA

[1] Mitchelle, J.; Kumpulan Data Jenis Kepribadian Myers-Briggs. Mencakup Banyak Jenis MBTI Orang dan Konten yang Ditulis oleh Mereka. Tersedia online: https://www.kaggle.com/datasnaek/mbti-type (diakses pada 4 Maret 2021).
[2] Cui B, Qi C. 2017. Analisis survei metode pembelajaran mesin untuk pemrosesan bahasa alami untuk prediksi tipe kepribadian MBTI. Tersedia di http://cs229.stanford. edu/ proj2017/final-reports/ 5242471.pdf.
[3] Bharadwaj S, Sridhar S, Choudhary R, Srinath R. 2018. Identifikasi ciri-ciri persona berdasarkan Indikator Tipe Myers-Briggs (MBTI) - pendekatan klasifikasi teks. Dalam: Konferensi Internasional 2018 tentang Kemajuan dalam Komputasi, Komunikasi dan Informatika (ICACCI). DOI 10.1109/icacci.2018.855482
[4] Li C, Hancock M, Bowles B, Hancock O, Perg L, Brown P , dkk, Wade R . 2018. Ekstraksi fitur dari postingan media sosial untuk pengetikan psikometri peserta. Augmented Cognition: Catatan Kuliah Teknologi Cerdas dalam Ilmu Komputer 267–286 DOI 10.1007/978-3-319-91470-1_23.
[5] Amirhosseini MH, Kazemian H. 2020. Pendekatan pembelajaran mesin untuk prediksi tipe kepribadian berdasarkan myers—briggs type indicator R. Teknologi Multimoda dan Interaksi 4(1):9 DOI 10.3390/mti4010009.
[6] Mehta Y, Fatehi S, Kazameini A, Stachl C, Cambria E, Eetemadi S. 2020. Bottom-up and top-down: memprediksi kepribadian dengan fitur psikolinguistik dan model bahasa. Di: Pada tahun 2020 konferensi internasional IEEE tentang penambangan data (ICDM). Piscataway: IEEE, 1184–1189.
[7] Keh SS, Cheng I. 2019. Klasifikasi kepribadian Myers-Briggs dan generasi bahasa spesifik kepribadian menggunakan model bahasa terlatih. Pracetak ArXiv. arXiv:1907.06333.
[8] Choong, En & Varathan, Kasturi. (2021). Memprediksi penilaian-persepsi Myers-Briggs Type Indicator (MBTI) di forum sosial online. TemanJ. 9.e11382. 10.7717/peerj.11382.
[9] Hernandez, R.; Knight, IS Memprediksi Indikator Tipe Myers-Bridge dengan klasifikasi teks. Dalam Prosiding Konferensi ke-31 tentang Sistem Pemrosesan Informasi Saraf, Long Beach, CA, AS, 4–9 Desember 2017. Tersedia online: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1184/ laporan/6839354.pdf (diakses pada 9 September 2018).
[10] Srilakshmi Bharadwaj, dkk. 2018. Identifikasi Sifat Persono Berdasarkan Myres-Briggs Type Indicator (MBTI) – Suatu Pendekatan Klasifikasi Teks. Konferensi Internasional 2018 tentang Kemajuan dalam Komputasi, Komunikasi dan Informatika, 19-22 September 2018, Banglore, India. hal.1076-1082.
[11] David Keirsey. Empat Temperamen. https://keirsey.com/temperamentoverview. Diakses tanggal 10 Februari 2020
[12] Pouria Kaviani, Sunita Dhorte. 2017. Survei Singkat Algoritma Naive bayes. International Journal of Advance Engineering and research Development, 4(11). hal.607-611.
[13] Bayu Yudha Pratama, Riyanarto Sarno. 2015. Klasifikasi Kepribadian Berdasarkan Teks Twitter Menggunakan Naive bayes, KNN, dan SVM. 2015. Konferensi Internasional Rekayasa Data dan Perangkat Lunak, 25-26. Nopember 2015, Yogyakarta, Indonesia. hal.170-174.
[14] Muhammad Fikry, Yusra. 2018. Ekstrover atau Introver : Klasifikasi Kepemilikan Pengguna Twitter Dengan Menggunakan Metode Support Vector Machine. Jurnal Sains, Teknologi, dan Industri, 16(1). hal.72-76.