Prediction of Personality Traits Using Social Media Data

Austin T Bomhold
School Of Engineering And Technology
University of Washington, Tacoma, Tacoma, Washington
USA

Johnalbert S Garnica
School Of Engineering And Technology
University of Washington, Tacoma, Tacoma, Washington
USA

Masumi Yano
School Of Engineering And Technology
University of Washington, Tacoma, Tacoma, Washington
USA

Martine De Cock
School Of Engineering And Technology
University of Washington, Tacoma, Tacoma, Washington
USA

Abstract

This study explores the prediction of personality traits, gender, and age using machine learning techniques applied to social media data. By leveraging multimodal data from users social media profiles, including images, text, and relational engagement, we developed and evaluated predictive models for each task. Profile images were processed using ResNet50, textual content was analyzed with advanced natural language processing techniques such as BERT, and relational data was transformed using TF-IDF vectorization. Our findings demonstrate the effectiveness of these models, with gender classification achieving 82 percent accuracy, age prediction reaching 78 percent, and personality trait estimation marginally surpassed baseline metrics with 72 percent. Challenges such as class imbalance, overfitting, and the inherent variability of social media data were addressed with preprocessing techniques and model optimizations. These results underline the potential of machine learning for behavioral analytics, with applications in personalized marketing and social research. Future work will focus on integrating multimodal approaches and addressing ethical considerations in data usag

CCS Concepts

• **Computing methodologies → Neural networks**; **Learning linear models**; **Information extraction**; • **Information systems → Personalization**; • **Human-centered computing** → *Human computer interaction (HCI).*

Keywords

Machine Learning, Social Media Analysis, Personality Prediction, Gender Classification, Age Prediction, Computer Vision, Natural Language Processing, Deep Learning, ResNet50, BERT, RoBERTa, Feature Engineering

ACM Reference Format:
Austin T Bomhold, Johnalbert S Garnica, Masumi Yano, and Martine De Cock. 2024. Prediction of Personality Traits Using Social Media Data. In . ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Machine learning has transformed the way patterns and insights are extracted from the digital footprints of modern human interactions. In this study, we applied machine learning techniques to analyze social media data to predict gender, age, and personality traits based on users' online profiles. Leveraging various frameworks and libraries only recently available through advancements in the machine learning field, each team member focused on a distinct aspect of the data: Masumi analyzed profile pictures to predict gender. Johnny used relational data to perform age prediction as a multiclass classification problem with four categories. Austin utilized textual content to estimate personality scores as a constrained regression problem with values ranging from 1 to 5.

The data was collected by a third-party organization that conducted a survey where users elected to turn over their profile pictures, posts they liked, and the contents of their messages.

To evaluate our approach, we implemented different standard machine-learning methods against an established baseline. Each individual component demonstrated performance improvements over the baseline using feature-specific techniques. Our work explored both theoretical machine learning concepts and their practical applications in real-world scenarios. In this report, we present the methodologies employed, the results achieved, and a comprehensive analysis of the outcomes.

2 Dataset and Metrics

The dataset utilized in this study involved roughly 10000 Facebook users. The pictures were provided in standard JPEG format but included all pictographic content, not simply portraits. The text files contained the aggregate status updates and post details. The text was primarily in English, with a minority in other languages like Spanish, and used locale-specific encoding. Finally, the relational data linking users to the posts they "liked" resulted in a network of content and related user engagement. Additionally, a total of 82 features were extracted from the status updates using a transparent text analysis tool that categorizes words into psychologically meaningful groups called Linguistic Inquiry and Word Count (LIWC)[12].

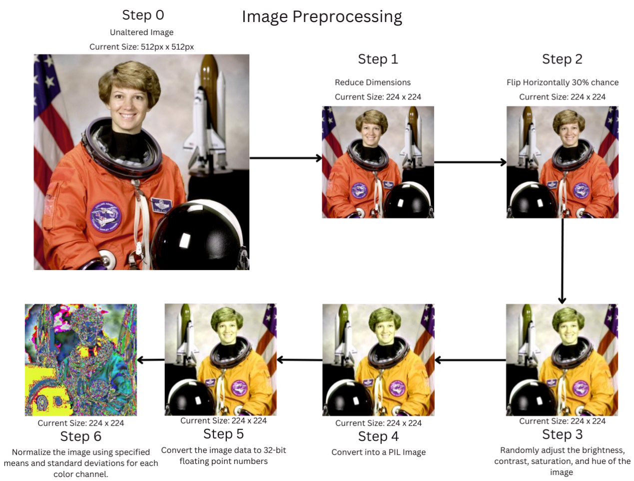

Figure 1: Image Preprocessing

3 Methods

3.1 Data Preprocessing

To achieve accurate model evaluation, a public test set of 334 users was used to conduct a preliminary assessment before a separate dataset, which was held by the University of Washington Tacoma and inaccessible to us during development, was used for the final evaluation. The structure provided a robust foundation for analyzing and predicting gender, age, and personality traits across the various modalities. As mentioned previously, our data was raw real-world data and, as such, required significant preprocessing to address issues that made it unsuitable for direct model training. To ensure quality and reliability for our predictive tasks, preprocessing was performed for each modality.

3.1.1 Images. The images were prepared for model training via several preprocessing steps. First, all images were resized to a uniform dimension of 224 by 224 pixels to ensure consistency, as the original dataset contained images of different sizes. Additionally, resizing the images was necessary to align the input dimensions with the requirements of the pre-trained model used in this study. Next, the pixel values were normalized to the range [-1, 1] in order to fit the parameters and improve the learning process by stabilizing gradient updates. A more even training process ensured faster and more efficient convergence and reduced the risk of instability. Data augmentations were applied using PyTorch's "transforms.ColorJitter" in order to increase reliablity [9]. The augmentations included variations in brightness, contrast, saturation, and hue. Distortions allowed the model to learn more generalized features and reduced the risk of overfitting to specific image characteristics. The preprocessing ensured that the image data was both suitable for training and capable of supporting a generalized model for predictions.

3.1.2 Text. The text data required extensive reworking for analysis. Several techniques were applied to standardize and clean the data. Inconsistencies in encoding were resolved by dynamically decoding files using Python's charset libraries, with any files of indeterminate encoding standardized to an ISO format. URLs within the text were replaced with tokens consisting of their root domains to preserve the contextual information. Emojis were not entirely discarded, but a dictionary was used to map these symbols to corresponding emotional labels such as "happy" or "sad," which were reincorporated into the text to reinforce their usage. Symbols such as extraneous punctuation and characters without semantic values were also removed. Repeated characters in words (e.g., "happpppp-pyyyyyy") were reduced to their standard forms (e.g., "happy"), and redundant whitespaces were eliminated to ensure consistency. Further preprocessing was conducted using natural language library toolkit (NLTK)[2]. The text was trimmed and stop words were removed. The remaining words were lemmatized to their root forms and finally tokenized into a list of small text clusters. These steps collectively ensured that the text data was consistent, meaningful, and suitable. Allowing the text to proceed to feature extraction and subsequent model training.

3.1.3 Likes. Getting the relation data ready required multiple preprocessing steps. Firstly, the post IDs were aggregated and sorted by total interactions. While initially using simple count vectorization, we settled with Term-Frequencey-Inverse-Document-Frequencey (TFiDF). TFiDF adds an extra normalization step that scales the frequency counts with respect to a document's overall usage of a token. Our implementation of TFiDF was from the standard SKLearn [3] library and allowed us to limit the maximum features to the 10K most relevant. Additionally, the count vectorization encoded both individual and consecutive pairs of connections between users and content by way of sklearn's n-gram feature, which was originally designed to link characters in text together. We removed links between users and posts that were in the top 5 percent as the universality did not help to classify between users. Similarly, the bottom 5 percent did not represent a sizeable enough influence given the scarcity. In total, this created a significantly more efficient feature space that maintained the majority of the behavioral patterns displayed by each user.

3.2 Models

Various machine learning models were implemented and evaluated for each task to achieve accurate predictions. These models were carefully selected to align with the characteristics of the data and the specific requirements of predicting gender, age, and personality traits. Below, we detail the models used for each data modality, including their configurations, training methodologies, and performance evaluation metrics.

3.2.1 Image. For the image-based predictions, a pre-trained VGG-16 [6] model was initially chosen as a feature extractor due to its simplicity and reliability. To streamline the training process, the convolutional layers of the network were frozen, as their primary role was to perform automated feature extraction. The actual prediction task was handled by a dense layer. This allows us to only train the predictive component of the model. This approach aimed to reduce computational complexity in order to focus on the classification.

However, this configuration presented several challenges, such as prolonged training times, unstable loss values, large model weight files, and susceptibility to overfitting. While the model achieved over 90 percent accuracy during local training, its performance dropped to 78 percent during the formal evaluation, highlighting issues with generalization.

Overcoming these limitations mandated changing to ResNet50 [11], which offered several advantages over VGG-16. First, ResNet50 uses residual blocks to mitigate gradient vanishing and exploding issues; this enables faster convergence of the loss function and is critical in training machine learning models because it reflects how well the model minimizes errors in predictions over time. A stable and efficiently converging loss function indicates that the optimization process is progressing toward a solution without issues such as oscillations or divergence. Second, it relies on a single dense layer for predictions, meaning that we can reduce the training time. Finally, the weight files for ResNet50 are around five times smaller compared to the VGG-16 model.

With improved storage efficiency and computational manageability, ResNet50 was a more suitable choice for this task. The cross-entropy loss was implemented as the objective function for binary classification, as it is better at penalizing large errors and promoting more viable decision boundaries. The Adam optimizer was chosen for its adaptive learning rate that updates the magnitude of the gradients for each parameter during the training. This ensures faster and more stable convergence while reducing the need for extensive manual tuning. These methodological improvements addressed the limitations encountered with the initial model and enhanced the overall performance of the predictive system

3.2.2 Text. Originally, a series of traditional machine learning techniques were tried while the data was evaluated for cleanliness. Linear regression was the first model to be tried, with the goal of predicting gender. This worked well for gender and when replaced with logistic regression saw accuracy for gender as the text was the most likely to contain the necessary information it was selected to evaluate personality traits. The text started to struggle with surpassing the baseline. The poor performance was expected as these models are known to be subpar for complex and high-dimensional hypothesis spaces. Textual data was the most difficult to process due to the lack of a clear structure and was by far the hardest to analyze.

After moving to traits, a change in the data was needed and decision trees with random forest were tried. Decision trees are a type of supervised learning algorithm that can be used for classification and regression tasks. They work by recursively splitting the data into subsets based on the values of input features, and then making a prediction based on the majority class or average value of the target variable in each subset. A random forest is a collection of decision trees that are trained on different subsets of the data. When used in such clusters, the resulting model is more robust to overfitting and can generalize better.

Additionally, while the text needed the most preprocessing, the dataset included a Linguistic Inquiry and Word Count (LIWC), a list of 82 personality traits, which were used to enhance the text data. LIWC works by scanning text and matching each word against its internal list of traits. The dictionary categories are then assigned to the word based on the most similar traits. LIWC then assigns a score to each word based on the number of traits it matched. The spread of the scores across the traits is then used to remove any LIWC features that have low variation, namely 'Seg'.

In an attempt to improve the model and include the LIWC features, a recurrent neural network was attempted. The model was trained using standard tensorflow dense layers, bidirectional layers with long short term memory, and various normalization methods [1]. The model was trained using the Adam optimizer with a dynamic learning rate and early stopping to prevent overfitting. A pre-trained BERT model was additionally tried as a method for extracting the text features and embedding them into a dense layer. Finally, the outputs of the model were split into the five personality traits.

However, all deployed models suffered from severe overfitting and generalization issues. Multiple factors contributed to this including the lack of a clear structure and high variability that made generalizations hard. Additionally, technical challenges prevented the model from being deployed in a timely manner and led to a lack of feedback about the real-world performance.

3.2.3 Likes. Initial efforts began with a Naive Bayes classifier to predict gender and linear regression for age and personality traits. The Naive Bayes approach, using the Multinomial variant from scikit-learn, achieved an accuracy of 76 percent in gender classification. Features were extracted through a CountVectorizer, analyzing user interactions such as likes. However, age prediction, treated as a regression problem, and personality trait predictions using linear regression yielded suboptimal results, with only minor improvements over baseline performance in terms of root mean squared error RMSE.

Recognizing the limitations of linear regression for capturing the complexity of the relationships within the data, the study incorporated regression models, including Ridge [7], Lasso [13], and ElasticNet [14]. These models introduced regularization penalties to better manage feature selection and multicollinearity. Ridge Regression introduced an L2 penalty to reduce the influence of less significant features while retaining all potentially useful predictors. Lasso Regression employed an L1 penalty to set certain coefficients to zero, effectively performing feature selection and identifying the most predictive likes for personality traits. ElasticNet combined L1 and L2 penalties, balancing feature selection with the capacity to handle correlated features, which was particularly relevant for the dataset's highly interrelated user likes.

Each model's performance was evaluated based on RMSE, with the best-performing model selected for each personality trait. While these models demonstrated significant improvements over linear regression during training, their evaluation performance remained modest, exceeding baseline performance by an average of 0.4 in RMSE.

3.3 Results and Analysis

Average of 10 fold cross validation to estimate how well each model can perform if were to have the proper evaluation.

3.3.1 Age. The age classification task employed the preprocessing and model configuration detailed in. Utilizing a multiclass logistic regression model with a one-vs-rest strategy, the model achieved an average accuracy of 78.17 percent across 10-fold cross-validation, surpassing the established baseline.

Analysis of the confusion matrix reveals that the "xx-24" age group had the highest number of correct predictions, with 441 true labels accurately classified as "xx-24." This indicates that the model effectively captures the distinguishing features associated with this age group. In contrast, the "50-xx" age group exhibited the lowest number of correct predictions, with only 26 true labels correctly identified. This discrepancy likely stems from an imbalanced dataset, where older age groups are underrepresented compared to younger cohorts such as "xx-24." Consequently, the model exhibits limited generalization capabilities for older age ranges.

The confusion matrix also uncovers specific patterns of misclassification. For instance, individuals within the "25-34" age group were frequently predicted as either "xx-24" or "35-49," suggesting a high degree of feature similarity between adjacent age ranges. This overlap is expected, as neighboring age groups often share similar characteristics and interaction patterns on social media platforms. Additionally, the relatively low misclassification rates for distant age groups (e.g., "xx-24" misclassified as "50-xx") indicate that the model is proficient in distinguishing between significantly different age categories.

Overall, the model demonstrates a robust ability to predict age categories with reasonable accuracy, particularly for well-represented groups in the dataset. However, performance declines for less prevalent age groups, highlighting the impact of class imbalance on predictive accuracy. Future improvements could be achieved by addressing this imbalance through techniques such as data augmentation, oversampling of minority classes, or the implementation of cost-sensitive training methods. These strategies would enhance the model's ability to generalize across all age ranges, thereby improving overall classification performance.

The model achieved an accuracy of 78.17 percent in 10 fold cross validation average and surpassed the baseline.

From the confusion matrix, among the "xx-24" age group, 441 of true labels were correctly predicted as the actual (xx-24) class. From this we can see that the model is capturing the features associated with each age group. Conversely, the model performed less accurately for the "50-xx" age group with only 26 correct predictions. This likely reflects an imbalance distribution in the dataset where older age groups are underrepresented compared to younger age groups like "xx-24". The model wasn't able to generalize to the older age ranges. The confusion matrix also reveals a pattern of misclassification. For example, individuals in the "25-34" age group were predicted as "xx-24" or "35-49" the most. This suggests the overlapping feature similarity between the adjacent age ranges. This is expected as neighboring age groups may share similar characteristics. Additionally, the relatively low misclassification rates for distant group (e.g., "xx-24" misclassified as "50-xx") indicates that the model is learning how to distinct features of each age range.

Overall, the model's performance demonstrates that it can successfully predict age categories with reasonable accuracy especially for well represented groups. Further improvement could be achieved by addressing the class imbalance through data augmentations, oversampling, or cost-sensitive training.

3.3.2 Gender. The gender classification model achieved an accuracy of 82 percent on the validation set. Out of 950 validation samples, the model produced 84 false positives (male images misclassified as female) and 86 false negatives (female images misclassified as male). The precision of the model was calculated to be 84.62 percent, indicating that the majority of images predicted as female were correctly classified. Assuming a balanced distribution of male and female samples in the training dataset, these results suggest that the model effectively learned distinguishing features for both classes. The relatively similar numbers of false positives and false negatives reflects a balanced performance across the gender classes. The F1 score, which balances precision and recall, was measured at 84.46 percent, further demonstrating the model's robust performance in classifying both male and female images. This balanced F1 score underscores the model's ability to maintain consistent performance across classes, avoiding significant bias toward either gender.

An analysis of the misclassified images reveals specific patterns contributing to classification errors. Among the false positives, many images contained multiple individuals with at least one female face visible. In such cases, the model may prioritize female features, leading to the incorrect classification of the entire image as female. This suggests a limitation in the feature extraction process, where the presence of female features in a multi-individual context can overshadow male features, resulting in incorrect classification. Conversely, the false negatives predominantly included non-human images, such as cartoon characters, hand-written illustrations, billboards, and other abstract representations. These images lack the distinct human facial features that the model relies on for accurate gender prediction. As a result, the model may default to predicting the more prevalent class in the training data when confronted with ambiguous or non-human imagery. This dependency highlights the model's reliance on specific facial features and its reduced capability to handle diverse or non-standard image types.

Overall, the gender classification model demonstrates strong performance, effectively distinguishing between male and female images with balanced precision and recall. However, the identified patterns in misclassifications indicate areas for potential improvement. Enhancing the feature extraction process to better handle images with multiple individuals and incorporating techniques to manage non-human or abstract representations could further refine the model's accuracy. Additionally, leveraging advanced methodologies, such as convolutional neural networks (CNNs) or incorporating contextual information through large language models (LLMs), may address the current limitations and enhance the model's robustness in diverse scenarios.

3.3.3 Big Five Personality traits. The test classification task aimed to evaluate the capability of our models in using natural language processing to solve prediction challenges. The best performing model was the random Forrest, but even this model could not reliably beat the baseline on all five traits at once.

In final, our model achieved:

- **Openness:** RMSE = 0.6659 ± 0.0226
- **Conscientiousness:** RMSE = 0.7347 ± 0.0176
- **Extroversion:** RMSE = 0.8237 ± 0.0194
- **Agreeableness:** RMSE = 0.6765 ± 0.0167

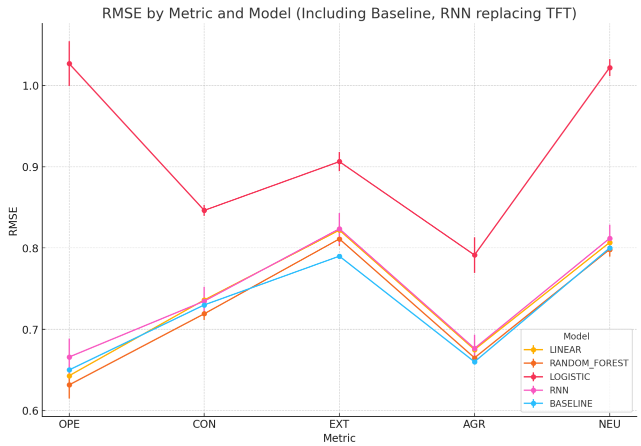

Figure 2: Text Models

- **Neuroticism:** RMSE = 0.8119 ± 0.0169

These scores indicate the model's ability to approximate personality traits, with performance slightly surpassing the baseline RMSE values in the case of most traits.

However, the performance of the model was not consistent across all traits, with some traits showing significantly higher RMSE values than others. Extraversion in particular, which is a trait that is often considered a good indicator of expressiveness, showed the highest RMSE value of 0.83 compared to the other traits.

Openness and Agreeability were predicted with the highest accuracy, with RMSE values of 0.66 and 0.67, respectively. The high dimensionality of the traits, as well as the poor attempt at multi task learning, may have hindered our model's ability.

This suggests that the model, as it stands, will not be able to accurately predict all five traits at once, and the task should have been subdivide among smaller more independent networks.

Despite this, text preprocessing steps, such as lemmatization and emoji mapping, played a critical role in ensuring data consistency and enriching contextual understanding.

Challenges such as subtle personality indicators and domain-specific language may have limited the model's performance for certain traits, particularly Neuroticism and Extroversion.

As the data clearly shows, the model was not able to predict the traits with a high degree of accuracy. While this might seem counter intuitive, given the advantages of deep learning, it is important to remember that such models are extremely sensitive to overfitting. Neural networks generally require precise tuning that made them ineffectual for the given time constraints. Similarly the need to present formal results necessitated testing that em-posed even stricter limitations.

4 Limitations

4.1 Images

While images proved effective for gender classification, their utility for predicting other characteristics, such as age and personality traits, were limited. Gender often has clear visual indicators, age and personality traits are not as easily inferred from facial features alone. Additionally, the nature of social media profile profile pictures introduces further challenges. Users have the freedom to set profile pictures that do not accurately represent their current age or personality. These images may include childhood photos, group photos, or even non-human photos such as cartoons, hand-drawn illustrations, memes, or landscapes. Such variability leads to ambiguity, which affects the model's ability to generalize effectively.

Deep learning models used in this study rely solely on the features present in the image itself without any access to the context behind the image. For example, when the model tries to predict age ranges based on facial features, a childhood photo as a profile picture would likely result in a misclassification into the youngest age group because the model identifies the features in the face without considering the user's actual age. This challenges is consistent with findings from Eidinger et al [4]. They explored age and gender estimation using image-based models. Their study achieved a maximum accuracy of 66.6 with an average of only 60.48 percent without face alignment and 56.7 percent with alignment. Here, the alignment process is designed to standardize the position of facial features, however, the process failed to improve performance significantly.

Similarly, for personality traits, facial features alone often fail to provide meaningful insights. Other features such as contextual cues, place of photo taken, activity depicted in the image could be far more informative. For instance, a photo taken outdoors might convey an impression of extroversion, while an indoor photo might suggest introversion or a sense of coziness. This limitation aligns with findings from the research conducted by Segalin et al [10], where researchers used models like AlexNet and VeryDeep-16 to predict personality traits. Despite using advanced architectures at the time of the research, their results remained modest, with average accuracies around 55.6 percent during training and 53 percent during testing for AlexNet, 56.2 percent during training and 53.4 percent during testing for VertDeep-16. These results suggest that even with advanced methodologies and dedication, predicting personality traits based on images remains challenging and often fails to surpass random guessing.

Addressing these issues may require integrating systems with pre-built contextual understanding, such as Vision-Language Models, which combine visual and textual information to provide more holistic analysis. Additionally, incorporating multi-modal approaches and leveraging metadata, captions, or social media activity patterns could enhance the predictive power of models for these complex traits.

4.2 Text

The text-based personality prediction approach encountered several significant challenges despite implementing increasingly sophisticated models. Initial experiments with linear and logistic regression demonstrated promising results for binary classification tasks like gender prediction, but struggled to surpass baseline performance for multi-dimensional personality trait prediction, aligning with findings from Farnadi et al. [5] regarding the limitations of traditional classifiers for personality prediction. This limitation was expected, as these models typically underperform in complex, high-dimensional hypothesis spaces characteristic of personality prediction.

The transition to more advanced models, including random forests and neural architectures, revealed deeper challenges in the data structure itself. While LIWC features provided 82 psychologically-meaningful dimensions to supplement raw text, similar to the approach used in [5], the inherent variability and lack of clear structure in social media text posed persistent challenges for feature extraction and model generalization. Our implementation of recurrent neural networks with LSTM layers and BERT embeddings, despite their theoretical capacity to capture complex linguistic patterns, suffered from significant overfitting issues.

These generalization problems stemmed from multiple factors. First, the high variability in social media text made it difficult to establish consistent patterns that could reliably indicate personality traits, a challenge also noted by Farnadi et al. [5] in their analysis of Facebook status updates. Second, while LIWC features offered a structured approach to text analysis, the removal of low-variance features like 'Seg' highlighted the sparsity issues in the feature space. Finally, the combination of BERT embeddings with dense layers, though theoretically promising, faced practical challenges in deployment and validation, limiting our ability to assess and improve real-world performance.

These limitations suggest that while modern deep learning architectures can process complex textual data, the fundamental challenges of extracting personality indicators from informal social media text persist regardless of model sophistication. This experience reinforces the need for robust preprocessing strategies and possibly alternative approaches to text representation for personality prediction tasks, extending beyond the traditional feature extraction methods discussed in [5].

4.3 Relation

The limitations of using relational data for personality prediction become evident when compared to direct physiological measurements. A study using electroencephalogram (EEG) achieved R^2 of 0.92 in personality prediction using just a basic two-channel system, compared to clinical systems with 64-128 channels [8]. This success highlights three fundamental challenges with like-based prediction:

First, relational data suffers from severe information loss. While EEG directly captures neural activity, social media likes compress rich human behavior into binary interactions. Even preprocessing techniques like TFiDFVectorizer, which attempts to recover information by weighting rare likes more heavily, cannot reconstruct lost contextual information about user motivation and engagement depth.

Second, the structural limitations of like data prevent capturing personality's multi-dimensional nature. The EEG study succeeded by combining ResNet for spatial patterns and LSTM for temporal sequences, matching personality's inherent complexity. In contrast, our relational data's flat binary structure (presence/absence of likes) eliminates interaction nuance. Despite implementing increasingly sophisticated regression models (linear, ridge, lasso, ElasticNet), we cannot manufacture these missing dimensions - analogous to reconstructing a 3D object from its shadow.

Finally, while EEG provides clean physiological signals in a controlled environment, like data represents indirect behavioral signals distorted by multiple factors: algorithmic content presentation, social conformity pressures, temporal trends, and platform dynamics. These distortions limit model performance regardless of sophistication. This suggests that the challenge lies not in technical implementation but in the fundamental nature of relational data itself, indicating future work should focus on integrating multiple data sources rather than more complex modeling approaches.

5 Conclusion and Future Work

This study employed machine learning techniques to analyze social media data, demonstrating their effectiveness in predicting gender, age, and personality traits based on users' digital profiles. The proposed approach outperformed the baseline in the evaluation, underscoring the potential of machine learning in multi-modal analysis. By allocating resources across team members to handle images, text, and relational data, theoretical concepts were effectively applied in practical contexts, allowing for meaningful analysis of the results. In a 10-fold cross-validation setup, the image-based model achieved an accuracy of 82 percent, the age prediction model attained 78 percent accuracy, and the personality traits model obtained an RMSE of .

Despite encountering technical challenges related to virtual machine configuration and limitations in resource connectivity, the team demonstrated individual expertise and effective collaboration. Although the final model could not proceed to formal evaluation, the comprehensive report and documentation reflect the significant effort and dedication invested in the project. The analysis highlights the potential of this work as a foundational tool for applications in marketing and behavioral analysis.

Future work could focus on increasing the accuracy of each model by utilizing more advanced architectures, such as transformer-based models, or by incorporating additional modalities, such as user interaction data or temporal patterns. Additionally, extending this research to address ethical considerations, including privacy-preserving machine learning and algorithmic fairness, is essential to ensure the responsible deployment of such systems. Building upon the solid foundation established in this project, further research can tackle more complex machine learning challenges, apply the findings to real-world problems, and continue advancing the field of predictive analytics in social media.

Acknowledgments

To Martine, for the dataset and great real world problem.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/ Software available from tensorflow.org.

[2] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

[3] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. 108–122.

[4] Eran Eidinger, Roee Enbar, and Tal Hassner. 2014. Age and Gender Estimation of Unfiltered Faces. *IEEE Transactions on Information Forensics and Security* 9, 12 (2014), 2170–2179. https://doi.org/10.1109/TIFS.2014.2359646

[5] Golnoosh Farnadi, Susana Zoghbi, Marie-Francine Moens, and Martine De Cock. 2013. Recognising Personality Traits Using Facebook Status Updates. In *Workshop on Computational Personality Recognition (Shared Task)* (Boston, MA, USA) *(AAAI Technical Report WS-13-01)*. Association for the Advancement of Artificial Intelligence, 14–18. https://www.aaai.org/ocs/index.php/WS/AAAIW13/paper/view/7047

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *arXiv preprint* abs/1512.03385 (2015). https://doi.org/10.48550/arXiv.1512.03385

[7] Arthur E. Hoerl and Robert W. Kennard. 2000. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* 42, 1 (2000), 80–86. https://doi.org/10.2307/1271436

[8] Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private Traits and Attributes are Predictable from Digital Records of Human Behavior. *Proceedings of the National Academy of Sciences* 110, 15 (April 2013), 5802–5805. https://doi.org/10.1073/pnas.1218772110

[9] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 8024–8035.

[10] Cristina Segalin, Dong Seon Cheng, and Marco Cristani. 2017. Social profiling through image understanding: Personality inference using convolutional neural networks. *Computer Vision and Image Understanding* 156 (2017), 34–50. https://doi.org/10.1016/j.cviu.2016.10.013

[11] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint* abs/1409.1556 (2015). https://doi.org/10.48550/arXiv.1409.1556

[12] Yla Tausczik and James Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology* 29 (March 2010), 24–54. https://doi.org/10.1177/0261927X09351676

[13] Robert Tibshirani. 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 58, 1 (1996), 267–288. https://doi.org/10.2307/2346178

[14] Hui Zou and Trevor Hastie. 2005. Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67, 2 (April 2005), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x