Proceedings of the of the 17th Conference on Computer Science and Intelligence Systems pp. 235–240
DOI: 10.15439/2022F133
ISSN 2300-5963 ACSIS, Vol. 30

# Personality Prediction from Social Media Posts using Text Embedding and Statistical Features

Seiyu Majima and Konstantin Markov
University of Aizu
Aizuwakamatsu, Fukushima, Japan
Email: m5241125, markov@u-aizu.ac.jp

***Abstract*—Recent advances in deep learning based language models have boosted the performance in many downstream tasks such as sentiment analysis, text summarization, question answering, etc. Personality prediction from text is a relatively new task that has attracted researchers' attention due to the increased interest in personalized services as well as the availability of social media data. In this study, we propose a personality prediction system where text embeddings from large language models such as BERT are combined with multiple statistical features extracted from the input text. For the combination, we use the self-attention mechanism which is a popular choice when several information sources need to be merged together. Our experiments with the Kaggle dataset for MBTI clearly show that adding text statistical features improves the system performance relative to using only BERT embeddings. We also analyze the influence of the personality type words on the overall results.**

## I. Introduction

PERSONALITY research has a long history of studies mainly in psychology where stable patterns of thoughts, feelings, and behaviors have been associated with the so called personality traits. They are useful indicators for describing individual's preferences in perceiving the world and making decisions [1].

Recent advances in natural language processing have made it possible to build machine learning models using online data on human behavior and preferences to automatically predict people's personality traits. Applications include wide variety of internet services including recommender systems [2], product personalization [3], social network and sentiment analysis [4], [5].

In psychological science, there are two widely adopted models for formal description of the personality traits. The five factor model (Big Five) [6] consists of five broad dimensions of personality - Openness, Conscientiousness, Agreeableness, Extraversion, and Neuroticism. Individual's scores on each of these dimensions is obtained using a standardized self-report questionnaires. In the other model, personality is formally described by 16 types known as MBTI (Myers-Briggs Type Indicator) [7]. MBTI is an introspection self-reported diagnostic test aimed at showing psychological preferences about how individuals perceive the world and make decisions. The subjects are classified into 16 personality types that are created from the combination of binary assignments to four dimensions: Introversion versus Extraversion (I/E), Sensing versus Intuiting (S/I), Thinking versus Feeling (T/F), and Judging versus Perceiving (J/P) as shown in Fig. 1. It is been long considered that personality is reflected in individual's use of language [8]. People with high score in extraversion use more positive emotion words while those higher in neuroticism favor first-person words such as "I", "my", and "me".

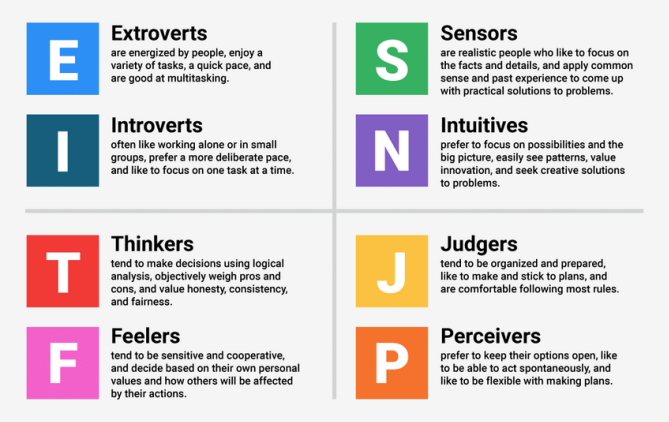


Fig. 1. MBTI type personality keys [7]

A significant number of studies have been dedicated to automatic personality prediction. As an input modality, text data are widely used because they are easy to collect, though video has also been used lately [9]. Some of the first works have focused on text features based on lexicon, syntax, etc., and investigated their correlation with the personality traits as well as their classification performance using shallow machine learning models [10], [11], [12], [13]. Others rely on the commonly used TF-IDF features, for example [14], where personality traits are predicted using an XGBoost classifier.

Some recent works utilize the achievements in the neural networks based text processing by using word embeddings from pre-trained Word2Vec [15] or GloVe [16] models as well as big language models such as BERT [17], [18]. In [17], BERT embeddings are compared with a set of psycho-linguistic features and has been found that the BERT derived features perform better.

In this study, as a starting point we also use BERT derived embeddings like in [17], but then we try to combine them with a set of different statistical features extracted from the input text documents. Those features include uni-gram and bi-gram histograms, topic distribution, post and word length statistics,

etc. In order to combine them with the BERT document embeddings in an efficient way, we use a self-attention based method.

## II. System Description

We assume that the text data consist of multiple posts from various users with known MBTI labels. The system takes all posts from a single user as input data and outputs four dimensional binary vector where each element corresponds to one of the four MBTI axes, i.e. E/I, N/S, T/F and P/J. For example, if the output is $[1, 0, 0, 1]$, the personality type is ESFP. Since there are only 16 possible personality types, we cast the personality prediction task as a classification task with 16 categories which are then projected onto the four MBTI axes.

Input text data from each user are transformed into a document vector using pre-trained BERT language model. In our base system, document vectors from all users are passed to a simple MLP classifier with 16 outputs. In the full system, in addition to the document vector, several statistical features are extracted from the input text data and linearly transformed to match the document vector's dimension. Then, using self-attention mechanism, all vectors are combined into a single final vector which is passed to the the same MLP classifier.

### A. Text Embedding

In natural language processing applications it has become popular to adopt the transfer learning approach where pre-trained language models build from large amounts of data, such as BERT or GPT, are fine tuned or used to extract features from text for further processing by smaller machine learning models.

In this study, we selected the BERT-large model since it provides 1024-dimensional vector representation, i.e. embedding, for each input word token. In order to obtain a single vector for the whole document, we aggregate the BERT outputs by taking their average as was proposed in [17]. Fig. 2 shows the document vector extraction procedure.

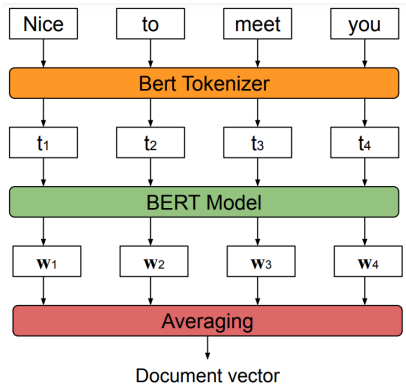


Fig. 2. Document vector extraction procedure.

Another popular way of obtaining representation vector for the whole input is to use the BERT output for the [CLS] token [19]. This, however, works when the number of input tokens is less than maximum input length of the language model which was not the case for the majority of the users data in our experiments.

After document vectors for all the users are obtained, we use a simple MLP network to classify them into one of the 16 MBTI categories. The winning category is then transformed into four dimensional MBTI axes vector. This is our base system and its block diagram is shown in Fig. 3. It is similar to the system investigated in [17], but the main difference is the way we aggregate the BERT outputs.

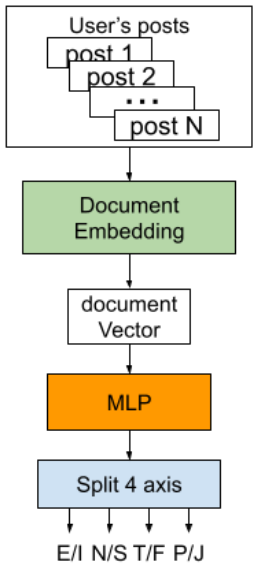


Fig. 3. Base system using only text embedding.

### B. Statistical Features

Although the BERT is a very powerful language model, it is designed to capture and learn dependencies mainly at word and sentence level. When the task is to extract information at a higher user level, as in our case, some useful user dependent characteristics of the input text may get "overlooked" by the BERT model. For example, the usage of some specific words or phrases, or special symbols like emojis is difficult to obtain from the language model. However, it is easy to obtain such information using statistical analysis of users text.

*1) Uni-gram and Bi-gram Histogram:* Different people tend to have their own vocabulary of most used words and phrases and we suppose that the personality plays some role in the formation of this vocabulary. In order to get user specific vocabulary representation we use a histogram of uni-grams (single words) and bi-grams (pairs of words) present in the use data.

First, we create a global vocabulary from all users text data. Then, for each user, the frequency of each uni-gram or bi-gram is obtained from the user's data forming a histogram feature vector. Stop words like "I", "and", "the" which are common and do not provide any discriminating information

are removed from the vocabulary. Rare words, i.e. words with frequency less than a specified threshold, are removed as well.

*2) Topic Distribution:* Another factor that may be influenced by the user's personality is the topic of the user's post. There various ways to determine the topic of a given text. Topic classification into pre-determined categories such as news, politics, sports, etc. has been studied for years [20]. In our case, however, we are more interested in the topic differences among the posts rather than their labels. Furthermore, the granularity level of the topic categories may result in quite different classification results. That is why we adopted an unsupervised topic learning approach.

First, each post from all users is transformed into a vector using the same approach as in our base system, i.e. using BERT language model. Post vectors obtained this way are clustered into several clusters with the K-means algorithm. Then for each user, cluster occupancy histogram of its posts is used as a topic distribution feature vector.

*3) Post and Word Length Statistics:* The number of words in a post can vary significantly depending on various factors one of which we assume is the personality type. If there is any correlation, it can be reviled by taking the first and second order statistics of the word number in a post. Extending this idea to the word length in letters, for each user we construct a feature vector from the mean and variance of word number per post and letter number per word.

*4) Emoticon Usage:* It's a common practice to use emoticons in users posts to express emotions. Some examples of most often used emoticons are given in Table I. We suppose that different people may use different sets of emoticons and the frequency of their usage may be related to personality types. In fact, in [21], emoji embeddings have been concatenated with word embeddings in an attention-based BiLSTM model.

Based on the set of emoticons found in the dataset, for each user we obtain a histogram of emoticons present in its posts which is used as a feature vector.

TABLE I
SOME FREQUENTLY USED EMOTICONS AND THEIR MEANING

| Emoticon | Meaning |
|---|---|
| :) | happy face |
| :D | laughing |
| :'( | crying |
| :-/ | annoyed |

### C. Self-Attention based Embedding and Feature Combination

As we described in Section II-A, our base system uses text embedding to predict user's personality type. By combining the text embedding with the feature vectors obtained from the statistical text analysis we aim at improving the system performance.

There are various way to combine vectors representing different information sources including simple concatenation, weighted sum, etc. In this study, we adopt the weighted sum approach where weights are calculated dynamically based on the current input and a learned stream importance. This approach is implemented using a self-attention network [22].

Given vectors $v_1, v_2, ..., v_N$, we first, pass them through a vector specific linear layer with sigmoid activation function

$$u_i = sig(W_i v_i + b_i) \tag{1}$$

The combined output vector $o$ is then calculated as a weighted sum of vectors $v_i$ as follows

$$o = \sum_i a_i v_i \tag{2}$$

where weights $a_i$ are obtained using softmax function

$$a_i = \frac{\exp u_i}{\sum_j \exp u_j} \tag{3}$$

We have to note that the original text embedding vector and the statistical feature vectors have different sizes. In order to equalize their dimensionality, each statistical feature vector is passed trough a linear layer with a proper weight matrix and no bias.

Fig. 4 shows the block diagram of our system where the output of the self-attention network $o$ is used as input to the personality prediction MLP classifier.

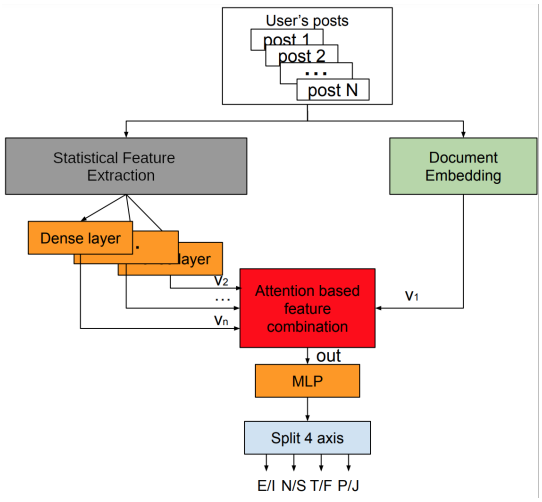


Fig. 4. Self-attention based text embedding and statistical features combination system.

## III. EXPERIMENTS

### A. Dataset

In this study we use the Kaggle MBTI personality type dataset [23]. The data were collected through the Personality-Cafe forum and include diverse selection of posts from people interacting in an informal online social setting.

This dataset contains records of the last 50 posts from 8600 PersonalityCafe users along with their MBTI binary personality type. The MBTI label distribution is given in Table II which shows that the data are quite unbalanced.

TABLE II
DISTRIBUTION OF THE MBTI TYPE LABELS.

| MBTI type label | Number of labels |
|---|---|
| E/I | 1999/6676 |
| N/S | 1197/7478 |
| T/F | 4694/3981 |
| P/J | 5241/3434 |

### B. Data Pre-processing

First, all text data were cleaned which is a standard text pre-processing practice. This includes case conversion, reducing repeated characters and punctuations, expanding contractions, removing numbers, etc. There was a substantial number of URL links which we replaced with a special token [URL]. Posts consisting of URLs only were removed.

### C. Model Training

For model training and evaluation we adopted a 10-fold cross-validation scheme shown in Fig. 5. In each fold, 10% of the data is reserved for testing. The rest is divided into training and validation sets with 9:1 proportion. This way, we train 10 different models and tune their hyper-parameters on the corresponding validation set. After that, each model is evaluated on the fold's test data and the results are averaged over the folds.

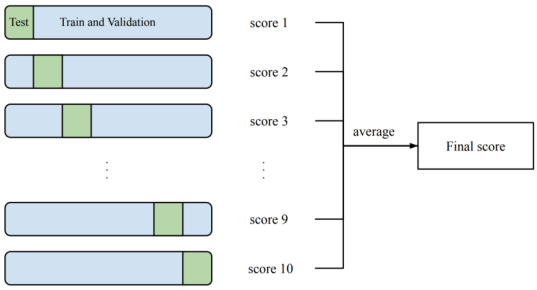


Fig. 5. 10-fold cross-validation scheme for model training and evaluation.

### D. Evaluation Metrics

For classification tasks, the standard evaluation metrics are *Accuracy* and/or *F1-score* and we are reporting all the results in term of those two metrics.

In order to be able to compare our results with those already published, we calculate the *Accuracy* and *F1-score* separately for each of the four MBTI axes, i.e. E/I, N/S, T/F, P/J and for the overall system performance we take their average.

### E. Results

How efficient would be a combination of several feature vectors depends not only on the way they are combined, but to a large extent to how much discriminating information each feature contributes. It is difficult to asses this contribution quantitatively, so we analysed the differences in feature vectors representing different users. For example, in the uni-gram case, these are the word usage histograms. A comparison of such histograms for two users is shown in Fig. 6. It is apparent that they are quite different and could be helpful for the user discrimination.

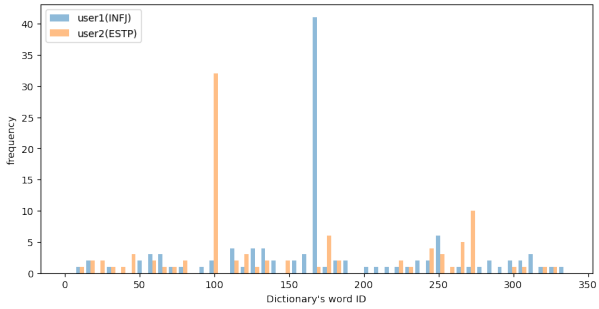


Fig. 6. Uni-gram histogram of two users.

For the topic distribution feature, we needed to tune the number of topic clusters used to create it. This number is a hyper-parameter which can be determined manually since the range of possible values is not that wide. We created topic models with 2 to 16 clusters and evaluated them by combining the topic feature vectors with document vectors from BERT. The obtained average Accuracy and F1-score are given in Fig. 7. It is clear that the optimum number of topic clusters is 4.

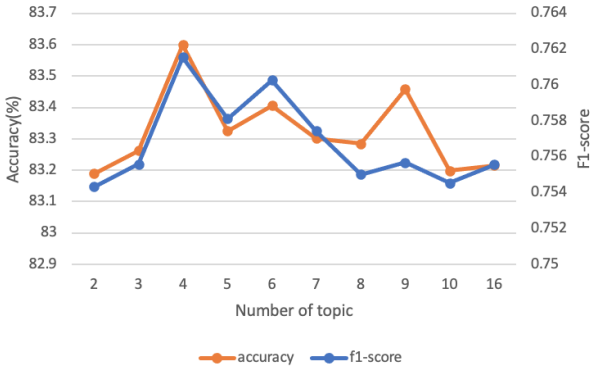


Fig. 7. Classification performance with respect of the number of topic clusters.

The performances of our base system, i.e. using only document embedding as feature vector, and the proposed system with self-attention based combination with statistical features are summarized in Table III and Table IV. The first row in each table shows the base system results. The following rows show the result when the document embedding vectors are combined with one of the uni-gram (uni), bi-gram (bi), topic distribution (t), post and word length (p), and emoticon usage (e) features. The last row shows the performance of the best multiple statistical features combination.

As can be seen from these tables, each additional statistical feature slightly improved the base system results. Uni-gram and bi-gram features scored almost the same which may be explained by the limited vocabulary and the amount of training data per user. The highest result, however, was obtained with multiple features combination.

Finally, in Fig. 8 we compare our best result with the results from some other studies where the same Kaggle MBTI dataset

TABLE III
ACCURACY (%) OF THE TEXT EMBEDDING (BASE) AND ITS COMBINATION WITH THE STATISTICAL FEATURE VECTORS.

| Features | E/I | N/S | T/F | P/J | Ave |
|---|---|---|---|---|---|
| base | 83.1 | 88.5 | 84.1 | 78.0 | 83.4 |
| base+uni | 84.4 | 88.7 | 84.5 | 78.5 | 84.0 |
| base+bi | 84.4 | 88.7 | 84.3 | 78.4 | 84.0 |
| base+t | 84.1 | 88.5 | 84.5 | 77.5 | 83.7 |
| base+p | 84.4 | 88.7 | 84.5 | 78.0 | 83.9 |
| base+e | 84.1 | 88.6 | 84.3 | 77.9 | 83.7 |
| base+uni+p+e | 84.6 | 88.8 | 84.5 | 78.7 | 84.2 |

TABLE IV
F1-SCORE OF THE TEXT EMBEDDING (BASE) AND ITS COMBINATION WITH THE STATISTICAL FEATURE VECTORS.

| Features | E/I | N/S | T/F | P/J | Ave |
|---|---|---|---|---|---|
| base | 0.753 | 0.664 | 0.839 | 0.764 | 0.755 |
| base+uni | 0.758 | 0.691 | 0.844 | 0.772 | 0.767 |
| base+bi | 0.761 | 0.689 | 0.841 | 0.771 | 0.766 |
| base+t | 0.756 | 0.685 | 0.841 | 0.761 | 0.761 |
| base+p | 0.762 | 0.693 | 0.844 | 0.766 | 0.766 |
| base+e | 0.758 | 0.690 | 0.841 | 0.765 | 0.763 |
| base+uni+p+e | 0.762 | 0.695 | 0.844 | 0.773 | 0.769 |

was used. In [14], the TF-IDF features are used with XGBoost classifier while in [17] BERT model is combined with an MLP network.

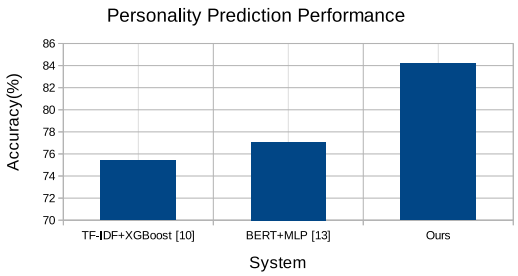


Fig. 8. Comparison of our and other published systems.

## IV. DISCUSSION

It is well known that "models are as good as the data" they are trained on. The data samples and truth labels quality plays essential role in the final systems performance. This is an important issue especially when the data are collected from social media.

During the analysis of the Kaggle MBTI dataset we noticed that all users refer to their own MBTI type more often than to the other personality types. Fig. 9 shows a heatmap of the frequency of the MBTI type word usage by each personality type users. It is clear that user's own MBTI type word is mentioned several times more frequently in their posts. This suggests that a histogram vector of personality type words usage can be highly discriminative for this dataset.

Thus, as an additional experiment, we trained an MLP classifier with only MBTI histogram feature vectors (one per user) as well as their combination with document embedding vectors. The results shown in Table V were surprising, though

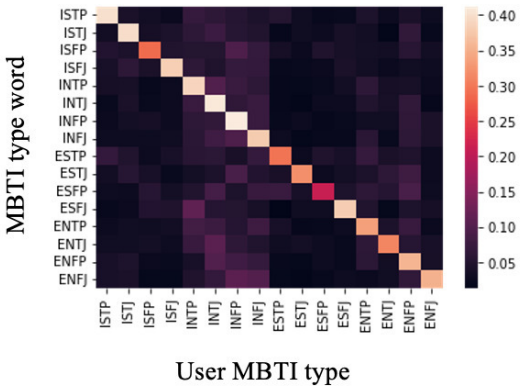


Fig. 9. Heatmap of MBTI type word usage by users of different personality types.

not unexpected. The MBTI feature alone outperformed our

TABLE V
ACCURACY (%) OF THE BASE SYSTEM, MBTI FEATURES AND THEIR COMBINATION.

| Features | E/I | N/S | T/F | P/J | Ave |
|---|---|---|---|---|---|
| base | 83.1 | 88.5 | 84.1 | 78.0 | 83.4 |
| mbti | 86.2 | 90.5 | 84.8 | 80.3 | 85.4 |
| base+mbti | 87.2 | 91.3 | 86.1 | 81.8 | 86.6 |

best system and when combined with the BERT document features improved the result by almost 4%. Of course, this outcome is specific to the Kaggle dataset and will not hold with other data collections. Nevertheless, it underlines the importance of the data when building machine learning systems.

## V. CONCLUSION

In this paper, we proposed a personality type prediction system which combines text document embedding vectors obtained from the BERT language model with statistical feature vectors extracted from the text data. Using powerful language models for downstream tasks has been proved quite effective and our results confirm this conclusion. However, there is always room for improvement when additional task specific knowledge is incorporated in the system as in our usage of statistical text information.

System evaluation with a single dataset reveals only the potential effect and efficiency of the proposed approach and further experimentation with other data collections is necessary in order to prove its merits.

## REFERENCES

[1] S. M. Sarsam, H. Al-Samarraie, and A. I. Alzahrani, "Influence of personality traits on users' viewing behaviour," *Journal of Information Science*, April 2021. [Online]. Available: https://doi.org/10.1177/0165551521998051

[2] S. C. Matz, M. Kosinski, G. Nave, and D. J. Stillwell, "Psychological targeting as an effective approach to digital mass persuasion," *Proceedings of the national academy of sciences*, vol. 114, no. 48, pp. 12 714–12 719, 2017.

[3] S. T. Völkel, R. Schödel, D. Buschek, C. Stachl, Q. Au, B. Bischl, M. Bühner, and H. Hussmann, "Opportunities and challenges of utilizing personality traits for personalization in hci," *Personalized Human-Computer Interaction*, vol. 31, 2019.

[4] J. M. Balmaceda, S. Schiaffino, and D. Godoy, "How do personality traits affect communication among users in online social networks?" *Online Information Review*, 2014.

[5] L. Yue, W. Chen, X. Li, W. Zuo, and M. Yin, "A survey of sentiment analysis in social media," *Knowledge and Information Systems*, vol. 60, no. 2, pp. 617–663, 2019.

[6] O. P. John, L. P. Naumann, and C. J. Soto, "Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues." 2008.

[7] P. D. Tieger, B. Barron, and K. Tieger, *Do what you are: Discover the perfect career for you through the secrets of personality type*. Hachette UK, 2014.

[8] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman *et al.*, "Personality, gender, and age in the language of social media: The open-vocabulary approach," *PloS one*, vol. 8, no. 9, p. e73791, 2013.

[9] C. Suman, S. Saha, A. Gupta, S. K. Pandey, and P. Bhattacharyya, "A multi-modal personality prediction system," *Knowledge-Based Systems*, vol. 236, p. 107715, 2022.

[10] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *Journal of artificial intelligence research*, vol. 30, pp. 457–500, 2007.

[11] G. Carducci, G. Rizzo, D. Monti, E. Palumbo, and M. Morisio, "Twitpersonality: Computing personality traits from tweets using word embeddings and supervised learning," *Information*, vol. 9, no. 5, p. 127, 2018.

[12] K. A. Nisha, U. Kulsum, S. Rahman, M. Hossain, P. Chakraborty, T. Choudhury *et al.*, "A comparative analysis of machine learning approaches in personality prediction using mbti," in *Computational Intelligence in Pattern Recognition*. Springer, 2022, pp. 13–23.

[13] A. Al Marouf, M. K. Hasan, and H. Mahmud, "Comparative analysis of feature selection algorithms for computational personality prediction from social media," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 3, pp. 587–599, 2020.

[14] M. H. Amirhosseini and H. Kazemian, "Machine learning approach to personality type prediction based on the myers–briggs type indicator®," *Multimodal Technologies and Interaction*, vol. 4, no. 1, 2020. [Online]. Available: https://www.mdpi.com/2414-4088/4/1/9

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[16] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[17] Y. Mehta, S. Fatehi, A. Kazameini, C. Stachl, E. Cambria, and S. Eetemadi, "Bottom-up and top-down: Predicting personality with psycholinguistic and language model features," in *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2020, pp. 1184–1189.

[18] H. Jun, L. Peng, J. Changhui, L. Pengzheng, W. Shenke, and Z. Kejia, "Personality classification based on bert model," in *2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT)*. IEEE, 2021, pp. 150–152.

[19] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[20] L. Xia, D. Luo, C. Zhang, and Z. Wu, "A survey of topic models in text classification," in *2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD)*. IEEE, 2019, pp. 244–250.

[21] L. Zhou, Z. Zhang, L. Zhao, and P. Yang, "Attention-based bilstm models for personality recognition from user-generated content," *Information Sciences*, vol. 596, pp. 460–471, 2022.

[22] B. Škrlj, S. Džeroski, N. Lavrač, and M. Petkovič, "Feature importance estimation with self-attention networks," *arXiv preprint arXiv:2002.04464*, 2020.

[23] M. J, "(MBTI) myers-briggs personality type dataset," 2017. [Online]. Available: https://www.kaggle.com/datasnaek/mbti-type