

Klasifikasi Myers-Briggs Type Indicator Pengguna Twitter Menggunakan Support Vector Machine dan Augmentasi Data

Sulthan Athallah¹, Herdi Ashaury², Puspita Nurul Sabrina³

^{1, 2, 3} Jurusan Teknik Informatika, Fakultas Sains dan Informatika
Universitas Jenderal Achmad Yani
Jl. Terusan Jenderal Sudirman, Cimahi
Email : sulthanathallah05@gmail.com

ABSTRAK

Kepribadian merupakan aspek yang melekat pada seseorang, salah satu cara menentukannya yaitu dengan Myers-Briggs Type Indicator (MBTI). Beberapa penelitian terdahulu sudah melakukan klasifikasi MBTI menggunakan beberapa metode data mining. Salah satunya merupakan Support Vector Machine (SVM). Karena akurasi penelitian ini saat eksperimen kecil, maka digunakannya metode tambahan untuk meningkatkan performa model SVM yaitu dengan menggunakan augmentasi data. Proses klasifikasi diawali dengan preprocessing data, yaitu data cleaning dan text preprocessing. Lalu data dibobotkan menggunakan Term-Frequency Inverse Document Frequency. Data yang sudah dibersihkan berjumlah 154.813 record. Kemudian membagi menjadi data latih dan uji dengan volume 70:30 untuk menghindari overfitting. Data latih dilatih dengan SVM dan dievaluasi, lalu data latih yang sudah melalui preprocessing diaugmentasi dengan synonym replacement sebanyak lima iterasi. Proses ini menghasilkan 1.083.658 record untuk data latih. Data hasil augmentasi diekstraksi dan klasifikasi SVM kembali. Kemudian model dengan data latih asli dan hasil augmentasi digunakan untuk klasifikasi data uji dan berhasil meningkatkan 6% performa model SVM untuk klasifikasi dimensi MBTI.

Kata kunci: *augmentasi, mbti, svm.*

ABSTRACT

Personality is an inherent aspect of an individual, and one way to determine it is through the Myers-Briggs Type Indicator (MBTI). Previous studies have classified MBTI using various data mining methods, including Support Vector Machine (SVM). Due to the low accuracy of this research during experiments, an additional method was employed to enhance the SVM model's performance, namely using data augmentation. The classification process begins with data preprocessing, involving data cleaning and text preprocessing. Then, the data is weighted using Term-Frequency Inverse Document Frequency. The cleaned data amounts to 154,813 records. Next, it is divided into training and testing sets with a 70:30 ratio to avoid overfitting. The training data is trained with SVM and evaluated. The preprocessed training data is then augmented with synonym

replacement through five iterations, resulting in 1,083,658 records for training. The augmented data is extracted and classified again using SVM. The model, trained with both the original and augmented data, is used to classify the test data, successfully improving the SVM model's performance by 6% for MBTI dimension classification.

Keywords: *augmentation, mbti, svm.*

1. PENDAHULUAN

Semakin luas dan mudahnya penggunaan media sosial di dunia ini membantu dalam terbangunnya jembatan komunikasi antar banyak orang yang menggunakan media sosial, salah satunya yaitu Twitter. Dengan memanfaatkan Twitter, proses komunikasi dan bertukar informasi dapat dilakukan dengan lebih cepat dan mudah (Muhammad et al., 2021). Kepribadian merupakan karakteristik yang dapat diperhatikan dari berbagai hal dan arah ketika berkomunikasi, seperti bagaimana cara individu tersebut dalam berperilaku dan bertutur kata. Dengan mengetahui kedua hal tersebut, maka kita dapat menghadapi dan mengatasi individu yang kita temui saat menggunakan sosial media (Brewer, 2019). Terdapat banyak cara untuk mengetahui kepribadian seseorang, salah satunya yaitu menggunakan metode Myers-Briggs *Type Indicator* (MBTI) yang berfokus pada 4 dimensi yaitu *Introvert-Extrovert* (EI), *Sensing-Intuition* (SN), *Thinking-Feeling* (TF), dan *Judging-Perceiving* (JP) yang menghasilkan salah satu dari 16 kemungkinan jenis MBTI (Hanbury, 2016).

Terdapat beberapa cara untuk melakukan klasifikasi MBTI seperti menggunakan *Traditional Learning* seperti yang dilakukan oleh (Sahono et al., 2020) dengan SVM untuk klasifikasi salah satu dimensi MBTI dari *dataset* yang berasal dari Kaggle yaitu *Introvert-Extrovert* dengan hasil akurasi 84.07%, atau menggunakan teknik *Deep Learning* seperti yang dilakukan oleh (Ontoum & Chan, 2022) yang menggunakan *Recurrent Neural Network* untuk klasifikasi keseluruhan kelas jenis MBTI dengan akurasi 49.75%. Peningkatan performa klasifikasi dapat dilakukan dengan memperbanyak data sehingga model dapat membaca pola yang bervariasi. Penambahan data pada penelitian ini dengan dilakukan menerapkan augmentasi pada data latih yang digunakan. Seperti yang dilakukan oleh (Harahap & Muslim, 2020) yang menggunakan *dataset* Kaggle, menerapkan augmentasi pada data sebelum diklasifikasi menggunakan *Random Forest*, berhasil meningkatkan 40% dari akurasi awal.

Penelitian ini menggunakan pendekatan pada penelitian (Sahono et al., 2020), yaitu mengklasifikasi dimensi MBTI dan menerapkan prosesnya pada semua dimensi MBTI dan tidak terfokus hanya kepada 1 dimensi saja. *Dataset* yang digunakan juga berbeda dari yang sudah digunakan pada penelitian sebelumnya yang menggunakan *dataset* Kaggle, penelitian ini menggunakan *dataset* dari data.world dengan *tweet* yang masing-masing terpisah untuk setiap akun. Proses augmentasi dilakukan dengan menerapkan *Synonym Replacement* dengan 5 iterasi terhadap data latih.

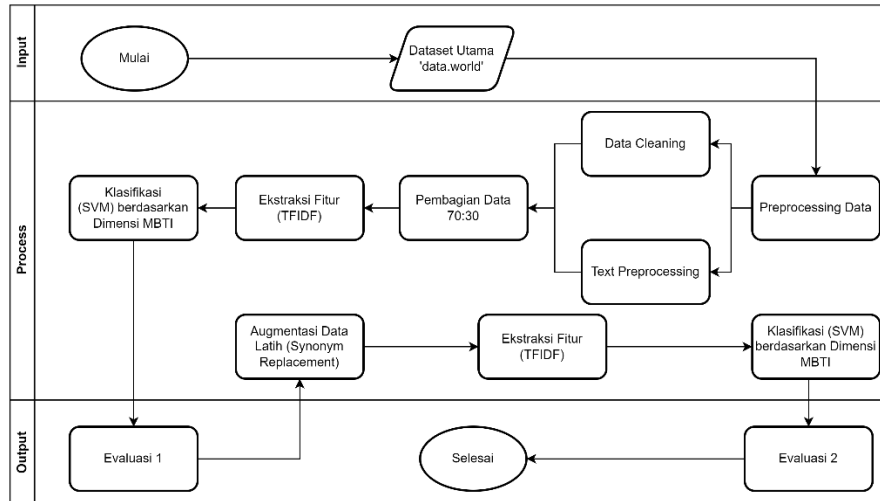
Berdasarkan latar belakang penelitian yang sudah dijelaskan secara singkat, terdapat sebuah rumusan masalah dalam penelitian ini yaitu, apakah performa klasifikasi dimensi MBTI menggunakan metode SVM dapat ditingkatkan dengan menerapkan augmentasi terhadap *dataset* yang digunakan. Dengan rumusan masalah yang ada, tujuan dilaksanakannya penelitian ini yaitu, meningkatkan performa dari model klasifikasi dimensi MBTI menggunakan metode SVM dengan menerapkan augmentasi data terhadap *dataset* yang digunakan selama proses klasifikasi berlangsung.

2. METODOLOGI

2.1 Alur Penelitian

Penelitian dilakukan dengan melalui beberapa tahapan, yaitu memasukkan data ke dalam sistem, *preprocessing* data dengan *data cleaning* dan *text preprocessing*, membagi data dengan dimensi 70:30, ekstraksi fitur menggunakan *Term Frequency – Inverse Document Frequency* (TF-IDF), pelatihan model klasifikasi SVM dan hasilnya akan masuk ke tahap

evaluasi 1. Setelah evaluasi 1, proses augmentasi dengan pendekatan dekatan *synonym replacement* akan diterapkan pada data latih dengan dimensi 70, ekstraksi fitur dengan TF-IDF, melatih model dengan data hasil augmentasi dan masuk ke tahap evaluasi 2. Untuk alur yang lebih jelas dapat dilihat pada Gambar 1.



Gambar 1. Alur Penelitian

2.2 Support Vector Machine

Boser, Guyon, dan Vapnik menggabungkan beberapa teori komputasi yang sudah terkenal seperti margin hyperplane, cover, kernel, dan lainnya pada tahun 1992 sehingga terciptalah metode SVM (Pisner & Schnyer, 2019). Dalam penelitian ini, SVM digunakan untuk klasifikasi 4 dimensi yang membentuk MBTI, di mana hyperplane akan memisahkan 2 kelas yang ada di masing-masing dimensi MBTI. Proses perhitungan SVM sendiri difokuskan untuk menentukan *hyperplane* dari masing-masing dimensi MBTI menggunakan model LinearSVC dengan parameter $C=0.1$, $penalty=l1$, dan $dual=False$.

Perlu diperhatikan bahwa proses pelatihan model SVM untuk klasifikasi dimensi MBTI dilakukan sebanyak 4 kali sesuai dengan banyak dimensi MBTI yaitu 4. Hasil dari keempat model ini akan digabungkan sehingga membentuk 16 jenis MBTI.

2.3 Augmentasi Data

Proses augmentasi data yang diterapkan pada data latih untuk klasifikasi dimensi MBTI pada penelitian ini merujuk pada (Harahap & Muslim, 2020) yang melakukan prediksi jenis MBTI dengan menerapkan *Random Deletion* terhadap *dataset* yang digunakan. Berbeda dengan penelitian tersebut, *Synonym Replacement* yang akan memilih n kata secara acak dari sebuah *tweet* yang kemudian akan diganti dengan sinonimnya dan setelah diganti sebanyak i iterasi maka i dokumen tersebut akan ditambahkan ke *dataset* yang digunakan (Wei & Zou, 2019). Proses augmentasi terhadap data latih penelitian ini dilakukan sebanyak 5 iterasi sehingga data latih sesudah augmentasi berjumlah 1.083.658 *record*. Sebagian hasil dari proses augmentasi dapat dilihat pada Tabel 1.

Tabel 1. Hasil Augmentasi Terhadap Data Latih

Ket	Teks
Asli	<i>discover decide interested rectum wan pic mouse</i>

Augmented	<i>notice make up one s mind worry rectum wan exposure sneak</i>
	<i>light upon settle concerned rectum wide area network flick creep</i>
	<i>key out determine interest rectum pallid film mouse</i>
	<i>name adjudicate concern rectum pallid exposure pussyfoot</i>
	<i>disclose adjudicate concern rectum pale moving picture black eye</i>

2.4 Confusion Matrix

Confusion Matrix digunakan pada tahap evaluasi untuk mengetahui seberapa baik model dalam menghadapi kelas selama proses klasifikasi berlangsung (R, J & R, 2016). *Confusion Matrix* melakukan perhitungan statistik berdasarkan 4 komponen yang dibentuk dengan hasil prediksi dan hasil sebenarnya seperti yang terlihat pada Tabel 2, dan menghasilkan akurasi model, presisi model, *recall* model, dan nilai F model yang digunakan untuk pelatihan.

Tabel 2. Confusion Matrix

	Predicted: No	Predicted: Yes
Actual: No	TN	FP
Actual: Yes	FN	TP

3. HASIL PENELITIAN DAN EVALUASI

3.1 Pengumpulan Data

Penelitian ini menggunakan data yang diambil dari salah satu *data bank* yaitu data.world yang diunggah oleh pengguna dengan nama Barret Brister, merupakan data berupa postingan atau *tweet* yang diambil dari salah satu *social media* yaitu Twitter. Data mentah yang diambil dari data.world dibagi menjadi 16 berkas yang masing-masing merepresentasikan 16 jenis MBTI yang ada, dengan total data ketika digabungkan menjadi 163.200 *record*. Contoh dari data yang berhasil diambil dapat dilihat pada Gambar 2.

Screen name	MBTI	Time	Tweet
girlgroupzhoel	ENFJ	2021-05-20 18:05:19	@kiiim93 Thank you so much ❤️💕I am going only ...
girlgroupzhoel	ENFJ	2021-05-20 16:46:23	Thank you guys for the supportive messages ❤️💕...
girlgroupzhoel	ENFJ	2021-05-20 15:20:04	I mean mcountdown 🤔Wtf is MC countdown wtaf
girlgroupzhoel	ENFJ	2021-05-20 15:17:52	Son: Why is my sister called Rose?🤔Dad: Becaus...
girlgroupzhoel	ENFJ	2021-05-20 13:41:25	Mayhaps I cried for an hour straight already....

Gambar 2. Data MBTI dari data.world

3.2 Preprocessing Data

Preprocessing data yang dilakukan pada penelitian ini dilakukan dengan 2 tahap, yaitu *data cleaning* dan *text preprocessing*. Kedua proses ini dilakukan dengan tujuan untuk membersihkan data dari *noise* dan kata-kata yang tidak memiliki *value*.

a. Data Cleaning

Proses *data cleaning* bertujuan untuk membersihkan *noise* yang ada pada data yaitu nilai kosong, duplikat, dan atribut atau fitur yang tidak digunakan.

b. Text Preprocessing

Proses *text preprocessing* bertujuan untuk menghapus *noise* yang ada di dalam teks data yang digunakan dalam proses pelatihan model nantinya. Tahapan yang ada pada proses ini antara lain:

Case folding mengubah huruf kapital menjadi huruf kecil, hasilnya dapat dilihat pada Gambar 3.

MBTI	Tweet
ENFJ	@kiiim93 thank you so much ❤️👉i am going only ...
ENFJ	thank you guys for the supportive messages ❤️👉...
ENFJ	i mean mcountdown 🤔wtf is mc countdown wtaf
ENFJ	son: why is my sister called rose?👉dad: becaus...
ENFJ	mayhaps i cried for an hour straight already....

Gambar 3. Hasil Case Folding

Url removal menghapus bentuk tautan yang ada di dalam *tweet*, hasilnya dapat dilihat pada Gambar 4.

what's up sexy mamas 🇲🇻	https://t.co/bd2yvjpvcy
why twitter always cutting my face outta my sel...	
@juggosipe0 it's just a suggestion but it's a s...	
can they whisper "yes baby yes"	https://t.co/cz...
she got there first 😊	tps://t.co/zca4q9qbl2 ...
what's up sexy mamas 🇲🇻	
why twitter always cutting my face outta my sel...	
@juggosipe0 it's just a suggestion but it's a s...	
can they whisper "yes baby yes"	
she got there first 😊	

Gambar 4. Hasil URL Removal

User and punctuation removal menghapus bentuk pemanggilan atau *mention* terhadap pengguna Twitter lainnya di dalam sebuah *tweet*, hasilnya dapat dilihat pada Gambar 5.

thank you so much	i am going only for one...
thank you guys for the supportive messages	...
i mean mcountdown	wtf is mc countdown wtaf
son why is my sister called rose	dad becaus...
mayhaps i cried for an hour straight already	...

Gambar 5. Hasil User and Punctuation Removal

Whitespace removal menghapus ruang putih yang ditinggalkan oleh proses-proses sebelumnya, hasilnya dapat dilihat pada Gambar 6.

thank you so much i am going only for one year ...
thank you guys for the supportive messages i wa...
i mean mcountdown wtf is mc countdown wtaf
son why is my sister called rose dad because yo...
mayhaps i cried for an hour straight already gu...

Gambar 6. Hasil Whitespace Removal

Stopword removal menghapus kata henti seperti *of*, *it*, *and*, dll. Proses ini dilakukan karena kata-kata itu tidak memiliki *value* apapun, hasilnya dapat dilihat pada Gambar 7.

year crazy kinda childhood dream coming wanted ye
guys supportive messages wan scream ngl
mcountdown wtf mc countdown wtaf
sister called rose dad mother loves roses dad d...
mayhaps cried hour straight guess

Gambar 7. Hasil Stopword Removal

Lemmatization mengubah bentuk kata yang ada di data menjadi bentuk asli dari kata tersebut, hasilnya dapat dilihat pada Gambar 8.

```
year crazy kinda childhood dream come want year...
guy supportive message wan scream ngl
mcountdown wtf mc countdown wtaf
sister call rise dad mother love rise dad dad p...
mayhaps cry hour straight guess
```

Gambar 8. Hasil Lemmatization

Non-english removal menghapus kata-kata yang bukan ada dalam bahasa Inggris, hasilnya dapat dilihat pada Gambar 9.

```
year crazy childhood dream come want year imagi...
guy supportive message wan scream
<null>
sister call rise dad mother love rise dad dad p...
cry hour straight guess
```

Gambar 9. Hasil Non-English Removal

Semua proses pembersihan data yang dilakukan menghasilkan data bersih sebanyak 154.813 *record*.

3.3 Pembagian Data

Setelah melalui tahap *preprocessing data*, maka selanjutnya adalah membagi data menjadi data latih dan uji. Data dibagi dengan dimensi data latih sebesar 70% dari data asli, dan 30% untuk data uji dari data asli. Dimensi pembagian ini digunakan untuk menghindari adanya *overfitting*, yaitu kondisi ketika model terlalu menghafal dan mengingat data latih sehingga ketika diberikan data yang serupa model akan menghasilkan akurasi yang sangat tinggi, dan menghasilkan akurasi yang sangat rendah ketika menghadapi data baru dengan kondisi data yang berbeda. Proses ini menghasilkan 108.368 *record* data latih dan 46.444 *record* data uji, dengan masing-masing distribusi data tiap dimensi dan kelas sebagai berikut:

Tabel 3. Distribusi Data per Kelas

Dimensi	Kelas	Data Latih	Data Uji
EI	E	54.199	23.387
	I	54.169	23.057
SN	S	54.031	23.331
	N	54.337	23.113
TF	T	54.392	23.212
	F	53.976	23.387
JP	J	54.283	23.272
	P	54.085	23.172

3.4 Ekstraksi Fitur

Term Frequency Inverse Document Frequency atau biasa disingkat sebagai TFIDF, merupakan metode pembobotan kata yang akan digunakan dengan menerapkannya terhadap setiap kata yang ada di dalam *tweet*. Proses ini akan melakukan tokenisasi

terhadap setiap kata, menghitung frekuensi kemunculannya di *tweet* yang ada, menghitung frekuensi kemunculannya di seluruh dokumen, lalu melakukan perkalian terhadap hasil perhitungan keduanya (Rofiqi et al., 2019). Perhitungannya sendiri menggunakan persamaan 1, 2, dan 3 berikut:

$$TF_{ij} = \frac{f_{ij}}{max_j} \quad (1)$$

Keterangan:

TF_{ij} : *Term Frequency* kata i di kalimat j

f_{ij} : Frekuensi kemunculan kata i pada kalimat j

max_j : Total kata di kalimat j

$$IDF_i = \log\left(\frac{D}{DF_i}\right) \quad (2)$$

Keterangan:

IDF_i : Bobot untuk kata i

D : Total dokumen

DF_i : Jumlah dokumen mengandung kata i

$$TFIDF_{ij} = TF_{ij} \cdot IDF_i \quad (3)$$

Proses ekstraksi fitur menghasilkan bobot untuk masing-masing kata yang ada di dalam data. Untuk hasilnya dapat dilihat pada Tabel 4 dengan sebagian data kata yang berhasil di ekstrak.

Tabel 4. Hasil Ekstraksi Fitur

Kata	Bobot
stress exhaustion	0.2880713411173808
blame stress	0.2880713411173808
exercise blame	0.2880713411173808
eat exercise	0.2880713411173808
health eat	0.2785417225087895
...	
discipline health	0.2880713411173808
motivate discipline	0.2880713411173808
time motivate	0.2785417225087895
hard time	0.20196690318047947
exhaustion	0.23171377308956284

3.5 Eksperimen

Penelitian ini melakukan 2 eksperimen untuk melakukan evaluasi model SVM dalam melakukan klasifikasi MBTI.

a. Eksperimen 1

Eksperimen pertama dilakukan untuk mengetahui performa model SVM terhadap data latih yang belum melalui tahap augmentasi. Berikut hasil eksperimen klasifikasi terhadap setiap dimensi MBTI yang dibagi menjadi hasil *Confusion Matrix* dan visualisasi *Hyperplane*.

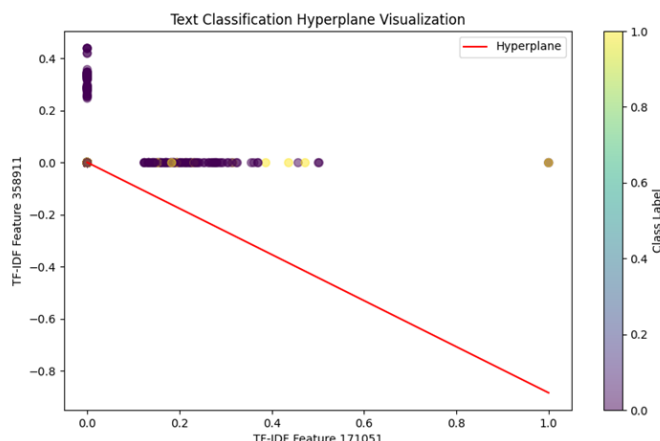
Dimensi EI

Tabel 5. Confusion Matrix Data Uji Dimensi EI

	Predicted: E	Predicted: I
Actual: E	14.907	8.480
Actual: I	12.519	10.538

Tabel 6. Perhitungan Confusion Matrix Data Uji Dimensi EI

	Precision	Recall	F1-Score	Akurasi
E	54%	64%	59%	55%
I	55%	46%	50%	



Gambar 10. Visualisasi Hyperplane Dimensi EI

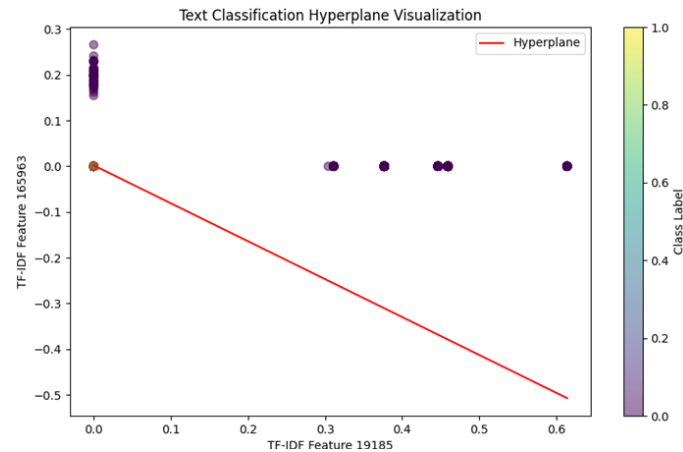
Dimensi SN

Tabel 7. Confusion Matrix Data Uji Dimensi SN

	Predicted: N	Predicted: S
Actual: N	14.561	8.552
Actual: S	11.932	11.399

Tabel 8. Perhitungan Confusion Matrix Data Uji Dimensi SN

	Precision	Recall	F1-Score	Akurasi
N	55%	63%	59%	56%
S	57%	49%	53%	



Gambar 11. Visualisasi Hyperplane Dimensi SN

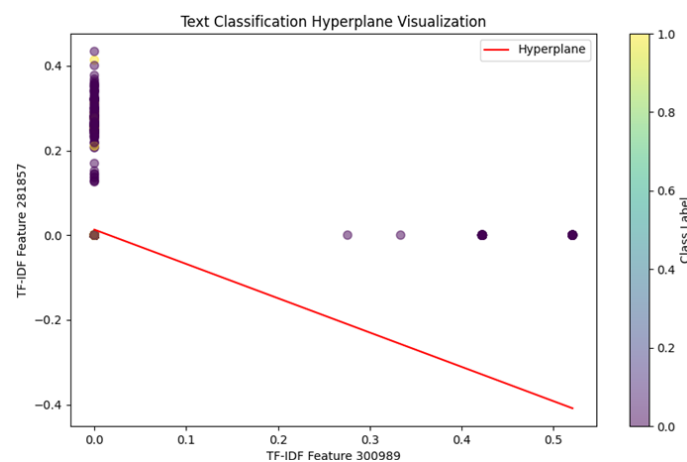
Dimensi TF

Tabel 9. Confusion Matrix Data Uji Dimensi TF

	Predicted: F	Predicted: T
Actual: F	11.625	11.607
Actual: T	9.114	14.098

Tabel 10. Perhitungan Confusion Matrix Data Uji Dimensi TF

	Precision	Recall	F1-Score	Akurasi
F	56%	50%	53%	55%
T	55%	61%	58%	



Gambar 12. Visualisasi Hyperplane Dimensi TF

Klasifikasi Myers-Briggs Type Indicator Pengguna Twitter Menggunakan Support Vector Machine dan Augmentasi Data

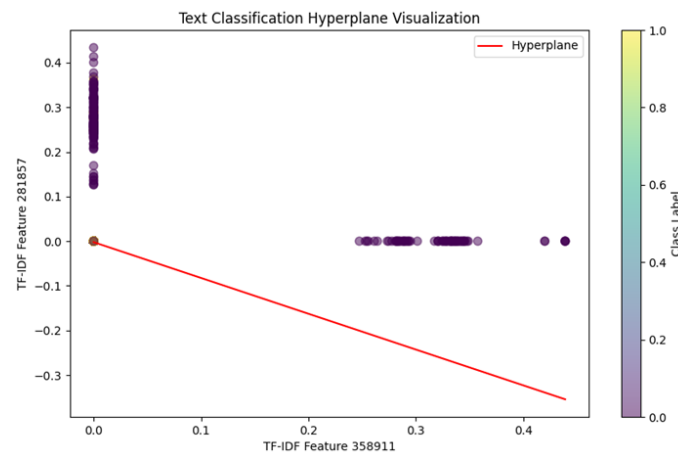
Dimensi JP

Tabel 11. Confusion Matrix Data Uji Dimensi JP

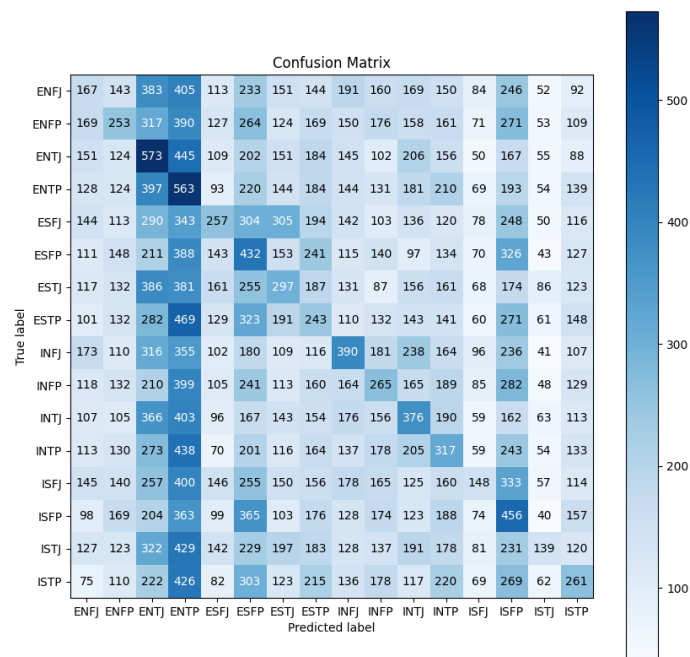
	Predicted: J	Predicted: P
Actual: J	10.938	12.334
Actual: P	8.189	14.983

Tabel 12. Perhitungan Confusion Matrix Data Uji Dimensi JP

	Precision	Recall	F1-Score	Akurasi
J	57%	47%	52%	56%
P	55%	65%	59%	



Gambar 13. Visualisasi Hyperplane Dimensi JP



Gambar 14. Confusion Matrix Seluruh Kelas MBTI

Tabel 13. Perhitungan Confusion Matrix Seluruh Kelas MBTI

	Precision	Recall	F1-Score	Akurasi
ENFJ	8%	6%	7%	11%
ENFP	12%	9%	10%	
ENTJ	11%	20%	14%	
ENTP	9%	19%	12%	
ESFJ	13%	9%	10%	
ESFP	10%	15%	12%	
ESTJ	12%	10%	11%	
ESTP	8%	8%	8%	
INFJ	15%	13%	14%	
INFP	11%	9%	10%	
INTJ	13%	13%	13%	
INTP	11%	11%	11%	
ISFJ	12%	5%	7%	
ISFP	11%	16%	13%	
ISTJ	15%	5%	7%	
ISTP	13%	9%	11%	

Dengan memperhatikan hasil dari eksperimen 1, dapat dilihat bahwa akurasi dari model pelatihan SVM terhadap klasifikasi MBTI terbilang rendah karena hanya mendapatkan sebesar 11%. Akan tetapi untuk akurasi klasifikasi setiap dimensi MBTI terbilang cukup baik dengan akurasi sebesar 55% hingga 56%. Adapun visualisasi *Hyperplane* yang digunakan menggambarkan fitur dari proses ekstraksi menggunakan TF-IDF berdasarkan koefisien SVM dari hasil klasifikasi yang dilakukan. Sumbu X dan Y pada visualisasi sendiri merupakan pencerminan frekuensi frasa yang menggambarkan dimensi MBTI. Garis *Hyperplane* membedakan frasa yang memiliki nilai indeks sama dengan atau kurang dari 0 merupakan frasa yang tidak mencerminkan dimensi MBTI terkait.

b. Eksperimen 2

Eksperimen dilakukan untuk mengetahui performa model SVM terhadap data latih yang sudah melalui tahap augmentasi dan menggunakan data uji yang sama dengan eksperimen 1.

Dimensi EI

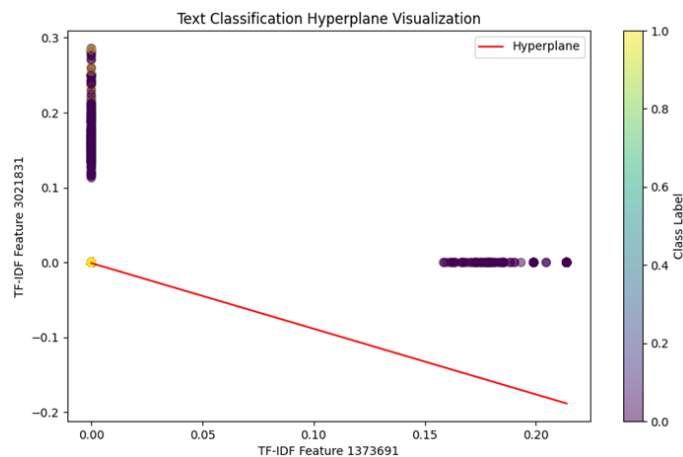
Tabel 14. Confusion Matrix Data Uji Dimensi EI

	Predicted: E	Predicted: I
Actual: E	13.829	9.558
Actual: I	8.337	14.720

Tabel 15. Perhitungan Confusion Matrix Data Uji Dimensi EI

	Precision	Recall	F1-Score	Akurasi
E	62%	59%	61%	61%
I	61%	64%	62%	

Klasifikasi Myers-Briggs Type Indicator Pengguna Twitter Menggunakan Support Vector Machine dan Augmentasi Data



Gambar 15. Visualisasi Hyperplane Dimensi EI

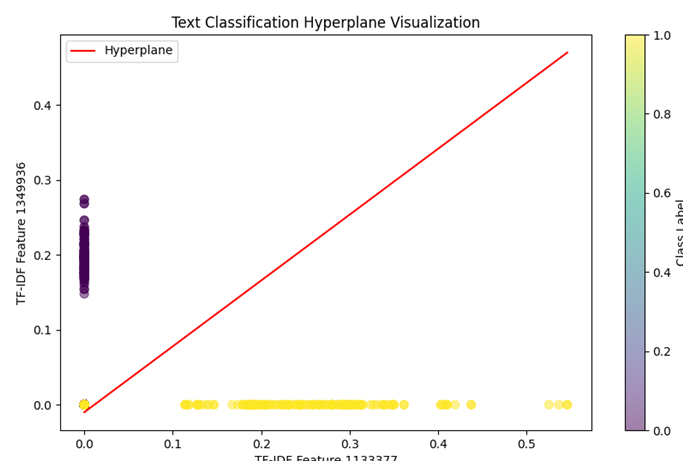
Dimensi SN

Tabel 16. Confusion Matrix Data Uji Dimensi SN

	Predicted: N	Predicted: S
Actual: N	13.490	9.623
Actual: S	7.860	15.471

Tabel 17. Perhitungan Confusion Matrix Data Uji Dimensi SN

	Precision	Recall	F1-Score	Akurasi
N	63%	58%	61%	62%
S	62%	66%	64%	



Gambar 16. Visualisasi Hyperplane Dimensi SN

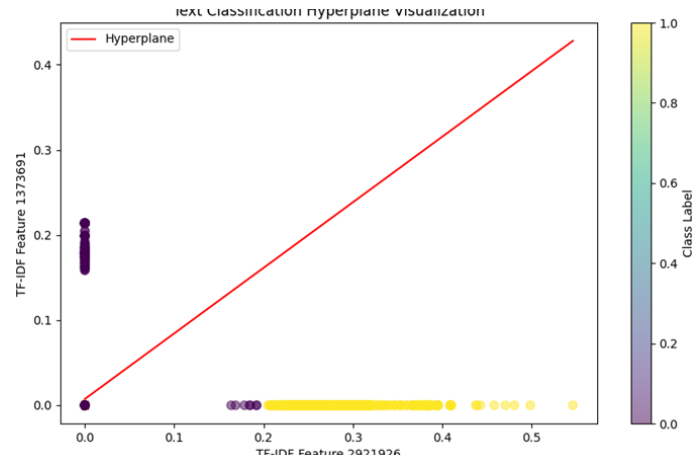
Dimensi TF

Tabel 18. Confusion Matrix Data Uji Dimensi TF

	Predicted: F	Predicted: T
Actual: F	14.365	8.867
Actual: T	9.192	14.020

Tabel 19. Perhitungan Confusion Matrix Data Uji Dimensi TF

	Precision	Recall	F1-Score	Akurasi
F	61%	62%	61%	61%
T	61%	60%	61%	



Gambar 17. Visualisasi Hyperplane Dimensi TF

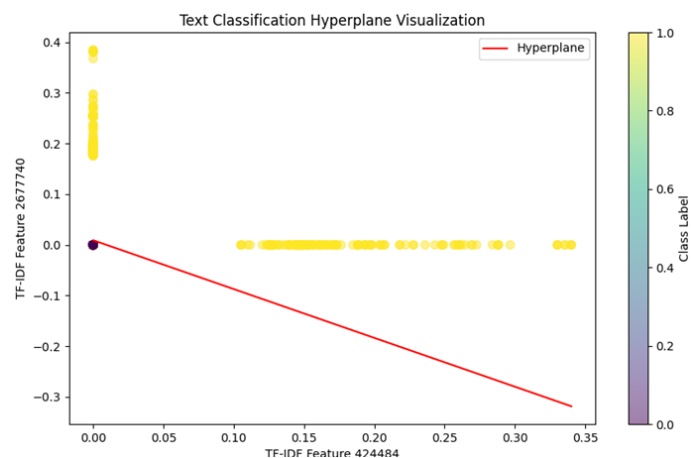
Dimensi JP

Tabel 20. Confusion Matrix Data Uji Dimensi JP

	Predicted: J	Predicted: P
Actual: J	13.787	9.485
Actual: P	8.148	15.024

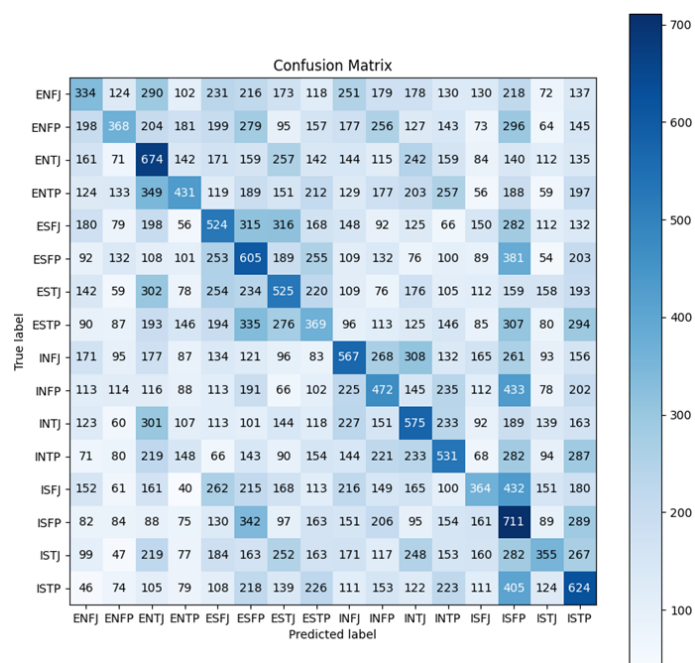
Tabel 21. Perhitungan Confusion Matrix Data Uji Dimensi JP

	Precision	Recall	F1-Score	Akurasi
J	63%	59%	61%	62%
P	61%	65%	63%	



Gambar 18. Visualisasi Hyperplane Dimensi JP

Klasifikasi Myers-Briggs Type Indicator Pengguna Twitter Menggunakan Support Vector Machine dan Augmentasi Data



Gambar 19. Confusion Matrix Seluruh Kelas MBTI

Tabel 22. Perhitungan Confusion Matrix Seluruh Kelas MBTI

	Precision	Recall	F1-Score	Akurasi
ENFJ	15%	12%	13%	17%
ENFP	22%	12%	16%	
ENTJ	18%	23%	20%	
ENTP	22%	14%	18%	
ESFJ	17%	18%	17%	
ESFP	16%	21%	18%	
ESTJ	17%	18%	18%	
ESTP	13%	13%	13%	
INFJ	19%	19%	19%	
INFP	16%	17%	17%	
INTJ	18%	20%	19%	
INTP	19%	19%	19%	
ISFJ	18%	12%	15%	
ISFP	14%	24%	18%	
ISTJ	19%	12%	15%	
ISTP	17%	22%	19%	

3.6 Perbandingan Hasil Eksperimen

Hasil eksperimen 1 dan 2 yang dievaluasi menggunakan *confusion matrix*, akan dibandingkan untuk melihat apakah performa model SVM dalam klasifikasi MBTI meningkat ketika menggunakan data latih yang sudah diaugmentasi atau tidak.

Dimensi EI

Tabel 23. Performa Model Sebelum dan Sesudah Augmentasi Data Latih Dimensi EI

	E	I
--	---	---

Precision	> 8%	> 6%
Recall	< 6%	> 16%
F1-Score	> 2%	> 12%
Akurasi	> 6%	

Dimensi SN

Tabel 24. Performa Model Sebelum dan Sesudah Augmentasi Data Latih Dimensi SN

	N	S
Precision	> 8%	> 5%
Recall	< 5%	> 17%
F1-Score	> 2%	> 11%
Akurasi	> 6%	

Dimensi TF

Tabel 25. Performa Model Sebelum dan Sesudah Augmentasi Data Latih Dimensi TF

	F	T
Precision	> 5%	> 6%
Recall	> 12%	< 1%
F1-Score	> 8%	> 3%
Akurasi	> 6%	

Dimensi JP

Tabel 26. Performa Model Sebelum dan Sesudah Augmentasi Data Latih Dimensi JP

	J	P
Precision	> 6%	> 6%
Recall	> 12%	> 0%
F1-Score	> 9%	> 4%
Akurasi	> 6%	

4. KESIMPULAN

Kesimpulan yang dihasilkan dari penelitian ini yaitu penerapan Augmentasi terhadap data latih yang digunakan untuk proses klasifikasi berhasil meningkatkan performa model dalam menghadapi kelas-kelasnya, dalam bagian dari dimensi MBTI. Performa yang berhasil ditingkatkan antara lain akurasi pada setiap model dimensi MBTI dengan peningkatan konsisten sebesar 6%, dan kemampuan model dalam mempelajari pola dan variasi setiap kelas dari dimensi MBTI, serta akurasi akhir untuk semua kelas MBTI yaitu peningkatan 6 % di mana akurasi awal 11% menjadi 17%. Peningkatan performa juga bisa dilihat melalui F1-Score setiap dimensi yang dipengaruhi oleh peningkatan representasi, variasi, serta generalisasi *record* setelah penerapan augmentasi pada data latih sehingga keseimbangan *precision* dan *recall* yang terpengaruh pada proses penerapan augmentasi tersebut. Saran untuk penelitian yang dilakukan menggunakan konsep dan konteks Augmentasi Data, SVM dan MBTI, dapat mencoba penggunaan berbagai macam teknik Augmentasi Data dan percobaan menggabungkannya dengan *kernel trick* SVM selain *kernel linear* seperti yang digunakan pada penelitian ini.

UCAPAN TERIMA KASIH

Saya Sulthan Athallah sebagai penulis mengucapkan rasa Syukur dan terima kasih yang besar kepada semua pihak yang membantu saya dalam bentuk apapun dalam pembuatan dan pelaksanaan penelitian "Klasifikasi Myers-Briggs Type Indicator Pengguna Twitter Menggunakan Support Vector Machine dan Augmentasi Data" hingga selesai.

DAFTAR RUJUKAN

- Brewer, L. (2019). General Psychology: Required Reading. *Personality Traits*, 1–324.
- Hanbury, D. (2016). Anthropomorphism: Understanding Personality and Individual Differences in Nonhuman Animals. In *Encyclopedia of Personality and Individual Differences*. https://doi.org/10.1007/978-3-319-28099-8_961-1
- Harahap, R. N., & Muslim, K. (2020). Peningkatan Akurasi pada Prediksi Kepribadian MbtI Pengguna Twitter Menggunakan Augmentasi Data. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, 7(4), 815. <https://doi.org/10.25126/jtiik.2020743622>
- Muhammad, W. Z., Erliana, Y. D., & Hakim, L. (2021). Hubungan Jenis Kepribadian (Ekstrovert & Introvert) Dengan Pengungkapan Diri (Self Disclosure) Pada Pengguna Media Sosial Instagram: Studi Kasus Pada Mahasiswa Fakultas Psikologi Universitas Teknologi Sumbawa. *Jurnal PSIMAWA*, 4(1), 13–18.
- Ontoum, S., & Chan, J. (2022). *Personality Type Based on Myers-Briggs Type Indicator with Text Posting Style by using Traditional and Deep Learning*. <https://doi.org/10.48550/arXiv.2201.08717>
- Pisner, D. A., & Schnyer, D. M. (2019). Support vector machine. *Machine Learning: Methods and Applications to Brain Disorders*, 101–121. <https://doi.org/10.1016/B978-0-12-815739-8.00006-7>
- R, J, M. N., & R, P. (2016). Performance Analysis of Text Classification Algorithms using Confusion Matrix. *International Journal of Engineering and Technical Research*, 0869(4), 75–78.
- Rofiqi, M. A., Fauzan, A. C., Agustin, A. P., & Saputra, A. A. (2019). Implementasi Term-Frequency Inverse Document Frequency (TF-IDF) Untuk Mencari Relevansi Dokumen Berdasarkan Query. *ILKOMNIKA: Journal of Computer Science and Applied Informatics*, 1(2), 58–64. <https://doi.org/10.28926/ilkomnika.v1i2.18>
- Sahono, M. N., Sidiastahta, F. U., Shidik, G. F., Fanani, A. Z., Muljono, Nuraisha, S., & Lutfina, E. (2020). Extrovert and Introvert Classification based on Myers-Briggs Type Indicator(MBTI) using Support Vector Machine (SVM). *Proceedings - 2020 International Seminar on Application for Technology of Information and Communication: IT Challenges for Sustainability, Scalability, and Security in the Age of Digital Disruption, ISemantic 2020*, 572–577. <https://doi.org/10.1109/iSemantic50169.2020.9234288>
- Wei, J., & Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 6382–6388. <https://doi.org/10.18653/v1/d19-1670>