Personality Detection on Cross-Platform Social Media Data

Ibrahim Mazlum
School of Electronic Engineering and Computer Science Queen Mary University of london
London, United Kingdom
i.mazlum@se21.qmul.ac.uk

***Abstract*— Automatic personality detection has been a topic of interest to researchers over the past two decades. In recent years, automatic personality prediction based on social media data has grown in popularity in computational linguistics and NLP communities. Myers-Briggs Type Indicator (MBTI) is popular on social media platforms and users share about their personality types on the online platforms. Therefore, it is possible to prepare self-reported dataset. We collected data from Twitter and created a self-reported dataset for this study. In addition to the Twitter dataset, we used two existing datasets from other social media platforms, and perform pre-processing and feature analysis for cross-platform personality prediction. Furthermore, we have extracted different additional features for each dimension of MBTI to increase performance. As a result of binary classification made with extra features for each dimension, we have achieved average weighted f1 scores range from 67% to 86% for the E/I dimension, 81% to 93% for the N/S dimension, 59% to 79% for the T/F dimension, and 50% to 67% for the J/P dimension. Combining the results from each dimension yielded a f1 score between 20% and 42% for predicting the MBTI personality type.**

Keywords—Natural Language Processing, Personality Detection, Text Classification, Social Media,

I. INTRODUCTION

Individual and consistent distinctions in distinctive patterns of thinking, feeling, and behaving are referred to as "personality" in this context [1]. Personality has been reported to influence language use [2] [3] [4]. As a consequence, language might provide clues about a person's character. Because a person's character is believed to be rather unchanging over lengthy duration of time, the correlation between personality type and linguistic use is likely to be analysed if adequate textual data is available.

On a regular basis, a significant number of people post information about their lives, including their ideas, feelings, and opinions, via a variety of online social media platforms. As discussed in section 2, numerous research has demonstrated that this kind of social media content may let computers make personality predictions. Automated personality profiling offers itself well to the development of applications that are useful in a variety of subjects. Automatic personality detection can serve as the foundation for a variety of applications, including analysing community behaviour [5] and recognising social issues, as well as creating recommendation systems [6] for social media marketing. Aside from business and psychology, health care is another field where personality prediction is valuable. Preotiuc-Pietro et al. (2015) explored the relation between personality types, social media behaviour, and psychiatric problems such as depression and posttraumatic stress disorder [7]. Specific personality characteristics were discovered to be predictor of personality disorder. Mitchell et al. (2015) demonstrate that linguistic characteristics are predictor of schizophrenia [8].

The Big Five traits and the Myers-Briggs Type Indicators (MBTI) are the two most extensively used personality models for making predictions about people's personalities. In The Big Five (Goldberg, 1990) model, personality qualities are classified into five categories: extraversion, agreeableness, conscientiousness, neuroticism, and openness [9]. On the other hand, The Myers-Briggs Type Indicator (MBTI) model (Myers et al, 1990) identifies 16 personality types that are divided into four dimensions: Extraversion-Introversion, iNtuition-Sensing, Feeling-Thinking, and Perceiving-Judging [10]. There has been research to indicate that the Big Five and MBTI models are related despite their differing theoretical foundations [11] [12]. MBTI personality model will be used in this project.

Although natural language processing has made significant advancements in recent years, the task of personality detection based on text remains a problematic one for several reasons. The most major issue is that there are insufficient labelled data on this topic, and a few existing datasets suffer from reliability because almost all of the MBTI datasets are labelled with users' self-reported personality types.

Datasets from the Personality Cafe, Reddit, and Twitter platforms will be used in this study. We created Twitter dataset for this research by using TwitterAPI. The Personality Cafe MBTI dataset taken from Personality Cafe website which is a forum dedicated to discussing personality types. The reddit dataset taken from Gjurkovic and Snajder (2018) [13]. This data was collected from subreddits which are related to personality types.

This project's goal is to develop an automatic personality detection approach that operates effectively across various platforms. We used datasets from three distinct platforms, one of which was the Twitter dataset that we created for this research. First, we experimented optional pre-processing techniques to extract similar features from different platform datasets in order to build a model that can perform well cross platform. Secondly, we extracted additional features from the datasets and ran the model with different combinations of these features to investigate whether they were predictive in each binary dimension of the MBTI model. Finally, we combined the best results from the binary classifiers to predict MBTI personality type and compared the results with the 16-class classifier.

In this project, Certain pre-processing techniques have been shown to be beneficial for a model that will run cross-

platform to tolerate differences in the dataset caused by platform rules and user behaviour. Furthermore, it has been experienced that certain extra features are predictive in the specific binary dimensions of the MBTI model. In addition, it has been shown that personality estimation can yield better results by combining binary classification results in cross platform personality detection.

The remaining parts of the paper are organised as described below. The following section outlines the MBTI personality model and briefly covers related research. Section 3 describes the methodology of research. Section 4 explains experiments and results. Section 5 concludes the paper and Section 6 discusses future studies.

II. Related Work

A. MBTI Personality Model

Currently, the Myers-Briggs Type Indicators (MBTI) is one of the most popular personality models with Big Five. Celli and Lepri (2018) in their study compared both model from the computational viewpoint and proved that modelling MBTI outperformed the modelling Big Five [13].

MBTI creates a binary categorization based on four different dimensions and produces 16 possible personality type depending on the combination of these four values.

- Introversion/Extraversion: The first dimension is connected to the person's energy. Extroverts prefer to put their energy into dealing with people, circumstances, or the outer world. Introverts prefer to focus their energies on dealing with ideas, facts, explanations, or beliefs, or the inner world.
- iNtuition /Sensing: The second dimension is about how information is processed. Sensing is the type of person who wants to look at facts, with what is known. A person with iNtuition tends to experiment with ideas, investigate the unknown.
- Feeling /Thinking: The third dimension is related to decision making. Thinking people make decisions based on objective reasoning and an independent perspective. The type of people who prefer to use values is Feeling.
- Perception /Judgement: The final dimension is related to the way of life chosen. Judgment is the type of person that prefers to have their lives planned and structured. Perception is the type of person that prefers to go with the flow, to be adaptable, and to respond to events as they occur.

Table I shows the 16 personality types available in MBTI model. Each personality type is named with 4 letters and each letter represent one the binary dimensions mentioned above.

TABLE I. MBTI Personality Types

ENFJ	INFJ	INTJ	ENTJ
ENFP	INFP	INTP	ENTP
ESFP	ISFP	ISTP	ESTP
ESFJ	ISFJ	ISTJ	ESTJ

B. Lexical Studies

A significant amount of study has been conducted over the past two decades on detecting personality type from text. The emergence of the Big Five personality model is also based on statistical analysis of language usage [14]. In addition, the relationship between personality type and language use was examined in studies conducted in different fields [15]. The findings of these studies had a positive impact on the automatic detection of personality from text.

Relatively small datasets were use in the early research. Argamon et al. launched the first study for this purpose using essays as a dataset [16]. They built a binary classifier to predict extraversion and neuroticism which are labels in Big Five model. Oberlander and Gill used mails as dataset and n-gram features to predict personality based on Big Five model and [17]. Mairesse et al. used dialogues extracted from the recordings as the dataset and showed in their study that personality can be predicted with better accuracy by using gold labels instead of self-reported ones [18]. They used naïve bayes and support vector machine models with features extracted by LIWC (Linguistic Inquiry and Word Count) and MRC Psycholinguistic database. In addition to these, Iacobelli at al. predict personality type with support vector machine model by extracting bigram features from blog data based on Big Five personality model [19]. Quercia et al. and Golbeck et al. perform a regression analysis on data from Twitter to develop and automatic personality detection model [20] [21].

Kosinki at al. used a large dataset (MyPersonality) for the first time for personality recognition in their project [22]. MyPersonality dataset included posts on Facebook and personality types of millions of users based on Big Five model.

The MBTI personality model has recently gained popularity among Twitter users, and the data gathered from Twitter has been used in academic research. Plank and Hovy created a dataset containing approximately 1.2 million tweets, user's gender and self-reported MBTI personality type and attempted to predict personality type in each dimension of MBTI [23]. They used logistic regression as a model and n-grams, and some metadata as features. Then, Verhoeven at al. used n-grams as a feature on TwiSty dataset to predict MBTI personality. They collected TwiSty dataset which contains user's self-reported MBTI personality type and their posts in six European languages [24]. Yamada at al. used Japanese tweets and they used linguistic features with user related meta features. They stated that user texts have higher correlation with personality type compared to user behaviours [25]. Gjurkovic and Šnajder published a large dataset (MBTI9k) with over 9k users collected from reddit [26]. They trained support vector machine and logistic regression with TF-IDF, LIWC and PSYC features to predict MBTI personality type. This dataset is also one the datasets used in this project.

III. Methodology

I will provide a brief overview of my study's methodology.

A. Datasets

Data from three different social media platforms were used in this research. These platforms are Personality Cafe website, Reddit and Twitter. The personality cafe forum is a forum where discussions on health, behaviour, care, testing and personality types. Personality Cafe data is taken from Kaggle. This data set consists of 8675 users' text postings on the

website and their self-reported MBTI personality types. The average word count of the raw texts in this data is about 1225.

Reddit is one of the world's largest, publicly accessible discussion platforms. Reddit facilitates discussions through subgroups which is called subreddit, each of which focuses on a certain topic. The reddit dataset (MBTI9k) was collected by Gjurkovic and Snajder from subreddits related to personality types [26]. Similar to the personality cafe dataset, the reddit dataset contains the posts shared by 9046 users and their self -reported MBTI personality types. The reddit dataset consists of longer posts compared to other datasets and the average word count is approximately 3500.

The twitter dataset was collected from Twitter by using TwitterAPI for this research. Firstly, some search sentences such as “I am an …”, “My MBTI is …”, “My personality type is …” were determined to label users. Afterwards the three points were filled for all personality types and queries were made on Twitter with TwitterAPI. It was observed that some of the results did not represent the personality type of the person who posted the tweet, since TwitterAPI tokenizes the search phrase and search for words. For example, TwitterAPI returns a tweet shared as “My mother’s personality type is INTJ” for the query “My personality type is INTJ”. Therefore, all tweets are checked manually and unreliable tweets and retweets are not included in the dataset. In the next stage, the tweets were collected by going the timelines of the users who were labelled with their self-reported personality type. Users who share less than 200 words are not included in the dataset. Tweets where users self-report their personality type is removed from the dataset, and the rest of the tweets of that user are used for the prediction. Twitter dataset consists of 7811 different users and average word count is almost 1300. Precision is prioritized at every stage of the data collection process.

The datasets collected from social media are not balanced in terms of the distribution of personality types, because personality types have an effect on the choice of whether to use social media or not. In other words, people with certain personality types might be more likely to use social media. Furthermore, the distribution of personality types across three platforms is very similar. Fig 1. depicts how users are distributed according to their MBTI personality types across all platforms and Fig 2. shows the distribution according to each MBTI dimension.

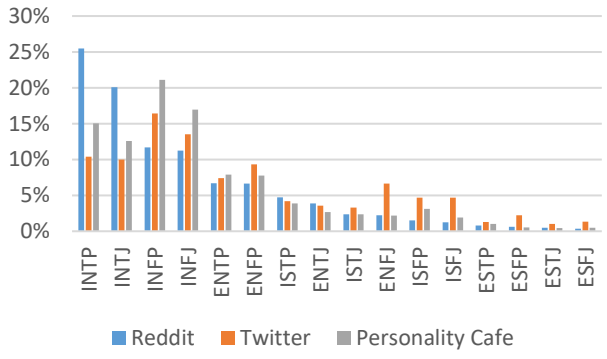

Fig. 1. Distribution of MBTI types across all platforms

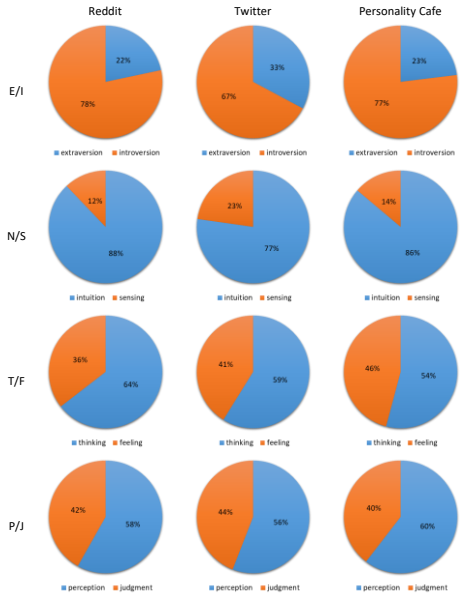

Fig. 2. Distribution of MBTI dimensions across all platforms

B. *Pre-processing Techniques*

In this study, three different social media platform data were used and the personality detection model developed on one platform data were performed and evaluated on the other platforms. The model was experimented by applying specific pre-processing techniques to all three datasets.

Firstly, Langdetect library was used for all datasets and examples with an English score below .999 were removed from the datasets in order to clean datasets. After eliminating non-English samples from the datasets, basic pre-processing techniques considered necessary were applied to all dataset. All words containing non-English characters were deleted from the texts in the datasets. In addition, all symbols sentimentally are not helpful in the text were removed. All personality types mentioned in the texts have been replaced with type-mention token. All texts have been converted to lowercase and unnecessary spaces have been deleted. All URLs were also removed from the datasets, but considering that the URL usage rate might have a relationship with the personality type, the number of URLs for each user saved for further investigation.

Plank and Hovy stated in their study that removing stop words negatively affected model performance [23]. Removing stop words is experimented in this study and similar results were obtained. Therefore, stop words were not removed.

Since this research focuses on finding the best approach for the model trained on dataset from one platform to perform well on other platform datasets, common features must be extracted from all datasets. Therefore, the at signs (@) and hash signs (#) have been deleted. These symbols are valid for only Twitter dataset and the at sign is used before usernames and the hash sign is used before hashtags. Only the symbols were removed from tokens containing these symbols to avoid losing any information, the remainder of the word was preserved.

Differences in platform policies and user actions also lead to differences in platform data. For instance, due to Twitter's character limit, the structure of the texts in the Twitter dataset differs from that of other datasets. Twitter dataset consist of

shorter texts and the use of emoji contractions and acronym words is higher in the Twitter dataset.

Emojis can carry important emotional sentiments [27]. Since personality types effect the emotions of individuals, emotions might be predictive for personality types. Therefore, pre-processing techniques were experimented optionally replaced emojis with their word forms or removed them from text. Additionally, in this study, we extracted extra features for 6 emotions using a specific pre-trained transformer. We took into account that emojis can also affect the performance of this transformer.

Furthermore, extra pre-processing techniques were applied for approximately 1500 acronym words to replace them with their long forms. For example, we replaced the word "idk" with "I do not know". We also applied the pre-processing technique similar to acronym words for almost 100 contractions.

C. Feature Extraction

In this study, Term Frequency-Inverse Document Frequency (TF-IDF) weighed word n-gram features were extracted as the basis to find best approach among the experiments for automatic personality detection. Total number of TF-IDF features limited to 10.000. The objective for utilising TF-IDF frequencies rather than raw token frequencies is to limit the influence of tokens that appear often in articles of all classes and are thus less valuable. The formula for calculating the TF-IDF for a term t in a document d in a group of documents is $TF - IDF(t,d) = TF(t,d)\ x\ IDF(t)$, and IDF is calculated as $IDF(t) = \log(\frac{(1+n)}{(1+DF(t)} + 1)$. DF(t) is document frequency of t and n is total number of documents. TF-IDF transformer was implemented from Sklearn library.

In addition to Term Frequency-Inverse Document Frequency (TF-IDF), extra features have been extracted with using some pre-trained transformers. All user's texts were split into sentences and each sentence was used as the input of relevant model and the outputs were saved as feature. The feature extraction process with transformers was performed on the processed text, which gives the highest accuracy rate with TF-IDF features among the experimented pre-processing techniques.

Distilbert-base-uncased transformer [28] which is finetuned on the emotion dataset employed to extract additional features in consideration of the possibility that user personality types might affect the emotion of the posts on platforms. This is a multi-labelled text classification model takes texts as input and returns the probability scores for each label. There are 6 labels which are sadness, joy, love, anger, fear, surprise. For each emotion, the average of the scores returned by the model were recorded as extra features. Total of 6 attributes have been added to experiment with this model.

We used the sentiment function of the TextBlob library to extract two additional features similar to the ones above. These two features are polarity and subjectivity. We used the average of these two features for each user. Subjectivity measures how much factual and personal idea in represented in the text and polarity measures positivity or negativity of text.

D. Experiment Method

In this project we are researching good approach to develop a model that can perform well across different platform data. Firstly, datasets belonging to 3 different platforms were collected and organized. Considering that using a large amount of data would increase the performance, the number of users in the Twitter dataset with the least number of users was determined as the minimum limit that can be taken from each dataset, and 7800 users were randomly selected from each dataset. This dataset, which comprises 7800 user data from each platform, was used for the study.

After the data set creation process was completed, different models were tried with TF-IDF features and the Linear Support Vector Classifier (LinearSVC) model (with standard parameters from the sklearn library) that showed the best performance in terms of both working time and accuracy was selected as the classification model to be used in the rest of the study. Similar to the support vector machine model, the purpose of the LinearSVC model is to return the optimal hyperplane that categorizes the provided data. However, LinearSVC's kernel allows it to scale a big amount of data.

Afterwards, LinearSVC model was run with TF-IDF features with 16 classes corresponding to 16 personality types. The pre-processing techniques mentioned in the section B were experienced and the best approach for pre-processing was determined. In this process, each platform data was split into training and test data with 80% and 20% ratio respectively. In order to examine the effect of optional preprocessing steps, the model trained on the training data from each platform was tested on a total of 3 test data belonging to all platforms and the accuracy rates were obtained. In summary, at each step, 3x3 matrices were created and examined.

After the preprocessing steps were decided, the extra features explained in the previous section were extracted by using the user texts processed with these techniques. In total, we have extracted 8 additional features, including 6 emotions (sadness, joy, love, anger, fear, surprise) and subjectivity and polarity. We ran the model for each dimension of the MBTI model by combining these features with TF-IDF features in various combinations. In the next section, we report the performance-enhancing features in each dimension.

Finally, we made an estimation of the MBTI personality model by combining the classification results we made in the binary dimension. We compared the performance of 4 binary classifiers with 16 class classifiers. We reported how performance changed for different personality types.

IV. Experiments and Results

Experiments for this study were conducted in three stages. Preprocessing strategies are evaluated on all datasets in the initial step. The evaluation was performed using the 16-class classifier (LinearSVC), and the findings were interpreted based on the classifier's accuracy and average weighted f1 score. Extra features were retrieved using preprocessed texts in the first step, after assessing the findings and deciding on the best approach for pre-processing. The effects of the extra features on the four dimensions of the MBTI personality model were investigated separately in the second stage. Finally, by combining the outputs of the binary classifier, personality prediction was achieved.

A. *Pre-processing Techniques*

First, the required pre-processing methods are carried out. All URLs, meaningless symbols, extraneous spaces, and non-English words and characters have been removed from the texts. The type-mention token has been used to replace the personality types mentioned. The entire text has been converted to lowercase letters. The stop words were not removed. To determine a reference point, the model was run for the texts with these preprocessing techniques performed. The LinearSVC model was ran with 10,000 TF-IDF weighted n-gram features, and 16 personality types were classified. After training on each platform's training data, the model was evaluated on three platforms' test data, and the accuracy rate and average weighted f1 score were reported. Table II shows the 16 class model accuracy rates and average weighted f1 scores for each test step after standard pre-processing.

TABLE II. Model Results with Standard Pre-processing

Train Data	Test Data	Accuracy	F1 Score
Reddit	Reddit	0.2775	0.3401
Reddit	Twitter	0.1564	0.1899
Reddit	Cafe	0.2544	0.2840
Twitter	Reddit	0.1846	0.2365
Twitter	Twitter	0.2333	0.2478
Twitter	Cafe	0.2012	0.2348
Cafe	Reddit	0.2666	0.3215
Cafe	Twitter	0.1346	0.1686
Cafe	Cafe	0.4012	0.4327

It can be clearly seen in Table II that the Personality Cafe dataset provided the best performance with 40% accuracy in 16 label classification. This is followed by Reddit dataset with 27% accuracy and Twitter dataset with 23% accuracy. This conclusion might be explained, as mentioned in the data set section, by the fact that personality cafe is a venue for discussion of personality types, and the data acquired from it might has bias. While collecting the Reddit dataset, the users who reported their personality type were labelled in the MBTI related subreddits. However, the posts of the users in these subreddits were not included in the dataset in order to reduce the bias effect [26]. In addition, since the Twitter dataset is generally collected from the Twitter stream, it can be thought that it does not contain a bias. Following the determination of the basic preprocessing steps, optional preprocessing steps were applied independently. The findings were evaluated by comparing the results of the standard preprocessing phase in Table II.

The emojis in the texts were then replaced with their word representations, and the same model was run again. Table III shows the accuracy rates and the average weighted f1 scores.

TABLE III. Model Results with Standard Pre-processing and Emoji Replacement

Train Data	Test Data	Accuracy	F1 Score
Reddit	Reddit	0.2775	0.3409
Reddit	Twitter	0.1583	0.1911
Reddit	Cafe	0.2564	0.2869
Twitter	Reddit	0.1955	0.2377
Twitter	Twitter	0.2404	0.2490
Twitter	Cafe	0.2006	0.2316
Cafe	Reddit	0.2641	0.3200
Cafe	Twitter	0.1423	0.1721
Cafe	Cafe	0.4026	0.4357

The table above illustrates that replacing emojis with word forms slightly improves accuracy for the majority. Knowing that the emoji usage rate in the Twitter dataset is higher than in other datasets, the increased performance the models using the Twitter dataset proves that the emojis have sentimental meanings that will improve model performance. In addition, emojis can also affect the performance of the model we use to extract the emotional features, which we explain in the next section. Due to these reasons, we chose to keep the word representation in the text instead of completely deleting the emojis.

The second preprocessing technique experienced in this research was to replace acronym words with their long forms. The model was run again by replacing about 1500 distinct acronym words with their long equivalents in all data sets.

Table IV. Model Results with Standard Pre-processing and Acronym Replacement

Train Data	Test Data	Accuracy	F1 Score
Reddit	Reddit	0.2808	0.3458
Reddit	Twitter	0.1596	0.1971
Reddit	Cafe	0.2513	0.2848
Twitter	Reddit	0.1820	0.2254
Twitter	Twitter	0.2365	0.2455
Twitter	Cafe	0.2045	0.2401
Cafe	Reddit	0.2673	0.3244
Cafe	Twitter	0.1442	0.1720
Cafe	Cafe	0.4020	0.4381

Table IV demonstrate, with a few exceptions, that replacing acronym words improves the performance of both cross-platform and on-platform models. The acronym word usage rate differs depending on the nature of the platform. Especially, the Twitter dataset contains a large amount of acronym words. Replacing words with their long form also helps to extract similar features from different platform data. The cross-platform performance improvement can be explained in this way.

Finally, we rerun the model for a preprocessing step that replaces 100 contractions with their long forms, similar to the acronym word replacing. As can be seen from the Table V, changing contractions had a similar effect to changing acronym words. Both preprocessing steps slightly increase model performance.

Table V. Model Results with Standard Pre-processing and Contractions Replacement

Train Data	Test Data	Accuracy	F1 Score
Reddit	Reddit	0.2808	0.3484
Reddit	Twitter	0.1628	0.2000
Reddit	Cafe	0.2538	0.2821
Twitter	Reddit	0.1878	0.2305
Twitter	Twitter	0.2429	0.2515
Twitter	Cafe	0.2077	0.2407
Cafe	Reddit	0.2616	0.3182
Cafe	Twitter	0.1423	0.1735
Cafe	Cafe	0.4083	0.4436

We experienced the optional preprocessing steps separately and saw that they enabled the model to work with

higher accuracy, and we applied these techniques to all the data before the next step.

B. Experiments with Extra Features

In the second stage of the experiments, all of the pre-processing steps were applied to all datasets. Extra features described in feature extraction section were obtained with help of certain transformers by using pre-processed data. Then, it was investigated whether extra features are predictive for each dimension of the MBTI model. The model was run by using the different combinations of extra features together with the TF-IDF features and results were examined.

The first dimension of the MBTI model (Introversion/Extraversion) is about how people manage their energies. Introvert people focus on beliefs and ideas, while extroverts spend their energy with other people. It can be clearly seen from the Fig 2, in all datasets we used, the number of introvert users is much higher than the extravert. Goby stated in his study that there is a strong relationship between first dimension of the MBTI model and people's online preferences [29]. Introvert people prefer to use online platforms more compared to extravert people.

We run LinearSVC model with only TF-IDF features for first dimension of MBTI model. Table VI shows the accuracy rates and f1 scores for the first dimension of MBTI model with only using TF-IDF features.

TABLE VI. Model Results in E/I Dimension with TF-IDF Features

Train Data	Test Data	Accuracy	F1 Score
Reddit	Reddit	0.7576	0.8574
Reddit	Twitter	0.6564	0.7206
Reddit	Cafe	0.7750	0.8281
Twitter	Reddit	0.7512	0.8343
Twitter	Twitter	0.6474	0.6753
Twitter	Cafe	0.7185	0.7481
Cafe	Reddit	0.7551	0.8340
Cafe	Twitter	0.6480	0.7035
Cafe	Cafe	0.7980	0.8342

He and Melo stated in their study that while extroverted people use laughing expressions and positive words more, introverts use uncertain words like probably [30]. We rerun the model by adding the polarity score as a feature which is about positivity or negativity of the text. Table VII shows the results. The polarity feature slightly improves all in-platform performance and many cross-platform performance for E/I dimension.

TABLE VII. Model Results in E/I Dimension with TF-IDF and Polarity Features

Train Data	Test Data	Accuracy	F1 Score
Reddit	Reddit	0.7590	0.8582
Reddit	Twitter	0.6635	0.7326
Reddit	Cafe	0.7686	0.8183
Twitter	Reddit	0.7506	0.8344
Twitter	Twitter	0.6481	0.6761
Twitter	Cafe	0.7186	0.7482
Cafe	Reddit	0.7551	0.8340
Cafe	Twitter	0.6468	0.7049
Cafe	Cafe	0.8000	0.8358

The second dimension of MBTI (iNtuition /Sensing) is about how to take information. Intuitive people focus on ideas, while sensing people focus on reality. Table VIII shows the results for second dimension of MBTI model with only using TF-IDF features.

TABLE VIII. Model Results in N/S Dimension with TF-IDF Features

Train Data	Test Data	Accuracy	F1 Score
Reddit	Reddit	0.8769	0.9344
Reddit	Twitter	0.7730	0.8691
Reddit	Cafe	0.8423	0.8775
Twitter	Reddit	0.8750	0.9303
Twitter	Twitter	0.7602	0.8037
Twitter	Cafe	0.8641	0.9252
Cafe	Reddit	0.8756	0.9312
Cafe	Twitter	0.7666	0.8550
Cafe	Cafe	0.8724	0.9181

Stajner and Yenikent stated that, while sensing individuals build shorter and simple sentences, intuitive people build more complex and longer sentences [31]. We added the average sentence length for each user as an extra feature. This feature was normalized by dividing the overall mean across the platforms to eliminate the difference between platforms. We observed that using only sentence length as a feature did not provide a performance increase in this dimension. An extra metric that can measure sentence complexity might help to predict personality type in this dimension more accurately. In addition, we combined TF-IDF features with joy, surprise and love features and run the model. We reached the results in table below. These features increase scores slightly in some tests.

TABLE IX. Model Results in N/S Dimension with TF-IDF and Additional Features Joy, Surprise, Love

Train Data	Test Data	Accuracy	F1 Score
Reddit	Reddit	0.8769	0.9344
Reddit	Twitter	0.7737	0.8701
Reddit	Cafe	0.8410	0.8735
Twitter	Reddit	0.8756	0.9313
Twitter	Twitter	0.7609	0.8052
Twitter	Cafe	0.8654	0.9260
Cafe	Reddit	0.8756	0.9313
Cafe	Twitter	0.7673	0.8560
Cafe	Cafe	0.8724	0.9181

The third dimension of MBTI (Feeling /Thinking) is about decision making. While thinking individuals give more importance to objective reasons, other types of people care more about feelings. Table X shows the scores for binary classifications for this dimension with only TF-IDF features.

TABLE X.. Accuracy Rates in T/F Dimension with TF-IDF Features

Train Data	Test Data	Accuracy	F1 Score
Reddit	Reddit	0.6923	0.7208
Reddit	Twitter	0.5967	0.6314
Reddit	Cafe	0.7102	0.7136
Twitter	Reddit	0.6435	0.6564
Twitter	Twitter	0.6192	0.6271
Twitter	Cafe	0.6153	0.6228
Cafe	Reddit	0.6929	0.7146
Cafe	Twitter	0.5762	0.5901
Cafe	Cafe	0.7878	0.7878

We have used sadness and surprise as additional features with TF-IDF features and experiment the model. Below table shows the results with these two extra features. model accuracy increased with the effect of extra features.

TABLE XI. Model Results in T/F Dimension with TF-IDF and Additional Features Sadness, Surprise

Train Data	Test Data	Accuracy	F1 Score
Reddit	Reddit	0.6936	0.7210
Reddit	Twitter	0.5987	0.6326
Reddit	Cafe	0.7147	0.7176
Twitter	Reddit	0.6436	0.6570
Twitter	Twitter	0.6199	0.6279
Twitter	Cafe	0.6154	0.6226
Cafe	Reddit	0.6917	0.7134
Cafe	Twitter	0.5769	0.5907
Cafe	Cafe	0.7878	0.7879

The final dimension of MBTI (Perception /Judgement) is about individuals life style. While judging people prefer organized life, perceiving people prefer spontaneity. Table XII below shows results for last dimension only with TF-IDF features.

TABLE XII. Model Results in J/P Dimension with TF-IDF Features

Train Data	Test Data	Accuracy	F1 Score
Reddit	Reddit	0.6121	0.6323
Reddit	Twitter	0.6006	0.6225
Reddit	Cafe	0.6012	0.6337
Twitter	Reddit	0.5025	0.5198
Twitter	Twitter	0.6205	0.6226
Twitter	Cafe	0.4955	0.5007
Cafe	Reddit	0.5935	0.6622
Cafe	Twitter	0.5814	0.6327
Cafe	Cafe	0.6653	0.6743

We experimented the model with different combinations of additional features in this dimension. In J/P dimension of the MBTI model, we observed that the love and subjectivity features increase the model performance. Table XIII shows the model results with additional features.

TABLE XIII. Model Results in J/P Dimension with TF-IDF and Additional Features Love, Subjectivity

Train Data	Test Data	Accuracy	F1 Score
Reddit	Reddit	0.6160	0.6365
Reddit	Twitter	0.6012	0.6135
Reddit	Cafe	0.6032	0.6384
Twitter	Reddit	0.5044	0.5153
Twitter	Twitter	0.6230	0.6251
Twitter	Cafe	0.5019	0.5023
Cafe	Reddit	0.5936	0.6623
Cafe	Twitter	0.5814	0.6327
Cafe	Cafe	0.6654	0.6744

After examining the additional features for each MBTI dimension separately, we found that specific features can be predictive for certain dimension. (Polarity feature in E/I dimension, joy, surprise, love features in N/S dimension, sadness and surprise features in T/F dimension, and love and subjectivity features in J/P dimension)

C. Combining all together

After selecting the pre-processing techniques provide the best performance and the extra features that increase performance in each dimension, we created a new dataset containing 23,000 users, with an equal number of samples from each platform. We split this dataset into training data and test data at 80% and 20%, respectively. First, we ran the 16-class LinearSVC model with TF-IDF and combination of additional features. Then, we run the four binary LinearSVC model for each MBTI dimension using the TF-IDF features and the extra features that are predictive for the corresponding dimension. We made personality prediction by combining the results of binary classifiers and compared the performance with the first scenario. Table XIV shows the average weighted f1 scores for both approaches.

Table XIV. F1 Score Comparison for 16 Class Classifier and 4 Binary Classifier

Train Data	Test Data	16 Class Classifier	4 Binary Classifier
Reddit	Reddit	0.3428	0.3594
Reddit	Twitter	0.2068	0.2108
Reddit	Cafe	0.2997	0.3100
Twitter	Reddit	0.2256	0.2617
Twitter	Twitter	0.2408	0.2340
Twitter	Cafe	0.2448	0.2092
Cafe	Reddit	0.3252	0.3409
Cafe	Twitter	0.1668	0.2119
Cafe	Cafe	0.4436	0.4243

It can be seen from the Table XIV that personality prediction after binary classification gives better results in most of the tests. Especially in cross-platform tests, the scenario of combining binary classifications performs better.

V. Conclusion and Discussion

We defined a new dataset obtained for this study from Twitter and two existing datasets collected from Reddit and Personality Cafe Forum. The datasets in total contain the shares of more than 25,000 users on the relevant platforms and their self-reported MBTI personality types.

This study investigated additional features that could be predictive for each MBTI personality dimension as well as the necessary pre-processing methods for a cross-platform automated MBTI personality detection model. In the first stage of experiments, it has been demonstrated that the data acquired from different platforms has certain variances as a result of platform policies and user behaviour. For example, character limitation on Twitter causes users to use more acronyms, contractions and emojis. We noticed that various pre-processing approaches were required to extract similar features from the datasets for cross-platform study, and we have shown that it improves model performance by experimenting pre-processing techniques such as emoji replacement and acronym word replacement.

In the second part stage of experiments, we examined at each dimension of the MBTI personality model separately. We were able to slightly improve performance in each dimension by applying the additional features. we have achieved average weighted f1 scores range from 67% to 86% for the E/I dimension, 81% to 93% for the N/S dimension, 59% to 79% for the T/F dimension, and 50% to 67% for the J/P dimension.

Finally, we compared the personality prediction approach with the 16-class model and the personality prediction approach by combining the binary classification results. We found that personality estimation using binary classification results provides higher accuracy in cross-platform tests.

VI. Future Work

Unlike other text classification applications in the field of natural language processing, it is not possible to predict personality type with high accuracy from short texts. Because it is not possible for individuals to give enough clues about personality types in short texts.

In this study, we extracted extra features with the help of pre-trained transformers, based on the fact that the objectivity, positivity, negativity and emotions of the individuals might be a reflection of the personality type. These extra features slightly increase the performance of the model for binary dimensions. These results are encouraging for future. A deep learning-based model trained on datasets containing longer texts can perform very well for this purpose, as it can extract more of these extra features in its architecture.

References

[1] Ryne A Sherman, Christopher S Nave, and David C Funder. 2013. Situatonal construal is related to personality and gender. *Journal of Research in Personality*, 47(1):1-14

[2] Andrew Schwartz, Johannes C Eichstaedt, Margeret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Selligman, et al. 2013b. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

[3] Richard Tucker. 1968. Judging personality from language usage: a flipino example. *Philippine Sociological Review*, 16(1/2):30-39

[4] Jocab B Hirsh and Jordan B Peterson. 2009. Personality and language use in self-narratives. *Journal of research in personality*, 43(3):524-527.

[5] Stepen J Guy, Sujeong Kim, Ming C Lin, and Dinesh Manocha. 2011. Simulating heterogeneous crowd behaviors using personality trait theory. In *Proceeding of the 2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA'11),* pages 43-52.

[6] Wen Wu, Li Chen, and Liang He. 2013. Using personality to adjust diversity in recommender systems. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media (HT'13),* pages 225-229.

[7] Daniel Preoutiuc-Pietro, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, Hansen Andrew Schwartz, and Lyle H Ungar. 2015. The role of personality, age and gender in tweeting about mental illnesses. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality,* NAACL.

[8] Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 2015. Quantifying the language of schizophrenia in social media. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Reality.*

[9] Goldberg. L. R. (1990). An alternative "description of personality": the Big-Five factor structure. *Journal of personality and social psychology*, 59(6):1216

[10] Isabel Briggs Myers, Mary H. McCaulley, and Allen L. Hammer. 1990. *Inroduction to Type: A description of the theory and applications of the Myers-Briggs type indicator*. Consulting Psychologist Press

[11] Robert R. McCrae and Paul T. Costa. 1989. Reinterpreting the Myers-Briggs type indicator from the perspective of the five-factor model of personality. *Journal of personality* 57(1):17-40.

[12] Adrian Furnham. 1996. The big five versus the big four:the relationship between the Myers-Briggs Type Indicator (MBTI) and NEO-PI five factor model of personality. *Personality and Individual Differences* 21(2):303-307

[13] Fabio Celli and Bruno Lepri. 2018. Is Big Five Better Than MBTI? A Personality Computing Challenge Using Twitter Data. In *CLiC-it.*

[14] John M. Digman. 1990. Personality structure: Emergence of the five-factor model. *Annual review of psychology* 41(1):417-440.

[15] James W. Pennebaker and Laura A. King. 1999. Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology* 51(1):547-577

[16] Shlomo Argamon, Sushant Dhawle, Mohle Koppel, and James Pennebaker. 2005. Lexical predictors of personality type. In *Proceedimgs of the joint Annual Meeting of the Interface and the Classification Society.* pages 1-16.

[17] Jon Oberlander and Alastair J. Gill. 2006. Language with character: A stratified corpus comparison of individual differences in e-mail communication. *Discourse Processes* 42(3):239-270

[18] François Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research* 30:457-500.

[19] Fracisco Iacobelli, Alastair J, Gill, Scott Nowson, and Jon Oberlander. 2011. Large scale personality classification of bloggers. In *Affective computing and intelligent interaction*, Springer, pages 568-577

[20] Daiele Quercia, Michal Konsinki, David Stillwell, and Jon Crowcroft. 2011. Our Twitter profiles, ourselves: Predicting personality with Twitter. In *Proceeding of 2011 IEEE International Conference on Privacyy, Security, Risk, and Trust (PASSAT) and IEEE International Conference on Social Computing (SocialCom).* pages 180-185.

[21] Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. 2011. Predicting personality from Twitter. In *Proceedings of 2011 IEEE International Conference on PRovacy, Security, Risk and Trust (PASSAT) and IEEE International Conference on Social Computing (SocialCom).* pages 149-156

[22] Michal Kosinki, Sandra C. Matz, Samuel D. Gosling, Vesselin Popov, and David Stillwell. 2015. Facebook as a research tool for the social sciences: Oppornuties, challenges, ethical considerations, and practical guidelines. *American Psychologist* 70(6):543.

[23] Barbara Plank and Dirk Hovy. 2015. Personality traits on twitter—or—how to get 1.500 personality tests in a week. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92-98, Lisbon, Portugal, September. Association for Computational Linguistics.

[24] Ben Verhoeven, Walter Daelemans, and Barbara Plank. 2016. TwiSty: A multilingual Twitter stylometry corpus for gender and personality profiling. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).* pages 1632-1637

[25] Kosuke Yamada, Ryohei Sasano, and Koichi Takeda. 2019. Incorporating textual information on user behavior for personality prediction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 177-182, Florence, Italy, July. Association for Computational Linguistics.

[26] Matej Gjurkovi´c and Jan Snajder. 2018. Reddit: A gold mine for personality prediction. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 87-97, New Orleans, Louisiana, USA. Association for Computational Linguistics.

[27] Felbo B. Mislove, A. Søgaard, A. Rahwan, and Lehmann S. 2017. Using millions of emoji occurrences to learn any-domain represantations for detecting sentiment, emotion and sarcasm. arXiv

[28] Victor Sanh, Lysandre Debut, Julien Chauumond, Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv

[29] Valerie Proscilla Goby. 2006. Personality and online/offline choices: Mbti profiles and favored communication modes in singapore study. *CyberPsychology & Behavior*, 9(1):5-13.

[30] Xiaoli He and Gerard de Melo. 2021. Personality Predictive Lexical Cues and Their Correlations. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 514–523, Held Online. INCOMA Ltd..

[31] Sanja Stajner and Seren Yenikent. 2021. Why Is MBTI Personality Detection from Texts a Difficult Task?. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3580–3589, Online. Association for Computational Linguistic