Traitlytics: Analysing Personality from LinkedIn Profile Data

Akshara Balasubramanian
ASU id: 1236173941
abalas47@asu.edu

Aman Pandey
ASU id: 1233641203
apand105@asu.edu

Jaydeep Patil
ASU id: 1229592431
jtpatil@asu.edu

Reuben Roy Kochukudiyil
ASU id: 1233723597
rkochuk1@asu.edu

ABSTRACT

In a number of domains, such as behavioural analysis, professional growth, and hiring, personality evaluation is essential. Conventional techniques for evaluating personality depend on self-reported questionnaires, which might be biased and have scalability problems. The growing popularity of professional networking sites such as LinkedIn provide an opportunity to investigate data-driven methods for predicting personality. This study looks on predicting personality qualities from LinkedIn profile data using machine learning and large language models (LLMs). Our goal is to provide a scalable and impartial approach to personality evaluation by examining textual data from profile summaries, endorsements, and activity patterns. This study examines different modelling techniques, assesses how well they work, and talks about possible practical uses for automated personality prediction.

So far, we have conducted a literature review, finalized the dataset sources, and started preprocessing LinkedIn-related textual data. Additionally, initial experimentation with traditional ML models has been performed using baseline features. Moving forward, we aim to refine feature extraction, incorporate network-based attributes, and fine-tune transformer-based models for improved accuracy. The final evaluation will benchmark our approach against existing research to ensure scalability and reliability in real-world applications.

INTRODUCTION

Problem statement:

Behavioural analysis, professional growth, and recruitment all make extensive use of personality assessments. Self-reported questionnaires, like the Big Five Personality Test, are the mainstay of traditional personality evaluation techniques. But these approaches lack scalability, are subjective, and are prone to bias. The emergence of online professional networks such as LinkedIn presents a chance to investigate different, data-based methods for predicting personality traits of different users.

The majority of personality prediction research to date has been on social networking sites like Facebook and Twitter, which offer unstructured, informal data. Nonetheless, LinkedIn offers organized professional data, such as endorsements, activity patterns, and profile summaries, which may provide more in-depth understanding of a person's professional characteristics. This project aims to address the limitations of traditional self-assessments by leveraging LinkedIn profile data to predict personality traits using machine learning (ML) and large language models (LLMs).

Project Objectives:

The primary objective of this project is to develop an automated personality prediction system using LinkedIn profile data. Specifically, we aim to:

1. **Extract and analyze textual data** from LinkedIn profiles, including summaries, skills, and endorsements, to predict the Big Five Personality Traits (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism).
2. **Compare traditional machine learning models** (e.g., Logistic Regression, Random Forest) with transformer-based models (e.g., BERT) to determine the most effective approach for personality classification.
3. **Explore a hybrid approach** by incorporating both textual features and network-based features (e.g., connections, activity patterns) to improve prediction accuracy.
4. **Evaluate model performance** using standard classification metrics such as accuracy, precision, recall, F1-score, and AUC-ROC, and analyze the semantic alignment between predicted and self-reported traits using **cosine similarity**.
5. **Ensure ethical considerations** in data collection and processing, following privacy guidelines while leveraging publicly available LinkedIn data for research purposes.

RELATED WORK

Research on predicting personalities from digital footprints has expanded recently, especially as social media platforms have grown in popularity. These platforms offer enormous volumes of personal information that can be examined to deduce different facets of an individual's conduct, inclinations, and character attributes. The information found in user profiles, such as activity patterns, endorsements, and text descriptions, offers a special chance to forecast personality traits in a work environment in the context of professional networks like LinkedIn.

While substantial work has been done in using social media data to predict personality, there are fewer studies which have focused on professional networks like LinkedIn, which provide a distinct set of cues and behaviors. LinkedIn data is often more structured and professional, potentially offering different linguistic patterns and personality indicators compared to more casual platforms. Moreover, few studies have explored the combination of textual features with network-based features, such as connections and activity patterns, to improve prediction accuracy.

We have referred few technical papers and summarized the methods used and addressed some gaps which are relevant to our chosen topic.

Paper 1: Niels van de Ven, Aniek Bogaert, Alec Serlie, and Mark J. Brandt. 2017. Personality perception based on LinkedIn profiles. *Journal of Managerial Psychology*, 32(2), 92-104. https://doi.org/10.1108/JMP-07-2016-0220

This paper investigated whether LinkedIn users' personality traits could be inferred accurately based on their profiles. Their study examined how observers formed personality perceptions based on profile elements like photos, endorsements, and textual descriptions. The findings suggest that while profile elements strongly influence impressions, they do not necessarily align with self-reported personality traits. A limitation of this study was the reliance on psychology students as raters, which may have introduced biases in personality assessment. Furthermore, the study did not explore machine learning-based approaches for personality prediction, which is a gap we address in our project.

Paper 2: Christian, H., Suhartono, D., Chowanda, A., & Zamli, K. Z. (2021). Text-based personality prediction from multiple social media data sources using pre-trained language model and model averaging. *Journal of Big Data*, 8(1), 1-14. https://doi.org/10.1186/s40537-021-00459-1

This paper leveraged pre-trained transformer-based models, such as BERT, to predict personality traits from text data extracted from multiple social media platforms. They demonstrated that transformer models outperform traditional machine learning models when analyzing textual data. However, their study focused on casual social media platforms like Twitter and Facebook rather than professional networks like LinkedIn. Additionally, they did not explore the combination of textual and network-based features, which is an area our project expands upon.

Paper 3: Fernandez, S., Stöcklin, M., Terrier, L., & Kim, S. (2021). Using available signals on LinkedIn for personality assessment. *Journal of Research in Personality, 93*, 104122.

This paper extended this work by systematically assessing personality traits using LinkedIn profile indicators. Their study applied regression and classification techniques to analyze 607 profiles based on 33 personality-related attributes. They found that LinkedIn profiles provide valid signals for traits like extraversion, conscientiousness, and openness, but neuroticism was difficult to infer. This study highlights the potential of structured LinkedIn features for personality assessment. However, it does not explore deep learning models or NLP-based approaches, which we incorporate in our methodology.

Paper 4: Dai, K., González Nespereira, C., Fernández Vilas, A., & Díaz Redondo, R. P. (2015). Scraping and clustering techniques for the characterization of LinkedIn profiles. *ArXiv preprint arXiv:1505.00989*

This paper applied NLP and clustering techniques to analyze 5.7 million LinkedIn profiles, uncovering trends in career paths and educational backgrounds. Their work focused on structuring LinkedIn data rather than personality prediction. While their approach effectively groups users based on career-related attributes, it does not integrate personality assessment models, leaving a gap in the direct application of machine learning for personality prediction.

Paper 5: Kashkin, V., & Paliy, V. (2024). Automated LinkedIn Analysis to Determine Psychometric Characteristics of a Client. Asian Social Science, 20(2), 35-48

This paper explored the application of AI-based psychometric tools, such as Crystal Personality Insights and Humantic AI, to analyze LinkedIn profiles. They evaluated how these tools use natural language processing (NLP) and machine learning to extract psychometric characteristics. While this study highlights the potential of AI in personality assessment, it primarily evaluates pre-built tools rather than developing new models. Additionally, the study does not address explainability in personality prediction, which we aim to tackle through SHAP and LIME in our project

Paper 6: Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., et al. (2013). "Personality, gender, and age in the language of social media: The Open-Vocabulary approach." PLOS ONE, 8(9), e73791.

This paper applied an open-vocabulary approach to analyze 700 million words from Facebook messages of 75,000 participants who also completed personality tests. Their study revealed correlations between linguistic patterns and personality traits. For

example, neurotic individuals frequently used words like "sick of" and "depressed," while conscientious people used words related to family and work. Although this study pioneered the application of NLP in personality prediction, it was limited to Facebook and did not explore modern deep learning approaches, which our project integrates.

METHODOLOGY

In this section we will describe our approach to predicting personality traits from LinkedIn profile data using machine learning (ML) and large language models (LLMs). Our methodology involves data collection, preprocessing, model selection, training, and evaluation.

Data Collection:

For our project, we are utilizing two primary sources of data: the **Big Five Personality Dataset** and publicly available LinkedIn profile data (either through datasets or web scraping, adhering to ethical guidelines).

The **Big Five Personality Dataset** will serve as a foundation for training and validating our personality prediction model. This dataset, available on Kaggle, contains over 1 million instances with 110 features, each corresponding to user personality traits. It is structured around self-reported responses to the Big Five personality test, with ratings from 1 to 5 for each personality trait (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism). The dataset includes both the **questions** related to each trait and the **responses** that allow us to label data with known personality traits.

Key features of the dataset include:

- Personality traits: Ratings for each of the Big Five traits (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism).
- User responses: 10 questions for each personality trait.
- Class distribution: The dataset contains a balanced set of responses, making it suitable for model training**.**

Data from LinkedIn profiles on the other hand, will be gathered using web scraping methods while respecting privacy and ethical standards in order to create a solid dataset for personality research. This includes metadata (such relationships and skills) as well as textual content (including job descriptions, endorsements, and summaries). The collection will also be supplemented by publicly accessible datasets pertaining to social media personality analysis.

For our initial research purpose, we used publicly available datasets which has LinkedIn profile data. Another alternative dataset which we found online for training purpose could be obtained from: https://github.com/jkwieser/personality-prediction-from-text

For Web scraping of LinkedIn profiles, we used Selenium browser. We wrote a python script to help automate the process of scraping data from the LinkedIn profiles relevant to our project. This data was successfully transferred to an Excel sheet which helped in organizing this data. We later on merged the scraped data with a public dataset after preprocessing the data.

As we don't have a dataset with text as well as labels assigned so we used **cosine similarity**, to assign the labels to the extract text. Using cosine similarity, we assigned each text a label of 1-5 for each of the 5 traits (1 being the lowest and 5 being the highest value).

Data Preprocessing:

The collected data will undergo preprocessing to ensure consistency and usability. This includes:

- **Text Cleaning:** Removing unnecessary characters, special symbols, and redundant spaces.
- **Tokenization & Lemmatization:** Breaking down text into meaningful units and standardizing word forms.
- **Feature Extraction:** Transforming raw textual data into structured representations using TF-IDF and word embeddings.
- **Handling Missing Data:** Addressing incomplete or inconsistent profile
- **Vectorization**: Vectorizing the data using **TF-IDF** Vectorization

We will perform Personality trait annotation which is the process of assigning the Big Five personality traits (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) to individual data samples based on user profile data.

Steps involved:

1. **Data Annotation:**

- We need to annotate each LinkedIn profile based on the personality traits of the user. For this, we use a **supervised learning** approach where each profile is labelled with its corresponding personality traits.
- We use the **Big Five Personality Traits dataset** (e.g., Kaggle dataset) that includes text samples and predefined labels for traits. These labels serve as our **ground truth** for training the model.

2. **Annotation Process**:

- Profile Descriptions: Text from user descriptions, such as summaries, posts, endorsements, etc., will be used to infer personality traits.
- Profile Labels: Each user profile will have tags or annotations indicating their personality traits.

- Manual Annotation: In some cases, we might need to manually label a small subset of to ensure the accuracy of the labels in real-world applications, especially for profiles that lack explicit labels.

3. **Trait Mapping**:

- Each profile will be mapped to one or more personality traits based on **textual content analysis** (e.g., NLP methods such as sentiment analysis, entity recognition, and text classification).

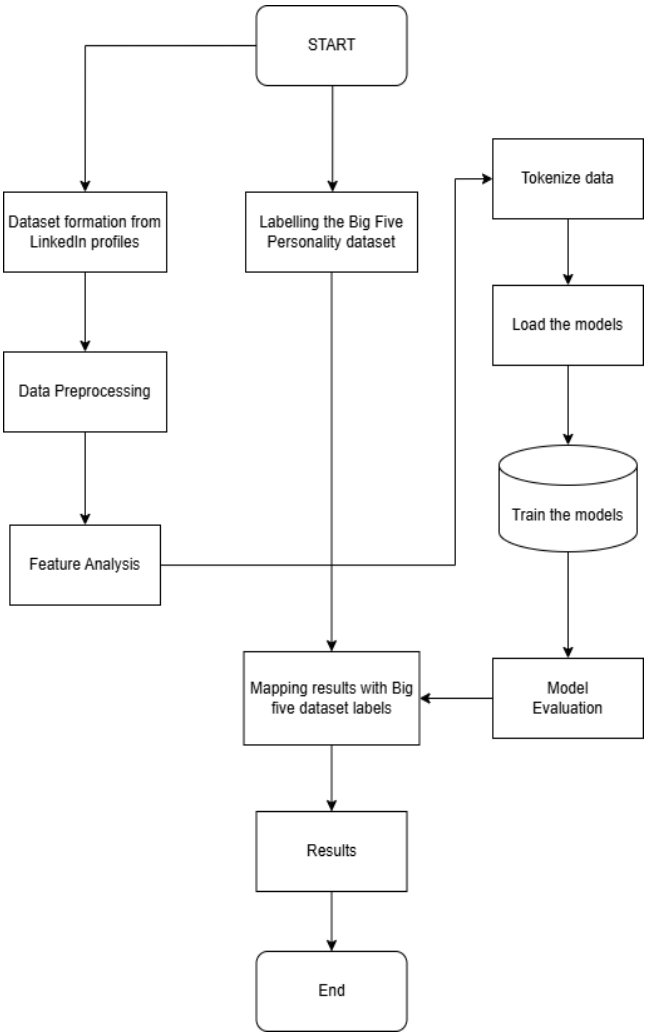

Figure 1: Flowchart of the system

4. Vectorization:

We used Term Frequency-Inverse Document Frequency (TF-IDF) vectorization to handle the textual data from the Big Five trait sentences and LinkedIn profiles. By assessing a word's significance to a document within a collection, this method transforms unprocessed text into numerical feature vectors. Words that appear frequently in a document but seldom across the entire corpus receive higher weights. This helps capture the distinctive and informative words that describe a user's profile or a trait description. The generated TF-IDF vectors reflect both profile texts and personality trait phrases in the same high-dimensional space, giving a consistent basis for comparison.

5. Assigning Labels:

After vectorizing the trait-defining sentences and the LinkedIn profile texts, we calculated the degree to which a profile matched each personality feature using cosine similarity. Cosine similarity calculates the cosine of the angle between two vectors in a multi-dimensional space, producing a similarity score between 0 (totally dissimilar) and 1 (same direction). We calculated the average cosine similarity between ten typical sentence vectors that corresponded to each personality feature and a user's profile vector. This successfully annotated the user's personality based on the content of their profile by producing a continuous score for each trait.

Methods used:

In the initial phase of the project, we planned to use classification models such as **Support Vector Machines (SVM)**, **Random Forest Classifier**, and fine-tuned **BERT-based language models** to predict the Big Five personality traits from LinkedIn profile data. The goal was to categorize profiles into specific trait classes (e.g., high, medium, low) based on textual patterns.
However, upon closer inspection of the Big Five Personality Dataset, we found that the trait labels were continuous numeric scores (decimal values ranging from 0 to 1) rather than categorical labels. Therefore, a regression-based approach was more appropriate than classification. As a result, we adapted our methodology accordingly and used the following machine learning models for our project:

- **Logistic Regression:**

It is a linear model for binary or multi-class classification is called logistic regression. It calculates the likelihood that an input falls into a specific class, in this example a personality trait.

The performance of more sophisticated models will be compared using this model as a baseline classifier. We will expand logistic regression to handle multi-class classification since personality qualities may be thought of as multi-class labels. It works effectively for figuring out how well basic linear models predict personality traits from textual data.

Individual characteristics such as conscientiousness, extraversion, agreeableness, neuroticism, and openness will be predicted using logistic regression based on textual elements taken from LinkedIn profiles.

- **SVM (Support Vector Machines):**

This is a supervised learning model that finds the hyperplane that best separates data points of different classes. It can be used for both classification and regression tasks. Profiles will be categorized using SVM according to personality attributes. It is an excellent choice for NLP-based tasks like personality prediction because of its exceptional efficacy in high-dimensional. It will assist in assessing how well models function on datasets where it is difficult to distinguish between different features.

Text from LinkedIn profiles will be categorized by the model into one of the five personality types. SVM's kernel trick will assist in managing non-linear distinctions between personality traits.

Root Mean Squared Error (RMSE) was used to evaluate the performance of an SVR with an RBF kernel that had been improved for each attribute. SVR's capacity to extract underlying personality signals from structured professional language was demonstrated by the results, which displayed differing levels of prediction accuracy across attributes. Lastly, by predicting the trait scores of a bespoke sample profile, we showed how useful the trained models are in the actual world. This experiment demonstrates the efficacy of using regression models in conjunction with traditional NLP techniques for psychological profiling from internet traces.

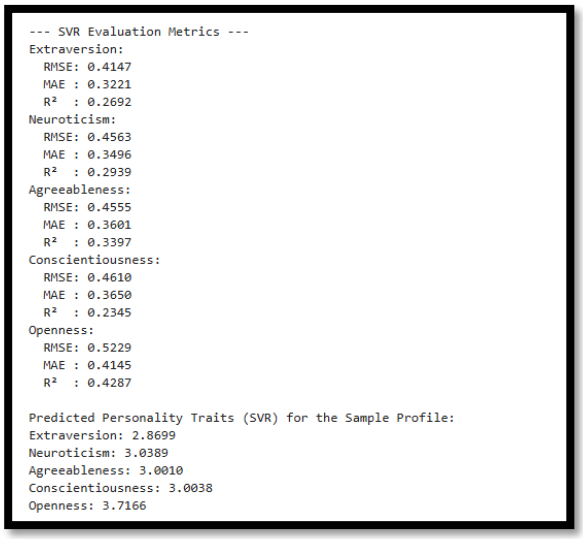

Figure 2: Training Support Vector Regressor

These were the values obtained while training the support vector regressor.

- **Random Forest Regressor:**

Random Forest Regressor is an ensemble learning method that builds multiple decision trees and merges their predictions to produce more accurate and stable results. Each tree in the forest is trained on a random subset of the training data, and at each split in the tree, only a random subset of features is considered. This approach reduces the risk of overfitting and improves generalization by combining the outputs of many diverse models. For regression tasks, the final prediction is typically the average of the predictions from all individual trees. Due to its robustness, ability to handle non-linear relationships, and resistance to overfitting, Random Forest is widely used in machine learning applications involving structured and unstructured data.

In this model, we trained a separate Random Forest Regressor model for each of the Big Five personality traits using structured profile data from LinkedIn. Each model was initialized with n_estimators = 100, meaning 100 decision trees were used in the ensemble. This value strikes a balance between performance and computational efficiency—enough trees to capture complex patterns without excessive training time. We also set a fixed random_state = 42 to ensure reproducibility of results. The data was split into training and testing sets using an 80/20 split via train_test_split, which is a standard practice to evaluate model generalization. The features used for training were generated by applying TF-IDF vectorization on cleaned profile text, limited to the top 5,000 terms (max_features = 5000) to reduce dimensionality and avoid overfitting. By keeping other Random Forest parameters at their defaults (like maximum depth or minimum samples per split), we relied on the algorithm's built-in robustness to manage overfitting and bias-variance trade-offs without manual tuning.

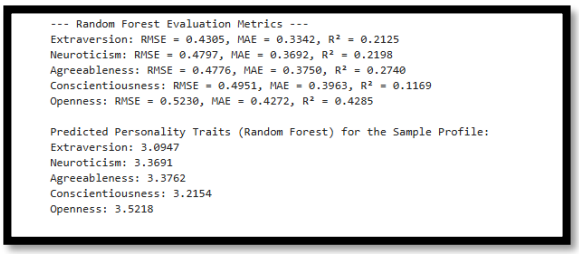

Figure 3:Training Random Forest Regressor

These are the values obtained after training the Random Forest Regressor.

- **BERT:**

We employed the BERT (Bidirectional Encoder Representations from Transformers) model, specifically bert-base-uncased, for personality trait prediction based on LinkedIn profile text. A regression approach was used by adapting the BertForSequenceClassification model to predict continuous scores for the Big Five personality traits.

The model utilizes the **bert-base-uncased** pre-trained weights, with the final classification layer adjusted to output five scores corresponding to the Big Five traits. A maximum token length of **128** was set for input sequences to balance computational efficiency and context retention. We used the **AdamW optimizer** with a **learning rate of 2e-5**, which is commonly recommended for fine-tuning transformer models. The model was trained over **3 epochs** using a **batch size of 8**, chosen to accommodate memory constraints while maintaining learning stability. The loss function used was **Mean Squared Error (MSE)**, suitable for continuous output regression tasks. Evaluation was done using **Root Mean Squared Error (RMSE)** across all five traits. We continuously updated the parameters in order to fine tune the trained model to get the best values possible.

```
Train Loss: 1.8433
Epoch 1/3 completed.
Train Loss: 0.3407
Epoch 2/3 completed.
Train Loss: 0.2569
Epoch 3/3 completed.

Evaluation Metrics for Each Personality Trait:

Extraversion:
  RMSE: 0.4132
  MAE : 0.3157
  R²  : 0.2744

Neuroticism:
  RMSE: 0.4665
  MAE : 0.3497
  R²  : 0.2620

Agreeableness:
  RMSE: 0.4492
  MAE : 0.3425
  R²  : 0.3580

Conscientiousness:
  RMSE: 0.4923
  MAE : 0.3962
  R²  : 0.1267

Openness:
  RMSE: 0.4900
  MAE : 0.3967
  R²  : 0.4983


Predicted Personality Traits for the Sample Profile:
Extraversion: 2.7711
Neuroticism: 3.0802
Agreeableness: 3.0439
Conscientiousness: 2.9000
Openness: 3.6727
```

Figure 4: Training BERT model

These were the outputs for the custom BERT model.

During the training of the BERT model for personality trait prediction, we monitored the Mean Squared Error (MSE) loss across epochs. The initial training loss began at **1.8433**, and showed a significant reduction over time, reaching **0.3407** by the end of the second epoch and further dropping to **0.2569** at the conclusion of the third epoch.

The model was learning efficiently and progressively reducing prediction errors with each epoch, as evidenced by the steady drop in loss. Proper convergence is seen in the downward trend of the loss values, which indicates that the model did not experience underfitting or overfitting during the brief training period.

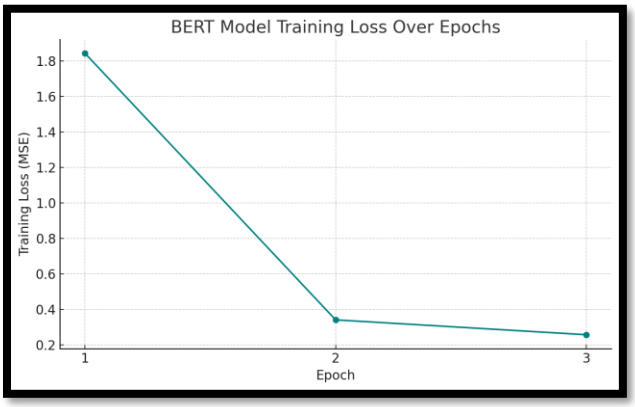

Figure 5: BERT model training loss

RESULTS

Since the prediction of personality traits was approached as a **regression problem**, where the goal was to estimate continuous scores for each of the Big Five traits. We used regression-based evaluation metrics. The following metrics were selected to evaluate model performance:

- **Root Mean Squared Error (RMSE)**
  Measures the square root of the average squared differences between predicted and actual trait scores. Penalizes larger errors more heavily, making it a good metric when large deviations are undesirable. Useful for understanding the typical size of prediction errors.

- **Mean Absolute Error (MAE)**
  Calculates the average absolute differences between predicted and actual values. Offers a more interpretable and balanced error measurement compared to RMSE, especially when outliers are not a major concern. Indicates the average magnitude of errors across all predictions.

- **R-squared Score (R^2)**
  Represents the proportion of variance in the actual trait scores that is explained by the model. Values closer to 1 indicate better predictive power, while values near 0 imply poor explanatory strength. Useful for assessing the overall goodness-of-fit of the model.

Each personality trait was modelled independently. We recorded RMSE, MAE, and R^2 for each trait to comprehensively evaluate performance. Classification-based metrics such as accuracy, precision, recall, F1-score, AUC-ROC, and cosine similarity were

not applicable in this regression setting and were therefore excluded from final evaluation.

Model Performance:

The main evaluation metric for our project is RMSE. So, we compared all the trained models and our BERT model and visualised the RMSE values for each trait.

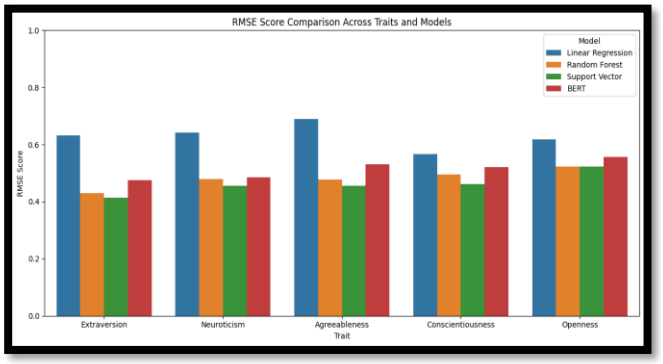

Figure 6: RMSE Comparison

After model training and testing we know that Logistic regression wasn't a good choice for predicting personality traits. While Support Vector and Random Forest gave good results in the training phase but couldn't perform very well in the testing phase. BERT was able to give more accurate results as compared to the other models.

Table 1: Comparison based on RMSE values

Trait	Linear Regression	Random Forest	Support Vector	BERT
Extraversion	0.6326	0.4305	0.4147	0.4762
Neuroticism	0.6410	0.4797	0.4563	0.4844
Agreeableness	0.6900	0.4776	0.4555	0.5308
Conscientiousness	0.5671	0.4951	0.4610	0.5208
Openness	0.6190	0.5230	0.5229	0.5557

This is how the model performed on a LinkedIn profile summary of a user:

Text: "A lifelong seeker of knowledge, I thrive in unstructured environments where creativity and imagination are valued over convention. My journey spans experimental art installations, indie game development, and philosophical blogging. I'm fascinated by abstract concepts, constantly exploring new frameworks in AI ethics, metaphysics, and the intersection of technology and human emotion. I tend to question norms, challenge groupthink, and avoid routine. While I'm not driven by strict deadlines or traditional career ladders, I bring depth, originality, and a vision that's often outside the mainstream."

Output:
Predicted Personality Traits for the provided Sample Profile:

Extraversion: 3.0074
Neuroticism: 2.9990
Agreeableness: 3.0713
Conscientiousness: 3.2820
Openness: 3.7746

CONCLUSION

The purpose of this study, Traitlytics: Analysing Personality from LinkedIn Profile Data, was to examine whether data-driven methods might be used to deduce the Big Five personality traits from LinkedIn profiles. We suggested a methodology that combined TF-IDF vectorization, cosine similarity-based label assignment, and regression-based machine learning models to predict continuous personality trait scores. This approach addressed the shortcomings of traditional self-report methods and capitalized on the richness of professional textual data.

Logistic Regression, SVR, Random Forest Regressor, and an improved BERT model were the four models we tested. Due to their ability to capture non-linearity, SVR and Random Forest both shown improvements in performance, whereas the linear baseline model, as anticipated, fared the worst. The improved BERT model fared better than the others, particularly when it came to predicting qualities like neuroticism and extraversion. This demonstrated how well transformer-based architectures comprehend the subtleties of professional language context. This trend was validated by evaluation using measures such as RMSE, MAE, and R2, although the absolute error rates were not insignificant, indicating the task's underlying complexity.

In the end, our research demonstrates that although professional profiles do include personality-related indicators, it is still difficult to adequately convey the complexity of human characteristics, particularly when relying solely on text. However, by providing a new approach to personality inference on professional platforms, this work adds to the expanding field of digital psychometrics. It establishes the framework for future studies that might include more signals for a more thorough and accurate personality assessment, such as behavioral data, social network structure, or multimodal inputs.

FUTURE WORK

The current study lays a foundation for automated personality prediction from LinkedIn profile data, but several avenues warrant further exploration to enhance the accuracy, robustness, and applicability of the proposed methodologies. Future research could focus on the following key areas:

1. **Data Enrichment and Diversity:** Expanding the dataset to include a larger and more diverse range of LinkedIn profiles is crucial. Investigating methods for collecting data from various sections of the profile beyond summaries and endorsements, such as posts, articles, comments, group memberships, and recommendations, could provide a more holistic view of an individual's professional online persona.

2. **Advanced Feature Engineering**: Moving beyond traditional TF-IDF and basic textual features, future work should delve into more sophisticated feature engineering techniques. This includes extracting and incorporating network-based features more comprehensively, such as the size and structure of a user's network, the nature and frequency of interactions, and the characteristics of their connectionsTemporal analysis of activity patterns over time might also reveal valuable insights into personality stability and change.

3. **Model Innovation and Hybrid Approaches**: Exploring more advanced Large Language Model architectures, potentially those specifically designed for regression tasks or multi-modal inputs, could yield improved performance. Fine-tuning strategies could be further optimized, and the potential of transfer learning from related domains should be investigated. Developing truly hybrid models that seamlessly integrate textual features, network features, and potentially other modalities within a unified framework is a promising direction.

4. **Refined Labelling and Ground Truth:** The reliance on cosine similarity for label assignment, while a practical initial approach, could be a source of error. Future work should explore more sophisticated weak supervision techniques or, ideally, incorporate a phase of expert human annotation on a subset of LinkedIn profiles to establish a more reliable ground truth for training and validation.

5. **Rigorous Evaluation and Interpretability**: Conducting more rigorous evaluation using diverse datasets and employing robust cross-validation techniques is essential. Exploring alternative evaluation metrics that are more aligned with the nuances of personality assessment, potentially including correlations with self-report questionnaires or observer ratings where feasible, would provide a more comprehensive understanding of model performance. Furthermore, enhancing the interpretability of the models using techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) is crucial for building trust and understanding which specific features or linguistic patterns contribute most to the personality predictions.

6. **Ethical Implications and Responsible Deployment**: A thorough examination of the ethical implications of automated personality prediction from professional networking data is paramount for responsible innovation. Future research must address potential biases in the models and data that could lead to unfair or discriminatory outcomes, particularly in high-stakes applications like recruitment. Exploring the potential positive applications of this technology, such as personalized career guidance, team building, and understanding professional dynamics, while actively mitigating risks related to privacy, surveillance, and misuse, should be a central focus.

By addressing these areas, future research can build upon the findings of this study to develop more accurate, reliable, interpretable sound systems for analysing personality from professional online data, unlocking its potential for various beneficial applications while navigating the complex ethical landscape

REFERENCES

1. N. van de Ven, A. Bogaert, A. Serlie, and M. J. Brandt, "Personality perception based on LinkedIn profiles," *Journal of Managerial Psychology*, vol. 32, no. 2, pp. 92–104, 2017. doi: 10.1108/JMP-07-2016-0220.
2. H. Christian, D. Suhartono, A. Chowanda, and K. Z. Zamli, "Text-based personality prediction from multiple social media data sources using pre-trained language model and model averaging," *Journal of Big Data*, vol. 8, no. 1, pp. 1–14, 2021. doi: 10.1186/s40537-021-00459-1.
3. S. Fernandez, M. Stöcklin, L. Terrier, and S. Kim, "Using available signals on LinkedIn for personality assessment," *Journal of Research in Personality*, vol. 93, p. 104122, 2021.
4. K. Dai, C. González Nespereira, A. Fernández Vilas, and R. P. Díaz Redondo, "Scraping and clustering techniques for the characterization of LinkedIn profiles," *ArXiv preprint arXiv:1505.00989*, 2015.
5. V. Kashkin and V. Paliy, "Automated LinkedIn analysis to determine psychometric characteristics of a client," *Asian Social Science*, vol. 20, no. 2, pp. 35–48, 2024.
6. H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, et al., "Personality, gender, and age in the language of social media: The Open-Vocabulary approach," *PLOS ONE*, vol. 8, no. 9, e73791, 2013. doi: 10.1371/journal.pone.0073791.