E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 04, Desember 2022 : 1088 - 1095

SOCIAL MEDIA USER PERSONALITY CLASSIFICATION BASED ON HOW USER LIVE AND MAKE DECISION

Chamadani Faisal Amri[1], Sri Suryani Prasetyowati[2], Yuliant Sibaroni[3]
[1, 2, 3]School of Computing, Telkom University
Jl. Telekomunikasi No. 1, Bandung, Indonesia
e-mail: daniamri@student.telkomuniversity.ac.id[1], srisuryani@telkomuniversity.ac.id[2], yuliant@telkomuniversity.ac.id[3]

ABSTRAK

Klasifikasi kepribadian merupakan salah satu cara dalam bidang Natural Language Processing (NLP) dengan kumpulan data-data yang dapat mendeskripsikan kepribadian pengguna melalui masukan set dokumen teks seperti unggahan status. Sosial media merupakan salah satu cara untuk berinteraksi secara online yang dapat memberikan kemudahan bagi penggunanya seperti, berinteraksi, mengeskpersikan diri, memperluas pertemanan, Unggahan status pada media sosial dapat diekstrak menjadi informasi yang berguna dalam proses klasifikasi kepribadian. Penelitian ini melakukan klasifikasi berdasarkan bagaimana pengguna media sosial menjalani hidup dan mengambil keputusan, yang merupakan representasi dari atribut kelas "Thinkers/Feelers" dan "Judgers/Perceivers" model Myers-Briggs Type Indicator (MBTI). Peneliti terdorong untuk mengembangkan sebuah sistem klasifikasi kepribadian dengan fitur ekstraksi yang dapat meningkatkan performa sistem. Dalam research ini, terdapat tiga eksperimen utama yang dilakukan, eksperimen menggunakan data dengan teknik oversample pada kelas Thinker/Feelers (TF) dan Judgers/Perceivers (JP) memberikan hasil yang terbaik dibandingan eksperimen lain dengan f1-score dan akurasi sebesar 92% menggunakan metode klasifikasi Random Forest dan Glove sebagai fitur eksraksinya.

***Kata Kunci**: Glove, Klasifikasi Kepribadian, Myers-Briggs Type Indicator (MBTI), Random Forest , Sosial Media.*

ABSTRACT

Personality classification is one of the ways in the field of Natural Language Processing (NLP) with a collection of data that can describe the user's personality through input sets of text documents such as status uploads. Social media is one way to interact online that can provide convenience for users, such as interacting, expressing themselves, and expanding friendships. Status posts on social media can be extracted into useful information in the personality classification process. This research performs classification based on how social media users live their lives and make decisions, which is a representation of the "Thinkers/Feelers" and "Judgers/Perceivers" class attributes of the Myers-Briggs Type Indicator (MBTI) model. Researchers are encouraged to develop a personality classification system with feature extraction that can improve system performance. In this research, there are three main experiments conducted, experiments using data with oversample techniques in the Thinker/Feelers (TF) and Judgers/Perceivers (JP) classes provide the best results compared to other experiments with f1-score and accuracy of 92% using the Random Forest classification method and Glove as the extraction feature.

***Keywords**: Glove, Personality Classification, Myers-Briggs Type Indicator (MBTI), Random Forest, Social Media.*

I. INTRODUCTION

PERSONALITY is a combination of the nature and attitude of a human being is dealing with different social conditions or situations [1]. Every human personality certainly has the characteristics of a different mindset, emotion, and behavior [2]. Some researchers believe that personality can be used as an effective reference in measuring and predicting academic performance [1]. The Myers-Briggs Type Indicator (MBTI) is the most frequently used approach in psychology in observing human behavior. MBTI describes personality with eight aspects: extrovert (E), introvert (I), sensor (S), intuitive (N), thinker (T), feel (F), judge (J), and perceive (P). [1], [3].

The rapid development of technology, the internet, and the popularity of social media tools can make it easier for researchers to analyze and classify users. One of the popular social media is Twitter. Many researchers who do text mining use social media such as Twitter as a medium to get information. Many Twitter users unwittingly provide information about their personality through tweets [4]. This can make it easier for researchers to analyze and classify personality based on Twitter user accounts.

Previous research experts have carried out personality classification using the Support Vector Machine (SVM) conducted by B. Y. Pratama [5] in 2015, R. I. Kurnia [6] in 2020, XGBoost conducted by V. Ong [4] in 2017, Multinomial Naïve Bayes conducted by Y. Artissa [7] in 2019, and *k*-Means clustering conducted by A. Talasbaek [1] in 2020. These studies use indicators of personality type Big Five and Myers Briggs Type Indicator

(MBTI). However, these studies still have weaknesses, one of which is the low accuracy of the results. In a study conducted by Y. Artissa [7], personality classification was carried out using the Multinomial Naïve Bayes method with the Big Five model. Then this study compares the accuracy of the classification with the process without changing the form of the word basis, classification using the stemming process, and classification using the lemmatization process. The best accuracy is obtained from the process. The classification using stemming is 59.9% because the stemming process changes the shape into a base.

B. Y Pratama [5] conducted a study using tweets uploaded on Twitter as the data used for personality classification. This study compares three methods, namely Naïve Bayes, K-nearest Neighbors (KNN), and Support Machine. Vector (SVM) uses the TF-IDF feature. The highest accuracy obtained in this study was 63% using the Naïve Bayes method, then SVM was 61%, and KNN was 60%. The SVM method was also compared with XGBoost in a study conducted by V. Ong [4]. SVM got an accuracy result of 76.2%, while XGBoost got the best accuracy result of 97.9%.

Based on previous research, This research performs classification based on how social media users live their lives and make decisions, which is a representation of the attributes of the "Thinkers/Feelers" and " Judgers/Perceivers" classes of the Myers-Briggs Type Indicator (MBTI) model [1], [3] with the Random Forest method[8] using GloVe to obtain vector representations for words [9]. The parameters of how users live and make decisions are obtained by looking at the user's mindset through tweets using machine learning. Based on the tweets, we can label the users as thinkers, feelers, judgers, or perceivers. The Random Forest method is used because the algorithm can provide high accuracy results [10], so this algorithm can provide better results than previous research. This study was designed to determine the accuracy obtained by using this method.

II. Research Method

This section describes the research methods used and the system workflow from start to finish which can be seen in Figure 1.

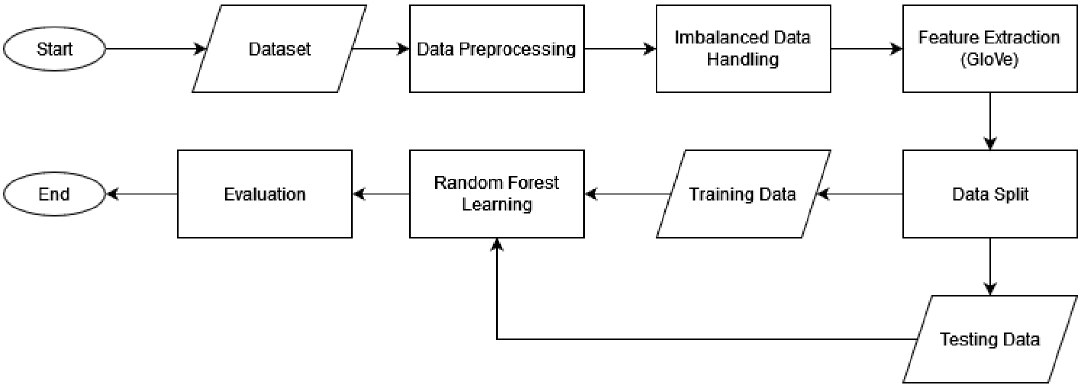

Fig. 1. System Workflow

A. Dataset

In this research, the dataset is obtained from the online community platform Kaggle (https://www.kaggle.com/datasnaek/mbti-type). This dataset consists of 2 columns with 8675 rows of data. The data used are tweets obtained from Twitter. Twitter is one of the popular social media for millions of people to interact, express thoughts, emotions, and beliefs by writing and uploading them on social media [12]. Table 1 is a description of the data that will be used in this research.

TABLE I
EXAMPLE OF PERSONALITY DATASET BASED ON THE TWEET

No	Type	Posts
1	ENFP	because I myself can lose touch with them. I'm not clear headed atm, but anyway. I believe in an inconceivable higher power.
2	ENTJ	Hello! I am working on a presentation by type. Part of each presentation is feedback from a range of people of the type being reviewed
3	INFJ	Trying not to feel totally worthless... Why do I have to be so sensitive ugh stupid brain.
4	INTP	Painting the world with the colors of my soul. Interpret it as you like - helping others by volunteering, teaching, giving everything you have.
5	ISFJ	I love feeling affectionate for the one I love and care for.

B. Data Preprocessing

Data preprocessing is a step to make data clean, efficient, and ready to be processed into the system. In the context of personality classification, preprocessing is carried out to clean texts that do not have a value to describe personality.

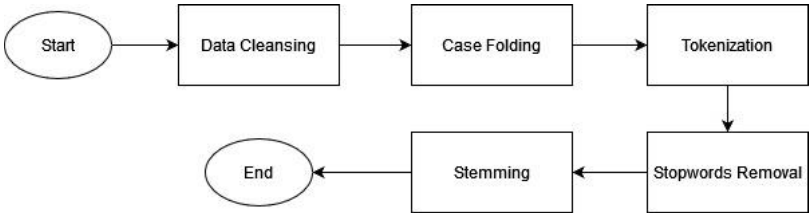

Fig. 2. Data Preprocessing

1) Data Cleansing

Processes such as removing numbers, punctuation, Uniform Resource Locator (URL), and emoji in a sentence are data cleansing processes. This process aims to ensure that there are no data types other than strings and objects in the dataset. This process, it can make the system easier to process the data. The data cleansing process can be seen in Table 2.

TABLE II
DATA CLEANSING PROCESS

No	Before	After
1	enfp and intj moments https://www.youtube.com/watch?v=iz7lE1g4XM4 sportscenter not top ten plays	enfp and intj moments sportscenter not top ten plays
2	|40386 but look at those eyes... i think dogs are excellent spirit animals.	but look at those eyes i think dogs are excellent spirit animals
3	I used to smoke weed alot, everyday from morning till night for 3/4 years	I used to smoke weed alot everyday from morning till night for years

2) Case Folding

Case folding is a process of changing all text or letters from a to z to lowercase [11]. The case folding process is carried out to make all structured and consistent data easy to process by the system. Table 3 is an example of a case folding process.

TABLE III
CASE FOLDING PROCESS

No	Before	After
1	because I myself can lose touch with them. I'm not clear headed atm, but anyway. I believe in an inconceivable higher power.	because i myself can lose touch with them im not clear headed atm, but anyway i believe in an inconceivable higher power
2	Hello! I am working on a presentation by type. Part of each presentation is feedback from a range of people of the type being reviewed	hello i am working on a presentation by type part of each presentation is feedback from a range of people of the type being reviewed
3	I love feeling affectionate for the one I love and care for.	i love feeling affectionate for the one i love and care for

3) Tokenization

Tokenization is a process for changing sentences into tokens in the form of words, characters, or subwords. This token helps in understanding the context or developing a model for the NLP system to make it easier to process. Table 4 is an example of the tokenization process.

TABLE IV
TOKENIZATION PROCESS

No	Before	After
1	because i myself can lose touch with them. im not clear headed atm, but anyway. i believe in an inconceivable higher power.	“because”, “i”, “myself”, “can”, “lose”, “touch”, “with”, “them”, “im”, “not”, “clear”, “headed”, “atm”, “but”, “anyway”, “i”, “believe”, “in”, “an”, “inconceivable”, “higher”, “power”
2	hello! i am working on a presentation by type. part of each presentation is feedback from a range of people of the type being reviewed	“hello”, “i”, “am”, “working”, “on”, “a”, “presentation”, “by”, “type”, “part”, “of”, “each”, “presentation”, “is”, “feedback”, “from”, “a”, “range”, “of”, “people”, “of”, “the”, “type”, “being”, “reviewed”
3	i love feeling affectionate for the one i love and care for.	“i”, “love”, “feeling”, “affectionate”, “for”, “the”, “one”, “i”, “love”, “and”, “care”, “for”

4) Stopwords Removal

Stopwords removal is a step to eliminate words that are so commonly used and have no value or meaning in a document. Thus, the word carries a bit of useful information. In this research, stopwords removal uses the English dictionary from Natural Language Toolkit (NLTK) version 3.7. Table 5 is an example of the stopwords removal process.

TABLE V
STOPWORDS REMOVAL PROCESS

No	Before	After
1	“because”, “i”, “myself”, “can”, “lose”, “touch”, “with”, “them”, “i'm”, “not”, “clear”, “headed”, “atm”, “but”, “anyway”, “i”, “believe”, “in”, “an”, “inconceivable”, “higher”, “power”	“lose”, “touch”, “clear”, “headed”, “atm”, “inconceivable”, “higher”, “power”
2	“hello”, “i”, “am”, “working”, “on”, “a”, “presentation”, “by”, “type”, “part”, “of”, “each”, “presentation”, “is”, “feedback”, “from”, “a”, “range”, “of”, “people”, “of”, “the”, “type”, “being”, “reviewed”	“working”, “presentation”, “type”, “presentation”, “feedback”, “range”, “people”, “type”, “reviewed”
3	“i”, “love”, “feeling”, “affectionate”, “for”, “the”, “one”, “i”, “love”, “and”, “care”, “for”	“love”, “feeling”, “affectionate”, “love”, “care”

5) Stemming

Stemming is the process of reducing or removing affixes contained in words into basic words. This process can affect the system in measuring accuracy. In this research, the stemming process will use a library from the Natural Language Toolkit (NLTK), namely Porter Stemmer. Table 6 is an example of the stemming process.

TABLE VI
STEMMING PROCESS

No	Before	After
1	“lose”, “touch”, “clear”, “headed”, “atm”, “inconceivable”, “higher”, “power”	“lose”, “touch”, “clear”, “headed”, “atm”, “inconceivable”, “high”, “power”
2	“working”, “presentation”, “type”, “presentation”, “feedback”, “range”, “people”, “type”, “reviewed”	“work”, “presentation”, “type”, “presentation”, “feedback”, “range”, “people”, “type”, “review”
3	“love”, “feeling”, “affectionate”, “love”, “care”	“love”, “feel”, “affectionate”, “love”, “care”

C. Imbalanced Data Handling

Imbalanced Data Handling is the process of adjusting the class distribution in a dataset when there is an uneven distribution of data. Uneven data distribution makes the system not work optimally and inaccurately. It can be observed in Figure 3 that the dataset used in this study has an uneven distribution of data. Therefore, it is necessary to handle imbalanced data with undersampling and oversampling techniques.

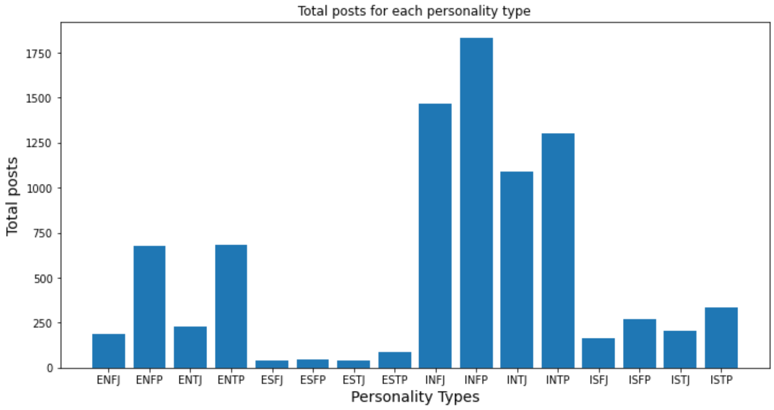

Fig. 3 Data distribution

D. *Feature Extraction*

Feature extraction is an important technique in performing dimensionality reduction to extract important features [12]. In text classification, feature extraction will structure previously unstructured text so that it can be processed in machine learning. GloVe (Global Vectors) is one of the word embedding methods that can perform and learn word representation [13] proposed by J. Pennington [9]. GloVe is a method that considers semantics and context. GloVe derives the semantic relationship between words based on the co-occurrence matrix contained in the corpus. So that the resulting vector representation in the document is the average of all vectorized word representations of words in the document [9].

The GloVe works by indicating the closest word in a sentence. Such as the example in Table 5, the word "love" is adjacent to the words "i", "feeling", and "and" so that it can be calculated that "love" with "i" has a correlation value of 2 and "and" with "feeling" has a correlation value of 1 with "love". Table 7 and Table 8 will illustrate how GloVe works.

TABLE VII
DOCUMENT EXAMPLE

Document
I love feeling affectionate for the one I love and care for

TABLE VIII
GLOVE

No	**i**	**love**	**feeling**	**affectionate**	**for**	**the**	**one**	**and**	**care**
i	0	2	0	0	0	0	1	0	0
love	2	0	1	0	0	0	0	1	0
feeling	0	1	0	1	0	0	0	0	0
affectionate	0	0	1	0	1	0	0	0	0
for	0	0	0	1	0	1	0	0	1
the	0	0	0	0	1	0	1	0	0
one	1	0	0	0	0	1	0	0	0
and	0	1	0	0	0	0	0	0	1
care	0	0	0	0	1	0	0	1	0

E. *Random Forest Learning*

Random Forest classification is an algorithm method with a supervised method [14]. The random forest consists of many basic classifiers, such as decision trees that are completely independent of each other. Input a test sample to a new classifier and the class label of the sample can be determined based on the voting results of every single class [8]. Voting results or the most votes from every single class [8]. The prediction obtained from the average of each tree is the final prediction of the random forest [15]. The more trees, the more the random forest algorithm provides a high accuracy [12].

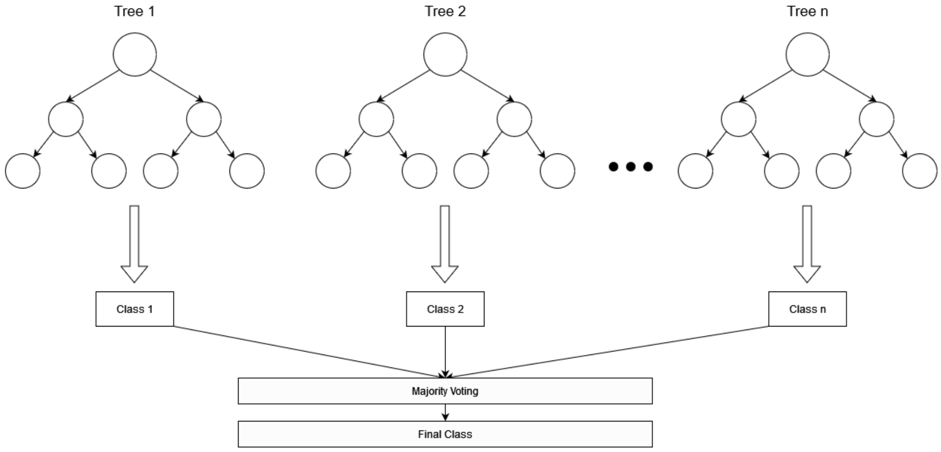

Fig. 4 Random Forest method illustration

Based on the illustration in Figure 4, it can be observed that the random forest method has the following brief steps:

1. A dataset that has k number of records takes n number of random records contained in the random forest.
2. A decision tree is built for each sample.
3. The random forest consists of many decision tree classifications that will produce output.
4. The final result will be determined based on the majority vote of every single class or average for classification and regression.

F. Performance Measure

To evaluate the model that has been created, this research uses the calculation of accuracy, precision, recall, and f1-score values.

$$Accucary = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Recall = \frac{TP}{FN + TP} \tag{2}$$

$$Precision = \frac{TP}{FP + TP} \tag{3}$$

$$F1 - Score = 1 + \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

A confusion matrix is a formula to calculate the performance of the classifier created. True Positive (TP) is the amount of data that is positive and correctly predicted as positive. False Positive (FP) is negative data but predicted as positive. False Negative (FN) is positive data but predicted as negative. True Negative (TN) is negative data and is correctly predicted as negative. The confusion matrix can be illustrated in Table 9.

TABLE IX
CONFUSION MATRIX

	Positive	**Negative**
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

III. RESULT & DISCUSSION

In this research, the class attributes that will be used are Thinkers/Feelers and Judgers/Perceivers from the Myers-Briggs Type Indicator (MBTI) model. The system will be tested with three experiments. The first experiment performs classification with pure data (baseline), the second experiment performs classification with undersample data and the last experiment performs classification with oversample data. The purpose of this experiment is to determine the impact given to pure data, undersample data, and oversample data on system performance. The results of the experiment can be seen in Table 10.

TABLE X
COMPARISON OF DATA DISTRIBUTION TECHNIQUE

Scenario	F1-Score			Accuracy		
Name	All Class	TF	JP	All Class	TF	JP
Data Normal	8.7%	66.56%	57.36%	26.82%	67.17%	63.60%
Undersample	12.04%	60.34%	55.59%	26.92%	60.4%	55.60%
Oversample	84.35%	93%	92%	84.65%	93%	92%

The table results show that the first experiment with data normal gets the best results on the thinker/feelers (TF) personality class with an f1-score value of 66.56% and an accuracy of 67.17%. The second experiment with undersample data gives the best results on the personality type with the Thinker/Feelers (TF) category with an f1-score value of 60.34% and an accuracy of 60.4%. Experiments with oversample data gave quite high results compared to previous experiments. The Thinker/Feelers (TF) and Judgers/Perceivers (JP) personality types gave the best and similar results with f1-score and accuracy of 92%-93%. On the other hand, the data with all tested classes gave good results with f1-score and accuracy of 84%.

With the results that have been obtained, it can be observed that the data processed using random forest classification with oversample techniques has excellent results with f1-score and accuracy values of 84%-92% compared to normal data and undersample data which only get an average f1-score of 43.4% and accuracy of 50%.

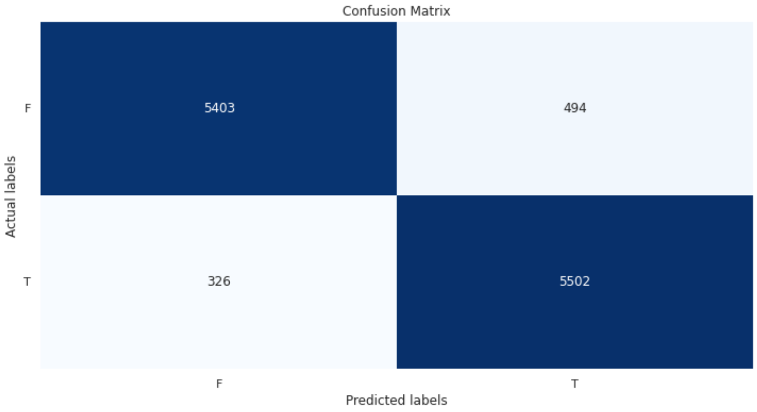

Fig. 5 Thinker/Feelers Confusion Matrix using Oversample Method (T: Thinker; F: Feeler)

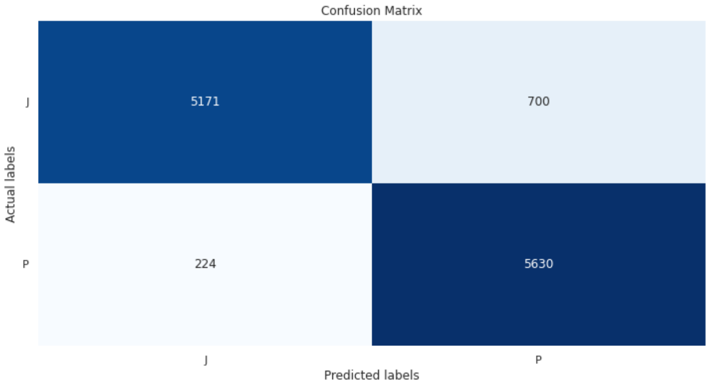

Fig. 6 Judger/Perceiver Confusion Matrix using Oversample Method (J:Judger; P:Perceiver)

In Figure 5, from 5,897 feelers data, 5,403 data were correctly classified as feeler class, and 494 data were incorrectly predicted as thinker class. On the other hand, 5,828 data from the thinker class, 5,502 data were correctly classified as the thinker class, and 326 data were incorrectly predicted as the feelers class. In Figure 6, from 5,871 judgers data, 5,171 data are correctly predicted as judgers class, and 700 data are incorrectly predicted as perceiver class. On the other hand, 5,854 data on perceivers, 5630 data were correctly classified as the perceiver class, and 224 were incorrectly classified as the judger class.

IV. CONCLUSION

In this research, a personality classification system has been built using random forest classification method and

Glove feature extraction based on how social media users live their lives and make decisions which are attribute representations of the Thinker/Feelers and Judger/Perceiver classes of the MBTI model with three experiments, namely normal data, undersample data and oversample data. The purpose of conducting these three experiments is to compare the techniques used to organize the class distribution on a dataset. Of the three experiments, the experiment with oversample data has high results compared to the other experiments. The personality type categories Thinkers/Feelers (TF) and Judgers/Perceivers (JP) gave the best and similar results with an f1-score and accuracy of 92%. On the other hand, for all classes, the results are quite good with an f1-score and accuracy of 84%.

For future work, this system can certainly be further developed by using other experiments comparing other feature extraction or other class distribution techniques. The dataset used in this research has uneven data distribution so that the results obtained in all classes are not maximized. However, if the dataset has an even distribution of data, it can improve the results of the system.

References

[1] A. Talasbek, A. Serek, M. Zhaparov, S.-M. Yoo, Y.-K. Kim, and G.-H. Jeong, *Personality Classification by Applying k-means Clustering*. 2020. doi: 10.1109/ICAIIC48513.2020.9065244.

[2] X. Wang, Y. Sui, K. Zheng, Y. Shi, and S. Cao, “Personality classification of social users based on feature fusion,” *Sensors*, vol. 21, no. 20, Oct. 2021, doi: 10.3390/s21206758.

[3] M. Carlyn, “An Assessment of the Myers-Briggs Type Indicator,” *Journal of Personality Assessment*, vol. 41, no. 5, pp. 461–473, 1977, doi: 10.1207/s15327752jpa4105_2.

[4] V. Ong *et al.*, “Personality prediction based on Twitter information in Bahasa Indonesia,” in *Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, FedCSIS 2017*, Nov. 2017, pp. 367–372. doi: 10.15439/2017F359.

[5] B. Y. Pratama and R. Sarno, “Personality classification based on Twitter text using Naive Bayes, KNN and SVM,” in *Proceedings of 2015 International Conference on Data and Software Engineering, ICODSE 2015*, Mar. 2016, pp. 170–174. doi: 10.1109/ICODSE.2015.7436992.

[6] R. I. Kurnia, Y. D. Tangkuman, and A. S. Girsang, “Classification of user comment using word2vec and SVM classifier,” *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 1, pp. 643–648, 2020, doi: 10.30534/ijatcse/2020/90912020.

[7] Y. B. N. D. Artissa, I. Arsor, and S. A. Faraby, “Personality Classification based on Facebook status text using Multinomial Naïve Bayes method,” 2019. doi: 10.1088/1742-6596/1192/1/012003.

[8] A. Parmar, R. Katariya, and V. Patel, “A Review on Random Forest An Ensemble Classifier,” pp. 1–6, 2019, doi: 10.1007/978-3-030-03146-6_86.

[9] J. Pennington, R. Socher, and C. D. Manning, “GloVe Global Vectors for Word Representation”.

[10] J. Song *et al.*, “The Random Forest Model Has the Best Accuracy Among the Four Pressure Ulcer Prediction Models Using Machine Learning Algorithms,” pp. 1–13, Feb. 2021, doi: https://doi.org/10.2147/RMHP.S297838.

[11] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, “Personality Predictions Based on User Behavior on the Facebook Social Media Platform,” pp. 1–11, Oct. 2018, doi: 10.1109/ACCESS.2018.2876502.

[12] R. N. Wykole and A. D. Thakare, “A REVIEW OF FEATURE EXTRACTION METHODS FOR TEXT CLASSIFICATION,” *International Journal of Advance Engineering and Research Development*, pp. 1–4, 2018.

[13] M. Naili, A. H. Chaibi, and H. H. ben Ghezala, “Comparative study of word embedding methods in topic,” *Procedia Computer Science*, pp. 1–10, 2017, doi: 10.1016/j.procs.2017.08.009.

[14] N. Haziqah *et al.*, “Improving Intelligent Personality Prediction using Myers-Briggs Type Indicator and Random Forest Classifier,” 2020. [Online]. Available: www.ijacsa.thesai.org

[15] “An Implementation and Explanation of the Random Forest in Python”, Accessed: Dec. 07, 2021. [Online]. Available: https://towardsdatascience.com/an-implementation-and-explanation-of-the-random-forest-in-python-77bf308a9b76