Pertanika J. Sci. & Technol. 34 (2): 1077 - 1105 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

TF-IDF Versus Linguistic Features: Evaluating Feature Sets for MBTI Personality Type Classification

Janice Allison Anak Sabang[1]*, Stephanie Chua[1], and Puteri Nor Ellyza Nohuddin[2]

[1]Data Engineering Programme, Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
[2]Faculty of Business, Higher Colleges of Technology, University City Road, PO Box 7947, Sharjah, United Arab Emirates

ABSTRACT

The Myers-Briggs Type Indicator (MBTI) is used to categorise individuals into one of the 16 types, using the acronym across the four binary personality trait divisions: Extraversion against Introversion (E/I), Intuition against Sensing (N/S), Feeling against Thinking (F/T), and Judging against Perceiving (J/P). While MBTI personality types are typically determined through questionnaire answering, the task of categorising an individual's MBTI personality type based on their written texts can be presented as a classification task that utilises machine learning and deep learning techniques. The objective of this paper is to compare and determine the best feature set for MBTI personality type classification. The methods involved in this study were text mining, feature generation, machine learning, and model evaluation. The feature generation approaches tested out were the statistical analysis approach and the semantic approach involving grammar class tagging and synonym generation. Different document-term matrix representations involving both standard and synset column representations for the semantic approach's synonyms generation are also experimented. This study found that the best MBTI personality type classification performance was obtained using the Logistic Regression model through the utilisation of the TF-IDF Top 10,000 nouns feature set from the statistical analysis with the semantic approach's grammar class tagging.

Keywords: Deep learning, feature set, machine learning, Myers-Briggs type indicator, prediction model

ARTICLE INFO

Article history:
Received: 10 October 2025
Accepted: 09 December 2026
Published: 30 April 2026

DOI: https://doi.org/10.47836/pjst.34.2.20

E-mail addresses:
23020097@siswa.unimas.my (Janice Allison Anak Sabang)
chlstephanie@unimas.my (Stephanie Chua)
pnohuddin@hct.ac.ae (Puteri Nor Ellyza Nohuddin)
* Corresponding author

INTRODUCTION

The personality types of categorisation philosophy were rooted in Carl Gustav Jung's theory of psychological types on the

e-ISSN: 2231-8526

concepts of Introversion and Extraversion, and cognitive functions. Katharine Cook Briggs and Isabel Briggs Myers, on the other hand, had contributed to the personality types of categorisations through the development of a four-letter acronym that detailed the order of a person's Jungian preferences (NERIS Analytics Limited, 2013). The four-letter acronyms were then specified by Briggs Myers and Kirby (2000) as a four binary personality traits division featuring the versus pairs of *Extraversion* (E) and *Introversion* (I), *Intuition* (N) and *Sensing* (S), *Feeling* (F) and *Thinking* (T), and lastly, *Judging* (J) and *Perceiving* (P). In total, there are 16 MBTI types from the combination of traits across the four binary personality trait divisions.

The MBTI personality type of a person is typically ascertained through answering a questionnaire that requires the respondent to input their choice tendency to a set of situations, and eventually, the accumulated answers will help to determine the MBTI personality type that this person belongs to. Instead of continuing down the manual path that involved questionnaires, machine learning and deep learning techniques can be utilised in discovering an individual's MBTI personality type. With Xue et al. (2017) observation on the presence of a connection between an individual's behaviour and their responses on social networks, this suggests that their chosen words may hint at their thought process when dealing with the various issues that arise on social media platforms. It was also discovered that the texts posted on social media may contain the emotional expressions of the said individual (Riza & Charibaldi, 2021). In Li (2021) research, the social media platforms were utilised by those with the *Extraversion* (E) trait to share information, personal information, as well as work-related topics, while those with the *Feeling* (F) trait would lean towards expressing emotions and wishes. From another observation from Akber et al. (2024) in the context of the link between Ekman emotions and MBTI personality types' traits, the researchers had categorised the "positive affection" to be portrayed by individuals with *Extraversion* (E), *Sensing* (S), and *Feeling* (F) traits, "negative aversion" to be portrayed by those with *Extraversion* (E), *Intuition* (N), *Thinking* (T), and *Perceiving* (P) traits, and lastly, the "negative distress" to be portrayed by those with *Introversion* (I), *Intuition* (N), *Thinking* (T), and *Judging* (J) traits. Hence, this opens the possibility of analysing and obtaining useful features from users with different MBTI personality types and the posts they wrote for machine learning and deep learning training purposes, enabling these models to perform predictions on the MBTI personality type of an individual through their written texts. The objective of this study is to compare and determine the best feature set and classification model for MBTI personality type identification.

LITERATURE REVIEW

Myers-Briggs Type Indicator (MBTI)

The development of Myers-Briggs Type Indicator (MBTI) was inspired by Carl Gustav Jung's theory of psychological types, where Geyer (2014), in his research, discovered that

Katharine Briggs had chosen to abandon her own personality typology in proceeding with Jung's theory after coming across Jung's book. From here on, a four-letter acronym that referred to the four binary personality traits of Extraversion (E) / Introversion (I), Intuition (N) / Sensing (S), Feeling (F) / Thinking (T), and Judging (J) / Perceiving (P) were formed (Briggs Myers & Kirby, 2000). Figure 1 shows the acronyms, NERIS Analytics Limited (2013) dubbed names, and traits composition for each of the 16 MBTI personality types.

ENFJ **The Protagonist** Extraverted Intuition Feeling Judging	**ENFP** **The Campaigner** Extraverted Intuition Feeling Perceiving	**ENTJ** **The Commander** Extraverted Intuition Thinking Judging	**ENTP** **The Debater** Extraverted Intuition Thinking Perceiving
ESFJ **The Consul** Extraverted Sensing Feeling Judging	**ESFP** **The Entertainer** Extraverted Sensing Feeling Perceiving	**ESTJ** **The Executive** Extraverted Sensing Thinking Judging	**ESTP** **The Entrepreneur** Extraverted Sensing Thinking Perceiving
INFJ **The Advocate** Introverted Intuition Feeling Judging	**INFP** **The Mediator** Introverted Intuition Feeling Perceiving	**INTJ** **The Architect** Introverted Intuition Thinking Judging	**INTP** **The Logician** Introverted Intuition Thinking Perceiving
ISFJ **The Defender** Introverted Sensing Feeling Judging	**ISFP** **The Adventurer** Introverted Sensing Feeling Perceiving	**ISTJ** **The Logistician** Introverted Sensing Thinking Judging	**ISTP** **The Virtuoso** Introverted Sensing Thinking Perceiving

Figure 1. 16 MBTI personality types of acronyms, dubbed names and their traits composition

Aligning with the initial purpose of the MBTI personality type questionnaire in assisting the suitability between an individual's personality type and their job during the World War II era, as highlighted by Geyer (2014), Fatima et al. (2022) direction of study was to suggest a list of suitable careers for individuals through text classification means as based by their determined personality profiles. Whereas in noting the trend that involved the usage of the MBTI personality types, Wang, C. et al. (2024) proved that the emergence of the MBTI personality type testing as a cultural and social phenomenon had led to an influence on the Chinese youth's behaviour and social interactions. The increase in the stated trend may be contributed by a few factors. Based on Hua and Zhou (2023), the increase in the *Barnum effect*, which was defined as "a psychological phenomenon in which people easily accept general and ambiguous personality interpretations", due to the vagueness that came with this personality testing had reduced the anxiety and depression levels among individuals, and thus left a positive effect on the adolescent mental health, as observed in this study's findings. In seeking self-recognition, the MBTI personality type testing has also helped in reaching great satisfaction in terms of an individual's personal identity and collective social interaction (Fan, 2024). In a more personal take, the MBTI personality type test allowed an individual to understand themselves along with people within their circle (Lee & Shin, 2024). Other advantages of discerning one's personality type include obtaining

better language-acquiring tactics as observed in Gu and Sharil (2023) research, as well as tailoring one's language learning method to maximise the learning task, as emphasised in Jiang (2024) research counterpart. A positive effect contributed by the MBTI personality types in the context of academic performance and learning motivation was also discovered in Ke (2024) study.

Myers-Briggs Type Indicator Classification

Dataset

In an earlier study on MBTI personality type classification by Keh and Cheng (2019), the researchers performed manual data scraping on the Personality Café forum for their research dataset. Instead of implementing the same method, a public dataset created by Mitchell (2017) on Kaggle containing over 8600 rows of data on a person's personality type and the last 50 things they wrote in Personality Café forum was mostly opted by several researchers through the expense of different years, namely, Amirhosseini and Kazemian (2020), Khan et al. (2020), Mushtaq et al. (2020), Vaddem and Agarwal (2020), Basto (2021), Choong and Varathan (2021), Kaushal et al. (2021), Maulidah and Pardede (2021), Ren et al. (2021), Ontoum and Chan (2022), Ryan et al. (2023), Zhang (2023), and Wang, Y. (2024). As based on Choong and Varathan (2021) emphasis, the factors that led to utilising the Mitchell (2017) dataset were due to its availability and size, not consisting of microblogs, and had been cited in other research at least twice to enable comparison. A concern from Choong and Varathan (2021) on a different handling of mircroblogs' texts would be due to a few factors emphasised by Stajner and Yenikent (2021) in their study, where these texts were composed of incorrect form of grammar as well as punctuations uses that caused the post author to appear *Extraverted* (E) as based on the text due to the repeated usage of exclamation marks in these platforms. Despite opting for another personality typing known as the Big Five Model, Garg and Garg (2021) had utilised Mitchell (2017) dataset in their research by mapping the MBTI personality type's binary traits to the four personality terms of the Big Five Model, with an exception to the *Neuroticism* trait. Alsini et al. (2024) had adapted the Mitchell (2017) dataset by focusing on the Big Five Model's *Agreeableness* trait by basing the trait on the *Feeling* and *Thinking* binary traits. Adawadkar and Gandhi (2023), on the other hand, chose a different public dataset from Kaggle, where an MBTI personality type-based KPMI dataset that contained survey responses reflecting the respondents' psychological choices was used instead. Some researchers, on the other hand, have utilised different datasets in their studies. A crowd-sourced database from Rev. Emmy Kegler's personality sets was used in Shafi et al. (2021) study, while a combination of datasets featuring Mitchell (2017) dataset, Reddit MBTI9K datasets, and PANDORA dataset was utilised in Cerkez et al. (2021) study. A summary of the datasets utilised by different authors mentioned is shown in Table 1.

In countering the concern about the presence of disproportioned MBTI personality types in classes, the undersampling technique was implemented by Ren et al. (2021),

while both undersampling and oversampling techniques were applied by Khan et al. (2020), Maulidah and Pardede (2021), and Ontoum and Chan (2022) in their respective research. Choosing a different option, Ryan et al. (2023) utilised the Synthetic Minority Oversampling Technique (SMOTE) in their study.

Table 1
Datasets used by different authors

Dataset	Author
Manual data scraping	(Keh & Cheng, 2019)
Single public dataset	(Amirhosseini & Kazemian, 2020) (Khan et al., 2020) (Mushtaq et al., 2020) (Vaddem & Agarwal, 2020) (Basto, 2021) (Choong & Varathan, 2021) (Kaushal et al., 2021) (Maulidah & Pardede, 2021) (Ren et al., 2021) (Shafi et al., 2021) (Ontoum & Chan, 2022) (Adawadkar & Gandhi, 2023) (Ryan et al., 2023) (Zhang, 2023) (Wang, Y., 2024)
Multiple public datasets	(Cerkez et al., 2021)

Data Preprocessing

All the research mentioned in the previous section had performed data preprocessing except for Basto (2021) due to the researcher's take on the limitation posed by data preprocessing and text representation tasks in identifying patterns within the abstracted data. Shafi et al. (2021) study did not detail any preprocessing steps done on the dataset either. A summary of the removed elements and varied preprocessing steps included in different research is shown in Figures 2 and 3, respectively.

Authors	Year	Removed Elements							
		Links	Stopwords	Special Characters	Emoticons	Numbers	One-letter words	Nouns	MBTI profile strings
Keh & Cheng	2019			✓					✓
Amirhosseini & Kazemian	2020	✓	✓						✓
Khan et al.			✓						
Mushtaq et al.		✓							✓
Vaddem & Agarwal		✓	✓	✓					✓
Cerkez et al.	2021	✓	✓	✓		✓	✓		
Choong & Varathan		✓	✓					✓	✓
Garg & Garg			✓						✓
Kaushal et al.		✓	✓						
Maulidah & Pardede		✓	✓	✓	✓				
Ren et al.		✓	✓	✓					
Ontoum & Chan	2022	✓	✓						
Ryan et al.	2023	✓	✓	✓					
Zhang			✓	✓					
Ashraf et al.	2024	✓	✓	✓	✓				
Y. Wang		✓	✓	✓		✓			
Alsini et al.		✓	✓	✓					
Hartono et al.	2025		✓						
Yang et al.			✓						

Figure 2. Removed elements in the preprocessing step

Authors	Year	Preprocessing Steps							
		Links & Usernames Conversion	Lowercase Conversion	Shortforms Conversion	Emoticons Conversion	Numeric Digits Conversion	Text Tokenisation	Text Lemmatisation	Text Stemming
Keh & Cheng	2019		✓	✓			✓		
Amirhosseini & Kazemian	2020							✓	
Khan et al.							✓		✓
Mushtaq et al.			✓					✓	
Vaddem & Agarwal								✓	
Choong & Varathan	2021				✓	✓	✓	✓	
Garg & Garg								✓	
Kaushal et al.							✓	✓	✓
Maulidah & Pardede			✓				✓	✓	
Ontoum & Chan	2022						✓	✓	
Ryan et al.	2023		✓				✓	✓	
Zhang		✓					✓		
Ashraf et al.	2024						✓	✓	
Y. Wang			✓				✓	✓	
Alsini et al.			✓				✓		
Hartono et al.	2025							✓	
Yang et al.			✓						

Figure 3. Further preprocessing steps taken

Feature Selection

The Term Frequency-Inverse Document Frequency (TF-IDF) technique allowed a vocabulary dictionary to be learned by the machine learning models from a token counts matrix, as well as output a term-document matrix (Amirhosseini & Kazemian, 2020). In highlighting this technique's usefulness, Khan et al. (2020) had implied on the TF-IDF score's ability to adjust the weights between the most regular, general words, and less utilised words, while Mushtaq et al. (2020) inputted on the contribution in terms of the capability of detecting words based on its importance to the increase in the machine learning models' performance efficiency. This technique was implemented by several researchers in their respective studies, namely, Amirhosseini and Kazemian (2020), Khan et al. (2020), Vaddem and Agarwal (2020), Mushtaq et al. (2020), Kaushal et al. (2021), Ontoum and Chan (2022), Zhang (2023), and Alsini et al. (2024). Ontoum and Chan (2022) had an additional utilisation of the BOW approach alongside TF-IDF.

Word embeddings were also applied in a few of the observed studies' feature selection step. Ryan et al. (2023) chose the Continuous Bag of Words (CBOW) model upon taking Mikolov et al. (2013) discovery on the quicker training time and better frequent words representation of the CBOW model into consideration. In a comparison between CBOW and Skip-gram models, Ryan et al. (2023) observed that the CBOW model utilised a series of context words in a phrase to predict a single word, while the opposite approach of utilising a single provided word in predicting a series of context words was done by

the Skip-gram model. Apart from that, Ren et al. (2021) had implemented a fine-tuned Bidirectional Encoder Representations from Transformers (BERT) model for a sentence-level embedding task, where multiple successive sentences from each sample data were included instead of only separating each sentence from the sample data as done in Keh and Cheng (2019) study. Adapting Keh and Cheng (2019) approach, Zhang (2023) and Ashraf et al. (2024) respective research had utilised the BERT model to perform word embedding, where the latter researchers observed that BERT had better performance than GloVe in the said task. Cerkez et al. (2021) had opted for FastText in performing the word embedding task in their research, while Maulidah and Pardede (2021) utilised an embedding layer with a vector length of 100 for word representation. Besides that, Alsini et al. (2024) experimented with Word2Vec, GloVe, and sentence embedding.

Table 2
Feature selection method utilised by different authors

Feature Selection Method	Author
Term Frequency-Inverse Document Frequency (TF-IDF)	(Amirhosseini & Kazemian, 2020) (Khan et al., 2020) (Vaddem & Agarwal, 2020) (Mushtaq et al., 2020) (Kaushal et al., 2021) (Ontoum & Chan, 2022) (Zhang, 2023) (Alsini et al., 2024)
Word embeddings	(Keh & Cheng, 2019) (Cerkez et al., 2021) (Maulidah & Pardede, 2021) (Ren et al., 2021) (Ryan et al., 2023) (Zhang, 2023) (Ashraf et al., 2024) (Alsini et al., 2024)
Sentiment analysis	(Basto, 2021) (Ren et al., 2021)
Grammar analysis	(Basto, 2021)

In a separate experiment by Basto (2021), the researcher implemented sentiment analysis to capture additional patterns in texts, aspect analysis to pinpoint critical aspects in the dataset, and grammar analysis to calculate the percentage of occurrences of syntactic terms in the personality types of traits. From this research, it was found that the sentiment analysis feature had helped in reaching the best accuracy score in the binary traits' classes of E/I, N/S, and F/T, with the scores of 78.96%, 85.95%, and 75.43%, respectively (Basto, 2021). Whereas for the J/P binary class in the same study by Basto (2021), the highest score of 64.29% was achieved through a feature generated by a BERT model. Similarly, Ren et al. (2021) had also implemented the sentiment analysis as part of the research's feature selection step through the utilisation of a common-sense knowledge extraction tool named SenticNet5.The feature selection methods discussed are summarised in Table 2.

Comparative Analysis

The MBTI personality type can be divided into binary classification and multiclass classification approaches, where the binary classification task was viewed in the context

of separated binary personality traits of E/I, N/S, F/T, and J/P, as based on Briggs Myers and Kirby (2000) research, while the multiclass classification task was viewed in the context where each MBTI personality type was captured as a single class, disregarding the composition of the four binary personality traits of the said MBTI personality type. Hence, unlike the binary classification's multiple outcomes that came from the four binary categories, there will only be one outcome from each model in the multiclass classification approach. An analysis of the classification models for both MBTI personality type binary classification and multiclass classification will be done to discover the best-performing model utilised in different research.

It was observed that the same dataset provided by Mitchell (2017) was consistently used in all the binary classification research discussed. In terms of the utilisation of the XGBoost classification model, Khan et al. (2020) model proved to obtain the best performance as compared to the researchers who utilised the same model, with an achieved score of over 90% in all binary personality traits categories. Whereas for a similar implementation of the RNN model across Ontoum and Chan (2022) and Amirhosseini and Kazemian (2020) studies, Ontoum and Chan (2022) model obtained a better score as compared to the latter. The LR model utilised in Zhang (2023) research scored better accuracy scores in comparison to Vaddem and Agarwal (2020), Basto (2021), Kaushal et al. (2021), Maulidah and Pardede (2021), and Wang, Y. (2024), respectively. Khan et al. (2020) SVM and RF models had proved to perform the best among the similar models implemented across different studies. Being the only research that utilised the CNN model for MBTI personality type classification, Ren et al. (2021) model scored an accuracy of over 80% for all categories. On the other hand, Basto (2021) BERT model had an overall better performance than the researcher's LSTM model. Focusing on the F/T category in observing the Big Five Model's equivalent of the *Agreeableness* trait, Alsini et al. (2024) found that the SVM model had outperformed the other classification models with scores of 84% in all evaluation metrics. Thus, it can be concluded that Khan et al. (2020) XGBoost model had the best accuracy performance in the binary MBTI personality type approach, with the respective scores of 0.9937, 0.9992, 0.9455, and 0.9553 in the E/I, N/S, F/T, and J/P categories. The accuracy scores from the discussed studies are all compiled in Table 3.

A few of the researchers had also included the F_1 score as their model evaluation metric alongside the accuracy score, while Ryan et al. (2023) had opted to only focus on the F_1 score metric in their research. In comparison, Ashraf et al. (2024) RF model was the best-performing classification model due to achieving better F_1 scores than Khan et al. (2020) XGBoost model, Maulidah and Pardede (2021) RF model, Ryan et al. (2023) LR model, Basto (2021) BERT model, and Alsini et al. (2024) LSTM model with sentence embeddings for the F/T category only. The F_1 scores from the mentioned researchers are compiled in Table 4.

Table 3
Accuracy scores compilation for the binary classification approach

Classification Models	Author	MBTI Binary Personality Traits			
		E/I	N/S	F/T	J/P
Extreme Gradient Boosting (XGBoost)	(Amirhosseini & Kazemian, 2020)	0.7817	0.8606	0.7178	0.6570
	(Khan et al., 2020)	0.9937	0.9992	0.9455	0.9553
	(Mushtaq et al., 2020)	0.8901	0.8593	0.8419	0.8542
	(Vaddem & Agarwal, 2020)	0.7621	0.8562	0.7501	0.6376
	(Maulidah & Pardede, 2021)	0.7192	0.7727	0.7050	0.6509
	(Alsini et al., 2024)	-	-	0.8300	-
Recurrent Neural Network (RNN)	(Amirhosseini & Kazemian, 2020)	0.6760	0.6200	0.7780	0.6370
	(Ontoum & Chan, 2022)	0.8359	0.9322	0.8000	0.7740
Logistic Regression (LR)	(Vaddem & Agarwal, 2020)	0.7907	0.8668	0.7787	0.6717
	(Basto, 2021)	0.7896	0.8595	0.7543	0.6333
	(Kaushal et al., 2021)	0.8225	0.8801	0.7935	0.7367
	(Maulidah & Pardede, 2021)	0.7625	0.8634	0.7233	0.6473
	(Zhang, 2023)	0.8937	0.9026	0.9023	0.8121
	(Wang, Y., 2024)	0.8580	0.8930	0.8620	0.7910
	(Alsini et al., 2024)	-	-	0.8400	-
Support Vector Machine (SVM)	(Vaddem & Agarwal, 2020)	0.7639	0.8658	0.7676	0.6842
	(Khan et al., 2020)	0.9594	0.9808	0.9263	0.9137
	(Kaushal et al., 2021)	0.8428	0.8820	0.8345	0.7856
	(Maulidah & Pardede, 2021)	0.8738	0.9759	0.7417	0.7191
	(Ontoum & Chan, 2022)	0.8215	0.8732	0.8049	0.7270
	(Alsini et al., 2024)	-	-	0.8400	-
Random Forest (RF)	(Khan et al., 2020)	0.9836	0.9945	0.8215	0.9162
	(Maulidah & Pardede, 2021)	0.9495	0.9893	0.7119	0.7406
	(Kaushal et al., 2021)	0.7709	0.8612	0.7635	0.6547
	(Alsini et al., 2024)	-	-	0.7900	-
Convolutional Neural Network (CNN)	(Ren et al., 2021)	0.8146	0.9251	0.8357	0.8236
BERT	(Basto, 2021)	0.7646	0.8636	0.6829	0.6429
LSTM	(Basto, 2021)	0.7669	0.8541	0.4897	0.6063
	(Alsini et al., 2024)	-	-	0.9017	-

As opposed to the binary classification, the researchers who attempted the multiclass classification approach had mainly utilised the accuracy score as the evaluation metric. It was also observed that there was a variation in the MBTI personality types of datasets used in different research for this approach. In one of the earliest studies done by Keh and Cheng (2019), the BERT classification model utilised in this research had only achieved an accuracy score of 0.4790 for a self-collected Personality Café forum dataset.

Table 4
F_1 scores compilation for the binary classification approach

Classification Models	Author	MBTI Binary Personality Traits			
		E/I	N/S	F/T	J/P
Extreme Gradient Boosting (XGBoost)	(Khan et al., 2020)	0.9856	0.9975	0.9472	0.9619
	(Maulidah & Pardede, 2021)	0.7182	0.7725	0.7051	0.6505
Logistic Regression (LR)	(Ryan et al., 2023)	0.8389	0.8821	0.8561	0.7578
	(Maulidah & Pardede, 2021)	0.6669	0.6939	0.7131	0.6314
Random Forest (RF)	(Maulidah & Pardede, 2021)	0.9450	0.9886	0.7306	0.7348
	(Ashraf et al., 2024)	0.9898	0.9975	0.9818	0.9823
BERT	(Basto, 2021)	0.3300	0.0000	0.7200	0.7200
Long-Short Term Memory (LSTM)	(Alsini et al., 2024)	-	-	0.9020	-

Table 5
Accuracy scores compilation for the multiclass classification approach

Author	Classification Models	Dataset	Accuracy
(Keh & Cheng, 2019)	BERT	Personality Café forum	0.4790
(Vaddem & Agarwal, 2020)	Extreme Gradient Boosting (XGBoost)	(Mitchell, 2017)	0.3117
(Shafi et al., 2021)	Ensemble Bagged Trees	Rev. Emmy Kegler's dataset	0.7075
(Cerkez et al., 2021)	Convolutional Neural Network (CNN)	(Mitchell, 2017), Reddit MBTI9k datasets & PANDORA dataset	0.6700
(Ontoum & Chan, 2022)	Naïve Bayes (NB)	(Mitchell, 2017)	0.4103
	Support Vector Machines (SVM)		0.4197
	Recurrent Neural Network (RNN)		0.4975
(Wang, Y., 2024)	Logistic Regression (LR)		0.6200

From the compilation of the multiclass MBTI personality type classification shown in Table 5, it was further observed that the highest accuracy was obtained by Shafi et al. (2021) Ensemble Bagged Trees model on Rev. Emmy Kegler's dataset with a score of 0.7075. In another comparison between the studies done on Mitchell (2017) dataset, the best performing model in terms of accuracy was the LR model in Wang, Y. (2024) with a score of 0.6200.

A study by Naz et al. (2025) offered an insight into the considerations on the variety of classification models chosen in the previous studies, where the researchers had discovered that the NB, LR, and SVM models had performed consistently in all four evaluation metrics, while the XGBoost model was noted to have contributed to better results through

overfitting reduction and generalisation enhancement. Besides that, the CNN and LSTM models obtained high scores in the precision and recall metrics, whereas another observation came from the BERT model, which showed to be able to offer variability in dependence on a given architecture and its implementation (Naz et al., 2025).

In gaining a deeper understanding of the MBTI personality types, the traits and psychological functions that make up each personality type were observed. The binary and multiclass MBTI personality type classification approaches were also discussed in depth in this section. As a conclusion, the best performing model in the binary and multiclass classification approaches through the utilisation of Mitchell (2017) dataset was Khan et al. (2020) XGBoost model and Wang, Y. (2024) LR model, respectively.

METHODOLOGY

The steps taken in this study's methodology were text mining, feature generation, machine learning, and model evaluation. The metrics involved in determining the best-performing classification model were accuracy, precision, recall, and F_1 score. Figure 4 shows the methodology processes involved in the study.

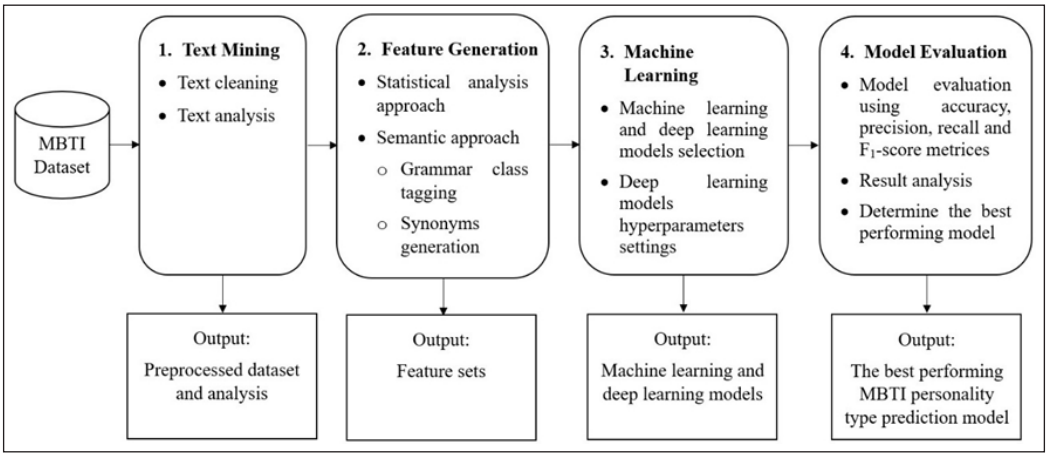

Figure 4. Methodology processes

MBTI Dataset

As done in most of the mentioned past research, this study similarly utilised the MBTI personality type dataset provided by Mitchell (2017) on a public data science platform named Kaggle. This dataset contained a total of 8 675 rows of data with two columns specifying the MBTI personality type of an individual, and their compilation of text posts on the Personality Café forum. The composition of each MBTI personality type's rows within Mitchell (2017) The dataset is shown in Table 6.

Text Mining

There were two steps included under the text mining process, namely, text cleaning and text analysis. In eliminating irrelevant elements from the raw dataset, text cleaning was done through the removal of delimiters, links or Universal Resource Locators (URLs), numbers, punctuations, stop words, and MBTI-related acronyms, lowercasing, shortened words or contractions expansion, lemmatisation, and WordNet library word existence checking. The removal of MBTI-related acronyms was made based on the consideration highlighted by Keh and Cheng (2019), where explicit mentions of the MBTI personality types may "distort the task or make the task too easy" for the classification model. The WordNet library word existence checking step, on the other hand, was taken to ensure that only English words, as referred from the WordNet library, were contained in the cleaned dataset.

Table 6
Composition of each MBTI personality type in Mitchell (2017) dataset

MBTI	Rows
ENFJ	190
ENFP	675
ENTJ	231
ENTP	685
ESFJ	42
ESFP	48
ESTJ	39
ESTP	89
INFJ	1 470
INFP	1 832
INTJ	1 091
INTP	1 304
ISFJ	166
ISFP	271
ISTJ	205
ISTP	337

For text analysis, the composition of total words and unique words in both the raw dataset and the pre-processed dataset will be compared. After going through the layers of preprocessing steps, the total number of words and unique words in the pre-processed dataset has drastically reduced in comparison to that of the raw dataset. The difference between the total words and unique words composition of the two versions of the dataset is shown in Table 7.

Table 7
Total words and total unique words in the raw dataset and pre-processed dataset

Dataset Version	Total Words	Total Unique Words
Raw Dataset	10 628 873	598 960
Pre-processed Dataset	4 391 333	30 625

Feature Generation

The feature generation approaches that were attempted under this process were a statistical analysis approach and a semantic approach, with the latter consisting of grammar class tagging and synonym generation. The details for each of the approaches were further discussed in the following subsections.

Statistical Analysis Approach

The specified statistical analysis approach was the Term Frequency-Inverse Document Frequency (TF-IDF), which was one of the commonly used approaches in the past research discussed. Due to this reason, the TF-IDF will be implemented as a baseline comparison to the other feature generation approaches experimented with in this study. The pseudocode for the steps taken in this approach is shown in Figure 5.

```
START

      IMPORT sklearn library's TfidfVectorizer

      Declare TfidfVectorizer as vectoriser

      READ preprocessed file

      Generate TF-IDF values for words in preprocessed file using vectoriser

      Filter TF-IDF top words

      WRITE TF-IDF top words into feature set file

END
```

Figure 5. Pseudocode for statistical analysis approach

Semantic Approach

There were two semantic approaches implemented under this section, namely, grammar class tagging and synonym generation.

Grammar Class Tagging

For grammar class tagging, only words with the same grammar class will be included within a feature set. With this being said, the grammar classes that will be focused on in this step would be the adjectives (a), verbs (v), and nouns (n). The grammar class filtering was done using WordNet library's Part of Speech (POS) tags, where each word's lemma form would be assigned a POS tag. Hence, relevant words will be accumulated based on their grammar class and will then be part of the feature set of that specific grammar class. The pseudocode for the steps taken in grammar class tagging is shown in Figure 6.

Synonyms Generation

The second approach will involve the generation of synonyms for keywords obtained in the grammar class tagging. By including synonyms as a possible part of the classification models' feature set, words that have a similar meaning to the keywords throughout the dataset would also be able to function as features. The synonyms generation step was also

```
START
      READ preprocessed file
      FOR word in file
            APPEND word into preprocessed word list
      END FOR
      SET preprocessed word list to only contain unique words
      FOR word in preprocessed word list
            Obtain POS tag for word from WordNet library's synsets
            IF POS tag equals to the sought grammar class' POS tag
                  APPEND word into finalised list
            END IF
      END FOR
      Remove empty strings from finalised list
      SET finalised list to only contain unique words
END
```

Figure 6. Pseudocode for grammar class tagging

```
START
      FOR word in finalised list
            APPEND word into synonyms list
            FOR synonyms in WordNet library's synsets
                  Obtain word's lemma
                  APPEND lemma into synonyms list
            END FOR
      END FOR
      Remove empty strings from synonyms list
      Lowercase all words in synonyms list
      SET synonyms list to only contain unique words
END
```

Figure 7. Pseudocode for synonyms generation

done through the utilisation of the WordNet library, and while the generated synonyms may overlap with either the keywords or the other keywords' synonyms, these repetitions will be removed in the finalised version of the feature set by only maintaining the unique words. The pseudocode for the steps taken in synonym generation is shown in Figure 7 below, as proceeded with the example of generated synonyms obtained from this approach in Figure 8.

Words	Synonyms
mythological	fabulous, mythic, mythologic, mythical
neurological	neurologic
prestigious	esteemed, honored
nice	gracious, squeamish, nice, prissy, skillful, overnice, courteous, decent, dainty
solvable	resolvable

Figure 8. Examples of generated synonyms

Combined Approaches

A combination of approaches mentioned in the previous sections into feature sets was also attempted. The variations of combined approaches include TF-IDF Top words with grammar class tagging filtering, as well as TF-IDF top words with grammar class tagging filtering + synonyms for both standard column representation and synset column representation.

TF-IDF Top Words with Grammar Class Tagging Filtering

For this approach, the words filtered out from the TF-IDF step will undergo another layer of filtering based on grammar class tagging, and thus, further reducing the feature set to only containing top words from TF-IDF that belonged to a certain grammar class. The pseudocode for this combination of approaches is shown in Figure 9.

```
START
      IMPORT sklearn library's TfidfVectorizer
      Declare TfidfVectorizer as vectoriser
      READ preprocessed file
      Generate TF-IDF values for words in preprocessed file using vectoriser
      Filter TF-IDF top words
      WRITE TF-IDF top words into TF-IDF feature set file
      READ TF-IDF feature set file
      FOR word in file
            APPEND word into TF-IDF word list
      END FOR
      FOR word in TF-IDF word list
            Obtain POS tag for word from WordNet library's synsets
            IF POS tag equals to the sought grammar class' POS tag
                  APPEND word into finalised list
            END IF
      END FOR
      Remove empty strings from finalised list
      SET finalised list to only contain unique words
END
```

Figure 9. Pseudocode for TF-IDF top words with grammar class tagging filtering

TF-IDF Top Words with Grammar Class Tagging Filtering + Synonyms

Extending the above combined approaches, words obtained from the synonyms generation were added into this variation, where words that were filtered by both the TF-IDF and grammar class tagging will undergo the synonyms generation step in expanding the words contained in the feature set. The pseudocode for this combined approach is shown in Figure 10.

```
START
      IMPORT sklearn library's TfidfVectorizer
      Declare TfidfVectorizer as vectoriser
      READ preprocessed file
      Generate TF-IDF values for words in preprocessed file using vectoriser
      Filter TF-IDF top words
      WRITE TF-IDF top words into TF-IDF feature set file
      READ TF-IDF feature set file
      FOR word in file
            APPEND word into TF-IDF word list
      END FOR
      FOR word in TF-IDF word list
            Obtain POS tag for word from WordNet library's synsets
            IF POS tag equals to the sought grammar class' POS tag
                  APPEND word into finalised list
            END IF
      END FOR
      Remove empty strings from finalised list
      SET finalised list to only contain unique words
      FOR word in finalised list
            APPEND word into synonyms list
            FOR synonyms in WordNet library's synsets
                  Obtain word's lemma
                  APPEND lemma into synonyms list
            END FOR
      END FOR
      Remove empty strings from synonyms list
      Lowercase all words in synonyms list
      SET synonyms list to only contain unique words
END
```

Figure 10. Pseudocode for TF-IDF top words with grammar class tagging filtering + synonyms

Machine Learning

Models Selection

Based on the MBTI personality type classifications observed in the previous research, there was a clear utilisation of machine learning and deep learning models. Taking the same

consideration into account, both machine learning and deep learning models were similarly utilised for the classification task. Table 8 shows the models that were implemented in this study.

Table 8
Machine learning and deep learning models utilised

Model Type	Utilised Models	Model / Layer Name
Machine Learning	Gaussian Naïve Bayes (GNB)	GaussianNB
	Multinomial Naïve Bayes (MNB)	MultinomialNB
	Support Vector Machine (SVM)	SVC
	Logistic Regression (LR)	LogisticRegression
Deep Learning	Convolutional Neural Network (CNN)	Conv1D
	Recurrent Neural Network (RNN), with Long-Short Term Memory (LSTM) layer	LSTM
	Distil Bidirectional Encoder Representations from Transformers (DistilBERT)	distilbert-base-uncased

Machine Learning Models Hyperparameters

The GNB and the MNB models had maintained the default hyperparameter settings, while some alterations were made on the SVM model's "gamma" hyperparameter and the LR models' "solver" and "random_state" hyperparameters. The hyperparameters for the GNB, MNB, SVM, and LR models utilised in this study are shown in the respective order of Tables 9, 10, 11, and 12.

Table 9
GNB model's hyperparameters details

Gaussian Naïve Bayes (GNB)	
Priors	None
Var_smoothing	1e-09

Table 10
MNB model's hyperparameters details

Multinomial Naïve Bayes (MNB)	
Alpha	1.0
Force_alpha	True
Fit_prior	True
Class_prior	None

Table 11
SVM model's hyperparameters details

Support Vector Machine (SVM)	
C	1.0
Kernel	rbf
Degree	3
Gamma	auto
Coef0	0.0
Shrinking	True
Probability	False
Tol	0.001
Cache_size	200
Class_weight	None
Verbose	False
Max_iter	-1
Decision_function_shape	ovr
Break_ties	False
Random_state	None

Deep Learning Models Hyperparameters

The hyperparameters for the LSTM and CNN models, as well as the DistilBERT model utilised in this study, are shown in Tables 13 and 14, respectively.

Training and Testing

The train_test_split function was implemented to set the size of the train and test sets from the dataset for the proceeding classification task. For this study, the train and test split proportion utilised was 80-20.

Model Evaluation

The performance of all implemented machine learning and deep learning models was evaluated using accuracy, precision, recall, and F_1 score metrics, as derived from the confusion matrix.

RESULTS AND DISCUSSION

Feature Sets

The feature sets obtained from the different feature generation approaches mentioned in the previous section had involved four feature sets from statistical analysis approach, three feature sets from semantic approach's grammar class tagging, three feature sets from semantic approach's synonyms generation with standard column representation, three feature sets from semantic approach's synonyms generation with synset column representation, with twelve feature sets for each of the combined approaches of statistical analysis approach with semantic approach's grammar class tagging filtering, statistical analysis

Table 12
LR model's hyperparameters details

Logistic Regression (LR)	
Penalty	l2
Dual	False
Tol	0.0001
C	1.0
Fit_intercept	True
Intercept_scaling	1
Class_weight	None
Random_state	1
Solver	liblinear
Max_iter	100
Multi_class	deprecated
Verbose	0
Warm_start	False
N_jobs	None
L1_ratio	None

Table 13
LSTM and CNN models' hyperparameter details

Hyperparameters	Long-Short Term Memory (LSTM)	Convolutional Neural Network (CNN)
Batch	448	448
Epoch	25	25
Layer name	LSTM	Conv1D
Layers	2 layers (384, 256)	3 layers (128, 64, 32)
Dropout layer	2 layers (0.25, 0.25)	-
Activation	Softmax	Sigmoid
Optimiser	Adam	Adam

Table 14
DistilBERT model's hyperparameters details

Distil Bidirectional Encoder Representations from Transformers (DistilBERT)	
Batch	32
Learning rate	10^{-5}
Early stopping epoch	5
Warm-up steps	100
Weight decay	0.01

approach with semantic approach's grammar class tagging filtering and synonyms generation with standard column representation, and statistical analysis approach with semantic approach's grammar class tagging filtering and synonyms generation with synset column representation. All of the stated feature sets are specified in Table 15.

Table 15
Feature sets and their count across different feature generation approaches

Feature Generation Approach	Feature Set
Statistical Analysis (TF-IDF)	Top 1 000 words (1 000 features)
	Top 5 000 words (5 000 features)
	Top 10 000 words (10 000 features)
	Top 25 000 words (25 000 features)
Semantic: Grammar Class Tagging	Adjectives (2 827 features)
	Verbs (3 199 features)
	Nouns (23 087 features)
Semantic: Synonyms Generation (Standard Representation)	Adjectives + Synonyms (6 934 features)
	Verbs + Synonyms (10 916 features)
	Nouns + Synonyms (35 736 features)
Semantic: Synonyms Generation (Synset Representation)	Adjectives + Synonyms (2 229 features)
	Verbs + Synonyms (2 354 features)
	Nouns + Synonyms (15 375 features)
Statistical Analysis + Semantic: Statistical Analysis (TF-IDF) Grammar Class Tagging Filtered	Top 1 000 Adjectives (122 features)
	Top 1 000 Verbs (117 features)
	Top 1 000 Nouns (629 features)
	Top 5 000 Adjectives (539 features)
	Top 5 000 Verbs (584 features)
	Top 5 000 Nouns (3 532 features)
	Top 10 000 Adjectives (977 features)
	Top 10 000 Verbs (1 170 features)
	Top 10 000 Nouns (7 268 features)
	Top 25 000 Adjectives (2 288 features)
	Top 25 000 Verbs (2 779 features)
	Top 25 000 Nouns (18 626 features)
Statistical Analysis + Semantic: Statistical Analysis (TF-IDF) Grammar Class Tagging Filtered + Synonyms Generation (Standard Representation)	Top 1 000 Adjectives Synonyms (1 007 features)
	Top 1 000 Verbs Synonyms (1 160 features)
	Top 1 000 Nouns Synonyms (4 222 features)
	Top 5 000 Adjectives Synonyms (2 611 features)
	Top 5 000 Verbs Synonyms (3 934 features)
	Top 5 000 Nouns Synonyms (12 449 features)
	Top 10 000 Adjectives Synonyms (3 666 features)

Table 15 *(continued)*

Feature Generation Approach	Feature Set
Statistical Analysis + Semantic: Statistical Analysis (TF-IDF) Grammar Class Tagging Filtered + Synonyms Generation (Synset Representation)	Top 10 000 Verbs Synonyms (6 202 features)
	Top 10 000 Nouns Synonyms (18 464 features)
	Top 25 000 Adjectives Synonyms (6 135 features)
	Top 25 000 Verbs Synonyms (10 177 features)
	Top 25 000 Nouns Synonyms (31 393 features)
	Top 1 000 Adjectives Synonyms (98 features)
	Top 1 000 Verbs Synonyms (94 features)
	Top 1 000 Nouns Synonyms (404 features)
	Top 5 000 Adjectives Synonyms (445 features)
	Top 5 000 Verbs Synonyms (483 features)
	Top 5 000 Nouns Synonyms (2 258 features)
	Top 10 000 Adjectives Synonyms (787 features)
	Top 10 000 Verbs Synonyms (950 features)
	Top 10 000 Nouns Synonyms (4 743 features)
	Top 25 000 Adjectives Synonyms (1 805 features)
	Top 25 000 Verbs Synonyms (2 092 features)
	Top 25 000 Nouns Synonyms (12 504 features)

Document-term Matrix Representation

Before proceeding with the classification step, a document-term matrix was generated for each of the feature sets involved, where the presence of a particular feature in each row of posts was marked with either "1" for a present word or "0" for an absent word. Since each column represented a single feature, the size of the document-term matrix will expand based on the size of the feature set itself. Thus, a feature set with a larger number of features will have a more extensive document-term matrix size, given that a single feature is represented in its respective column within the matrix.

To reduce the dimensionality of the document-term matrix, a word vector based on synset column representation for the semantic approach's synonyms generation feature sets was experimented with in this study. For this take, each column of the document-term matrix will represent the presence of a feature and its synonyms instead of having one column representing only one feature's presence throughout the dataset, as done in the standard representation previously explained. Hence, the classification runs for both the standard and synset column representations for the semantic approach's synonyms generation feature sets will be included in the following section as well.

Model Evaluation

The best classification results done on the statistical analysis approach, semantic approach's grammar class tagging, semantic approach's synonyms generation with standard representation, semantic approach's synonyms generation with synset column representation, alongside the attempts done on a combination of feature generation approaches, namely, statistical analysis with semantic approach's grammar class tagging filtering, statistical analysis with semantic approach's grammar class tagging filtering + synonyms generation with standard column representation, and statistical analysis with semantic approach's grammar class tagging filtering + synonyms generation with synset column representation, are as shown in Table 16**.** Due to the limitation presented by the extensive size of the document-term matrix involved, the classification attempt on the semantic approach's nouns + synonyms feature set with standard column representation, and the combined feature generation approaches of TF-IDF Top 25 000 nouns + synonyms feature set with standard column representation had not be carried out.

Table 16
The best machine learning models' classification results from all feature generation approaches

Feature Approach	Features	Model	Accuracy	Precision	Recall	F_1 Score
TF-IDF Top 10, 000	10 000	LR	0.80	0.78	0.80	0.78
TF-IDF Top 25, 000	25 000	LR	0.89	0.88	0.89	0.88
Nouns Only	23 087	LR	0.89	0.88	0.89	0.88
Verbs + Synonyms	10 916	LR	0.87	0.86	0.87	0.86
Nouns + Synonyms	15 375	LR	0.77	0.76	0.77	0.76
TF-IDF Top 10,000 Nouns	7 268	LR	0.89	0.88	0.89	0.88
TF-IDF Top 25, 000 Nouns	18 626	LR	0.89	0.88	0.89	0.88
TF-IDF Top 5, 000 Nouns + Synonyms	12 449	LR	0.89	0.88	0.89	0.88
TF-IDF Top 10, 000 Nouns + Synonyms	18 464	LR	0.89	0.88	0.89	0.88
TF-IDF Top 10, 000 Nouns + Synonyms	4 743	LR	0.75	0.74	0.75	0.74
TF-IDF Top 25, 000 Nouns + Synonyms	12 504	GNB	0.75	0.73	0.75	0.72

Proceeding from the machine learning models, Table 17 shows the classification results for the deep learning models' MBTI personality type classifications. The DistilBERT model had outperformed the CNN and RNN models in all the tested evaluation metrics.

From the machine learning classification runs done, the highest score across all evaluation metrices was obtained by the LR model through the utilisation of TF-IDF Top 10 000, TF-IDF Top 25 000, nouns only, TF-IDF Top 10 000 nouns, TF-IDF Top 25 000 nouns,

Table 17
Deep learning models' classification result

Model	Accuracy	Precision	Recall	F_1 Score
CNN	0.86	0.86	0.86	0.85
RNN (LSTM layer)	0.86	0.85	0.86	0.86
DistilBERT	0.88	0.87	0.88	0.87

TF-IDF Top 5 000 nouns + synonyms with standard column representation, and TF-IDF Top 10 000 nouns + synonyms with standard column representation feature sets, where all had a shared score of 0.89 for accuracy, 0.88 for precision, 0.89 for recall, and 0.88 for F_1 score. Though in terms of feature set size, the TF-IDF Top 10 000 nouns from the statistical analysis with the semantic approach's grammar class tagging filtering had the smallest feature set size. Thus, the best feature set and model combination from the machine learning classification experimented in this study would be the TF-IDF Top 10 000 nouns feature set through the utilisation of the LR model. Whereas for deep learning, DistilBERT achieved the best classification result as compared to the other tested models with a score of 0.88 for accuracy, 0.87 for precision, 0.88 for recall, and 0.87 for F_1 score.

DISCUSSION

Feature Sets Interpretability Comparison

In comparing the feature sets that contribute to the classification model's high performance in the context of human interpretability and understanding, part of the statistical analysis, the semantic analysis' grammar class tagging only, and the combined approaches of statistical analysis with grammar class tagging filtering feature sets' content are accumulated in Table 18.

Since TF-IDF had relied on the statistical calculations regarding the presence of words within the dataset, the words accumulated through this approach may contain random words that had high TF-IDF values, and thus, the top words from the statistical approach

Table 18
Part of the content from TF-IDF top 1000, adjectives only, and TF-IDF top 1000 adjectives' feature sets

Feature Set	Content
TF-IDF Top 1 000	like, think, people, know, one, get, really, time, feel, thing, make, say, much, well, want, love, type, good, way, see, go, also, lot, even, always, …
Adjectives Only	ingenuous, evangelical, neutral, graphical, rendezvous, unsharpened, fallacious, unsympathetic, theocratic, untreated, antic, supranational, …
TF-IDF Top 1000 Adjectives	related, particular, green, little, normal, huge, strong, black, open, romantic, impossible, live, special, entire, intuitive, new, several, …

had been based on the ranking for each word's calculated value. In a way, the computed TF-IDF values may help the classification model to capture patterns statistically and thus utilise the words included in the feature set to perform classification.

Whereas for the semantic analysis approach, the feature sets under this approach had focused on the classification of words based on the English grammar classes. Taking the adjectives-only feature set as an example, the words accumulated in this feature set all fall under the category of the same grammar class, namely "adjectives", with the characteristic of the words contained being "a word that describes a person or thing" (Oxford University Press, n.d.). Since this relates to the basic human understanding of the English language, the grouping of words through the semantic analysis' grammar class tagging would be clearer to the human understanding as compared to the calculations of values implemented in the statistical analysis approach. Despite this, the TF-IDF feature sets had obtained a much better classification result as compared to those of the grammar class tagging's feature sets, and thus, this led to the preferred implementation of the statistical approach in classification tasks to achieve a higher classification performance over better human interpretability.

Since the domain of this research focuses on the MBTI personality types of individuals, which involves the patterns discerned from human traits, relying solely on the statistical approach may not help in gaining a better understanding of the connection that may exist between the words used by individuals and their respective MBTI personality type. Seizing the chance of compromising between the statistical approach and the semantic approach's grammar class tagging, feature sets composed of these two approaches were tested in this study. The accuracy result achieved between the utilisation of the TF-IDF top words and the TF-IDF top words with grammar class tagging filtering feature sets were the same, which was 0.89. Further comparison between the feature set sizes of the two approaches had proved that the latter had achieved the same high performance with a much smaller feature set size, where the TF-IDF Top 10 000 nouns filtered feature set contained 7 268 features, while the TF-IDF Top 10 000 feature set contained 10 000 features. Hence, instead of solely relying on the calculated values from the statistical analysis approach, combining it with the semantic approach's grammar class tagging filtering has helped to reduce the size of the document-term matrix in achieving a good performance in the MBTI personality type classification.

Linear Regression and DistilBERT Models Comparison

The comparisons on the confusion matrices and the Receiver Operating Characteristic (ROC) curve graphs for the LR model with TF-IDF Top 10 000 nouns feature set and DistilBERT model, and the top 4 wrong MBTI personality types of predictions across the LR and DistilBERT models are shown in Figure 11, Figure 12, Figure 13, Figure 14, and Table 15, respectively. As observed from the confusion matrices comparison in Figures 11 and 12, both models seemed to similarly encounter low correct predictions on the INFJ, INFP, INTJ, and INTP personality types. Similarly, the ROC curve graphs for the LR model and the DistilBERT model in Figure 13 and Figure 14, respectively, showed that

the four stated MBTI personality types plotted lines had steered the closest to the diagonal line that indicated the neutrality between a true and a false prediction. From here, it can be concluded that both models encountered confusion when predicting the INFJ, INFP, INTJ, and INTP personality types. Since the four stated MBTI personality types were ranked with the highest number of posts contained in the dataset, this may be a contributing factor to the confusion in the models' predictions. While that, the predictions done on the MBTI personality types with a lesser number of posts were solidified through the utilisation of the oversampling technique.

From the wrong predictions for both models, as shown in Table 19, it can be concluded that most prediction confusion by the models had mostly centred on the MBTI personality types with the Intuition (N) trait and thus caused the mislabels to exist between personality types with the N trait.

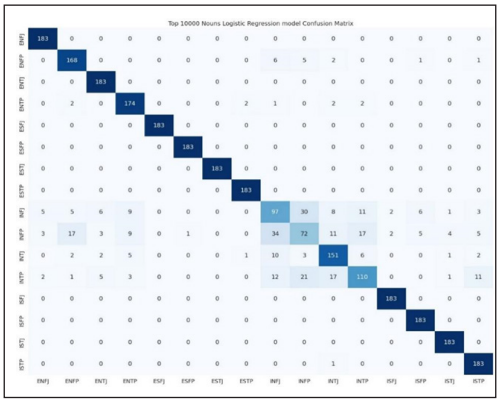

Figure 11. Logistic Regression model confusion matrix

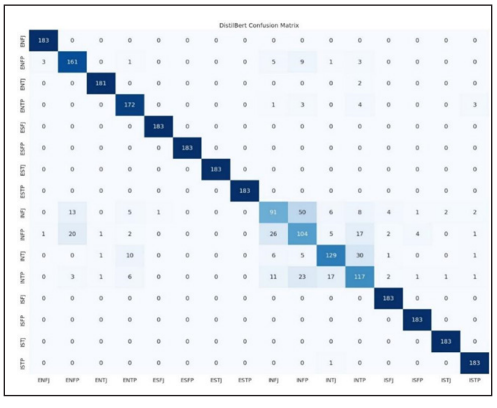

Figure 12. DistilBert model confusion matrix

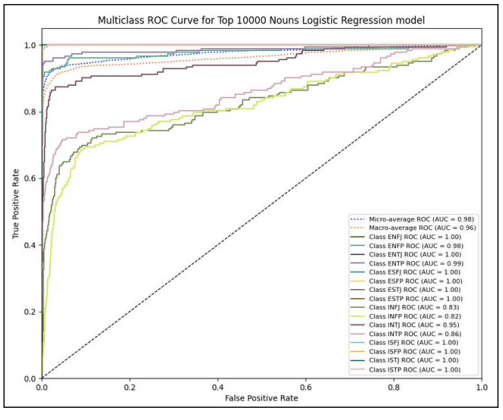

Figure 13. Logistic Regression model ROC curve

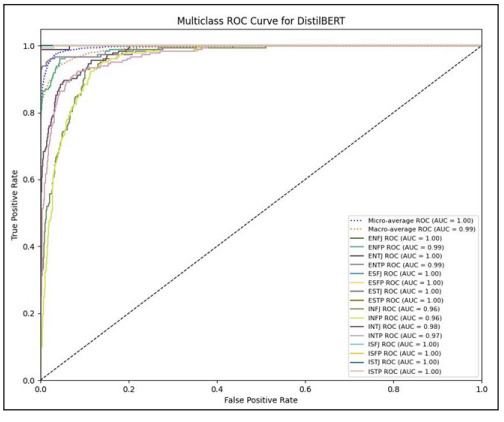

Figure 14. DistilBERT model ROC curve

Table 19
Comparison of the top 4 wrong predicted MBTI personality types across LR and DistilBERT models

Classification Model	Actual Label	Wrong Prediction
Linear Regression Model (LR)	INFJ	INFP, INTP, ENTP
	INFP	INFJ, ENFP, INTP
	INTJ	INFJ, INTP, ENTP
	INTP	INFP, INTJ, INFJ
DistilBERT Model	INFJ	INFP, ENFP, INTP
	INFP	INFJ, ENFP, INTP
	INTJ	INTP, ENTP, INFJ
	INTP	INFP, INTJ, INFJ

CONTRIBUTIONS

The contributions for this study include:

1. Features the synonyms generation

   As part of the feature generation step, this study has introduced an extension of features in the synonym generation approach through the inclusion of words' synonyms, as obtained from the WordNet library, alongside the features obtained from the grammar class tagging approach, to capture the variation of words that share the same meaning.

2. Document-term matrix size reduction

   An attempt to reduce the size of the document-term matrix was also experimented with in the synset column representation for the semantic approach's synonyms generation, where each column of the document-term matrix was represented by the presence of a feature and its synonyms, as opposed to having one column representing only a single feature's presence as done in the standard column representation of the matrix.

CONCLUSION

Based on the comparison of different feature sets used for the MBTI personality type classification done in this study, it can be concluded that the best performing classification model in terms of accuracy, precision, recall, and F_1 score metrices was the Logistic Regression (LR) model through the utilisation of the statistical analysis approach's TF-IDF Top 10 000 and Top 25 000 words, grammar class tagging's nouns only, statistical analysis with grammar class tagging filtering's TF-IDF Top 10 000 nouns and TF-IDF Top 25 000 nouns, statistical analysis with grammar class tagging filtering + synonyms generation with standard column representation's TF-IDF Top 5 000 nouns + synonyms and TF-IDF Top 10 000 nouns + synonyms as features.

The best classification result achieved from the deep learning model's utilisation had come from the DistilBERT model, with slightly lower scores across the evaluation metrics in comparison to the LR model. In an overall view, both the LR model and the DistilBERT model obtained a high score of above 80% for accuracy, precision, recall, and F_1 score metrics for the MBTI personality type classification task.

Due to the limitations encountered in this study, a possible direction to be pursued for future works related to the MBTI personality type classification would be to explore the utilisation of all nouns and their synonyms as features. Apart from that, focusing on the possible alternatives of reducing the document-term matrix size that contains many features without negatively affecting the performance of the classification model would benefit in improving the time taken for the model to perform the classification task.

ACKNOWLEDGEMENT

We extend our sincere gratitude to the Faculty of Computer Science and Information Technology (FCSIT) at Universiti Malaysia Sarawak (UNIMAS) for the facilities provided throughout this research. We would also like to thank Universiti Malaysia Sarawak (UNIMAS) for supporting the fees for this article. We confirm that this research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

LIST OF ABBREVIATIONS

ENFJ : Extraverted-Intuition-Feeling-Judging Type
ENFP : Extraverted-Intuition-Feeling-Perceiving Type
ENTJ : Extraverted-Intuition-Thinking-Judging Type
ENTP : Extraverted-Intuition-Thinking-Perceiving Type
ESFJ : Extraverted-Sensing-Feeling-Judging Type
ESFP : Extraverted-Sensing-Feeling-Perceiving Type
ESTJ : Extraverted-Sensing-Thinking-Judging Type
ESTP : Extraverted-Sensing-Thinking-Perceiving Type
INFJ : Introverted-Intuition-Feeling-Judging Type
INFP : Introverted-Intuition-Feeling-Perceiving Type
INTJ : Introverted-Intuition-Thinking-Judging Type
INTP : Introverted-Intuition-Thinking-Perceiving Type
ISFJ : Introverted-Sensing-Feeling-Judging Type
ISFP : Introverted-Sensing-Feeling-Perceiving Type
ISTJ : Introverted-Sensing-Thinking-Judging Type
ISTP : Introverted-Sensing-Thinking-Perceiving Type
MBTI : Myers-Briggs Type Indicator

REFERENCES

Adawadkar, K., & Gandhi, V. (2023). Comparative analysis of classification algorithms for classifying psychotypes. *Proceedings of the 2023 3rd International Conference on Smart Data Intelligence (ICSMDI 2023)*, 483-488. https://doi.org/10.1109/ICSMDI57622.2023.00091

Akber, M. A., Ferdousi, T., Ahmed, R., Asfara, R., Rab, R., & Zakia, U. (2024). Personality and emotion: A comprehensive analysis using contextual text embeddings. *Natural Language Processing Journal, 9*, Article 100105.https://doi.org/10.1016/j.nlp.2024.100105

Alsini, R., Naz, A., Khan, H. U., Bukhari, A., Daud, A., & Ramzan, M. (2024). Using deep learning and word embeddings for predicting human agreeableness behaviour. *Scientific Reports, 14*(1), Article 29875. https://doi.org/10.1038/s41598-024-81506-8

Amirhosseini, M. H., & Kazemian, H. (2020). Machine learning approach to personality type prediction based on the Myers–Briggs Type Indicator®. *Multimodal Technologies and Interaction, 4*(1). https://doi.org/10.3390/mti4010009

Ashraf, N., Ahmad, R. S., Bano, S., Azeem, H. M., & Naz, S. (2024). Enhancing MBTI personality prediction from text data with an advanced word embedding technique. *VFAST Transactions on Software Engineering, 12*(3), 35-43. https://doi.org/10.21015/vtse.v12i3.1864

Basto, C. (2021). Extending the abstraction of personality types based on MBTI with machine learning and natural language processing (NLP). *arXiv*. https://doi.org/10.48550/arXiv.2105.11798

Briggs Myers, I., & Kirby, L. K. (2000). *Introduction to type: A guide to understanding your results on the Myers-Briggs Type Indicator*. Consulting Psychologists Press.

Cerkez, N., Vrdoljak, B., & Skansi, S. (2021). A method for MBTI classification based on the impact of class components. *IEEE Access, 9*, 146550-146567. https://doi.org/10.1109/ACCESS.2021.3121137

Choong, E. J., & Varathan, K. D. (2021). Predicting judging-perceiving of Myers-Briggs Type Indicator (MBTI) in online social forum. *PeerJ, 9*, Article e11382. https://doi.org/10.7717/peerj.11382

Fan, X. (2024). Research on audience psychological communication in the era of new media-Taking the MBTI phenomenon as an example. *SHS Web of Conferences, 199*, Article 02018. https://doi.org/10.1051/shsconf/202419902018

Fatima, N., Gul, S., Ahmed, J., Khand, Z. H., & Mujtaba, G. (2022). A rule-based machine learning model for career selection through MBTI personality. *Mehran University Research Journal of Engineering and Technology, 41*(2), 185-196. https://doi.org/10.22581/muet1982.2202.18

Garg, S., & Garg, A. (2021). Comparison of machine learning algorithms for content-based personality resolution of tweets. *Social Sciences and Humanities Open, 4*(1), Article 100178. https://doi.org/10.1016/j.ssaho.2021.100178

Geyer, P. (2014). *Isabel Briggs Myers, psychological type, and the spirit of C.G. Jung*.

Gu, Y., & Sharil, W. N. E. H. (2023). Study on the effect of personality type on the language learning strategies of non-English major students through MBTI test. *Educational Administration Theory and Practice Journal, 29*(4), 1-19. https://doi.org/10.52152/kuey.v29i4.756

Hua, J., & Zhou, Y. X. (2023). Personality assessment usage and mental health among Chinese adolescents: A sequential mediation model of the Barnum effect and ego identity. *Frontiers in Psychology, 14*, Article 1097068. https://doi.org/10.3389/fpsyg.2023.1097068

Jiang, H. (2024). The impact of personality types on second language vocabulary acquisition of college students: Based on MBTI personality categorisation. *Journal of Education, Humanities and Social Sciences IMPES, 26*, 704-710. https://doi.org/10.54097/c5r4zm20

Kaushal, P., B. P., N. B., M. S., P., S., K., & Koundinya, A. K. (2021). Myers–Briggs personality prediction and sentiment analysis of Twitter using machine learning classifiers and BERT. *International Journal of Information Technology and Computer Science, 13*(6), 48-60. https://doi.org/10.5815/ijitcs.2021.06.04

Ke, M. (2024). The influence of MBTI personality types on college students' academic performance: The mediating role of learning motivation. *Journal of Education, Humanities and Social Sciences PSHE, 29*, 449-454.https://doi.org/10.54097/vysk5519

Keh, S. S., & Cheng, I.-T. (2019). Myers-Briggs personality classification and personality-specific language generation using pre-trained language models. *arXiv*. https://doi.org/10.48550/arXiv.1907.06333

Khan, A. S., Ahmad, H., Asghar, M. Z., Saddozai, F. K., Arif, A., & Khalid, H. A. (2020). Personality classification from online text using machine learning approach. *International Journal of Advanced Computer Science and Applications, 11*(3), 460-476. https://doi.org/10.14569/IJACSA.2020.0110358

Lee, H., & Shin, Y. (2024). A study on MBTI perceptions in South Korea: Big data analysis from the perspective of applying MBTI to contribute to the sustainable growth of communities. *Sustainability, 16*(10), Article 4152. https://doi.org/10.3390/su16104152

Li, W. (2021). *Predicting MBTI personality type of Twitter users* [Doctoral thesis, Rutgers University]. Rutgers University Libraries. https://doi.org/doi:10.7282/t3-75wc-2x18

Maulidah, M., & Pardede, H. F. (2021). Prediction of Myers–Briggs Type Indicator personality using long short-term memory. *Jurnal Elektronika dan Telekomunikasi, 21*(2), 104-111. https://doi.org/10.14203/jet.v21.104-111

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *arXiv*. https://doi.org/10.48550/arXiv.1310.4546

Mitchell, J. (2017). *(MBTI) Myers-Briggs personality type dataset* [Data set]. https://www.kaggle.com/datasets/datasnaek/mbti-type

Mushtaq, Z., Ashraf, S., & Sabahat, N. (2020). Predicting MBTI personality type with K-means clustering and gradient boosting. *Proceedings of the 2020 23rd IEEE International Multi-Topic Conference (INMIC 2020)*. https://doi.org/10.1109/INMIC50486.2020.9318078

Naz, A., Khan, H. U., Bukhari, A., Alshemaimri, B., Daud, A., & Ramzan, M. (2025). Machine and deep learning for personality traits detection: A comprehensive survey and open research challenges. *Artificial Intelligence Review, 58*(8), Article 11245. https://doi.org/10.1007/s10462-025-11245-3

NERIS Analytics Limited. (2013). *Our framework*.

Ontoum, S., & Chan, J. H. (2022). Personality type based on Myers-Briggs Type Indicator with text posting style using traditional and deep learning. *arXiv*. https://doi.org/10.48550/arXiv.2201.08717

Oxford University Press. (n.d.). *Adjective*. In *Oxford Advanced Learner's Dictionary*. Retrieved March 3, 2024, from https://www.oxfordlearnersdictionaries.com/definition/english/adjective

Ren, Z., Shen, Q., Diao, X., & Xu, H. (2021). A sentiment-aware deep learning approach for personality detection from text. *Information Processing & Management, 58*(3), Article 102532. https://doi.org/10.1016/j.ipm.2021.102532

Riza, M. A., & Charibaldi, N. (2021). Emotion detection in Twitter social media using long short-term memory (LSTM) and fastText. *International Journal of Artificial Intelligence & Robotics, 3*(1), 15-26. https://doi.org/10.25139/ijair.v3i1.3827

Ryan, G., Katarina, P., & Suhartono, D. (2023). MBTI personality prediction using machine learning and SMOTE for balancing data based on statement sentences. *Information, 14*(4), Article 217.https://doi.org/10.3390/info14040217

Shafi, H., Sikender, A., Jamal, I. M., Ahmad, J., & Aboamer, M. A. (2021). A machine learning approach for personality type identification using MBTI framework. *Journal of Independent Studies and Research Computing, 19*(2). https://doi.org/10.31645/JISRC.43.19.2.2

Stajner, S., & Yenikent, S. (2021). Why is MBTI personality detection from texts a difficult task? In P. Merlo, J. Tiedemann, & R. Tsarfaty (Eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 3580-3589). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.312

Vaddem, N., & Agarwal, P. (2020). Myers-Briggs personality prediction using machine learning techniques. *International Journal of Computer Applications, 175*(23), 41-44. https://doi.org/10.5120/ijca2020920764

Wang, C., Gao, Y., & Xie, Y. (2024a). Analysing the dissemination of the Myers-Briggs Type Indicator (MBTI) in China: A case study of Weibo and Xiaohongshu texts. *EAI Endorsed Transactions on Smart Cities*. https://doi.org/10.4108/eai.15-3-2024.2346532

Wang, Y. (2024b). Logistic regression model for personality type prediction based on the Myers-Briggs Type Indicator. *Transactions on Computer Science and Intelligent Systems Research, 7*, 206-215.https://doi.org/10.62051/4d9gv137

Xue, D., Hong, Z., Guo, S., Gao, L., Wu, L., Zheng, J., & Zhao, N. (2017). Personality recognition on social media with label distribution learning. *IEEE Access, 5*, 13478-13488. https://doi.org/10.1109/ACCESS.2017.2719018

Zhang, H. (2023). MBTI personality prediction based on BERT classification. *Highlights in Science, Engineering and Technology CSIC, 34*, 368-374. https://doi.org/10.54097/hset.v34i.5497