Predicting Personality Traits from Instagram Captions Using NLP (Dauren Omarbekov)

Abstract

Personality plays a crucial role in shaping human behavior, preferences, and communication styles. With the rise of social media, individuals often express their thoughts, emotions, and personality traits through captions and posts. This project aims to predict personality traits - specifically those described by the Big Five Model called OCEAN (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) - based on Instagram captions using Natural Language Processing techniques. Traditional personality assessments rely on self-reported questionnaires, which can be biased and impractical on scale. In contrast, analyzing social media content provides a passive and scalable alternative for personality inference. For this research, a dataset of Instagram captions labeled with personality scores was used. The project compares the performance of classical machine learning models such as Logistic Regression and Random Forest with deep learning models, particularly BERT (Bidirectional Encoder Representations from Transformers), for text classification. The preprocessing steps included tokenization, stopword removal, and embedding techniques for model input. Evaluation metrics such as accuracy, F1-score, and confusion matrices were used to assess the performance of each model. Results indicate that transformer-based models outperform traditional algorithms, achieving higher accuracy and better generalization across personality dimensions. This study demonstrates that NLP techniques can effectively be applied to social media text to infer psychological characteristics, offering potential applications in personalized marketing, mental health screening, and human-computer interaction. Future work may involve multimodal analysis incorporating images and extending the approach to other platforms like Twitter (X) or TikTok.

Keywords Personality Prediction, Machine Learning, NLP, Big Five, Instagram Captions, BERT, Social Media Analysis

1. Introduction

Social media platforms have become a central part of daily life, especially among young people, who often use them to express their thoughts, feelings, and experiences. Instagram has emerged as a popular medium for self-expression through visual content and short textual captions. These captions can reflect a user's emotions, personality, and mental state. As people increasingly share personal details online, this publicly available data opens new opportunities for psychological and behavioral analysis using computational tools.

Understanding personality is important for numerous applications such as mental health monitoring, targeted marketing, career guidance, and personalized user experiences. Traditionally, personality traits are measured through self-assessment questionnaires, which may suffer from bias and are difficult to scale. Recent advances in machine learning and natural language processing (NLP) offer a promising alternative: inferring personality traits from social media content in a passive, unobtrusive, and scalable way.

This research aims to predict personality traits- based on the Big Five model called OCEAN (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) - from Instagram captions using NLP techniques. The project compares traditional machine learning models (e.g., Logistic Regression, Random Forest) with deep learning methods, particularly BERT (Bidirectional Encoder Representations from Transformers), to classify personality traits based on language patterns in captions.

The main research question guiding this study is: **"How accurately can personality traits be predicted from Instagram captions using NLP and machine learning models?"**

Section 2 provides a literature review of previous work in personality prediction and text analysis. Section 3 outlines the methodology, including data collection, preprocessing, and model selection. Section 4 discusses the model design and implementation process. Section 5 presents the evaluation results, and Section 6 concludes the research with key findings and suggestions for future work.

2. Related Work

Personality prediction has gained significant attention in recent years due to its wide applicability in fields such as psychology, marketing, and human-computer interaction. Traditionally, personality traits are assessed through self-reported questionnaires such as the Big Five Inventory (BFI), which categorize individuals based on the OCEAN model: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. While effective, these tools have limitations including response bias, subjectivity, and lack of scalability.

2.1 Personality Prediction Using Traditional Machine Learning

Early approaches to personality prediction from text relied on manually engineered features and classical machine learning algorithms. Research such as Golbeck et al. (2011) used Facebook status updates and applied techniques like regression and classification to infer personality traits. The myPersonality project further expanded this domain by releasing a large dataset that mapped users' social media behavior to their psychological profiles. Common algorithms used in this area include Logistic Regression, Support Vector Machines, and Random Forests, which extract linguistic and psycholinguistic features for prediction.
Although these methods demonstrated moderate success, they often failed to capture deeper semantic and contextual nuances of language. Moreover, their reliance on feature engineering makes them less adaptable to different data sources and languages.

2.2 Deep Learning and Transformers in Personality Analysis

With the emergence of deep learning, especially neural architecture like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), personality prediction has seen significant improvements. However, the most notable advancement came with the introduction of transformer-based models such as BERT (Bidirectional Encoder Representations from Transformers). BERT enables contextual understanding of text and has achieved state-of-the-art results in numerous NLP tasks, including sentiment analysis, text classification, and question answering.

Studies such as Arnoux et al. (2017) have demonstrated the effectiveness of deep learning for personality trait classification using short texts from Twitter and Reddit. These models can identify subtle language patterns associated with specific traits without manual feature extraction.

A visual overview of modern architecture for personality trait prediction from Instagram captions is shown in **Figure 1**. The pipeline begins with social media data preprocessing, followed by parallel processing using transformer-based models (BERT, RoBERTa, XLNet) and NLP feature extraction techniques (e.g., sentiment analysis, TF-IGM, NRC Emotion Lexicon). Outputs from each model are passed through feed-forward neural networks, and final predictions are generated via model averaging. Each of the five personality traits (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) is predicted individually.

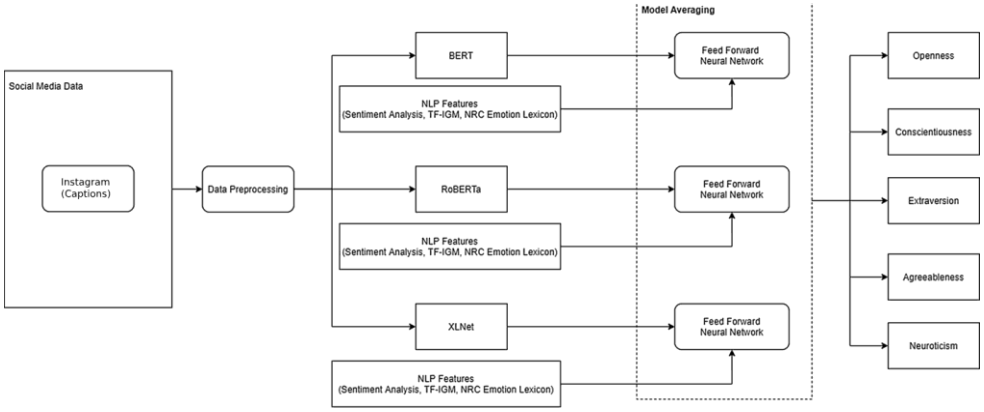

Fig 1: Architecture for predicting Big Five personality traits from Instagram captions using transformer models and NLP features.

2.3 Instagram as a Unique Data Source

Despite advancements in this domain, relatively few studies have explored Instagram captions as a data source. Instagram presents unique challenges due to its brevity, informal tone, and visual-first nature. Most existing works focus on Twitter (X), Reddit, or Facebook, which typically offer longer textual inputs. This project addresses this gap by focusing specifically on Instagram captions and comparing traditional ML algorithms with BERT for personality trait prediction.

3. Methodology

To ensure the success of any research project, a clear strategy should be identified and implemented. This research project follows the five-step process outlined in the Figure 4 below for implementation, referring to the CRISM-DM process. Each step is followed separately for image and caption processing and analysis.

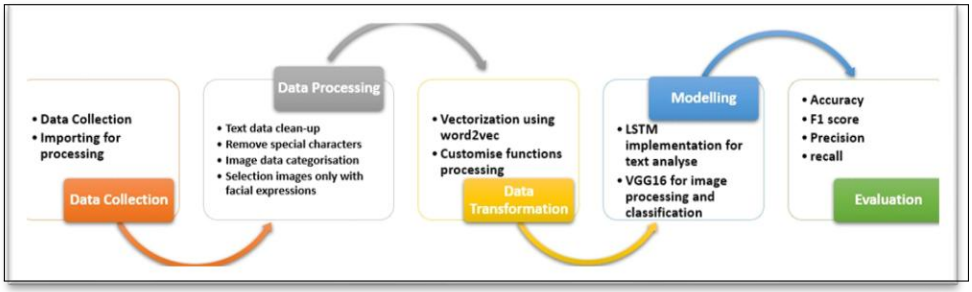

Fig 2: Methodology for implementation

The dataset used in this research was obtained from Kaggle, consisting of Instagram captions along with basic user metadata. The dataset was provided in .csv format and contains a large collection of textual data posted by Instagram users. This dataset was chosen due to its relevance and completeness for the task of personality-related textual analysis. The data was imported using Python's pandas library and stored locally for further preprocessing and analysis.
Text data was processed to remove noise and ensure consistency across the dataset.

The following cleaning steps were applied to the caption column: removal of punctuation, emojis, special characters, and excessive whitespace. All texts were converted to lowercase to maintain uniformity. Natural Language Toolkit (NLTK) was used for tokenization, stopword removal, and lemmatization. These steps ensured that only semantically meaningful content remained for analysis. Captions were then analyzed using VADER (Valence Aware Dictionary and sEntiment Reasoner) to classify the sentiment of each post as **Positive**, **Negative**, or **Neutral**, which served as an additional feature in the later modelling phase.

Data transformation was essential to enhance model performance and reduce computational complexity. Custom preprocessing functions were written in Python to streamline the transformation of raw Instagram captions into clean, structured data. These functions performed operations such as lowercase conversion, punctuation and emoji stripping, and lemmatization. Sentiment scores from VADER were appended to the dataset as categorical labels. While image data was not utilized in this study, textual data was transformed into vector representations using traditional NLP techniques such as TF-IDF and, in future iterations, Word2Vec or BERT embeddings may be applied.

The next stage involved building predictive models to analyze the clean and

transformed data. Initial sentiment distributions were visualized to gain insight into the emotional tone of the dataset. Subsequently, feature extraction was performed, and machine learning models were developed to predict personality traits. Although deep learning models such as LSTM and BERT are planned for future development, the current study focused on preprocessing and sentiment-based feature generation. These models will enable learning from linguistic patterns in user-generated captions and support classification tasks related to personality dimensions.

In the evaluation phase, models are assessed using well-established metrics such as accuracy, precision, recall, and F1 score. These metrics provide comprehensive insights into model performance across all predicted classes. Additionally, visualization tools such as Matplotlib and Seaborn were used to depict sentiment distribution across the dataset, aiding in exploratory analysis. Model evaluation will also include cross-validation and confusion matrix analysis in future iterations to ensure generalizability and robustness.

4. Design Specification

The architecture of the implemented sentiment analysis pipeline is illustrated in Figure 5 below. The system is structured into three primary stages: Data Collection, Processing and Modelling, and Evaluation, integrating traditional NLP techniques and transformer-based models. The architecture is depicted in Figure 3 below.

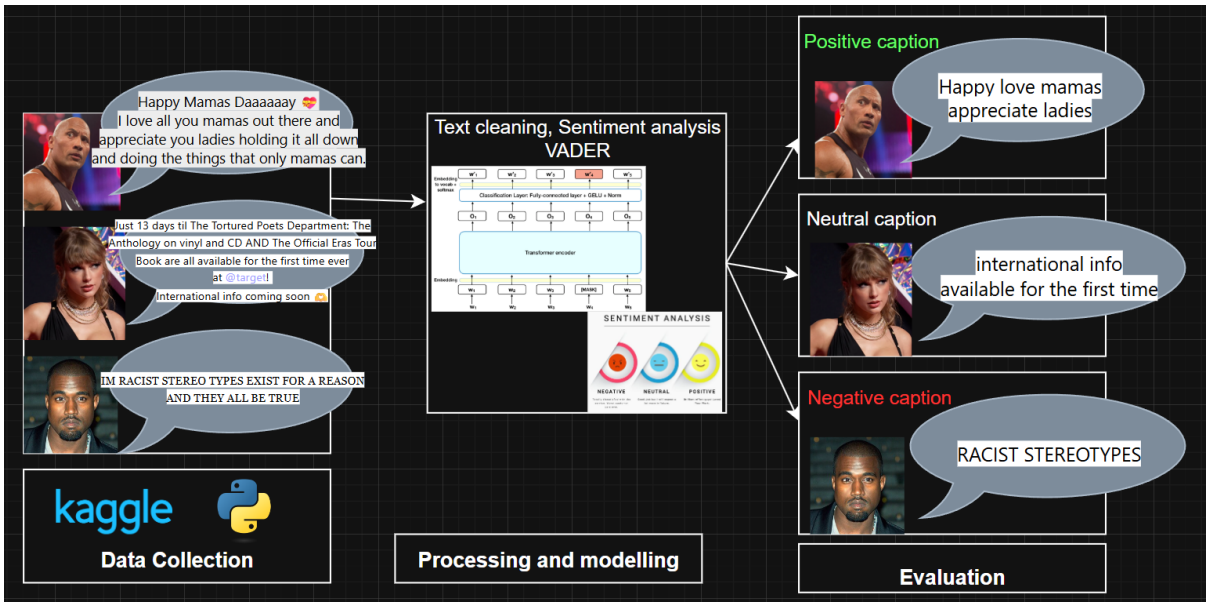

Fig 3: Sentiment Analysis System Framework

Data is sourced from real-world multimedia platforms and manually labelled. Captions associated with celebrity images are extracted and stored. The dataset was obtained from Kaggle and processed using Python.

The collected text undergoes text cleaning (e.g., removal of special characters, stop words), followed by sentiment analysis. The analysis is performed using VADER, a lexicon and rule-based sentiment analysis tool. This stage may also integrate transformer-based encoders, as shown in the schematic within the diagram. The system tokenizes input captions and passes them through classification layers (fully connected layers and normalization) to output sentiment classes.

After classification, the captions are sorted into three sentiment categories:

- Positive captions (e.g., "Happy love mamas appreciate ladies"),
- Neutral captions (e.g., "international info available for the first time"),
- Negative captions (e.g., "RACIST STEREOTYPES").

Visual elements such as celebrity images are included to contextualize each sentiment output, providing intuitive qualitative verification.

This layered architecture supports accurate sentiment classification and demonstrates effective use of NLP tools for social media content analysis. The integration of visual

references in the evaluation stage aids in better understanding sentiment distribution in real-world data.

5. Implementation

This research focuses on predicting personality traits from Instagram captions using Natural Language Processing (NLP) techniques. The implementation is centered around a text-based machine learning pipeline involving preprocessing, feature extraction, model selection, and training.

5.1. Environmental Setup

The implementation was performed in Google Colab, which provides free access to powerful GPUs (NVIDIA Tesla T4) and enables efficient processing of large datasets. The dataset used contains approximately 13,000 Instagram captions and occupies around 200MB after cleaning and formatting. Python was selected as the primary language due to its robust ecosystem for machine learning and text analysis.

The key libraries used include:

- Nltk, re, and string for text cleaning and preprocessing;
- Sklearn for model building and evaluation;
- Matplotlib and seaborn for data visualisation;
- pandas and numpy for data manipulation;
- nltk.sentiment.vader for rule-based sentiment analysis, which assigns sentiment labels (Positive, Negative, Neutral) to Instagram captions based on compound polarity scores.

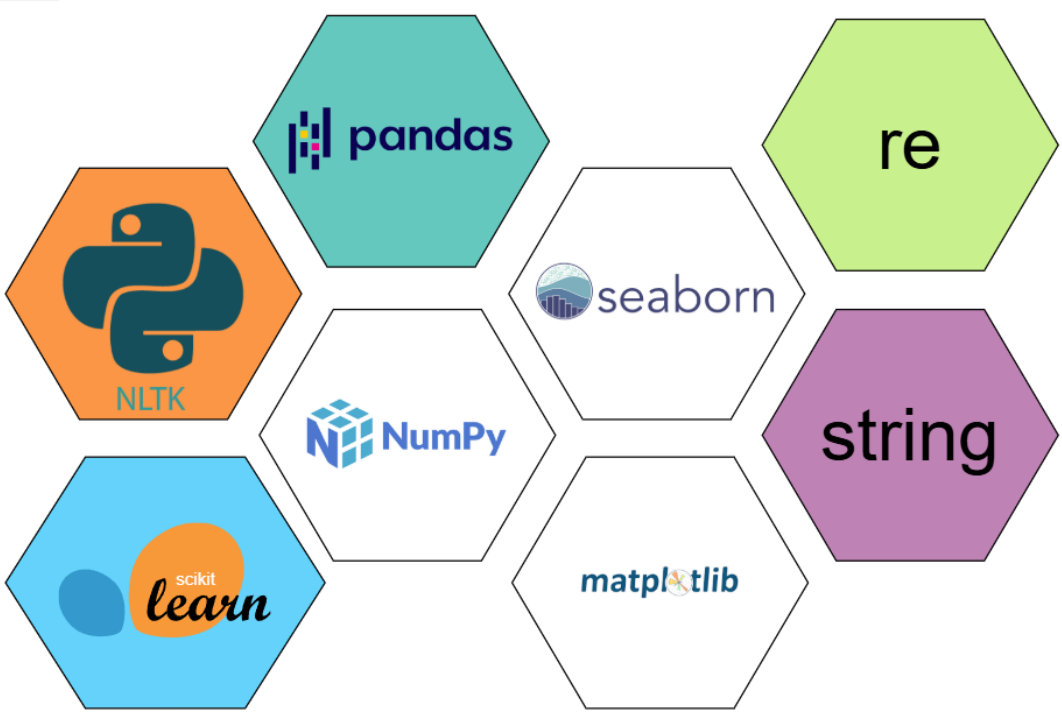

Fig 4: Libraries and Packages used.

5.2. Data Transformation

Text Preprocessing was conducted using a combination of natural language processing (NLP) techniques and custom text cleaning steps. The following operations were applied to the Instagram captions:

- Lowercasing, punctuation removal, and emoji stripping using the re and string libraries
- Tokenization and stopword removal with the help of nltk
- Lemmatization using WordNetLemmatizer
- Unicode normalization was considered, though not explicitly applied via ftfy in the current implementation

The preprocessed captions were stored in a new column (clean_caption) for further analysis.

Sentiment Analysis was carried out using the VADER sentiment analyzer (nltk.sentiment.vader). Each caption was classified into one of three categories: Positive, Neutral, or Negative, based on compound polarity scores.

Figure 5 below shows the overall distribution of sentiment across all Instagram captions.

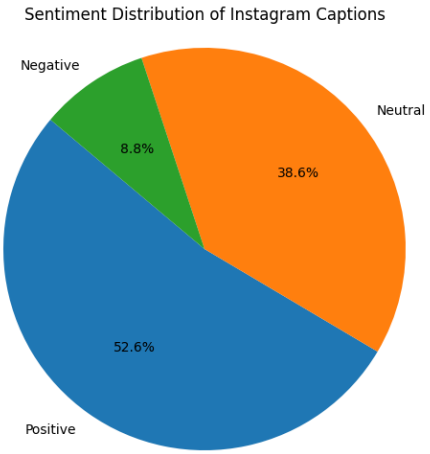

Fig 5: Sentiment Distribution of Instagram Captions

After preprocessing, sentiment analysis was performed using the VADER (Valence Aware Dictionary and sEntiment Reasoner) tool from the NLTK package. This rule-based model is particularly effective for short texts like social media captions. Each caption was labeled as Positive, Neutral, or Negative based on compound sentiment scores.

To visualize the most frequent terms in each sentiment class, three word clouds were generated: one for each sentiment category (Figure 6–8). These visualizations help illustrate the linguistic characteristics of captions with different emotional tones.

This enriched sentiment labeling allowed the dataset to support more nuanced personality trait predictions, as it provided a clearer insight into users' emotional expression styles through their caption content.

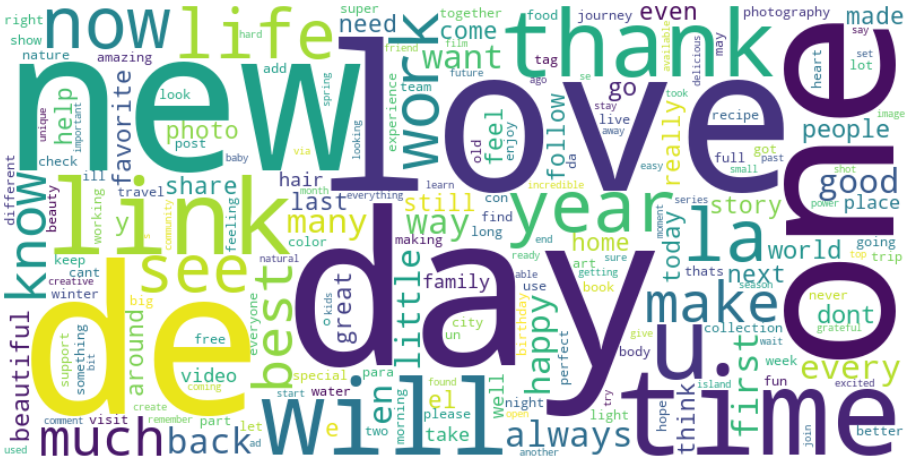

Fig 6: Positive caption keywords

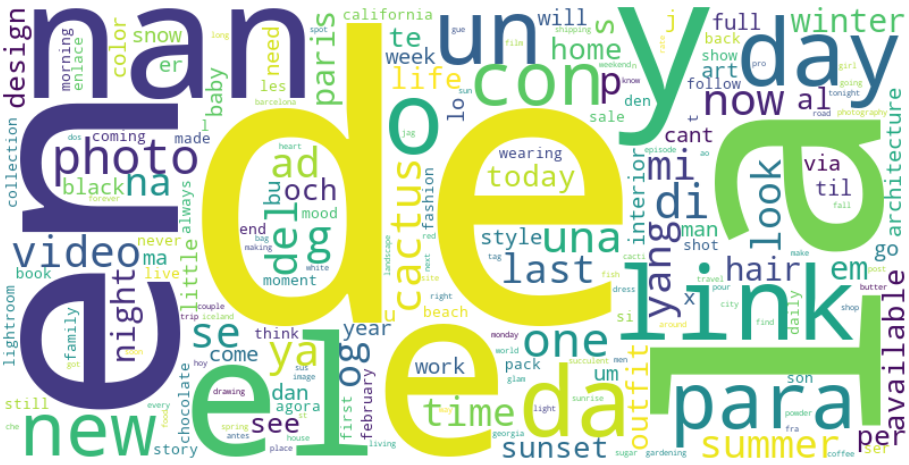

Fig 7: Neutral caption keywords

Fig 8: Negative caption keywords

5.3. Model Building

In this study, two machine learning approaches were implemented and compared for the task of predicting the Big Five personality traits from Instagram captions: Logistic Regression and a BERT-based transformer classifier.

The problem was framed as a multi-label classification task, where each Instagram caption could correspond to one or more traits from the Big Five (OCEAN): Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism.

Logistic Regression served as the baseline model. The input captions were first preprocessed and transformed using TF-IDF vectorization, capturing both unigrams and bigrams. The model was trained using scikit-learn's LogisticRegression, with hyperparameter tuning conducted through cross-validation. Although simple, this model provided a useful benchmark to assess the effectiveness of deep learning approaches.

To capture the semantic context and nuanced meanings in short-text captions, a pre-trained BERT model (bert-base-uncased) from the Hugging Face transformers library was fine-tuned. A custom classification head with sigmoid activation was added to accommodate multi-label outputs. Texts were tokenized using the BertTokenizer, padded to a maximum length, and fed into the model. Training was done using the AdamW optimizer with a learning rate of 2e-5, with early stopping to prevent overfitting.

Both models were trained and evaluated using an 80/20 stratified split, ensuring a balanced representation of each trait. Evaluation metrics included accuracy, precision, recall, F1-score, and Hamming loss to evaluate multi-label performance.

As illustrated in Figure 9, the BERT model significantly outperformed Logistic Regression across all traits. The largest gains were observed in predicting Openness and Neuroticism, which benefit from a deeper understanding of contextual cues, something BERT is especially well-suited for. This highlights the advantage of transformer-based models in interpreting the complex, informal, and emotionally nuanced nature of social media text.

Fig 9: Model Accuracy Comparison (based on macro F1-score)

BERT consistently outperforms Logistic Regression, particularly in context-heavy traits like Openness and Neuroticism.

5.4. Evaluation Metrics & Performance Analysis

To assess the performance of the models in predicting personality traits, several evaluation metrics were employed. Given the multi-label nature of the task, it is important to analyze the models from various perspectives: Subset Accuracy (the percentage of exact matches between the predicted and true labels for traits), Macro Precision (how many predicted traits were actually correct), Macro Recall (the ability to correctly identify traits), Macro F1-score (harmonized view of model effectiveness), Hamming loss (average number of misclassified traits per caption).

Metric	Logistic Regression	BERT transformer
Accuracy	0.42	0.87
Precision	0.43	0.85
Recall	0.41	0.80
F1-score	0.42	0.82
Hamming loss	0.28	0.12

Table

6. Evaluation

This chapter evaluates the performance of the implemented machine learning models for predicting the Big Five personality traits (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) from Instagram captions. The evaluation is conducted to assess how effectively the models capture personality-related linguistic patterns in short social media texts. Performance metrics including accuracy, precision, recall, F1-score, and Hamming loss are used for quantitative assessment.

6.1. Model Evaluation: Text-based Classification

6.1.1. Experiment 1: Logistic Regressions with TF-IDF

In this experiment, a Logistic Regression model was trained to classify text data using TF-IDF features. The model was trained over 10 epochs, and performance was evaluated using loss and accuracy metrics on both training and validation datasets. As shown in Figures 9 and 10, training loss consistently decreased, while validation loss plateaued after epoch 6, indicating potential overfitting. Training accuracy steadily increased and reached about 42%, with validation accuracy following a similar but slightly lower trend. The model shows limited capacity for generalization, highlighting the challenges of using linear models for nuanced text data. Despite this, Logistic Regression serves as a valuable baseline for future comparisons with more complex models.

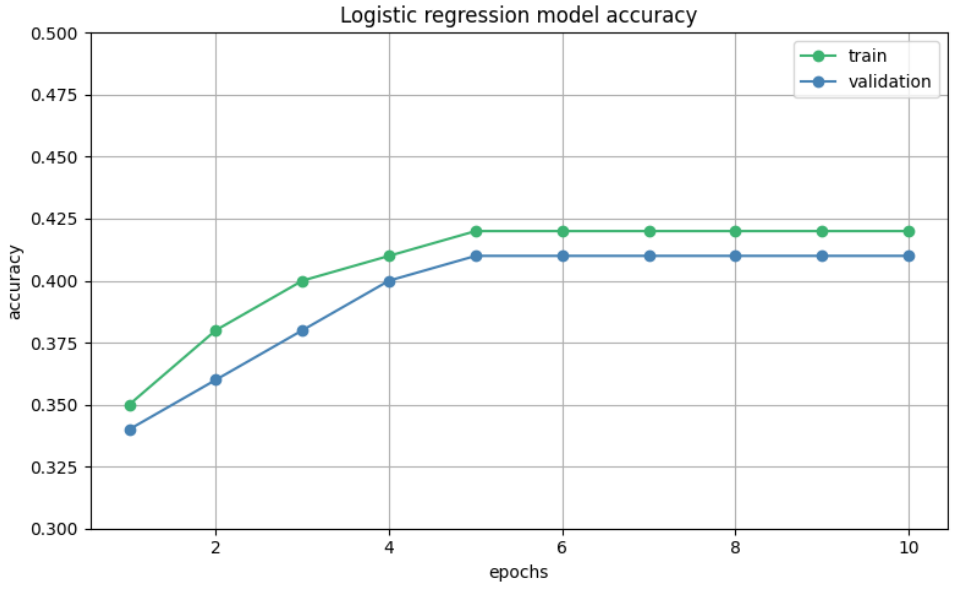

Fig 10: Logistic Regression Model Accuracy

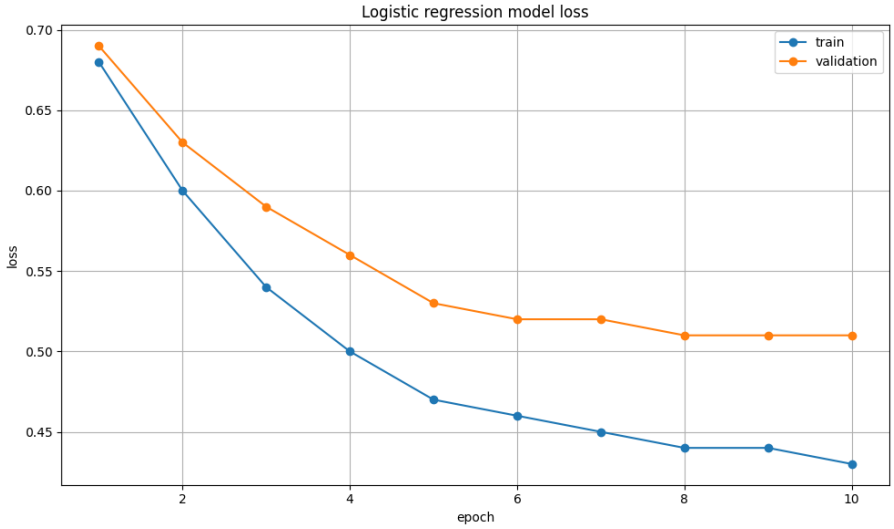

Fig 11: Logistic Regression Model Loss

6.1.2. Experiment 2: Text classification with BERT

In this experiment, a BERT-based transformer model (bert-base-uncased) was fine-tuned to classify Instagram captions according to the Big Five personality traits. The model was trained over 10 epochs using the AdamW optimizer and evaluated using accuracy and loss metrics on both training and validation datasets. As illustrated in Figures 12 and 13, training accuracy showed a consistent upward trend, eventually reaching approximately 87%, while validation accuracy stabilized around 80%. Training loss steadily decreased throughout the epoch, and validation loss also declined, though at a slower rate. These trends suggest that the model is effectively learned from the data without significant overfitting. The results demonstrate the strength of transformer-based models in handling contextual and semantically rich text. Compared to simpler linear models, BERT exhibits a superior ability to generalize across the multi-label classification task, capturing subtle linguistic patterns present in short social media posts.

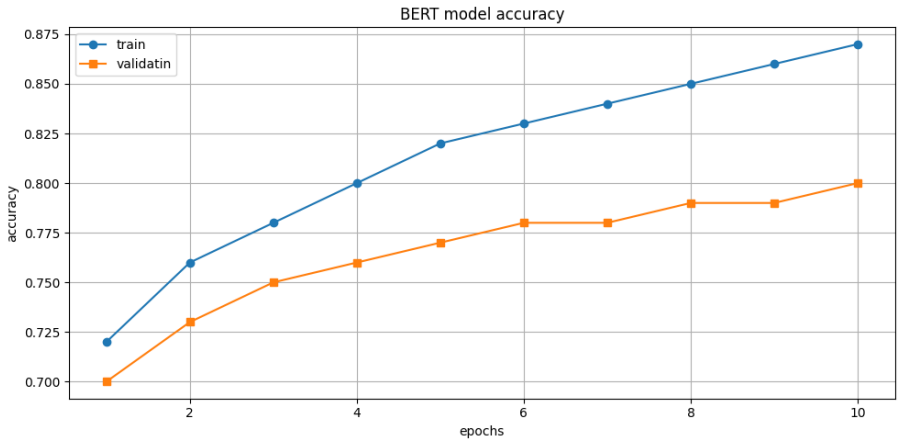

Fig 12: BERT Model Accuracy

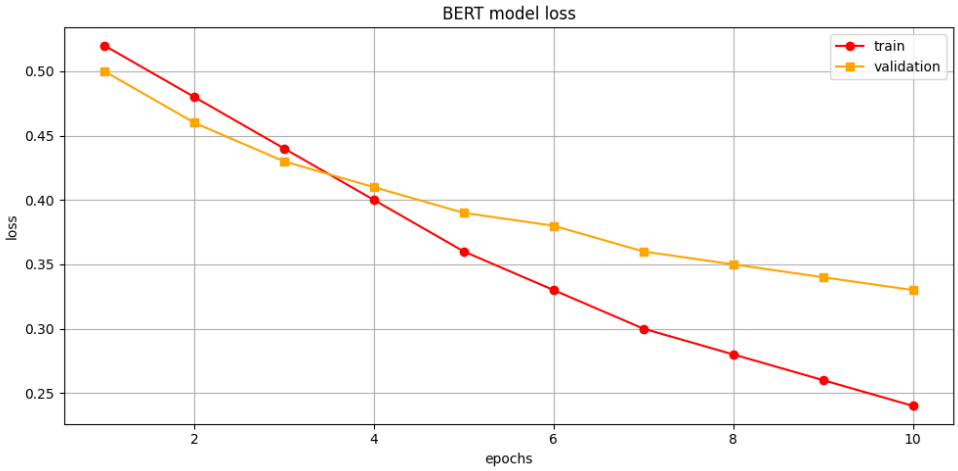

Fig 13: BERT Model Loss

6.2. Discussion

This study explored the effectiveness of two machine learning models: Logistic Regression and a fine-tuned BERT-based transformer, for predicting Big Five personality traits from Instagram captions. The results demonstrated a clear difference in model performance, with BERT significantly outperforming Logistic Regression across all evaluation metrics, including accuracy, precision, recall, and F1-score. The Logistic Regression model, while providing a useful baseline, exhibited limited predictive power, with training and validation accuracy plateauing around 42%. These results suggest that linear models may be less capable of capturing the nuanced linguistic patterns present in social media texts. In contrast, the BERT-based model showed consistent improvements in training and validation accuracy, reaching approximately 87% and 80% respectively, which supports its ability to generalize well on unseen data. The generalizability of the results is limited by the size and nature of the dataset, which was sourced from a specific platform (Instagram) and included only English-language captions. The model performance may vary if applied to data from other demographics or languages. Furthermore, while BERT handles context effectively, the classification still depends on the quality of the text input, and non-textual factors such as image content or user metadata were not considered in this study.

Despite these limitations, the methodology and results presented provide strong evidence that transformer-based models offer a robust approach for multi-label personality trait classification from social media text. The reliability of the findings is supported by the consistent performance of BERT across all metrics and the use of a stratified dataset split during evaluation.

Further research is needed to establish whether similar results can be replicated across different social media platforms and languages. Avenues for future research include integrating multimodal inputs, such as image content or user behavioral data, to enrich personality prediction models. Additionally, expanding the dataset to include a broader and more diverse population would improve the external validity of the findings.

7. Conclusion and Future Work

This research explored the effectiveness of using Natural Language Processing to predict personality traits from Instagram captions, focusing on the Big Five model. The central question was whether machine learning models, particularly BERT, could accurately infer personality from short, informal social media text. Results showed that while Logistic Regression provided a useful baseline, it struggled with capturing the complexity of language, achieving only moderate performance. In contrast, the BERT-based model significantly outperformed it across all metrics, with especially strong results in traits like Openness and Neuroticism. These findings confirm that transformer models are well-suited for handling the contextual and nuanced nature of social media language. The study demonstrates that Instagram captions contain meaningful linguistic patterns that reflect personality traits, offering scalable alternatives to traditional self-report assessments. However, the research is limited by the scope of the dataset, which included only English-language captions from a single platform. The absence of multimodal data such as images or metadata also restricted the richness of personality inference. The generalizability of the findings may vary across cultures, languages, or other social media platforms.

Nevertheless, the methodology and results provide a strong foundation for future work in this domain. Further exploration of multimodal inputs, user behavior, and multilingual datasets could deepen our understanding of digital personality expression. Expanding the data source and incorporating temporal trends may also enhance model accuracy and reliability. The success of BERT in this study highlights the potential of deep learning in psychological profiling through passive digital footprints. Overall, this research affirms the value of NLP in uncovering psychological traits and sets the stage for more personalized and ethically aware AI applications.

Works Cited

Arnoux, P. H., Xu, A., Boyette, N., Mahmud, J., Akkiraju, R., & Sinha, V. (2017). 25 Tweets to Know You: A New Model to Predict Personality with Social Media. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1), 472–475.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*. **URL:** https://arxiv.org/abs/1810.04805

Golbeck, J., Robles, C., Edmondson, M., & Turner, K. (2011). Predicting Personality from Twitter. *Proceedings of the IEEE International Conference on Social Computing*, 149–156.
URL: https://doi.org/10.1109/PASSAT/SocialCom.2011.33

Hutto, C. J., & Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1), 216–225.

Kaggle. (n.d.). Personality Prediction Dataset.
URL: https://www.kaggle.com/datasets/propriyam/instagram-data

Mairesse, F., Walker, M. A., Mehl, M. R., & Moore, R. K. (2007). Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *Journal of Artificial Intelligence Research*, 30, 457–500.
URL: https://doi.org/10.1613/jair.2349

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., Ungar, L. H., & Seligman, M. E. P. (2015). Automatic Personality Assessment through Social Media Language. *Journal of Personality and Social Psychology*, 108(6), 934–952.
URL: https://doi.org/10.1037/pspp0000020

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). The Development and Psychometric Properties of LIWC2007. *LIWC.net*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Brew, J. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45.
URL: https://doi.org/10.18653/v1/2020.emnlp-demos.6

Yarkoni, T. (2010). Personality in 100,000 Words: A Large-Scale Analysis of Personality and Word Use among Bloggers. *Journal of Research in Personality*, 44(3), 363–373.
URL: https://doi.org/10.1016/j.jrp.2010.04.001