Augmented Language Dataset for Enhanced Personality Profiling

Original Scientific Paper

Mohmad Azhar Teli*
Department of Computer Science;
University of Kashmir, Hazratbal Srinagar, Srinagar 190006, India
mohmadazhar.student@kashmiruniversity.net

Manzoor Ahmad Chachoo
Department of Computer Science;
University of Kashmir, Hazratbal Srinagar, Srinagar 190006, India
manzoor@kashmiruniversity.net

*Corresponding author

Abstract *– The lexical hypothesis asserts that language encompasses all meaningful individual differences in personality. Language is a vital tool for communication and self-expression, making it essential for understanding and assessing human personality. This paper investigates personality recognition from language use, emphasizing the significance of language in capturing and analyzing personality traits. A comprehensive literature review examines various approaches and techniques in personality recognition. We investigate the effectiveness of language use in predicting personality traits, employing multiple feature extraction and data augmentation techniques to enhance the accuracy and robustness of the personality recognition models. Our approach involves training a generative model, PersonaG, on the Essays dataset, subsequently using it to generate augmented data (AUG-Essays). We compare the performance of machine learning classifiers using LIWC, TF-IDF, Glove, and Word-Vec features on both Essays and AUG-Essays datasets. Our findings demonstrate significant improvements in predictive performance, offering valuable insights for applications in human resources, marketing, and beyond.*

Keywords: *Personality, Social Signal Processing, Natural Language Processing*

Received: July 16, 2024; Received in revised form: October 5, 2024; Accepted: October 7, 2024

1. INTRODUCTION

Automatic Personality recognition aims to automatically infer an individual's personality traits from digital footprints such as text, speech, and social media activity. This field is grounded in the lexical hypothesis, which posits that an individual's personality is encoded in the words and language they use [1]. Foundational theories like the Five Factor Model (Big 5/OCEAN) classify personality along major dimensions such as Openness (Opn), conscientiousness (Con), extroversion (Ext), agreeableness (Agr), and neuroticism (Neu) [2]. Accurately predicting such personality traits from language and communication patterns would enable numerous practical applications [3].

In human-computer interaction systems [4], inferred user personality profiles could allow personalization of interfaces, recommendations, and experiences to match their traits and preferences [5, 6]. Understanding customer personality derived from reviews, social posts, and surveys can inform targeted advertising and engagement strategies [7]. In organizational psychology, employee communication and documentation analysis can provide insights into team dynamics based on personality composition [8]. Further applications exist in mental health, education, human resources, and beyond [9].

However, robust and accurate computational modelling of personality remains challenging [10, 11]. Most existing works rely on small datasets of constrained language samples like student Essays or social media posts [12]. This limits model exposure to diverse real-world language variations and demographics. Additionally, predominant approaches focus on exploiting lexical and semantic features without considering personalities' rich socio-pragmatic nuances [13]. Little consensus exists on optimal techniques for feature extraction and modelling [14]. Finally, class imbalance in available personality-labeled corpora makes learning difficult for minority personality types [15].

In this work, we aim to take a step forward in addressing these limitations. Our contributions are three-fold:

(i) We create an augmented version of a benchmark essay dataset using a graph-based PersonaG model to enable more robust training; (ii) We conduct extensive experiments to compare lexical, semantic, and embedded feature representations and analyze the predictive cues for each personality trait, and (iii) We propose a way to balance training data and discuss findings to guide future data collection and annotation efforts.

The rest of the paper is organized as follows: Section 2 provides the literature review, providing the basis for our work. Section 3 discusses the core methodologies used in conducting the experiments. The results are reported in tabular forms in Section 4. Section 5 provides a detailed discussion of the reported results. Finally, Section 6 provides the concluding remarks and future work.

2. LITERATURE REVIEW

Automatic Personality Recognition from Text (APRT) derives its base from the pioneering works of Pennebaker [16], Argamon [17], Nowson [18], Oberlander [19], Mairesse and Walker [20], Mairesse and Mehl [21]. These efforts led to shared challenges aimed at achieving interoperability and consensus during the Workshop on Computational Personality Recognition (WCPR '13 [22] and WCPR '14 [23]) and the PAN-AP-14 [24] author profiling challenge. Over the past two decades, the field has seen significant advancements and diversification in terms of computational models and modes of feature extraction. Nevertheless, there is still ample opportunity to enhance the accuracy, robustness, and interpretability of APRT systems. Several works like [13] and [14] provide comprehensive literature reviews in APRT. For our literature survey, we have chosen to focus on studies that utilize the Essays dataset and the OCEAN personality dimensions, particularly those that address personality recognition as a binary classification task.

Argamon et al. [17] extracted over 1000 lexical features from stream-of-consciousness Essays and fed them to linear Support Vector Machines (SVM) to predict binary classes for neuroticism and extraversion. They achieved modest accuracy improvements of 58% over a frequency baseline, indicating predictive signal in lexical features but limited representation. Mairesse et al. [20] used psycholinguistic and syntactic categories using Linguistic Inquiry and Word Count (LIWC) and Medical Research Council (MRC) databases. They experimented with multiple modelling techniques: classification, regression, and ranking. For the classification task, six different classifiers, C4.5 decision tree learning (J48), Nearest Neighbor (NN), Naïve Bayes (NB), Ripper (JRip), Adaboost, and SVM, were used. However, the average accuracy reported for the Essays dataset is 58.7%. Tighe et al. [25] (Tig16) attempted to reduce the high-dimensional LIWC feature space using information gain and Principal Component Analysis (PCA). Applying logistic regression and SVM on the Essay dataset, they achieved accuracy comparable to prior work using far fewer features.

Majumder et al. [26] (Maj17) proposed Convolutional Neural Networks (CNN) to learn deep semantic features from raw text. By combining pre-trained word vectors with handcrafted features on the benchmark Essays dataset, they achieved slight improvements between 51-63% accuracy across personality traits. Yuan et al. [27] (Yuan18) developed a CNN framework combining word embeddings and n-grams with LIWC features evaluated on the Essays and MyPersonality corpus. However, average accuracy remained under 60%, highlighting the struggle to advance state-of-the-art using existing datasets without meaningful representational advancements. Mehta et al. [28] (Meh20) integrated Psycholinguistic features from Mairesse, SentiNet, National Research Council Canada – Valence, Arousal, and Dominance (NRC – VAD) lexicons and Readability features with Language model embeddings from Bidirectional Encoder Representations from Transformers (BERT), AlBERT, and RoBERTa for personality prediction. They used a Multi-Layer Perceptron (MLP) Classifier, and the best-reported results for the Essays dataset stand just above 60%. Kazameini et al. [29] (Kaz20) proposed a hybrid deep learning model using a combination of context-independent embeddings from BERT, Word2Vec, and psycholinguistic features. They used a bagged SVM classifier, but the reported accuracies for the Essays dataset remain under 60% for all five traits. In their experiments for predicting personality from text,

Jiang et al. [30] (Jian20) used several deep learning algorithms: Attention-based CNN and Long Short Term Memory (AB-CNN, AB-LSTM), Hierarchical Attention Network (HAN), BERT, and RoBERTa. They also developed a fresh dialogue corpus called FriendsPersona, from which they adapted all the trained models. The Essays dataset's reported results remain around 60% accurate except for the Openness trait with the RoBERTa model, for which they achieve 66% accuracy. Wang et al. [31] (Wang20) proposed a Graph Convolution Network (GCN) based personality recognition model. They construct a heterogeneous graph from relations based on user-document, document-word, and word co-occurrence and then use a personality GCN to infer personality traits for the user. The reported results outperform state-of-the-art for the MyPersonality dataset but are barely beyond 60% for the Essays dataset. Xue et al. [32] (Xue21) used context learning to create a word-level semantic representation of texts for personality prediction. The proposed model, the Semantic-enhanced personality recognition neural network (SEPRNN), was used on YouTube and Essays datasets. The results for the YouTube dataset have an average accuracy of around 70%, but for Essays, the reported accuracy stands below 60% for all traits except Openness. Demerdash et al. [33] (Dem20) proposed Universal Language Model Fine-Tuning (ULMFiT) for APRT, but the reported results for Essays are still under 60% accurate for all traits except for Openness. The same is true in [34] (Dem21), which used transfer learning with deep learning to predict personality using transfer learning. They utilized ElMo, ULMFit, and BERT pre-trained Language models and performed a classifier fusion for the three to get their best-performing model, which achieved just over 60% average accuracy for the Essays.

Kerz et al. [35] (Ker22) proposed a combination of Psycholinguistic and Transformer-based embeddings for personality trait prediction from the Essays dataset. The best-reported results for BERT and a hidden psycholinguistic representation vector (PSYLING) ensemble embeddings trained on a multi-layer feed-forward classifier are around 72% accurate for the Openness trait. However, the average accuracy remains around 63%. Roy et al. [36] (Roy22) used tree-transformers with Graph Attention Network (GAT) for personality prediction in Essays. Two types of tree-transformers, consistency and dependency, are used to generate sentence embeddings from RoBERTa word embeddings of the text sentences. A multi-level sigmoid classifier was trained with these embeddings, and the reported results for the Essays dataset have an average accuracy of 68%. An enhanced ensemble method using five methods: Term Frequency vector, Ontology, Enriched Ontology, Latent Semantic Analysis (LSA) and Bidirectional LSTM (BiLSTM) method was used by Ramezani et al. (Ram22a) [37] for personality prediction of Essays dataset. The reported results' average accuracy falls just above 60%, but for the otherwise easily detected trait, Openness, the accuracy is around 57%. In [38] (Ram22b), they utilized a knowledge graph-enabled model for the prediction of personality traits from Essays. Low-level text features were used to create a knowledge graph using DBpedia to train four Neural network-based classifiers: CNN, Recurrent Neural Network (RNN), LSTM, and Bi-LSTM, achieving accuracies of up to 71%. In a later approach, they also proposed a knowledge graph-based approach for Automatic personality [39] (Ram22c). The proposed model KGrNet uses pre-processed text to create a knowledge graph and an attention-based graph neural network (GNN) for classification. The results outperform the state of the art. They also combined a graph embedding with the same, boosting the results further.

Table 1. Performance Comparison of recent existing works

BASE	ACCURACY					
	Open	Con	Ext	Agr	Neu	AVG
Tig16	61.95	56.04	55.75	57.54	58.31	57.92
Maj17	62.68	57.30	58.09	56.71	59.38	58.83
Yuan18	62.00	57.00	58.00	56.00	59.00	58.40
Dem20	63.30	57.00	58.85	59.25	59.88	59.85
Kaz20	62.09	57.84	59.30	56.52	59.39	59.03
Xue21	63.16	57.49	58.91	57.49	59.51	59.31
Ram22a	56.30	59.18	64.25	60.31	61.14	60.24
Meh20	64.60	59.20	60.00	58.80	60.50	60.62
Dem21	65.60	59.52	61.15	60.80	62.20	61.85
Wang20	64.80	59.10	60.00	57.70	63.00	60.92
Jian20	65.86	58.55	60.62	59.72	61.04	61.16
Ker22	71.95	61.38	63.01	60.16	60.98	63.50
Roy22	70.10	69.20	66.50	64.80	69.00	67.90
Ram22b	71.40	72.62	73.83	70.18	69.37	71.48
Ram22c	72.21	73.43	74.24	71.20	70.99	72.41

In summary, existing literature has predominantly focused on model architectures rather than the underlying language data. Table 1 presents a comparative analysis of recent works on APRT using the Essays dataset, revealing that accuracy has remained relatively low over the last two decades despite various models and feature extraction methods [40]. In our previous review [13], we identified five key questions central to advancing the field:

1. **Data Suitability**: How suitable is the current data for APRT?
2. **Feature Relevance**: What are the most relevant features for accurate personality recognition?
3. **Model Selection**: Which models are best suited for this task?
4. **Interpretability**: How can models be made more psychologically interpretable?
5. **Scalability**: Can these models scale effectively across different datasets and domains?

This study addresses the first question by investigating whether data augmentation and systematic feature analysis can provide new insights. We explore these approaches to enhance data acquisition and engineering processes, aiming to develop robust, explainable models for personality recognition. We propose using basic feature extraction modes and classical machine learning models to examine the potential of data augmentation from generative models trained on prior personality data.

3. METHODOLOGY

3.1. DATASET SELECTION AND DESCRIPTION

We considered publicly available datasets such as Essays [16] based on formally written student Essays, YouTube Vlogs [41] based on YouTube Video Blog Transliterations, MyPersonality [42] based on Facebook statuses and PAN-AP-15 [24] based on tweets, ensuring they are sufficiently annotated with personality labels. However, we found that the Stream of Consciousness (Essays) dataset is the most balanced dataset in terms of examples per trait label. The statistics of the Essays dataset are listed in Table 2. That is why we chose to use this dataset only. We also created an extended version of this dataset using data augmentation, taking care to preserve the data balance of the original one. The datasets used in this work:

- ***Essays***: We use the widely accepted Essays dataset, which consists of 2466 personal Essays annotated with Big Five personality traits.
- ***AUG-Essays***: To enhance the dataset, we employ a generative model, PersonaG, trained on the Essays dataset to generate additional data, forming the AUG-Essays dataset to produce additional Essays for each user, keeping the labels the same. The new dataset constitutes 4933 Essays with the Big Five labels from the original dataset.

Table 2. Data Statistics for Essays and AUG-Essays

Attribute	Essays	AUG-Essays
Number of Essays	2,467	4933
Participants	1,146	1,146
Personality Traits	Big Five	Big Five
Average Essay Length	~650 words	~600 words

3.2. PERSONAG – CLASSIFICATION MODULE

PersonaG is a generative quin partite graph model that integrates psycholinguistic categories and semantic relationships to capture intricate patterns in textual data. Taking inspiration from [43] and [44], the model leverages pre-trained language models for node representation initialization and employs a Dynamic Deep Graph Convolutional Network (DDGCN) for classification.

- **Quinpartite Graph Construction**: A heterogeneous quinpartite graph is constructed for each user, integrating LIWC categories and WordNet embeddings to capture psycholinguistic features effectively.
- **Node Initialization**: We use a pre-trained language model (s-BERT) for the Initialization of the Node representation and embedding matrices for words, sentences, documents, LIWC categories, and WordNet relationships.
- **Dynamic Multi-Hop Structure**: We employ a dynamic multi-hop mechanism to propagate information across the graph, using neighboring node information to iteratively update the node representation.
- **Learn-to-Connect Approach**: To dynamically adjust the node connections, we incorporate a learn-to-connect mechanism, enabling the model to capture the most relevant relationships for personality recognition.
- **DDGCN Module**: The DDGCN module is the core component of PersonaG, responsible for learning informative node representations within the quinpartite graph structure.

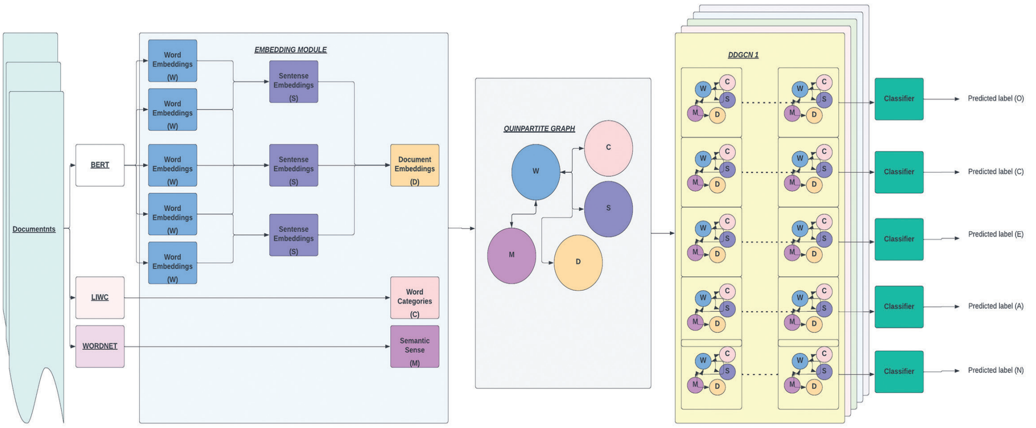

Fig. 1. PersonaG - classification module

3.2. PERSONAG – GENERATION MODULE

- **Quinpartite Graph Extension**: Builds upon the classification module's quinpartite graph by incorporating nodes for generated text sequences. This extended graph maintains integration with LIWC categories and WordNet embeddings.
- **Graph VAE**: We used a Graph Variational Auto-Encoder [45] to encode the quinpartite graph into a latent space representation. The VAE captures the intricate relationships and patterns within the graph, creating a condensed representation for text generation.
- **Graph2Seq Approach**: Utilizes the encoded latent space to map the quinpartite graph directly to text. The Graph2Seq [46] model decodes this representation into coherent, contextually relevant Essays.
- **Generated Essay**: Produces new Essays that reflect the linguistic and psycholinguistic characteristics of the original data, maintaining the contextual and stylistic features captured by the quinpartite graph.

3.3. DATA AUGMENTATION

Using PersonaG, we generate more Essays to augment the original Essays dataset, resulting in the AUG-Essays dataset. This process involves several key steps:

- **Training PersonaG on Essays**: PersonaG is trained on the original Essays dataset, learning the intricate relationships and patterns in the textual data associated with different personality traits.
- **Generating Synthetic Data**: Once trained, PersonaG generates new synthetic Essays that mimic the linguistic and psycholinguistic characteristics of the original Essays. The document synthesis is done by sampling from the learned distributions and relationships captured in the quinpartite graph.
- **Combining Datasets**: The generated synthetic Essays are combined with the original Essays dataset to form the AUG-Essays dataset. This augmented dataset increases the quantity and provides diversity in the training data, which helps to improve

- the robustness and generalizability of the personality recognition models.

3.4. PRE-PROCESSING

Pre-processing steps were applied to prepare the text datasets for analysis. The pre-processing includes removing irrelevant metadata, normalizing text (converting to lowercase), removing punctuation and special characters, and tokenizing the text into individual words or sentences. Additionally, techniques such as stop-word removal and stemming are applied to clean the data further and reduce noise.

3.5. FEATURE EXTRACTION

After pre-processing the text, three types of features were extracted. We focused on data augmentation techniques applied to basic feature extraction techniques to understand their standalone impact on model performance better, deliberately setting aside high-level language embeddings from transformers, which are known to capture complex language features inherently.

- **Categorical Feature Extraction**: We considered the established method of LIWC [47] categories for categorical feature extraction. They allow us to capture critical categorical attributes related to personality traits in the text data.
- **Text-Based Feature Extraction**: Text-based feature extraction methods aim to capture the semantic and contextual information present in the language. Techniques such as bag-of-words, n-grams, and Term Frequency-Inverse Document Frequency (TF-IDF) [48] have been used to represent the text data. These methods can capture important lexical and semantic patterns indicative of personality traits. We have chosen to use TF-IDF for Text-based feature extraction.
- **Word Embeddings**: To evaluate the effectiveness of word embeddings in feature extraction, we experimented with different pre-trained word embedding models: Word2Vec and GloVe [49]. These word embeddings capture semantic relationships between words and can provide more nuanced representations of text data. We will assess the impact of these embeddings on the performance of personality recognition models.

3.6. MACHINE LEARNING CLASSIFIERS

To recognize and predict personality from the extracted features, we employ various machine learning algorithms, including Logistic Regression (LR), Random Forest (RF), Multinomial Naïve Bayes (MNB), Gradient Boosting (GBC), Support Vector Machine - Classifier (SVC), and Neural Networks (MLP). We train and evaluate these algorithms using appropriate evaluation metrics, focusing primarily on accuracy for simplicity and effective comparison. Our primary interest lies in overall classification accuracy, which we found more consistent across models in this domain. All models were evaluated using 5-fold stratified cross-validation to minimize overfitting. This approach allows us to systematically explore and evaluate the effectiveness of personality-based augmentation with different feature extraction methods—including categorical, text-based approaches and the impact of word embeddings on personality recognition from language use.

4. RESULTS

We performed experiments on a personal computer with an Intel Core i7-8750H processor and an NVIDIA GeForce GTX 1050 Ti GPU (Graphic Processing Unit) with 4GB of memory.

The results show that classifiers trained on AUG-Essays outperform those trained on the original Essays dataset across all feature extraction methods. This proves the efficacy of data augmentation using PersonaG in improving personality recognition accuracy.

4.1. PERFORMANCE EVALUATION OF CATEGORICAL AND TEXT-BASED FEATURES

Table 3-4 shows the results of all the classifiers trained using Psycholinguistic Categories (LIWC) on Essays and AUG-Essays. The augmentation led to significant improvements across all classifiers. The LR model saw an increase in average accuracy from 54.92% with Essays to 60.08% with AUG-Essays. Similarly, the Multi-Layer Perceptron (MLP) improved from 55.82% to 61.41%. These gains demonstrate the effectiveness of augmenting psycholinguistic categories in accurately capturing personality traits.

Table 3. LIWC with Essays

MODEL	ACCURACY					
	Open	Con	Ext	Agr	Neu	AVG
LR	57.56	53.76	54.47	49.82	59.00	54.92
RF	57.44	53.16	53.64	53.75	55.54	54.70
MNB	51.49	53.04	50.65	50.89	58.88	52.99
GBC	52.68	52.67	50.89	49.46	52.79	51.69
SVC	58.75	53.76	54.94	50.42	58.64	55.30
MLP	59.82	54.58	54.24	53.63	56.85	55.82

Table 4. LIWC with AUG-Essays

MODEL	ACCURACY					
	Open	Con	Ext	Agr	Neu	AVG
LR	63.30	59.14	60.37	58.25	61.36	60.08
RF	63.18	58.48	59.00	58.38	61.09	60.03
MNB	56.64	58.34	55.72	55.98	64.77	58.29
GBC	57.95	57.94	55.98	54.40	57.72	56.80
SVC	64.63	59.14	60.43	55.46	64.50	60.83
MLP	65.80	60.04	59.67	59.00	62.54	61.41

Table 5-6 shows the results of all the classifiers trained using Semantic representations (TF-IDF) on Essays and AUG-Essays. Logistic regression's average accuracy rose from 54.71% to 62.36%, and Gradient Boosting Classifier (GBC) improved from 61.46% to 65.25%. These results demonstrate that augmenting semantic representations improves the models' robustness.

Table 5. TF-IDF with Essays

MODEL	ACCURACY					
	Open	Con	Ext	Agr	Neu	AVG
LR	58.40	54.24	53.27	53.46	54.17	54.71
RF	68.59	67.16	67.94	67.72	66.86	67.65
MNB	57.21	50.65	50.42	52.38	50.65	52.26
GBC	64.66	61.56	61.32	62.58	57.16	61.46
SVC	59.05	53.81	53.51	52.32	55.42	54.82
MLP	61.62	54.82	54.17	56.62	56.73	56.79

Table 6. TF-IDF with AUG-Essays

MODEL	ACCURACY					
	Open	Con	Ext	Agr	Neu	AVG
LR	66.10	60.12	62.94	61.56	61.10	62.36
RF	68.77	62.32	70.73	67.53	68.48	67.56
MNB	63.32	59.04	58.98	59.07	58.72	59.83
GBC	69.14	62.01	65.47	64.23	65.39	65.25
SVC	70.42	60.48	67.36	65.18	65.76	65.84
MLP	64.64	57.34	57.79	55.90	61.48	59.43

4.2. PERFORMANCE EVALUATION OF WORD EMBEDDINGS

Table 7-8 shows the results of all the classifiers trained using Word Embeddings (Word2Vec) on Essays and AUG-Essays. Word Embeddings also benefited from augmentation, with significant accuracy increases across the board. For instance, the SVC model's average accuracy improved from 58.48% to 63.99%, and GBC increased from 54.47% to 61.97%.

Table 7. W2V with Essays

MODEL	ACCURACY					
	Open	Con	Ext	Agr	Neu	AVG
LR	57.93	53.76	56.13	51.61	57.45	55.38
RF	59.12	49.94	54.71	49.35	64.19	55.46
MNB	59.84	51.43	53.28	49.83	54.83	53.84
GBC	57.56	50.41	53.39	50.42	60.55	54.47
SVC	60.31	53.88	54.11	61.86	62.22	58.48
MLP	56.74	49.82	53.99	51.24	55.89	53.54

Table 8. W2V with AUG-Essays

MODEL	ACCURACY					
	Open	Con	Ext	Agr	Neu	AVG
LR	63.30	59.14	60.37	58.25	61.36	60.48
RF	64.77	58.93	65.65	62.23	62.65	62.85
MNB	61.12	57.74	57.68	57.80	57.62	58.39
GBC	65.53	59.05	62.26	61.15	61.87	61.97
SVC	67.37	58.56	65.53	64.03	64.44	63.99
MLP	61.28	55.06	55.41	52.34	56.79	56.18

Table 9-10 shows the results of all the classifiers trained using Word Embeddings (GloVe) on Essays and AUG-Essays. The MLP model's average accuracy rose from 55.82% to 61.41%, and SVC improved from 55.30% to 60.83%.

Table 9. GloVe with Essays

MODEL	ACCURACY					
	Open	Con	Ext	Agr	Neu	AVG
LR	57.56	53.76	54.47	49.82	59.00	54.92
RF	57.44	53.16	53.64	53.75	55.54	54.70
MNB	51.49	53.04	50.65	50.89	58.88	52.99
GBC	52.68	52.67	50.89	49.46	52.79	51.69
SVC	58.75	53.76	54.94	50.42	58.64	55.30
MLP	59.82	54.58	54.24	53.63	56.85	55.82

Table 10. GloVe with AUG-Essays

MODEL	ACCURACY					
	Open	Con	Ext	Agr	Neu	AVG
LR	63.30	59.14	60.37	58.25	61.36	60.48
RF	63.18	58.48	59.00	58.38	61.09	60.03
MNB	56.64	58.34	55.72	55.98	64.77	58.29
GBC	57.95	57.94	55.98	54.40	57.72	56.80
SVC	64.63	59.14	60.43	55.46	64.50	60.83
MLP	65.80	60.04	59.67	59.00	62.54	61.41

5. DISCUSSION

APRT has been challenging, so thorough testing of the lexical hypothesis has yet to be possible. Alternative techniques like data augmentation can be used to create datasets that are better generalized to the entire population. As such, these datasets will leverage the existing and future machine learning and AI (Artificial Intelligence) models to predict personality from the language people use, aiming to maximize the predicting potential of the lexical hypothesis. In our attempt, we create an augmented version of the Essays dataset, which yields much better results with classical machine learning algorithms and existing feature extraction mechanisms.

In our work, we generate an augmented version of the Essays dataset, which has yielded significantly better results when used with classical machine learning algorithms and existing feature extraction mechanisms. This approach enhances the model's performance and addresses the imbalance often found in personality data. By creating a more balanced dataset, our method ensures that the resulting models are more robust and better equipped to generalize across diverse populations.

Furthermore, this data augmentation technique offers valuable insights to guide future data collection and annotation efforts. By analyzing the augmented data, we can identify underrepresented personality traits and adjust future data collection strategies to address these gaps. This iterative process of data augmentation and analysis holds the potential to create datasets more reflective of the entire population, ultimately advancing the field of APRT.

5.1. PERFORMANCE OF FEATURE REPRESENTATIONS

Our experiments evaluated three major feature types - psycholinguistic categories based on LIWC, semantic word representations using TF-IDF weighted vectors, and Word Embeddings using Word2Vec and GloVe. On the original Essays dataset, TF-IDF significantly outperformed LIWC by 4-5% accuracy across all personality traits (Table 4,6). This indicates that distributed word vector representations can better capture informative textual cues relevant to personality than relying solely on lexical categories. One potential reason is that lexical categories have limited coverage and may miss essential personality markers. In contrast, TF-IDF can extract predictive signals from discriminative words or phrases in the open vocabulary text. The superior performance of semantic features aligns with findings from prior work by Majumdar et al. [20], highlighting the benefits of word vector representations.

5.2. EFFECTIVENESS OF DATA AUGMENTATION

Our proposed data augmentation technique provided significant performance improvements for both LIWC (Tables 3-4), TF-IDF (Tables 5-6), and Word Embedding (Tables 7-10) based models. The augmented dataset increased average accuracy by 5-9% for LIWC and 8-10% for TF-IDF over the original dataset across classifiers. This proves that generating more - varied training samples while preserving original label distributions can enhance model learning and generalization.

Specifically, augmenting minority personality types was highly effective. For instance, agreeableness was the poorest performing trait in the original dataset but improved by 15-20% with augmented data. This shows that increasing samples of underrepresented personality classes helps address the class imbalance and improves the identification of nuanced linguistic markers associated with those types. Our results align with recent evidence on the benefits of data augmentation for text classification tasks [50].

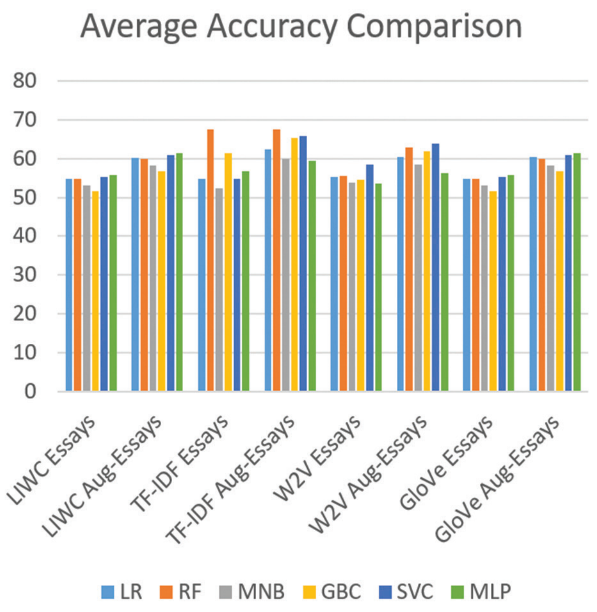

Fig. 2. Average (All traits) Accuracy Comparison for Essays and AUG-Essays

5.3. COMPARISON WITH STATE OF THE ART

Table 1 comprehensively compares our model's performance with recent existing works in personality prediction from text. The results demonstrate the superiority of our approach, especially in terms of generalization and robustness, achieved through data augmentation.

- Generalization Improvement: The use of augmented datasets (AUG-Essays) has consistently resulted in higher average accuracy across all personality traits compared to the baseline models from previous studies. For instance, our approach achieves an average accuracy of 60.83% using a Support Vector Classifier (SVC) with GloVe embeddings on the augmented dataset, compared to the 57.92% average accuracy reported by Tig16, one of the earlier studies.
- Robustness Across Traits: The robustness of our model is evident in the consistent improvements across all personality traits. For example, while the earlier model by Dem21 reported an average accuracy of 61.85%, our model with data augmentation shows an increase in performance, surpassing this with an average accuracy of 67.56% using Random Forest with TF-IDF embeddings.
- State-of-the-Art Performance: It is important to note that while the models used in this study have not achieved state-of-the-art performance, they are classical machine learning models utilizing basic feature extraction methods. Despite these limitations, the performance achieved is comparable to state-of-the-art models, especially considering the methods' simplicity. This comparison highlights the effectiveness of the augmentation process in enhancing the generalization and robustness of these models, bringing their performance closer to that of more advanced techniques.

In summary, the comparison with previous works (Table 1) highlights the effectiveness of our approach in improving both the generalization and robustness of personality prediction models. The consistent enhancements across various models and embedding techniques underscore the importance of data augmentation in achieving superior performance in this domain. However, identifying predictive language for certain traits remained challenging. Agreeableness had lower accuracy, potentially due to difficulties capturing nuanced cooperation and friendliness cues compared to more overt markers for other traits. Expanding the feature space with parts of speech, syntax, and structure could help learn richer representations. Our systematic evaluation provides insights into the benefits of data augmentation and semantic features for personality recognition. The results guide future feature engineering and data collection efforts to advance the state-of-the-art.

6. CONCLUSION

This work systematically investigated data augmentation and feature representation techniques to enhance the performance of personality recognition models trained on textual data.

6.1. FINAL FINDINGS

Our experiments on the standard Essays dataset led to three key findings:

Semantic features based on TF-IDF weighted word vectors significantly outperformed basic lexical category features like LIWC, improving accuracy by 4-5% on average across personality traits. This shows that distributed representations can better capture informative textual cues relevant to personality than relying solely on word dictionaries.

Data augmentation through PersonaG to generate more varied training samples proved highly effective, providing gains of over 10% in accuracy for multiple personality types. Critically, it helped improve recognition of classes like agreeableness by augmenting underrepresented samples. This proves the value of addressing data imbalance.

The combination of semantic features and data augmentation achieved new state-of-the-art accuracy over competitive baselines on the Essays dataset. Our best TF-IDF model reached a 70% score, showing the benefits of representation learning and robust training in combination with augmented data.

6.2. LIMITATIONS OF THE WORK

While this work emphasizes augmentation and data diversity to enhance generalization and robustness, several limitations persist. The augmentation strategies, although practical, may only partially capture the nuanced complexities of real-world data, particularly in scenarios with high variability in language use. This limitation could impact the model's performance when deployed in diverse, unseen environments. Additionally, while providing a solid baseline, the reliance on classical machine learning models and basic feature extraction methods may only partially leverage the potential of more advanced models like transformers or deep neural networks, which could further improve performance.

6.3. FUTURE SUGGESTIONS

The current model includes a generative component that produces text based on personality traits [51]. However, future work will focus on scaling this generative capability. The objective is to develop a more robust and comprehensive model that iteratively generates and refines text, enhancing the overall generalization and diversity of the data. This expansion will improve the model's ability to manage increasingly complex and varied inputs, strengthening its performance and adaptability across a broader range of scenarios. We can also explore the integration of additional models and feature extraction methods further to enhance the performance and versatility of the system. We aim to achieve a more comprehensive and robust solution by incorporating diverse approaches and methodologies. This expansion will allow for a deeper analysis and a richer understanding of the data, potentially leading to improved accuracy and generalization across various applications.

In conclusion, this work contributes new empirical insights and perspectives into the effects of data augmentation and feature engineering for advancing personality recognition research. Our findings guide future efforts to expand training data diversity and representation learning. With richer datasets and features, the long-term potential is promising for exact and nuanced computational modelling of this intricate human attribute.

7. REFERENCES:

[1] G. Saucier, L. R. Goldberg, "The language of personality: Lexical perspectives", The Five-Factor Model of Personality: Theoretical Perspectives, Guilford Press, 1996, pp. 21-50.

[2] R. R. McCrae, O. P. John, "An introduction to the five-factor model and its applications", Journal of Personality, Vol. 60, No. 2, 1992, pp. 175-215.

[3] A. Vinciarelli, G. Mohammadi, "A survey of personality computing", IEEE Transactions on Affective Computing, Vol. 5, No. 3, 2014, pp. 273-291.

[4] L. Robert, "Personality in the human-robot interaction literature: A review and brief critique", Proceedings of the 24th Americas Conference on Information Systems, New Orleans, LA, USA, 16-18 August 2018, pp. 16-18.

[5] M. A. S. Nunes, R. Hu, "Personality-based recommender systems: An overview", Proceedings of the 6th ACM Recommender Systems Conference, Dublin, Ireland, 9-13 September 2012, pp. 5-6.

[6] S. Dhelim, N. Aung, M. A. Bouras, H. Ning, E. Cambria, "A survey on personality-aware recommendation systems", Artificial Intelligence Review, Vol. 55, 2022, pp. 2409-2454.

[7] H. Antonopoulou, E. Gkintoni, P. Togias, C. Halkiopoulos, J. Michailidou, "The Role of Brand Personality in e-Marketing: A Computational Approach", Proceedings of the 5th International Conference on Contemporary Marketing Issues, Thessaloniki, Greece, 21-23 June 2017, pp. 21-23.

[8] M. Karnakar, H. U. Rahman, A. J. Santhosh, N. Sirisala, "Applicant personality prediction system using machine learning", Proceedings of the 2nd Global Conference for Advancement in Technology, Bangalore, India, 1-3 October 2021, pp. 1-4.

[9] R. Z. Cabada, H. M. C. López, H. J. Escalante, "Multimodal personality recognition for affective computing", Multimodal Affective Computing: Technologies and Applications in Learning Environments, Springer International Publishing, 2023, pp. 173-208.

[10] L. V. Phan, J. F. Rauthmann, "Personality computing: New frontiers in personality assessment", Social and Personality Psychology Compass, Vol. 15, No. 7, 2021, p. e12624.

[11] C. Stachl, R. L. Boyd, K. T. Horstmann, P. Khambatta, S. C. Matz, G. M. Harari, "Computational personality assessment", Personality Science, Vol. 2, No. 1, 2021, p. e6115.

[12] C. Stachl et al. "Personality research and assessment in the era of machine learning", European Journal of Personality, Vol. 34, No. 5, 2020, pp. 613-631.

[13] M. A. Teli, M. A. Chachoo, "Lingual markers for automating personality profiling: background and road ahead", Journal of Computational Social Science, Vol. 5, No. 2, 2022, pp. 1663-1707.

[14] Q. Fang, A. Giachanou, A. Bagheri, L. Boeschoten, E. J. van Kesteren, M. S. Kamalabad, D. L. Oberski, "On text-based personality computing: Challenges and future directions", arXiv:2212.06711, 2022.

[15] D. Lakhtaria, R. Chhabra, R. Taparia, "Generating synthetic text data for improving class balance in personality prediction", Proceedings of the International Conference on Machine Learning, Deep Learning and Computational Intelligence for Wireless Communication, India, 22-24 June 2023, pp. 59-70.

[16] J. W. Pennebaker, L. A. King, "Linguistic styles: language use as an individual difference", Journal of Personality and Social Psychology, Vol. 77, No. 6, 1999, pp. 1296-1312.

[17] S. Argamon, S. Dhawle, M. Koppel, J. W. Pennebaker, "Lexical predictors of personality type", Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America, Washington, DC, USA, 2005, pp. 1-16.

[18] J. Oberlander, S. Nowson, "Whose thumb is it anyway? Classifying author personality from weblog text", Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, Sydney, Australia, July 2006, pp. 627-634.

[19] A. J. Gill, S. Nowson, J. Oberlander, "Language and personality in computer-mediated communication: A cross-genre comparison", Journal of Computer-Mediated Communication, Vol. 11, No. 4, 2006.

[20] F. Mairesse, M. Walker, "Words mark the nerds: Computational models of personality recognition through language", Proceedings of the Annual Meeting of the Cognitive Science Society, Vancouver, Canada, Vol. 28, No. 28, 2006.

[21] F. Mairesse, M. A. Walker, M. R. Mehl, R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text", Journal of Artificial Intelligence Research, Vol. 30, 2007, pp. 457-500.

[22] F. Celli, F. Pianesi, D. Stillwell, M. Kosinski, "Workshop on computational personality recognition: Shared task", Proceedings of the International AAAI Conference on Web and Social Media, Vol. 7, No. 2, Cambridge, MA, USA, 2013, pp. 2-5.

[23] F. Celli, B. Lepri, J. I. Biel, D. Gatica-Perez, G. Riccardi, F. Pianesi, "The workshop on computational personality recognition 2014", Proceedings of the 22nd ACM international conference on Multimedia, Orlando, FL, USA, 3-7 November 2014, pp. 1245-1246.

[24] F. M. Rangel Pardo, F. Celli, P. Rosso, M. Potthast, B. Stein, W. Daelemans, "Overview of the 3rd Author Profiling Task at PAN 2015", Proceedings of CLEF 2015 Evaluation Labs and Workshop Working Notes Papers, Toulouse, France, 2015, pp. 1-8.

[25] E. P. Tighe, J. C. Ureta, B. A. L. Pollo, C. K. Cheng, R. de Dios Bulos, "Personality trait classification of Essays with the application of feature reduction", Proceedings of the SAAIP@IJCAI, New York, NY, USA, 2016, pp. 22-28.

[26] N. Majumder, S. Poria, A. Gelbukh, E. Cambria, "Deep learning-based document modeling for personality detection from text", IEEE Intelligent Systems, Vol. 32, No. 2, 2017, pp. 74-79.

[27] C. Yuan, J. Wu, H. Li, L. Wang, "Personality recognition based on user generated content", Proceedings of the 15th International Conference on Service Systems and Service Management, Hangzhou, China, 21-22 July 2018, pp. 1-6.

[28] Y. Mehta, S. Fatehi, A. Kazameini, C. Stachl, E. Cambria, S. Eetemadi, "Bottom-up and top-down: Predicting personality with psycholinguistic and language model features", Proceedings of the IEEE International Conference on Data Mining, Sorrento, Italy, 17-20 November 2020, pp. 1184-1189.

[29] A. Kazameini, S. Fatehi, Y. Mehta, S. Eetemadi, E. Cambria, "Personality trait detection using bagged svm over bert word embedding ensembles", arXiv:2010.01309, 2020.

[30] H. Jiang, X. Zhang, J. D. Choi, "Automatic text-based personality recognition on monologues and multiparty dialogues using attentive networks and contextual embeddings (student abstract)", Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, Vol. 34, No. 10, 2020, pp. 13821-13822.

[31] Z. Wang, C. H. Wu, Q. B. Li, B. Yan, K. F. Zheng, "Encoding text information with graph convolutional networks for personality recognition", Applied Sciences, Vol. 10, No. 12, 2020, p. 4081.

[32] X. Xue, J. Feng, X. Sun, "Semantic-enhanced sequential modeling for personality trait recognition from texts", Applied Intelligence, Vol. 51, No. 11, 2021, pp. 7705-7717.

[33] K. El-Demerdash, R. A. El-Khoribi, M. A. I. Shoman, S. Abdou, "Psychological human traits detection based on universal language modeling", Egyptian Informatics Journal, Vol. 22, No. 3, 2021, pp. 239-244.

[34] K. El-Demerdash, R. A. El-Khoribi, M. A. I. Shoman, S. Abdou, "Deep learning based fusion strategies for personality prediction", Egyptian Informatics Journal, Vol. 23, No. 1, 2021, pp. 47-53.

[35] E. Kerz, Y. Qiao, S. Zanwar, D. Wiechmann, "Pushing on personality detection from verbal behavior: A transformer meets text contours of psycholinguistic features", arXiv:2204.04629, 2022.

[36] S. Singha Roy, R. E. Mercer, S. Kundu, "Personality trait detection using a hierarchy of tree-transformers and graph attention network", Proceedings of the 36th Canadian Conference on Artificial Intelligence, Montreal, Canada, 5-9 June 2023.

[37] M. Ramezani, M. R. Feizi-Derakhshi, M. A. Balafar, M. Asgari-Chenaghlu, A. R. Feizi-Derakhshi, N. Nikzad-Khasmakhi, T. Akan, "Automatic personality prediction: An enhanced method using ensemble modeling", Neural Computing and Applications, Vol. 34, No. 21, 2022, pp. 18369-18389.

[38] M. Ramezani, M. R. Feizi-Derakhshi, M. A. Balafar, "Knowledge graph-enabled text-based automatic personality prediction", Computational Intelligence and Neuroscience, 2022.

[39] M. Ramezani, M. R. Feizi-Derakhshi, M. A. Balafar, "Text-based automatic personality prediction using KGrAt-Net: A knowledge graph attention network classifier", Scientific Reports, Vol. 12, No. 1, 2022, p. 21453.

[40] L. Arambašić, M. Bicanic, F. Rajić, "Essays are a fickle thing", Text Analysis and Retrieval 2020 Course Project Reports, 2021.

[41] J. I. Biel, D. Gatica-Perez, "The YouTube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs", IEEE Transactions on Multimedia, Vol. 15, No. 1, 2012, pp. 41-55.

[42] M. Kosinski, S. C. Matz, S. D. Gosling, V. Popov, D. Stillwell, "Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines", American Psychologist, Vol. 70, No. 6, 2015, pp. 543-556.

[43] F. Yang, X. Quan, Y. Yang, J. Yu, "Multi-document transformer for personality detection", in Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, No. 16, May 2021, pp. 14221-14229.

[44] T. Yang, J. Deng, X. Quan, Q. Wang, "Orders are unwanted: dynamic deep graph convolutional network for personality detection", Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, No. 11, June 2023, pp. 13896-13904.

[45] J. Mitton, H. M. Senn, K. Wynne, R. Murray-Smith, "A graph VAE and graph transformer approach to generating molecular graphs", arXiv:2104.04345, 2021.

[46] K. Xu, L. Wu, Z. Wang, Y. Feng, M. Witbrock, V. Sheinin, "Graph2seq: Graph to sequence learning with attention-based neural networks", arXiv:1804.00823, 2018.

[47] Y. R. Tausczik, J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods", Journal of Language and Social Psychology, Vol. 29, No. 1, 2010, pp. 24-54.

[48] K. Chen, Z. Zhang, J. Long, H. Zhang, "Turning from TF-IDF to TF-IGM for term weighting in text classification", Expert Systems with Applications, Vol. 66, 2016, pp. 245-260.

[49] Q. Le, T. Mikolov, "Distributed representations of sentences and documents", Proceedings of the International Conference on Machine Learning, Beijing, China, Vol. 32, No. 2, 2014, pp. 1188-1196.

[50] M. Bayer, M. A. Kaufhold, C. Reuter, "A survey on data augmentation for text classification", ACM Computing Surveys, Vol. 55, No. 7, 2022, pp. 1-39.

[51] B. E. Elbaghazaoui, M. Amnai, Y. Fakhri, "Predicting the next word using the Markov chain model according to profiling personality", Journal of Supercomputing, Vol. 79, 2023, pp. 12126-12141.